Tight Measurement Bounds for Exact Recovery of Structured Sparse Signals

Nikhil Rao[1] Benjamin Recht[2] Robert Nowak[1]

[1] Department of Electrical and Computer Engineering
[2] Department of Computer Sciences
University of Wisconsin-Madison

Abstract

Standard compressive sensing results state that to exactly recover an s sparse signal in $\mathbb{R}^p$, one requires $\mathcal{O}(s \cdot \log p)$ measurements. While this bound is extremely useful in practice, often real world signals are not only sparse, but also exhibit structure in the sparsity pattern. We focus on group-structured patterns in this paper. Under this model, groups of signal coefficients are active (or inactive) together. The groups are predefined, but the particular set of groups that are active (i.e., in the signal support) must be learned from measurements. We show that exploiting knowledge of groups can further reduce the number of measurements required for exact signal recovery, and derive near optimal bounds for the same. The number of measurements needed only depends on the number of groups under consideration, and not the particulars of the groups (e.g., compositions, sizes, extents, overlaps, etc.). The results are also shown to predict experimental performance quite well.

1 Introduction

In many fields such as genetics, image processing, and machine learning, one is faced with the task of recovering very high dimensional signals from relatively few measurements. In general this is not possible, but fortunately many real world signals are, or can be transformed to be, sparse, meaning that only a small fraction signal coefficients are non-zero. Compressed Sensing [3, 7] allows us to recover sparse, high dimensional signals with very few measurements. In fact, the results indicate that one only needs $\mathcal{O}(s \cdot \log p)$ random measurements to exactly recover an s sparse signal of length p. Moreover, greedy and convex procedures ([1, 10]) exist to efficiently recover the signal.

In many applications however, one not only has knowledge about the sparsity of the signal, but some additional information about the structure of the sparsity pattern as well:

- In genetics, the genes are arranged into pathways, and genes belonging to the same pathway are highly correlated with each other [22].
- In image processing, the wavelet transform coefficients can be modeled as belonging to a tree, with parent-child coefficients exhibiting similar properties [13, 6, 16].
- In wideband spectrum sensing applications, the spectrum typically displays clusters of non-zero frequency coefficients, each corresponding to a narrowband transmission [12]

In cases such as these, the sparsity pattern can be represented as a union of certain groups of coefficients (e.g., coefficients in certain pathways, tree branches, or clusters). This knowledge about the signal structure can help further reduce the number of measurements one needs to exactly recover the signal. Indeed, the

authors in [5] derive information theoretic bounds for the number of measurements needed for a variety of signal ensembles, including trees. In, [17, 11], the authors show that one needs far fewer measurements when the signal can be expressed as lying in a union of subspaces, and explicit bounds are derived for the tree-structured case. In this paper, we derive bounds on the number of random iid gaussian measurements needed to exactly recover a sparse signal when its pattern of sparsity is a union of a subset of groups.

We show that the number of measurements needed only depends on the number of groups under consideration, and not the particulars of the groups (e.g., compositions, sizes, extents, etc.). In particular, the groups need not be disjoint. The degree to which groups overlap, remarkably, has no effect on our bounds. This is somewhat surprising since overlapping groups are strongly coupled in the observations, tempting one to suppose that overlap may make recovery more challenging.

We analyze the group-structured sparse recovery problem using a random Gaussian measurement model. We emphasize that although the derivation assumes the measurement matrix to be Gaussian, it can be extended to **any** subgaussian case, by paying a small constant penalty, as shown in [19]. We restrict ourselves to the Gaussian case here since it highlights the main ideas and keeps the analysis as simple as possible.

In this paper, we aim to contribute the following:

- We show that the group lasso with overlap [8] method lends itself well to the task of structured signal recovery under CS measurements.
- We derive near optimal measurement bounds for group-sparse signals, under no assumptions about the nature of overlap between groups.
- We show how, using these bounds, we can recover structured sparse signals with far fewer measurements than the standard number of compressive measurements.

To the best of our knowledge, these results are new and distinct from prior theoretical characterizations of group lasso methods. In [15, 14], the authors derive consistency results for the group lasso under arbitrary groupings of variables. In [4], the authors consider the groups to *partition* the space, and derive conditions for recovery using the group lasso [9]. In this paper, we focus on deriving non-asymptotic sample bounds rather than consistency, and make **no** assumptions about the nature of groups. We appeal to the notion of restricted minimum singular values of an operator, which is a much weaker assumption than the restricted isometry conditions required in [17, 4].

The main result shows that signals with support on k of M possible groups, exact recovery is possible from $O(k[\log(M-k)+1])$ measurements using an overlapping group lasso algorithm. Also, the bounds we derive is near optimal in the sense that the constants cannot be made much smaller. Note that the bound does not depend on the size of the support (sparsity level s), and so the bound can be much smaller than the number of measurements required by methods that assume a simple, unstructured model of sparsity such as the lasso.

Our proof derives from the techniques developed in [24]. The rest of this paper is organized as follows: in section 2, we lay the groundwork for the main contribution of the paper, *viz.* bounding the gaussian width of the Normal Cone at a point under the atomic norm of group sparse signals. We describe the theory and reasoning behind this approach. In section 3 we derive bounds on the gaussian width of the aforementioned normal cone, and show how this allows us to bound the number of random *i.i.d.* gaussian measurements needed to be taken for exact recovery of group sparse signals. Section 4 outlines the experiments we performed and the corresponding results obtained. We conclude our paper in section 5.

1.1 Notations

We first introduce notations that we will use for the rest of the paper. Consider a signal of length p, that is s sparse. Note here that in case of multidimensional signals like images, we assume they are vectorized to have length p. The coefficients of the signal are grouped into sets $\{G_i\}_{i=1}^{M}$, such that $\forall i \in \{1, 2, \cdots, M\}, G_i \subset$

$\{1, 2, \cdots, p\}$. We denote the set of groups by $\mathcal{G} = \{G_i\}_{i=1..M}$, and $|\cdot|$ denotes the cardinality of a set. We let x^* be the (sparse) signal to be recovered, whose non zero coefficients lie in k of the M groups $\mathcal{G}^* \subset \mathcal{G}$. Clearly, $|\mathcal{G}^*| = k \leq M = |\mathcal{G}|$. We let $\Phi_{n\times p}$ be a measurement matrix consisting of *i.i.d.* gaussian entries of mean 0 and unit variance so that every column is a realization of an *i.i.d.* gaussian length n vector with covariance I. For any vector $x \in \mathbb{R}^p$, we denote by x_G the subvector of x such that $(x_G)_i = x_i$ if $i \in G$, and 0 otherwise. We denote the observed vector by $y \in \mathbb{R}^n \quad : y = \Phi x^*$. The absence of a subscript following a norm $||\cdot||$ implies the $\ell 2$ norm.

2 Preliminaries

In this section, we will set up the problem that we wish to solve in this paper. We will argue as to why exact recovery of the signal corresponds to the minimization of the atomic norm of the signal, with the atoms obeying certain properties governed by the signal structure.

2.1 Atoms and the atomic set

To begin with, let us formalize the notion of atoms and the atomic norm of a signal (or vector). We will restrict our attention to group-sparse signals in $\mathbb{R}^p$, though the same concepts can be extended to other spaces as well. We assume that $x \in \mathbb{R}^p$ can be decomposed as :

$$x = \sum_{i=1}^{k} c_i a_i \quad c_i \geq 0$$

The vectors a_i are called *atoms*, and form the basic building blocks of any signal, which can be represented as a conic combination of the atoms. We call $\mathcal{A} = \{a_i\}$ to be the *atomic set*. Given a vector $x \in \mathbb{R}^p$ and an atomic set, we define the *atomic norm* as

$$||x||_{\mathcal{A}} = \inf \left\{ \sum_{a\in\mathcal{A}} c_a : x = \sum_{a\in\mathcal{A}} c_a a \quad c_a \geq 0 \quad \forall a \in \mathcal{A} \right\} \tag{1}$$

The atomic decomposition of the signal has been known to be the simplest representation of the signal in some sense. Hence, to obtain a "simple" representation of a vector, we look to minimize the atomic norm subject to constraints (equation (2)):

$$\hat{x} = \underset{x\in\mathbb{R}^p}{\operatorname{argmin}} ||x||_{\mathcal{A}} \quad \textbf{s.t.} \quad y = \Phi x^* \tag{2}$$

Indeed, when the atoms are merely the canonical basis in $\mathbb{R}^p$, the atomic norm reduces to the standard ℓ_1 norm, and minimization of the atomic norm yields the well known *lasso* procedure [23].

Assuming we are aware of the group structure $\mathcal{G}$, we now proceed to define the atomic set and the corresponding atomic norm for our framework:

$$\forall G \in \mathcal{G}, \text{ let } A_G = \{a^G \in \mathbb{R}^p : ||(a^G)_G||_2 = 1, (a^G)_{G^c} = 0\}$$

$$\mathcal{A} = \{A_G\}_{G\in\mathcal{G}} \tag{3}$$

We now show that the atomic norm of a vector $x \in \mathbb{R}^p$ under the atomic set defined in equation (3) is equivalent to the overlapping group lasso norm defined in [8], a special case of which is the standard group lasso norm [9]. Thus, minimizing the atomic norm in this case is exactly the same as the group lasso with overlapping groups.

Lemma 2.1 *Given any arbitrary set of groups $\mathcal{G}$, we have*

$$||x||_{\mathcal{A}} = \Omega^{\mathcal{G}}_{overlap}(x)$$

and hence any algorithm that minimizes one norm will equivalently minimize the other.

Proof In equation (1), we can substitute $v_G = c_G a$, giving us $c_G = |c_G| \cdot ||a|| = ||c_G a|| = v_G$. Hence,

$$\begin{aligned} ||x||_{\mathcal{A}} &= inf \left\{ \sum_{a \in \mathcal{A}} c_a : x = \sum_{a \in \mathcal{A}} c_a a \;\; c_a \geq 0 \forall a \in \mathcal{A} \right\} \\ &= inf \left\{ \sum_{G \in \mathcal{G}} ||v_G|| \quad : x = \sum_{G \in \mathcal{G}} v_G \right\} \\ &= \Omega^{\mathcal{G}}_{overlap}(x) \quad \blacksquare \end{aligned}$$

Corollary 2.2 *Under the atomic set defined in 3 , when $\mathcal{G}$ partitions $\mathbb{R}^p$,*

$$||x||_{\mathcal{A}} = \sum_{G \in \mathcal{G}} ||x_G||$$

Proof $\Omega^{\mathcal{G}}_{overlap} = \sum_{G \in \mathcal{G}} ||x_G||$ in the non overlapping case. $\blacksquare$

Thus, equation (2) yields:

$$\hat{x} = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \;\; \Omega^{\mathcal{G}}_{overlap}(x) \;\; \textbf{s.t.} \;\; y = \Phi x^* \tag{4}$$

which can be solved using the overlapping group lasso [8].

Also note that we can directly compute the dual of the atomic norm from the set of atoms

$$\|u\|^*_{\mathcal{A}} = \sup_{a \in \mathcal{A}} \langle a, u \rangle = \max_{G \in \mathcal{G}} ||u_G|| \tag{5}$$

The dual norm will be useful in our derivations below.

2.2 Gaussian Widths and Exact Recovery

We define the *tangent cone* and *normal cone* at x^* with respect to $conv(\mathcal{A})$ under $||x||_{\mathcal{A}}$ as [18]:

$$\mathcal{T}_{\mathcal{A}}(x^*) = \text{cone}\{z - x^* \quad : ||z||_{\mathcal{A}} \leq ||x^*||_{\mathcal{A}}\} \tag{6}$$

$$\begin{aligned} \mathcal{N}_{\mathcal{A}}(x^*) &= \{u \;:\; \langle u, z \rangle \leq 0, \;\; \forall z \in \mathcal{T}_{\mathcal{A}}(x^*)\} \\ &= \{u \;:\; \langle u, x^* \rangle = t\|x\|_{\mathcal{A}} \text{ and } \|u\|^*_{\mathcal{A}} \leq t \text{ for some } t \geq 0\} \end{aligned} \tag{7}$$

We note that, from [24] (Prop. 2.1), $\hat{x} = x^*$ (equation (2)) is unique *iff*

$$null(\Phi) \cap \mathcal{T}_{\mathcal{A}}(x^*) = \{0\} \tag{8}$$

Hence, we require that the tangent cone at x^* intersects the nullspace of Φ only at the origin, to guarantee exact recovery.

Before we state the main recovery result from [24], we define the *gaussian width* of a set:

Definition Let $\mathbb{S}^{p-1}$ denote the unit sphere in $\mathbb{R}^p$. The Gaussian width $w(S)$ of a set $S \in \mathbb{S}^{p-1}$ is

$$w(S) = \mathbb{E}_g \left[\sup_{z \in S} g^t z \right]$$

where $g \sim \mathcal{N}(0, I)$

Gordon uses the Gaussian width to provide bounds on the probability that a random subspace of a certain dimension misses a subset of the sphere [2]. In [24], these results are specialized to the case of atomic norm recovery. In particular, we will make use of the following

Proposition 2.3 *[[24], Corollary 3.2] Let $\Phi : \mathbb{R}^p \to \mathbb{R}^n$ be a random map with i.i.d. zero-mean Gaussian entries having variance $1/n$. Further let $\Omega = T_{\mathcal{A}}(x^*) \cap \mathbb{S}^{p-1}$ denote the spherical part of the tangent cone $T_{\mathcal{A}}(x^*)$. Suppose that we have measurements $y = \Phi x^*$, and we solve the convex program (2). Then x^* is the unique optimum of (2) with high probability provided that*

$$n \geq w(\Omega)^2 + \mathcal{O}(1).$$

To complete our problem setup we will also restate proposition 3.6 in [24] :

Proposition 2.4 *(Proposition 3.6 in [24]) Let C be any non-empty convex cone in $\mathbb{R}^p$, and let $g \sim \mathcal{N}(0, I)$ be a Gaussian vector. Then:*

$$w(C \cap \mathbb{S}^{p-1}) \leq \mathbb{E}_g[dist(g, C^*)] \tag{9}$$

where $dist(.,.)$ denotes the Euclidean distance between a point and a set, and C^* is the dual cone of C

We can then square equation (9) use Jensen's inequality to obtain

$$w(C \cap \mathbb{S}^{p-1})^2 \leq \mathbb{E}_g[dist(g, C^*)^2] \tag{10}$$

We note here that the dual cone of the Tangent cone is the Normal cone, and vice-versa.

Thus, to derive measurement bounds, we only need to calculate the square of the gaussian width of the intersection of the tangent cone at x^* with respect to the atomic norm and the unit sphere. This value can be bounded by the distance of a gaussian random vector to the Normal cone at the same point, as implied by equation (10). In the next section, we derive bounds on this quantity.

3 Gaussian Width of the Normal Cone of the Group Sparsity Norm

For generic groups $\mathcal{G}$ $v \in \mathcal{N}_{\mathcal{A}}(x^*) \Leftrightarrow \exists t \geq 0 \quad : \langle v, x^* \rangle = t\|x^*\|_{\mathcal{A}}, \quad \|v_G\| = 1$ if $G \in \mathcal{G}^*, \quad \|v_G\| \leq t$ if $G \notin \mathcal{G}^*$. It is not hard to see that, in the case of disjoint groups,

$$\mathcal{N}_{\mathcal{A}}(x^*) = \{z \in \mathbb{R}^p : z_i = t\frac{(x^*_G)_i}{||x^*_G||} \quad \forall G \in \mathcal{G}^*, \quad ||z_G|| \leq t \quad \forall G \notin \mathcal{G}^*, t > 0\} \tag{11}$$

However, in the case of overlapping groups, we do not know how to obtain such a closed from.

We now prove the main result of this paper, a near optimal bound on the number of gaussian measurements needed to recover a group-sparse signal:

Theorem 3.1 *To exactly recover a k- group sparse signal decomposed into M groups in $\mathbb{R}^p$, $\mathcal{O}(k[log(M - k) + 1])$ iid gaussian measurements are sufficient.*

To prove this result, we need two lemmas:

Lemma 3.2 *Let $q_1, \ldots, q_L$ be L, χ-squared random variables with d-degrees of freedom. Then*

$$\mathbb{E}[\sup_{1 \leq i \leq L} q_i] \leq 2\log(L) + d + 1.$$

Proof This result is well-known, but we provide a simple proof for completeness. $\|x\|_2$ is Lipschitz with Lipschitz constant 1. Thus we can apply concentration of measure for Gaussian random variables. First, bound the expectation as

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq i\leq L} q_i\right] &= \int_0^\infty \mathbb{P}\left[\max_{1\leq i\leq L} q_i \geq t\right] dt \\ &\leq \delta + \int_\delta^\infty \mathbb{P}\left[\max_{1\leq i\leq L} q_i \geq t\right] dt \\ &\leq \delta + L\int_\delta^\infty \mathbb{P}\left[q_1 \geq t\right] dt\end{aligned}$$

Setting $\delta = 2\log(L) + d$ gives

$$\begin{aligned}\delta + L\int_\delta^\infty \mathbb{P}\left[q_1 \geq t\right] dt &= 2\log(L) + d + L\int_{2\log(L)+d}^\infty \mathbb{P}\left[q_1 \geq t\right] dt \\ &\leq 2\log(L) + d + n\int_{2\log(L)}^\infty \mathbb{P}\left[q_1 \geq (\sqrt{d}+\sqrt{t})^2\right] dt \\ &\leq 2\log(L) + d + \frac{L}{2}\int_{2\log(L)}^\infty \exp\left(-\tfrac{t}{2}\right) dt \\ &= 2\log(L) + d + L\exp\left(-\tfrac{t}{2}\right)\Big|_{t=2\log(L)} dt \\ &= 2\log(L) + d + 1\end{aligned}$$

where the first inequality follows by the triangle inequality and the second inequality is concentration of measure. ∎

Lemma 3.3 *Suppose $v \in \mathbb{R}^p$ is supported on some set of groups $\mathcal{G}^* \subset \mathcal{G}$*

$$\|v\| \leq \sqrt{|\mathcal{G}^*|}\|v\|_{\mathcal{A}}^* .$$

Proof By duality, it suffices to show that $\|z\|_{\mathcal{A}} \leq \sqrt{|\mathcal{G}|}\|z\|$ for all z. For any z, there exists a representation $z = \sum_{G\in\mathcal{G}^*} w_G$ where none of the w_G overlap. It then follows that

$$\|z\|_{\mathcal{A}} \leq= \sum_{G\in\mathcal{G}^*} \|w_G\| \leq \sqrt{|\mathcal{G}^*|}\left(\sum_{G\in\mathcal{G}^*} \|w_G\|^2\right)^{1/2} = \sqrt{|\mathcal{G}^*|}\|z\|$$

completing the proof. ∎

Proof of Theorem 3.1 Let $v \in \mathcal{N}_{\mathcal{A}}(x^*)$ with $\|v\|_{\mathcal{A}}^* = 1$. Since $x_G = 0$ for $G \notin \mathcal{G}^*$, we may assume that $v_G = 0$ for $G \notin \mathcal{G}^*$.

We need to compute an upper bound on the expected distance of a normal random vector to a point in the normal cone at x^*. To proceed, let $w \sim \mathcal{N}(0, I_p)$ be a vector with iid gaussian entries. Let $t(w) = \max_{G\notin\mathcal{G}^*} \|w_G\|$ and construct a vector $r \in \mathcal{N}_{\mathcal{A}}(x^*)$ by setting:

$$r_G = \begin{cases} tv_G & \text{if } G \in \mathcal{G}^* \\ w_G & \text{otherwise} \end{cases}$$

It is clear that r_G satisfies all the necessary properties for membership in $\mathcal{N}_{\mathcal{A}}(x^*)$. Referring to equation

(10), we now consider the expected squared distance between r and w:

$$\begin{aligned}
\mathbb{E}[||r-w||^2] &= \mathbb{E}\sum_{i=1}^{|\mathcal{G}|}(r_i - w_i)^2 \\
&= \mathbb{E}[\sum_{g\in\mathcal{G}^*}(r_G - w_G) \quad + \quad \sum_{G\notin\mathcal{G}^*}(r_G - w_G)] \\
&\overset{(i)}{=} \mathbb{E}[\sum_{G\in\mathcal{G}}(t\cdot v_G - w_G)^2] \\
&\overset{(ii)}{=} \mathbb{E}[t(w)^2]\left\|\sum_{G\in\mathcal{G}^*} v_G\right\|^2 + \mathbb{E}\left[\sum_{G\in\mathcal{G}^*}||w_G||^2\right] \\
&\overset{(iii)}{\leq} (2\log(M-k)+B+1)\left\|\sum_{G\in\mathcal{G}^*} v_G\right\|^2 + kB \\
&\overset{(iv)}{\leq} k(2\log(M-k)+B+1)+kB \\
&= 2k(\log(M-k)+B)+k
\end{aligned}$$

Where (i) follows from that fact that we considered part of the same gaussian vector to lie in the normal cone. (ii) follows because t is independent from w_G. B is the maximum over all group sizes, and (ii) uses this to bound the expectation of the χ^2 random variable $||w_G||^2$. (iii) follows from Lemma 3.2 as $|\mathcal{G}\setminus\mathcal{G}^*| = M-k$. (iv) follows from Lemma 3.3, giving us the final result. ∎

Thus, we see that we require $2k(\log(M-k)+B)+k$ gaussian measurements to exactly recover a $M-$ group sparse signal, out of which k groups are active. It is worthwhile to note that the constant we obtain (2) is **near optimal**, in that the bound on the maximum of χ^2 variables cannot be made much tighter. If the groups are disjoint to begin with , the the normal cone will be given by equation (11), and we would have an equality in (iv). Thus, we see that we do not pay an additional penalty when the groups overlap. This fact is surprising, since one would expect that one would need more measurements to effectively capture the dependencies among the overlapping groups.

We note that although we pay no extra price to measure the signal when the groups overlap, there is an additional cost in the recovery process of the signal, in that the groups need to first be separated by replication of the coefficients [8], or resort to a primal-dual method to solve the problem [20].

4 Experiments and Results

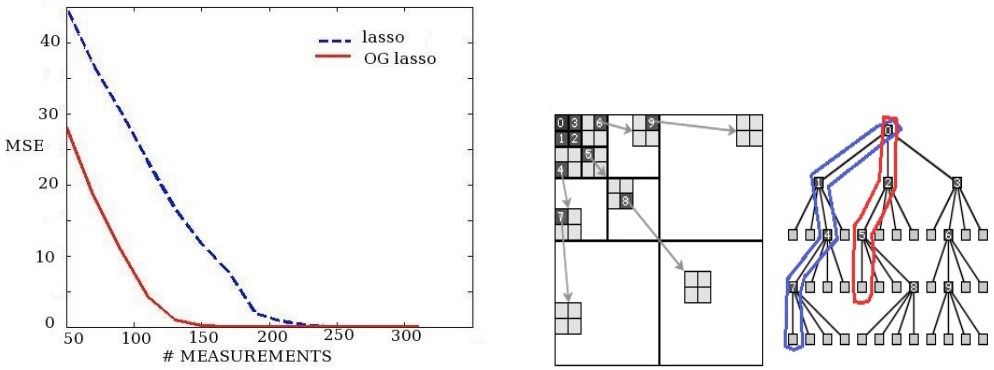
We extensively tested our method against the standard lasso procedure . In the case where the groups overlap, we use the replication method outlined in [8], to reduce the problem to that of non overlapping groups. Note that, [20] develop an algorithm to solve the overlapping group lasso problem without replication. Although this is better in terms of memory requirements, the replication method is faster, and we prefer to use it. We will denote the group lasso with overlap by "OG lasso".

4.1 Comparison with lasso

We compare the number of measurements needed for our method with that needed for the lasso. For the lasso, we use the bound derived in [24] , *viz.* $(2s+1)log(p-s)$. We generate toy signals with $p = 1024$. We consider $M = 32$ non overlapping groups for simplicity, all of the same size $B = 32$. We let $k = 2$, which gives an upper bound on the sparsity as $s \leq k \times B = 64$. However, we randomly assign non zero coefficient

to one of these locations in the signal with probability 0.7. This yields $s \sim 45$ in expectation. The active groups are assigned uniformly at random.

For the lasso , we use SpaRSA [21] , learning the regularization parameter λ over a logarithmic grid. The same grid is used to learn λ for the group lasso, for which we use SpaRSA again, after the replication step. Fig 1(a) displays the mean reconstruction error $||\hat{x} - x^*||_2^2$ as a function of the number of random measurements taken. The errors have been averaged over 200 tests, and each time a new random signal was generated with the above mentioned parameters.

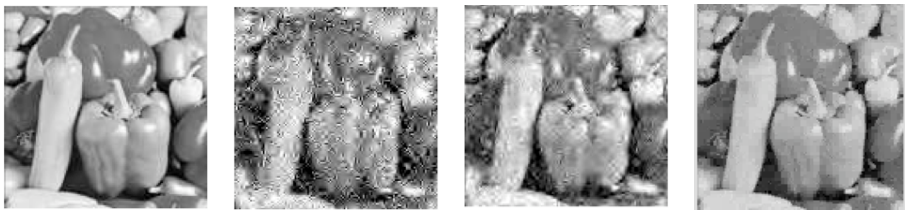


(a) (b)

Figure 1: Comparison of the reconstruction error of the lasso and the group lasso, with increasing measurements. (a) Lasso achieves near perfect reconstruction when $\#measurements \sim 225$, whereas the same error is achieved by OG lasso when $\#measurements \sim 150$. (b) shows the modeling of DWT coefficients along paths on a tree

4.2 Wavelet transform coefficients modeled as overlapping groups

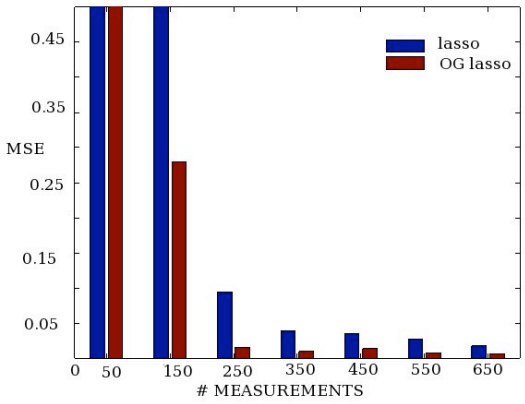
Wavelet transform coefficients of images/signals obey a group sparse structure, due to the parent-child dependencies across scales [16].These dependencies can be modeled as groups along paths in the wavelet tree from root to leaf (Figure 1(b)). We consider the "peppers" image, sized to 128 X 128. We vectorize the image to length 16384, and take 5000 gaussian measurements. We considered this image so as to compare our results with those in [17]. We see that (Figure 2), we can recover the signal accurately, with very few measurements. The figure also shows the results obtained by [17], where the authors also take 5000 measurements.



(a) Peppers (b) CoSaMP (MSE = 22.8) (c) Model-Based (MSE = 11.1) (d) OG Lasso (MSE = 7.8)

Figure 2: Performance on the peppers image

Along the same lines, we consider the DWT coefficients of a piecewise constant, length 2048 signal, with 5 "jumps" located randomly in the signal. Considering the wavelet tree structure again [17, 16], we have $k = 5, M = 1024, B = 11, s \leq 55, p = 2048$. These numbers indicate that we need atleast 185 measurements for recovery using the group lasso with overlap, as opposed to nearly 800 measurements for recovery using the lasso. Figure 3(a) shows that this is indeed true, validating the bound we obtained in theorem 3.1. The y-axis has been cropped at 0.5 for representation purposes. Also, we note that we require only around 550 measurements to recover the signal using the lasso. The disparity between this and the theoretical bound in [24] arises because we can only upper bound s in our experiments, and it is more than likely that $s << 55$.



(a)

Measurements	MSE OG lasso	MSE lasso
100	204.31	481.61
200	73.92	238.54
300	26.97	118.95
400	9.89	58.75
500	3.46	29.45
600	1.25	14.44
700	0.45	7.12
800	0.15	3.51
900	0.05	1.71

(b)

Figure 3: Performance on DWT coeffecients (a) and the "metastasis" part of the gene data (b)

4.3 Gene expression data, overlapping pathways

Finally, we test our recovery bounds on gene expression data for breast cancer detection. The dataset we use is the same as that used in [8]. The genes are grouped into 639 pathways, [22]. Like [8], we restrict our attention to only 3310 genes that are at least in one pathway.

We try to predict the values of the sparse gene expression data, based on observing a limited number of compressive measurements. For the sake of consistency, we limit ourselves to only one class among the 2 in the dataset. We note that the gene expression data is nearly sparse, meaning a very large fraction of elements are close to zero. We retain the 20 "most active" groups (in terms of magnitude), to sparsify the data. We use a randomly selected set of 50 measurements for training λ, and the remaining for testing, from the section of the data that corresponds to tumors leading to metastasis. Figure 3(b) shows the mean reconstruction error over the test set, for a range of random measurements considered for every vector.

5 Conclusion

We showed that, when additional structure about the support of the signal to be estimated is known, we can recover the signal in much fewer measurements that what would be needed in the standard compressed sensing framework. Also, we showed that we surprisingly do not pay an extra penalty when the groups overlap each other. Moreover, the bound holds for arbitrary group structures, and can be used in a variety of compressed-sensing applications. The bounds we derive are tight, and can be extended to subgaussian measurement matrices by incurring a constant penalty.

References

[1] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Scientific Computing*, 20(1):33–61, 1998.

[2] Y. Gordon. On milmans inequality and random subspaces which escape through a mesh in Rn. *Geometric aspects of functional analysis, Isr. Semin.*, 1317:84–106, 1986 - 87.

[3] Candes E. J., Romberg J., and Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Information Theory*, 52:489–509, 2006.

[4] Huang J. and Zhang T. The benefit of group sparsity. *Technical report, arXiv:0901.2962. Preprint available at http://arxiv.org/pdf/0903.2962v2*, May 2009.

[5] Huang J., Zhang T., and Metaxas D. Learning with structured sparsity. *Technical report, arXiv:0903.3002. Preprint available at http://arxiv.org/pdf/0903.3002v2*, May 2009.

[6] Romberg J.K, Choi H., and Baraniuk R.G. Bayesian tree structured image modeling using wavelet domain hidden markov models. *Transactions on Image Processing*, March 2000.

[7] Donoho D. L. Compressed sensing. *IEEE Trans. Information Theory*, 52:1289–1306, 2006.

[8] Jacob L., Obozinski G., and Vert J.P. Group lasso with overlap and graph lasso. *Proceedings of the 26th International Conference on machine Learning*, 2009.

[9] Yuan M. and Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the royal statistical society. Series B*, 68:49–67, 2006.

[10] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Processing*, pages 3397–3415, Dec 1993.

[11] Duarte M.F., Cevher V., and Baraniuk R.G. Model-based compressive sensing for signal ensembles. *Allerton*, 2009.

[12] M. Mishali and Y. Eldar. Blind multi-band signal reconstruction: compressed sensing for analog signals. *IEEE Trans. Signal Processing*, 57(30):993–1009, March 2009.

[13] Crouse M.S., Nowak R.D., and Baraniuk R.G. Wavelet based statistical signal processing using hidden markov models. *Transactions on Signal Processing*, 46(4):886–902, 1998.

[14] Jenatton R., Mairal J., Obozinski G., , and Bach F. Proximal methods for hierarchical sparse coding. *Technical report, arXiv:1009.3139. submitted*, 2010.

[15] Jenatton R., Audibert J.Y., and Bach F. Structured variable selection with sparsity inducing norms. *Technical report, arXiv:0904.3523. Preprint available at http://arxiv.org/pdf/0904.3523v3*, Sep 2009.

[16] N. Rao, R. Nowak, S. Wright, and N. Kingsbury. Convex approaches to model wavelet sparsity patterns. *IEEE International Conference on Image Processing*, 2011.

[17] Baraniuk R.G., Cevher V., M.F. Duarte, and Hegde C. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 2010.

[18] T. Rockafellar and J. B. Wets. Variational analysis. *Springer Series of Comprehensive Studies in Mathematics*, 317, 1997.

[19] Mendelson S., Pajor A., and Tomczak-Jaegermann N. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248 – 1282, 2006.

[20] Mosci S., Villa S., Verri A., and Rosasco L. A primal-dual algorithm for group sparse regularization with overlapping groups. *Neural Information Processing Systems*, 2010.

[21] Wright S.J., Nowak R.D., and Figueiredo M.A.T. Sparse reconstruction by separable approximation. *Transactions on Signal Processing*, 57:2479–2493, 2009.

[22] A. Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression proles. *National Academy of Sciences*, 102:1554515550, 2005.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

[24] Chandrasekaran V., Recht B., Parrilo P. A., and Willsky A. The convex geometry of linear inverse problems. *preprint arXiv:1012.0621v1*, 2010.