INFORMATION THEORETIC LIMITS FOR STANDARD AND ONE-BIT COMPRESSED SENSING WITH GRAPH-STRUCTURED SPARSITY

Adarsh Barik, Jean Honorio

Purdue University
Department of Computer Science
West Lafayette, Indiana, USA

ABSTRACT

In this paper, we analyze the information theoretic lower bound on the necessary number of samples needed for recovering a sparse signal under different compressed sensing settings. We focus on the weighted graph model, a model-based framework proposed by [1], for standard compressed sensing as well as for one-bit compressed sensing. We study both the noisy and noiseless regimes. Our analysis is general in the sense that it applies to any algorithm used to recover the signal. We carefully construct restricted ensembles for different settings and then apply Fano's inequality to establish the lower bound on the necessary number of samples. Furthermore, we show that our bound is tight for one-bit compressed sensing, while for standard compressed sensing, our bound is tight up to a logarithmic factor of the number of non-zero entries in the signal. *A full version of this paper is accessible at:* `https://www.cs.purdue.edu/homes/abarik/abarik_cs_icassp_full.pdf`

Index Terms— Compressive sensing, weighted graph model, information theoretic bounds

1. INTRODUCTION

Sparsity has been a useful tool to tackle high dimensional problems in many fields such as compressed sensing, machine learning and statistics. Several naturally occurring and artificially created signals manifest sparsity in their original or transformed domain. For instance, sparse signals play an important role in applications such as medical imaging, geophysical and astronomical data analysis, computational biology, remote sensing as well as communications.

In compressed sensing, sparsity of a high dimensional signal allows for the efficient inference of such a signal from a small number of observations. The true high dimensional sparse signal $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is not observed directly. Instead, a low dimensional linear transformation $\mathbf{y} \in \mathbb{R}^n$ is observed along with a design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{e} \;, \tag{1}$$

where $\mathbf{e} \in \mathbb{R}^d$ is a zero mean independent additive noise. The true high dimensional sparse signal $\boldsymbol{\beta}^*$ is then inferred from observations $(\mathbf{X}, \mathbf{y})$. We refer to the setup in equation 1 as "standard compressed sensing". Many signal acquisition settings such as magnetic resonance imaging [9] use standard compressed sensing as their underlying model. One-bit compressed sensing [10, 6] is a generalization of the setup presented in (1), in the sense that it considers quantizing the measurements to one bit, i.e.,

$$\mathbf{y} = \mathrm{sign}(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}) \;, \tag{2}$$

where the function sign acts on individual entries of $\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}$ and returns their sign. This kind of quantization is particularly appealing for hardware implementations.

The learning problem in compressed sensing is to recover a signal which is a *good* approximation of the true signal. The goodness of approximation can be measured by either a prespecified distance between the inferred and the true signal, e.g., ℓ_0-norm, ℓ_1-norm., or by the similarity of their support (i.e., the indices of their non-zero entries). The algorithms for compressed sensing try to provide performance guarantees for either one or both of these measures. For instance, [11], [12], [13], [14] and [15] provide performance guarantees in terms of distance, while [16] and [17] provide performance guarantees in terms of support recovery for standard compressed sensing. The authors in [8] provide guarantees in terms of both distance and support for one-bit compressed sensing. All the aforementioned works use deterministic sparse signals. There have been some empirical studies [18, 19, 20] where Bayesian priors have been used to model stochastic sparse signals but in this paper, we do not focus on the Bayesian framework.

The authors in [2] initially proposed a model-based sparse recovery framework. Under this framework, they showed that the sufficient number of samples for correct recovery is logarithmic with respect to the cardinality of the sparsity model (the sparsity model is defined only by the support of the sparse signal $\boldsymbol{\beta}^*$), i.e., the number of supports in the sparsity model. The model of [2] considered signals with common sparsity structure and small cardinality. Later, [1] proposed a weighted

Table 1: Sample Complexity Results for Structured Sparsity Models (d is the dimension of the true signal, s is the signal sparsity, i.e., the number of non-zero entries, g is the number of connected components, $\rho(G)$ is the maximum weight degree of graph G, B is the weight budget in the weighted graph model, K is the block sparsity, J is the number of entries in a block and N is the total number of blocks in the block structured sparsity model

	Standard Compressed Sensing		One-bit Compressed Sensing	
Sparsity Structure	Our Lower Bound	Upper Bound	Our Lower Bound	Upper Bound
Weighted Graph Model	$\tilde{\Omega}(s(\log \rho(G) \frac{B}{s}) + g \log \frac{d}{g})$	$O(s(\log \rho(G) \frac{B}{s}) + g \log \frac{d}{g})$ [1]	$\Omega(s(\log \rho(G) \frac{B}{s}) + g \log \frac{d}{g})$	NA
Tree Structured	$\tilde{\Omega}(s)$	$O(s)$ [2]	$\Omega(s)$	NA
Block Structured	$\tilde{\Omega}(KJ + K \log \frac{N}{K})$	$O(KJ + K \log \frac{N}{K})$ [2]	$\Omega(KJ + K \log \frac{N}{K})$	NA
Regular s-sparsity	$\tilde{\Omega}(s \log \frac{d}{s})$	$O(s \log \frac{d}{s})$ [3, 4, 5]	$\Omega(s \log \frac{d}{s})$	$O(s \log \frac{d}{s})$[6, 7, 8]

graph model for graph-structured sparsity and accompanied it with a nearly linear time recovery algorithm. They also analyzed the sufficient number of samples for efficient recovery. Notice that regular sparse signals, i.e., signals without any additional sparsity structure can be easily modeled using the weighted graph model. Independently, there has been a long line of work ([5, 4, 3]) which deals with algorithms with support recovery guarantees for regular sparse signals.

In this paper, we analyze the necessary condition on the sample complexity for exact sparse recovery. While our proof techniques can also be applied to any model-based sparse recovery framework, we apply our method to get the necessary number of samples to perform efficient recovery on a weighted graph model. We provide results for both the noisy and noiseless regimes of compressed sensing. We also extend our results to one-bit compressed sensing. We note that a lower bound on sample complexity was previously provided in [21] when the observer has access to only the measurements $\mathbf{y}$. Here, we analyze the more relevant setting in which the observer has access to the measurements $\mathbf{y}$ along with the design matrix $\mathbf{X}$. Compared to [21], we additionally analyze one-bit compressed sensing in detail. Table 1 shows a comparison of our information theoretic lower bounds on sample complexity under different settings with the existing upper bounds available in the literature. Note that our bounds for one-bit compressed sensing match upper bounds by [6]. The author in [22] provided information-theoretic lower bounds for regular sparsity in standard compressed sensing. For this case, along with other instances of standard compressed sensing our bounds are tight up to a factor of $\log s$, where s is the number of non-zero entries in $\boldsymbol{\beta}^*$.

2. PROBLEM DESCRIPTION

In this section, we introduce the observation model and later specialize it for specific problems such as standard compressed sensing and one-bit compressed sensing.

2.1. Notation

In what follows, we list down the notations which we use throughout the paper. The unobserved true d-dimensional signal is denoted by $\boldsymbol{\beta}^* \in \mathbb{R}^d$. The inferred signal is represented by $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$. We call a signal $\boldsymbol{\beta} \in \mathbb{R}^d, s < d$ an s-sparse signal if $\boldsymbol{\beta}$ contains only s non-zero entries. The n-dimensional observations are denoted by $\mathbf{y} \in \mathbb{R}^n, n \ll d$. We denote the design matrix by $\mathbf{X} \in \mathbb{R}^{n \times d}$. The $(i, j)^{\text{th}}$ element of the design matrix is denoted by $\mathbf{X}_{ij}, \forall\ 1 \leq i \leq n, 1 \leq j \leq d$. The i^{th} row of $\mathbf{X}$ is denoted by $\mathbf{X}_{i.}, \forall\ 1 \leq i \leq n$, and the j^{th} column of $\mathbf{X}$ is denoted by $\mathbf{X}_{.j}, \forall\ 1 \leq j \leq d$,. We assume that the true signal $\boldsymbol{\beta}^*$ belongs to a set $\mathcal{F}$, which is defined more formally later. The number of elements in a set A is denoted by $|A|$. The measurement vector $\mathbf{y} \in \mathbb{R}^n$ is a function $f(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e})$ of $\mathbf{X}, \boldsymbol{\beta}^*$ and $\mathbf{e}$ where $\mathbf{e} \in \mathbb{R}^n$ is Gaussian noise with i.i.d. entries, each with mean 0 and variance σ^2. For brevity, we use function notation f while keeping in mind that f acts on each row of $\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}$. The probability of the occurrence of an event E is denoted by $\mathbb{P}(E)$. The shorthand notation $[p]$ is used to denote the set $\{1, 2, \dots, p\}$. Other notations specific to weighted graph models are defined later in Section 3. For brevity and comparisons, we use the order notations $\tilde{o}$ and $\tilde{\Omega}$ which ignore logarithmic factors of s.

2.2. Observation Model

We define a general observation model. The learning problem is to estimate the unobserved true s-sparse signal $\boldsymbol{\beta}^*$ from noisy observations. Since $\boldsymbol{\beta}^*$ is a high dimensional signal, we do not sample it directly. Rather, we observe a function of its inner product with the rows of a randomized matrix $\mathbf{X}$. Formally, the i^{th} measurement $\mathbf{y}_i$ comes from the model, $\mathbf{y}_i = f(\mathbf{X}_{i.}\boldsymbol{\beta}^* + \mathbf{e}_i)$, where $f : \mathbb{R} \to \mathbb{R}$ is a fixed function. We observe n such i.i.d. samples and collect them in the measurement vector $\mathbf{y} \in \mathbb{R}^n$. We can express this mathematically by,

$$\mathbf{y} = f(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}) \,. \tag{3}$$

Our task is to recover an estimate $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ of $\boldsymbol{\beta}^*$ from the observations $(\mathbf{X}, \mathbf{y})$. By choosing an appropriate function f, we can describe specific instances of compressed sensing.

2.2.1. *Standard Compressed Sensing*

The standard compressed sensing is a special case of equation (3) by choosing $f(x) = x$. Then we simply have,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{e} \; . \tag{4}$$

2.2.2. *One-bit Compressed Sensing*

The problem of signal recovery in one-bit compressed sensing has been introduced recently [10]. In this setup, we do not have access to linear measurements but rather observations come in the form of a single bit. This can be modeled by choosing $f(x) = \mathrm{sign}(x)$. In other words, we have,

$$\mathbf{y} = \mathrm{sign}(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}) \; . \tag{5}$$

Note that we lose lot of information by limiting the observations to a single bit. It is known that for the noiseless case, unlike standard compressed sensing, one can only recover $\boldsymbol{\beta}^*$ up to scaling[6].

2.3. Problem Setting

We assume that the nature picks a true s-sparse signal $\boldsymbol{\beta}^*$ uniformly at random from a set of signals $\mathcal{F}$. Then observations are generated using the model described in equation (3). The function f is chosen appropriately for different settings. We also assume that the observer has access to the design matrix $\mathbf{X}$. Thus, observations are denoted by $(\mathbf{X}, \mathbf{y})$. This procedure can be interpreted as a Markov chain which is described as $\boldsymbol{\beta}^* \to \big(\mathbf{X}, \mathbf{y} = f(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e})\big) \to \hat{\boldsymbol{\beta}}$. We use this Markov chain in our proofs. We assume that the true signal $\boldsymbol{\beta}^*$ comes from a weighted graph model. We state our results for standard sparse compressed sensing and one-bit compressed sensing. We note that our arguments for establishing information theoretic lower bounds are not algorithm specific. We use Fano's inequality [23] to prove our result by carefully constructing restricted ensembles. Any algorithm which infers $\boldsymbol{\beta}^*$ from this particular ensemble would require a minimum number of samples. The use of restricted ensembles is customary for information theoretic lower bounds [24, 25].

3. WEIGHTED GRAPH MODEL (WGM)

We assume that the true s-sparse signal comes from a weighted graph model. This encompasses many commonly seen sparsity patterns in signals such as tree structured sparsity, block structured sparsity as well as the regular s-sparsity without any additional structure. Next, we introduce the Weighted Graph Model (WGM) which was proposed by [1].

The Weighted Graph Model is defined on an underlying graph $G = (V, E)$ whose vertices represent the coefficients of the unknown $s-$sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$ i.e. $V = [d] = \{1, 2, \ldots, d\}$. Moreover, the graph is weighted and thus we introduce a weight function $w : E \to \mathbb{N}$. Borrowing some notations from [1], $w(F)$ denotes the sum of edge weights in a forest $F \subseteq G$, i.e., $w(F) = \sum_{e \in F} w_e$. We also assume an upper bound on the total edge weight which is called the weight budget and is denoted by B. The number of non-zero coefficients of $\boldsymbol{\beta}^*$ is denoted by the sparsity parameter s. The number of connected components in a forest F is denoted by $\gamma(F)$. The weight-degree $\rho(v)$ of a node $v \in V$ is the largest number of adjacent nodes connected by edges with the same weight, i.e., $\rho(v) = \max_{b \in \mathbb{N}} |\{(v', v) \in E \mid w(v', v) = b\}|$. We define the weight-degree $\rho(G)$ of G to be the maximum weight-degree of any $v \in V$. Next, we define the Weighted Graph Model on coefficients of $\boldsymbol{\beta}^*$ as follows:

Definition 1 ([1])**.** *The $(G, s, g, B) - WGM$ is the set of supports defined as $\mathbb{M} = \{S \subseteq [d] \mid |S| = s$ and $\exists\, F \subseteq G$ with $V_F = S, \gamma(F) = g,\ w(F) \leq B\}$.*

4. MAIN RESULTS

In this section, we state our results for the standard compressed sensing and one-bit compressed sensing. We consider both the noisy and noiseless cases. We establish an information theoretic lower bound on the sample complexity for signal recovery on a WGM (See Appendix A for detailed proofs).

4.1. Results for Standard Compressed Sensing

For standard compressed sensing, the recovery is not exact for the noisy case but it is sufficiently close to the true signal in ℓ_2-norm with respect to the noise. Our setup, in this case, uses a Gaussian design matrix. The formal statement of our result is as follows.

Theorem 1 (Standard Compressed Sensing, Noisy Case)**.** *There exist a (G, s, g, B)-WGM $\mathbb{M}$ and a finite set of weights ψ such that if nature picks $\boldsymbol{\beta}^*$ uniformly at random from $\mathcal{F} = \{\boldsymbol{\beta} \mid \boldsymbol{\beta}_i = 0, \mathit{if}\, i \notin S, \boldsymbol{\beta}_i \in \psi, \mathit{if}\, i \in S, S \in \mathbb{M}\}$ and draws $n \in \tilde{o}(\log|\mathcal{F}|) = \tilde{o}((s-g)(\log\rho(G) + \log\frac{B}{s-g}) + g\log\frac{d}{g} + (s-g)\log\frac{g}{s-g} + s\log 2)$ i.i.d. observations from a family of distributions $\mathcal{D}$ indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ following equation (4), i.e., $(\mathbf{X}_{i.}, \mathbf{y}_i) \sim D(\boldsymbol{\beta})$ which represents $\mathbf{X}_{i.} \sim \mathcal{N}(0, \frac{1}{n})^d, \mathbf{y}_i = \mathbf{X}_{i.}\boldsymbol{\beta} + \mathbf{e}_i, \mathbf{e}_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \{1 \ldots n\}$ then,*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{D \in \mathcal{D}} \mathop{\mathbb{P}}_{\substack{\boldsymbol{\beta}^* \sim \mathrm{Unif}(\mathcal{F}) \\ (\mathbf{X}, \mathbf{y}) \sim D(\boldsymbol{\beta}^*)^n}} (\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})\| \geq C\|\mathbf{e}\|) \geq \frac{1}{10}$$

for $0 < C \leq C_0$, where $\hat{\boldsymbol{\beta}} : \mathbb{R}^{n \times d} \times \mathbb{R}^n \to \mathbb{R}^d$ is any procedure used to infer $\boldsymbol{\beta}^*$, and $\mathcal{D}$ is the family of all distributions indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ with support on $\mathbb{R}^{n \times d} \times \mathbb{R}^n$.

We provide a similar result for the noiseless case. In this case recovery is exact. We use a Bernoulli design matrix for our proofs. Note that a Bernoulli random variable denoted by $\mathrm{Bernoulli}(0.5)$ takes a value -1 with probability 0.5 and $+1$ with probability 0.5. In what follows, we state our result.

Theorem 2 (Standard Compressed Sensing, Noiseless Case)**.** *There exist a (G, s, g, B)-WGM $\mathbb{M}$ and a finite set of weights ψ such that if nature picks $\boldsymbol{\beta}^*$ uniformly at random from $\mathcal{F} = \{\boldsymbol{\beta} \mid \boldsymbol{\beta}_i = 0, \textit{if } i \notin S, \boldsymbol{\beta}_i \in \psi, \textit{if } i \in S, S \in \mathbb{M}\}$ and draws $n \in \tilde{o}(\log|\mathcal{F}|) = \tilde{o}((s-g)(\log\rho(G) + \log\frac{B}{s-g}) + g\log\frac{d}{g} + (s-g)\log\frac{g}{s-g} + s\log 2)$ i.i.d. observations from a family of distributions $\mathcal{D}$ indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ following equation (4) with zero noise, i.e., $(\mathbf{X}_{i.}, \mathbf{y}_i) \sim D(\boldsymbol{\beta})$ which represents $\mathbf{X}_{i.} \sim \frac{1}{\sqrt{n}}\mathrm{Bernoulli}(0.5)^d, \mathbf{y}_i = \mathbf{X}_{i.}\boldsymbol{\beta}, \forall i \in \{1 \ldots n\}$ then,*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{D \in \mathcal{D}} \underset{\substack{\boldsymbol{\beta}^* \sim \mathrm{Unif}(\mathcal{F}) \\ (\mathbf{X},\mathbf{y}) \sim D(\boldsymbol{\beta}^*)^n}}{\mathbb{P}} (\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})) \geq \frac{1}{2}$$

where $\hat{\boldsymbol{\beta}} : \mathbb{R}^{n \times d} \times \mathbb{R}^n \to \mathbb{R}^d$ is any procedure used to infer $\boldsymbol{\beta}^*$, and $\mathcal{D}$ is the family of all distributions indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ with support on $\mathbb{R}^{n \times d} \times \mathbb{R}^n$.

We note that when $s \gg g$ and $B \geq s - g$ then $\tilde{\Omega}((s-g)(\log\rho(G) + \log\frac{B}{s-g}) + g\log\frac{d}{g} + (s-g)\log\frac{g}{s-g} + s\log 2)$ is roughly $\tilde{\Omega}(s(\log\rho(G) + \log\frac{B}{s}) + g\log\frac{d}{g})$.

4.2. Results for One-bit Compressed Sensing

In this setting, we provide a construction which works for both the noisy and noiseless case. We use a Bernoulli design matrix for both of these setups. Our first result in this setting shows that even in our restricted ensemble recovering the true signal exactly is difficult.

Theorem 3 (One-bit Compressed Sensing, Exact Recovery)**.** *There exist a (G, s, g, B)-WGM $\mathbb{M}$ and a finite set of weights ψ such that if nature picks $\boldsymbol{\beta}^*$ uniformly at random from $\mathcal{F} = \{\boldsymbol{\beta} \mid \boldsymbol{\beta}_i = 0, \textit{if } i \notin S, \boldsymbol{\beta}_i \in \psi, \textit{if } i \in S, S \in \mathbb{M}\}$ and draws $n \in o(\log|\mathcal{F}|) = o((s-g)(\log\rho(G) + \log\frac{B}{s-g}) + g\log\frac{d}{g} + (s-g)\log\frac{g}{s-g} + s\log 2)$ i.i.d. observations from a family of distributions $\mathcal{D}$ indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ following equation (5), i.e., $(\mathbf{X}_{i.}, \mathbf{y}_i) \sim D(\boldsymbol{\beta})$ which represents $\mathbf{X}_{i.} \sim \frac{1}{\sqrt{n}}\mathrm{Bernoulli}(0.5)^d, \mathbf{y}_i = \mathrm{sign}(\mathbf{X}_{i.}\boldsymbol{\beta} + \mathbf{e}_i), \mathbf{e}_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \{1 \ldots n\}$ then,*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{D \in \mathcal{D}} \underset{\substack{\boldsymbol{\beta}^* \sim \mathrm{Unif}(\mathcal{F}) \\ (\mathbf{X},\mathbf{y}) \sim D(\boldsymbol{\beta}^*)^n}}{\mathbb{P}} (\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})) \geq \frac{1}{2}$$

where $\hat{\boldsymbol{\beta}} : \mathbb{R}^{n \times d} \times \{-1, +1\}^n \to \mathbb{R}^d$ is any procedure used to infer $\boldsymbol{\beta}^*$, and $\mathcal{D}$ is the family of all distributions indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ with support on $\mathbb{R}^{n \times d} \times \{-1, +1\}^n$.

Our second result provides a bound on the necessary number of samples for approximate signal recovery, which we state formally below.

Theorem 4 (One-bit Compressed Sensing, Approximate Recovery)**.** *There exist a (G, s, g, B)-WGM $\mathbb{M}$ and a finite set of weights ψ such that if nature picks $\boldsymbol{\beta}^*$ uniformly at random from $\mathcal{F} = \{\boldsymbol{\beta} \mid \boldsymbol{\beta}_i = 0, \textit{if } i \notin S, \boldsymbol{\beta}_i \in \psi, \textit{if } i \in S, S \in \mathbb{M}\}$ and draws $n \in o(\log|\mathcal{F}|) = o((s-g)(\log\rho(G) + \log\frac{B}{s-g}) + g\log\frac{d}{g} + (s-g)\log\frac{g}{s-g} + s\log 2)$ i.i.d. observations from family of distributions $\mathcal{D}$ indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ following equation (5), i.e., $(\mathbf{X}_{i.}, \mathbf{y}_i) \sim D(\boldsymbol{\beta})$ which represents $\mathbf{X}_{i.} \sim \frac{1}{\sqrt{n}}\mathrm{Bernoulli}(0.5)^d, \mathbf{y}_i = \mathrm{sign}(\mathbf{X}_{i.}\boldsymbol{\beta} + \mathbf{e}_i), \mathbf{e}_i \sim \mathcal{N}(0, \sigma^2), \forall i \in \{1 \ldots n\}$ then,*

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{D \in \mathcal{D}} \underset{\substack{\boldsymbol{\beta}^* \sim \mathrm{Unif}(\mathcal{F}) \\ (\mathbf{X},\mathbf{y}) \sim D(\boldsymbol{\beta}^*)^n}}{\mathbb{P}} \left(\left\| \frac{\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})}{\|\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|} \right\| \geq \epsilon \right) \geq \frac{1}{2}$$

for some $\epsilon > 0$, where $\hat{\boldsymbol{\beta}} : \mathbb{R}^{n \times d} \times \{-1, +1\}^n \to \mathbb{R}^d$ is any procedure used to infer $\boldsymbol{\beta}^*$, and $\mathcal{D}$ is the family of all distributions indexed by $\boldsymbol{\beta} \in \mathbb{R}^d$ with support on $\mathbb{R}^{n \times d} \times \{-1, +1\}^n$.

Our proof techniques can be applied to prove lower bounds of the sample complexity for several specific sparsity structures as long as one can bound the cardinality of the model. For well-known sparsity structures such that tree-structured sparsity, block sparsity and regular s-sparsity our lower bounds for standard compressed sensing match with respective upper bounds up to a factor of $\log s$. For one-bit compressed sensing, we provide novel lower bounds for tree-structured sparsity and block sparsity. The use of the model-based framework for one-bit compressed sensing remains an open area of research and our information theoretic lower bounds on sample complexity may act as a baseline comparison for the algorithms proposed in the future. In the case of regular s-sparsity, the bound on the sample complexity for one-bit compressed sensing is tight as it matches the current upper bound [6] (See Table 1 and Appendix B for details).

5. REFERENCES

[1] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt, "A nearly-linear time framework for graph-structured sparsity," in *International Conference on Machine Learning*, 2015, pp. 928–937.

[2] Richard G Baraniuk, Volkan Cevher, Marco F Duarte, and Chinmay Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[3] Emmanuel J Candes and Terence Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[4] Mark Rudelson and Roman Vershynin, "Geometric Approach to Error-Correcting Codes and Reconstruction of Signals," *International mathematics research notices*, vol. 2005, no. 64, pp. 4019–4041, 2005.

[5] Martin J Wainwright, "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using L1-Constrained Quadratic Programming (Lasso)," *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[6] Yaniv Plan and Roman Vershynin, "Robust 1-Bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 482–494, 2013.

[7] Ankit Gupta, Robert Nowak, and Benjamin Recht, "Sample Complexity for 1-Bit Compressed Sensing and Sparse Classification," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1553–1557.

[8] Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori, "One-bit compressed sensing: Provable support and vector recovery," in *International Conference on Machine Learning*, 2013, pp. 154–162.

[9] Michael Lustig, David Donoho, and John M Pauly, "Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[10] Petros T Boufounos and Richard G Baraniuk, "1-Bit Compressive Sensing," in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*. IEEE, 2008, pp. 16–21.

[11] Deanna Needell and Joel A Tropp, "CoSamp: Iterative Signal Recovery From Incomplete and Inaccurate Samples," *Communications of the ACM*, vol. 53, no. 12, pp. 93–100, 2010.

[12] Wei Dai and Olgica Milenkovic, "Subspace Pursuit for Compressive Sensing: Closing the Gap Between Performance and Complexity," Tech. Rep., ILLINOIS UNIV AT URBANA-CHAMAPAIGN, 2008.

[13] Sujit Kumar Sahoo and Anamitra Makur, "Signal recovery from random measurements via extended orthogonal matching pursuit," *IEEE Transactions on Signal Processing*, vol. 63, no. 10, pp. 2572–2581, 2015.

[14] Gabriel Peyre, "Best basis compressed sensing," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2613–2622, 2010.

[15] Thomas Blumensath and Mike E Davies, "Iterative Hard Thresholding for Compressed Sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[16] Amin Karbasi, Ali Hormati, Soheil Mohajer, and Martin Vetterli, "Support recovery in compressed sensing: An estimation theoretic approach," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*. IEEE, 2009, pp. 679–683.

[17] Xiao Li, Sameer Pawar, and Kannan Ramchandran, "Sub-linear time compressed sensing using sparse-graph codes," in *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 1645–1649.

[18] Shihao Ji, Ya Xue, and Lawrence Carin, "Bayesian compressive sensing," *IEEE Transactions on signal processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[19] Shihao Ji, David Dunson, and Lawrence Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2008.

[20] Lihan He, Haojun Chen, and Lawrence Carin, "Tree-structured compressive sensing with variational bayesian analysis," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 233–236, 2009.

[21] Adarsh Barik, Jean Honorio, and Mohit Tawarmalani, "Information theoretic limits for linear prediction with graph-structured sparsity," in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 2348–2352.

[22] Martin J Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5728–5741, 2009.

[23] Thomas M Cover and Joy A Thomas, "Elements of Information Theory 2nd Edition," 2006.

[24] Narayana P Santhanam and Martin J Wainwright, "Information-Theoretic Limits of Selecting Binary Graphical Models in High Dimensions," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4117–4134, 2012.

[25] Wei Wang, Martin J Wainwright, and Kannan Ramchandran, "Information-Theoretic Bounds on Model Selection for Gaussian Markov Random Fields," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1373–1377.

[26] John Duchi, "Global fano method," 2016, https://web.stanford.edu/class/stats311/Lectures/lec-06.pdf.

A. PROOF OF THEOREMS 1, 2, 3, 4

Proof. In this section, we prove our main results stated in Section 4.

A.1. Proof Sketch

We use the Markov chain described in subsection 2.3 in our proofs. We assume that nature picks a true s-sparse signal, $\boldsymbol{\beta}^* \in \mathbb{R}^d$, uniformly at random from a family of signals, $\mathcal{F}$. The definition of $\mathcal{F}$ varies according to the specific setups. Specifically, for the noisy case of standard compressed sensing, our restricted ensemble $\mathcal{F}_1$ is defined as:

$$\begin{aligned}\mathcal{F}_1 = \{&\boldsymbol{\beta} \mid S \in \mathbb{M} \text{ such that } \boldsymbol{\beta}_i = 0, \text{ if } i \notin S, \boldsymbol{\beta}_i \in \\ &\left\{ \left\{ \frac{C_0\sigma\sqrt{n}}{\sqrt{2(1-\epsilon)}}, \frac{C_0\sigma\sqrt{n}}{\sqrt{2(1-\epsilon)}} + \frac{C_0\sigma\sqrt{n}}{\sqrt{(1-\epsilon)}} \right\}, \text{ if } i \in S \right\}\end{aligned} \tag{6}$$

for some $0 < \epsilon < 1$ and $\mathbb{M}$ is as in Definition 1 in our restricted (G, s, g, B)-WGM. For the noiseless case, we simplify our ensemble as follows:

$$\begin{aligned}\mathcal{F}_2 = \{&\boldsymbol{\beta} \mid S \in \mathbb{M} \text{ such that } \boldsymbol{\beta}_i = 0, \text{ if } i \notin S, \ \boldsymbol{\beta}_i \in \{-1, 1\}, \\ &\text{if } i \in S\} \ ,\end{aligned} \tag{7}$$

where $\mathbb{M}$ is as in Definition 1 in our restricted (G, s, g, B)-WGM and for one-bit compressed sensing, we define our ensemble in the following way:

$$\begin{aligned}\mathcal{F}_3 = &\{\boldsymbol{\beta} \mid S \in \mathbb{M} \text{ such that } \boldsymbol{\beta}_i = \begin{cases} -\epsilon, \text{ if } i \in A \\ \sqrt{\frac{2}{s}} + \epsilon, \text{ if } i \in B \\ 0, \text{ otherwise} \end{cases} \\ &, |A| = |B|, S = A \cup B, A \cap B = \phi\},\end{aligned} \tag{8}$$

for some $\epsilon > 0$ and $\mathbb{M}$ is as in Definition 1 in our restricted (G, s, g, B)-WGM.

Nature then generates independent and identically distributed samples using the true $\boldsymbol{\beta}^*$. These samples are of the form $(\mathbf{X}, \mathbf{y} = f(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}))$. We choose an appropriate f and noise $\mathbf{e}$ for the specific setups under analysis. Similarly, the design matrix $\mathbf{X}$ also varies according to the specific settings. Although, we note that choice of $\boldsymbol{\beta}^*$ and $\mathbf{X}$ are marginally independent in all the settings. Next, we obtain an upper bound on the mutual information between the true signal $\boldsymbol{\beta}^*$ and the observations $(\mathbf{X}, \mathbf{y})$ using the following results. We provide three lemmas, one for each of the restricted ensembles defined in equations (6), (7) and (8). First, we analyze noisy case of standard compressed sensing.

Lemma 1. *Let $\boldsymbol{\beta}^*$ be chosen uniformly at random from $\mathcal{F}_1$ defined in equation* (6). *Then,* $\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \leq \frac{n}{2} \log(1 + \frac{s}{2}k_1^2 + \frac{s}{2}k_2^2 - \frac{s^2}{d}\frac{k_1^2}{4} - \frac{s^2}{d}\frac{k_2^2}{4} - \frac{s^2}{d}\frac{k_1 k_2}{2})$*, for some constants k_1 and k_2 independent of n, d and s and where $(\mathbf{X}, \mathbf{y})$ is generated using the noisy setup defined in equation* (4).

(See Appendix D for the detailed proof.)

Next, we analyze the noiseless case of standard compressed sensing.

Lemma 2. *Let $\boldsymbol{\beta}^*$ be chosen uniformly at random from $\mathcal{F}_2$ defined in equation* (7). *Then,* $\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \leq 3n \log \frac{es}{27}$ *, where $(\mathbf{X}, \mathbf{y})$ is generated using the noiseless setup, i.e., $\mathbf{e} = \mathbf{0}$ defined in equation* (4).

(See Appendix E for the detailed proof.)

The following lemmas provide an upper bound on the mutual information between the true signal $\boldsymbol{\beta}^*$ and observed samples $(\mathbf{X}, \mathbf{y} = \mathrm{sign}(\mathbf{X}\boldsymbol{\beta}^* + \mathbf{e}))$ as in one-bit compressed sensing.

Lemma 3. *Let $\boldsymbol{\beta}^*$ be chosen uniformly at random from $\mathcal{F}_3$ defined in equation* (8). *Then,* $\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \leq 2n \log 2$ *, where $(\mathbf{X}, \mathbf{y})$ is generated using the setup defined in equation* (5) *for either the noisy or noiseless case.*

(See Appendix F for the detailed proof.)

Finally, we use Fano's inequality [23] to prove our main results. We explain each of these steps in detail in Appendix A.

A.2. Restricted ensemble

First, we need to construct a weighted graph G to define our family of sparse signals $\mathcal{F}$. Our construction of weighted graph follows [21]. We construct the underlying graph G for the WGM using the following steps:

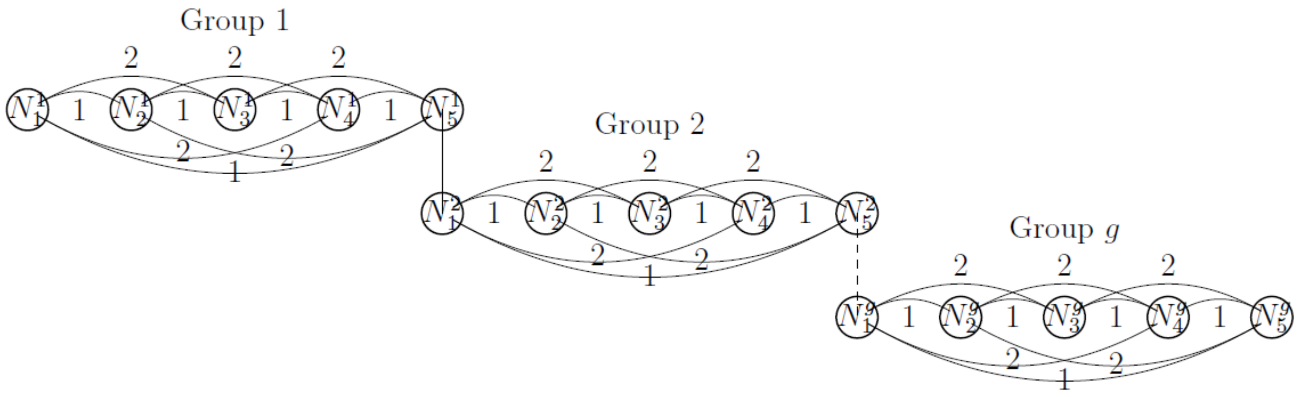


Fig. 1: An example of constructing an underlying graph for $\rho(G) = 2$ and $\frac{B}{s-g} = 2$

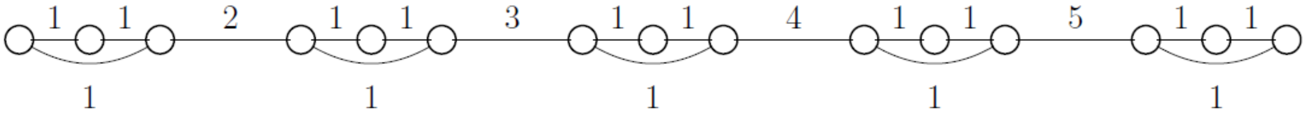


Fig. 2: An example of an underlying graph G for $(G, s, g, B) - WGM$ with parameters $d = 15, s = 10, g = 5, B = 5, \rho(G) = 2$

- We split d nodes equally into g groups with each group having $\frac{d}{g}$ nodes.
- For each group j, we denote a node by N_i^j where j is the group index and i is the node index. Each group j, contains nodes from N_1^j to $N_{\frac{d}{g}}^j$.
- We allow for circular indexing within a group, i.e., for any group j, a node N_i^j is equivalent to node $N^j_{i \bmod \frac{d}{g}}$.
- For each $p = 1, \ldots, \frac{B}{s-g}$, node N_i^j has an edge with nodes $N^j_{i+(p-1)\frac{\rho(G)}{2}+1}$ to $N^j_{i+p\frac{\rho(G)}{2}}$ with weight p.
- Cross edges between nodes in two different groups are allowed as long as edge weights are greater than $\frac{B}{s-g}$ and this does not affect $\rho(G)$.

Figure 1 shows an example of a graph constructed using the above steps. Furthermore, the parameters of our WGM satisfy the following requirements:

R1 $\frac{d}{g} \geq \frac{\rho(G)B}{s-g} + 1$,

R2 $\frac{\rho(G)B}{2(s-g)} \geq \frac{s}{g} - 1$,

R3 $B \geq s - g$.

These are quite mild requirements on the parameters and are easy to be fulfilled. To that end, we state the following proposition from [21] (full version).

Proposition 1 (Proposition 2 in [21])**.** *Given any value of s, g that satisfy R3 (i.e., $B \geq s - g$), there are infinitely many choices for $\rho(G)$ and d that satisfy R1 and R2 and hence, there are infinitely many (G, s, g, B)-WGM which follow our construction.*

Figure 2 shows one graph which follows our construction and additionally fulfills R1, R2 and R3. Now that we have defined the underlying weighted graph G for our WGM, we next define the possible coefficients for the true signal in our (G, s, g, B)-WGM. We define the restricted ensemble for each setting in a different fashion. In particular, restricted ensemble for noisy standard compressed sensing $\mathcal{F}_1$ is defined in equation 6, restricted ensemble for noiseless standard compressed sensing $\mathcal{F}_2$ is defined in equation 7 and restricted ensemble for one-bit compressed sensing $\mathcal{F}_3$ is defined in equation 8.

Next, we count the number of elements in $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$. We provide the results in the following lemmas.

Lemma 4 (Cardinality of Restricted Ensemble for Standard Compressed Sensing - See equation (13) in [21])**.** *For any $\mathcal{F} \in \{\mathcal{F}_1, \mathcal{F}_2\}$ as defined in equations* (6), (7)*, we have that $|\mathcal{F}| \geq 2^s (\frac{d}{g})^g (\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)}$.*

Lemma 5 (Cardinality of Restricted Ensemble for One-bit Compressed Sensing)**.** *For $\mathcal{F}_3$ as defined in equation* (8)*, $|\mathcal{F}_3| \geq 2^{\frac{s}{2}} (\frac{d}{g})^g (\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)}$.*

Proof. The proof follows a similar approach as in Lemma 4. We use the following counting argument:

1. We choose one node from each of the g groups in the underlying graph G to be the root of a connected component. Each group has $\frac{d}{g}$ possible candidates for being the root and hence we can choose them in $(\frac{d}{g})^g$ possible ways.

2. Since we are interested only in establishing a lower bound on $|\mathcal{F}|$, we only consider the cases where each connected component has $\frac{s}{g}$ nodes. Given a root node N_i^j in group j, we choose the remaining $\frac{s}{g} - 1$ nodes to be connected to the root only from nodes N_{i+1}^j to $N_{i+\frac{B\rho(G)}{2(s-g)}}^j$ (using circular indices if needed). Note that at least till the last $\frac{\rho(G)B}{2(s-g)}$ nodes, we always include node N_i^j and we never include $N_r^j, r \leq i-1$ in our selection. Furthermore, R1 guarantees that we have enough nodes to avoid any possible repetitions due to circular indices for the last $\frac{\rho(G)B}{2(s-g)}$ nodes and R2 ensures that we have enough nodes to form a connected component. This guarantees that all the supports are unique. Hence, given a root node N_i^j we have $\binom{\frac{\rho(G)B}{2(s-g)}}{\frac{s}{g}-1}$ choices. . Finally, considering all the groups, the number of choices is $(\binom{\frac{\rho(G)B}{2(s-g)}}{\frac{s}{g}-1})^g$.

3. We choose $\frac{s}{2}$ entries in the support of $\boldsymbol{\beta}$ in $\binom{s}{\frac{s}{2}}$ ways.

Steps 1 and 2 are borrowed from [21] and written here for completion. Thus,

$$\begin{aligned}
|\mathcal{F}_3| &\geq \binom{s}{\frac{s}{2}} \frac{d^g}{g^g} \binom{\frac{\rho(G)B}{2(s-g)}}{\frac{s}{g}-1}^g \\
&\geq 2^{\frac{s}{2}} \frac{d^g}{g^g} \left(\frac{\rho(G)Bg}{2(s-g)^2} \right)^{s-g} .
\end{aligned}$$

□

A.3. Bound on Mutual Information

We utilize results from three lemmas, Lemma 1, 2 and 3, one for each of the restricted ensembles defined in equations (6), (7) and (8) to get an upper bound on the mutual information between the s-sparse signal $\boldsymbol{\beta}^*$ and the observations $(\mathbf{X}, \mathbf{y})$.

A.4. Bound on the Inference Error

In this subsection, we analyze the inference error using the results from the previous sections and Fano's inequality. If nature chooses $\boldsymbol{\beta}^*$ from a restricted ensemble $\mathcal{F}$ uniformly at random then for the Markov chain described in Section 2, Fano's inequality [23] can be written as

$$\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq 1 - \frac{\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) + \log 2}{\log |\mathcal{F}|} . \tag{9}$$

Since, we have already established bounds on $\mathrm{I}(\boldsymbol{\beta}^*, (\mathbf{X}, \mathbf{y}))$ and $\log |\mathcal{F}|$, equation (9) readily provides a bound on the error of exact recovery. We can use this directly to prove some of our theorems.

A.4.1. Proof of Theorem 2 - Noiseless Case in Standard Compressed Sensing

Using the results from Lemmas 2 and subsection A.2 along with equation (9), we get:

$$\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq 1 - \frac{3n \log \frac{es}{27} + \log 2}{\log \left(2^s (\frac{d}{g})^g (\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)}\right)}$$

It follows that $\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq \frac{1}{2}$ as long as $n \leq \frac{1}{6} \frac{(s-g)(\log \frac{\rho(G)}{2} + \log \frac{B}{s-g}) + g \log \frac{d}{g} + (s-g) \log \frac{g}{s-g} + s \log 2}{\log \frac{es}{27}} - \frac{\log 2}{3 \log \frac{es}{27}}$. This proves Theorem 2.

A.4.2. Proof of Theorem 3 - Exact Recovery in One-bit Compressed Sensing

Using the results from Lemmas 5 and 3 along with equation (9), we get:

$$\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq 1 - \frac{2n \log 2 + \log 2}{\log \left(2^{\frac{s}{2}} (\frac{d}{g})^g (\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)}\right)}$$

It follows that $\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq \frac{1}{2}$ as long as $n \leq \frac{1}{2} \frac{(s-g)(\log \frac{\rho(G)}{2} + \log \frac{B}{s-g}) + g \log \frac{d}{g} + (s-g) \log \frac{g}{s-g} + \frac{s}{2} \log 2}{2 \log 2} - \frac{1}{2}$. This proves Theorem 3.

Now we prove the remaining results, which require additional arguments.

A.4.3. Proof of Theorem 1 - Noisy Case in Standard Compressed Sensing

For noisy setups as described in equation (4), we are interested in the error bound in terms of the ℓ_2-norm. We obtain this bound by using the following rule:

$$\begin{aligned} \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq C\|\mathbf{e}\|) &\geq \mathbb{P}(\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\| \geq C\|\mathbf{e}\|, \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \\ &= \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq C\|\mathbf{e}\| \mid \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*)\, \mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \end{aligned} \tag{10}$$

We bound the terms $\mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq C\|\mathbf{e}\| \mid \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*)$ and $\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*)$ separately.

Lemma 6 (See Lemma 1, 2 and Corollary 1 in [21]). *If $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$ come from a family of signals $\mathcal{F}_1$ as defined in equation* (6) *then*

1. *For some $C_0 \geq C > 0$*

$$\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\| \leq \frac{C_0 \sigma \sqrt{n}}{\sqrt{(1-\epsilon)}} \iff \boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}\,. \tag{11}$$

2. *For some $0 < \epsilon < 1$,*

$$\mathbb{P}\left(\|\mathbf{e}\|^2 \leq \sigma^2 \frac{n}{1-\epsilon}\right) \geq 1 - \exp\left(-\frac{\epsilon^2 n}{4}\right). \tag{12}$$

3. *If the above two claims hold then,*

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq C\|\mathbf{e}\| \mid \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*\right) \geq 1 - \exp\left(-\frac{\epsilon^2 n}{4}\right). \tag{13}$$

Using the results from Lemmas 1 and 4 along with equation (9), we get:

$$\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq 1 - \frac{\frac{n}{2} \log(1 + \frac{s}{2} k_1^2 + \frac{s}{2} k_2^2 - \frac{s^2}{d} \frac{k_1^2}{4} - \frac{s^2}{d} \frac{k_2^2}{4} - \frac{s^2}{d} \frac{k_1 k_2}{2}) + \log 2}{\log \left(2^s (\frac{d}{g})^g (\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)}\right)}$$

It follows that $\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq \frac{1}{2}$ as long as $n \leq \frac{(s-g)(\log \frac{\rho(G)}{2}+\log \frac{B}{s-g})+g\log \frac{d}{g}+(s-g)\log \frac{g}{s-g}+s\log 2}{\log(1+\frac{s}{2}k_1^2+\frac{s}{2}k_2^2-\frac{s^2}{d}\frac{k_1^2}{4}-\frac{s^2}{d}\frac{k_2^2}{4}-\frac{s^2}{d}\frac{k_1k_2}{2})} - \frac{\log 2}{\log(1+\frac{s}{2}k_1^2+\frac{s}{2}k_2^2-\frac{s^2}{d}\frac{k_1^2}{4}-\frac{s^2}{d}\frac{k_2^2}{4}-\frac{s^2}{d}\frac{k_1k_2}{2})}$. Now let $n \in \tilde{o}((s-g)(\log \rho(G) + \log \frac{B}{s-g}) + g \log \frac{d}{g} + (s-g)\log \frac{g}{s-g} + s\log 2)$ so that $\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \geq \frac{1}{2}$. Using result from Lemma 6 and equation (10) we can write

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq C\|\mathbf{e}\|\right) \geq \left(1 - \exp\left(-\frac{\epsilon^2 n}{4}\right)\right)\frac{1}{2} \,.$$

We know that $n \geq 1$ and if we choose $\epsilon \geq \sqrt{-4\log 0.8} \sim 0.9448$, then we can write inequality (10) as,

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq C\|\mathbf{e}\|\right) \geq \frac{1}{10} \,.$$

This completes the proof of Theorem 1.

A.4.4. Proof of Theorem 4 - Approximate Recovery in One-bit Compressed Sensing

First we note that,

$$\|\boldsymbol{\beta}\| = 1 \,, \forall \boldsymbol{\beta} \in \mathcal{F}_3,$$

and consequently,

$$\left\| \frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|} \right\| = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \,.$$

Therefore,

$$\begin{aligned} \mathbb{P}\left(\left\| \frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|} \right\| \geq \epsilon \right) &= \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq \epsilon) \\ &\geq \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq \epsilon, \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \\ &= \mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq \epsilon \mid \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*)\,\mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) \end{aligned} \tag{14}$$

Next we prove the following lemma.

Lemma 7. *If $\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^* \in \mathcal{F}_3$ then $\mathbb{P}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \geq \epsilon \mid \hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*) = 1$.*

Proof. We prove that two arbitrarily chosen $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ such that $\boldsymbol{\beta}^1, \boldsymbol{\beta}^2 \in \mathcal{F}_3$ as defined in equation (8) and $\boldsymbol{\beta}^1 \neq \boldsymbol{\beta}^2$ then $\|\boldsymbol{\beta}^1 - \boldsymbol{\beta}^2\| \geq \epsilon$.

$\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ have the same support. Since we assume that $\boldsymbol{\beta}^1 \neq \boldsymbol{\beta}^2$, both vectors must differ in at least one coefficient on their support. Let i be such a coefficient. Then,

$$\begin{aligned} \|\boldsymbol{\beta}^1 - \boldsymbol{\beta}^2\| &\geq |\boldsymbol{\beta}^1_i - \boldsymbol{\beta}^2_i| \\ &= \left| \sqrt{\frac{2}{s}} + 2\epsilon \right| \\ &\geq \epsilon \end{aligned}$$

$\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ have different supports When $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ have different supports then we can always find i and j such that $i \in S_1, i \notin S_2$ and $j \notin S_1, j \in S_2$ where S_1 and S_2 are supports of $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$ respectively. Then,

$$\begin{aligned} \|\boldsymbol{\beta}^1 - \boldsymbol{\beta}^2\| &\geq \sqrt{(\boldsymbol{\beta}^1_i)^2 + (\boldsymbol{\beta}^2_j)^2} \\ &\geq \sqrt{\epsilon^2 + \epsilon^2} \\ &\geq \epsilon \,. \end{aligned}$$

Since the above is true for any two arbitrarily chosen $\boldsymbol{\beta}^1$ and $\boldsymbol{\beta}^2$, this holds for $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$ as well. This proves the lemma. □

Substituting results from Lemma 7 into equation (14), we get,

$$\mathbb{P}\left(\left\|\frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|}\right\| \geq \epsilon\right) = \mathbb{P}(\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}^*)$$

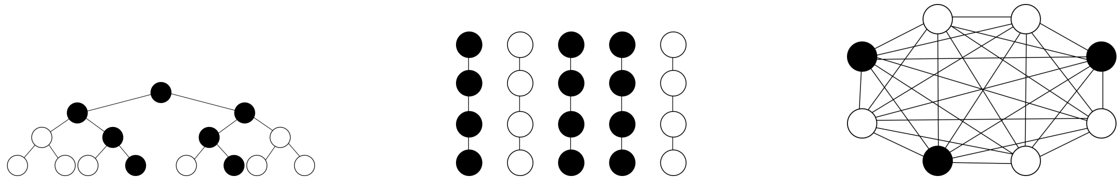
Using the results from Lemmas 3 and 5 along with equation (9), we get:

$$\mathbb{P}\left(\left\|\frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|}\right\| \geq \epsilon\right) \geq 1 - \frac{2n\log 2 + \log 2}{\log\left(2^{\frac{s}{2}}(\frac{d}{g})^g(\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)}\right)}$$

It follows that as long as $n \leq \frac{1}{2}\frac{(s-g)(\log\frac{\rho(G)}{2}+\log\frac{B}{s-g})+g\log\frac{d}{g}+(s-g)\log\frac{g}{s-g}+\frac{s}{2}\log 2}{2\log 2} - \frac{1}{2}$, we get $\mathbb{P}\left(\left\|\frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|}\right\| \geq \epsilon\right) \geq \frac{1}{2}$. This completes the proof of Theorem 4. □

B. SPECIFIC EXAMPLES

Our proof techniques can be applied to prove lower bounds of the sample complexity for several specific sparsity structures as long as one can bound the cardinality of the model. Below, we provide information theoretic lower bounds on sample complexity for some well-known sparsity structures.



(a) Tree-structured sparsity

(b) Block sparsity

(c) Regular s-sparsity

Fig. 3: Different sparsity structures as weighted graph models: nodes are variables, black nodes are selected variables, each edge has unit weight. Tree-structured sparsity: $d = 15$ nodes, $s = 7$ selected nodes, $g = 1$ connected component, $B = s - g = 6$. Block sparsity: $N = 5$ columns, $K = 3$ selected columns, $J = 4$ rows, $d = JN = 20$ nodes, $s = JK = 12$ selected nodes, $g = 3$ connected components, $B = s - g = 9$ Regular s-sparsity: $d = 8$ nodes, $s = 3$ selected nodes, $g = 1$ connected component, $B = s - g = 2$

B.1. Tree-structured sparsity model

The tree-sparsity model [2] is used in several applications such as wavelet decomposition of piecewise smooth signals and images. In this model, one assumes that the coefficients of the $s-$sparse signal form a k-ary tree and the support of the sparse signal form a rooted and connected sub-tree on s nodes in this $k-$ary tree. The arrangement is such that if a node is part of this subtree then the parent of such node is also included in the subtree. Let T be a rooted binary tree on $[d]$ nodes. For any node i, $\pi_T(i)$ denotes the parent of node i in T. Then, a tree-structured sparsity model $\mathbb{M}_{\text{tree}}$ on the tree T is the set of supports defined as

$$\mathbb{M}_{\text{tree}} = \{S \subseteq [d] \mid |S| = s, \text{ and } \forall i \in S,\ \pi_T(i) \in S\}. \tag{15}$$

The following corollary provides information theoretic lower bounds on the sample complexity.

Corollary 1. *There exists a binary tree-structured sparsity model, such that*

1. *If $n \in \tilde{o}(s)$ then $\mathbb{P}(\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\| \geq C\|e\|) \geq \frac{1}{10}$ for noisy standard compressed sensing.*
2. *If $n \in \tilde{o}(s)$ then $\mathbb{P}(\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}) \geq \frac{1}{2}$ for noiseless standard compressed sensing.*
3. *If $n \in o(s)$ then $\mathbb{P}\left(\|\frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|}\| \geq \epsilon\right) \geq \frac{1}{2}$ for one-bit compressed sensing.*
4. *If $n \in o(s)$ then $\mathbb{P}(\boldsymbol{\beta} \hat{\neq} \boldsymbol{\beta}^*) \geq \frac{1}{2}$ for one-bit compressed sensing.*

(See Appendix C for the detailed proof.)

B.2. Block sparsity model

In the block sparsity model [2], an $s-$sparse signal, $\boldsymbol{\beta} \in \mathbb{R}^{J \times N}$, i.e., $d = JN$ can be represented as a matrix with J rows and N columns. The index ij denotes the index of the entry of $\boldsymbol{\beta}$ at i^{th} row and j^{th} column. The support of $\boldsymbol{\beta}$ comes from K columns of this matrix such that $s = J \times K$. More precisely, a block sparsity model $\mathbb{M}_{\text{block}}$ is a set of supports defined as

$$\begin{aligned} \mathbb{M}_{\text{block}} = &\{S \subseteq [J \times N] \mid \forall i \in [J], j \in L, ij \in S \text{ and} \\ &L \subseteq \{1, \dots, N\}, \ |L| = K\} \ . \end{aligned} \tag{16}$$

The above can be modeled as a weighted graph model. In particular, we can construct a graph G over all the entries in $\boldsymbol{\beta}$ by treating nodes in the column of the matrix as connected nodes (see Figure 3b). The following corollary provides information theoretic lower bounds on the sample complexity.

Corollary 2. *There exists a block structured sparsity model, such that*

1. *If* $n \in \tilde{o}(KJ + K \log \frac{N}{K})$ *then* $\mathbb{P}(\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\| \geq C\|e\|) \geq \frac{1}{10}$ *for noisy standard compressed sensing.*
2. *If* $n \in \tilde{o}(KJ + K \log \frac{N}{K})$ *then* $\mathbb{P}(\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}) \geq \frac{1}{2}$ *for noiseless standard compressed sensing.*
3. *If* $n \in o(KJ + K \log \frac{N}{K})$ *then* $\mathbb{P}\left(\|\frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|}\| \geq \epsilon\right) \geq \frac{1}{2}$ *for one-bit compressed sensing.*
4. *If* $n \in o(KJ + K \log \frac{N}{K})$ *then* $\mathbb{P}(\boldsymbol{\beta} \hat{\neq} \boldsymbol{\beta}^*) \geq \frac{1}{2}$ *for one-bit compressed sensing.*

(See Appendix C for the detailed proof.)

B.3. Regular s-sparsity model

When the model does not have any additional structure besides sparsity, we call it a regular s-sparsity model. That is, a regular s-sparsity model $\mathbb{M}_{\text{regular}}$ is a set of supports defined as

$$\mathbb{M}_{\text{regular}} = \{S \subseteq [d] | |S| = s\} \tag{17}$$

The following corollary provides information theoretic lower bounds on the sample complexity.

Corollary 3. *There exists a regular s-sparsity model, such that*

1. *If* $n \in \tilde{o}(s \log \frac{d}{s})$ *then* $\mathbb{P}(\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\| \geq C\|e\|) \geq \frac{1}{10}$ *for noisy standard compressed sensing.*
2. *If* $n \in \tilde{o}(s \log \frac{d}{s})$ *then* $\mathbb{P}(\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}) \geq \frac{1}{2}$ *for noiseless standard compressed sensing.*
3. *If* $n \in o(s \log \frac{d}{s})$ *then* $\mathbb{P}\left(\|\frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|} - \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}^*\|}\| \geq \epsilon\right) \geq \frac{1}{2}$ *for one-bit compressed sensing.*
4. *If* $n \in o(s \log \frac{d}{s})$ *then* $\mathbb{P}(\boldsymbol{\beta} \hat{\neq} \boldsymbol{\beta}^*) \geq \frac{1}{2}$ *for one-bit compressed sensing.*

(See Appendix C for the detailed proof.)

C. PROOF OF COROLLARIES 1, 2 AND 3

Until now, the restricted ensembles used in our proofs are defined on a general weighted graph model. These proofs can be instantiated directly to commonly used sparsity structures such as tree structured sparsity, block sparsity and regular s-sparsity by defining $\mathcal{F}_1, \mathcal{F}_2$ and $\mathcal{F}_3$ defined in equations (6), (7), (8), on the models $\mathbb{M}_{\text{tree}}, \mathbb{M}_{\text{block}}$ and $\mathbb{M}_{\text{regular}}$ defined in equations (15), (16) and (17). We only need to bound the cardinality of these restricted ensembles to extend our proofs to specific sparsity structures.

Proof of Corollary 1. We take a restricted ensemble $\mathcal{F} \in \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3\}$ which is defined on $\mathbb{M}_{\text{tree}}$. For such a restricted ensemble, we have that $\log |\mathcal{F}| \geq cs$ for an absolute constant $c > 0$. This follows from the fact that we have at least 2^s different choices of $\boldsymbol{\beta}^*$ for standard compressed sensing and $2^{\frac{s}{2}}$ for one-bit compressed sensing. Using the below results from [2], we note that this is not a weak bound as the upper bound is of the same order, i.e.,

$$|\mathcal{F}| \leq \begin{cases} 2^s \frac{2^{2s+8}}{se^2}, s \geq \log d \\ 2^s \frac{(2e)^s}{s+1}, s < \log d \end{cases} , \forall \mathcal{F} \in \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3\}$$

From the above and using Theorems 1, 2, 3, and 4, we prove our claim.

Proof of Corollary 2. We take a restricted ensemble $\mathcal{F} \in \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3\}$ which is defined on $\mathbb{M}_{\text{block}}$. For such a restricted ensemble, one can bound $|\mathcal{F}|$ by choosing K connected components from N. It is easy to see that the number of possible signals in this model $\mathcal{F}$, would be, $|\mathcal{F}| \geq 2^{\frac{KJ}{2}} \binom{N}{K} \geq 2^{\frac{KJ}{2}} (\frac{N}{K})^K$. Given this and Theorems 1, 2, 3, and 4, we prove our claim.

Proof of Corollary 3. In this case, we take a restricted ensemble $\mathcal{F} \in \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3\}$ which is defined on $\mathbb{M}_{\text{regular}}$. We can bound the cardinality of such a restricted ensemble $\mathcal{F}$ in the following way.

$$\begin{aligned} |\mathcal{F}| &\geq 2^{\frac{s}{2}} \binom{d}{s}, \quad \forall \mathcal{F} \in \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3\} \\ &\geq 2^{\frac{s}{2}} \frac{d^s}{s^s} \end{aligned}$$

Given the above and Theorems 1, 2, 3, and 4, we prove our claim.

D. PROOF OF LEMMA 1

Proof. Note that $\mathbf{y}_i = \mathbf{X}_{i.} \boldsymbol{\beta} + \mathbf{e}_i, \forall i \in \{1, \dots, n\}$. Furthermore, in the noisy setup we choose the Gaussian design matrix. That is, each $\mathbf{X}_{ij}, \forall i \in [n], j \in [d]$ is drawn independently from $\mathcal{N}(0, \frac{1}{n})$. First, we show that for a given $\boldsymbol{\beta}$, $(\mathbf{X}_{i.}, \mathbf{y}_i) \in \mathbb{R}^{d+1}, \forall i \in \{1, \dots, n\}$ follows multivariate normal distribution $\mathcal{P}_{(\mathbf{X}_i, \mathbf{y}_i)|\boldsymbol{\beta}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ where,

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \begin{bmatrix} \frac{1}{n} & 0 & \dots & \frac{\boldsymbol{\beta}_1}{n} \\ 0 & \frac{1}{n} & \dots & \frac{\boldsymbol{\beta}_2}{n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\boldsymbol{\beta}_1}{n} & \frac{\boldsymbol{\beta}_2}{n} & \dots & \frac{\|\boldsymbol{\beta}\|_2^2}{n} + \sigma^2 \end{bmatrix} \end{aligned} \tag{18}$$

We denote the distribution over all the samples as $\mathcal{P}_{(\mathbf{X}, \mathbf{y})|\boldsymbol{\beta}}$. It can be easily verified that $\mathbf{a}^\intercal (\mathbf{X}_{i.}, \mathbf{y}_i)$ follows normal distribution for any given $\mathbf{a} \in \mathbb{R}^{d+1}$. This implies that $(\mathbf{X}_i, \mathbf{y}_i)$ follows a multivariate normal distribution. Second, since each $\mathbf{X}_{ij} \sim \mathcal{N}(0, \frac{1}{n})$ and $\mathbf{e}_i \sim \mathcal{N}(0, \sigma^2)$, thus $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}_i, \mathbf{y}_i) = \mathbf{0} \in \mathbb{R}^{d+1}$. To compute the covariance matrix, we recall that $\mathbf{X}_{ij}$ are independently distributed. Therefore $\text{Cov}(\mathbf{X}_{ij}, \mathbf{X}_{ik}) = 0, \forall j \neq k$. Thus,

$$\begin{aligned} \text{Cov}(\mathbf{X}_{ij}, \mathbf{y}_i) = \text{Cov}(\mathbf{X}_{ij}, \sum_{j=1}^{d} \boldsymbol{\beta}_j \mathbf{X}_{ij} + \mathbf{e}_i) &= \frac{\boldsymbol{\beta}_j}{n} \\ \text{Cov}(\mathbf{y}_i, \mathbf{y}_i) &= \frac{\|\boldsymbol{\beta}\|_2^2}{n} + \sigma^2 \end{aligned}$$

Note that for any arbitrary distribution $\mathcal{Q}$ over $(\mathbf{X}, \mathbf{y})$, the following inequality holds (See equation 5.1.4 in [26]).

$$\text{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \leq \frac{1}{|\mathcal{F}_1|} \sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \text{KL}(\mathcal{P}_{(\mathbf{X}, \mathbf{y})|\boldsymbol{\beta}} \| \mathcal{Q}) \tag{19}$$

We choose a $\mathcal{Q}$ which decomposes in the following way:

$$\mathcal{Q} = Q^n$$

Using the independence of samples and factorization of $\mathcal{Q}$, we can write equation (19) as,

$$\text{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \leq \frac{n}{|\mathcal{F}_1|} \sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \text{KL}(\mathcal{P}_{(\mathbf{X}_i, \mathbf{y}_i)|\boldsymbol{\beta}} \| Q) \tag{20}$$

Recall that $(\mathbf{X}_i, \mathbf{y}_i)|\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \forall \boldsymbol{\beta} \in \mathcal{F}_1$ where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is computed according to equation (18). Let $Q \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. By the KL divergence between two multivariate normal distributions, we can write equation (20) as,

$$\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \le \frac{n}{2} \frac{1}{|\mathcal{F}_1|} \sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \left(\log \frac{\det(\boldsymbol{\Sigma})}{\det(\boldsymbol{\Sigma}_{\boldsymbol{\beta}})} - d - 1 + \mathrm{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \right) \tag{21}$$

We then choose the covariance matrix $\boldsymbol{\Sigma}$ which minimizes equation (21).

$$\boldsymbol{\Sigma} = \arg\min_{\boldsymbol{\Sigma}} \frac{n}{2} \frac{1}{|\mathcal{F}_1|} \sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \left(\log \frac{\det(\boldsymbol{\Sigma})}{\det(\boldsymbol{\Sigma}_{\boldsymbol{\beta}})} - d - 1 + \mathrm{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \right)$$

We solve the above equation for a positive definite covariance matrix $\boldsymbol{\Sigma}$. This can be easily done by taking the derivative of the equation and equating it to zero. That is,

$$\sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{\Sigma}^{-1} = \mathbf{0}$$

$$\boldsymbol{\Sigma} \left(\sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{\Sigma}^{-1} \right) \boldsymbol{\Sigma} = \mathbf{0}$$

and therefore:

$$\boldsymbol{\Sigma} = \frac{1}{|\mathcal{F}_1|} \sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \tag{22}$$

Substituting value of $\boldsymbol{\Sigma}$ from equation (22) to equation (21), we get:

$$\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \le \frac{n}{2} \left(\log\det \left(\frac{1}{|\mathcal{F}_1|} \sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \right) - \frac{1}{|\mathcal{F}_1|} \sum_{\boldsymbol{\beta} \in \mathcal{F}_1} \log\det(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \right) \tag{23}$$

Next, we compute the determinant of the above covariance matrices.

Computing determinant of covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. Note that for a block matrix,

$$\det \left(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \right) = \det(\mathbf{A}) \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}) \tag{24}$$

provided that $\mathbf{A}$ is invertible. Note that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ where,

$$\begin{aligned} \mathbf{A} &= \mathrm{diag}(\frac{1}{n}) \in \mathbb{R}^{d \times d} \\ \mathbf{B} &= \begin{bmatrix} \frac{\boldsymbol{\beta}_1}{n} \\ \vdots \\ \frac{\boldsymbol{\beta}_d}{n} \end{bmatrix} \in \mathbb{R}^{d \times 1} \\ \mathbf{C} &= \mathbf{B}^\intercal \in \mathbb{R}^{1 \times d} \\ \mathbf{D} &= \frac{\|\boldsymbol{\beta}\|_2^2}{n} + \sigma^2 \in \mathbb{R}^{1 \times 1} \end{aligned}$$

Using equation (24), it follows that,

$$\begin{aligned} \det(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) &= \frac{1}{n^d} \left(\frac{\|\boldsymbol{\beta}\|_2^2}{n} + \sigma^2 - \frac{\|\boldsymbol{\beta}\|_2^2}{n} \right) \\ &= \frac{\sigma^2}{n^d} \end{aligned}$$

We can simplify equation (23):

$$\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \le \frac{n}{2}\left(\log\det\left(\frac{1}{|\mathcal{F}_1|}\sum_{\boldsymbol{\beta}\in\mathcal{F}_1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) - \log\frac{\sigma^2}{n^d}\right) \tag{25}$$

Computing determinant of covariance matrix $\boldsymbol{\Sigma}$. Now note that,

$$\boldsymbol{\Sigma} = \frac{1}{|\mathcal{F}_1|}\sum_{\boldsymbol{\beta}\in\mathcal{F}_1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \begin{bmatrix} \frac{1}{n} & 0 & \dots & \frac{\bar{\boldsymbol{\beta}}_1}{n} \\ 0 & \frac{1}{n} & \dots & \frac{\bar{\boldsymbol{\beta}}_2}{n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\bar{\boldsymbol{\beta}}_1}{n} & \frac{\bar{\boldsymbol{\beta}}_2}{n} & \dots & \frac{\sum_{\boldsymbol{\beta}\in\mathcal{F}_1}\|\boldsymbol{\beta}\|_2^2}{|\mathcal{F}_1|n} + \sigma^2 \end{bmatrix}$$

where $\bar{\boldsymbol{\beta}}_i = \frac{\sum_{\boldsymbol{\beta}\in\mathcal{F}_1}\boldsymbol{\beta}_i}{|\mathcal{F}_1|}, \forall i \in [d]$. Using the same approach as equation (24), we can compute the determinant of $\boldsymbol{\Sigma}$:

$$\det(\boldsymbol{\Sigma}) = \frac{\sigma^2 + \frac{\sum_{\boldsymbol{\beta}\in\mathcal{F}_1}\|\boldsymbol{\beta}\|_2^2}{|\mathcal{F}_1|n} - \frac{\sum_{i=1}^d \bar{\boldsymbol{\beta}}_i^2}{n}}{n^d} \tag{26}$$

Each $\boldsymbol{\beta}_i \in \{k_1\sigma\sqrt{n},\ k_2\sigma\sqrt{n}\forall i \in [d]\}$ for some constants k_1 and k_2 with equal probability. Each $\boldsymbol{\beta}$ is s-sparse and all the $\boldsymbol{\beta}_i$ are treated equally. That is overall there should be $s|\mathcal{F}_1|$ non-zero coefficients, half of them are $k_1\sigma\sqrt{n}$ and the other half are $k_2\sigma\sqrt{n}$. Using this, equation (26) implies:

$$\begin{aligned}\det(\boldsymbol{\Sigma}) &= \frac{\sigma^2 + \frac{\frac{s|\mathcal{F}_1|}{2}k_1^2\sigma^2 n + \frac{s|\mathcal{F}_1|}{2}k_2^2\sigma^2 n}{|\mathcal{F}_1|n} - \frac{\sum_{i=1}^d(\frac{\frac{s}{d}|\mathcal{F}_1|\frac{k_1+k_2}{2}}{|\mathcal{F}_1|})^2}{n}}{n^d} \\ &= \frac{\sigma^2}{n^d}\left(1 + \frac{s}{2}k_1^2 + \frac{s}{2}k_2^2 - \frac{s^2}{d}\frac{k_1^2}{4} - \frac{s^2}{d}\frac{k_2^2}{4} - \frac{s^2}{d}\frac{k_1k_2}{2}\right)\end{aligned}$$

Substituting this in equation (25), we get

$$\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \le \frac{n}{2}\log\left(1 + \frac{s}{2}k_1^2 + \frac{s}{2}k_2^2 - \frac{s^2}{d}\frac{k_1^2}{4} - \frac{s^2}{d}\frac{k_2^2}{4} - \frac{s^2}{d}\frac{k_1k_2}{2}\right)$$

□

E. PROOF OF LEMMA 2

Proof. We make use of the bound from equation (20). For the noiseless case, we assume that the entries of $\mathbf{X}_i$ follow a Bernoulli distribution with $\mathbf{X}_{ij} \in \{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}$. Now, since $\boldsymbol{\beta}$ is an $s - sparse$ vector with binary non-zero entries $\mathbf{y}_i = \sum_{j=1}^d \boldsymbol{\beta}_j\mathbf{X}_{ij}$ takes values in a finite set which we denote as $\mathcal{Y}$. We can compute the size of $\mathcal{Y}$ in the following way. First, note that if $\boldsymbol{\beta}_i \in \{0, a, b\}$ then

$$\begin{aligned}&\mathcal{Y} = \left\{\mathbf{y}_i | \mathbf{y}_i = \alpha a + \beta b - \gamma a - \tau b;\ \alpha + \beta + \gamma + \tau = s;\ \alpha, \beta, \gamma, \tau \in \mathbb{Z}_{\ge 0}\right\} \\ &|\mathcal{Y}| = \binom{s+3}{3} \le \frac{e^3 s^3}{3^3}\end{aligned}$$

Now we assume that $Q_{(\mathbf{X}_i,\mathbf{y}_i)} = \mathcal{P}_{\mathbf{X}_i}Q_{\mathbf{y}_i}$. Furthermore, let $Q_{\mathbf{y}_i}$ be a discrete uniform distribution on $\mathcal{Y}$. Recall that $\mathbf{X}_i$ and $\boldsymbol{\beta}^*$ are marginally independent. Thus, $\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i,\boldsymbol{\beta})} = \mathcal{P}_{\mathbf{X}_i}\mathcal{P}_{\boldsymbol{\beta}}\mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}}$. We bound the KL divergence between $\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}}$ and Q as follows:

$$\begin{aligned}
\mathrm{KL}(\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}} \| Q_{(\mathbf{X}_i,\mathbf{y}_i)}) &= -\sum_{\mathbf{X}_i,\mathbf{y}_i} \mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}} \log \frac{Q_{(\mathbf{X}_i,\mathbf{y}_i)}}{\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i} \mathcal{P}_{\mathbf{X}_i}\mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}} \log \frac{Q_{(\mathbf{X}_i,\mathbf{y}_i)}}{\mathcal{P}_{\mathbf{X}_i}\mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i=\boldsymbol{\beta}^\intercal\mathbf{X}_i} \mathcal{P}_{\mathbf{X}_i} \log \frac{\mathcal{P}_{\mathbf{X}_i} Q_{\mathbf{y}_i}}{\mathcal{P}_{\mathbf{X}_i}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i=\boldsymbol{\beta}^\intercal\mathbf{X}_i} \mathcal{P}_{\mathbf{X}_i} \log Q_{\mathbf{y}_i} \\
&= \log \frac{e^3 s^3}{3^3}
\end{aligned}$$

Thus,

$$\mathrm{I}(\boldsymbol{\beta}^*, (\mathbf{X}, \mathbf{y})) \leq n \log \frac{e^3 s^3}{3^3}$$

□

F. PROOF OF LEMMA 3

Proof. Again, we make use of the bound from equation (20). We first analyze the noisy case. **Noisy Case.** In this case, $\mathbf{y}_i \in \mathcal{Y}, \forall i \in [d]$ can take two possible values and hence $\mathcal{Y} = \{+1, -1\}$ where $\mathcal{Y}$ is the set of all possible values of $\mathbf{y}_i$. Thus $|\mathcal{Y}| = 2$. Let $\mathbb{P}_{\mathbf{y}_i|\mathbf{X}_{i.},\boldsymbol{\beta}}(\mathbf{y}_i = +1) = \mathbb{P}(\mathbf{e}_i > -\mathbf{X}_{i.}^\intercal \boldsymbol{\beta}) = p$ and $\mathbb{P}_{\mathbf{y}_i|\mathbf{X}_{i.},\boldsymbol{\beta}}(\mathbf{y}_i = -1) = 1 - p$. We choose $Q_{(\mathbf{X}_i,\mathbf{y}_i)} = \mathcal{P}_{\mathbf{X}_i} Q_{\mathbf{y}_i}$, where $Q_{\mathbf{y}_i} = \text{Bernoulli}(0.5)$. Similar to the proof of Lemma 2, we can bound the KL divergence between $\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}}$ and $Q_{(\mathbf{X}_i,\mathbf{y}_i)}$ as follows:

$$\begin{aligned}
\mathrm{KL}(\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}} \| Q_{(\mathbf{X}_i,\mathbf{y}_i)}) &= -\sum_{\mathbf{X}_i,\mathbf{y}_i} \mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}} \log \frac{Q_{(\mathbf{X}_i,\mathbf{y}_i)}}{\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i} \mathcal{P}_{\mathbf{X}_i}\mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}} \log \frac{Q_{(\mathbf{X}_i,\mathbf{y}_i)}}{\mathcal{P}_{\mathbf{X}_i}\mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i\in\{+1,-1\}} \mathcal{P}_{\mathbf{X}_i}\mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}} \log \frac{\mathcal{P}_{\mathbf{X}_i} Q_{\mathbf{y}_i}}{\mathcal{P}_{\mathbf{X}_i}\mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}}} \\
&= p \log 2p + (1-p) \log 2(1-p) \\
&\leq 2 \log 2
\end{aligned}$$

Thus,

$$\mathrm{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \leq 2n \log 2$$

Next we analyze the noiseless case.

Noiseless Case. Again, the number of values that $\mathbf{y} \in \mathcal{Y}$ can take is two, i.e., $|\mathcal{Y}| = 2$. We choose $Q_{(\mathbf{X}_i,\mathbf{y}_i)} = \mathcal{P}_{\mathbf{X}_i} Q_{\mathbf{y}_i}$, where $Q_{\mathbf{y}_i} = \text{Bernoulli}(0.5)$. Then using the same approach as above, we get

$$\begin{aligned}
\text{KL}(\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}} \| Q_{(\mathbf{X}_i,\mathbf{y}_i)}) &= -\sum_{\mathbf{X}_i,\mathbf{y}_i} \mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}} \log \frac{Q_{(\mathbf{X}_i,\mathbf{y}_i)}}{\mathcal{P}_{(\mathbf{X}_i,\mathbf{y}_i)|\boldsymbol{\beta}}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i} \mathcal{P}_{\mathbf{X}_i} \mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}} \log \frac{Q_{(\mathbf{X}_i,\mathbf{y}_i)}}{\mathcal{P}_{\mathbf{X}_i} \mathcal{P}_{\mathbf{y}_i|\mathbf{X}_i,\boldsymbol{\beta}}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i=\text{sign}(\boldsymbol{\beta}^\intercal \mathbf{X}_i)} \mathcal{P}_{\mathbf{X}_i} \log \frac{Q_{\mathbf{X}_i} Q_{\mathbf{y}_i}}{\mathcal{P}_{\mathbf{X}_i}} \\
&= -\sum_{\mathbf{X}_i,\mathbf{y}_i=\text{sign}(\boldsymbol{\beta}^\intercal \mathbf{X}_i)} \mathcal{P}_{\mathbf{X}_i} \log Q_{\mathbf{y}_i} \\
&= 2 \log 2
\end{aligned}$$

Thus,

$$\text{I}(\boldsymbol{\beta}^*; (\mathbf{X}, \mathbf{y})) \leq 2n \log 2$$

□