The Performance of Group Lasso for Linear Regression of Grouped Variables

Marco F. Duarte,† Waheed U. Bajwa,† and Robert Calderbank*

Technical Report TR-2010-10
Department of Computer Science
Duke University

February 15, 2011

1 Introduction

The lasso [19] and group lasso [23] are popular algorithms in the signal processing and statistics communities. In signal processing, these algorithms allow for efficient sparse approximations of arbitrary signals in overcomplete dictionaries. In statistics, they facilitate efficient variable selection and reliable regression under the linear model assumption. In both cases, there is now ample empirical evidence to suggest that an appropriately regularized group lasso can outperform the lasso whenever there is a natural grouping of the dictionary elements/regression variables in terms of their contributions to the observations [1, 23].

Our goal in this technical report is to analytically characterize the regression performance of the group lasso algorithm using ℓ_1/ℓ_2 regularization for the case in which one can have far more regression variables than observations. Analytical characterization of group lasso in this "underdetermined" setting has received some attention lately in the statistics literature [1, 14–17]. However, prior analytical work on the performance of group lasso either studies an asymptotic regime [1, 15–17], focuses on random design matrices [1, 16], and/or relies on metrics that are computationally expensive to evaluate [14, 15, 17]. Recently, Candés and Plan [4] successfully circumvented somewhat similar shortcomings

*MFD is with the Department of Computer Science, WUB is with the Department of Electrical and Computer Engineering, and RC is with the Departments of Computer Science, Electrical and Computer Engineering, and Mathematics at Duke University, Durham, NC 27708. This work was supported by grants ONR N00014-08-1-1110 and AFOSR FA9550-09-1-0422, FA9550-09-1-0643, and FA9550-05-0443. MFD was also supported by NSF Supplemental Funding DMS-0439872 to UCLA-IPAM, P.I. R. Caflisch.

† Equal contribution to the research; author ordering was determined by a fair coin toss.

of the performance analysis for the lasso by imposing a probabilistic model on the vector of regression coefficients. Specifically, [4] showed that under mild, computable conditions on arbitrary (random or deterministic) design matrices, the lasso can perform near-optimally in terms of the regression error with very high probability for the following model: (i) locations of the nonzero regression coefficients are chosen uniformly at random; (ii) "signs" of nonzero regression coefficients are statistically independent; and (iii) nonzero regression coefficients have zero median.

In this technical report, we study the regression performance of the group lasso algorithm using ℓ_1/ℓ_2 regularization in the underdetermined case under a generalization of the probabilistic framework of [4] to the group case. Specifically, our framework assumes that: (i) locations of the groups of nonzero regression coefficients are chosen uniformly at random; (ii) "directions" of the groups of nonzero regression coefficients are statistically independent; and (iii) nonzero regression coefficients have zero median. Our main contribution here is proving under this model that the group lasso[1] can also perform near-optimally in terms of the regression error with very high probability under mild, computable conditions on arbitrary design matrices. To the best of our knowledge, these are the first results for group lasso that are non-asymptotic in nature, applicable to arbitrary design matrices through easily computable metrics, and still allow for near-optimal scaling of the number of observations with the number of groups of nonzero regression coefficients. Our proof techniques are natural extensions of the ones used in [4] for the lasso and rely on our recent result concerning the conditioning of random block-subdictionaries of matrices [2], an extension of a result by Tropp [21] that facilitated the analysis in [4].

This technical report is organized as follows. Section 2 provides background and notation. Section 3 provides our result, with the proof provided in the appendix, and Section 4 contrasts our result with related prior work.

2 Background and Notation

We consider a vector of observations $y \in \mathbb{R}^n$ corresponding to the classical linear model $y = X\beta + z$, where X denotes the design matrix containing one regression variable per column, β denotes the vector of regression coefficients for these variables, and z denotes the modeling error. Here, we assume (without loss of generality) that X has unit-norm columns and we treat z as an independent and identically distributed (i.i.d.) Gaussian vector with variance σ^2.

The key distinguishing feature of our model is that we assume there is a natural grouping of the regression variables. For the sake of this exposition, we consider p equal-sized groups of the regressors, which leads to the following block representation of β:

$$\beta = [\beta_1^T \ \beta_2^T \ \ldots \ \beta_p^T]^T,$$

[1] We refer to the group lasso algorithm using ℓ_1/ℓ_2 regularization as "group lasso" throughout the rest of the technical report for brevity.

where $\beta_i \in \mathbb{R}^m,\ 1 \leq i \leq p$, denote different groups of size m in β. We can now define the $\ell_{q,r}$ norm of a vector $\beta \in \mathbb{R}^{pm}$ containing p blocks of size m entries each as

$$\|\beta\|_{q,r} = \left(\sum_{i=1}^{p} \|\beta_i\|_q^r \right)^{1/r},$$

with the standard modification for $q, r = \infty$. The group lasso solution for estimating β from y under this setup can then be written as [23]

$$\widehat{\beta} = \arg \min_{\beta \in \mathbb{R}^{pm}} \frac{1}{2} \|y - X\beta\|_2^2 + 2\lambda\sigma\sqrt{m}\|\beta\|_{2,1}. \tag{1}$$

3 Regression Performance of Group Lasso

In this section, we provide performance guarantees for group lasso for the underdetermined case, $n < pm$, using the metric of regression error: $\|X\beta - X\widehat{\beta}\|_2$. In order to make this problem well-posed and tractable, we assume that the vector of regression coefficients $\beta \in \mathbb{R}^{pm}$ is k-block sparse with $\#\{i : \beta_i \neq \mathbf{0}\} = k \ll p$ and we impose a statistical prior on β. Specifically, we assume that: (i) block support of β, $I = \{i : \beta_i \neq \mathbf{0}\}$, has a uniform distribution over all k-subsets of $\{1, \ldots, p\}$; (ii) "directions" of the nonzero blocks of β are statistically independent: $\mathbb{P}\left(\bigcap_{i \in I} \overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right) = \prod_{i \in I} \mathbb{P}\left(\overline{\text{sign}}(\beta_i) \in \mathcal{A}_i\right)$, where $\overline{\text{sign}}(\beta_i) = \beta_i / \|\beta_i\|_2$ denotes the unit-norm vector pointing in the direction of β_i in $\mathbb{R}^m$; and (iii) nonzero regression coefficients have zero median: $\mathbb{E}(\text{sign}(\beta)) = \mathbf{0}$, where $\text{sign}(\cdot)$ denotes the entry-wise sign operator.

The main result of this technical report relies on three easily computable metrics of the design matrix, namely, coherence, block coherence, and spectral norm of X. The coherence of a matrix $X \in \mathbb{R}^{n \times pm}$ with unit norm columns is defined as

$$\mu = \max_{1 \leq i,i' \leq p, 1 \leq j,j' \leq m, (i,j) \neq (i',j')} |\langle X_{ij}, X_{i'j'} \rangle|,$$

where X_{ij} denotes the j^{th} column of the i^{th} block of $X = [X_1\ X_2\ \ldots\ X_p]$. Similarly, the block coherence of X is defined as

$$\mu_B = \max \left\{ \max_{1 \leq i,i' \leq p, i \neq i'} \|X_i^* X_{i'}\|_2, \max_{1 \leq i \leq p} \|X_i^* X_i - I\|_2 \right\},$$

where X_i denotes the i^{th} block of X. Note here that X_i^* denotes the adjoint of X_i rather than a submatrix of X^*. We now state our main theorem, which is motivated by the analysis in [4]; its proof is given in the appendix.

Theorem 1. *Suppose that the vector of regression coefficients β is drawn according to the statistical model described earlier. If $k \leq C_0 p / \|X\|_2^2 \log(pm)$, and the matrix X satisfies $\mu \leq 1/m$ and $\mu_B \leq C_1 / \log(pm)$ for some positive numerical constants C_0 and C_1, then the group lasso estimate $\widehat{\beta}$ computed with $\lambda = \sqrt{2 \log(pm)}$ obeys*

$$\|X\beta - X\widehat{\beta}\|_2^2 \leq Cmk\sigma^2 \log(pm)$$

with probability at least $1-(pm)^{-1}(2\pi\log(pm))^{-1/2}-8(pm)^{-2\log 2}$. *Here,* $C>0$ *is a constant independent of the problem parameters.*

4 Discussion and Related Work

Note that since β has mk nonzero regression coefficients, Theorem 1 states that the group lasso results in near-optimal regression error (modulo the logarithmic factor) of $O(mk\sigma^2\log(pm))$ provided the coherence and block coherence of the design matrix are not too high. Equally importantly, the theorem states that if the design matrix is an approximately tight frame, $\|X\|_2^2 \approx \frac{pm}{n}$, then this regression error can be achieved as long as the number of nonzero regression coefficients satisfies $mk = O(n/\log(pm))$. Summarizing, our result establishes that the group lasso performs near-optimal regression even when the number of nonzero regression coefficients scales almost linearly with the number of observations, provided X is an approximately tight frame and its coherence and block coherence are not too high. Some examples of design matrices satisfying these requirements include random Gaussian matrices and deterministic matrices designed from Grassmanian packings [3].

In terms of relation with previous work, there have been other efforts in the recent past to establish near-optimal performance of the group lasso in the underdetermined setting [1, 14–17]. However, there are three key aspects of our work that set it apart from these and similarly related works. First, our results are completely non-asymptotic in nature. Second, our results are applicable to arbitrary design matrices through the metrics of coherence, block coherence, and spectral norm, all of which are easily computable in polynomial time. Last, our results allow for near-optimal scaling of the number of observations with the number of groups of nonzero regression coefficients for matrices that are approximately tight frames. It is also worth noting here that the key enabling factor that makes our results possible is a weak statistical prior on the vector of regression coefficients β, which is in contrast with prior work on the group lasso where the focus tends to be on deterministic β.

In addition to the literature on group linear regression, there is also a line of work in compressive sensing and sparse approximation literature that can be thought of as a special case of the problem studied here. In that work, termed the multiple measurement vector (MMV) [6] or multivariate linear regression [18] problem, it is assumed that a total of m correlated, sparse vectors $B = [\beta_1\ \beta_2\ \ldots\ \beta_m]$ are observed using a single design matrix $X \in \mathbb{R}^{d\times p}$ to obtain a set of observation vectors $Y = [y_1\ y_2\ \ldots\ y_m]$: $Y = XB + Z$, where $Z \in \mathbb{R}^{d\times m}$ denotes the observation noise. The key distinguishing feature of the MMV setup is the assumption that correlation across the vectors $\{\beta_i\}$ imposes the constraint that they share approximately the same support, so that B has only a small number, k, of nonzero rows.

Interestingly, it is possible to express the MMV problem in terms of a group linear regression problem studied in this technical report. In order to do so, we use vect(A) to denote a column vector obtained by stacking the columns

of a matrix A. Next, we define $y' = \text{vect}(Y^T)$, $\beta' = \text{vect}(B^T)$, and $z' = \text{vect}(Z^T)$ as the vectorized versions of the observations, the sparse vectors, and the noise, respectively. Additionally, we define the expanded design matrix $X' = X^T \otimes I_{m \times m}$, where $I_{m \times m}$ denotes the $m \times m$ identity matrix and $\otimes$ denotes the standard Kronecker product. It then follows from a simple linear algebra identity involving the Kronecker product that the MMV problem can be equivalently expressed as $y' = X'\beta' + z'$, where $\beta' \in \mathbb{R}^{pm}$ has a total of k nonzero blocks.

The preceding analysis shows that the MMV problem can be viewed as a special case of the group linear regression problem where the design matrix X' has a particular Kronecker structure. It is a simple exercise then to tweak the proof of Theorem 1 in order to account for the special structure of the design matrix and obtain even stronger results, as desired in the MMV literature. More specifically, note that the proof of Theorem 1 relies on our recent result concerning the conditioning of random block-subdictionaries of matrices [2], but conditioning of random block-subdictionaries of $X' = X^T \otimes I_{m \times m}$ is trivially guaranteed by the random subdictionaries result of Tropp [21] because of the special structure of X', which leads to a stronger variant of Theorem 1.

The analysis in this technical report, therefore, can also be thought of as specifying the performance of the group lasso for the MMV problem using the metric of regression error. On the other hand, while there exists a significant body of literature on the MMV problem [7–12, 18, 20, 22], these works differ from the MMV interpretation of our analysis in two key aspects. First, the focus in most of these works is either on model selection (support detection) or on estimation (reconstruction) error. Second, and most importantly, similar to the case of previous work on group linear regression, most of these works also either study an asymptotic regime, focus on random design matrices, or rely on metrics that are either computationally expensive to evaluate or which do not allow for near-optimal scaling of the number of observations with the number of nonzero rows of B. The notable exception to this is a recent work by Eldar and Rauhut [10], which studies the MMV problem in a noiseless setting (Z being an all-zeros matrix) and provides guarantees for exact recovery of B from Y. However, note that analyzing the regression error, which is the focus of this technical report, is a vacuous problem in a noiseless setting.

We conclude by noting that for $m = 1$ in the group linear regression problem, in which case the group lasso reduces to the standard lasso, Theorem 1 reduces to that obtained in [4] for the lasso. Further, while it is possible to make use of [2] together with the analysis in [4] to analyze the performance of the standard lasso for group regression error, this would require imposing statistical independence on the signs of nonzero regression coefficients even within the groups in β. We plan to highlight these and other subtle, but important, differences between the lasso and the group lasso in the future.

A Proof of Theorem 1

We mirror the procedure of the proof of Theorem 1.2 in [4]. The proof uses the following two lemmas.

Lemma 1. *The group lasso estimate obeys*

$$\|X^*(y - X\widehat{\beta})\|_{2,\infty} \leq 2\lambda\sigma\sqrt{m}.$$

Proof. Since $\widehat{\beta}$ minimizes the objective function over β, then 0 must be a subgradient of the objective function at $\widehat{\beta}$. The subgradients of the group lasso objective function are of the form [23]

$$X_i^*(X\beta - y) + 2\lambda\sigma\sqrt{m}\epsilon_i = 0,\ i = 1, \ldots, p,$$

where $\epsilon_i \in \mathbb{R}^m$ is of the form $\epsilon_i = \overline{\text{sign}}(\beta_i)$ if $\beta_i = 0$ and $\|\epsilon_i\|_2 \leq 1$ otherwise. Here we denote $\overline{\text{sign}}(\beta_i) = \beta_i/\|\beta_i\|_2$, i.e., the unit-norm vector pointing in the direction of β_i in $\mathbb{R}^m$. We also extend this notation to higher-dimensional vectors in a block-wise fashion. Hence, since 0 is a subgradient at $\widehat{\beta}$, there exists $\epsilon = [\epsilon_1^T\ \ldots\ \epsilon_p^T]^T$ such that

$$X^*(X\widehat{\beta} - y) = -2\lambda\sigma\sqrt{m}\epsilon.$$

The conclusion follows from $\|\epsilon\|_{2,\infty} \leq 1$. □

We also borrow the following theorem from [2].

Theorem 2. *Define random variables $\delta_1, \ldots, \delta_p$ that are independent and identically distributed (i.i.d.) Bernoulli with parameter $\delta := k/p$, and form a block subdictionary $X_{I'} = [X_i : \delta_i = 1]$. Then, for $q = 2\log(pm)$, we have the bound*

$$[\mathbb{E}\|X_{I'}^* X_{I'} - \text{Id}\|_2^q]^{1/q} \leq 20\mu_B \log(pm) + 9\sqrt{\delta \log(pm)(1 + (m-1)\mu)}\|X\|_2 + \delta\|X\|_2^2,$$

where Id *denotes the identity matrix of appropriate size.*

We assume that $\sigma = 1$ without loss of generality and establish three conditions that together imply the theorem:

- *Invertibility condition.* The submatrix $X_I^* X_I$ is invertible and obeys

$$\|(X_I^* X_I)^{-1}\|_2 \leq 2.$$

- *Orthogonality condition.* The vector z obeys $\|X^* z\|_{2,\infty} \leq \sqrt{2m} \cdot \lambda$.

- *Complementary size condition.* The following inequality holds:

$$\|X_{I^C}^* X_I (X_I^* X_I)^{-1} X_I^* z\|_{2,\infty} + 2\lambda\sqrt{m}\|X_{I^C}^* X_I (X_I^* X_I)^{-1} \overline{\text{sign}}(\beta_I)\|_{2,\infty} \leq (2 - \sqrt{2})\lambda\sqrt{m}.$$

We assume that the three conditions hold. Since $\widehat{\beta}$ minimizes the group lasso objective function, we must have

$$\frac{1}{2}\|y - X\widehat{\beta}\|_2^2 + 2\lambda\sqrt{m}\|\widehat{\beta}\|_{2,1} \leq \frac{1}{2}\|y - X\beta\|_2^2 + 2\lambda\sqrt{m}\|\beta\|_{2,1}.$$

Set $h = \widehat{\beta} - \beta$, and note that

$$\|y - X\widehat{\beta}\|_2^2 = \|(y - X\beta) - Xh\|_2^2 = \|Xh\|_2^2 + \|y - X\beta\|_2^2 - 2\langle Xh, y - X\beta\rangle.$$

Plugging this identity with $z = y - X\beta$ into the above inequality and rearranging the terms gives

$$\frac{1}{2}\|Xh\|_2^2 \leq \langle Xh, z\rangle + 2\lambda\sqrt{m}(\|\beta\|_{2,1} - \|\widehat{\beta}\|_{2,1}).$$

Next, break up h into h_I and $h_{I^C} = \widehat{\beta}_{I^C}$ and rewrite the above equation as

$$\frac{1}{2}\|Xh\|_2^2 \leq \langle h, X^*z\rangle + 2\lambda\sqrt{m}(\|\beta_I\|_{2,1} - \|\beta_I + h_I\|_{2,1} - \|h_{I^C}\|_{2,1}). \tag{2}$$

For each $i \in I$, we have

$$\begin{aligned}\|\widehat{\beta}_i\|_2 = \|\beta_i + h_i\|_2 &\geq \|\beta_i\|_2 - \frac{\langle h_i, \beta_i\rangle}{\|\beta_i\|_2} \geq \|\beta_i\|_2 - \left\langle h_i, \frac{\beta_i}{\|\beta_i\|_2}\right\rangle, \\ &\geq \|\beta_i\|_2 - \langle h_i, \overline{\text{sign}}(\beta_i)\rangle.\end{aligned}$$

This is due to the projection of h_i on $\text{span}\{\beta_i\}$ having magnitude $\frac{\langle h_i, \beta_i\rangle}{\|\beta_i\|_2}$. Thus, we can write $\|\widehat{\beta}_I\|_{2,1} \geq \|\beta_I\|_{2,1} - \langle h_I, \overline{\text{sign}}(\beta_I)\rangle$. Merging this inequality with (2) gives us

$$\begin{aligned}\frac{1}{2}\|Xh\|_2^2 &\leq \langle Xh, z\rangle + 2\lambda\sqrt{m}(\langle h_I, \overline{\text{sign}}(\beta_I)\rangle - \|h_{I^C}\|_{2,1}), \\ &= \langle h, X^*z\rangle + 2\lambda\sqrt{m}(\langle h_I, \overline{\text{sign}}(\beta_I)\rangle - \|h_{I^C}\|_{2,1}), \\ &= \langle h_I, X_I^*z\rangle + \langle h_{I^C}, X_{I^C}^*z\rangle + 2\lambda\sqrt{m}(\langle h_I, \overline{\text{sign}}(\beta_I)\rangle - \|h_{I^C}\|_{2,1}). \end{aligned} \tag{3}$$

We will need a brief lemma extending Hölder's inequality to the block norms defined earlier. Its proof is a simple exercise.

Lemma 2. *For all $a, b \in \mathbb{R}^{pm}$, $\langle a, b\rangle \leq \|a\|_{2,1}\|b\|_{2,\infty}$.*

The orthogonality condition and the lemma above implies

$$\langle h_{I^C}, X_{I^C}^*z\rangle \leq \|h_{I^C}\|_{2,1}\|X_{I^c}z\|_{2,\infty} \leq \sqrt{2m}\cdot\lambda\|h_{I^C}\|_{2,1}.$$

Merging this result with (3) results in

$$\frac{1}{2}\|Xh\|_2^2 \leq \langle h_I, v\rangle - (2 - \sqrt{2})\lambda\sqrt{m}\|h_{I^C}\|_{2,1}, \tag{4}$$

where $v = X_I^* z - 2\lambda\sqrt{m} \cdot \overline{\text{sign}}(\beta_I)$. We aim to bound each of the terms on the right hand side independently. For the first term, we have

$$\begin{aligned}\langle h_I, v\rangle &= \langle (X_I^* X_I)^{-1} X_I^* X_I h_I, v\rangle = \langle X_I^* X_I h_I, (X_I^* X_I)^{-1} v\rangle \\ &= \langle X_I^* X_I h, (X_I^* X_I)^{-1} v\rangle - \langle X_I^* X_I h_{I^C}, (X_I^* X_I)^{-1} v\rangle.\end{aligned}$$

Denote the two terms on the right hand side as A_1 and A_2, respectively. For A_1 we use Lemma 2 to obtain

$$A_1 \leq \|(X_I^* X_I)^{-1} v\|_{2,1} \|X_I^* X_I h\|_{2,\infty}.$$

Now we bound these two terms. For the first term,we get

$$\|(X_I^* X_I)^{-1} v\|_{2,1} \leq \sqrt{k}\|(X_I^* X_I)^{-1} v\|_2 \leq \sqrt{k}\|(X_I^* X_I)^{-1}\|_2 \|v\|_2 \leq 2k\|v\|_{2,\infty},$$

due to the invertibility condition. Using the orthogonality condition, we get

$$\|v\|_{2,\infty} = \|X_I^* z - 2\lambda\sqrt{m} \cdot \overline{\text{sign}}(\beta_I)\|_{2,\infty} \leq \|X_I^* z\|_{2,\infty} + 2\lambda\sqrt{m} \leq (2+\sqrt{2})\lambda\sqrt{m}.$$

For the second term, we use Lemma 1 and the orthogonality condition to get

$$\begin{aligned}\|X_I^* X_I h\|_{2,\infty} &\leq \|X_I^*(X\beta - y)\|_{2,\infty} + \|X_I^*(y - X\widehat{\beta})\|_{2,\infty} \\ &\leq \|X_I^* z\|_{2,\infty} + \|X_I^*(y - X\widehat{\beta})\|_{2,\infty} \leq (2+\sqrt{2})\lambda\sqrt{m}.\end{aligned}$$

So we get $A_1 \leq 2(2+\sqrt{2})^2\lambda^2 mk$. For A_2, we have from Lemma 2 that

$$|A_2| \leq \|h_{I^C}\|_{2,1} \|X_{I^C}^* X_I (X_I^* X_I)^{-1} v\|_{2,\infty} \leq (2-\sqrt{2})\lambda\sqrt{m}\|h_{I^C}\|_{2,1},$$

because of the complementary size condition. Using now these bounds on A_1, A_2, we have

$$\langle h_I, v\rangle \leq 2(2+\sqrt{2})^2\lambda^2 mk + (2-\sqrt{2})\lambda\sqrt{m}\|h_{I^C}\|_{2,1}.$$

Plugging this into (4) gives

$$\frac{1}{2}\|X(\beta - \widehat{\beta})\|_2^2 \leq 2(2+\sqrt{2})^2\lambda^2 mk,$$

proving the theorem. □

A.1 Invertibility condition

Define $Z = \|X_I^* X_I - \text{Id}\|_2$, where Id denotes the identity matrix. To study the distribution of Z, we first study the distribution of the sister random variable $Z' = \|X_{I'}^* X_{I'} - \text{Id}\|_2$, with I' following the probability model of Theorem 2. Note that under this modified model, $|I'|$ is a binomial random variable with parameter $\delta = k/p$. Using Markov's inequality and Theorem 2, we can show that

$$\begin{aligned}\mathbb{P}(Z' > 1/2) &\leq (1/2)^{-q}\mathbb{E}(Z'^q) \\ &\leq 2^q \left(20\mu_B \log(pm) + 9\sqrt{\delta \log(pm)(1+(m-1)\mu)}\|X\|_2 + \delta\|X\|_2^2\right)^q,\end{aligned}$$

where $q = 2\log(pm)$. Next, we will show that for all $t > 0$,

$$\mathbb{P}(Z > t) \leq 2\mathbb{P}(Z' > t), \tag{5}$$

an argument dubbed *Poissonization* in [4]. We write

$$\begin{aligned}
\mathbb{P}(\|X_{I'}^* X_{I'} - \mathrm{Id}\|_2 > t) &= \sum_{\ell=0}^{p} \mathbb{P}\left(\|X_{I'}^* X_{I'} - \mathrm{Id}\|_2 > t | |I'| = \ell\right) \mathbb{P}(|I'| = \ell) \\
&\geq \sum_{\ell=k}^{p} \mathbb{P}\left(\|X_{I'}^* X_{I'} - \mathrm{Id}\|_2 > t | |I'| = \ell\right) \mathbb{P}(|I'| = \ell) \\
&= \sum_{\ell=k}^{p} \mathbb{P}\left(\|X_{I_\ell}^* X_{I_\ell} - \mathrm{Id}\|_2 > t\right) \mathbb{P}(|I'| = \ell), \qquad (6)
\end{aligned}$$

where I_ℓ is selected uniformly at random from the set of subsets of $\{1, \ldots, p\}$ of cardinality ℓ. We now make two observations: (i) since $|I'|$ is a binomial random variable with parameters $(p, k/p)$, its median is simply given by k and therefore $\mathbb{P}(|I'| \geq k) \geq 1/2$; and (ii) $\mathbb{P}(\|X_{I_\ell}^* X_{I_\ell} - \mathrm{Id}\|_2 > t)$ is a nondecreasing function of ℓ, since $X_{I'_\ell}^* X_{I'_\ell} - \mathrm{Id}$ is a submatrix of $X_{I_\ell}^* X_{I_\ell} - \mathrm{Id}$ for $1 \leq \ell' \leq \ell$ and the spectral norm of a matrix is always greater than that of its submatrix. Therefore we can write

$$\begin{aligned}
\mathbb{P}(\|X_{I'}^* X_{I'} - \mathrm{Id}\|_2 > t) &\geq \mathbb{P}\left(\|X_{I_k}^* X_{I_k} - \mathrm{Id}\|_2 > t\right) \sum_{\ell=k}^{p} \mathbb{P}(|I'| = \ell) \\
&\geq \mathbb{P}\left(\|X_{I_k}^* X_{I_k} - \mathrm{Id}\|_2 > t\right) \mathbb{P}(|I'| \geq k) \\
&\geq \frac{1}{2}\mathbb{P}\left(\|X_{I_k}^* X_{I_k} - \mathrm{Id}\|_2 > t\right) \\
&= \frac{1}{2}\mathbb{P}\left(\|X_I^* X_I - \mathrm{Id}\|_2 > t\right), \qquad (7)
\end{aligned}$$

since I_k and I have the same probability distribution. Thus, by merging (6) and (7), we obtain

$$\mathbb{P}\left(Z > \frac{1}{2}\right) \leq 2^{q+1}[10\mu_B \log(pm) + 9\sqrt{\delta \log(pm)(1 + (m-1)\mu)}\|X\|_2 + \delta\|X\|_2^2]^q,$$

We can make the constants C_0, C_1 large enough so that the term inside brackets on the right hand side is bounded by $1/4$, giving

$$\mathbb{P}(Z > 1/2) \leq 2(1/2)^{2\log(pm)} = 2(pm)^{-2\log 2} \leq 2(pm)^{-2\log 2}.$$

A.2 Orthogonality condition

Note that $\|X^* z\|_{2,\infty} \leq \sqrt{2} \cdot \lambda\sqrt{m}$ is implied by $\|X^* z\|_\infty \leq \sqrt{2} \cdot \lambda$, which matches the orthogonality condition of Theorem 1.2 of [4] with only the number of columns changing from p to pm. Therefore, the condition holds with probability at least $1 - (pm)^{-1}(2\pi \log(pm))^{-1/2}$.

A.3 Complementary size condition

We partition the complementary size condition into two statements:

$$\|X_{I^C}^* X_I (X_I^* X_I)^{-1} \overline{\text{sign}}(\beta_I)\|_{2,\infty} \leq \frac{1}{4}, \tag{8}$$

$$\|X_{I^C}^* X_I (X_I^* X_I)^{-1} X_I^* z\|_{2,\infty} \leq \left(\frac{3}{2} - \sqrt{2}\right) \lambda \sqrt{m}. \tag{9}$$

We begin with the first inequality (8). Denote the vector

$$Z_{0,i} = X_i^* X_I (X_I^* X_I)^{-1} \overline{\text{sign}}(\beta_I)$$

for each $i \in I^C$. Further denote

$$Z_0 = \max_{i \notin I} \|Z_{0,i}\|_2 = \|X_{I^C}^* X_I (X_I^* X_I)^{-1} \overline{\text{sign}}(\beta_I)\|_{2,\infty}.$$

We then simply need to show that with large probability $Z_0 \leq 1/4$. Define the matrix $W_i = (X_I^* X_I)^{-1} X_I^* X_i$ for $i \notin I$. Further, denote by W_i^j, $1 \leq j \leq |I|$, the submatrix of W_i containing its j^{th} block of rows. We can then write $Z_{0,i} = \sum_{j=1}^{|I|} W_i^{j*} \overline{\text{sign}}(\beta_j)$. where W_i^{j*} is the adjoint of W_i^j. The sum terms have norms bounded by

$$\left\| W_i^{j*} \overline{\text{sign}}(\beta_j) \right\|_2 \leq \left\| W_i^j \right\|_2 \left\| \overline{\text{sign}}(\beta_j) \right\|_2 = \left\| W_i^j \right\|_2.$$

At this point we make use of the vector Bernstein inequality from [5, 13].

Lemma 3. *Let $\{v_k\} \in \mathbb{R}^m$ be a finite sequence of independent random vectors. Suppose that $\mathbb{E}(v_k) = 0$ and $\|v_k\|_2 \leq B$ almost surely, and put $\sigma^2 \geq \sum_k \mathbb{E}\|v_k\|_2^2$. Then for all $0 \leq t \leq \sigma^2/B$,*

$$\mathbb{P}\left(\left\| \sum_k v_k \right\|_2 \geq t \right) \leq e^{-\frac{t^2}{8\sigma^2} + \frac{1}{4}}.$$

We use the lemma by setting $B = \max_{1 \leq j \leq |I|} \|W_i^j\|_2$ and $\sigma^2 = \sum_{j=1}^{|I|} \|W_i^j\|_2^2 = \|W_i\|_2^2$ to obtain

$$\mathbb{P}(|Z_{0,i}| > t) \leq 2e^{-t^2/8 \max_{j \in I} \|W_i\|_2^2}$$

for $0 \leq t \leq 1$ (as $\sigma^2 > B$). A union bound then gives us $\mathbb{P}(Z_0 > t) \leq 2pme^{-t^2/8\kappa^2}$, where $\kappa > \max_{i \notin I} \|W_i\|_2$. We can see that under the invertibility condition,

$$\max_{i \notin I} \|W_i\|_2 = \max_{i \notin I} \|(X_I^* X_I)^{-1} X_I^* X_i\|_2 \leq 2 \max_{i \notin I} \|X_I^* X_i\|_2 = 2\|X_I^* X_{I^C}\|_{B,1}.$$

Thus, conditioned on a bound

$$\gamma > \|X_I^* X_{I^C}\|_{B,1} \tag{10}$$

and the invertibility condition, and replacing $t = 1/4$, the probability of the first inequality failing to hold is at most $2pme^{-1/512\gamma^2}$.

The second condition (9) is implied by

$$\|X_{I^C}^* X_I (X_I^* X_I)^{-1} X_I^* z\|_\infty \leq \left(\frac{3}{2} - \sqrt{2}\right) \lambda. \tag{11}$$

To prove this equivalent condition, we use the second half of [4, Lemma 3.3], restated below.

Lemma 4. *Let $(W_j')_{j\in J}$ be a fixed collection of vectors in $\mathbb{R}^n$ and set $Z_1 = \max_{j\in J} |\langle W_j', z\rangle|$. We then have $\mathbb{P}(Z_1 \geq t) \leq 2|J|e^{-t^2/2(\kappa')^2}$ for any $\kappa' \geq \max_{j\in J} \|W_j'\|_2$.*

We denote $W_{ij}' = X_I(X_I^* X_I)^{-1} X_I^* X_{ij}$ for $i \notin I$, $1 \leq j \leq m$. Then we can write

$$Z_1 = \|X_{I^C}^* X_I (X_I^* X_I)^{-1} z\|_\infty = \max_{i\notin I, 1\leq j\leq m} |\langle W_{i,j}', z\rangle|.$$

To use Lemma 4 in this case, we assume that the invertibility condition holds and search for a bound on κ':

$$\begin{aligned}\kappa' &= \max_{i\notin I, 1\leq j\leq m} \|X_I(X_I^* X_I)^{-1} X_I^* X_{ij}\|_2 \leq \sqrt{2} \max_{i\notin I, 1\leq j\leq m} \|X_I^* X_{ij}\|_2 \\ &\leq \sqrt{2} \max_{i\notin I} \|X_I^* X_i\|_2 = \sqrt{2}\|X_I^* X_{I^C}\|_{B,1} \leq \sqrt{2}\gamma.\end{aligned}$$

Thus, we have that conditioned on the bound (10) and the invertibility condition, (11) holds except with probability at most $2pme^{-(3/2-\sqrt{2})^2\lambda^2/4\gamma^2}$.

To finalize, we define the event

$$E = \{Z \leq 1/2\} \cup \{\|X_I^* X_{I^C}\|_{B,1} \leq \gamma\}.$$

Then we have that the probability P of the complementary size condition not being met is upper bounded by

$$\begin{aligned}P &\leq \mathbb{P}(\{Z_0 > 1/4\} \cup \{Z_1 \geq (3/2 - \sqrt{2})\lambda\}|E) + \mathbb{P}(E^C) \\ &\leq 2pme^{-1/512\gamma^2} + 2pme^{-(3/2-\sqrt{2})^2\lambda^2/4\gamma^2} + \mathbb{P}(Z > 1/2) + \mathbb{P}(\|X_I^* X_{I^C}\|_{B,1} > \gamma) \\ &\leq 2pme^{-1/512\gamma^2} + 2pme^{-(3/2-\sqrt{2})^2\lambda^2/4\gamma^2} + 2(pm)^{-2\log 2} + \mathbb{P}(\|X_I^* X_{I^C}\|_{B,1} > \gamma).\end{aligned}$$

We set $\gamma = C_2/\sqrt{\log(pm)}$ so that each of the first two terms of the right hand side is upper bounded by $2(pm)^{-2\log 2}$. To get the probability of the bound (10) being valid, we appeal to [2, Lemma 5] together with the Markov inequality and a Poissonization argument (see (5) and (7) for an example) to obtain

$$\begin{aligned}\mathbb{P}(\|X_I^* X_{I^C}\|_{B,1} > \gamma) &\leq 2\gamma^{-q}\mathbb{E}(\|X_I^* X_{I^C}\|_{B,1}^q) \\ &\leq 2\gamma^{-q}(2^{1.5}\sqrt{q}\mu_B + \sqrt{\delta(1+(m-1)\mu)}\|X\|_2 + \delta\|X\|_2^2)^q\end{aligned}$$

where $q = 2\log(pm)$. We replace the value of γ and q selected above as well as the bounds on k, μ and μ_B from the theorem to obtain

$$\mathbb{P}\left(\|X_I^* X_{I^C}\|_{B,1} > \frac{C_2}{\sqrt{m\log(pm)}}\right) \leq 2\left(\frac{8C_1}{C_2} + \frac{2}{C_2}\sqrt{\frac{2C_0}{m}} + \frac{2C_0}{C_2}\right)^{2\log(pm)}.$$

By picking the constants C_0, C_1, C_2 small enough so that the base of the exponential term on the right hand side is less than $1/2$, we get $\mathbb{P}(\|X_I^* X_{I^C}\|_{B,1} > C_2/\sqrt{\log(pm)}) < (pm)^{-2\log 2}$. Thus, the complementary size condition holds with probability at least $1 - 8(pm)^{-2\log 2}$.

By joining the three conditions (noting that the third condition already accounts for the first), we have that Theorem 1 holds with probability at least $1 - 8(pm)^{-2\log 2} - (pm)^{-1}(2\pi\log(pm))^{-1/2}$. □

References

[1] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Machine Learning Research*, 9(6):1179–1225, June 2008.

[2] W. Bajwa, R. Calderbank, and M. F. Duarte. On the conditioning of random block subdictionaries. Technical Report TR-2010-06, Duke University, Department of Computer Science, Durham, NC, Sept. 2010.

[3] R. Calderbank, R. H. Hardin, E. M. Rains, P. W. Shor, and N. J. A. Sloane. A group-theoretic framework for the construction of packings in Grassmanian spaces. *J. Algebraic Combinatorics*, 9(2):129–140, Mar. 1999.

[4] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics*, 37(5A):2145–2177, Oct. 2009.

[5] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. 2010. Preprint.

[6] S.F. Cotter, B.D. Rao, E. Kjersti, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, 53(7):2477–2488, July 2005.

[7] M. E. Davies and Y. C. Eldar. Rank awareness in joint sparse recovery. Apr. 2010. submitted to *IEEE Trans. Info. Theory*.

[8] Y. C. Eldar, P. Kuppinger, and H. Bölcksei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Signal Processing*, 58(6):3042–3054, June 2010.

[9] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Info. Theory*, 55(11):5302–5316, 2009.

[10] Y. C. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans. Info. Theory*, 6(1):505–519, Jan. 2010.

[11] M. Fornasier and H. Rauhut. Recovery algorithms for vector valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613, 2008.

[12] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *J. Fourier Anal. Appl.*, 14(5):655–687, 2008.

[13] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Info. Theory*, 2010. To appear.

[14] J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38(4):1978–2004, Aug. 2010.

[15] H. Liu and J. Zhang. Estimation consistency of the group lasso and its applications. In *Int. Conf. Artificiall Intelligence and Statistics (AISTATS)*, pages 376–383, Clearwater Beach, FL, Apr. 2009.

[16] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *J. Royal Statist. Soc. B*, 70(1):53–71, Jan. 2008.

[17] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Statistics*, 2:605–633, 2008.

[18] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39(1):1–47, Jan. 2011.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc B*, 58(1):267–288, 1996.

[20] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86, Apr. 2006.

[21] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Acad. Sci. Paris, Ser. I*, 346(23–24):1271–1274, 2008.

[22] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86:572–588, Apr. 2006.

[23] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. B*, 68(1):49–67, Feb. 2006.