ALGORITHMS AND BOUNDS FOR SENSING CAPACITY AND COMPRESSED SENSING WITH APPLICATIONS TO LEARNING GRAPHICAL MODELS

Shuchin Aeron, Manqi Zhao, Venkatesh Saligrama

Boston University
Electrical and Computer Engineering Department
8 Saint Mary's street, MA 02215

ABSTRACT

We consider the problem of recovering sparse phenomena from projections of noisy data, a topic of interest in compressed sensing. We describe the problem in terms of sensing capacity, which we define as the supremum of the ratio of the number of signal dimensions that can be identified per projection. This notion quantifies minimum number of observations required to estimate a signal as a function of sensing channel, SNR, sensed environment(sparsity) as well as desired distortion up to which the sensed phenomena must be reconstructed. We first present bounds for two different sensing channels: (A) i.i.d. Gaussian observations (B) Correlated observations. We then extend the results derived for the correlated case to the problem of learning sparse graphical models. We then present convex programming methods for the different distortions for the correlated case. We then comment on the differences between the achievable bounds and the performance of convex programming methods.

keywords: Sensing capacity; SNR level; Compressed Sensing; LASSO; LP; Learning Graphical Models.

I. INTRODUCTION

In many real world sensor network applications, the problem of recovering the signal from noisy projections is of interest. For example, consider SNET localization, where we need to localize target positions in the field given noisy observations made by the sensors. Although the dimension of the field is very large, the number of active targets is often very small. Another example is to reconstruct natural images or CAT scans from noisy observations. Here one exploits the fact that since images usually have large variations only at the boundaries, the total variation should be approximately sparse. The issue of sparse reconstruction arises in bio-informatics as well, namely, reverse engineering of gene interaction networks [5], where the edge density is known to be scale free.

A fundamental aspect of these scenarios is that the underlying signal of interest is sparse in some representation, i.e., the signal (or the number of edges) has small support. Another important aspect of these problems is the desired objective. In field and image reconstruction problems one usually seeks estimates that have small average mean squared error (MSE). In contrast for localization, communications, and learning graphical models one usually requires that the support of the estimate match the true signal.

The problem is essentially one of solving a large-scale under-determined system of equations in a noisy environment. There are two different aspects to the problem. From the perspective of algorithms, researchers have proposed convex optimization algorithms(e.g., Basis Pursuit, Lasso) to analyze their performances(see [6], [8], [9], [14], [15], [10]). On the other hand fundamental information-theoretic bounds that are algorithm independent have been presented in [2] [1]. Here the authors propose a quantity, named sensing capacity, to incorporate the effects of distortion metric, sensing modality, sensing environment and signal-to-noise (SNR) level into a single metric.

In this paper we first present fundamental information theoretic bounds for two different sensing channels: (A) i.i.d. Gaussian observations (B) Correlated observations. We then extend the results derived for the correlated case to the problem of learning sparse graphical models. We then present convex programming methods for the different distortions for the correlated case. Finally we provide comparisons between the achievable information theoretic bounds and the performance of convex programming methods such as LASSO, which is an ℓ_1-constrained quadratic programming problem for reconstructing signals from noisy observations([7], [12], [13], [16], [11]).

II. PROBLEM SET-UP

Let $\mathbf{Y} \in \mathbb{R}^{m \times n}$ be a measurement matrix observed under a given signal-to-noise ratio(SNR) model given by,

$$\mathbf{Y} = \mathbf{G}\mathbf{X} + \mathbf{N}/\sqrt{SNR}$$

where $\mathbf{X} \in \mathbb{R}^{n \times n}$ is a matrix, in general, modeling the ambient domain of observation (field) and $\mathbf{N} \in \mathbb{R}^{m \times 1}$ is an

AWGN vector with unit variance. The matrix $\mathbf{G} \in \mathbb{R}^{m \times n}$ is a measurement or sensing matrix.

Let $\mathbf{X}_i, i = 1, 2, ..., n$ denote the columns of the matrix $\mathbf{X}$. We assume that each $\mathbf{X}_i$ is a sparse vector, i.e. consists of few non-zero components. This sparsity can be modeled for each $\mathbf{X}_i$ using priors that induce sparsity. We can consider both non-random and random models for $\mathbf{X}$. In the non-random case $\mathbf{X}$ is assumed to be arbitrary except for a sparsity constraint:

$$\mathbf{X} \in \{Z \in \mathbb{R}^{m \times n} \mid \|Z\|_0 \leq \alpha mn\}$$

where, $\|\cdot\|_0$ refers to the the ℓ_0 norm.

In the random case we consider the matrix $\mathbf{X}$ to have i.i.d. components. Each component is distributed according to a mixture distribution with a singular measure of weight $(1-\alpha)$ at zero. Both mixtures of discrete and continuous densities can be considered.

We also consider two different models for sensing matrices.

1. Each element of the matrix $\mathbf{G}_{ij}$ is drawn i.i.d. according to a Gaussian distribution, $\sim \mathcal{N}(0, \frac{1}{n})$, where our normalization is with respect to the signal dimension. In the literature a different normalization with respect to the number of observations is considered [14]. Specifically, $\mathbf{G}_{ij} \sim \mathcal{N}(0, \frac{1}{m})$. The different normalizations turn out to be insignificant when sparsity level α is held constant. However, the results need to be re-interpreted for vanishing sparsity regimes. The i.i.d. choice of sensing matrix is motivated by the *compressed sensing* problem that deals with efficient and non-adaptive sampling of sparse signals.
2. Row vectors $\mathbf{G}_i,\ i = 1, 2, ..., m$ of the sensing matrix $\mathbf{G}$ are distributed i.i.d. according to a multivariate normal distribution $\sim \mathcal{N}(0, \Sigma)$. The covariance matrix Σ characterizes the components of each row vector. This choice of sensing matrix is motivated by the problem of learning graphical models.

We use the notion of sensing capacity to characterize the performance of the above set-up. More elaborate notion of sensing capacity in relation to applications arising in sensor networks is described in [1],[2] and the reader is referred to these papers for more details. To this end we denote the ratio $\frac{n}{m}$ as the *Sensing Rate* for the set-up. We consider asymptotic situation letting both the ambient dimension, n, and the number of sensors approach infinity. For this we need a sequence of n-fold probability distribution, $\mathcal{P}_n$, a sequence of sensing matrices, $\mathbf{G}$ and sequence of estimators, that maps the m-dimensional observation to the solution. Then we define the sensing capacity as follows,

Definition 2.1: Given this set-up as outlined above we define an ϵ *sensing capacity* as the supremum over all the sensing rates such that for the sequence of n-dimensional realizations over the sensing domains and the sequence of sensing matrices $\mathbf{G}$, $\exists$ a sequence of reconstruction operators such that the probability that the distortion in reconstruction is below d_0 is greater than $1-\epsilon$, i.e.,

$$C_\epsilon(\Theta, \mathbf{G}, d_0) = \limsup_{m,n} \left\{ \frac{n}{m} : Pr(d(\mathbf{X}, \hat{\mathbf{X}} \geq d_0) \leq \epsilon | \mathbf{G}) \to 0 \right\}$$

where Θ is the parameter(s) governing the sensing domain, e.g. sparsity in the case under consideration and where $\hat{\mathbf{X}}$ denotes the reconstruction/estimate of $\mathbf{X}$ from $\mathbf{Y}$.

This leads to the following definition of Sensing Capacity,

Definition 2.2: Sensing Capacity is defined as,

$$C = \lim_{\epsilon \to 0} C_\epsilon(\Theta, \mathbf{G}, d_0)$$

The problem of finding the sensing capacity as defined above is a function of the sensing channel(s) Φ, the complexity of the solution objective defined by the mapping f, the parameter Θ governing the probability distribution(s) $\mathcal{P}_n$ and the distortion measure $d(.,.)$ with the desired QOS d_0 in reconstruction.

III. SUPPORT RECOVERY

III-A. IID Sensing Channel

We are given m observations, $\mathbf{Y}_j,\ j = 1, 2, \ldots, m$ with

$$\mathbf{Y}_j = \sum_{i=1}^{n} G_{ji}\mathbf{X}_i + N_j/\sqrt{SNR}$$

where, $\mathbf{X}$ is an arbitrary real-valued vector of length n. We assume that elements $G_{ji} \sim \mathcal{N}(0, 1/n)$ of matrix G are distributed i.i.d. and $N_j \sim \mathcal{N}(0, 1)$ are i.i.d. Gaussian random variables.

Lower Bound: We first prove a lower bound for sensing capacity for binary signals $\mathbf{X}$. Let $\mathbf{X}$ belong to the set of k-sparse sequences in $\{0, 1\}^n$. Let $\alpha = \frac{k}{n}$ denote the sparsity ratio. Then the probability of error in detecting $\mathbf{X}_1$ from a sequence $\mathbf{X}_2$ that differs from $\mathbf{X}_2$ in 2 places is given by, the standard Q function,

$$P_e = Q\left(\frac{||\mathbf{G}(\mathbf{X}_1 - \mathbf{X}_2)||\sqrt{SNR}}{2}\right) \geq \frac{1}{2\sqrt{2}} \exp\left\{-\frac{||\mathbf{G}(\mathbf{X}_1 - \mathbf{X}_2)||^2 SNR}{4}\right\}$$

where the inequality follows from the following lemma,

Lemma 3.1: For all $x > 0$

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{y^2}{2}} dy \geq \frac{1}{2\sqrt{2}} e^{-x^2} \quad (1)$$

Proof: See Appendix. ■

Now taking the expectation with respect to $\mathbf{G}$ and noting that $||\mathbf{G}(X_1 - X_2)||^2$ is a χ^2 random variable with m degrees of freedom we have

$$P_e \geq \frac{1}{2\sqrt{2}} \left(\frac{1}{1 + \frac{2SNR}{n}} \right)^{m/2}$$
$$= \frac{1}{2\sqrt{2}} 2^{-\frac{m}{2}\log(1+\frac{2SNR}{n})} \approx \frac{1}{2\sqrt{2}} 2^{-\frac{mSNR}{n}}$$

for sufficiently large n, m. The binary signal case can be readily generalized to arbitrary real valued signals, $\mathbf{X}$, which are bounded from below. To this end let $\beta = \inf_k |\mathbf{X}_k| > 0$. Then

$$P_e \geq \frac{1}{2\sqrt{2}} 2^{-\frac{m}{2}\log(1+\frac{2\beta SNR}{n})} \approx \frac{1}{2\sqrt{2}} 2^{-\frac{m\beta SNR}{n}} \quad (2)$$

for sufficiently large n, m. It can be seen from the above expression that in order for the probability of error to go to zero at a fixed SNR, $\frac{m}{n}$ has to go to ∞ or $\frac{n}{m}$ has to go to zero. Thus in this case sensing capacity is zero.

Upper Bound: Note that from our lower bound it is unclear at what rate the sensing capacity $\frac{n}{m}$ approaches zero. We will now derive an upper bound to the sensing capacity $\frac{n}{m}$ and establish the rate of approach to zero. To this end we have the following lemma.

Lemma 3.2: For the set-up under consideration $m = \mathcal{O}(\alpha n \log n)$ sensors suffice for exact recovery. Alternatively, an SNR of $\mathcal{O}(\log n)$ with $m = \mathcal{O}(\alpha n)$ sensors is sufficient for exact recovery.

Proof: See appendix (section V-B). ■

This implies that for perfect recovery sensing capacity goes to zero at rate $\mathcal{O}(\frac{1}{\log n})$ for fixed SNR level.

Remark 3.1: Although we assume binary valued $\mathbf{X}$ in the proof of the above lemma, the main step (see proof of Lemma 3.2 in the appendix) only requires the application of restricted isometry property (RIP). This property applies equally well to arbitrary $\mathbf{X}$. For this reason with a slight modification we can extend the proof to cover real-valued vector $\mathbf{X}$ as well. We state this below without proof:

Theorem 3.1: Suppose $\mathbf{X}$ is a real-valued parameter bounded from below, i.e., $\beta = \inf_k |\mathbf{X}_k| > 0$. Consider the setup of Lemma 3.2. It follows that $m = \mathcal{O}(\alpha n \log n)$ measurements are sufficient for support recovery.

Remark 3.2: Note that support recovery for matrix valued signal, $\mathbf{X} \in \mathbb{R}^{n \times n}$, can be decomposed into recovery for each column through a union bound. Indeed it suffices to have $m = \mathcal{O}(\alpha n \log^2 n)$ or $m = \mathcal{O}(\alpha n)$ and $SNR = \mathcal{O}(\log^2 n)$ for perfect reconstruction.

III-B. Correlated Channel

We are given m observations, $\mathbf{Y}_j$, $j = 1, 2, \ldots, m$ with

$$\mathbf{Y}_j = \sum_{i=1}^{n} G_{ji}\mathbf{X}_i + N_j/\sqrt{SNR}$$

where, $\mathbf{X}$ is an arbitrary real-valued vector of length n. We assume that each row vector $G_j = [G_{j1}, G_{j2} \ldots, G_{jn}] \sim \mathcal{N}(0, \Sigma)$ is a Gaussian random vector and different rows of matrix G are distributed i.i.d. Suppose, G_j is normalized such that $\sigma_{\max}(\Sigma) = 1$. Also, let N_j be a sequence of i.i.d. zero mean unit variance gaussian random variables.

Lower Bound: Let $\Sigma = U\Sigma_D U^*$ be the singular value decomposition of Σ and $\sigma_{\min}$ be the minimum singular value. Consider the transformation $\tilde{G} = GU\Sigma_D^{-1/2}$ and signal transformation $\tilde{X} = \Sigma_D^{1/2}U^*X$. The row vector $\tilde{G}_j = G_j U\Sigma_D^{-1/2}$ is a Gaussian row vector with i.i.d. components. This implies that the matrix $\tilde{G}$ has i.i.d. components. Therefore,

$$G(X_1 - X_2) = \tilde{G}(\tilde{X}_1 - \tilde{X}_2)$$

Note that $\|\tilde{X}_1 - \tilde{X}_2\| \geq \sigma_{\min}\|X_1 - X_2\|$. Consequently, if the space of signals, $\mathbf{X}$ are bounded from below by $\beta = \min_k |X(k)| > 0$, the following lower bound follows along the lines leading upto Equation 2.

$$P_e \geq \frac{1}{2\sqrt{2}} 2^{-\frac{m}{2}\log(1+\sigma_{\min}\frac{2\beta SNR}{n})} \approx \frac{1}{2\sqrt{2}} 2^{-\frac{m\sigma_{\min}\beta SNR}{n}}$$

Upper Bound: For the upper bound the main step we need to check is the RIP property for correlated Gaussian channels. The rest of the steps will follow identically along the lines of the upper bound for i.i.d. channel.

Lemma 3.3: Consider the correlated sensing channel G given above. It follows that for any $\delta > 0$ there exists a γ such that for any $k \leq \gamma n$ sparse signal a modified RIP property holds, i.e.,

$$(1-\delta)\sigma_{\min}(\Sigma)\|\mathbf{X}\| \leq \|G\mathbf{X}\| \leq (1+\delta)\sigma_{\max}(\Sigma)\|\mathbf{X}\|$$

We provide an outline of the proof here. To this end, let $\pi(l)$, $l = 1, 2, \ldots, \binom{n}{k}$ be the different choices of k-sparse vectors. The k-sparse real-valued vectors are $\mathbf{X}_{\pi(l)}$, $l = 1, 2, \ldots, \binom{n}{k}$. Suppose, $G_{\pi(l)}$, $l = 1, 2, \ldots, \binom{n}{k}$ are the corresponding sub-matrices obtained by selecting k columns of G. Let $G_{\pi(l)} = U_l\Sigma_l U_l^*$ be the corresponding singular value decomposition. Consider the transformations $\tilde{G}_{\pi(l)} = G_{\pi(l)}U_l\Sigma_l^{-1/2}$ and signal transformation $\tilde{\mathbf{X}}_{\pi(l)} = \Sigma_l^{1/2}U_l^*\mathbf{X}_{\pi(l)}$. Then it follows that $G_{\pi(l)}\mathbf{X}_{\pi(l)} = \tilde{G}_{\pi(l)}\tilde{\mathbf{X}}_{\pi(l)}$

By noting that $G_{\pi(l)}$ is a matrix with i.i.d. components it follows from an application of Johnson-Lindenstrauss theorem (see [3]) that there exists a number δ for sufficiently large, n such that,

$$(1-\delta)\sigma_{\min}(\Sigma_l)\|\mathbf{X}_{\pi(l)}\| \leq (1-\delta)\|\tilde{\mathbf{X}}_{\pi(l)}\| \leq \|\tilde{G}_{\pi(l)}\tilde{\mathbf{X}}_{\pi(l)}\| \leq (1-\delta)\|\tilde{\mathbf{X}}_{\pi(l)}\| \leq (1+\delta)\sigma_{\max}(\Sigma_l)\|\mathbf{X}_{\pi(l)}\|$$

with overwhelming probability. This implies that with the same overwhelming probability the following holds:

$$(1-\delta)\sigma_{\min}(\Sigma_l)\|\mathbf{X}_{\pi(l)}\| \leq \|G_{\pi(l)}\mathbf{X}_{\pi(l)}\| \leq (1+\delta)\sigma_{\max}(\Sigma_l)\|\mathbf{X}_{\pi(l)}\|$$

Next we note that, $\sigma_{\min}(\Sigma_l) \leq \sigma_{\min}(\Sigma)$ and $\sigma_{\max}(\Sigma_l) \leq \sigma_{\max}(\Sigma)$

Now using a standard union bounding argument over the $\left(\begin{array}{c} n \\ k \end{array}\right)$ sets (see [3]) it follows that the RIP property holds as well for all k sparse sequences, $\mathbf{X}$ where $k \leq \gamma n$ for sufficiently small γ.

Based on the above analysis we have the following result for correlated channels:

Theorem 3.2: Consider the aforementioned correlated channel. Suppose $\mathbf{X}$ is a αn sparse real-valued parameter bounded from below, i.e., $\beta = \inf_k |\mathbf{X}_k| > 0$ with α sufficiently small to satisfy the RIP property. Let $\sigma_{\max}$ and $\sigma_{\min}$ be the maximum and minimum singular values of any row vector of correlated channel G. It follows that $m = \mathcal{O}(\alpha n \log n \frac{\sigma_{\max}^2}{\sigma_{\min}^2})$ measurements are sufficient for support recovery.

IV. LEARNING GAUSSIAN GRAPHICAL MODELS

Markov random fields (MRFs) are special random fields that can be associated with some graph $G = (V, E)$. Namely, potentials of a MRF satisfy certain properties that lead to conditional independence relations with respect to cutsets of the associated graph, as illustrated in Figure IV. Here we adopt such models with the interpretation that each node of the graph denotes a random quantity that pertains to a sensor measurement or observation, and the graph structure connecting the nodes reflects first-order dependencies between the measurements. In this paper we deal with Gaussian graphical models. A Gaussian graphical model is given by a Gaussian distribution, namely, if the vector $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{|V|})$ denotes the random vector corresponding to the different observations at the different nodes then:

$$\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$$

The concentration matrix $\Lambda^0 = \Sigma^{-1} \doteq [\lambda_{uv}^0]$ reflects the graphical structure, namely, if the element λ_{ik}^0 is zero it implies that there is no edge connecting the ith and jth observation. Furthermore, observation at any node, v, can

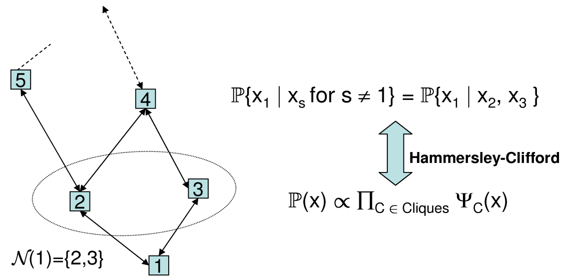

Fig. 1. Markov Random Field and Characterization in terms of Potentials

be written as a linear superposition of the realization at the neighboring nodes of the graph:

$$\mathbf{X}_v = \sum_{(v,w)\in E} \lambda_{v,w}\mathbf{X}_w + N_v$$

where, $N_v \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable independent of $\mathbf{X}_w$ for $w \neq v$. The matrix $\Lambda \doteq [\lambda_{uv}]$ is related to Λ^0 by the following expression:

$$\Lambda = D^{-1/2}\Lambda^0 D^{-1/2}$$

where, D is the diagonal of Λ_0.

Our goal is to determine the edge connectivity namely, the matrix Λ from a set of T i.i.d. observations of variables at each node. Denote the observations at node v as $\mathbf{X}_v(1), \mathbf{X}_v(2), \ldots, \mathbf{X}_v(T)$.

Note that the setup falls into the correlated channel case considered in the previous section. Therefore, we can prove the following theorem which we state without proof:

Theorem 4.1: Consider the setup above. Suppose each column of Λ is a d (i.e., the nodal degree) sparse real-valued parameter bounded from below, i.e., $\beta = \inf |\lambda_{uv}| > 0$ with $\alpha = \frac{d}{|V|}$ sufficiently small to satisfy the RIP property for each column. Let $\sigma_{\max}$ and $\sigma_{\min}$ be the maximum and minimum singular values of Σ. It follows that $T = \mathcal{O}(d \log^2 |V| \frac{\sigma_{\max}^2}{\sigma_{\min}^2})$ measurements are sufficient for support recovery.

Remark 4.1: The $\log^2 |V|$ term appears from the requirement for simultaneous recovery of connectivity pattern for all the nodes. This introduces an extra $\log$ factor.

Remark 4.2: Note that if the degree is constant one only needs $\log^2 |V|$ measurements for support recovery. Thus if one has 20000 nodes (usually found in gene-networks) and there are at most 20 edges for each node it follows that we approximately need only 400 measurements.

Finally, we present convex programming methods for recovery of the normalized concentration matrix, Λ^0 using LASSO. Suppose the Frobenius norm of noise satisfies

$$\|N\|_F \leq \eta$$

We define the following optimization problem:

$$\min_{\lambda_{u,v}} \sum_{u,v} |\lambda_{u,v}|$$

subject to:

$$\sum_{t=1}^{T} \sum_{v} (\mathbf{X}_v(t) - \sum_{w} \lambda_{v,w} \mathbf{X}_w(t))^2 \leq \eta$$

The solution to the above optimization problem approximates Λ in the ℓ_2 norm:

Theorem 4.2: Let $\hat{\Lambda}$ be the solution to the above optimization problem. It follows that if the degree d at each node satisfies $d/T \leq C \log(|V|/T)$ then,

$$\|\hat{\Lambda} - \Lambda\|_F \leq \frac{\|N\|_F}{\sigma_{\min}(\Sigma_0)(1 - \sqrt{\frac{d}{\rho T}})^{0.5} - \sqrt{\rho}\sigma_{\max}(\Sigma_0)(1 + \sqrt{\frac{d}{\rho T}})^{0.5}} \quad (3)$$

whenever the denominator in the above expression is positive for some $0 < \rho \leq \gamma T$ and $\gamma T > d$. Here γ is a sufficiently small number so that every $T \times \gamma T$ submatrix of a $T \times |V|$ i.i.d. Gaussian matrix satisfies the RIP property with $\delta \approx \sqrt{\gamma}$.

The proof follows by straightforward application of Theorem 1 of [4]. The main step is that the intersection of all the following constraints ensures good approximation.

1) Cone Constraint: This follows from the observation that the solution to the optimization problem must satisfy $\|\hat{\Lambda}\|_1 \leq \|\Lambda\|_1$ where the ℓ_1 norm here denotes the absolute sum of elements of the matrix. Consider J to be the set of indices of Λ that are non-zero and J^c the complementary set. Let $\hat{\Lambda}_J$ be the matrix obtained from $\hat{\Lambda}$ by restricting its support to J. Then we have that,

$$\|\hat{\Lambda}_{J^c}\|_1 \leq \|\hat{\Lambda}_J - \Lambda\|_1$$

2) Tube Constraint:

$$\sum_{t=1}^{T} \sum_{v} (\sum_{w} \hat{\lambda}_{v,w} \mathbf{X}_w(t) - \sum_{w} \lambda_{v,w} \mathbf{X}_w(t))^2 \leq 2\eta$$

3) Restricted Isometry Property: This is the property we derived in the previous section which we recall here:

$$(1-\delta)\sigma_{\min}(\Sigma)\|\Lambda\|_F \leq \|\mathbf{X}\Lambda\|_F \leq (1+\delta)\sigma_{\max}(\Sigma)\|\Lambda\|_F$$

for sufficiently sparse matrix Λ. Here the matrix $\mathbf{X} = [\mathbf{X}_{tv}]$.

The main implications of the result is that we need the denominator of Equation 3 to be positive. This holds if

$$\frac{\sigma_{\min}^2}{\sigma_{\max}^2} \approx C_0 d/T$$

where C_0 is some positive constant. This implies that the number of measurements T must scale with the condition number. This is the scaling suffered in the maximum likelihood bound derived in Theorem 4.1 for exact support recovery. Hence LASSO appears to follow a similar scaling for the number of measurements. However, we point out that ℓ_2 recovery does not imply support recovery. Indeed for support recovery we must have $\|N\|_F$ to be essentially a constant together with non-zero components of Λ strictly bounded away from zero. Consequently this implies that SNR must scale with graph size. Therefore, although the performance of LASSO appears to scale optimally with respect to number of measurements there appears to be a gap in terms of SNR.

V. APPENDIX

V-A. Proof of lemma 3.1

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{y^2}{2}} dy$$

By the change of variables $y = x + z$ we have,

$$\begin{aligned} Q(x) &= \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{y^2}{2}} dy \\ &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{(x+z)^2}{2}} dz \\ &= e^{-x^2/2} \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{z^2}{2}+xz} dz \end{aligned}$$

Since $\frac{x^2+z^2}{2} \geq xz$ for $x, z \geq 0$, we have

$$Q(x) \geq e^{-x^2} \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-z^2} dz$$

Since $Q(0) = 1/2$ we have

$$Q(x) \geq \frac{1}{2\sqrt{2}} e^{-x^2}$$

V-B. Proof of lemma 3.2

To this end, let $\mathbf{X}_0$ be the true signal. We define the error events,

$$E_l = \bigcup_{\mathbf{X}: d_H(\mathbf{X}_0, \mathbf{X}) = l} \{\mathbf{X}_0 \to \mathbf{X}\}$$

i.e., E_l denotes the union of error events with a hamming error of l from the true signal $\mathbf{X}_0$. It can be shown that under the AWGN noise $\mathbf{N} \in \mathbb{R}^m$,

$$E_l = \bigcup_{\mathbf{X}: d_H(\mathbf{X}_0, \mathbf{X}) = l} \left\{ \mathbf{N}^T(\mathbf{G}\mathbf{X}_0 - \mathbf{G}\mathbf{X}) \geq \frac{1}{2} ||\mathbf{G}\mathbf{X}_0 - \mathbf{G}\mathbf{X}||^2 \right\}$$

and the probability of error is given by,

$$P_e = Pr\left(\bigcup_l E_l\right)$$

To this end we have the following corollary.

Corollary 5.1: Restricted Isometry Property (RIP): Given matrix $\mathbf{G} \in \mathbb{R}^{m\times n}$, such that each element of the matrix is drawn i.i.d. as $\mathbf{G}_{ij} \sim \mathcal{N}(0, \frac{SNR}{n})$. Then for $0 < \delta < 1$, for all $l \leq \frac{c_1(\delta)m}{\log n}$, $c_1(\delta) > 0$ and for all $\mathbf{X}_0, \mathbf{X} \in \{0,1\}^n : d_H(\mathbf{X}_0, \mathbf{X}) = l$,

$$(1-\delta)\frac{SNRm}{n}||\mathbf{X}_0 - \mathbf{X}||^2 \leq ||\mathbf{G}(\mathbf{X}_0 - \mathbf{X})||^2 \leq (1+\delta)\frac{SNRm}{n}||\mathbf{X}_0 - \mathbf{X}||^2 \quad (4)$$

with probability exceeding $1 - e^{-c_2(\delta)m}$, $c_2(\delta) > 0$.

Proof: The result follows directly from Theorem 5.2 in [**?**]. ■

To this end assume that the matrix $\mathbf{G}$ satisfies a restricted isometry property of order l_0, where l_0 is the maximum allowable hamming error. This implies that for all $\mathbf{X}_0, \mathbf{X} : d_H(\mathbf{X}_0, \mathbf{X}) = l$, $l \leq l_0$, and $\delta \in (0,1)$ equation (4) holds true. In the following we will use a shorthand notation $\mathbf{G} \in$ RIP(δ) to denote that $\mathbf{G}$ satisfies equation (4)for $l \leq l_0$.

Now *conditioned* on $\mathbf{G} \in$ RIP(δ) and the fact that $\mathbf{X}_0, \mathbf{X} \in \{0,1\}^n$, it is straightforward to show that for all $1 < l \leq l_0$,

$$E_l \subset E_l^\delta = \bigcup_{\mathbf{X}:d_H(\mathbf{X}_0,\mathbf{X})=l}\left\{\mathbf{N}^T(\mathbf{G}\mathbf{X}_0 - \mathbf{G}\mathbf{X}) \geq \frac{1}{2}\frac{SNRm}{n}(1-\delta)||\mathbf{X}_0 - \mathbf{X}||^2\right\}$$

To this end let $\mathbf{X}^l \in \{0,1\}^n$ denote a vector that is at a hamming distance of l from $\mathbf{X}_0$. Now note that for each set of type,

$$S_l = \left\{\mathbf{N}^T(\mathbf{G}\mathbf{X}_0 - \mathbf{G}\mathbf{X}^l) \geq \frac{SNRm}{2n}(1-\delta)||\mathbf{X}_0 - \mathbf{X}^l||^2\right\} \quad (5)$$

comprising the union in E_l^δ the R.H.S can be written as the following superposition,

$$\mathbf{N}^T(\mathbf{G}\mathbf{X}_0 - \mathbf{G}\mathbf{X}^l) = \sum_{j=1}^{l}\mathbf{N}^T\mathbf{G}(\mathbf{X}_0 - \mathbf{X}_j^1)$$

for some $\mathbf{X}_j^1 \in \{0,1\}^n$, $j = 1,2,...,l$ is at a hamming distance of 1 from $\mathbf{X}_0$. Similarly the L.H.S can be written as the following superposition,

$$\frac{SNRm}{2n}(1-\delta)||\mathbf{X}_0 - \mathbf{X}^l||^2 = \frac{SNRm}{2n}(1-\delta)\sum_{j=1}^{l}||(\mathbf{X}_0 - \mathbf{X}_j^1)||^2$$

Thus the set in equation (5) can be written as,

$$S_l = \left\{\sum_{j=1}^{l}\mathbf{N}^T\mathbf{G}(\mathbf{X}_0 - \mathbf{X}_j^1) \geq \frac{SNRm}{2n}(1-\delta)\sum_{j=1}^{l}||(\mathbf{X}_0 - \mathbf{X}_j^1)||^2\right\}$$

From the above equation and from corollary 5.1 it follows that

$$S_l \subset \left\{\sum_{j=1}^{l}\mathbf{N}^T\mathbf{G}(\mathbf{X}_0 - \mathbf{X}_j^1) \geq \frac{1}{2}\frac{1-\delta}{1+\delta}\sum_{j=1}^{l}||\mathbf{G}(\mathbf{X}_0 - \mathbf{X}_j^1)||^2\right\}$$

From above it follows that every set of the type in equation 5 comprising the union in E_l^δ is contained in the following union,

$$A^\delta = \bigcup_{\mathbf{X}:d_H(\mathbf{X}_0,\mathbf{X})=1}\left\{\mathbf{N}^T\mathbf{G}(\mathbf{X}_0 - \mathbf{X}) \geq \frac{1}{2}\frac{1-\delta}{1+\delta}||\mathbf{G}(\mathbf{X}_0 - \mathbf{X})||^2\right\}$$

Thus conditioned on $\mathbf{G} \in$ RIP(δ), it implies that $E_l^\delta \subset A^\delta$ for all $l > 1$. Also since $\frac{1-\delta}{1+\delta} < 1, \forall 0 < \delta < 1$ it implies that $E_1^\delta \subset A^\delta$. Therefore we have,

$$\begin{aligned} P_e\left(\bigcup_l E_l\right) &\leq P\left(\bigcup_l E_l^\delta\right) \\ &= P\left(A^\delta|\mathbf{G} \in \text{RIP}(\delta)\right) + P\left(\bigcup_l E_l^\delta|\mathbf{G} \notin \text{RIP}(\delta)\right) \\ &\leq (1 - e^{-mc_2(\delta)})P(A^\delta) + e^{-mc_2(\delta)} \end{aligned}$$

Now upper bounding $P(A^\delta)$ using the standard union bound over the n possible sequences at Hamming distance of 1 from $\mathbf{X}_0$ and using the standard upper bound to the error function, $Q(x) \leq e^{-\frac{x^2}{2}}$ we have,

$$P_e \leq (1 - e^{-mc_2(\delta)})\exp\left\{-\frac{(1-\delta)^2||\mathbf{G}(\mathbf{X}_0 - \mathbf{X}^1)||^2}{8(1+\delta)^2}\right\}e^{\log n} + e^{-mc_2(\delta)}$$

Now taking the expectation over $\mathbf{G}$ we get the following upper bound,

$$P_e \leq (1 - e^{-mc_2(\delta)})e^{-\frac{m}{2}\log\left(1+\frac{(1-\delta)^2 SNR}{4n(1+\delta)^2}\right)}e^{\log n} + e^{-mc_2(\delta)}$$

Now for $k = \alpha n$ sparse sequences the maximal allowable distortion is $l_0 \leq 2\alpha n$. Therefore in order that corollary 5.1 to hold true and to make the probability of error to go zero it is sufficient to choose $m = \frac{8(1+\delta)^2 \alpha n(\log n+\eta)}{(1-\delta)^2 SNR}$ for any $\eta > 0$. Thus a scaling of $m = \mathcal{O}(\alpha n \log n)$ is sufficient for exact support recovery.

One can also choose $m = \frac{8(1+\delta)^2 \alpha n(1+\eta)}{(1-\delta)^2}, \eta > 0$ and SNR = $\log n$ with maximal allowable hamming distortion $l_0 \leq \frac{2\alpha n}{\log n}$ satisfying corollary 5.1.

VI. REFERENCES

[1] S. Aeron, M. Zhao, and V. Saligrama, *Information theoretic bounds to sensing capacity of sensor networks under fixed SNR*, IEEE Information Theory Workshop (Lake Tahoe, CA), Sept. 2-6 2007.

[2] S. Aeron, M. Zhao, and V. Saligrama, *On sensing capacity of sensor networks for a class of linear observation models*, IEEE Statistical Signal Processing Workshop, Wisconsin-Madison, WI, August 26-29 2007.

[3] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A Simple Proof of the Restricted Isometry Property for Random Matrices*, to appear in Constructive Approximation

[4] E. J. Candes, J. Romberg and T. Tao. *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 1207-1223.

[5] K. Basso et. al., "Reverse Engineering of Regulatory Networks in Human B cells," *Nature Genetics* 2005.

[6] E. Candes, J. Romberg, and T. Tao, *Robust uncertainity principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Transactions on Inforamtion Theory **52** (2006), no. 2, 489–509.

[7] E. Candes and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies?*, preprint (2004).

[8] D. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory **52** (2006), no. 4, 1289–1306.

[9] D. Donoho, *For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparset solution*, Communications on Pure and Applied Mathematics **59** (2006), no. 6, 797–829.

[10] M.F. Duarte, M.A. Davenport, and R.G. Baraniuk, *Sparse signal detection from incoherent projections*, Intl. Conf. on Acoustics Speech and Signal Processing, May 2006.

[11] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association **96** (2001), no. 456, 1138–1360.

[12] J. Haupt and R. Nowak, *Signal reconstruction from noisy random projections*, IEEE Transactions on Information Theory **52** (2006), no. 9, 4036–4068.

[13] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society **58** (1996), no. 1, 267–288.

[14] J. A. Tropp, *Recovery of short linear combinations via l_1 minimization*, IEEE Transactions on Inforamtion Theory **51** (2005), no. 4, 1568–1570.

[15] J. A. Tropp, *Just relax: Convex programming methods for identifying sparse signals*, IEEE Transactions on Inforamtion Theory **51** (2006), no. 3, 1030–1051.

[16] M. J. Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programs*, Allerton Conference on Communication, Control and Computing, 2006.