Information Theoretic Limits for Linear Prediction with Graph-Structured Sparsity

Adarsh Barik
Krannert School of Management
Purdue University
West Lafayette, Indiana 47906
Email: abarik@purdue.edu

Jean Honorio
Department of Computer Science
Purdue University
West Lafayette, Indiana 47906
Email: jhonorio@purdue.edu

Mohit Tawarmalani
Krannert School of Management
Purdue University
West Lafayette, Indiana 47906
Email: mtawarma@purdue.edu

Abstract—We analyze the necessary number of samples for sparse vector recovery in a noisy linear prediction setup. This model includes problems such as linear regression and classification. We focus on structured graph models. In particular, we prove that sufficient number of samples for the weighted graph model proposed by Hegde and others [2] is also necessary. We use the Fano's inequality [11] on well constructed ensembles as our main tool in establishing information theoretic lower bounds.

Index Terms—Compressive sensing, Linear Prediction, Classification, Fano's Inequality, Mutual Information, Kullback Leibler divergence.

I. INTRODUCTION

Sparse vectors are widely used tools in fields related to high dimensional data analytics such as machine learning, compressed sensing and statistics. This makes estimation of sparse vectors an important field of research. In a compressive sensing setting, the problem is to closely approximate a $d-$dimensional signal by an $s-$sparse vector without losing much information. For regression, this is usually done by observing the inner product of the signal with a design matrix. It is a well known fact that if the design matrix satisfies the Restricted Isometry Property (RIP) then estimation can be done efficiently with a sample complexity of $O(s \log \frac{d}{s})$. Many algorithms such as CoSamp [5], Subspace Pursuit (SP) [4] and Iterative Hard Thresholding (IHT) [3] provide high probability performance guarantees. Baraniuk and others [1] came up with a model based sparse recovery framework. Under this framework, the sufficient number of samples for correct recovery is logarithmic with respect to the cardinality of the sparsity model.

A major issue with the model based framework is that it does not provide any recovery algorithm on its own. In fact, it is some times very hard to come up with an efficient recovery algorithm. Addressing this issue, Hegde and others [2] came up with a weighted graph model for graph structured sparsity and provided a nearly linear time recovery algorithm. They also analyzed the sufficient number of samples for efficient recovery. In this paper, we will provide the necessary condition on the sample complexity for sparse recovery on a weighted graph model. We will also note that our information theoretic lower bound can be applied not only to linear regression but also to other linear prediction tasks such as classification.

The paper is organized as follows. We describe our setup in Section II. Then we briefly describe the weighted graph model in Section III. We state our results in Section IV. In Section V, we apply our technique to some specific examples. At last, we provide some concluding remarks in Section VI. A detailed version of the same paper containing all the proofs is available at [15].

II. LINEAR PREDICTION MODEL

In this section, we introduce the observation model for linear prediction and later specify how to use it for specific problems such as linear regression and classification. Formally, the problem is to estimate an $s-$sparse vector $\bar{\beta}$ from noisy observations of the form,

$$z = f(X\bar{\beta} + e) , \tag{1}$$

where $z \in \mathbb{R}^n$ is the observed output, $X \in \mathbb{R}^{n \times d}$ is the design matrix , $e \in \mathbb{R}^n$ is a noise vector and $f : \mathbb{R}^n \to \mathbb{R}^n$ is a fixed function. Our task is to recover $\bar{\beta} \in \mathbb{R}^d$ from the observations z.

A. Linear Regression

Linear regression is a special case of the above by choosing $f(x) = x$. Then we simply have,

$$z = X\bar{\beta} + e . \tag{2}$$

Prior work analyzes the sample complexity of sparse recovery for the linear regression setup. In particular, if the design matrix X satisfies the Restricted Isometry Property (RIP) then algorithms such as CoSamp [5], Subspace Pursuit (SP) [4] and Iterative Hard Thresholding (IHT) [3] can recover $\bar{\beta}$ quite efficiently and in a stable way with a sample complexity of $O(s \log \frac{d}{s})$. Furthermore, it is known that Gaussian random matrices (or sub-Gaussian in general) satisfy RIP [6]. If we choose our design matrix to be a Gaussian matrix and we have a good sparsity model that incorporates extra information on the sparsity structure then we can reduce the sample complexity to $O(\log m_s)$ where m_s is number of possible supports in the sparsity model, i.e., the cardinality of the sparsity model [1]. In the same line of work, Hegde and others [2] proposed a weighted graph based sparsity model to efficiently learn $\bar{\beta}$.

B. Classification

We can model binary classification problems by choosing $f(x) = \text{sign}(x)$ or in other words, we can have,

$$z = \text{sign}(X\bar{\beta} + e) . \tag{3}$$

Similar to the linear regression setup, there is also prior work [9], [12], [10], on analyzing the sample complexity of sparse recovery for binary classification problem (also known as 1-bit compressed sensing).

Since arguments for establishing information theoretic lower bounds are not algorithm specific, we can extend our basic argument to the both settings mentioned above. For comparison, we will use the results by Hegde and others [2] in a linear regression setup.

III. WEIGHTED GRAPH MODEL (WGM)

In this section, we introduce the Weighted Graph Model (WGM) and formally state the sample complexity results from [2]. The Weighted Graph Model is defined on an underlying graph $G = (V, E)$ whose vertices are on the coefficients of the unknown $s-$sparse vector $\bar{\beta} \in \mathbb{R}^d$ 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 [2], for a forest $F \subseteq G$ we denote $\sum_{e \in F} w_e$ as $w(F)$. B denotes the weight budget, s denotes the sparsity (number of non-zero coefficients) of $\bar{\beta}$ and g denotes the number of connected components in 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.,

$$\rho(v) = \max_{b \in N} |\{(v', v) \in E \mid w(v', v) = b\}| . \tag{4}$$

We define the weight-degree of G, $\rho(G)$ to be the maximum weight-degree of any $v \in V$. Next, we define the Weighted Graph Model on coefficients of $\bar{\beta}$ as follows:

Definition 1 (Definition 1 in [2]): The $(G, s, g, B) - WGM$ is the set of supports defined as

$$\mathbb{M} = \{S \subseteq [d] \mid |S| = s \text{ and } \exists F \subseteq G \text{ with } V_F = S,$$
$$\gamma(F) = g, \ w(F) \leq B\} ,$$

where $\gamma(F)$ is number of connected components in a forest F. Authors in [2] provide the following sample complexity result for linear regression under their model:

Theorem 1 (Theorem 3 in [2]): Let $\bar{\beta} \in \mathbb{R}^d$ be in the $(G, s, g, B) - WGM$. Then

$$n = O(s(\log \rho(G) + \log \frac{B}{s}) + g \log \frac{d}{g}) \tag{5}$$

i.i.d. Gaussian observations suffice to estimate $\bar{\beta}$. More precisely, let $e \in \mathbb{R}^n$ be an arbitrary noise vector from equation (2) and X be an i.i.d. Gaussian matrix. Then we can efficiently find an estimate $\hat{\beta}$ such that

$$\|\bar{\beta} - \hat{\beta}\| \leq C\|e\| , \tag{6}$$

where C is a constant independent of all variables above.

Notice that in the noiseless case $(e = 0)$, we recover the exact $\bar{\beta}$. We will prove that information-theoretically, the bound on the sample complexity is tight and thus the algorithm of [2] is statistically optimal.

IV. MAIN RESULTS

In this section, we will state our results for both the noiseless and the noisy case. We establish an information theoretic lower bound on linear prediction problem defined on WGM. We use Fano's inequality [11] to prove our result by carefully constructing an ensemble, i.e., a WGM. Any algorithm which infers $\bar{\beta}$ from this particular WGM would require a minimum number of samples. Note that the use of restricted ensembles is customary for information-theoretic lower bounds [13] [14]. It follows that in the case of linear regression, the upper bound on the sample complexity by Hegde and others [2] is indeed tight.

A. Noiseless Case

Here, we provide a necessary condition on the sample complexity for exact recovery in the noiseless case. More formally,

Theorem 2: There exists a particular $(G, s, g, B) - WGM$, and a particular set of weights for the entries in the support of $\bar{\beta}$ such that if we draw a $\bar{\beta} \in \mathbb{R}^d$ uniformly at random and we have a data set $\mathcal{S}$ of $n \in o((s - g)(\log \rho(G) + \log \frac{B}{s-g}) + g \log \frac{d}{g} + (s-g) \log \frac{g}{s-g} + s \log 2)$ i.i.d. observations as defined in equation (1) with $e = 0$ then $P(\bar{\beta} \neq \hat{\beta}) \geq \frac{1}{2}$ irrespective of the procedure we use to infer $\hat{\beta}$ on $(G, s, g, B) - WGM$ from $\mathcal{S}$.

Proof sketch: We use Fano's inequality [11] on a carefully chosen restricted ensemble to prove our theorem. A detailed proof can be found in [15]. ∎

B. Noisy Case

A similar result can be stated for the noisy case. However, in this case, recovery is not exact but is sufficiently close in l_2-norm with respect to noise in the signal. Another thing to note is that in [2] inferred $\hat{\beta}$ can come from a slightly bigger WGM model but here we actually infer $\hat{\beta}$ from the same WGM.

Theorem 3: There exists a particular $(G, s, g, B) - WGM$, and a particular set of weights for the entries in the support of $\bar{\beta}$ such that if we draw a $\bar{\beta} \in \mathbb{R}^d$ uniformly at random and we have a data set $\mathcal{S}$ of $n \in o((s - g)(\log \rho(G) + \log \frac{B}{s-g}) + g \log \frac{d}{g} + (s-g) \log \frac{g}{s-g} + s \log 2)$ i.i.d. observations as defined in equation (1) with $e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma), \forall i \in \{1 \ldots n\}$ then $\mathbb{P}(\|\bar{\beta} - \hat{\beta}\| \geq C\|e\|) \geq \frac{1}{10}$ for $0 < C \leq C_0$ irrespective of the procedure we use to infer $\hat{\beta}$ on $(G, s, g, B) - WGM$ from $\mathcal{S}$.

Remark 1: Note that when $s \gg g$ and $B \geq s - g$ then $\Omega((s - g)(\log \rho(G) + \log \frac{B}{s-g}) + g \log \frac{d}{g} + (s-g) \log \frac{g}{s-g} + s \log 2)$ is roughly $\Omega(s(\log \rho(G) + \log \frac{B}{s}) + g \log \frac{d}{g})$.

Proof: We will prove this result in three steps. First, we will carefully construct an underlying graph G for the WGM. Second, we will bound mutual information between $\bar{\beta}$ and $\mathcal{S}$ by bounding the Kullback-Leibler (KL) divergence. Third, we

will bound the size of properly defined restricted ensemble to complete our proof.

a) Constructing an underlying graph G for the WGM: We construct an underlying graph for the WGM using the following steps:

- Divide 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, i.e., a node N_i^j where $i > \frac{d}{g}$ is same as node $N_{i-\frac{d}{g}}^j$.
- For each $p = 1, \ldots, \frac{B}{s-g}$, node N_i^j has an edge with nodes $N_{i+(p-1)\frac{\rho(G)}{2}+1}^j$ to $N_{i+p\frac{\rho(G)}{2}}^j$ with weight p.
- Cross edges between nodes in two different groups are allowed as long as they have edge weights greater than $\frac{B}{s-g}$ and they do not affect $\rho(G)$.

Figure 1 shows an example of a graph constructed using the above steps. Furthermore, parameters for 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 (see [15]) on the parameters and are easy to fulfill. Figure 2 shows one graph which follows our construction and also fulfills R1, R2 and R3. We define our restricted ensemble $\mathcal{F}$ on G as:

$$\mathcal{F} = \left\{ \beta \mid \beta_i = 0, \text{ if } i \notin S, \ \beta_i \in \left\{ \frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}}, \right. \right.$$
$$\left. \left. \frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}} + \frac{C_0 \sigma \sqrt{d}}{\sqrt{(1-\epsilon)}} \right\}, \text{ if } i \in S, \ S \in \mathbb{M} \right\} , \quad (7)$$

for some $0 < \epsilon < 1$ and $\mathbb{M}$ is as in Definition 1.
Our true $\bar{\beta}$ is picked uniformly at random from the above restricted ensemble. We will prove that on this restricted ensemble, our Theorem 3 holds. We will make use of following lemmas for our proof:

Lemma 1: Given the restricted ensemble $\mathcal{F}$,

$$\|\bar{\beta} - \hat{\beta}\| \leq \frac{C_0 \sigma \sqrt{d}}{\sqrt{(1-\epsilon)}} \iff \bar{\beta} = \hat{\beta} .$$

We are dealing with high dimensional cases, hence moving forward we will assume that $n < d$. We state another lemma:

Lemma 2: For some $0 < \epsilon < 1$,

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

From Lemma 1, Lemma 2 and using the fact that $d > n$ and $C \leq C_0$, the corollary below follows:

Corollary 1:

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

b) Bound on the mutual information: We will assume that the elements of design matrix X have been chosen at random and independently from $\mathcal{N}(0,1)$. The linear prediction problem from Section II can be described by the following Markov's chain:

$$\bar{\beta} \to y = X\bar{\beta} + e \to z = f(y) \to \hat{\beta} . \quad (8)$$

Lets say $\mathcal{S}$ contains n i.i.d. observations of z and $\mathcal{S}'$ contains n i.i.d. observations of y. Then using the data processing inequality [11] we can say that,

$$\mathbb{I}(\bar{\beta}, \mathcal{S}) \leq \mathbb{I}(\bar{\beta}, \mathcal{S}') . \quad (9)$$

Hence, for our purpose it suffices to have an upper bound on $\mathbb{I}(\bar{\beta}, \mathcal{S}')$. Now we can bound the mutual information by the following [8]:

$$\mathbb{I}(\bar{\beta}, \mathcal{S}') \leq \frac{1}{|\mathcal{F}|^2} \sum_{\beta \in \mathcal{F}} \sum_{\beta' \in \mathcal{F}} \mathbb{KL}(\mathbb{P}_{\mathcal{S}'|\beta} \| \mathbb{P}_{\mathcal{S}'|\beta'}) , \quad (10)$$

where $\mathbb{KL}$ is the Kullback-Leibler divergence. Note that $\mathcal{S}'$ consists of n i.i.d. observations of y. Hence,

$$\mathbb{I}(\bar{\beta}, \mathcal{S}') \leq \frac{n}{|\mathcal{F}|^2} \sum_{\beta \in \mathcal{F}} \sum_{\beta' \in \mathcal{F}} \mathbb{KL}(\mathbb{P}_{y_i|\beta} \| \mathbb{P}_{y_i|\beta'}) \quad (11)$$

Furthermore, from equation (8) and noting that the elements of X come independently from $\mathcal{N}(0,1)$,

$$y_i = X_i \beta + e_i$$
$$y_i|\beta \sim \mathcal{N}(0, \|\beta\|^2 + \sigma^2)$$
$$y_i|\beta' \sim \mathcal{N}(0, \|\beta'\|^2 + \sigma^2) .$$

We can bound the Kullback-Leibler divergence between $\mathbb{P}_{y_i|\beta}$ and $\mathbb{P}_{y_i|\beta'}$ as follows:

$$\mathbb{KL}(\mathbb{P}_{y_i|\beta} \| \mathbb{P}_{y_i|\beta'}) = \frac{1}{2}\left(\frac{\|\beta\|^2 + \sigma^2}{\|\beta'\|^2 + \sigma^2} - 1 \right.$$
$$\left. - \log \frac{\|\beta\|^2 + \sigma^2}{\|\beta'\|^2 + \sigma^2} \right)$$
$$\leq \frac{1}{2}\left(\frac{\|\beta\|^2 + \sigma^2}{\|\beta'\|^2 + \sigma^2} - 1 - 1 + \frac{\|\beta'\|^2 + \sigma^2}{\|\beta\|^2 + \sigma^2} \right)$$
$$\leq \frac{1}{2}\left(\frac{(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}} + \frac{C_0 \sigma \sqrt{d}}{\sqrt{(1-\epsilon)}})^2 s + \sigma^2}{(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}})^2 s + \sigma^2} \right.$$
$$\left. + \frac{(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}} + \frac{C_0 \sigma \sqrt{d}}{\sqrt{(1-\epsilon)}})^2 s + \sigma^2}{(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}})^2 s + \sigma^2} - 2 \right)$$
$$\leq \frac{1}{2}\left(2\frac{(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}} + \frac{C_0 \sigma \sqrt{d}}{\sqrt{1-\epsilon}})^2 s + \sigma^2}{(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}})^2 s + \sigma^2} - 2 \right)$$
$$\leq \frac{1}{2}\left(2\frac{(\sqrt{2}+1)^2 (\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}})^2 s + \sigma^2}{(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}})^2 s + \sigma^2} - 2 \right)$$
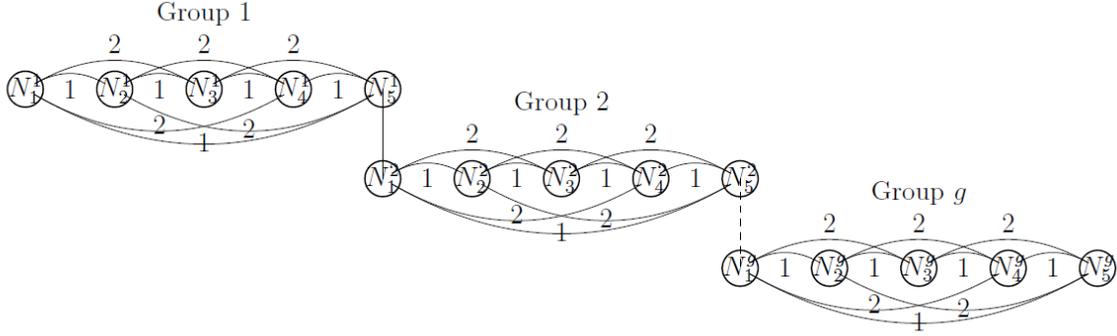$$\leq (\sqrt{2}+1)^2 - 1$$
$$\leq 5 .$$

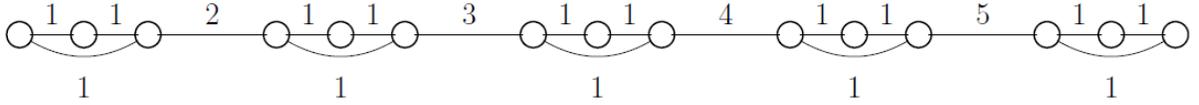Fig. 1. An example of constructing an underlying graph for $\rho(G) = 2$ and $\frac{B}{s-g} = 2$

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

The first inequality holds because $1 - \frac{1}{x} \leq \log x, \forall x > 0$, the second inequality holds by taking the largest value of numerators and the smallest value of denominators. The other inequalities follow from simple algebraic manipulation. Substituting $\mathbb{KL}(\mathbb{P}_{y_i|\beta} \| \mathbb{P}_{y_i|\beta'})$ in equation (11) we get,

$$\mathbb{I}(\bar{\beta}, \mathcal{S}') \leq 5n . \tag{12}$$

c) Bound on $|\mathcal{F}|$: Now we will count elements in $\mathcal{F}$ to complete our proof. We present the following counting argument to establish a lower bound on all the possible supports for our restricted ensemble:

1) We choose one node from each of the g groups in underlying graph G to be root of a connected component. Each group has $\frac{d}{g}$ possible candidates for 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 will only consider the cases where each connected component has $\frac{s}{g}$ nodes. Moreover, given a root node N_i^j in group j, we will choose the remaining $\frac{s}{g} - 1$ nodes connected with the root only from the nodes N_{i+1}^j to nodes $N_{i+\frac{B\rho(G)}{2(s-g)}}^j$ (using circular indices if needed). Construction of the graph G allows us to do this. 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 which across all the groups comes out to be $(\binom{\frac{\rho(G)B}{2(s-g)}}{\frac{s}{g} - 1}))^g$.

3) Each entry in the support of β can take two values which can either be $\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}}$ or $\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}} + \frac{C_0 \sigma \sqrt{d}}{\sqrt{(1-\epsilon)}}$.

It should be noted that any support chosen using the above steps satisfies constraint on weight budget, i.e., $w(F) \leq B$ as the maximum edge weight in any connected component will always be less than or equal to $\frac{B}{s-g}$. Combining all the above steps together we get:

$$\begin{aligned}
|\mathcal{F}| &\geq 2^s (\frac{d}{g})^g (\binom{\frac{\rho(G)B}{2(s-g)}}{\frac{s}{g} - 1})^g \\
&\geq 2^s (\frac{d}{g})^g (\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)} .
\end{aligned} \tag{13}$$

Using Fano's inequality [11] and results from equation (12) and equation (13), it is easy to prove the following lemma,

Lemma 3: If $n \in o(\log |\mathcal{F}|)$ then $\mathbb{P}(\hat{\beta} \neq \bar{\beta}) \geq \frac{1}{2}$.

By using Bayes' Theorem and combining Corollary 1 and Lemma 3,

$$\begin{aligned}
\mathbb{P}(\|\bar{\beta} - \hat{\beta}\| \geq C\|e\|) &\geq \mathbb{P}(\|\bar{\beta} - \hat{\beta}\| \geq C\|e\|, \bar{\beta} \neq \hat{\beta}) \\
&= \mathbb{P}(\|\bar{\beta} - \hat{\beta}\| \geq C\|e\| \mid \bar{\beta} \neq \hat{\beta}) \\
&\quad \mathbb{P}(\bar{\beta} \neq \hat{\beta}) \\
&\geq (1 - \exp(-\frac{\epsilon^2 n}{4})) \frac{1}{2} .
\end{aligned} \tag{14}$$

The last inequality (14) holds when n is $o(\log |\mathcal{F}|)$. We also know that $n \geq 1$ and if we choose $\epsilon \geq \sqrt{-4 \log 0.8} \sim 0.9448$,

then we can write inequality (14) as,

$$\mathbb{P}(\|\bar{\beta} - \hat{\beta}\| \geq C\|e\|) \geq \frac{1}{10} . \qquad (15)$$

∎

V. SPECIFIC EXAMPLES

Here, we will provide counting arguments for some of the well-known sparsity structures, such as tree sparsity and block sparsity models. It should be noted that barring the count of possible supports in the specific model our technique can be used to prove lower bounds of the sample complexity for other sparsity structures.

A. Tree-structured sparsity model

The tree-sparsity model [1], [7] is used in many applications such as wavelet decomposition of piecewise smooth signals and images. In this model, we assume 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 its parent is also included in it. Here, we will discuss the case of a binary tree which can be generalized to a $k-$ary tree. In particular, the following proposition provides a lower bound on the number of possible supports of an $s-$sparse signal following a binary tree-structured sparsity model.

Proposition 1: In a binary tree-structured sparsity model $\mathcal{F}$, $\log|\mathcal{F}| \geq cs$ for some $c > 0$.
The proof of the proposition 1 follows from the fact that we have at least 2^s different choices of $\bar{\beta}$ in our restricted ensemble. From the above and following the same proof technique as before, it is easy to prove the following corollary for the noisy case (a similar result holds for the noiseless case as well).

Corollary 2: In a binary tree-structured sparsity model, if $n \in o(s)$ then $\mathbb{P}(\|\bar{\beta} - \hat{\beta}\| \geq C\|e\|) \geq \frac{1}{10}$.
Essentially, Corollary 2 proves that the $O(s)$ sample complexity achieved in [2] is optimal for the tree-sparsity model.

B. Block sparsity model

In the block sparsity model, [1], an $s-$sparse signal, $\beta \in \mathbb{R}^{J \times N}$, can be represented as a matrix with J rows and N columns. The support of β comes from K columns of this matrix such that $s = JK$. More precisely,

Definition 2 (Definition 11 in [1]):

$$\mathcal{S}_K = \big\{ \beta = [\beta_1 \ldots \beta_N] \in \mathbb{R}^{J \times N} \ such \ that$$
$$\beta_n = 0 \ for \ n \notin L, \ L \subseteq \{1, \ldots, N\}, \ |L| = K \big\} .$$

The above can be modeled as a graph model. In particular, we can construct a graph G over all the elements in β by treating nodes in the column of the matrix as connected nodes (see Fig. 3) and then our problem is to choose K connected components from N.

It is easy to see that the number of possible supports in this model, $\mathcal{F}$, would be, $|\mathcal{F}| = 2^{KJ}\binom{N}{K} \geq 2^{KJ}(\frac{N}{K})^K$.
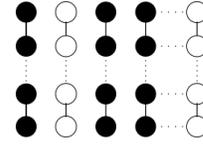

Fig. 3. Block sparsity structure as a graph model: nodes are variables, black nodes are selected variables

Correspondingly the necessary number of samples for efficient signal recovery comes out to be $\Omega(KJ + K \log \frac{N}{K})$. An upper bound of $O(KJ + K \log \frac{N}{K})$ was derived in [1] which matches our lower bound.

VI. CONCLUDING REMARKS

We proved that the necessary number of samples required to efficiently recover a sparse vector in the weighted graph model is of the same order as the sufficient number of samples provided by Hegde and others [2]. Moreover, our results not only pertain to linear regression but also apply to linear prediction problems in general.

REFERENCES

[1] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.
[2] C. Hegde, P. Indyk, and L. Schmidt, "A nearly-linear time framework for graph-structured sparsity," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 928–937, 2015.
[3] Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.
[4] Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing: Closing the gap between performance and complexity. Technical report, DTIC Document, 2008.
[5] Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.
[6] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.
[7] Chinmay Hegde, Piotr Indyk, and Ludwig Schmidt. A fast approximation algorithm for tree-sparse recovery. In *2014 IEEE International Symposium on Information Theory*, pages 1842–1846. IEEE, 2014.
[8] B. Yu. Assoaud, Fano, and Le Cam. In Torgersen E. Pollard D. and Yang G., editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435. Springer New York, 1997.
[9] Ankit Gupta, Robert D Nowak, and Benjamin Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *ISIT*, pages 1553–1557, 2010.
[10] Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya V Nori. One-bit compressed sensing: Provable support and vector recovery. In *ICML (3)*, pages 154–162, 2013.
[11] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006.
[12] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.
[13] Narayana P Santhanam and Martin J Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.
[14] 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*, pages 1373–1377. IEEE, 2010.
[15] Adarsh Barik, Jean Honorio, and Mohit Tawarmalani. Information theoretic limits for linear prediction with graph-structured sparsity. *CoRR, http://arxiv.org/abs/1701.07895*, 2017.

PROOF OF LEMMA 1

Proof: First note that when $\bar{\beta} = \hat{\beta}$ then its obvious that Lemma 1 holds. Here, we will prove that two arbitrarily chosen β_1 and β_2 such that $\beta_1, \beta_2 \in \mathcal{F}$ where $\beta_1 \neq \beta_2$ then $\|\beta_1 - \beta_2\| \geq \frac{C_0 \sigma \sqrt{d}}{\sqrt{1-\epsilon}}$. $\mathcal{F}$ is as defined in equation (7).

a) β_1 *and* β_2 *have the same support:* Since we assume that $\beta_1 \neq \beta_2$, thus they must differ in at least one position on their support. Lets say that one such position is i. Then,

$$\|\beta_1 - \beta_2\| \geq |\beta_{1i} - \beta_{2i}|$$
$$= \frac{C_0 \sigma \sqrt{d}}{\sqrt{1-\epsilon}} .$$

b) β_1 *and* β_2 *have different supports:* When β_1 and β_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 β_1 and β_2 respectively. Then,

$$\|\beta_1 - \beta_2\| \geq \sqrt{\beta_{1i}^2 + \beta_{2j}^2}$$
$$\geq \sqrt{\left(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}}\right)^2 + \left(\frac{C_0 \sigma \sqrt{d}}{\sqrt{2(1-\epsilon)}}\right)^2}$$
$$= \frac{C_0 \sigma \sqrt{d}}{\sqrt{1-\epsilon}} .$$

Since this is true for any two arbitrarily chosen β_1 and β_2, hence it holds for $\bar{\beta}$ and $\hat{\beta}$ as well. This proves the lemma. ∎

PROOF OF LEMMA 2

Proof:

$$\mathbb{P}\left(\|e\|^2 \geq \sigma^2 \frac{n}{1-\epsilon}\right) = \mathbb{P}\left(\exp\left(\frac{\lambda}{2}\|\frac{e}{\sigma}\|^2\right) \geq \exp\left(\frac{\lambda}{2}\frac{n}{1-\epsilon}\right)\right)$$
$$\leq \frac{\mathbb{E}\left[\exp\left(\frac{\lambda}{2}\|\frac{e}{\sigma}\|^2\right)\right]}{\exp\left(\frac{\lambda}{2}\frac{n}{1-\epsilon}\right)}$$
$$= \exp\left(\frac{-\lambda}{2}\frac{n}{1-\epsilon}\right)\left(\frac{1}{1-\lambda}\right)^{\frac{n}{2}} .$$

The first equality holds for any $\lambda > 0$, we take $0 < \lambda < 1$. The second inequality comes from Markov's inequality. The last equality follows since $\frac{e_i}{\sigma} \overset{iid}{\sim} \mathcal{N}(0, 1)$. Now, by taking $\lambda = \epsilon$,

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

The last inequality holds because for $0 < \epsilon < 1$, $\frac{\epsilon}{1-\epsilon} + \log(1-\epsilon) \geq \frac{\epsilon^2}{2}$. This proves our lemma,

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

∎

PROOF OF LEMMA 3

Proof: Using Fano's inequality [11], we can say that,

$$\mathbb{P}(\hat{\beta} \neq \bar{\beta}) \geq 1 - \frac{\mathbb{I}(\bar{\beta}, \mathcal{S}) + \log 2}{\log |\mathcal{F}|}$$
$$\geq 1 - \frac{\mathbb{I}(\bar{\beta}, \mathcal{S}') + \log 2}{\log |\mathcal{F}|}$$
$$\geq 1 - \frac{5n + \log 2}{\log |\mathcal{F}|} .$$

The first inequality follows from equation (9) and the second inequality follows from the upper bound on the mutual information established in equation (12). Now, we want $\mathbb{P}(\hat{\beta} \neq \bar{\beta}) \leq \frac{1}{2}$, then it follows that n must be,

$$n \geq \frac{1}{10} \log |\mathcal{F}| - \frac{1}{5} \log 2 . \tag{16}$$

This proves the lemma.

∎

PROOF OF THEOREM 2

Proof:

c) *Constructing an underlying graph* G: We assume that our underlying graph G fulfills all the properties mentioned while proving Theorem 3. On this underlying graph G, we define our restricted ensemble $\mathcal{F}$ as:

$$\mathcal{F} = \{\beta_i \in \{1, -1\}, \; if \; i \in S, \; else \; \beta_i = 0, \; S \in \mathbb{M}\} ,$$

where $\mathbb{M}$ is as in Definition 1.

d) *Bound on the mutual information:* We will assume that the elements of design matrix X have been chosen at random and independently from $\mathcal{N}(\frac{1}{s\sqrt{2}}, \frac{1}{s})$. As in the proof of Theorem 3, we can describe noiseless linear prediction problem as the following Markov's chain:

$$\bar{\beta} \to y = X\bar{\beta} \to z = f(y) \to \hat{\beta} . \tag{17}$$

Lets say $\mathcal{S}$ contains n i.i.d. observations of z and $\mathcal{S}'$ contains n i.i.d. observations of y. Then using the data processing inequality [11], we can say that,

$$\mathbb{I}(\bar{\beta}, \mathcal{S}) \leq \mathbb{I}(\bar{\beta}, \mathcal{S}') . \tag{18}$$

Hence, for our purpose it suffices to have an upper bound on $\mathbb{I}(\bar{\beta}, \mathcal{S}')$. Now using results from [8],

$$\mathbb{I}(\bar{\beta}, \mathcal{S}') \leq \frac{1}{|\mathcal{F}|^2} \sum_{\beta \in \mathcal{F}} \sum_{\beta' \in \mathcal{F}} \mathbb{KL}(\mathbb{P}_{\mathcal{S}'|\beta} \| \mathbb{P}_{\mathcal{S}'|\beta'}) .$$

where $\mathbb{KL}$ is the Kullback-Leibler divergence. Note that $\mathcal{S}'$ consists of n i.i.d. observations of y. Hence,

$$\mathbb{I}(\bar{\beta}, \mathcal{S}') \leq \frac{n}{|\mathcal{F}|^2} \sum_{\beta \in \mathcal{F}} \sum_{\beta' \in \mathcal{F}} \mathbb{KL}(\mathbb{P}_{y_i|\beta} \| \mathbb{P}_{y_i|\beta'}) . \tag{19}$$

Furthermore from equation (17) and noting that the elements of X come independently from $\mathcal{N}(\frac{1}{s\sqrt{2}}, \frac{1}{s})$,

$$y_i = X_i\beta$$

$$y_i|\beta \sim \mathcal{N}(\frac{\sum_{k=1}^{d}\beta_k}{s\sqrt{2}}, 1)$$

$$y_i|\beta' \sim \mathcal{N}(\frac{\sum_{k=1}^{d}\beta_k'}{s\sqrt{2}}, 1) \ .$$

We can bound $\mathbb{KL}(\mathbb{P}_{y_i|\beta}\|\mathbb{P}_{y_i|\beta'})$ by,

$$\mathbb{KL}(\mathbb{P}_{y_i|\beta}\|\mathbb{P}_{y_i|\beta'}) = \frac{1}{2}(\frac{\sum_{k=1}^{d}(\beta_k - \beta_k')}{s\sqrt{2}})^2$$

$$\leq 1 \ .$$

Substituting $\mathbb{KL}(\mathbb{P}_{y_i|\beta}\|\mathbb{P}_{y_i|\beta'})$ in equation (19) we get,

$$\mathbb{I}(\bar{\beta}, \mathcal{S}') \leq n \ . \tag{20}$$

e) Bound on $|\mathcal{F}|$: Using a similar counting logic used in Theorem 3, we can get:

$$|\mathcal{F}| \geq 2^s(\frac{d}{g})^g(\frac{\rho(G)Bg}{2(s-g)^2})^{(s-g)} \ . \tag{21}$$

We prove the theorem by substituting the mutual information from equation (20) and $|\mathcal{F}|$ from equation (21) in the Fano's inequality [11]. ∎

DISCUSSION ON THE REQUIREMENTS FOR THE UNDERLYING GRAPH G

We mentioned before that R1, R2 and R3 are quite mild requirements on the parameters. In fact, it is easy to see that,

Proposition 2: Given any value of s, g and $B \geq s - g$, 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.

Proof: R3 is readily satisfied if each edge has at least unit edge weight and we are not forced to choose isolated nodes in support. Most of the graph-structured sparsity models fulfill this requirement. R2 gives us a lower bound on the choice of $\rho(G)$,

$$\rho(G) \geq \frac{2(s-g)^2}{Bg} \ .$$

Similarly, given a value of $\rho(G)$, R1 just provides a lower bound on choice of d,

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

Clearly, there is an infinite number of combinations for $\rho(G)$ and d. ∎