Limits on Sparse Support Recovery via Linear Sketching with Random Expander Matrices

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

Abstract

Linear sketching is a powerful tool for the problem of sparse signal recovery, having numerous applications such as compressive sensing, data stream computing, graph sketching, and routing. Motivated by applications where the *positions* of the non-zero entries in a sparse vector are of primary interest, we consider the problem of *support recovery* from a linear sketch taking the form $\mathbf{Y} = \mathbf{X}\beta+\mathbf{Z}$. We focus on a widely-used expander-based construction in the columns of the measurement matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ are random permutations of a sparse binary vector containing $d \ll n$ ones and $n-d$ zeros. We provide a sharp characterization of the number of measurements required for an information-theoretically optimal decoder, thus permitting a precise comparison to the i.i.d. Gaussian construction. Our findings reveal both positive and negative results, showing that the performance nearly matches the Gaussian construction at moderate-to-high noise levels, while being worse by an arbitrarily large factor at low noise levels.

1 Introduction

In recent years, there has been a tremendous amount of research in linear sketching and sparse signal recovery, in which the goal is to recover a sparse vector $\beta \in \mathbb{R}^p$ based on n noisy observations of the form

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a known measurement matrix, and $\mathbf{Z} \in \mathbb{R}^n$ is additive noise. This simple model itself comes with a large number of applications, such as compressive sensing, data stream computing, graph sketching, routing, and wireless communication. Perhaps the most common performance criterion is to characterize a distance between β and an estimate $\hat{\beta}$, such as $\|\beta-\hat{\beta}\|_1$ or $\|\beta-\hat{\beta}\|_2$ [1–3].

In many applications, the primary interest is not in the values of β, but rather in the locations of the non-zero entries: In cognitive radio settings [4], one may seek to determine which channels are in use and which are available; in routing and data stream problems [5], one may be interested in determining a sparse set of IP addresses that have been accessed (or more generally, accessed an unusually large number of times, *cf.*, the *heavy hitters* problem [5,6]). This motivates the study of the *support recovery* problem, where the goal is to recover the non-zero indices $S := \{i : \beta_i \neq 0\}$. As well as being of direct interest, this can lead to other estimation guarantees, such ℓ_2 error [7].

The measurement matrix $\mathbf{X}$ plays a key role in determining both the required number of measurements for successful recovery, and the ability to perform the estimation efficiently. For example, i.i.d. Gaussian matrices tend to provide the best theoretical guarantees [1], but can pose significant limitations in terms of storage and computation. On the other hand, constructions based on structured matrices (Hadamard, Fourier, etc.) permit efficient operations such as matrix multiplications without needing explicit storage, but often at the cost of requiring more measurements.

Constructions based on *expanders* [8–11] have proved to be a promising approach for achieving the best of both worlds. Such constructions are highly sparse and thus permit efficient storage and computation. Moreover, they have been shown to achieve order-optimal results in terms of the ℓ_1-ℓ_1 error guarantee, expressing the difference $\|\hat{\beta}-\beta\|_1$ as a function of the noise norm $\|\mathbf{Z}\|_1$ [8]. However, to our knowledge, precise

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

characterizations of how these compare to the popular i.i.d. Gaussian construction have remained elusive. While the latter are less practical due to their storage and computation requirements, they generally come with the best known theoretical guarantees.

In this paper, we study the performance of expander-based measurements in the context of support recovery, and provide several results that allow for rigorous comparisons to the i.i.d. Gaussian construction. As formalized below, we focus on the most widely-used class of random matrices for constructing expanders, where each column is a random permutation of a sparse binary vector. Such constructions are particularly common in computer science applications. We study both upper and lower bounds on the number of measurements that hold regardless of the computational complexity. A key finding is that when the noise level is not too low, the sample complexity is nearly identical to that of Gaussian constructions, even up to the constant factors. In contrast, we show that Gaussian constructions may require significantly fewer measurements at low noise levels.

Information-theoretic studies of support recovery can be found in [7, 12–17]. The sharp thresholds for the linear model with i.i.d. Gaussian matrices that provide the baseline comparison for the present paper were first given in [15], though our analysis follows a more general approach for possibly non-linear models performed in [17]. The analysis therein heavily relies on the measurement matrix being i.i.d. (not necessarily Gaussian). A key contribution of the present work is the development of *change-of-measure* techniques in which the error probability under a non-i.i.d. distribution is bounded in terms of that of an i.i.d. distribution. We also address a second difficulty, namely, characterizing the concentration of relevant sums of non-independent random variables, as opposed to the i.i.d. sums arising in [17]. Although we do not consider it here, our techniques can directly be applied to an analogous group testing problem [18] to show the asymptotic optimality of our non-i.i.d. construction in the case that the number of defective items is $\Theta(1)$.

Notation We use bold symbols for collections of n scalars or vectors; vectors of other sizes may still have the regular font. We write β_S (respectively, $\mathbf{X}_S$ to denote the subvector of β (respectively, submatrix of $\mathbf{X}$) containing the entries (respectively, columns) indexed by S. We write $\beta_i := \beta_{\{i\}}$ and $\mathbf{X}_i := \mathbf{X}_{\{i\}}$, and let X_{ij} denote element (i,j) of $\mathbf{X}$. The complement with respect to $\{1,\dots,p\}$ is denoted by $(\cdot)^c$. For a given joint distribution P_{XY}, the corresponding marginal distributions are denoted by P_X, P_Y, $P_{Y|X}$, and so on. We use standard notations for the entropy and mutual information (e.g., $H(X)$, $I(X;Y|Z)$), and we define the binary entropy function $H_2(\epsilon) := -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)$. 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.

1.1 Definition and Probabilistic Constructions of Expander Matrices

In the following, we formally define the notion of an expander matrix $\mathbf{X} \in \{0,1\}^{n\times p}$. For a given subset $S \subseteq \{1,\dots,p\}$, we let $\mathcal{N}_{\mathbf{X}}(S) := \{i \in \{1,\dots,n\} : X_{ij} = 1 \text{ for some } j \in S\}$; interpreting $\mathbf{X}$ as representing a bipartite graph from $\{1,\dots,p\}$ to $\{1,\dots,n\}$, we can view $\mathcal{N}_{\mathbf{X}}(S)$ as being the neighborhood set of S.

Definition 1. A matrix $\mathbf{X} \in \{0,1\}^{n\times p}$ containing exactly d ones per column is a *lossless* (k,d,ϵ)*-expander matrix* if, for any $S \subseteq \{1,\dots,p\}$ with $|S| \le k$, we have $|\mathcal{N}_{\mathbf{X}}(S)| \ge (1-\epsilon)d|S|$.

One can think of this definition as stating that the bipartite graph representing $\mathbf{X}$ is "well-connected" in the sense that every "small" subset of the left-vertices has a number of neighbors within a factor of $1-\epsilon$ of the highest number possible.

There has been an extensive amount of work on both existence proofs for expander matrices with given parameters (p,n,k,d,ϵ) via the probabilistic method, and also explicit constructions; see [8, 19] and the references therein. In this paper, we consider a standard probabilistic construction in which each column $\mathbf{X}_i$ of $\mathbf{X}$ is a uniformly random permutation of a fixed vector containing d ones and $n-d$ zeros [8, 20]. We define the ratio $\rho := \frac{d}{n}$, and since we are interested in sparse matrices, we limit our attention to the case that $\rho \le \frac{1}{2}$.

The following result, proved in Appendix A, is obtained by starting with a general bound of Berinde [8] and then bounding it in a manner that is more specific to the scaling regimes considered in this paper.

Lemma 1. *For the preceding construction with* $k = \Theta(1)$, $\frac{d}{n} = \Theta(1)$ *and* $\epsilon = \Theta(1)$, *the random matrix* $\mathbf{X}$ *is a* (k,d,ϵ)*-expander matrix with probability approaching one as* $p \to \infty$ *provided that*

$$\epsilon \log\frac{n}{dk} - H_2(\epsilon) > \frac{\log p}{d}(1+\eta) \tag{2}$$

for some $\eta > 0$.

This result will be used in our numerical evaluations in Section 3 to check when the random matrices that we consider indeed form expander matrices. Note that our results will be concerned with the scaling $n = \Theta(k\log p) = \Theta(\log p)$; since we are also considering $\frac{d}{n} = \Theta(1)$, this means that $d = \Theta(\log p)$.

We briefly mention that under the scaling laws in Lemma 1, the storage of $\mathbf{X}$ requires roughly $npH_2(\rho)$ bits; for small values of ρ, this is far below the np real numbers required to store an i.i.d. Gaussian matrix.

1.2 Support Recovery: Problem Statement

The support set S is assumed to be equiprobable on $\mathcal{S}$, defined to contain the $\binom{p}{k}$ subsets of $\{1,\dots,p\}$ with cardinality k. 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 the expander-based distribution given in the previous subsection. We write $\mathbf{X}_i$ to denote the random i-th column; generically denoting one such column by $\mathbf{X}_0$, the corresponding distribution can be written as

$$P_{\mathbf{X}_0}(\mathbf{x}_0) = \frac{1}{\mu_n}\prod_{i=1}^{n} P_X(x_{0,i})\mathbb{1}\{N_1(\mathbf{x}_0)=d\}, \tag{3}$$

where $P_X \sim \text{Bernoulli}\big(\frac{d}{n}\big)$, $N_1(\mathbf{x}_0)$ is the number of ones in $\mathbf{x}_0$, and μ_n is a normalizing constant. Since all binary vectors with d ones are equiprobable under this distribution, it simply amounts to a random permutation, as desired. We let $P^{\ell}_{\mathbf{X}_0}$ denote the ℓ-fold product of $P_{\mathbf{X}_0}$, so that

$$\mathbf{X} \sim P^{p}_{\mathbf{X}_0}. \tag{4}$$

The i-th row of $\mathbf{X}$ is denoted by $X^{(i)}$.

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

$$Y^{(i)} = \langle X^{(i)}, \beta\rangle + Z^{(i)}, \tag{5}$$

where $Z^{(i)} \sim N(0,\sigma^2)$ is additive Gaussian noise. The corresponding conditional probability density function (PDF) is denoted by $P_{Y|X_S\beta_S}$, and we let $P^n_{Y|X_S\beta_S}(\cdot|\cdot,b_s)$ denote the n-fold product of $P_{Y|X_S\beta_S}(\cdot|\cdot,b_s)$. Note that here and subsequently, $P_{Y|X_S\beta_S}$ denotes the Gaussian conditional PDF, whereas $P_{\mathbf{X}_0}$ is a probability mass function (PMF). We allow β_S to be discrete or continuous, meaning P_{β_S} may be a PMF or PDF.

Given $\mathbf{X}$ and $\mathbf{Y}$, a *decoder* forms an estimate $\hat{S}$ of S. We assume that the decoder knows the system model (including k, P_{β_S} and $P_{Y|X_S\beta_S}$); that is, it knows the relevant probability distributions, but not the specific realizations. The error probability is given by

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

and is taken over the realizations of S, β, $\mathbf{X}$, and $\mathbf{Y}$ (the decoder is assumed to be deterministic).

1.3 Joint Distributions and Properties

It will prove convenient to work with random variables that are implicitly conditioned on $S=s=\{1,\dots,k\}$. 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 $S=s$, and we define

$$P_{\beta_s\mathbf{X}_s\mathbf{Y}}(b_s,\mathbf{x}_s,\mathbf{y}) := P_{\beta_s}(b_s)P^k_{\mathbf{X}_0}(\mathbf{x}_s)P^n_{Y|X_s\beta_s}(\mathbf{y}|\mathbf{x}_s,b_s). \tag{7}$$

It will also prove useful to also introduce the following counterpart for matrices that are i.i.d. on $P_X \sim \text{Bernoulli}\big(\frac{d}{n}\big)$; letting $P_X^{n\times p}$ denote the corresponding i.i.d. matrix distribution, we have

$$\widetilde{P}_{\beta_s\mathbf{X}_s\mathbf{Y}}(b_s,\mathbf{x}_s,\mathbf{y}) := P_{\beta_s}(b_s)P_X^{n\times k}(\mathbf{x}_s)P^n_{Y|X_s\beta_s}(\mathbf{y}|\mathbf{x}_s,b_s). \tag{8}$$

It is easy to show (e.g., see Appendix D) that any given symbol in a vector with distribution $P_{\mathbf{X}_0}$ (*cf.*, (3)) has distribution P_X. It follows that the marginal distribution corresponding to a single measurement is the same in both (7) and (8), and is given by

$$P_{\beta_sX_sY}(b_s,x_s,y) := P_{\beta_s}(b_s)P^k_X(x_s)P_{Y|X_s\beta_s}(y|x_s,b_s). \tag{9}$$

In accordance with the above definitions, we make use of the following random variables:

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

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

A key idea that we will use in our analysis is a *change of measure* from the expander-based column distribution $P_{\mathbf{X}_0}$ to the i.i.d. distribution P^n_X. Formally, we have the following.

Lemma 2. *For any choices of n and d, we have for all sequences $\mathbf{x}_0\in\{0,1\}^n$ that*

$$P_{\mathbf{X}_0}(\mathbf{x}_0) \le (n+1)P^n_X(\mathbf{x}_0). \tag{12}$$

Proof. This result will follow upon proving that $\mu_n \ge \frac{1}{n+1}$ in (3). By definition, μ_n is the probability that $N_1(\mathbf{X}_0)=d$ under $\mathbf{X}_0\sim\prod_{i=1}^n P_X(x_{0,i})$. Since $N_1(\mathbf{x}_0)$ can only take one of $n+1$ values, the inequality $\mu_n\ge\frac{1}{n+1}$ follows by noting that $N_1(\mathbf{X}_0)\sim\text{Binomial}\big(n,\frac{d}{n}\big)$, and hence its most probable value is d. □

Changes of measure of this type hold in greater generality beyond permutations of binary vectors [21, Ch. 2], and also beyond discrete measures (e.g., one can similarly change measure from uniform on a sphere to i.i.d. Gaussian [22]). Analyses based on such arguments are common in information-theoretic studies of

channel coding, but we are not aware of any previous works applying them to support recovery problems.

As in [16,17,23], we consider partitions of the support set $s \in \mathcal{S}$ into two sets $s_{\rm dif} \neq \emptyset$ and $s_{\rm eq}$; these can be thought of as corresponding to $s \backslash \bar{s}$ and $s \cap \bar{s}$ for some incorrect support set $\bar{s}$. For fixed $s \in \mathcal{S}$ and a corresponding pair $(s_{\rm dif}, s_{\rm eq})$, we introduce the notation

$$P_{Y|X_{s_{\rm dif}} X_{s_{\rm eq}} \beta_s}(y|x_{s_{\rm dif}}, x_{s_{\rm eq}}, b_s) := P_{Y|X_s \beta_s}(y|x_s, b_s) \tag{13}$$

$$P_{Y|X_{s_{\rm eq}} \beta_s}(y|x_{s_{\rm eq}}, b_s) := \sum_{x_{s_{\rm dif}}} P_X^{\ell}(x_{s_{\rm dif}}) \times P_{Y|X_{s_{\rm dif}} X_{s_{\rm eq}} \beta_s}(y|x_{s_{\rm dif}}, x_{s_{\rm eq}}, b_s), \tag{14}$$

where $\ell := |s_{\rm dif}|$. The key quantities in our bounds are the following conditional mutual informations computed with respect to (10), (13) and (14):

$$I_{s_{\rm dif}, s_{\rm eq}}(b_s) := I(X_{s_{\rm dif}}; Y | X_{s_{\rm eq}}, \beta_s = b_s) \tag{15}$$

$$= I\bigg(\sum_{i \in s_{\rm dif}} b_i X_i ; \sum_{i \in s_{\rm dif}} b_i X_i + Z \,\Big|\, \beta_s = b_s \bigg), \tag{16}$$

where (16) follows from (5) by noting that conditioning on $X_{s_{\rm eq}}$ amounts to subtracting it from the signal.

2 Main Results

We provide two main results; the first considers a discrete distribution on the non-zero entries β_S, and the second considers a continuous distribution.

2.1 Discrete Non-zero Entries

We first consider the distribution of β_s used in [15] and [17, Sec. IV-B], where β_s is a uniformly random permutation of a fixed vector $b_s = (b_1, \dots, b_k)$. We assume that k, $(b_1, \dots, b_k)$, ρ and σ^2 are fixed, i.e., they do not scale with p.

Theorem 1. *Under the preceding setup with fixed values of k, $b_s = (b_1, \dots, b_k)$, and σ^2, and using the random measurement matrix (4) with $\rho := \frac{d}{n} = \Theta(1)$, there exists a decoder such that $P_{\rm e} \to 0$ as $p \to \infty$, provided that*

$$n \geq \max_{(s_{\rm dif}, s_{\rm eq}) \,:\, s_{\rm dif} \neq \emptyset} \frac{|s_{\rm dif}| \log p}{I_{s_{\rm dif}, s_{\rm eq}}(b_s)} (1+\eta) \quad \text{(Achievability)}. \tag{17}$$

Conversely, any decoder satisfies $P_{\rm e} \to 1$ as $p \to \infty$ whenever

$$n \leq \max_{(s_{\rm dif}, s_{\rm eq}) \,:\, s_{\rm dif} \neq \emptyset} \frac{|s_{\rm dif}| \log p}{I_{s_{\rm dif}, s_{\rm eq}}(b_s)} (1-\eta) \quad \text{(Converse)} \tag{18}$$

for some $\eta > 0$.

Proof. See Section 4. □

Remark 1. Under i.i.d. Gaussian measurement matrices with $N(0, \sigma_X^2)$ entries, analogous bounds are known to hold with $I_{s_{\rm dif}, s_{\rm eq}}(b_s) = \frac{1}{2} \log\big(1 + \frac{\sigma_X^2}{\sigma^2} \sum_{i \in s_{\rm dif}} b_i^2\big)$ [15, 17]. In order to equalize the signal power in the two cases, one should set $\sigma_X^2 = \rho$, the non-centralized second moment arising in the expander-based distribution. In Section 2.2 we analytically compare these results in the limits of small and large noise levels, and in Section 3 we numerically compare the bounds for a wide range of finite noise levels.

We emphasize that the converse result not only gives conditions under which $P_{\rm e} \not\to 0$, but also under which $P_{\rm e} \to 1$. The latter is stronger, not only stating the "some" realizations of the random matrix $\mathbf{X}$ "sometimes" lead to errors, but instead that "almost all" realizations "almost always" lead to errors.

The mutual informations in our bounds can easily be computed for fixed $(s_{\rm dif}, s_{\rm eq})$ by evaluating the relevant integrals corresponding to differential entropies. While the optimization over $(s_{\rm dif}, s_{\rm eq})$ may be difficult for large k, this can be potentially simplified by further bounding the mutual information (e.g., see [24]). Moreover, a valid converse can be obtained without performing the full optimization, and we found numerically that the maximum is usually achieved by one of the extreme cases $|s_{\rm dif}| = 1$ or $|s_{\rm dif}| = k$.

2.2 Low-SNR and High-SNR Asymptotics

Theorem 1 provides an exact threshold on the number of measurements in terms of the vector b_s, noise level σ^2 and parameter $\rho := \frac{d}{n}$. The key quantities are the mutual informations $I_{s_{\rm dif}, s_{\rm eq}}(b_s)$, which we can study using the form given in (16). The threshold in (17)–(18) in fact has the same form as that for i.i.d. Gaussian matrices [15, 17], except that the mutual information is computed with respect to a different distribution P_X. Since we have assumed Gaussian noise, we obtain from a well-known saddlepoint property of the mutual information [25, Ex. 9.21] that the mutual informations are strictly higher for Gaussian measurements, and thus fewer measurements are required.

We proceed by showing that the corresponding gap is "small" at low signal-to-noise ratios (SNRs), and "large" at high SNRs; see Section 3 for numerical evaluations supporting this. Here the SNR is the ratio of the signal power to the noise power, given by $\mathrm{SNR} := \frac{k\rho}{\sigma^2}$.

First consider the low-SNR regime, where $(b_1, \dots, b_k)$ and ρ are fixed but $\sigma^2 \ll 1$. Using the formula $I(X; X+Z) = \frac{\mathrm{Var}[X]}{2\sigma^2} + O\big(\frac{1}{\sigma^4}\big)$ for the low-SNR asymptotics of a fixed random variable X (e.g., see [26,

Eq. (50)]; we have an extra factor of $\frac{1}{2}$ since we work in $\mathbb{R}$ rather than $\mathbb{C}$), we obtain

$$I_{s_{\rm dif},s_{\rm eq}}(b_s) = \frac{1}{2\sigma^2}\rho(1-\rho)\sum_{i\in s_{\rm dif}} b_i^2 + O\Big(\frac{1}{\sigma^4}\Big), \quad (19)$$

since the variance of Bernoulli(ρ) is $\rho(1-\rho)$. Similarly, when the measurement matrix is i.i.d. on $P_X \sim N(0,\rho)$ (see Remark 1), we obtain (19) with $\rho(1-\rho)$ replaced by ρ. Thus, *for the same average SNR, the expander distribution only requires a factor of* $\frac{1}{1-\rho}$ *more measurements asymptotically.* This is typically very close to one, since we want ρ small to ensure that $\mathbf{X}$ is sparse. In fact, this gap can be removed altogether by letting the non-zero entries of each column be ± 1 with equal probability, rather than always being one.

In contrast, the behaviors of the two ensembles differ significantly at high SNRs. The easiest way to see this is by trivially upper bounding (15) by the entropy: $I_{s_{\rm dif},s_{\rm eq}}(b_s) \le H\big(\sum_{i\in s_{\rm dif}} b_i X_i\big)$, is uniformly bounded with respect to the SNR since k is fixed.[1] In contrast, using the Gaussian mutual information formula $I(X;X+Z) = \frac{1}{2}\log\big(1+\text{SNR}\big)$, we see that with Gaussian measurements each term $I_{s_{\rm dif},s_{\rm eq}}(b_s)$ grows logarithmically in the SNR, and can be arbitrarily large.

It was shown in [27] that under Gaussian measurements and the polynomial-time LASSO decoder, the required number of measurements tends to $2k\log p$ in the limit of high SNR. Thus, unlike the optimal decoder, the performance saturates for the LASSO when Gaussian measurements are used. We found numerically that this saturation point may be above or below that of the expander ensemble with optimal decoding, depending on the choice of ρ. Specifically, the LASSO with Gaussian measurements may outperform the optimal decoder with expander-based measurements if ρ is small, whereas the opposite may be true if ρ is large. It remains an open question as to how well the LASSO performs with expander-based measurements.

2.3 Continuous Non-zero Entries

We now turn to the case that β_s is i.i.d. on $N(0,\sigma_\beta^2)$ for some variance σ_β^2. As noted in [15], one cannot expect to achieve $P_{\rm e} \to 0$ in this case, since the non-zero entries can be arbitrarily small. Nevertheless, we can characterize the required number of measurements to achieve $P_{\rm e} \to \delta > 0$, as the following result shows.

Theorem 2. *Under the preceding setup for the linear model with fixed values of k and σ^2, a distribution P_{β_s} i.i.d. on $N(0,\sigma_\beta^2)$, and the random measurement matrix in (4) with $\rho := \frac{d}{n} = \Theta(1)$, the optimal decoder (i.e. the decoder minimizing $P_{\rm e}$ in the absence of computational limitations) satisfies*

$$P_e = \mathbb{P}\Bigg[n \le \max_{(s_{\rm dif},s_{\rm eq})\,:\,s_{\rm dif}\neq\emptyset} \frac{|s_{\rm dif}|\log p}{I_{s_{\rm dif},s_{\rm eq}}(\beta_s)}\Bigg] + o(1). \quad (20)$$

Proof. The proof is similar to that of Theorem 1; the differences are outlined in Appendix C. □

Observe that the threshold in the probability in (20) coincides with that in (17)–(18). Thus, this result provides a natural counterpart to Theorem 1 for the case that the values of the non-zero entries are not fixed in advance. Similarly to Remark 1, an analogous bound holds in the case of i.i.d. Gaussian measurements upon a suitable modification of the mutual information terms $I_{s_{\rm dif},s_{\rm eq}}(\cdot)$.

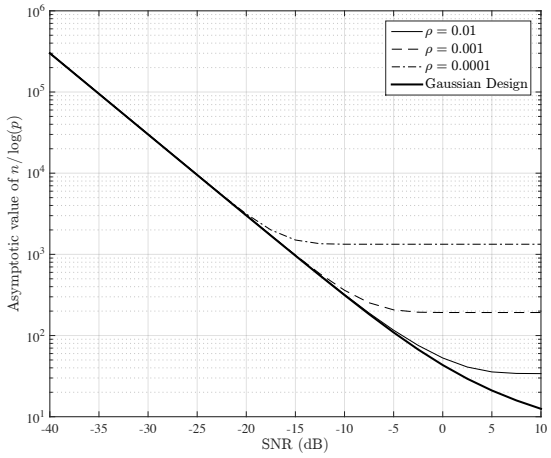

Figure 1: Asymptotic number of measurements for various $\rho = \frac{d}{n}$, and for the i.i.d. Gaussian construction.

3 Numerical Evaluations and Comparisons

In this section, we numerically compare the exact thresholds in Theorem 1 with those of i.i.d. Gaussian measurements with the same power (having the same form but with $I_{s_{\rm dif},s_{\rm eq}}(b_s) = \frac{1}{2}\log\big(1+\frac{\rho}{\sigma^2}\sum_{i\in s_{\rm dif}} b_i^2\big)$ [15, 17]). Note that we focus on the large-p asymptotics, replacing the arbitrarily small constant η in (17)–(18) by zero. For simplicity, we set $k = 15$ and $b_s = (b_0,\dots,b_0)$ (i.e., all of the non-zero entries are equal); we observed similar behavior when this was replaced by a vector with half $+b_0$'s and half $-b_0$'s, and when different values of k were used.

Figure 1 plots the asymptotic thresholds on the number of measurements for various values of ρ, as a func-

[1] For example, an upper bound independent of the SNR is obtained by upper bounding the entropy by the logarithm of the number of possible values.

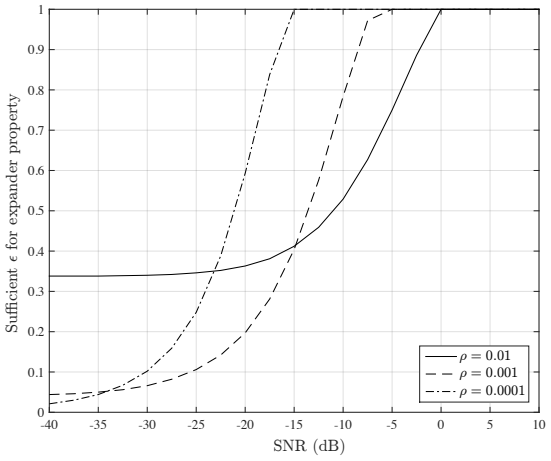

Figure 2: Infimum of ϵ such that (2) holds for some $\eta > 0$ when the number of measurements coincides with the threshold in Theorem 1.

tion of the per-sample SNR in dB:

$$\mathrm{SNR}_{\mathrm{dB}} := 10 \log \frac{k\rho}{\sigma^2}. \tag{21}$$

Note that the Gaussian ensemble depends only on the SNR and not on ρ, and hence there is only one curve for this ensemble, despite there being three for the expander-based ensemble.

As predicted by our findings in Section 2.2, the gap to the performance of Gaussian measurements is insignificant at low SNRs, but grows at higher SNRs. In fact, each of the curves exhibits a rapid transition from the regime where the performance nearly matches the Gaussian curve to that where the number of measurements saturates. The main effect of increasing ρ is in lowering the final saturated value, whereas the behaviors at low SNR are all similar. In fact, although it is not visible at the scale shown, smaller values of ρ are slightly more favorable at low SNR, yielding a smaller multiplicative factor $\frac{1}{1-\rho}$ (see Section 2.2).

Next, we check in which cases the parameters yield expander matrices with high probability. Specifically, Figure 2 plots the infima of ϵ such that the sufficient condition (2) holds for some $\eta > 0$, under the values of the SNR and number of measurements shown in Figure 1. In all cases, the expander property holds with small-to-moderate values of ϵ when the SNR is not too high. In contrast, at high SNRs, the value of ϵ approaches one, meaning that the expander property fails to hold. Interestingly, there appears to be a close correspondence between the expansion property holding and the support recovery performance being similar to the Gaussian design. Note, however, that we are only plotting a sufficient condition in Figure 1, so smaller values of ϵ may be possible.

4 Proofs

Here we provide the main steps of the proof of Theorem 1, deferring several technical details to the appendices in the supplementary material.

4.1 Preliminary Definitions

Along with the definitions in Section 1.3, we introduce the joint distributions

$$P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}}) := P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{y}|\mathbf{x}_s) \tag{22}$$

$$P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}) := \sum_{\mathbf{x}_{s_{\mathrm{dif}}}} P^{\ell}_{\mathbf{X}_0}(\mathbf{x}_{s_{\mathrm{dif}}}) P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}}) \tag{23}$$

$$\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}) := \sum_{\mathbf{x}_{s_{\mathrm{dif}}}} P_X^{n\times\ell}(\mathbf{x}_{s_{\mathrm{dif}}}) P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}}), \tag{24}$$

where $P_{\mathbf{Y}|\mathbf{X}_s}$ is the marginal distribution of (7). Moreover, we define

$$\tilde{\imath}(\mathbf{x}_{s_{\mathrm{dif}}}; \mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}) := \log \frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}})}, \tag{25}$$

and similarly

$$i^n(\mathbf{x}_{s_{\mathrm{dif}}}; \mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}, b_s) := \sum_{i=1}^{n} \imath(x^{(i)}_{s_{\mathrm{dif}}}; y^{(i)}|x^{(i)}_{s_{\mathrm{eq}}}, b_s) \tag{26}$$

$$\imath(x_{s_{\mathrm{dif}}}; y|x_{s_{\mathrm{eq}}}, b_s) := \log \frac{P_{Y|X_{s_{\mathrm{dif}}}X_{s_{\mathrm{eq}}}\beta_s}(y|x_{s_{\mathrm{dif}}}, x_{s_{\mathrm{eq}}}, b_s)}{P_{Y|X_{s_{\mathrm{eq}}}\beta_s}(y|x_{s_{\mathrm{eq}}}, b_s)}. \tag{27}$$

Quantities of this form are often referred to as *information densities* [17, 22]. We note, however, that the denominator in (25) contains the i.i.d.-based distribution $\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}$, as opposed to the true marginal distribution $P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}$. The idea is that, when equipped with the change-of-measure result in Lemma 2, the former is easier to analyze, since it permits a reduction to a summation over the measurements as in (26). Such "modified" information densities have also appeared in the channel coding literature (e.g., see [28,29]) for handling non-i.i.d. random codebook constructions.

Observe that averaging (27) with respect to the random variables in (10) conditioned on $\beta_s = b_s$ yields the conditional mutual information in (15).

4.2 Non-asymptotic Bounds

We begin by providing non-asymptotic upper and lower bounds on the error probability in terms of tail probabilities of the above information densities.

Theorem 3. *For any constants $\delta_1 > 0$ and γ, there exists a decoder such that*

$$P_{\mathrm{e}} \le \mathbb{P}\Bigg[\bigcup_{\substack{(s_{\mathrm{dif}}, s_{\mathrm{eq}}) \\ s_{\mathrm{dif}} \neq \emptyset}} \Bigg\{ \imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}}, \beta_s) \le \log \binom{p-k}{|s_{\mathrm{dif}}|} + 2\log\left(\frac{k}{\delta_1}\binom{k}{|s_{\mathrm{dif}}|}(n+1)^{|s_{\mathrm{dif}}|}\right) + \gamma \Bigg\} \Bigg] + P_0(\gamma) + 2\delta_1, \tag{28}$$

where

$$P_0(\gamma) := \mathbb{P}\left[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s, \beta_s}(\mathbf{Y}|\mathbf{X}_s, \beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} > \gamma \right]. \tag{29}$$

Proof. The proof follows the approach of [17, Thm. 1], but uses the modified information density in (25), and performs additional steps involving change of measure from the random permutation distribution to an i.i.d. distribution. See Appendix B for details. □

Theorem 4. *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 $s_{\mathrm{dif}} \neq \emptyset$) depending on $b_s \in \mathbb{R}^k$. For any decoder, we have*

$$P_{\mathrm{e}} \ge \mathbb{P}\Bigg[\imath^n(\mathbf{X}_{s_{\mathrm{dif}}(\beta_s)}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}(\beta_s)}, \beta_s) \le \log \binom{p-k+|s_{\mathrm{dif}}(\beta_s)|}{|s_{\mathrm{dif}}(\beta_s)|} + \log \delta_1 \Bigg] - \delta_1. \tag{30}$$

Proof. The proof is identical to that of [17, Thm. 2]; while an i.i.d. measurement matrix distribution was used therein, an inspection of the proof reveals that having i.i.d. columns is enough. □

While the preceding theorems resemble their counterparts for i.i.d. measurement matrices [17], there are two notable differences. First, $\imath^n$ is no longer an i.i.d. summation in the present setting, and it is thus more difficult to characterize the corresponding deviations from the mean. Moreover, the change of measure in Lemma 2 plays a key role in the proof of Theorem 3, and leads to the additional factor $(n+1)^{|s_{\mathrm{dif}}|}$ in (28).

4.3 Simplifications of Theorems 3 and 4

Next, we reduce the preceding theorems to a form that is more amenable to inferring bounds on the required number of measurements. We start with the achievability part:

1. Observe that, conditioned on $\beta_s = b_s$, the mean of $\imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}}, \beta_s)$ is $nI_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)$ (*cf.*, (15)); this is due to to the fact the marginal distribution corresponding to any single measurement is given by (9).

2. Fix $\delta_2 \in (0,1)$ and $\kappa > 0$, and suppose that for a fixed value b_s of β_s, we have for all $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ that

$$n \ge \frac{\log\binom{p-k}{|s_{\mathrm{dif}}|} + 2\log\left(\frac{k}{\delta_1}\binom{k}{|s_{\mathrm{dif}}|}(n+1)^{|s_{\mathrm{dif}}|}\right) + \gamma}{I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)(1-\delta_2)} \tag{31}$$

$$\frac{\mathrm{Var}\big[\imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}}, \beta_s) \,|\, \beta_s = b_s\big]}{I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)^2} \le n\kappa. \tag{32}$$

   Combining these conditions with the previous step, the union bound, and Chebyshev's inequality, we obtain

$$\mathbb{P}\Bigg[\bigcup_{\substack{(s_{\mathrm{dif}}, s_{\mathrm{eq}}) \\ s_{\mathrm{dif}} \neq \emptyset}} \Bigg\{ \imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}}, \beta_s) \le \log \binom{p-k}{|s_{\mathrm{dif}}|} + 2\log\left(\frac{k}{\delta_1}\binom{k}{|s_{\mathrm{dif}}|}(n+1)^{|s_{\mathrm{dif}}|}\right) + \gamma \Bigg\} \,\Big|\, \beta_s = b_s \Bigg] \le (2^k - 1)\frac{\kappa}{n\delta_2^2}, \tag{33}$$

   since there are exactly $2^k - 1$ ways of splitting s into $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ with $s_{\mathrm{dif}} \neq \emptyset$.

3. Combining the above observations gives

$$P_{\mathrm{e}} \le \mathbb{P}\big[\beta_s \notin \mathcal{B}(\delta_1, \delta_2, \gamma)\big] + (2^k - 1)\frac{\kappa}{n\delta_2^2} + P_0(\gamma) + 2\delta_1, \tag{34}$$

   where

$$\mathcal{B}(\delta_1, \delta_2, \gamma) := \big\{ b_s \,:\, (31) \text{ and } (32) \text{ hold for all } (s_{\mathrm{dif}}, s_{\mathrm{eq}}) \text{ with } s_{\mathrm{dif}} \neq \emptyset \big\}. \tag{35}$$

The simplification of Theorem 4 is done using similar steps. Fix $\delta_2 > 0$, and suppose that, for a fixed value b_s of β_s, the pair $(s_{\mathrm{dif}}, s_{\mathrm{eq}}) = (s_{\mathrm{dif}}(b_s), s_{\mathrm{eq}}(b_s))$ is such that

$$n \le \frac{\log\binom{p-k+|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|} - \log\delta_1}{I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)(1+\delta_2)} \tag{36}$$

and (32) holds. Combining these with Chebyshev's inequality, we find that the first probability in (30), with an added conditioning on $\beta_s = b_s$, is lower bounded by $1 - \frac{\kappa}{n\delta_2^2}$. Recalling that the partition $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ is an arbitrary function of β_s, we can ensure that (36) coincides with

$$n \le \max_{(s_{\mathrm{dif}}, s_{\mathrm{eq}}) \,:\, s_{\mathrm{dif}} \neq \emptyset} \frac{\log\binom{p-k+|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|} - \log\delta_1}{I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)(1+\delta_2)} \tag{37}$$

by choosing each pair $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ as a function of b_s to achieve this maximum. Combining these observations, we obtain

$$P_{\mathrm{e}} \ge \mathbb{P}\big[\beta_s \in \mathcal{B}'(\delta_1, \delta_2)\big]\Big(1 - \frac{\kappa}{n\delta_2^2}\Big) - \delta_1, \tag{38}$$

where

$$\mathcal{B}'(\delta_1,\delta_2) := \big\{b_s : \text{(32) and (36) hold for all } (s_{\rm dif}, s_{\rm eq}) \text{ with } s_{\rm dif} \neq \emptyset\big\}. \quad (39)$$

4.4 Bounding the Variance

In order to apply (35) and (39), we must characterize the variance appearing in (32). Under the setting of i.i.d. measurement matrices studied in [17], the variance of $\imath^n(\mathbf{X}_{s_{\rm dif}};\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},b_s)$ simplifies to n times the variance of a single-letter information density $\imath(X_{s_{\rm dif}};Y|X_{s_{\rm eq}},b_s)$. This is not the case in the present setting, as the columns of $\mathbf{X}_s$ are generated according to the non-i.i.d. distribution $P_{\mathbf{X}_0}$. However, we can still characterize each such variance accurately, as shown in the following.

Lemma 3. *For any fixed k, $b_s = (b_1,\dots,b_k)$ and $(s_{\rm dif}, s_{\rm eq})$, we have*

$$\mathrm{Var}\big[\imath^n(\mathbf{X}_{s_{\rm dif}};\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)\,|\,\beta_s = b_s\big] = nV_{s_{\rm dif},s_{\rm eq}}(b_s) + O(1), \quad (40)$$

where

$$V_{s_{\rm dif},s_{\rm eq}}(b_s) := \mathrm{Var}\big[\imath(X_{s_{\rm dif}};Y|X_{s_{\rm eq}},\beta_s)\,|\,\beta_s = b_s\big] - \sum_{i\in s}\mathrm{Var}\Big[\mathbb{E}[\imath(X_{s_{\rm dif}};Y|X_{s_{\rm eq}},\beta_s)\,|\,X_{s\setminus\{i\}}]\,\Big|\,\beta_s = b_s\Big]. \quad (41)$$

Proof. This follows by writing the variance as a sum of covariance terms, and then characterizing the joint distribution between two different columns of the measurement matrix $\mathbf{X}$. See Appendix D for details. □

It follows immediately from Lemma 3 that (32) holds for all b_s within an arbitrary set $\mathcal{B}_0$ upon setting $\kappa = \sup_{(s_{\rm dif},s_{\rm eq}),b_s\in\mathcal{B}_0}\frac{V_{s_{\rm dif},s_{\rm eq}}(b_s)}{I_{s_{\rm dif},s_{\rm eq}}(b_s)^2}(1+o(1))$. Note however, that in order to ensure that κ is finite, one should avoid cases in which $V_{s_{\rm dif},s_{\rm eq}}(b_s)=\infty$ or $I_{s_{\rm dif},s_{\rm eq}}(b_s)=0$.

4.5 Proof of Theorem 1

Theorem 1 follows without difficulty from the preceding two subsections. For the achievability part, we first observe from (29) that

$$\gamma = \log\frac{1}{\min_{b_s}P_{\beta_s}(b_s)} \implies P_0(\gamma)=0, \quad (42)$$

which follows since $P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{y}|\mathbf{x}_s) = \sum_{b_s}P_{\beta_s}(b_s)P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{y}|\mathbf{x}_s,b_s)$. Since k and $(b_1,\dots,b_k)$ are fixed (not scaling with p) in the theorem statement, we have $\gamma=\Theta(1)$. Similarly, and since $\rho=\Theta(1)$ by assumption, we have $I_{s_{\rm dif},s_{\rm eq}}(b_s)=\Theta(1)$ for all $(s_{\rm dif},s_{\rm eq})$. Moreover, since P_{β_s} is simply a random permutation, the variance in (32) is the same for all b_s on its support, and behaves as $O(n)$ by Lemma 3. This implies that $\kappa=\Theta(1)$ in (32). Hence, and by letting δ_1 vanish slowly as a function of p, we see that the second, third and fourth terms in (34) vanish provided that $\delta_2\to 0$ sufficiently slowly.

The first term in (34) also vanishes provided that n satisfies (31). By Stirling's approximation and the fact that $\delta_2\to 0$, we can readily simplify this to

$$n \geq \frac{|s_{\rm dif}|\log p + 2|s_{\rm dif}|\log n}{I_{s_{\rm dif},s_{\rm eq}}(b_s)}(1+o(1)). \quad (43)$$

Considering without loss of generality the case that (17) holds with equality, we see that (43) is indeed satisfied for sufficiently large p, since $\log n = O(\log\log p) = o(\log p)$.

For the converse part, we may also assume without loss of generality that (17) holds with equality, as the decoder can always choose to ignore measurements. We obtain the desired result be applying similar (yet simpler) steps to those above starting from (38).

5 Conclusion

We have provided exact thresholds on the required number of measurements for support recovery via linear sketching, using random constructions known to produce expander matrices with high probability. A key tool in our analysis was a change-of-measure technique, which may prove useful in handling other random constructions. We obtained bounds nearly matching those of Gaussian measurements at low SNRs, while showing that the gap between the two is significant at high SNRs. An important extension of this work is handling the case that k scales with p, which appears to require more sophisticated concentration inequalities in place of Chebyshev's inequality. Unfortunately, Bernstein's inequality (used in [17]) does not appear to be suitable, since we need to deal with sums of *non-independent* random variables. Finally, our results motivate the study of practical decoding methods for obtaining the optimal thresholds, or characterizing how the thresholds change when one switches to practical decoders such as the LASSO.

Acknowledgment

This work was supported in part by the European Commission under Grant ERC Future Proof, SNF 200021-146750 and SNF CRSII2-147633, and by the 'EPFL Fellows' programme under Horizon2020 Grant agreement no. 665667.

References

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

[2] K. Do Ba, P. Indyk, E. Price, and D. P. Woodruff, "Lower bounds for sparse recovery," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2010, pp. 1190–1197.

[3] E. Price and D. P. Woodruff, "$(1+\epsilon)$-approximate sparse recovery," in *IEEE Symp. Found. Comp. Sci. (FOCS)*, 2011.

[4] E. Axell, G. Leus, E. Larsson, and H. Poor, "Spectrum sensing for cognitive radio : State-of-the-art and recent advances," *IEEE Sig. Proc. Mag.*, vol. 29, no. 3, pp. 101–116, May 2012.

[5] G. Cormode, F. Korn, S. Muthukrishnan, and D. Srivastava, "Finding hierarchical heavy hitters in data streams," in *Proc. Int. Conf. Very Large Data Bases*, 2003.

[6] G. Cormode and S. Muthukrishnan, "What's hot and what's not: Tracking most frequent items dynamically," *ACM Trans. Database Sys.*, vol. 30, no. 1, pp. 249–278, March 2005.

[7] 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.

[8] R. Berinde, "Advances in sparse signal recovery methods," Master's thesis, MIT, 2009.

[9] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4299–4308, Sep. 2009.

[10] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," *Proc. IEEE*, vol. 98, no. 6, pp. 937–947, June 2010.

[11] B. Bah, L. Baldassarre, and V. Cevher, "Model-based sketching and recovery with expanders," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2014, pp. 1529–1543.

[12] 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.

[13] 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.

[14] G. Reeves, "Sparsity pattern recovery in compressed sensing," Ph.D. dissertation, Duke University, 2011.

[15] 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.

[16] 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.

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

[18] ——, "Phase transitions in group testing," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

[19] S. Hoory, N. Linial, and A. Wigderson, "Expander graphs and their applications," *Bulletin Amer. Math. Soc.*, vol. 43, no. 4, pp. 439–561, Oct. 2006.

[20] M. S. Pinsker, "On the complexity of a concentrator," in *7th Int. Teleg. Conf.*, 1973.

[21] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[22] 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.

[23] 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.

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

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2001.

[26] V. Prelov and S. Verdu, "Second-order asymptotics of mutual information," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1567–1580, Aug. 2004.

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

[28] V. Y. F. Tan, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Found. Trend. Comms. Inf. Theory*, vol. 11, no. 1-2, pp. 1–184, 2014.

[29] J. Scarlett, A. Martinez, and A. Guillen i Fabregas, "Second-order rate region of constant-composition codes for the multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 157–172, Jan 2015.

Supplementary Material for "Limits on Sparse Support Recovery via Linear Sketching with Random Expander Matrices"

(AISTATS 2016, Jonathan Scarlett and Volkan Cevher)

Note that all citations here are to the bibliography in the main document, and similarly for many of the cross-references.

A Proof of Lemma 1

In the notation of Definition 1, let $\mathcal{E}_\ell$ ($\ell = 1, \dots, k$) be the event that some set S of cardinality ℓ fails to satisfy the expansion property, i.e., $|\mathcal{N}_{\mathbf{X}}(S)| < (1-\epsilon)d|S|$. We start with the following non-asymptotic bound given in [8]:

$$\mathbb{P}[\mathcal{E}_\ell] \le \binom{p}{\ell}\binom{d\ell}{\epsilon d\ell}\Big(\frac{d\ell}{n}\Big)^{\epsilon d\ell}. \tag{44}$$

Applying the bounds $\log\binom{p}{\ell} \le \ell \log p$ and $\log\binom{d\ell}{\epsilon d\ell} \le d\ell H_2(\epsilon)$, we obtain

$$\log \mathbb{P}[\mathcal{E}_\ell] \le \ell \log p + d\ell H_2(\epsilon) + \epsilon d\ell \log \frac{d\ell}{n} \tag{45}$$

$$= \ell \log p - d\ell\Big(\epsilon \log \frac{n}{d\ell} - H_2(\epsilon)\Big) \tag{46}$$

Since $k = \Theta(1)$, we obtain from the union bound that $\mathbb{P}\big[\cup_{\ell=1,\dots,k} \mathcal{E}_s\big] \to 0$ provided that (46) tends to $-\infty$ for all ℓ. This is true provided that the second inequality in (2) holds; the dominant condition is the one with $\ell = k$.

B Proof of Theorem 3

Recall the definitions of the random variables in (10)–(11), and the information densities in (25)–(27). 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

$$\tilde{\imath}(\mathbf{x}_{s_{\rm dif}}; \mathbf{y}|\mathbf{x}_{s_{\rm eq}}) > \gamma_{|s_{\rm dif}|} \tag{47}$$

for all $2^k - 1$ partitions $(s_{\rm dif}, s_{\rm eq})$ of s with $s_{\rm dif} \neq \emptyset$. If no such s exists, or if multiple exist, then an error is declared.

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 1.2), and the decoder that we have chosen exhibits a similar symmetry, we can condition on $S = s = \{1, \dots, k\}$. By the union bound, the error probability is upper bounded by

$$P_{\rm e} \le \mathbb{P}\Big[\bigcup_{(s_{\rm dif}, s_{\rm eq})} \Big\{\tilde{\imath}(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}) \le \gamma_{|s_{\rm dif}|}\Big\}\Big] + \sum_{\bar{s} \in \mathcal{S}\setminus\{s\}} \mathbb{P}\Big[\tilde{\imath}(\mathbf{X}_{\bar{s}\setminus s}; \mathbf{Y}|\mathbf{X}_{\bar{s}\cap s}) > \gamma_{|s_{\rm dif}|}\Big], \tag{48}$$

where here and subsequently we let the condition $s_{\rm dif} \neq \emptyset$ remain implicit. 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 with the information density corresponding to $s_{\rm dif} = \bar{s}\setminus s$ and $s_{\rm eq} = s \cap \bar{s}$.

As mentioned previously, a key tool in the proof is the following change of measure (with $\ell := |s_{\rm dif}|$):

$$P_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}}) = \sum_{\mathbf{x}_{s_{\rm dif}}} \Big(\prod_{i \in s_{\rm dif}} P_{\mathbf{X}_0}(\mathbf{x}_i)\Big) P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{s_{\rm dif}}, \mathbf{x}_{s_{\rm eq}}) \tag{49}$$

$$\le (n+1)^\ell \sum_{\mathbf{x}_{s_{\rm dif}}} \Big(\prod_{i \in s_{\rm dif}} P_X^n(\mathbf{x}_i)\Big) P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{s_{\rm dif}}, \mathbf{x}_{s_{\rm eq}}) \tag{50}$$

$$= (n+1)^\ell \widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}}), \tag{51}$$

where we have used the definitions in (23)–(24), and (50) follows from (12). By an identical argument, we have

$$P_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}}, b_s) \le (n+1)^\ell \widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}}, b_s), \tag{52}$$

where $\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s} := P^n_{Y|X_{s_{\rm eq}}\beta_s}$ has an i.i.d. law.

We can weaken the second probability in (48) as follows (with $\ell := |\bar{s}\backslash s|$):

$$\mathbb{P}\Big[\tilde{\imath}(\mathbf{X}_{\bar{s}\backslash s};\mathbf{Y}|\mathbf{X}_{\bar{s}\cap s}) > \gamma_\ell\Big]$$

$$= \sum_{\mathbf{x}_{\bar{s}\cap s},\mathbf{x}_{\bar{s}\backslash s}} P^{k-\ell}_{\mathbf{X}_0}(\mathbf{x}_{\bar{s}\cap s}) P^{\ell}_{\mathbf{X}_0}(\mathbf{x}_{\bar{s}\backslash s}) \int_{\mathbb{R}^n} d\mathbf{y}\, P_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\cap s}) \mathbb{1}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\backslash s},\mathbf{x}_{\bar{s}\cap s})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\cap s})} > \gamma_\ell\bigg\} \tag{53}$$

$$\le (n+1)^\ell \sum_{\mathbf{x}_{\bar{s}\cap s},\mathbf{x}_{\bar{s}\backslash s}} P^{k-\ell}_{\mathbf{X}_0}(\mathbf{x}_{\bar{s}\cap s}) P^{\ell}_{\mathbf{X}_0}(\mathbf{x}_{\bar{s}\backslash s}) \int_{\mathbb{R}^n} d\mathbf{y}\, \widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\cap s}) \mathbb{1}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\backslash s},\mathbf{x}_{\bar{s}\cap s})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\cap s})} > \gamma_\ell\bigg\} \tag{54}$$

$$\le (n+1)^\ell \sum_{\mathbf{x}_{\bar{s}\cap s},\mathbf{x}_{\bar{s}\backslash s}} P^{k-\ell}_{\mathbf{X}_0}(\mathbf{x}_{\bar{s}\cap s}) P^{\ell}_{\mathbf{X}_0}(\mathbf{x}_{\bar{s}\backslash s}) \int_{\mathbb{R}^n} d\mathbf{y}\, P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\backslash s},\mathbf{x}_{\bar{s}\cap s}) e^{-\gamma_\ell} \tag{55}$$

$$= (n+1)^\ell e^{-\gamma_\ell}, \tag{56}$$

where in (53) 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, (54) follows from (51), and (55) follows by bounding $\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}$ using the event within the indicator function, and then upper bounding the indicator function by one. Substituting (56) into (48) gives

$$P_{\rm e} \le \mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \Big\{\tilde{\imath}(\mathbf{X}_{s_{\rm dif}};\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}) \le \gamma_\ell\Big\}\bigg] + \sum_{\ell=1}^{k}\binom{p-k}{\ell}\binom{k}{\ell}(n+1)^\ell e^{-\gamma_\ell}, \tag{57}$$

where the combinatorial terms arise from a standard counting argument [7].

We now fix the constants $\gamma'_1,\dots,\gamma'_k$ arbitrarily, and recall the following steps from [17] (again writing $\ell := |s_{\rm dif}|$):

$$\mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \Big\{\tilde{\imath}(\mathbf{X}_{s_{\rm dif}};\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}) \le \gamma_\ell\Big\}\bigg]$$

$$= \mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm dif}},\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}})} \le \gamma_\ell\bigg\}\bigg] \tag{58}$$

$$\le \mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm dif}},\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}})} \le \gamma_\ell \cap \log\frac{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)} \le \gamma'_\ell\bigg\}\bigg]$$
$$+ \mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \bigg\{\log\frac{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)} > \gamma'_\ell\bigg\}\bigg] \tag{59}$$

$$\le \mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm dif}},\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)} \le \gamma_\ell + \gamma'_\ell\bigg\}\bigg]$$
$$+ \mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \bigg\{\log\frac{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)} > \gamma'_\ell\bigg\}\bigg]. \tag{60}$$

The second term in (60) is upper bounded as

$$\mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})} \bigg\{\log\frac{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)} > \gamma'_\ell\bigg\}\bigg]$$

$$\le \sum_{(s_{\rm dif},s_{\rm eq})} \mathbb{P}\bigg[\log\frac{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)} > \gamma'_\ell\bigg] \tag{61}$$

$$= \sum_{(s_{\rm dif},s_{\rm eq})} \sum_{b_s,\mathbf{x}_{s_{\rm eq}}} P_{\beta_s}(b_s) P^{k-\ell}_{\mathbf{X}_0}(\mathbf{x}_{s_{\rm eq}}) \int_{\mathbb{R}^n} d\mathbf{y}\, P_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}},b_s) \mathbb{1}\bigg\{\log\frac{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}},b_s)} > \gamma'_\ell\bigg\} \tag{62}$$

$$\leq (n+1)^\ell \sum_{(s_{\rm dif},s_{\rm eq})} \sum_{b_s,\mathbf{x}_{s_{\rm eq}}} P_{\beta_s}(b_s) P_{\mathbf{X}_0}^{k-\ell}(\mathbf{x}_{s_{\rm eq}}) \int_{\mathbb{R}^n} d\mathbf{y}\, \widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}},b_s) \mathbb{1}\Big\{ \log \frac{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}})}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}},b_s)} > \gamma'_\ell \Big\} \tag{63}$$

$$\leq (n+1)^\ell \sum_{(s_{\rm dif},s_{\rm eq})} \sum_{b_s,\mathbf{x}_{s_{\rm eq}}} P_{\beta_s}(b_s) P_{\mathbf{X}_0}^{k-\ell}(\mathbf{x}_{s_{\rm eq}}) \int_{\mathbb{R}^n} d\mathbf{y}\, \widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}}(\mathbf{y}|\mathbf{x}_{s_{\rm eq}}) e^{-\gamma'_\ell} \tag{64}$$

$$= (n+1)^\ell \sum_{\ell=1}^k \binom{k}{\ell} e^{-\gamma'_\ell}, \tag{65}$$

where (61) follows from the union bound, and the remaining steps follow the arguments used in (53)–(56) (with (52) used in place of (51)).

We now upper bound the first term in (60), again following [17]. The numerator in the first term in (60) equals $P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)$ for all $(s_{\rm dif}, s_{\rm eq})$ (recall the definition in (22)), and we can thus write the overall term as

$$\mathbb{P}\Big[\log P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s) \leq \max_{(s_{\rm dif},s_{\rm eq})} \big\{ \log \widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s) + \gamma_\ell + \gamma'_\ell \big\} \Big]. \tag{66}$$

Using the same steps as those used in (58)–(60), we can upper bound this by

$$\mathbb{P}\Big[\log P_{\mathbf{Y}|\mathbf{X}_s\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s) \leq \max_{(s_{\rm dif},s_{\rm eq})} \big\{ \log \widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s) + \gamma_\ell + \gamma'_\ell + \gamma \big\} \Big] + \mathbb{P}\Big[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} > \gamma \Big] \tag{67}$$

for any constant γ. Reversing the step in (66), this can equivalently be written as

$$\mathbb{P}\Big[\bigcup_{(s_{\rm dif},s_{\rm eq})} \Big\{ \log \frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\rm dif}}\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm dif}},\mathbf{X}_{s_{\rm eq}},\beta_s)}{\widetilde{P}_{\mathbf{Y}|\mathbf{X}_{s_{\rm eq}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},\beta_s)} \leq \gamma_\ell + \gamma'_\ell + \gamma \Big\} \Big] + \mathbb{P}\Big[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} > \gamma \Big]. \tag{68}$$

The first logarithm in the first term is the information density in (26). Moreover, the choices

$$\gamma_\ell = \log\Big(\frac{k}{\delta_1} \binom{p-k}{\ell} \binom{k}{\ell} (n+1)^\ell \Big) \tag{69}$$

$$\gamma'_\ell = \log\Big(\frac{k}{\delta_1} \binom{k}{\ell} (n+1)^\ell \Big) \tag{70}$$

make (65) and the second term in (57) be upper bounded by δ_1 each. Hence, and combining (60) with (65) and (68), and recalling that $\ell = |s_{\rm dif}|$, we obtain (28).

C Proof of Theorem 2

Fix $0 < b_{\min} < b_{\max} < \infty$, and let $\mathcal{B}_0 := \{b_s : \min_i |b_i| \geq b_{\min} \cap \max_i |b_i| \leq b_{\max}\}$. The main step in proving Theorem 2 is in extending the arguments of Section 4.5 to show that

$$P_e \leq \mathbb{P}\Big[n \leq \max_{(s_{\rm dif},s_{\rm eq}):\, s_{\rm dif}\neq\emptyset} \frac{|s_{\rm dif}| \log p}{I_{s_{\rm dif},s_{\rm eq}}(\beta_s)} (1+\eta) \cap \beta_s \in \mathcal{B}_0 \Big] + \mathbb{P}\big[\beta_s \notin \mathcal{B}_0\big] + o(1), \tag{71}$$

and

$$P_e \geq \mathbb{P}\Big[n \leq \max_{(s_{\rm dif},s_{\rm eq}):\, s_{\rm dif}\neq\emptyset} \frac{|s_{\rm dif}| \log p}{I_{s_{\rm dif},s_{\rm eq}}(\beta_s)} (1-\eta) \cap \beta_s \in \mathcal{B}_0 \Big] + o(1), \tag{72}$$

Before proving these, we show how they yield the theorem. Using (16), it is readily verified that each $I_{s_{\rm dif},s_{\rm eq}}(\beta_s)$, with an i.i.d. Gaussian vector β_s, is a continuous random variable having no mass points. By taking $\eta \to 0$ sufficiently slowly and noting that we have restricted β_s to the set $\mathcal{B}_0$ (within which all of the $I_{s_{\rm dif},s_{\rm eq}}(\beta_s)$ are

bounded away from zero and infinity), we conclude that (71)–(72) remain true when η is replaced by zero, and its contribution is factored into the $o(1)$ terms. Hence, we obtain Theorem 2 by (i) dropping the condition $\beta_s \in \mathcal{B}_0$ from the first probability in (71); (ii) using the identity $\mathbb{P}[\mathcal{A}_1 \cap \mathcal{A}_2] \geq \mathbb{P}[\mathcal{A}_1] - \mathbb{P}[\mathcal{A}_2]$ to remove the same condition from the first probability in (72); (iii) noting that the remainder term $\mathbb{P}\big[\beta_s \notin \mathcal{B}_0\big]$ can be made arbitrarily small by choosing $b_{\min}$ sufficiently small and $b_{\max}$ sufficiently large.

It remains to establish (71)–(72). Recall the value of κ given following Lemma 3. The above choice of $\mathcal{B}_0$ ensures that all of the non-zero entries are bounded away from 0 and ∞, so that the mutual informations $I_{s_{\text{dif}},s_{\text{eq}}}(\beta_s)$ and variances $V_{s_{\text{dif}},s_{\text{eq}}}(\beta_s)$ are bounded away from zero and infinity, and hence $\kappa = \Theta(1)$.

Since P_{β_s} is continuous, we must choose γ and handle P_0 in (29) differently to the above. Similarly to the analysis of Gaussian measurements in [17], we fix $\delta_0 > 0$ and note that Chebyshev's inequality implies

$$\gamma = I_0 + \sqrt{\frac{V_0}{\delta_0}} \implies P_0(\gamma) \leq \delta_0, \tag{73}$$

where

$$I_0 := I(\beta_s; \mathbf{Y}|\mathbf{X}_s) \tag{74}$$

$$V_0 := \text{Var}\bigg[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)}\bigg]. \tag{75}$$

The following is a straightforward extension of [17, Prop. 4] to expander-based measurements.

Proposition 1. *The quantities I_0 and V_0 defined in (74)–(75) satisfy*

$$I_0 \leq \frac{k}{2}\log\bigg(1 + \frac{d\sigma_\beta^2}{\sigma^2}\bigg) \tag{76}$$

$$V_0 \leq 2n. \tag{77}$$

Proof. See Appendix E. □

We can now obtain (71)–(72) using the steps of the previous subsection; the condition $\mathbb{P}[\beta_s \in \mathcal{B}_0]$ arises in (35) and (39) due to the fact that this condition was used to establish (32), and the first two probabilities in (71) arise from the identity $\mathbb{P}[\mathcal{A}_1 \cup \mathcal{A}_2] \leq \mathbb{P}[\mathcal{A}_1 \cup \mathcal{A}_2^c] + \mathbb{P}[\mathcal{A}_2]$. The only additional step is in showing that we can simultaneously achieve $\gamma = o(\log p)$ and $P_0(\gamma) = o(1)$ in the achievability part whenever $n = \Theta(\log p)$, in the same way that we showed $2|s_{\text{dif}}|\log n = o(\log p)$ in the previous subsection. This immediately follows by substituting (76)–(77) into (73) (along with $d = O(n) = O(\log p)$) to obtain $\gamma = O(\log\log p) + \sqrt{\log p}$ for any $\delta_0 > 0$, and noting that δ_0 (and hence $P_0(\gamma)$) in (73) can be arbitrarily small.

D Proof of Lemma 3

We prove the lemma by characterizing the variance of a general function of $(\mathbf{X}_s, \mathbf{Y})$ of the form $f^n(\mathbf{X}_s, \mathbf{Y}) := \sum_{i=1}^n f(X_s^{(i)}, Y^{(i)})$. Clearly all of the quantities $\imath^n$ for the various $(s_{\text{dif}}, s_{\text{eq}})$ can be written in this general form. We have

$$\text{Var}\big[f^n(\mathbf{X}_s, \mathbf{Y})\big] = \text{Var}\bigg[\sum_{i=1}^n f(X_s^{(i)}, Y^{(i)})\bigg] \tag{78}$$

$$= \sum_{i=1}^n \sum_{j=1}^n \text{Cov}\Big[f(X_s^{(i)}, Y^{(i)}), f(X_s^{(j)}, Y^{(j)})\Big] \tag{79}$$

$$= n\text{Var}\Big[f(X_s, Y)\Big] + (n^2 - n)\text{Cov}\Big[f(X_s, Y), f(X_s', Y')\Big], \tag{80}$$

where (X_s, Y) and (X_s', Y') correspond to two different indices in $\{1, \cdots, n\}$; here (80) follows by simple symmetry considerations for the cases $i = j$ and $i \neq j$.

To compute the covariance of term in (80), we first find the joint distribution of (X_s, Y) and (X'_s, Y'). As noted in [29, Sec. IV-B], a uniform permutation of a vector with d ones and $n-d$ zeros can be interpreted as successively performing uniform sampling from a collection of symbols without replacement (n times in total), where the initial collection contains d ones and $n-d$ zeros. By considering the first two steps of this procedure, we obtain

$$\mathbb{P}[X_i = x_i] = P_X(x_i) \tag{81}$$

$$\mathbb{P}[X'_i = x'_i | X_i = x_i] = \frac{nP_X(x'_i) - \mathbb{1}\{x_i = x'_i\}}{n-1} \tag{82}$$

for $\nu = 1, 2$, where $P_X(1) = 1 - P_X(0) = \frac{d}{n}$. Denoting the right-hand side of (82) by $P'_X(x'_i|x_i)$, and writing $\mu_f := \mathbb{E}[f(X_s, Y)]$, the covariance in (80) is given by

$$\begin{aligned}
&\mathrm{Cov}\Big[f(X_s, Y), f(X'_s, Y')\Big] \\
&\quad = \mathbb{E}\Big[\big(f(X_s, Y) - \mu_f\big)\big(f(X'_s, Y') - \mu_f\big)\Big] &(83)\\
&\quad = \sum_{x_s} P_X^k(x_s) \sum_{x'_s} \bigg(\prod_{i \in s} P'_X(x'_i|x_i)\bigg) \mathbb{E}\Big[\big(f(x_s, Y) - \mu_f\big)\big(f(x'_s, Y') - \mu_f\big) \,\Big|\, X_s = x_s, X'_s = x'_s\Big]. &(84)
\end{aligned}$$

We now consider the various terms arising by substituting (82) into (84) and performing a binomial-type expansion of the product:

- There is a single term of the form (84) with each $P'_x(x'_i|x_i)$ replaced by $\frac{nP_X(x'_i)}{n-1}$. This yields an average of $\big(f(X_s, Y) - \mu_f\big)\big(f(X'_s, Y') - \mu_f\big)$ over *independent* random variables X_s and X'_s, and therefore evaluates to zero.
- There are k terms in which one value $P'_x(x'_i|x_i)$ in (84) is replaced by $\frac{-\mathbb{1}\{x_i = x'_i\}}{n-1}$ and the other $k-1$ are replaced by $\frac{nP_X(x'_i)}{n-1}$. Each such term can be written as $-\frac{n}{(n-1)^2}\mathrm{Var}\big[\mathbb{E}[f(X_s, Y) \,|\, X_{s\setminus\{i\}}]\big]$, which in turn behaves as $-\frac{1}{n}\mathrm{Var}\big[\mathbb{E}[f(X_s, Y) \,|\, X_{s\setminus\{i\}}]\big] + O(1)$.
- All of the remaining terms replace $P'_x(x'_i|x_i)$ in (84) by $\frac{-\mathbb{1}\{x_i = x'_i\}}{n-1}$ for at least two values of i. All such terms are easily verified to behave as $O\big(\frac{1}{n^2}\big)$, and the number of such terms is finite and does not scale with n (recall that k is fixed by assumption).

Substituting these cases into (84) and recalling that $k = \Theta(1)$ and $\frac{d}{n} = \Theta(1)$, we obtain (40).

E Proof of Proposition 1

Here we characterize I_0 and V_0, defined in (74)–(75), via an extension of the analysis given in [17, App. B]. Since $\mathbf{Y} = \mathbf{X}_s\beta_s + \mathbf{Z}$, we have

$$\begin{aligned}
I_0 = I(\beta_s; \mathbf{Y}|\mathbf{X}_s) &= H(\mathbf{Y}|\mathbf{X}_s) - H(\mathbf{Y}|\mathbf{X}_s, \beta_s) &(85)\\
&= H(\mathbf{X}_s\beta_s + \mathbf{Z}|\mathbf{X}_s) - H(\mathbf{Z}). &(86)
\end{aligned}$$

From [25, 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 (86) gives

$$\begin{aligned}
I_0 &= \frac{1}{2}\mathbb{E}\Big[\log\det\Big(\mathbf{I}_n + \frac{\sigma_\beta^2}{\sigma^2}\mathbf{X}_s\mathbf{X}_s^T\Big)\Big] &(87)\\
&= \frac{1}{2}\mathbb{E}\Big[\log\det\Big(\mathbf{I}_k + \frac{\sigma_\beta^2}{\sigma^2}\mathbf{X}_s^T\mathbf{X}_s\Big)\Big] &(88)\\
&= \frac{1}{2}\sum_{i=1}^{k}\mathbb{E}\Big[\log\Big(1 + \frac{\sigma_\beta^2}{\sigma^2}\lambda_i(\mathbf{X}_s^T\mathbf{X}_s)\Big)\Big] &(89)
\end{aligned}$$

$$\leq \frac{k}{2}\log\Big(1+\frac{d\sigma_\beta^2}{\sigma^2}\Big), \tag{90}$$

where (88) follows from the identity $\det(\mathbf{I}+\mathbf{AB})=\det(\mathbf{I}+\mathbf{BA})$, (89) follows by writing the determinant as a product of eigenvalues (denoted by $\lambda_i(\cdot)$), and (90) follows from Jensen's inequality and the following calculation:

$$\frac{1}{k}\mathbb{E}\Big[\sum_{i=1}^{k}\lambda_i(\mathbf{X}_s^T\mathbf{X}_s)\Big] = \frac{1}{k}\mathbb{E}[\mathrm{Tr}(\mathbf{X}_s^T\mathbf{X}_s)] = \mathbb{E}[\mathbf{X}_1^T\mathbf{X}_1] = d, \tag{91}$$

since the norm of $\mathbf{X}_1$ is d almost surely. This concludes the proof of (76).

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

$$\log\frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} = \log\frac{P_{\mathbf{Z}}(\mathbf{Z})}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{X}_s\beta_s+\mathbf{Z}|\mathbf{X}_s)} \tag{92}$$

$$= I_0 - \frac{1}{2\sigma^2}\mathbf{Z}^T\mathbf{Z} + \frac{1}{2}(\mathbf{X}_s\beta_s+\mathbf{Z})^T\big(\sigma^2\mathbf{I}+\sigma_\beta^2\mathbf{X}_s\mathbf{X}_s^T\big)^{-1}(\mathbf{X}_s\beta_s+\mathbf{Z}), \tag{93}$$

where $P_{\mathbf{Z}}$ is the density of $\mathbf{Z}$, and (93) 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 (93) 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 $\mathbf{M}=\sigma^2\mathbf{I}+\sigma_\beta^2\mathbf{X}_s\mathbf{X}_s^T$, we similarly observe that the final term in (93) is a sum of χ^2 variables (this is true conditioned on any $\mathbf{X}_s=\mathbf{x}_s$, and hence also true unconditionally), again yielding a variance of $\frac{n}{2}$. We thus obtain (77) using the identity $\mathrm{Var}[A+B]\leq\mathrm{Var}[A]+\mathrm{Var}[B]+2\max\{\mathrm{Var}[A],\mathrm{Var}[B]\}$.