Classification with Sparse Overlapping Groups

Nikhil S. Rao NRAO2@WISC.EDU
Robert D. Nowak NOWAK@ECE.WISC.EDU
Department of Electrical and Computer Engineering
University of Wisconsin-Madison

Christopher R. Cox CRCOX@WISC.EDU
Timothy T. Rogers TTROGERS@WISC.EDU
Department of Psychology
University of Wisconsin-Madison

Editor:

Abstract

Classification with a sparsity constraint on the solution plays a central role in many high dimensional machine learning applications. In some cases, the features can be grouped together, so that entire subsets of features can be selected or not selected. In many applications, however, this can be too restrictive. In this paper, we are interested in a less restrictive form of structured sparse feature selection: we assume that while features can be grouped according to some notion of similarity, not all features in a group need be selected for the task at hand. When the groups are comprised of disjoint sets of features, this is sometimes referred to as the "sparse group" lasso, and it allows for working with a richer class of models than traditional group lasso methods. Our framework generalizes conventional sparse group lasso further by allowing for overlapping groups, an additional flexiblity needed in many applications and one that presents further challenges. The main contribution of this paper is a new procedure called *Sparse Overlapping Group (SOG) lasso*, a convex optimization program that automatically selects similar features for classification in high dimensions. We establish model selection error bounds for SOGlasso classification problems under a fairly general setting. In particular, the error bounds are the first such results for classification using the sparse group lasso. Furthermore, the general SOGlasso bound specializes to results for the lasso and the group lasso, some known and some new. The SOGlasso is motivated by multi-subject fMRI studies in which functional activity is classified using brain voxels as features, source localization problems in Magnetoencephalography (MEG), and analyzing gene activation patterns in microarray data analysis. Experiments with real and synthetic data demonstrate the advantages of SOGlasso compared to the lasso and group lasso.

1. Introduction

Binary classification plays a major role in many machine learning and signal processing applications. In many modern applications where the number of features far exceeds the number of samples, one typically wishes to select only a few features, meaning only a few coefficients are non zero in the solution [1]. This corresponds to the case of searching for

[1] a zero coefficient in the solution implies the corresponding feature is not selected

sparse solutions. The notion of sparsity prevents over-fitting and leads to more interpretable solutions in high dimensional machine learning, and has been extensively studied in (Bach, 2010; Plan and Vershynin, 2013; Negahban et al., 2012; Bunea, 2008), among others.

In many applications, we wish to impose structure on the sparsity pattern of the coefficients recovered. In particular, often it is known a priori that the optimal sparsity pattern will tend to involve clusters or groups of coefficients, corresponding to pre-existing groups of features. The form of the groups is known, but the subset of groups that is relevant to the classification task at hand is unknown. This prior knowledge reduces the space of possible sparse coefficients thereby potentially leading to better results than simple lasso methods. In such cases, the group lasso, with or without overlapping groups (Yuan and Lin, 2006) is used to recover the coefficients. The group lasso forces all the coefficients in a group to be active at once: if a coefficient is selected for the task at hand, then all the coefficients in that group are selected. When the groups overlap, a modification of the penalty allows one to recover coefficients that can be expressed as a union of groups (Jacob et al., 2009; Obozinski et al., 2011).

While the group lasso has enjoyed tremendous success in high dimensional feature selection applications, we are interested in a much less restrictive form of structured feature selection for classification. Suppose that the features can be arranged into (possibly) *overlapping* groups based on some notion of similarity, depending on the application. The notion of similarity can be loosely defined, and it is used to reflect the prior knowledge that if a feature is relevant for the learning task at hand, then features similar to it may also be relevant. It is known that while many features may be similar to each other, not all similar features are relevant for the specific learning problem. We propose a new procedure called Sparse Overlapping Group (SOG) lasso to reflect this form of structured sparsity.

As an example, consider the task of identifying relevant genes that play a role in predicting a disease. Genes are organized into pathways (Subramanian et al., 2005), but not every gene in a pathway might be relevant for prediction. At the same time, it is reasonable to assume that if a gene from a particular pathway is relevant, then other genes from the same pathway may also be relevant. In such applications, the group lasso may be too constraining while the lasso may be too under-constrained.

1.1 Model and Results

We first present the main results of this paper at a glance. Uppercase and lowercase bold letters indicate matrices and vectors respectively. We assume a sparse learning framework, with a feature matrix $\mathbf{\Phi} \in \mathbb{R}^{n \times p}, \quad n \ll p$. We assume each element of $\mathbf{\Phi}$ to be distributed as a standard Gaussian random variable. Assuming the data to arise from a Gaussian distribution simplifies analysis, and allows us to leverage tools from existing literature. Later in the paper, we will allow for correlations in the features as well, reflecting a more realistic setting. In the results that follow, C is a positive constant, the value of which can be different from one result to the other.

We focus on binary classification settings, and assume that each observation $\boldsymbol{y}_i \in \{-1, +1\}, \quad i = 1, 2, \ldots, n$ are randomly distributed according to the model (Plan and Vershynin, 2013)

$$\mathbb{E}[\boldsymbol{y}_i | \boldsymbol{\phi}_i] = f(\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle), \tag{1}$$

where $\boldsymbol{\phi}_i$ is the i^{th} row of $\boldsymbol{\Phi}$ corresponding to the features of data i, $\boldsymbol{x}^\star$ is the true coefficient vector of interest, and f is a function mapping from $\mathbb{R}$ to $[-1,+1]$. The argument of f is the Euclidean inner product: $\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle = \boldsymbol{\phi}_i^T \boldsymbol{x}^\star$. The function f need not be known precisely. We only assume that it satisfies for $g \sim \mathcal{N}(0,1)$

$$\mathbb{E}[f(g)g] > 0 \ . \tag{2}$$

Without loss of generality, we assume $\boldsymbol{x}^\star$ to have unit Euclidean norm, since the normalization can be absorbed into the function f. The value

$$\sigma_f := 1/\mathbb{E}[f(g)g] \tag{3}$$

quantifies the strength of the correlation between the labels y_i and inner products $\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle$. It plays the role of the noise level and will appear in our error bounds, but it need not be known to compute our proposed estimator of $\boldsymbol{x}^\star$.

This set-up allows for the consideration of a very general setting for classification, and subsumes many interesting cases. For example, the logistic model

$$\mathbb{P}\left(\boldsymbol{y}_i = 1\right) = \frac{\exp(\beta \left\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \right\rangle)}{1 + \exp(\beta \left\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \right\rangle)} \ , \tag{4}$$

is equivalent to

$$f(\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle) = \tanh(\beta \langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle),$$

for any constant $\beta > 0$, yielding a corresponding [2]

$$\sigma_f = \frac{2}{\beta} \mathbb{E}[\text{sech}^2(\beta \boldsymbol{g}/2)]^{-1} \leq \frac{6}{\min\{\beta, 1\}}.$$

The constant β accounts for the fact that we consider $\|\boldsymbol{x}^\star\| = 1$. Indeed, for a general vector $\boldsymbol{z}$, we can write $\langle \boldsymbol{\phi}_i, \boldsymbol{z} \rangle = \beta \langle \boldsymbol{\phi}_i, \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|} \rangle$, where $\beta = \|\boldsymbol{z}\|$. The second argument in the inner product is now a vector of unit norm, and it gives rise to the expression in (4).

The framework also allows for the quantized 1-bit measurement model

$$f(\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle) = \text{sign}(\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle) \ ,$$

which can be seen as the limiting case of the logistic model as $\beta \to \infty$, and with $\sigma_f = \sqrt{\frac{\pi}{2}}$

We work with this general formulation since for classification problems of interest, the logistic (or any other) model may be chosen somewhat arbitrarily. Existing theoretical results often apply only to the chosen model (for example, (Negahban et al., 2012)). Our estimator only requires the observations are correlated with the features, in the sense of (3). In any such case, the underlying form of f need not be known to compute our estimator of $\boldsymbol{x}^\star$ and will enter in the error bounds only through σ_f.

We are interested in the following form of structured sparsity. Assume that the features can be organized into K *possibly overlapping* groups, each consisting of L features, based on a user-defined measure of similarity, depending on the application. Moreover, assume that if a certain feature is relevant for the learning task at hand, then features similar to it may

[2] See Corollary 3.3 in (Plan and Vershynin, 2013) for a derivation

also be relevant. Note that we assume groups of equal size L for convenience. It is easy to relax this assumption. These assumptions suggest a structured pattern of sparsity in the coefficients wherein a subset of the groups are relevant to the learning task, and within the relevant groups a subset of the features are selected. In other words, $\boldsymbol{x}^\star \in \mathbb{R}^p$ has the following structure:

- its support is localized to a union of a subset of the groups, and
- its support is localized to a sparse subset within each such group

Armed with these preliminaries, we state the theoretical sample complexity bounds proved later in the paper.

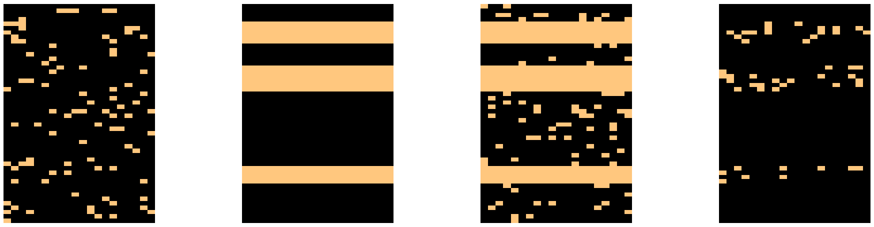
If $\boldsymbol{x}^\star \in \mathbb{R}^p$ has $k \leq K$ non-zero groups and $l \leq L$ coefficients non-zero within each non-zero group, then $\mathcal{O}(\sigma_f^2\, k \left[\log(\frac{K}{k}) + l \log(\frac{L}{l}) + l\right])$ independent Gaussian measurements of the form (1) are sufficient to accurately estimate $\boldsymbol{x}^\star$ by solving a convex program.

The statement above merits further explanation. We show in Lemma 12 (Section 4.1) via a combinatorial argument that to estimate any vector with parameters k, K, l, L as stated above, $\mathcal{O}(\sigma_f^2 k \left[\log(\frac{K}{k}) + l \log(\frac{L}{l}) + l\right])$ samples would suffice. However, looking for vectors with these properties amounts to solving a non-convex program. When the groups do not overlap, we show that the solution to a *convex* program also succeeds in accurately estimating $\boldsymbol{x}^\star$ using the same number of measurements. When the groups overlap, we show that the measurement bound holds with an additional factor of R^2, where $R \geq 1$ is the maximum number of groups that contain any one feature. In most applications of interest (e.g., the fMRI example discussed below), R is small. Nonetheless, we also show that no matter how many groups a particular coefficient belongs to, the sample complexity of our proposed estimator is never greater than that of the standard lasso or the overlapping group lasso. These statements will be made more precise in the sequel.

1.2 Motivation: The SOGlasso for Multitask Learning

The SOG lasso is motivated in part by multitask learning applications. The group lasso is a commonly used tool in multitask learning, and it encourages the same set of features to be selected across all tasks. We wish to focus on a less restrictive version of multitask learning, where the main idea is to encourage selection of features that are similar, but not identical, across tasks. This is accomplished by defining subsets of similar features and searching for solutions that select only a few subsets (common across tasks) and a sparse number of features within each subset (possibly different across tasks). Figure 1 shows an example of the patterns that typically arise in sparse multitask learning applications, along with the one we are interested in.

A major application that we are motivated by is the analysis of multi-subject fMRI data, where the goal is to predict a cognitive state from measured neural activity using voxels as features. Because brains vary in size and shape, neural structures can be aligned only crudely. Moreover, neural codes can vary somewhat across individuals (Feredoes et al., 2007). Thus, neuroanatomy provides only an approximate guide as to where relevant information is located across individuals: a voxel useful for prediction in one participant suggests



(a) Sparse

(b) Group sparse

(c) Group *plus* sparse

(d) Group *and* sparse

Figure 1: A comparison of different sparsity patterns in the multitask learning setting. Figure (a) shows a standard sparsity pattern. An example of group sparse patterns promoted by Glasso (Yuan and Lin, 2006) is shown in Figure (b). In Figure (c), we show the patterns considered in (Jalali et al., 2010). Finally, in Figure (d), we show the patterns we are interested in this paper. The groups are sets of rows of the matrix, and can overlap with each other

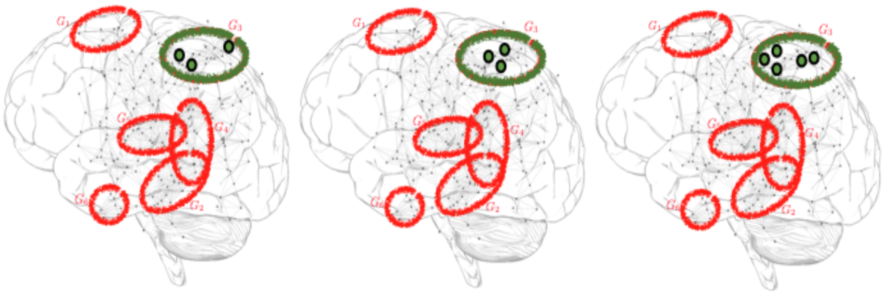


Figure 2: SOGlasso for fMRI inference. The figure shows three brains, and voxels in a particular anatomical region are grouped together, across all individuals (red and green ellipses). For example, the green ellipse in the brains represents a single group. The groups denote anatomically similar regions in the brain that may be co-activated. However, within activated regions, the exact location and number of voxels may differ, as seen from the green spots.

the general anatomical neighborhood where useful voxels may be found, but not the precise voxel. Past work in inferring sparsity patterns across subjects has involved the use of groupwise regularization (van Gerven et al., 2009), using the logistic lasso to infer sparsity patterns without taking into account the relationships across different subjects (Ryali et al., 2010), or using the elastic net penalty to account for groupings among coefficients (Rish et al., 2012). These methods do not exclusively take into account both the common macrostructure and the differences in microstructure across brains, and the SOGlasso allows one to model both the commonalities and the differences across brains. Figure 2 sheds light on the motivation, and the grouping of voxels across brains into overlapping groups

1.3 Our Contributions

In this paper, we consider binary classification with a constraint on the structure of the sparsity pattern of the coefficients. We assume that the coefficients can be arranged (according to a predefined notion of similarity) into (overlapping) groups. Not only are only a few groups selected, but the selected groups themselves are also sparse. In this sense, our constraint can be seen as an extension of sparse group lasso (Simon et al., 2013) for overlapping groups where the sparsity pattern lies in a union of groups. We are mainly interested in classification problems, but the method can also be applied to regression settings, by making an appropriate change in the loss function of course. We consider a union-of-groups formulation as in (Jacob et al., 2009), but with an additional sparsity constraint on the selected groups. To this end, we analyze the Sparse Overlapping Sets (SOG) lasso, where the overlapping groups might correspond to coefficients of features arbitrarily grouped according to the notion of similarity. We also consider a very general classification setting, and do not make restrictive assumptions on the observation model.

We introduce a constraint that promotes sparsity patterns that can be expressed as a union of sparsely activated groups. We show that the constraint is a tight convex relaxation of the set of coefficients having the sparsity pattern we are interested in. The main contribution of this paper is a theoretical analysis of the model selection consistency of the SOGlasso estimator, under a very general binary classification setting. Based on certain parameter values, our method reduces to other known cases of penalization for sparse high dimensional recovery. Specifically, under the logistic regression model, our method generalizes the group lasso (Meier et al., 2008; Jacob et al., 2009), and also extends to handle groups that can arbitrarily overlap with each other. We also recover results for the lasso for logistic regression (Bunea, 2008; Negahban et al., 2012; Plan and Vershynin, 2013; Bach, 2010). In this sense, our work unifies the lasso, the group lasso as well as the sparse group lasso for to handle overlapping groups. At the same time, our methods apply to settings beyond the logistic regression, to include a far richer class of models. To the best of our knowledge, this is the first paper that provides such a unified theory and sample complexity bounds for all these methods.

In the case of linear regression and multitask learning, the authors in (Sprechmann et al., 2010, 2011), consider a similar situation with non overlapping subsets of features. We assume that the features can arbitrarily overlap. When the groups overlap, the methods mentioned above suffer from a drawback: entire groups are set to zero, in effect zeroing out many coefficients that might be relevant to the tasks at hand. This has undesirable effects in many applications of interest, and the authors in (Jacob et al., 2009) propose a version of the group lasso to circumvent this issue.

We also test our regularizer on both toy and real datasets. Our experiments reinforce our theoretical results, and demonstrate the advantages of the SOGlasso over standard lasso and group lasso methods, when the features can indeed be grouped according to some notion of similarity. We show that the SOGlasso is especially useful in multitask Functional Magnetic Resonance Imaging (fMRI) applications, and gene selection applications in computational biology.

To summarize, the main contributions of this paper are the following:

1. **New regularizers for structured sparsity:** We propose the Sparse Overlapping Group (SOG) lasso, a convex optimization problem that encourages the selection of coefficients that are both within-and across- group sparse. The groups can arbitrarily overlap, and the pattern obtained can be represented as a union of a small number of groups. This generalizes other known methods, and provides a common regularizer that can be used for any structured sparse problem with two levels of hierarchy [3]: groups at a higher level, and singletons at the lower level.

2. **New theory for classification with structured sparsity:** We provide a theoretical analysis for the model selection consistency of the SOGlasso estimator, for binary classification. The general results we obtain specialize to the lasso, the group lasso (with or without overlapping groups) and the sparse group lasso. We also make minimal assumptions on the measurement model, allowing for theory that is applicable in a wide range of classification settings. We obtain a bound on the sample complexity of the SOGlasso under both independent and correlated Gaussian measurement designs, and this in turn also translates to corresponding results for the lasso and the group lasso. In this sense, we obtain a unified theory for performing structured feature selection in high dimensions.

3. **Applications:** A major motivating application for this work is the analysis of multi-subject fMRI data. We apply the SOGlasso to fMRI data, and show that the results we obtain not only yield lower errors on hold-out test sets compared to previous methods, but also lead to more interpretable results. To show it's applicability to other domains, we also apply the method to breast cancer data to detect genes that are relevant in the prediction of metastasis in breast cancer tumors.

In (Rao et al., 2013), the authors introduced the SOGlasso problem emphasizing the motivating fMRI application. The authors also derived theoretical consistency results under the *linear regression* setting. This paper gives further embellishes the reasons for considering the SOGlasso penalty, and derives consistency results for *classification*. We also present some novel applications in computational biology where similar notions can be applied to achieve significant gains over existing methods. Our work here presents novel results for the group lasso with potentially overlapping groups as well as the sparse group lasso for classification settings, as special cases of the theory we develop.

1.4 Past Work

When the groups of features do not overlap, (Simon et al., 2013) proposed the Sparse Group Lasso (SGL) to recover coefficients that are both within- and across- group sparse. SGL and its variants for multitask learning has found applications in character recognition (Sprechmann et al., 2011, 2010), climate and oceanology applications (Chatterjee et al., 2011), and in gene selection in computational biology (Simon et al., 2013). In (Jenatton et al., 2011), the authors extended the method to handle tree structured sparsity patterns, and showed that the resulting optimization problem admits an efficient implementation in terms of proximal point operators. Along related lines, the exclusive lasso (Zhou et al.,

[3] Further levels can also be added as in (Jenatton et al., 2011), but that is beyond the scope of this paper.

2010) can be used when it is explicitly known that features in certain groups are negatively correlated. When the groups overlap, (Jacob et al., 2009; Obozinski et al., 2011) proposed a modification of the group lasso penalty so that the resulting coefficients can be expressed as a union of groups. They proposed a replication-based strategy for solving the problem, which has since found application in computational biology (Jacob et al., 2009) and image processing (Rao et al., 2011), among others. The authors in (Mosci et al., 2010) proposed a method to solve the same problem in a primal-dual framework, that does not require coefficient replication. Risk bounds for problems with structured sparsity inducing penalties (including the lasso and group lasso) were obtained by (Maurer and Pontil, 2012) using Rademacher complexities. Sample complexity bounds for model selection in linear regression using the group lasso (with possibly overlapping groups) also exist (Rao et al., 2012). The results naturally hold for the standard group lasso (Yuan and Lin, 2006), since non overlapping groups are a special case. For the non overlapping case, (Stojnic et al., 2009) characterized the sample complexity of the group lasso, and also gave a semidefinite program to solve the group lasso under a block sparsity setting.

For logistic regression, (Bach, 2010; Bunea, 2008; Negahban et al., 2012; Plan and Vershynin, 2013) and references therein have extensively characterized the sample complexity of identifying the correct model using ℓ_1 regularized optimization. The authors in (Negahban et al., 2012) extended their results to include Generalized Linear Models as well (GLM's). In (Plan and Vershynin, 2013), the authors introduced a new framework to solve the classification problem: minimize [4] a linear cost function subject to a constraint on the ℓ_1 norm of the solution.

1.5 Organization

The rest of the paper is organized as follows: in Section 2, we formally state our structured sparse feature selection problem and the main results of this paper. Then in Section 3, we argue that the regularizer we propose does indeed help in recovering coefficient sparsity patterns that are both within-and across group sparse, even when the groups overlap. In Section 4, we leverage ideas from (Plan and Vershynin, 2013) and derive measurement bounds and consistency results for the SOGlasso under a logistic regression setting. We also extend these results to handle data with correlations in their entries. We perform experiments on real and toy data in Section 5, before concluding the paper and mentioning avenues for future research in Section 6.

2. Main Results: Classification with Structured Sparsity

We now return to the problem that we wish to solve in this paper, and state our main results in a more formal way. Recall that we are interested in recovering a coefficient vector $\boldsymbol{x}^\star$, from (corrupted) linear observations of the form $\langle \boldsymbol{\phi}_i, \boldsymbol{x}^\star \rangle$

The coefficient vector of interest is assumed to have a special structure. Specifically, we assume that $\boldsymbol{x}^\star \in \mathcal{C} \subset B_2^p$, where B_2^p is the unit euclidean ball in $\mathbb{R}^p$. This motivates the

[4] The authors in (Plan and Vershynin, 2013) write the problem as a maximization. We minimize the negative of the same function

following optimization problem (Plan and Vershynin, 2013):

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{n} -\boldsymbol{y}_i \langle \boldsymbol{\phi}_i, \boldsymbol{x} \rangle \quad \textbf{s.t.} \quad \boldsymbol{x} \in \mathcal{C}. \tag{5}$$

The function to be optimized has a very natural interpretation: We assume without loss of generality that the observations are positively correlated [5] with the inner products between the features and the coefficient vector. Hence, a natural thing to do would be to maximize the number of "sign agreements" between $\boldsymbol{y}_i$ and $\langle \boldsymbol{\phi}_i, \boldsymbol{x} \rangle$. The objective function in (5) maximizes the product of the two terms, a linear relaxation of the quantity we wish to optimize.

The statistical accuracy of $\widehat{x}$ can be characterized in terms of the *mean width* of $\mathcal{C}$, which is defined as follows

Definition 1 *Let $\boldsymbol{g} \in \mathcal{N}(0, \boldsymbol{I})$. The mean width of a set $\mathcal{C}$ is defined as*

$$\omega(\mathcal{C}) = \mathbb{E}_{\boldsymbol{g}} \left[\sup_{\boldsymbol{x} \in \mathcal{C} - \mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{g} \rangle \right],$$

where $\mathcal{C} - \mathcal{C}$ denotes the Minkowski set difference.

We now restate a result from (Plan and Vershynin, 2013)

Theorem 2 (Corollary 1.2 in (Plan and Vershynin, 2013)) *Let $\boldsymbol{\Phi} \in \mathbb{R}^{n \times p}$ be a matrix with i.i.d. standard Gaussian entries, and let $\mathcal{C} \subset B_2^p$. Fix $\boldsymbol{x}^\star \in \mathcal{C}$, and assume the observations follow the model (1) above. Then, for $\epsilon > 0$, if*

$$n \geq C \left(\frac{\sigma_f^2}{\epsilon^2} \right) \omega(\mathcal{C})^2,$$

then with probability at least $1 - 8\exp(-c(\frac{\epsilon}{\sigma_f})^2 n)$, the solution $\widehat{\boldsymbol{x}}$ to the problem (5) satisfies

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|^2 \leq \epsilon$$

with σ_f defined in (3).

We abuse notation and define the mean width as

$$\omega(\mathcal{C}) = \mathbb{E}_{\boldsymbol{g}} \left[\sup_{\boldsymbol{x} \in \mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{g} \rangle \right]. \tag{6}$$

The quantity defined above is a constant multiple of that in the original definition for centrally symmetric sets $\mathcal{C}$ (Plan and Vershynin, 2013), which will be the case for the remainder of this paper.

In this paper, we construct a new penalty that produces a *convex* set $\mathcal{C}$ that encourages structured sparsity in the solution of (5). We show that the resulting optimization can be efficiently solved. We bound the mean width of the set, which yields new bounds for classification with structured sparsity, via Theorem 2. We state the main results in this section, and defer the proofs to Section 4

[5] If the correlation is negative, the signs of $\boldsymbol{y}_i$ can be reversed

2.1 A New Penalty for Structured Sparsity

Assume that the features can be grouped according to similarity into K (possibly overlapping) groups $\mathcal{G} = \{G_1, G_2, \ldots, G_K\}$ with the largest group being of size L and consider the following definition of structured sparsity.

Definition 3 *We say that a vector $\boldsymbol{x}$ is (k, l)-group sparse if $\boldsymbol{x}$ is supported on at most $k \leq K$ groups and at most l elements in each active group are non zero.*

Note that $l = 0$ corresponds to $\boldsymbol{x} = \mathbf{0}$.

To encourage such sparsity patterns we define the following penalty. Given a group $G \in \mathcal{G}$, we define the set

$$\mathcal{W}_G = \{\boldsymbol{w} \in \mathbb{R}^p : \quad \boldsymbol{w}_i = 0 \ \ \textbf{if} \ \ i \notin G\} .$$

We can then define

$$\mathcal{W}(\boldsymbol{x}) = \left\{ \boldsymbol{w}_{G_1} \in \mathcal{W}_{G_1}, \quad \boldsymbol{w}_{G_2} \in \mathcal{W}_{G_2}, \ldots, \boldsymbol{w}_{G_M} \in \mathcal{W}_{G_M} : \quad \sum_{G \in \mathcal{G}} \boldsymbol{w}_G = \boldsymbol{x} \right\} .$$

That is, each element of $\mathcal{W}(x)$ is a set of vectors, one from each $\mathcal{W}_G$, such that the vectors sum to $\boldsymbol{x}$. As shorthand, in the sequel we write $\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})$ to mean a set of vectors that form an element in $\mathcal{W}(\boldsymbol{x})$

For any $\boldsymbol{x} \in \mathbb{R}^p$, define

$$h(\boldsymbol{x}) \ := \inf_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} \left(\alpha_G \|\boldsymbol{w}_G\|_2 + \beta_G \|\boldsymbol{w}_G\|_1 \right) , \tag{7}$$

where the $\alpha_G, \beta_G > 0$ are constants that tradeoff the contributions of the ℓ_2 and the ℓ_1 norm terms per group, respectively. The *SOGlasso* is the optimization in (5) with $h(\boldsymbol{x})$ as defined in (7) determining the structure of the constraint set $\mathcal{C}$, and hence the form of the solution $\widehat{\boldsymbol{x}}$. The ℓ_2 penalty promotes the selection of only a subset of the groups, and the ℓ_1 penalty promotes the selection of only a subset of the features within a group.

To keep the exposition simple, we will work with the following definition of $h(\boldsymbol{x})$ in the rest of the paper:

$$h(\boldsymbol{x}) \ := \inf_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} \left(\|\boldsymbol{w}_G\|_2 + \frac{\lambda_1}{\sqrt{l}} \|\boldsymbol{w}_G\|_1 \right) . \tag{8}$$

Note that the value of $\lambda_1 \geq 0$ can be varied to both emphasize or de emphasize the ℓ_1 penalty. In almost all applications of interest, the value of l will obviously be unknown, and the quantity $\frac{\lambda_1}{\sqrt{l}}$ needs to be tuned via cross validation. However, for the sake of proving our theorems, we will assume that the quantity is known (our goal is to recover vectors that are (k, l)- group sparse). This is consistent with other results in the literature where it is assumed that the parameters are known.

Definition 4 *We say the set of vectors $\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})$ is an optimal representation of $\boldsymbol{x}$ if they achieve the* inf *in (8).*

The objective function in (8) is convex and coercive. Hence, $\forall \boldsymbol{x}$, an optimal representation always exists.

The function $h(\boldsymbol{x})$ yields a convex relaxation for (k,l)-group sparsity. Define the constraint set

$$\mathcal{C} = \{\boldsymbol{x} : h(\boldsymbol{x}) \leq \sqrt{k}\,(1+\lambda_1)\,, \quad \|\boldsymbol{x}\|_2 \leq 1\}\ . \tag{9}$$

We show that $\mathcal{C}$ is convex and contains all (k,l)-group sparse vectors. We compute the mean width of $\mathcal{C}$ in (9), and subsequently obtain the following result:

Theorem 5 *Suppose there exists a coefficient vector $\boldsymbol{x}^\star$ that is (k,l)-group sparse. Suppose the data matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n\times p}$ and observation model follow the setting in Theorem 2. Suppose we solve (5) for the constraint set given by (9). For $\epsilon > 0$ and a constant C, if the number of measurements satisfies*

$$n \ \geq \ C\frac{\sigma_f^2}{\epsilon^2}k \ \min\left\{(1+\lambda_1)^2(\log(K)+L), \left(\frac{1+\lambda_1}{\lambda_1}\right)^2 l\log(p)\right\},$$

then with high probability, the solution of the SOGlasso satisfies

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2^2 \ \leq \epsilon.$$

Remarks

The results of Theorem 5 yield new results for classification with structured sparsity under the general binary observation setting (1). Specifically, note that the SOGlasso interpolates between the standard ℓ_1 regularized (lasso) and the group ℓ_1 regularized (group lasso) classification techniques:

- When $\lambda_1 = 0$, we obtain results for the group lasso. The result remains the same whether or not the groups overlap. The bound is given by

  $$n \geq C\delta^{-2}k(\log(K)+L).$$

  Note that this result is similar to that obtained for the linear regression case by the authors in (Rao et al., 2012).

- When all the groups are singletons, $(L = l = 1)$, the bound reduces to that for the standard lasso, with $KL = p$ being the ambient dimension. In this case, the signal sparsity $s := kl$ and the bound becomes:

  $$n \geq C\delta^{-2}kl\log(p).$$

The SOGlasso generalizes the lasso and the group lasso, and allows one to recover signals that are sparse, group sparse, or a combination of the two structures.

Moreover, since the model we consider subsumes the logistic regression setting, we obtain results for logistic regression with a general structured sparsity constraint on the solution. To the best of our knowledge, these are the first known sample complexity bounds for the group lasso for logistic regression with overlapping groups, and the sparse group lasso, both of which arise as special cases of the SOGlasso.

Problem (5) admits an efficient solution. Specifically, we can use the feature replication strategy as in (Jacob et al., 2009) to reduce the problem to a sparse group lasso, and use proximal point methods to recover the coefficient vector. We elaborate this in more detail later in the paper.

Theorem 5 bounds the number of measurements sufficient for accurate estimation using the overlapping group lasso and the lasso. A natural question to then ask is whether one can do better when it is known that the vectors we are interested in are further constrained by the number of non zero entries in active groups. When it is known that each coefficient belongs to at most R groups [6], we obtain the following sample complexity bound

Theorem 6 *Suppose the coefficient vector is (k, l) group sparse, with everything else the same as in Theorem 5. Suppose that each coefficient belongs to at most R groups. Then, with high probability, if the number of measurements satisfies*

$$n \geq C \frac{\sigma_f^2}{\epsilon^2} R^2 k \left[\log\left(\frac{K}{k}\right) + l \log\left(\frac{L}{l}\right) + l + 2 \right],$$

we have

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\star\|^2 \leq \epsilon.$$

The above result yields a tight bound when the groups do not overlap. Indeed, when $R = 1$ we see that the sample complexity bound is a function of the logarithm of the number of groups, and the overall sparsity of the signal kl. This is a tighter result that the bound obtained for the group lasso, where the second term would be $\mathcal{O}(kL)$. We pay a price of R^2 for overlapping groups, but in most practical examples we are interested in, R is typically small, and is a constant.

3. Analysis of the SOGlasso Penalty

Recall the definition of $h(\boldsymbol{x})$, from (8):

$$h(\boldsymbol{x}) = \inf_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} \|\boldsymbol{w}_G\|_2 + \mu \|\boldsymbol{w}_G\|_1 \tag{10}$$

where we set $\mu = \frac{\lambda_1}{\sqrt{l}}$

Remarks :

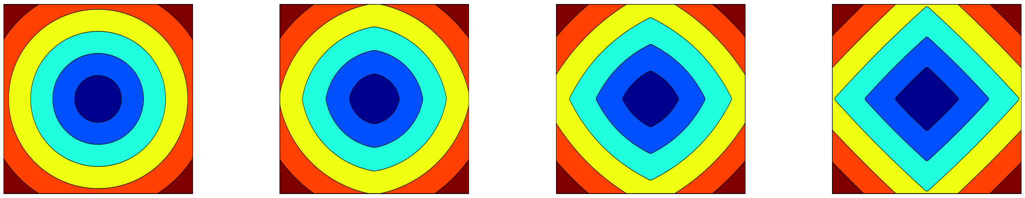
The SOGlasso penalty can be seen as a generalization of different penalty functions previously explored in the context of sparse linear regression and/or classification:

- If each group in $\mathcal{G}$ is a singleton, then the SOGlasso penalty reduces to the standard ℓ_1 norm, and the problem reduces to the lasso (Tibshirani, 1996; Bunea, 2008)
- if $\lambda_1 = 0$ in (8), then we are left with the latent group lasso (Jacob et al., 2009; Obozinski et al., 2011; Rao et al., 2012). This allows us to recover sparsity patterns that can be expressed as lying in a union of groups. If a group is selected, then all the coefficients in the group are selected.

[6] The value of R will always be known, since we assume that the groups $\mathcal{G}$ are known

- If the groups $G \in \mathcal{G}$ are non overlapping, then (8) reduces to the sparse group lasso (Simon et al., 2013). Of course, for non overlapping groups, if $\lambda_1 = 0$, then we get the standard group lasso (Yuan and Lin, 2006).

Figure 3 shows the effect that the parameter μ has on the shape of the "ball" $\|\boldsymbol{w}_G\| + \mu\|\boldsymbol{w}_G\|_1 \leq \delta$, for a single two dimensional group G.



(a) $\mu = 0$ (b) $\mu = 0.2$ (c) $\mu = 1$ (d) $\mu = 10$

Figure 3: Effect of μ on the shape of the set $\|\boldsymbol{w}_G\| + \mu\|\boldsymbol{w}_G\|_1 \leq \delta$, for a two dimensional group G. $\mu = 0$ (a) yields level sets for the ℓ_2 norm ball. As the value of μ in increased, the effect of the ℓ_1 norm term increases (b) (c). Finally as μ gets very large, the level sets resemble that of the ℓ_1 ball (d).

3.1 Properties of SOGlasso Penalty

The example in Table 1 gives an insight into the kind of sparsity patterns preferred by the function $h(\boldsymbol{x})$. We will tend to prefer solutions that have a small value of $h(\cdot)$. Consider 3 instances of $\boldsymbol{x} \in \mathbb{R}^{10}$, and the corresponding group lasso, ℓ_1 norm, and $h(\boldsymbol{x})$ function values. The vector is assumed to be made up of two groups, $G_1 = \{1, 2, 3, 4, 5\}$ and $G_2 = \{6, 7, 8, 9, 10\}$. $h(\boldsymbol{x})$ is smallest when the support set is sparse within groups, and also when only one of the two groups is selected (column 5). The ℓ_1 norm does not take into account sparsity across groups (column 4), while the group lasso norm does not take into account sparsity within groups (column 3). Since the groups do not overlap, the latent group lasso penalty reduces to the group lasso penalty and $h(\boldsymbol{x})$ reduces to the sparse group lasso penalty.

Support	**Values**	$\sum_G \|\boldsymbol{x}_G\|$	$\|\boldsymbol{x}\|_1$	$\sum_G (\|\boldsymbol{x}_G\| + \|\boldsymbol{x}_G\|_1)$
$\{1, 4, 9\}$	$\{3, 4, 7\}$	12	14	26
$\{1, 2, 3, 4, 5\}$	$\{2, 5, 2, 4, 5\}$	8.602	18	26.602
$\{1, 3, 4\}$	$\{3, 4, 7\}$	8.602	14	22.602

Table 1: Different instances of a 10-d vector and their corresponding norms.

The next table shows that $h(\boldsymbol{x})$ indeed favors solutions that are not only group sparse, but also exhibit sparsity within groups when the groups overlap. Consider again a 10-dimensional vector $\boldsymbol{x}$ with three overlapping groups $\{1, 2, 3, 4\}$, $\{3, 4, 5, 6, 7\}$ and $\{7, 8, 9, 10\}$. Suppose the vector $\boldsymbol{x} = [0\ \ 0\ \ 1\ \ 0\ \ 1\ \ 0\ \ 1\ \ 0\ \ 0\ \ 0]^T$. From the form of the function in (8), we see that the vector can be seen as a sum of three vectors $\boldsymbol{w}_i,\ \ i = 1, 2, 3$, corresponding to the three groups listed above. Consider the following instances of the $\boldsymbol{w}_i$ vectors, which are all feasible solutions for the optimization problem in (10):

1. $\boldsymbol{w}_1 = [0 \;\; 0 \;\; -1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_2 = [0 \;\; 0 \;\; 1 \;\; 1 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_3 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$
2. $\boldsymbol{w}_1 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_2 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_3 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0]^T$
3. $\boldsymbol{w}_1 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_2 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_3 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0]^T$
4. $\boldsymbol{w}_1 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_2 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0]^T$, $\boldsymbol{w}_3 = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]^T$

In the above list, the first instance corresponds to the case where the support is localized to two groups, and one of these groups (group 2) has only one zero. The second case corresponds to the case where all 3 groups have non zeros in them. The third case has support localized to two groups, and both groups are sparse. Finally, the fourth case has only the second group having non zero coefficients, and this group is also sparse. Table 2 shows that the smallest value of the sum of the terms is achieved by the fourth decomposition, and hence that will correspond to the optimal representation.

$A = \|\boldsymbol{w}_1\| + \mu\|\boldsymbol{w}_1\|_1$	$B = \|\boldsymbol{w}_2\| + \mu\|\boldsymbol{w}_2\|_1$	$C = \|\boldsymbol{w}_3\| + \mu\|\boldsymbol{w}_3\|_1$	$A + B + C$
$1 + \mu$	$2 + 4\mu$	0	$3 + 5\mu$
$1 + \mu$	$1 + \mu$	$1 + \mu$	$3 + 3\mu$
0	$\sqrt{2} + 2\mu$	$1 + \mu$	$1 + \sqrt{2} + 3\mu$
0	$\sqrt{3} + 3\mu$	0	$\sqrt{3} + 3\mu$

Table 2: Values of the sum of the ℓ_1 and ℓ_2 norms corresponding to the decompositions listed above. Note that the optimal representation corresponds to the case $\boldsymbol{w}_1 = \boldsymbol{w}_3 = \boldsymbol{0}$, and $\boldsymbol{w}_2$ being a sparse vector.

Lastly, we can show that $h(\boldsymbol{x})$ is a norm. This will imply that $h(\boldsymbol{x})$ is convex, and hence the penalty we consider will be convex. This will then mean that the optimization we are interested in solving is a convex program.

Lemma 7 *The function*

$$h(\boldsymbol{x}) = \inf_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} (\|\boldsymbol{w}_G\|_2 + \mu\|\boldsymbol{w}_G\|_1)$$

is a norm

Proof It is trivial to show that $h(\boldsymbol{x}) \geq 0$ with equality *iff* $\boldsymbol{x} = 0$. We now show positive homogeneity. Suppose $\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})$ is an optimal representation (Definition 4) of $\boldsymbol{x}$, and let $\gamma \in \mathbb{R}\backslash\{\mathbf{0}\}$. Then, $\sum_{G\in\mathcal{G}} \boldsymbol{w}_G = \boldsymbol{x} \quad \Rightarrow \sum_{G\in\mathcal{G}} \gamma\boldsymbol{w}_G = \gamma\boldsymbol{x}$. This leads to the following set of inequalities:

$$h(\boldsymbol{x}) = \sum_{G\in\mathcal{G}} (\|\boldsymbol{w}_G\|_2 + \mu\|\boldsymbol{w}_G\|_1) = \frac{1}{|\gamma|}\sum_{G\in\mathcal{G}} (\|\gamma\boldsymbol{w}_G\|_2 + \mu\|\gamma\boldsymbol{w}_G\|_1) \geq \frac{1}{|\gamma|}h(\gamma\boldsymbol{x}) \tag{11}$$

Now, assuming $\{\boldsymbol{v}_G\} \in \mathcal{W}(\gamma\boldsymbol{x})$ is an optimal representation of $\gamma\boldsymbol{x}$, we have that $\sum_{G\in\mathcal{G}} \frac{\boldsymbol{v}_G}{\gamma} = \boldsymbol{x}$, and we get

$$h(\gamma\boldsymbol{x}) = \sum_{G\in\mathcal{G}} (\|\boldsymbol{v}_G\|_2 + \mu\|\boldsymbol{v}_G\|_1) = |\gamma|\sum_{G\in\mathcal{G}} \left(\left\|\frac{\boldsymbol{v}_G}{\gamma}\right\|_2 + \mu\left\|\frac{\boldsymbol{v}_G}{\gamma}\right\|_1\right) \geq |\gamma|h(\boldsymbol{x}) \tag{12}$$

Positive homogeneity follows from (11) and (12). The inequalities are a result of the possibility of the vectors not corresponding to the respective optimal representations.

For the triangle inequality, again let $\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x}), \{\boldsymbol{v}_G\} \in \mathcal{W}(\boldsymbol{y})$ correspond to the optimal representation for $\boldsymbol{x}, \boldsymbol{y}$ respectively. Then by definition,

$$\begin{aligned} h(\boldsymbol{x}+\boldsymbol{y}) &\leq \sum_{G\in\mathcal{G}} (\|\boldsymbol{w}_G + \boldsymbol{v}_G\|_2 + \mu\|\boldsymbol{w}_G + \boldsymbol{v}_G\|_1) \\ &\leq \sum_{G\in\mathcal{G}} (\|\boldsymbol{w}_G\|_2 + \|\boldsymbol{v}_G\|_2 + \mu\|\boldsymbol{w}_G\|_1 + \mu\|\boldsymbol{v}_G\|_1) \\ &= h(\boldsymbol{x}) + h(\boldsymbol{y}) \end{aligned}$$

The first and second inequalities follow by definition and the triangle inequality respectively. ■

3.2 Solving the SOGlasso Problem

We solve the Lagrangian version of the SOGlasso problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \left(\sum_{i=1}^{n} -\boldsymbol{y}_i \langle\boldsymbol{\phi}_i, \boldsymbol{x}\rangle\right) + \eta_1 h(\boldsymbol{x}) + \eta_2\|\boldsymbol{x}\|^2,$$

where $\eta_1 > 0$ controls the amount by which we regularize the coefficients to have a structured sparsity pattern, and $\eta_2 > 0$ prevents the coefficients from taking very large values. We use the "covariate duplication" method of (Jacob et al., 2009) to first reduce the problem to the non overlapping sparse group lasso in an expanded space. One can then use proximal methods to recover the coefficients.

Proximal point methods progress by taking a step in the direction of the negative gradient, and applying a shrinkage/proximal point mapping to the iterate. This mapping can be computed efficiently for the non overlapping sparse group lasso, as it is a special case of general hierarchical structured penalties (Jenatton et al., 2011). The proximal point mapping can be seen as the composition of the standard soft thresholding and the group soft thresholding operators:

$$\begin{aligned}\tilde{\boldsymbol{w}} &= \operatorname{sign}(\boldsymbol{w}_{\nabla})\,[|\boldsymbol{w}_{\nabla}| - \eta_1\mu]^+ \\ (\boldsymbol{w}_{t+1})_G &= \frac{(\tilde{\boldsymbol{w}})_G}{\|(\tilde{\boldsymbol{w}})_G\|}\,[\|(\tilde{\boldsymbol{w}}_G\| - \eta_1]^+ \quad \textbf{if } \|(\tilde{\boldsymbol{w}})_G\| \neq 0 \\ (\boldsymbol{w}_{t+1})_G &= 0 \quad \textbf{otherwise}\end{aligned}$$

where $\boldsymbol{w}_{\nabla}$ corresponds to the iterate after a gradient step and $[\cdot]^+ = \max(0, \cdot)$. Once the solution is obtained in the duplicated space, we then recombine the duplicates to obtain the solution in the original space. Finally, we perform a debiasing step to obtain the final solution.

4. Proof of Theorem 5, Theorem 6 and Extensions to Correlated Data

In this section, we compute the mean width of the constraint set $\mathcal{C}$ in (9), which will be used to prove Theorems 5 and 6. First we define the following function, analogous to the ℓ_0 pseudo-norm:

Definition 8 *Given a set of K groups $\mathcal{G}$, for any vector $\boldsymbol{x}$ and its optimal representation $\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})$, noting that $x = \sum_{G\in\mathcal{G}} \boldsymbol{w}_G$, define*

$$\|\boldsymbol{x}\|_{\mathcal{G},0} = \sum_{G\in\mathcal{G}} \mathbb{1}_{\{\|\boldsymbol{w}_G\|\neq 0\}}.$$

In the above definition, $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

Define the set

$$\mathcal{C}_{nc}(k,l) = \left\{\boldsymbol{x} : \boldsymbol{x} = \sum_{G\in\mathcal{G}} \boldsymbol{w}_G, \quad \|\boldsymbol{x}\|_{\mathcal{G},0} \leq k, \quad \sum_{G\in\mathcal{G}} \|\boldsymbol{w}_G\|_0 \leq kl \quad \forall G \in \mathcal{G}\right\}. \tag{13}$$

We see that $\mathcal{C}_{nc}(k,l)$ contains (k,l)-group sparse signals (Definition 3). From the above definitions and our problem setup, our aim is to ideally solve the following optimization problem

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{n} -\boldsymbol{y}_{ti}\,\langle \boldsymbol{\phi}_{ti}, \boldsymbol{x}_t\rangle \quad \textbf{s.t.} \quad \boldsymbol{x} \in \mathcal{C}_{nc}(k,l) \tag{14}$$

However, the set $\mathcal{C}_{nc}(k,l)$ is not convex, and hence solving (14) will be hard in general. We instead consider a convex relaxation of the above problem. The convex relaxation of the (overlapping) group ℓ_0 pseudo-norm is the (overlapping) group ℓ_1/ℓ_2 norm. This leads to the following result:

Lemma 9 *The SOGlasso penalty (10) admits a convex relaxation of $\mathcal{C}_{ideal}(k,\alpha)$. Specifically, we can consider the set*

$$\mathcal{C}(k,l) = \{\boldsymbol{x} : h(\boldsymbol{x}) \leq \sqrt{k}(1+\lambda_1)\|\boldsymbol{x}\|_2\}$$

as a tight convex relaxation containing the set $\mathcal{C}_{nc}(k,l)$.

Proof Consider a (k,l)-group sparse vector $\boldsymbol{x} \in \mathcal{C}_{nc}(k,l)$. For any such vector, there exist vectors $\{\boldsymbol{v}_G\} \in \mathcal{W}(\boldsymbol{x})$ such that the supports of $\boldsymbol{v}_G$ do not overlap. We then have the following set of inequalities

$$\begin{aligned}
h(\boldsymbol{x}) &= \inf_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} \|\boldsymbol{w}_G\|_2 + \frac{\lambda_1}{\sqrt{l}} \|\boldsymbol{w}_G\|_1 \\
&\overset{(i)}{\leq} \sum_{G \in \mathcal{G}} \|\boldsymbol{v}_G\|_2 + \frac{\lambda_1}{\sqrt{l}} \sum_{G \in \mathcal{G}} \|\boldsymbol{v}_G\|_1 \\
&\overset{(ii)}{\leq} \sum_{G \in \mathcal{G}} \|\boldsymbol{v}_G\|_2 + \frac{\lambda_1}{\sqrt{l}} \sqrt{l} \sum_{G \in \mathcal{G}} \|\boldsymbol{v}_G\|_2 \\
&= (1 + \lambda_1) \sum_{G \in \mathcal{G}} \|\boldsymbol{v}_G\|_2 \\
&\overset{(iii)}{\leq} \sqrt{k}\,(1 + \lambda_1) \left(\sum_{G \in \mathcal{G}} \|\boldsymbol{v}_G\|_2^2 \right)^{\frac{1}{2}} \\
&= \sqrt{k}\,(1 + \lambda_1)\, \|\boldsymbol{x}\|_2
\end{aligned}$$

where (i) follows from the definition of the function $h(\boldsymbol{x})$ in (8), and (ii) and (iii) follow from the fact that for any vector $\mathbf{v} \in \mathbb{R}^d$ we have $\|\mathbf{v}\|_1 \leq \sqrt{d}\,\|\mathbf{v}\|_2$.This, coupled with the fact that $h(\boldsymbol{x})$ is a norm (Lemma 7) ensures that the set $\mathcal{C}(k,l)$ is convex.

To show that the relaxation is tight, we will consider a (k,l) sparse vector $\boldsymbol{x}$ and show that the inequality in the definition of the set holds with equality. Specifically, let $\boldsymbol{x} \in \mathbb{R}^p$ with non overlapping groups, and let the first k $\boldsymbol{w}_G$s in it's representation be active. Moreover, suppose the first l entries in each of these $\boldsymbol{w}_G$s are non zero. Let the non zero entries all be equal to $\frac{1}{\sqrt{kl}}$. Then $\|\boldsymbol{x}\| = 1$, $\sum_G \|\boldsymbol{w}_G\|_2 = \sqrt{k}$ and $\sum_G \|\boldsymbol{w}_G\|_1 = \sqrt{kl}$. The result follows. ■

4.1 Mean Widths for the SOGlasso

We see that, the mean width of the constraint set plays a crucial role in determining the consistency of the solution of the optimization problem. We now aim to find the mean width of the constraint set in (9), and as a result of it, prove Theorems 5 and 6. Before we do so, we restate Lemma 3.2 in (Rao et al., 2012) for the sake of completeness:

Lemma 10 *Let $q_1, \ldots, q_K$ be K, χ-squared random variables with d-degrees of freedom. Then*

$$\mathbb{E}[\max_{1 \leq i \leq K} q_i] \leq (\sqrt{2\log(K)} + \sqrt{d})^2.$$

First, we prove Theorem 6.

Lemma 11 *Suppose that it is known that each coefficient is part of at most R groups, and suppose we let*

$$h(\boldsymbol{x}) = \inf_{\{\boldsymbol{w}_{\boldsymbol{x}}\} \in \mathcal{W}} \sum_{G in \mathcal{G}} \|\boldsymbol{w}_G\|_2 + \frac{\lambda_1}{\sqrt{l}} \|\boldsymbol{w}_G\|_1$$

Then the mean width of the set

$$\mathcal{C} = \{\boldsymbol{x} : h(\boldsymbol{x}) \leq \sqrt{k}\,(1+\lambda_1)\,, \quad \|\boldsymbol{x}\|_2 \leq 1\}$$

is bounded as

$$\omega(\mathcal{C})^2 \leq CR^2 k \left[\log\left(\frac{K}{k}\right) + l\log\left(\frac{L}{l}\right) + l + 2\right].$$

Proof The intuition behind this proof is as follows: We first consider a non convex set, which is the "ideal" set of (k,l) sparse vectors that we are interested in. We then show that $\mathcal{C}$ is contained in the scaled convex hull of the non convex set, and hence by the properties of the mean width, $\omega(\mathcal{C})$ can be bounded by a scaling of that of the non convex set [7].

To this end, let us consider the following non convex ideal set,

$$\mathcal{C}_{nc} = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq 1, \|\boldsymbol{x}\|_{\mathcal{G},0} \leq k, \sum_{G\in\mathcal{G}} \|\boldsymbol{w}_G\|_0 \leq kl\}. \tag{15}$$

Consider $\boldsymbol{x} \in \mathcal{C}$. We now define vectors $\boldsymbol{x}_i$ as follows: $\boldsymbol{x}_1 = \sum_{r=1}^{k} \boldsymbol{w}_1^r$, where the vectors $\boldsymbol{w}_1$ are the k vectors $\boldsymbol{w}_G$ with largest norm. Along these lines, we define $\boldsymbol{x}_i = \sum_{r=1}^{k} \boldsymbol{w}_i^r$.

For a fixed i, let $\boldsymbol{x}_{i1}$ be the vector containing the top kl entries of $\boldsymbol{x}_i$ by magnitude, and define a general $\boldsymbol{x}_{ij}$ in this manner as well.

Note that $\boldsymbol{x}_i = \sum_j \boldsymbol{x}_{ij}$ and $\boldsymbol{x} = \sum_i \boldsymbol{x}_i$. Also, note that $\frac{\boldsymbol{x}_{ij}}{\|\boldsymbol{x}_{ij}\|} \in \mathcal{C}_{nc}$ since it has at most k active groups, and at most kl non zero elements.

Finally, note the following: By construction, we have for a fixed i, and $j > 1$

$$\|\boldsymbol{x}_{ij}\|_2 \leq \frac{1}{\sqrt{kl}}\|\boldsymbol{x}_{ij-1}\|_1. \tag{16}$$

This follows since each element of $\boldsymbol{x}_{ij}$ is smaller than the average of the entries of the vector $\boldsymbol{x}_{ij-1}$.

Using the exact same argument, we also have for $i > 1$

$$\left(\sum_{r=1}^{k} \|\boldsymbol{w}_i^r\|_2^2\right)^{\frac{1}{2}} \leq \frac{1}{\sqrt{k}} \sum_{r=1}^{k} \|\boldsymbol{w}_{i-1}^r\|_2. \tag{17}$$

[7] Lemma 9 showed that the set $\mathcal{C}$ is a tight convex outer relaxation of the non convex set.

Now,

$$\begin{aligned}
\sum_{ij} \|\boldsymbol{x}_{ij}\| &= \|\boldsymbol{x}_{11}\|_2 + \sum_{i>1} \|\boldsymbol{x}_{i1}\|_2 + \sum_{i} \sum_{j>1} \|\boldsymbol{x}_{ij}\|_2 \\
&\leq \|\boldsymbol{x}_{11}\|_2 + \sum_{i>1} \|\boldsymbol{x}_{i1}\|_2 + \sum_{i} \sum_{j>1} \frac{1}{\sqrt{kl}} \|\boldsymbol{x}_{ij-1}\|_1 \quad \text{from (16)} \\
&\leq \|\boldsymbol{x}_{11}\|_2 + \sum_{i>1} \|\boldsymbol{x}_{i1}\|_2 + \sum_{i} \sum_{j} \frac{1}{\sqrt{kl}} \|\boldsymbol{x}_{ij}\|_1 \\
&\leq \|\boldsymbol{x}_{11}\|_2 + \sum_{i>1} \|\boldsymbol{x}_{i1}\|_2 + \frac{1}{\sqrt{kl}} \sum_{i} \|\boldsymbol{x}_{i}\|_1 \quad \text{since the indices for } j \text{ are disjoint} \\
&= \|\boldsymbol{x}_{11}\|_2 + \sum_{i>1} \|\boldsymbol{x}_{i1}\|_2 + \frac{1}{\sqrt{kl}} \sum_{i} \left\| \sum_{r=1}^{k} \boldsymbol{w}_i^r \right\|_1 \\
&\leq \|\boldsymbol{x}_{11}\|_2 + \sum_{i>1} \|\boldsymbol{x}_{i1}\|_2 + \frac{1}{\sqrt{kl}} \sum_{i} \sum_{r=1}^{k} \|\boldsymbol{w}_i^r\|_1 \quad \text{triangle inequality} \\
&= \|\boldsymbol{x}_{11}\|_2 + \sum_{i>1} \|\boldsymbol{x}_{i1}\|_2 + \frac{1}{\sqrt{kl}} \sum_{G} \|\boldsymbol{w}_G\|_1 \qquad (18)
\end{aligned}$$

For $i > 1$, we have the following bound:

$$\begin{aligned}
\|\boldsymbol{x}_{i1}\|^2 &\leq \left\| \sum_{r=1}^{k} \boldsymbol{w}_i^r \right\|^2 \\
&= \left[\sum_{m=1}^{p} \left(\sum_{r=1}^{k} (\boldsymbol{w}_i^r)_m \right)^2 \right] \\
&\leq \left[\sum_{m=1}^{p} R^2 \max_r (\boldsymbol{w}_i^r)_m^2 \right] \\
&\leq R^2 \sum_{m=1}^{p} \sum_{r=1}^{k} (\boldsymbol{w}_i^r)_m^2 \\
&= R^2 \sum_{r=1}^{k} \|\boldsymbol{w}_i^r\|_2^2
\end{aligned}$$

the above inequality and (17) combine to give

$$\|\boldsymbol{x}_{i1}\|_2 \leq \frac{R}{\sqrt{k}} \sum_{r=1}^{k} \|\boldsymbol{w}_{i-1}^r\|_2. \qquad (19)$$

Substituting this in (18) and noting that $\|\boldsymbol{x}_{11}\|_2 \leq \|\boldsymbol{x}\|_2 \leq 1$, we have

$$\sum_i \sum_j \|\boldsymbol{x}_{ij}\|_2 \leq 1 + \frac{1}{\sqrt{k}} \left(R \sum_G \|\boldsymbol{w}_G\|_2 + \frac{1}{\sqrt{l}} \sum_G \|\boldsymbol{w}_G\|_1 \right) \tag{20}$$

$$= 1 + \frac{R}{\sqrt{k}} \left(\sum_G \|\boldsymbol{w}_G\|_2 + \frac{1}{R\sqrt{l}} \sum_G \|\boldsymbol{w}_G\|_1 \right) \tag{21}$$

If $\lambda_1 \geq \frac{1}{R}$ then the term in the parentheses is bounded by $\sqrt{k}(1 + \frac{\lambda_1}{\sqrt{\alpha}})$, and if not, then it is bounded by $\sqrt{k}(1 + \frac{1}{R})$. This gives :

$$\sum_i \sum_j \|\boldsymbol{x}_{ij}\|_2 \leq 1 + R + \max(1, R\lambda_1).$$

The following argument finishes the proof. Setting $\eta = 1 + R + \max(1, R\lambda_1)$

1. By construction, we have $\boldsymbol{x} = \sum_i \sum_j \boldsymbol{x}_{ij}$.
2. Also by construction, $\frac{\boldsymbol{x}_{ij}}{\|\boldsymbol{x}_{ij}\|_2} \in \mathcal{C}_{nc}$.
3. Now, letting $\lambda_{ij} = \frac{\|\boldsymbol{x}_{ij}\|_2}{\eta}$, we showed

$$\frac{\boldsymbol{x}}{\eta} = \sum_i \sum_j \lambda_{ij} \frac{\boldsymbol{x}_{ij}}{\|\boldsymbol{x}_{ij}\|_2}$$

4. We showed that $\sum_i \sum_j \lambda_{ij} \leq 1$, so that $\boldsymbol{x}$ can be written as a convex combination of the $\frac{\boldsymbol{x}_{ij}}{\|\boldsymbol{x}_{ij}\|_2}$, which are elements in $\mathcal{C}_{nc}$. This means that $\frac{\boldsymbol{x}}{\eta} \in conv(\mathcal{C}_{nc})$.

We then have the following bound for the mean width of $\mathcal{C}$:

$$\omega(\mathcal{C})^2 \leq CR^2 \omega(\mathcal{C}_{nc})^2.$$

It now remains to compute $\omega(\mathcal{C}_{nc})$. Lemma 12 yields the desired result.

■

Lemma 12 *For*

$$\mathcal{C}_{nc} = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq 1, \|\boldsymbol{x}\|_{\mathcal{G},0} \leq k, \sum_{G \in \mathcal{G}} \|\boldsymbol{w}_G\|_0 \leq kl\},$$

we have

$$\omega(\mathcal{C}_{nc})^2 \leq Ck \left[\log\left(\frac{K}{k}\right) + l \log\left(\frac{L}{l}\right) + l + 2 \right].$$

We prove this in Appendix A. We make use of Lemma 10 to obtain the result.

We now proceed to prove Theorem 5. To do so, we adopt a different strategy than the one used to prove Theorem 6. Instead of considering the non convex ideal set, we directly consider the convex set $\mathcal{C}$ and show that it is a subset of appropriately scaled versions of the overlapping group lasso or the lasso balls. The result then follows.

Lemma 13 *Consider the same set as that considered in Lemma 11. The mean width of the set can also be shown to satisfy:*

$$\omega(\mathcal{C})^2 \leq Ck \min\{\log K + L, l \log(p)\}.$$

Proof Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and for a given $\boldsymbol{x}$, let $\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})$ be its optimal representation (Definition 4). Since $\boldsymbol{x} = \sum_{G \in \mathcal{G}} \boldsymbol{w}_G$, we have

$$\begin{aligned}
\max_{\boldsymbol{x} \in \mathcal{C}} \boldsymbol{g}^T \boldsymbol{x} &= \max_{\boldsymbol{x} \in \mathcal{C}} \boldsymbol{g}^T \sum_{G \in \mathcal{G}} \boldsymbol{w}_G \\
&= \max_{\boldsymbol{x} \in \mathcal{C}} \sum_{G \in \mathcal{G}} \boldsymbol{g}^T \boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \boldsymbol{x} = \sum_{G \in \mathcal{G}} \boldsymbol{w}_G \\
&= \max_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} \boldsymbol{g}^T \boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \sum_{G \in \mathcal{G}} \|\boldsymbol{w}_G\|_2 + \frac{\lambda_1}{\sqrt{l}} \|\boldsymbol{w}_G\|_1 \leq \sqrt{k}(1 + \lambda_1) &\quad (22) \\
&\overset{(i)}{\leq} \max_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} \boldsymbol{g}^T \boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \sum_{G \in \mathcal{G}} (1 + \frac{\lambda_1}{\sqrt{l}}) \|\boldsymbol{w}_G\|_2 \leq \sqrt{k}(1 + \lambda_1) \\
&= \max_{\{\boldsymbol{w}_G\} \in \mathcal{W}(\boldsymbol{x})} \sum_{G \in \mathcal{G}} \boldsymbol{g}^T \boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \sum_{G \in \mathcal{G}} \|\boldsymbol{w}_G\|_2 \leq \frac{\sqrt{kl}(1 + \lambda_1)}{(\sqrt{l} + \lambda_1)} &\quad (23) \\
&\overset{(ii)}{=} \frac{\sqrt{kl}(1 + \lambda_1)}{\sqrt{l} + \lambda_1} \quad \max_{G \in \mathcal{G}} \|\boldsymbol{g}_G\|_2 \\
&\leq \sqrt{k}\,(1 + \lambda_1) \quad \max_{G \in \mathcal{G}} \|\boldsymbol{g}_G\|_2 &\quad (24)
\end{aligned}$$

where we define $\boldsymbol{g}_G$ to be the sub vector of $\boldsymbol{g}$ indexed by group G. (i) follows since the constraint set is a superset of the constraint in the expression above it, from the fact that $\|\boldsymbol{a}\|_2 \leq \|\boldsymbol{a}\|_1 \ \ \forall \boldsymbol{a}$, and (ii) is a result of simple convex analysis.

The mean width is then bounded as

$$\omega(\mathcal{C}) \leq \sqrt{k}\,(1 + \lambda_1) \quad \mathbb{E}\left[\max_{G \in \mathcal{G}} \|\boldsymbol{g}_G\|_2\right]. \tag{25}$$

Squaring both sides of (25), we get

$$\begin{aligned}
\omega(\mathcal{C})^2 &\leq k\,(1 + \lambda_1)^2 \left[\mathbb{E}[\max_{G \in \mathcal{G}} \|\boldsymbol{g}_G\|_2]\right]^2 \\
&\overset{(iii)}{\leq} k\,(1 + \lambda_1)^2 \quad \mathbb{E}\left[\left(\max_{G \in \mathcal{G}} \|\boldsymbol{g}_G\|_2\right)^2\right] \\
&\overset{(iv)}{=} k\,(1 + \lambda_1)^2 \quad \mathbb{E}\left[\max_{G \in \mathcal{G}} \|\boldsymbol{g}_G\|_2^2\right]
\end{aligned}$$

where (*iii*) follows from Jensen's inequality and (iv) follows from the fact that the square of the maximum of non negative numbers is the same as the maximum of the squares. Now, note that since $\boldsymbol{g}$ is Gaussian, $\|\boldsymbol{g}_G\|^2$ is a χ^2 random variable with at most B degrees of freedom. From Lemma 10, we have

$$\omega(\mathcal{C})^2 \leq k\,(1+\lambda_1)^2\,(\sqrt{2\log(K)}+\sqrt{L})^2. \tag{26}$$

This gives us one of the two terms in the $\min\{\cdot,\cdot\}$ in the statement of the Lemma. Since α is bounded away from 0, we can treat the term in the parenthesis as a constant. For the second term, let us revisit (22), and obtain the following inequalities:

$$\begin{aligned}
\max_{\boldsymbol{x}\in\mathcal{C}} \boldsymbol{g}^T\boldsymbol{x} &= \max_{\{\boldsymbol{w}_G\}\in\mathcal{W}(\boldsymbol{x})} \sum_{G\in\mathcal{G}} \boldsymbol{g}^T\boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \sum_{G\in\mathcal{G}} \|\boldsymbol{w}_G\|_2 + \frac{\lambda_1}{\sqrt{l}}\|\boldsymbol{w}_G\|_1 \leq \sqrt{k}(1+\lambda_1) \\
&\overset{(v)}{\leq} \max_{\{\boldsymbol{w}_G\}\in\mathcal{W}(\boldsymbol{x})} \sum_{G\in\mathcal{G}} \boldsymbol{g}^T\boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \sum_{G\in\mathcal{G}} \frac{1}{\sqrt{L}}\|\boldsymbol{w}_G\|_1 + \frac{\lambda_1}{\sqrt{l}}\|\boldsymbol{w}_G\|_1 \leq \sqrt{k}(1+\lambda_1) \\
&= \max_{\{\boldsymbol{w}_G\}\in\mathcal{W}(\boldsymbol{x})} \sum_{G\in\mathcal{G}} \boldsymbol{g}^T\boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \left(\frac{\sqrt{\alpha}+\lambda_1}{\sqrt{l}}\right) \sum_{G\in\mathcal{G}} \|\boldsymbol{w}_G\|_1 \leq \sqrt{k}(1+\lambda_1) \\
&= \max_{\{\boldsymbol{w}_G\}\in\mathcal{W}(\boldsymbol{x})} \sum_{G\in\mathcal{G}} \boldsymbol{g}^T\boldsymbol{w}_G \ \ \textbf{s.t.} \ \ \sum_{G\in\mathcal{G}} \|\boldsymbol{w}_G\|_1 \leq \frac{\sqrt{kl}(1+\lambda_1)}{\sqrt{\alpha}+\lambda_1} \\
&= \frac{\sqrt{kl}(1+\lambda_1)}{\sqrt{\alpha}+\lambda_1} \max_{G\in\mathcal{G}} \max_{i\in G} |(\boldsymbol{g}_G)_i| \\
&\leq \frac{\sqrt{kl}(1+\lambda_1)}{\lambda_1} \max_i |\boldsymbol{g}_i|
\end{aligned}$$

Where the constraint set in (v) is a superset of that in the statement above it. Again, after squaring both sides, taking expectations and applying Jensen's inequality,

$$\omega(\mathcal{C})^2 \leq kl\left(\frac{1+\lambda_1}{\lambda_1}\right)^2 \mathbb{E}\left[\max_i \boldsymbol{g}_i^2\right]$$

The quantity inside the expectation is a χ^2 variable with one degree of freedom, and from Lemma 10, we obtain

$$\omega(\mathcal{C})^2 \leq Ckl\log{(p)}.$$

This gives the second term in the $\min\{\cdot,\cdot\}$, and finishes the proof

■

Lemma 13 and Theorem 2 lead directly to Theorem 5.

The results in the proof above shed some more light on our regularizer $h(\boldsymbol{x})$. If $\lambda_1 = 0$, then the problem reduces to that of classification using the overlapping group lasso penalty, and we obtain the corresponding sample complexity bound. For simple sparsity without any structure, we would want λ_1 to be large, in which case $\frac{1+\lambda_1}{\lambda_1} \to 1$, and $(1+\lambda_1) \to \infty$. This would then entail the bounds for the ℓ_1 regularized problem taking over, keeping all other parameters k, K, l, L fixed.

4.2 Extensions to Data with Correlated Entries

The results we proved above can be extended to data $\mathbf{\Phi}$ with correlated Gaussian entries as well (see (Raskutti et al., 2010) for results in linear regression settings). Indeed, in most practical applications we are interested in, the features are expected to contain correlations. For example, in the fMRI application that is one of the major motivating applications of our work, it is reasonable to assume that voxels in the brain will exhibit correlation amongst themselves at a given time instant. This entails scaling the number of measurements by the condition number of the covariance matrix $\mathbf{\Sigma}$, where we assume that each row if the measurement matrix $\mathbf{\Phi}$ is sampled from a Gaussian $(0, \mathbf{\Sigma})$ distribution. Specifically, we obtain a generalization of the result in (Plan and Vershynin, 2013) for the SOGlasso with a correlated Gaussian design.

We now consider the following constraint set:

$$\mathcal{C}_{corr} = \{\boldsymbol{x} : h(\boldsymbol{x}) \leq \frac{1}{\sigma_{min}(\mathbf{\Sigma}^{\frac{1}{2}})} \sqrt{k}(1 + \lambda_1), \|\mathbf{\Sigma}^{\frac{1}{2}} \boldsymbol{x}\| \leq 1\}. \tag{27}$$

We consider the set $\mathcal{C}_{corr}$ and not $\mathcal{C}$ in (9), since we require the constraint set to be a subset of the unit Euclidean ball. In the proof of Corollary 14 below, we will reduce the problem to an optimization over variables of the form $\boldsymbol{z} = \mathbf{\Sigma}^{\frac{1}{2}} \boldsymbol{x}$, and hence we require $\|\mathbf{\Sigma}^{\frac{1}{2}} \boldsymbol{x}\|_2 \leq 1$. Enforcing this constraint leads to the corresponding upper bound on $h(\boldsymbol{x})$.

We now obtain the following generalization of Theorem 5, for correlated data

Corollary 14 *Let the entries of the data matrix $\mathbf{\Phi}$ be sampled from a $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ distribution. Suppose the measurements follow the model in (1). Suppose we wish to recover a $(k, l)-$ group sparse vector from the set $\mathcal{C}_{corr}$ in (27). Suppose the true coefficient vector $\boldsymbol{x}^\star$ satisfies $\|\mathbf{\Sigma}^{\frac{1}{2}} \boldsymbol{x}^\star\| = 1$. Then, so long as the number of measurements n satisfies*

$$n \geq C \frac{\sigma^2}{\epsilon^2} \kappa(\mathbf{\Sigma}) k \min\{\log(K) + L, l \log(p)\},$$

the solution to (5) satisfies

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\star\|^2 \leq \frac{\epsilon}{\sigma_{min}(\mathbf{\Sigma})}.$$

where $\sigma_{min}(\cdot)$, $\sigma_{max}(\cdot)$ and $\kappa(\cdot)$ denote the minimum and maximum singular values and the condition number of the corresponding matrices respectively.

We prove this result in Appendix B. The proof is a straightforward modification of the proof of Theorem 5. A similar result along the lines of Theorem 6 can also be proved.

5. Applications and Experiments

In this section, we perform experiments on both real and toy data, and show that the function proposed in (8) indeed recovers the kind of sparsity patterns we are interested in in this paper. First, we experiment with some toy data to understand the properties of the function $h(\boldsymbol{x})$ and in turn, the solutions that are yielded from the optimization problem (5). Here, we take the opportunity to report results on linear regression problems as well. We then present results using two datasets from cognitive neuroscience and computational biology.

5.1 The SOGlasso for Multitask Learning

The SOG lasso is motivated in part by multitask learning applications. The group lasso is a commonly used tool in multitask learning, and it encourages the same set of features to be selected across all tasks. As mentioned before, we wish to focus on a less restrictive version of multitask learning, where the main idea is to encourage sparsity patterns that are similar, but not identical, across tasks. Such a restriction corresponds to a scenario where the different tasks are related to each other, in that they use similar features, but are not exactly identical. This is accomplished by defining subsets of similar features and searching for solutions that select only a few subsets (common across tasks) and a sparse number of features within each subset (possibly different across tasks). Figure 1 shows an example of the patterns that typically arise in sparse multitask learning applications, along with the one we are interested in. We see that the SOGlasso, with it's ability to select a few groups and only a few non zero coefficients within those groups lends itself well to the scenario we are interested in.

In the multitask learning setting, suppose the features are give by $\boldsymbol{\Phi}_t$, for tasks $t = \{1, 2, \ldots, \mathcal{T}\}$, and corresponding sparse vectors $\boldsymbol{x}_t^\star \in \mathbb{R}^p$. These vectors can be arranged as columns of a matrix $\boldsymbol{X}^\star$. Suppose we are now given M groups $\tilde{\mathcal{G}} = \{\tilde{G}_1, \tilde{G}_2, \ldots\}$ with maximum size $\tilde{B}$. Note that the groups will now correspond to sets of rows of $\boldsymbol{X}^\star$.

Let $\boldsymbol{x}^\star = [\boldsymbol{x}_1^{\star T} \quad \boldsymbol{x}_2^{\star T} \quad \ldots \boldsymbol{x}_{\mathcal{T}}^{\star T}]^T \in \mathbb{R}^{\mathcal{T}p}$, and $\boldsymbol{y} = [\boldsymbol{y}_1^T \quad \boldsymbol{y}_2^T \quad \ldots \boldsymbol{y}_{\mathcal{T}}^T]^T \in \mathbb{R}^{\mathcal{T}n}$. We also define $\mathcal{G} = \{G_1, G_2, \ldots, G_M\}$ to be the set of groups defined on $\mathbb{R}^{\mathcal{T}p}$ formed by aggregating the rows of $\boldsymbol{X}$ that were originally in $\tilde{\mathcal{G}}$, so that $\boldsymbol{x}$ is composed of groups $G \in \mathcal{G}$, and let the corresponding maximum group size be $B = \mathcal{T}\tilde{B}$. By organizing the coefficients in this fashion, we can reduce the multitask learning problem into the standard form as considered in (5). Hence, all the results we obtain in this paper can be extended to the multitask learning setting as well.

5.1.1 Results on fMRI dataset

In this experiment, we compared SOGlasso, lasso, standard multitask Glasso (with each feature grouped across tasks), the overlapping group lasso (Jacob et al., 2009) (with the same groups as in SOGlasso) and the Elastic Net (Zou and Hastie, 2005) in analysis of the star-plus dataset (Wang et al., 2003). 6 subjects made judgements that involved processing 40 sentences and 40 pictures while their brains were scanned in half second intervals using fMRI[8]. We retained the 16 time points following each stimulus, yielding 1280 measurements at each voxel. The task is to distinguish, at each point in time, which kind of stimulus a subject was processing. (Wang et al., 2003) showed that there exists cross-subject consistency in the cortical regions useful for prediction in this task. Specifically, experts partitioned each dataset into 24 non overlapping regions of interest (ROIs), then reduced the data by discarding all but 7 ROIs and, for each subject, averaging the BOLD response across voxels within each ROI. With the resulting data, the authors showed that a classifier trained on data from 5 participants generalized above chance when applied to data from a 6th–thus proving some degree of consistency across subjects in how the different kinds of information were encoded.

[8] Data and documentation available at http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/

We assessed whether SOGlasso could leverage this cross-individual consistency to aid in the discovery of predictive voxels without requiring expert pre-selection of ROIs, or data reduction, or any alignment of voxels beyond that existing in the raw data. Note that, unlike (Wang et al., 2003), we do not aim to learn a solution that generalizes to a withheld subject. Rather, we aim to discover a group sparsity pattern that suggests a similar set of voxels in all subjects, before optimizing a separate solution for each individual. If SOGlasso can exploit cross-individual anatomical similarity from this raw, coarsely-aligned data, it should show reduced cross-validation error relative to the lasso applied separately to each individual. If the solution is sparse within groups and highly variable across individuals, SOGlasso should show reduced cross-validation error relative to Glasso. Finally, if SOGlasso is finding useful cross-individual structure, the features it selects should align at least somewhat with the expert-identified ROIs shown by (Wang et al., 2003) to carry consistent information.

We trained the 5 classifiers using 4-fold cross validation to select the regularization parameters, considering all available voxels without preselection. We group regions of $5 \times 5 \times 1$ voxels and considered overlapping groups "shifted" by 2 voxels in the first 2 dimensions.[9]

Figure 4 shows the prediction error (misclassification rate) of each classifier for every individual subject. SOGlasso shows the smallest error. The substantial gains over lasso indicate that the algorithm is successfully leveraging cross-subject consistency in the location of the informative features, allowing the model to avoid over-fitting individual subject data. We also note that the SOGlasso classifier, despite being trained without any voxel pre-selection, averaging, or alginment, performed comparably to the best-performing classifier reported by Wang et al. (2003), which was trained on features average over 7 expert pre-selected ROIs

To assess how well the clusters selected by SOGlasso align with the anatomical regions thought a-priori to be involved in sentence and picture representation, we calculated the proportion of selected voxels falling within the 7 ROIs identified by (Wang et al., 2003) as relevant to the classification task (Table 3). For SOGlasso an average of 61.9% of identified voxels fell within these ROIs, significantly more than for lasso, group lasso (with or without overlap) and the elastic net. The overlapping group lasso, despite returning a very large number or predictors, hardly overlaps with the regions of interest to cognitive neuroscientists. The lasso and the elastic net make use of the fact that a separate classifier can be trained for each subject, but even in this case, the overlap with the regions of interest is low. The group lasso also fares badly in this regard, since the same voxels are forced to be selected across individuals, and this means that the regions of interest which will be misaligned across subjects will not in general be selected for each subject. All these drawbacks are circumvented by the SOGlasso. This shows that even without expert knowledge about the relevant regions of interest, our method partially succeeds in isolating the voxels that play a part in the classification task.

We make the following observations from Figure 4 and Figure 5

- The overlapping group lasso (Jacob et al., 2009) is ill suited for this problem. This is natural, since the premise is that the brains of different subjects can only be crudely aligned, and the overlapping group lasso will force the same voxel to be selected across

[9] The irregular group size compensates for voxels being larger and scanner coverage being smaller in the z-dimension (only 8 slices relative to 64 in the x- and y-dimensions).

Method	Avg. Overlap with ROI %
OGlasso	27.18
ENet	43.46
Lasso	41.51
Glasso	47.43
SOGlasso	61.90

Table 3: Mean Sparsity levels of the methods considered, and the average overlap with the precomputed ROIs in (Wang et al., 2003)

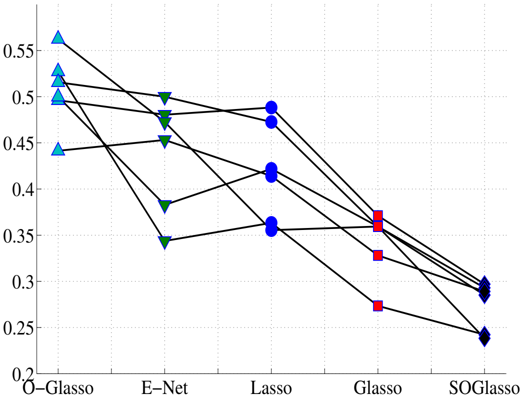


Figure 4: Misclassification error on a hold out set for different methods, on a per subject basis. Each solid line connects the different errors obtained for a particular subject in the dataset.

all individuals. It will also force all the voxels in a group to be selected, which is again undesirable from our perspective. This leads to a high number of voxels selected, and a high error.

- The elastic net (Zou and Hastie, 2005) treats each subject independently, and hence does not leverage the inter-subject similarity that we know exists across brains. The fact that all correlated voxels are also picked, coupled with a highly noisy signal means that a large number of voxels are selected, and this not only makes the result hard to interpret, but also leads to a large generalization error.

- The lasso (Tibshirani, 1996) is similar to the elastic net in that it does not leverage the inter subject similarities. At the same time, it enforces sparsity in the solutions, and hence a fewer number of voxels are selected across individuals. It allows any task correlated voxel to be selected, regardless of its spatial location, and that leads to a highly distributed sparsity pattern (Figure 5(a)). It leads to a higher cross-validation error, indicating that the ungrouped voxels are inferior predictors. Like the elastic net, this leads to a poor generalization error (Figure 4). The distributed sparsity pattern, low overlap with predetermined Regions of Interest, and the high error on the hold out set is what we believe makes the lasso a suboptimal procedure to use.

- The group lasso (Lounici et al., 2009) groups a single voxel across individuals. This allows for taking into account the similarities between subjects, but not the minor differences across subjects. Like the overlapping group lasso, if a voxel is selected for one person, the same voxel is forced to be selected for all people. This means, if a voxel encodes picture or sentence in a particular subject, then the same voxel is forced to be selected across subjects, and can arbitrarily encode picture or sentence. This gives rise to a purple haze in Figure 5(b), and makes the result hard to interpret. The purple haze manifests itself due to the large number of ambiguous voxels in Figure 5(d).

- Finally, the SOGlasso as we have argued helps in accounting for both the similarities and the differences across subjects. This leads to the learning of a code that is at the same time very sparse and hence interpretable, and leads to an error on the test set that is the best among the different methods considered. The SOGlasso (Figure 5(c)) overcomes the drawbacks of lasso and Glasso by allowing different voxels to be selected per group. This gives rise to a spatially clustered sparsity pattern, while at the same time selecting a negligible amount of voxels that encode both picture and sentences (Figure 5(d)). Also, the resulting sparsity pattern has a larger overlap with the ROI's than other methods considered.

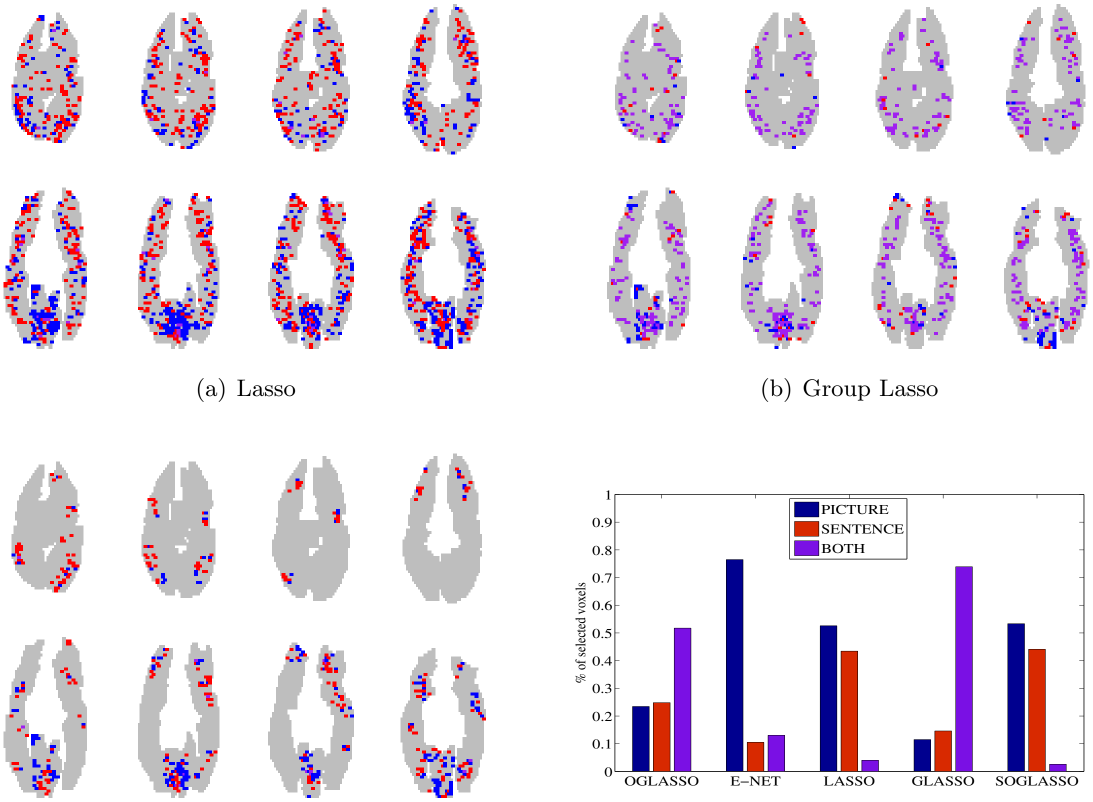
5.2 Toy Data, Linear Regression

Although not the primary focus of this paper, we show that the method we propose can also be applied to the linear regression setting. To this end, we consider simulated data and a multitask linear regression setting, and look to recover the coefficient matrix. We also use the simulated data to study the properties of the function we propose in (8).

The toy data is generated as follows: we consider $\mathcal{T} = 20$ tasks, and consider overlapping groups of size $B = 6$. The groups are defined so that neighboring groups overlap ($G_1 = \{1, 2, \ldots, 6\}$, $G_2 = \{5, 6, \ldots, 10\}$, $G_3 = \{9, 10, \ldots, 14\}$, $\ldots$). We consider a case with $M = 100$ groups, We set $k = 10$ groups to be active. We vary the sparsity level of the active groups α and obtain $m = 100$ Gaussian linear measurements corrupted with Additive White Gaussian Noise of standard deviation $\sigma = 0.1$. We repeat this procedure 100 times and average the results. To generate the coefficient matrices $\boldsymbol{X}^{\star}$, we select k groups at random, and within the active groups, only retain fraction α of the coefficients, again at random. The retained locations are then populated with uniform $[-1, 1]$ random variables.

The regularization parameters were clairvoyantly picked to minimize the Mean Squared Error (MSE) over a range of parameter values. The results of applying lasso, standard latent group lasso (Jacob et al., 2009), Group lasso where each group corresponds to a row of the sparse matrix, (Lounici et al., 2009) and our SOGlasso to these data are plotted in Figures 6(a), varying α.

Figure 6(a) shows that, as the sparsity within the active group reduces (i.e. the active groups become more dense), the overlapping group lasso performs progressively better. This is because the overlapping group lasso does not account for sparsity within groups,



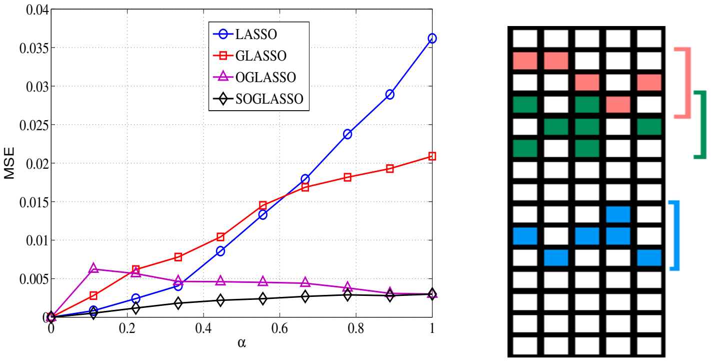
(a) Lasso

(b) Group Lasso

(c) SOG Lasso

(d) Voxel Encodings (%)

Figure 5: [Best seen in color]. Aggregated sparsity patterns across subjects per brain slice. All the voxels selected across subjects in each slice are colored in red, blue or purple. Red indicates voxels that exhibit a picture response in at least one subject and never exhibit a sentence response. Blue indicates the opposite. Purple indicates voxel that exhibited a a picture response in at least one subject and a sentence response in at least one more subject. (d) shows the percentage of selected voxels that encode picture, sentence or both.



(a) Varying α (b) Sample pattern

Figure 6: Figure (a) shows the result of varying α. The SOGlasso accounts for both inter and intra group sparsity, and hence performs the best. The Glasso achieves good performance only when the active groups are non sparse. Figure (b) shows a toy sparsity pattern, with different colors and brackets denoting different overlapping groups

and hence the resulting solutions are far from the true solutions for small values of α. The SOGlasso however does take this into account, and hence has a lower error when the active groups are sparse. Note that as $\alpha \to 1$, the SOGlasso approaches O-Glasso (Jacob et al., 2009). The Lasso (Tibshirani, 1996) does not account for group structure at all and performs poorly when α is large, whereas the Group lasso (Lounici et al., 2009) does not account for overlapping groups, and hence performs worse than O-Glasso and SOGlasso.

5.3 SOGlasso for Gene Selection

As explained in the introduction, another motivating application for the SOGlasso arises in computational biology, where one needs to predict whether a particular breast cancer tumor will lead to metastasis or not, from gene expression profiles. We used the breast cancer dataset compiled by (Van De Vijver et al., 2002) and grouped the genes into pathways as in (Subramanian et al., 2005). To make the dataset balanced, we perform a 3-way replication of one of the classes as in (Jacob et al., 2009), and also restrict our analysis to genes that are atleast in one pathway. Again as in (Jacob et al., 2009), we ensure that all the replicates are in the same fold for cross validation. We do not perform any preprocessing of the data, other than the replication to balance the dataset. We compared our method to the standard lasso, and the overlapping group lasso. The standard group lasso (Yuan and Lin, 2006) is ill-suited for this experiment, since the groups overlap and the sparsity pattern we expect is a union of groups, and it has been shown that the group lasso method will not recover the signal in such cases.

We trained a model using 4-fold cross validation on 80% of the data, and used the remaining 20% as a final test set. Table 4 shows the results obtained. We see that the SOGlasso penalty leads to lower classification errors as compared to the lasso or the latent group lasso. The errors reported are the ones obtained on the final (held out) test set. We refrain from performing simple ridge regression (Hoerl and Kennard, 1970) since the

Method	Misclassification Rate
lasso	0.42
OGlasso (Jacob et al., 2009)	0.39
SOGlasso	0.33

Table 4: Misclassification Rate on the test set for the different methods considered. The SOGlasso obtained better error rates as compared to the other methods.

data is high dimensional, and by not enforcing sparsity in the solution, the result will be un-interpretable.

6. Conclusions

In this paper, we introduced a function that can be used to constrain solutions of high dimensional feature selection problems so that they display both within and across group sparsity. We generalized the sparse group lasso to cases with arbitrary overlap, and proved consistency results in a classification setting. Our results unify the results between the lasso and the group lasso (with or without overlap), and reduce to those cases as special cases. We also outlined the use of the function in multitask fMRI and computational biology problems. Moreover, we make minimal assumptions on the model the generates data, and hence our results can be seen in a very general light.

From an algorithmic standpoint, when the groups overlap a lot, the replication procedure used in this paper might not be memory efficient. Future work involves designing algorithms that preclude replication, while at the same time allowing for the SOG- sparsity patterns to be generated.

From a cognitive neuroscience point of view, future work involves grouping the voxels in more intelligent ways. Our method to group spatially co-located voxels yields results that are significantly better than traditional lasso-based methods, but it remains to be seen whether there are better motivated ways to group them. For example, one might consider grouping voxels based on functional connectivities, or take into account the geodesic distance on the brain surface.

Appendix A.

We bound the mean width of the set in (15),
Proof [Proof of Lemma 12] Since $\|\boldsymbol{x}\|_2 \leq 1$,

$$\max \boldsymbol{g}^T \boldsymbol{x} = \max_{S \in \mathcal{S}} \|\boldsymbol{g}_S\|_2, \tag{28}$$

where $\mathcal{S}$ is the set of indices given by

$$\mathcal{S} = \{S_i \subset \{1, 2, \cdots, p\} : \quad j \in S_i \textbf{ if } \boldsymbol{x} \in \mathcal{C}_{nc}, \quad \boldsymbol{x}_j \neq 0\}.$$

The cardinality of $\mathcal{S}$ is bounded as

$$\begin{aligned}|\mathcal{S}| &= \binom{K}{k}\binom{kL}{kl} \\ &\leq \left(\frac{eK}{k}\right)^{k}\left(\frac{eL}{l}\right)^{kl} \\ \Rightarrow \log(|\mathcal{S}|) &\leq k\left[\log\left(\frac{K}{k}\right) + l\log\left(\frac{L}{l}\right) + 2\right]. \qquad (29)\end{aligned}$$

From (28) we have

$$\begin{aligned}\mathbb{E}[\max \boldsymbol{g}^T\boldsymbol{x}]^2 &= \mathbb{E}\max_{S\in\mathcal{S}} \|\boldsymbol{g}_S\|^2 \\ &\leq \left(\sqrt{\log|\mathcal{S}|} + \sqrt{|S|}\right)^2 \\ &\leq Ck\left[\log\left(\frac{K}{k}\right) + l\log\left(\frac{L}{l}\right) + l + 2\right], \qquad (30)\end{aligned}$$

where the first inequality follows from Lemma 3.2 in (Rao et al., 2012) and the last inequality follows from (29). ■

Appendix B.

We prove the result in Corollary 14. Before we do so, we state and prove a Lemma

Lemma 15 *Suppose $\boldsymbol{A} \in \mathbb{R}^{s\times t}$, and let $\boldsymbol{A}_G \in \mathbb{R}^{|G|\times t}$ be the sub matrix of $\boldsymbol{A}$ formed by retaining the rows indexed by group $G \in \mathcal{G}$. Suppose $\sigma_{max}(\boldsymbol{A})$ is the maximum singular value of $\boldsymbol{A}$, and similarly for $\boldsymbol{A}_G$. Then*

$$\sigma_{max}(\boldsymbol{A}) \geq \sigma_{max}(\boldsymbol{A}_G) \quad \forall G \in \mathcal{G}.$$

Proof Consider an arbitrary vector $\boldsymbol{x} \in \mathbb{R}^p$, and let $\bar{G}$ be the indices that are to indexed by G. We then have the following:

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{x}\|^2 &= \left\|\begin{bmatrix}\boldsymbol{A}_G\boldsymbol{x} \\ \boldsymbol{A}_{\bar{G}}\boldsymbol{x}\end{bmatrix}\right\|^2 \\ &= \|\boldsymbol{A}_G\boldsymbol{x}\|^2 + \|\boldsymbol{A}_{\bar{G}}\boldsymbol{x}\|^2 \\ \Rightarrow \|\boldsymbol{A}\boldsymbol{x}\|^2 &\geq \|\boldsymbol{A}_G\boldsymbol{x}\|^2 \qquad (31)\end{aligned}$$

We therefore have

$$\begin{aligned}\sigma_{max}(\boldsymbol{A}) &= \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\| \\ &\geq \sup_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}_G\boldsymbol{x}\| \\ &= \sigma_{max}(A_G)\end{aligned}$$

where the inequality follows from (31). ■

We now proceed to prove Corollary 14.

Proof Since the entries of the data matrices are correlated Gaussians, the inner products in the objective function of the optimization problem (5) can be written as

$$\langle \boldsymbol{\Phi}_i, \boldsymbol{x} \rangle = \langle \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\Phi}_i', \boldsymbol{x} \rangle = \langle \boldsymbol{\Phi}_i', \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{x} \rangle,$$

where $\boldsymbol{\Phi}_i' \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Hence, we can replace $\boldsymbol{x}$ in our result in Theorem 5 by $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{x}$, and make appropriate changes to the constraint set.

We then see that the optimization problem we need to solve is

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} - \sum_{i=1}^{n} \boldsymbol{y}_i \langle \boldsymbol{\Phi}_i', \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{x} \rangle \quad \textbf{s.t.} \quad \boldsymbol{x} \in \mathcal{C}_{corr}.$$

Defining $\boldsymbol{z} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{x}$, we can equivalently write the above optimization as

$$\hat{\boldsymbol{z}} = \arg\min - \sum_{i=1}^{n} \boldsymbol{y}_i \langle \boldsymbol{\Phi}_i', \boldsymbol{z} \rangle \quad \textbf{s.t.} \quad \boldsymbol{z} \in \boldsymbol{\Sigma}^{\frac{1}{2}} \mathcal{C}_{corr}, \tag{32}$$

where we define $\Sigma^{\frac{1}{2}}\mathcal{C}$ to be the set $\mathcal{C}v$, with each element multiplied by $\boldsymbol{\Sigma}^{\frac{1}{2}}$. We see that (32) is of the same form as (5), with the constraint set "scaled" by the matrix $\boldsymbol{\Sigma}^{\frac{1}{2}}$. We now need to bound the mean width of the set $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathcal{C}_{corr}$.

We then have

$$\begin{aligned} \max_{\boldsymbol{z} \in \boldsymbol{\Sigma}^{\frac{1}{2}} \mathcal{C}_{corr}} \boldsymbol{g}^T \boldsymbol{z} &= \max_{\boldsymbol{x} \in \mathcal{C}} \boldsymbol{g}^T \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{x} \\ &= \max_{\boldsymbol{x} \in \mathcal{C}_{corr}} (\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{g})^T \boldsymbol{x} \\ &\leq \frac{\sqrt{k}(1+\lambda_1)}{\sigma_{min}(\boldsymbol{\Sigma}^{\frac{1}{2}})} \max_{G \in \mathcal{G}} \|\boldsymbol{\Sigma}_G^{\frac{1}{2}} \boldsymbol{g}\| \end{aligned}$$

where the final inequality follows from the exact same arguments used to obtain (24). By $\boldsymbol{\Sigma}_G^{\frac{1}{2}}$, we mean the $|G| \times p$ sub matrix of $\boldsymbol{\Sigma}^{\frac{1}{2}}$ obtained by retaining rows indexed by group G.

To compute the mean width, we need to find $\mathbb{E}[\max_{G \in G} \|\boldsymbol{\Sigma}_G^{\frac{1}{2}} \boldsymbol{g}\|^2]$. Now, since $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I})$, $\boldsymbol{\Sigma}_G^{\frac{1}{2}} \boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_G^{\frac{1}{2}} (\boldsymbol{\Sigma}_G^{\frac{1}{2}})^T)$. Hence, $\|\boldsymbol{\Sigma}_G^{\frac{1}{2}} \boldsymbol{g}\|^2 \leq \sigma_{max}(\boldsymbol{\Sigma}_G^{\frac{1}{2}}) \|\boldsymbol{c}\|^2$ where $\boldsymbol{c} \sim \mathcal{N}(0, \boldsymbol{I}_{|G|})$. $\|\boldsymbol{c}\|^2 \sim \chi^2_{|G|}$, and we can again use Lemma 10 to obtain the following bound for the mean width:

$$\begin{aligned} \omega(\boldsymbol{\Sigma}^{\frac{1}{2}} \mathcal{C})^2 &\leq \frac{k(1+\lambda_1)^2}{\sigma_{min}(\boldsymbol{\Sigma}^{\frac{1}{2}})^2} (\sqrt{2\log(K)} + \sqrt{L})^2 \left[\max_{G \in \mathcal{G}} \sigma_{\max}(\boldsymbol{\Sigma}_G) \right] \\ &\leq \sigma_{max}(\boldsymbol{\Sigma}) \frac{k(1+\lambda_1)^2}{\sigma_{min}(\boldsymbol{\Sigma})} (\sqrt{2\log(K)} + \sqrt{L})^2 \\ &\leq C\kappa(\boldsymbol{\Sigma}) k (\log(K) + L) \end{aligned} \tag{33}$$

Similarly, following a procedure similar to that used to prove Theorem 5, we obtain

$$\omega(\boldsymbol{\Sigma}^{\frac{1}{2}}\mathcal{C})^2 \leq C\kappa(\boldsymbol{\Sigma})k\min\{\log(K) + L, l\log(p)\}, \tag{34}$$

where the last inequality follows from Lemma 15.

We then have that so long as the number of measurements n is larger than $C\frac{\sigma^2}{\epsilon^2}$ times the quantity in (34),

$$\|\hat{\boldsymbol{z}} - \boldsymbol{z}^{\star}\|^2 = \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\hat{\boldsymbol{x}} - \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{x}^{\star}\right\|^2 \leq \epsilon.$$

However, note that

$$\sigma_{min}(\boldsymbol{\Sigma})\|\hat{\boldsymbol{x}} - \boldsymbol{x}^{\star}\|^2 \leq \left\|\boldsymbol{\Sigma}^{\frac{1}{2}}\hat{\boldsymbol{x}} - \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{x}^{\star}\right\|^2. \tag{35}$$

(34) and (35) combine to give the final result. Note that for the sake of keeping the exposition simple, we have used Lemma 15 and bounded the number of sufffieicnt measurements as a function of the maximum singular value of $\boldsymbol{\Sigma}$. However, the number of measurements only depends on $\max_{G\in\mathcal{G}}\sigma_{max}(\boldsymbol{\Sigma}_G)$, which is typically much lesser.

■

References

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

Soumyadeep Chatterjee, Arindam Banerjee, Snigdhansu Chatterjee, and Auroop R Ganguly. Sparse group lasso for regression on land climate variables. In *ICDM Workshops*, pages 1–8, 2011.

Eva Feredoes, Giulio Tononi, and Bradley R Postle. The neural bases of the short-term storage of verbal information are anatomically variable across individuals. *The Journal of Neuroscience*, 27(41):11003–11008, 2007.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM, 2009.

Ali Jalali, Pradeep D Ravikumar, Sujay Sanghavi, and Chao Ruan. A dirty model for multi-task learning. 3:7, 2010.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*, 12:2297–2334, 2011.

Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

Andreas Maurer and Massimiliano Pontil. Structured sparsity and generalization. *The Journal of Machine Learning Research*, 13:671–690, 2012.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

Sofia Mosci, Silvia Villa, Alessandro Verri, and Lorenzo Rosasco. A primal-dual algorithm for group sparse regularization with overlapping groups. In *Advances in Neural Information Processing Systems*, pages 2604–2612, 2010.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on*, 59(1):482–494, 2013.

Nikhil Rao, Christopher Cox, Rob Nowak, and Timothy T Rogers. Sparse overlapping sets lasso for multitask learning and its application to fmri analysis. pages 2202–2210, 2013.

Nikhil S Rao, Robert D Nowak, Stephen J Wright, and Nick G Kingsbury. Convex approaches to model wavelet sparsity patterns. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 1917–1920. IEEE, 2011.

Nikhil S Rao, Ben Recht, and Robert D Nowak. Universal measurement bounds for structured sparse signal recovery. In *International Conference on Artificial Intelligence and Statistics*, pages 942–950, 2012.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Irina Rish, Guillermo A Cecchia, Kyle Heutonb, Marwan N Balikic, and A Vania Apkarianc. Sparse regression analysis of task-relevant information distribution in the brain. In *Proceedings of SPIE*, volume 8314, page 831412, 2012.

Srikanth Ryali, Kaustubh Supekar, Daniel A Abrams, and Vinod Menon. Sparse logistic regression for whole brain classification of fmri data. *NeuroImage*, 51(2):752, 2010.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina Eldar. Collaborative hierarchical sparse modeling. In *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, pages 1–6. IEEE, 2010.

Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina C Eldar. C-hilasso: A collaborative hierarchical sparse modeling framework. *Signal Processing, IEEE Transactions on*, 59(9):4183–4198, 2011.

Mihailo Stojnic, Farzad Parvaresh, and Babak Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *Signal Processing, IEEE Transactions on*, 57(8):3075–3085, 2009.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Marc J Van De Vijver, Yudong D He, Laura J van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

Marcel van Gerven, Christian Hesse, Ole Jensen, and Tom Heskes. Interpreting single trial data using groupwise regularisation. *NeuroImage*, 46(3):665–676, 2009.

Xuerui Wang, Tom M Mitchell, and Rebecca Hutchinson. Using machine learning to detect cognitive states across multiple subjects. *CALD KDD project paper*, 2003.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Yang Zhou, Rong Jin, and Steven Hoi. Exclusive lasso for multi-task feature selection. In *International Conference on Artificial Intelligence and Statistics*, pages 988–995, 2010.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.