The Group Dantzig Selector

Han Liu
Machine Learning Department
Carnegie Mellon University

Jian Zhang
Statistics Department
Purdue University

Xiaoye Jiang
Computational Mathematics
Stanford University

Jun Liu
Computer Science Department
Arizona State University

Abstract

We introduce a new method — the *group Dantzig selector* — for high dimensional sparse regression with group structure, which has a convincing theory about why utilizing the group structure can be beneficial. Under a *group restricted isometry condition*, we obtain a significantly improved nonasymptotic ℓ_2-norm bound over the basis pursuit or the Dantzig selector which ignores the group structure. To gain more insight, we also introduce a surprisingly simple and intuitive *sparsity oracle condition* to obtain a block ℓ_1-norm bound, which is easily accessible to a broad audience in machine learning community. Encouraging numerical results are also provided to support our theory.

1 INTRODUCTION

Grouped variables appear naturally in high dimensional statistical learning problems. For example, in data mining applications, categorical features are usually encoded via a set of dummy variables and as a result such dummy variables form a group. Another example is learning sparse additive models, where each component function can be represented using basis expansions and thus can be treated as a group. For such problems, it is more natural and suitable to select groups of variables instead of individual ones when a sparse model is preferred in statistical inference. These problems motivate the introduction of the group Lasso (Yuan and Lin, 2006), which extends the popular Lasso method (Tibshirani, 1996; Chen et al., 1998) by replacing the ℓ_1-norm regularization with a sum of ℓ_2-norm (or, block ℓ_1-norm) regularization. Such a regularization has the effect of "grouping" all the variables within each group so that their resulting estimates tend to zeroes or nonzeroes simultaneously.

Some recent work has addressed certain statistical properties of the group Lasso. Under a fixed group size assumption, Bach (2008) provides both necessary and sufficient conditions for variable selection consistency using the random design. Meier et al. (2007) provide a risk consistency result. Chesneau and Hebiri (2008) provide a sparsity oracle inequality. Under fixed design conditions, some asymptotic properties like estimation consistency and risk consistency have been shown in (Ravikumar et al., 2007) and (Nardi and Rinaldo, 2008). Liu and Zhang (2009) also provide a fixed design analysis of the ℓ_2-norm consistency of the group Lasso and apply it in fitting sparse additive models. However, none of these work demonstrates under what conditions the group Lasso can be superior to the Lasso. A satisfactory answer to this question recently appears in (Huang and Zhang, 2009). The key observation is a phenomenon called group noise condition, which says that for the ℓ_2-norm of the projected noise on the column span of variables within each group, the concentration term does not increase with the group size. Under a group sparse eigenvalue assumption, they show that the group Lasso can achieve a superior sample complexity and is more robust than the Lasso when correct group structure is availabe. Some related work also appear in the multi-task learning lietrature, see (Lounici et al., 2009; Obozinski and Wainwright, 2008).

In this paper, we introduce a novel sparse learning method which can also take advantage of the group structure. Our method is called the *group Dantzig selector*, which is an extension of the Dantzig selector (Candes and Tao, 2007) such that the group information can be explicitly encoded in a convex optimization problem. We show that it achieves similar theoretical performance as the group Lasso (Huang and Zhang, 2009) under a weaker and more understandable condition. Just like the Dantzig selector is a cousin of the Lasso (Meinshausen et al., 2007), the group Dantzig selector is a cousin of the group Lasso. The contri-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

butions of this paper include: (i) the formulation of a novel convex optimization problem for sparse learning with group structure, (ii) a convincing theory showing why the group Dantzig selector can effectively utilize the group information using the classical ℓ_2-norm as evaluation metric. (iii) a simple and clean analytical framework under which the effectiveness of the group Dantzig selector can be demonstrated using the block ℓ_1-norm as evaluation metric. In summary, our paper presents a novel method with strong theoretical guarantees and the results are easily accessible to a wide audience in machine learning community.

The rest of this paper is organized as follows. Section 2 introduces necessary notations and problem formulations. Section 3 presents in detail the group Dantzig selector. Section 4 studies the ℓ_2-theory. Section 5 studies the ℓ_1-theory. Section 6 reports empirical results of the group Dantzig selector and their comparisons with the Lasso and group Lasso. Section 7 provides some summary conclusions.

2 BACKGROUND

We start with some notations. Consider a $n \times p$ design matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ where $\mathbf{x}_j \in \mathbf{R}^n$ for all $j \in \{1, \ldots, p\}$. The response vector $\mathbf{y} = (y^1, \ldots, y^n) \in \mathbf{R}^n$ is assumed to be generated from the linear model $\mathbf{y} = X\beta + \epsilon$ with $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$. Although we assume ϵ is Gaussian, it should be clear from the analysis that the result can be straightforwardly extended to the sub-Gaussian family. We define the support of β to be $\mathrm{supp}(\beta) = \{j : \beta_j \neq 0, j = 1, \ldots, p\}$.

Given $M \subset \{1, \ldots, p\}$, we denote β_M to be the subvector of β with elements indexed by M. Similarly, X_M is the $n \times |M|$ submatrix of X with columns indexed by elments in M. Given a vector $v \in \mathbf{R}^n$, let $1 < q < \infty$, we denote $\|v\|_q = (v_1^q + \ldots + v_n^q)^{1/q}$, and we use $\|v\|_\infty = \max_{1 \leq j \leq n} |v_i|$ and $\|v\|_0 = |\mathrm{supp}(v)|$. Without loss of generality, we assume $\|\mathbf{x}_j\|_2 = 1$ in this paper.

Two very popular methods for sparse learning are the Lasso and the Dantzig selector, which are formulated as the solutions of the following convex optimization problems:

$$\text{(Lasso)} \qquad \widehat{\beta}^L = \arg\min_\beta \|\beta\|_1 \quad \text{s.t. } \|\mathbf{y} - X\beta\|_2 \leq \eta. \tag{1}$$

$$\text{(Dantzig selector)} \qquad \widehat{\beta}^D = \arg\min_\beta \|\beta\|_1 \quad \text{s.t. } \|X^T(\mathbf{y} - X\beta)\|_\infty \leq \lambda. \tag{2}$$

where η and λ are positive regularization parameters.

There are cases where we know a group structure on β in advance, and the variables belonging to the same group should be simultaneously included in or excluded from the model. In this case, we assume $\{1, \ldots, p\} = \cup_{j=1}^d G_j$ is partitioned into d nonoverlap groups $G_1, \ldots, G_d$. Let $M \subset \{1, \ldots, d\}$ be an index set of groups, we denote $G_M = \cup_{j \in M} G_j$. We also denote $\mathrm{gsupp}(\beta) = \{j : \|\beta_{G_j}\|_2 \neq 0, j = 1, \ldots, d\}$. In this paper, we always denote $\mathrm{gsupp}(\beta) = F$. The Group Lasso can be applied to solve the group sparse learning problem:

$$\text{(group Lasso)} \qquad \widehat{\beta}^{gL} = \arg\min_\beta \sum_{j=1}^d \|\beta_{G_j}\|_2 \quad \text{s.t. } \|\mathbf{y} - X\beta\|_2 \leq \eta_g. \tag{3}$$

where η_g is a positive regularization parameter.

In the sequel, to simplify notations and convey the key message clear, we only consider the case when $p, d \gg n$. We also assume that all the groups have equal size: $\forall j \in \{1, \ldots, d\}$, $|G_j| = k_0$. This is not a restriction of our method, as our analysis can be easily extended to handle the uneven group cases as in (Huang and Zhang, 2009). A full treatment of uneven sized groups will be reported in a longer technical report.

3 THE GROUP DANTZIG SELECTOR

In this section, we first introduce a new estimator called the group Dantzig selector as follows:

Definition 1. (The group Dantzig selector) *For sparse learning problems with given group structure, the group Dantzig estimator $\widehat{\beta}^{gD}$ is a solution to the following convex optimization problem*

$$\widehat{\beta}^{gD} = \arg\min_\beta \sum_{j=1}^d \|\beta_j\|_2 \quad \text{s.t. } \max_{1 \leq j \leq d} \|X_{G_j}^T (\mathbf{y} - X\beta)\|_2 \leq \lambda_g, \tag{4}$$

where $\lambda_g \geq 0$ is a regularization parameter.

The problem in (4) is convex, thus can be easily solved by any off-the-shelf convex programming solver. Indeed, our implementation of the group Dantzig selector is based on a variant of the spectral gradient projection method, which achieves a comparable speed and scalability as the coordinate descent methods of the group Lasso . Therefore, instead of emphasizing too much on the computation, we mainly consider motivation and theoretical advantages of the group Dantzig selector in the rest of this paper.

At the first sight of (4), people might wonder why we adopt $\max_{1 \leq j \leq d} \|X_{G_j}^T (\mathbf{y} - X\beta)\|_2 \leq \lambda_g$ as the constraint term instead of $\|X^T(\mathbf{y} - X\beta)\|_\infty$, since the

latter one also seems a natural candidate for what is being called the "group" Dantzig selector since the group information is already encoded in the block ℓ_1-nrom objective function. To answer this question, we first present a proposition which provides an equivalent form of the group Lasso problem defined in (3). Since the group Lasso becomes the Lasso when the group size equals one, the result also illustrates an interesting relationship between the Lasso and the Dantzig selector. Based on this connection, by "mimicking" how the Dantzig selector "modifies" the Lasso, we modify the group Lasso to obtain the group Dantzig selector.

Proposition 1. *$\widehat{\beta}^{gL}$ is a solution to the group Lasso problem defined in* (3) *with the regularization parameter η_g, if and only if there exists a nonnegative number λ_g, such that it is also the solution of the following optimization problem:*

$$\begin{aligned}\widehat{\beta}^{gL} &= \arg\min_{\beta} \|X\beta\|_2 \qquad (5)\\ &\quad \text{s.t.} \quad \max_{1\le j\le d} \|X_{G_j}^T (\mathbf{y} - X\beta)\|_2 \le \lambda_g.\end{aligned}$$

Proof. From the Lagrangian duality, for each η_g there exists a nonnegative λ_g such that each $\widehat{\beta}^{gL}$ solves (3) also solves the unconstrained optimization problem:

$$\min_{\beta} F(\beta) \qquad (6)$$

$$\text{where} \quad F(\beta) = \frac{1}{2}\|\mathbf{y} - X\beta\|_2^2 + \lambda_g \sum_{j=1}^{d} \|\beta_{G_j}\|_2.$$

It then suffices to show that (6) is equivalent to (5). This follows by a standard variational duality argument. Define

$$H(\beta, v) = \frac{1}{2}\|X\beta\|_2^2 - \mathbf{y}^T X\beta + \lambda \sum_{j=1}^{d} v_{G_j}^T \beta_{G_j},$$

it's obvious that

$$F(\beta) = \max_{\forall j, \|v_{G_j}\|_2 \le 1} H(\beta, v).$$

By strong duality, the desired result follows from the chain

$$\begin{aligned}\min_{\beta} F(\beta) &= \min_{\beta} \sup_{\forall j, \|v_{G_j}\|_2 \le 1} H(\beta, v)\\ &= \sup_{\forall j, \|v_{G_j}\|_2 \le 1} \min_{\beta} H(\beta, v)\\ &= \sup_{\forall j, \|v_{G_j}\|_2 \le 1, v_{G_j} = \frac{1}{\lambda_g} X_{G_j}^T (\mathbf{y} - X\beta)} \min_{\beta} H(\beta, v)\\ &= \sup_{\forall j, \|X_{G_j}^T (\mathbf{y} - X\beta)\|_2 \le \lambda_g} -\frac{1}{2}\|X\beta\|_2^2,\end{aligned}$$

where the third equality utilizes the saddle point conditions. □

Given the above proposition, setting the group size to be 1, i.e. $k_0 = 1$, we see that $\widehat{\beta}^L$ is a solution to the Lasso problem in (1) with the regularization parameter η, if and only if there exists a nonnegative number λ, such that

$$\widehat{\beta}^L = \arg\min_{\beta} \|X\beta\|_2 \ \text{ s.t. } \|X^T (\mathbf{y} - X\beta)\|_\infty \le \lambda. \quad (7)$$

Comparing (7) with (2), the Danzig selector simply modifies an equivalent form of the Lasso by replacing $\|X\beta\|_2$ with $\|\beta\|_1$. Using the same strategy, if we want a "group" version Dantzig selector, we need to change the objective function $\|X\beta\|_2$ of the equivalent form of the group Lasso in (5). The most natural candidate would be the block ℓ_1-norm. Therefore the definition in (4) naturally comes out.

Under certain conditions, the solution $\widehat{\beta}^D$ to the dantzig selector in (2) can be shown to be identical to the Lasso solution $\widehat{\beta}^L$ defined in (7) using the same regularization parameter λ (James et al., 2009). A similar argument is also true for the group Dantzig selector. A necessary and sufficient condition that characterizes the equality of the group Lasso and the Dantzig selector will be reported in the full version. The next proposition only considers the simplest orghogonal group design case, to shed light on the relationship between the group Lasso and the group Dantzig selector. The proof is omitted due to a space limit.

Proposition 2. *Let $\widehat{\beta}^{gD}$ be the unique group Dantzig selector solution in* (4) *and $\widehat{\beta}^{gL}$ be the unique group Lasso solution in* (5)*, both using the same regularization parameter λ_g. If for all $j, k \in \{1, \ldots, d\}$, $j \neq k$,*

$$X_{G_j}^T X_{G_j} = I_{k_0} \ \text{ and } \ X_{G_j}^T X_{G_k} = \mathbf{0},$$

where I_{k_0} is the k_0-dimensional identity matrix, then $\widehat{\beta}^{gD} = \widehat{\beta}^{gL}$.

4 ℓ_2-THEORY

In this section, we study the theoretical property of the group Dantzig selector using the ℓ_2-norm criteria $\|\widehat{\beta}^{gD} - \beta\|_2$. The first result of this paper is that the group Dantzig selector can be much more accurate than the Lasso or Dantzig selector when suitable group structure is available. Our analysis is based on a technique in a recent Lasso and Dantzig selector analysis in Cai et al. (2009) and extends the framework of the *Restricted Isometry Property* (RIP) introduced by Candes and Tao (2005) to the group version. Two key quantities are needed to establish the conditions:

Definition 2. (m-group restricted isometry coefficient) *: For $1 \le m \le d$, the m-group restricted isome-*

try coefficient δ_m for X is defined as

$$\begin{aligned} \delta_m \quad &= \quad \inf_{\delta \geq 0} \delta \qquad (8) \\ &\text{s.t.} \quad \sqrt{1-\delta}\|\gamma\|_2 \leq \|X\gamma\|_2 \leq \sqrt{1+\delta}\|\gamma\|_2 \\ &\text{for all} \quad \gamma \in \mathbf{R}^p,\ |\text{gsupp}(\gamma)| \leq m. \end{aligned}$$

Definition 3. (m, m'-group restricted orthogonality coefficient)*: For $1 \leq m, m' \leq d$, the m, m'-group restricted orthogonality coefficient $\theta_{m,m'}$ of X is defined as the smallest number that satisfies*

$$|\gamma^T X^T X \gamma'| \leq \theta_{m,m'} \|\gamma\|_2 \|\gamma'\|_2 \qquad (9)$$

where $\gamma, \gamma' \in \mathbf{R}^p$ are arbitrary, satisfying $|\text{gsupp}(\gamma)| \leq m$, $|\text{gsupp}(\gamma')| \leq m'$ with disjoint support.

From the above definitions, when there is no group structure available, i.e. when $k_0 = 1$, the definitions of δ_m and $\theta_{m,m'}$ exactly reduce to be the classical restricted isometry coefficient and m, m'-restricted isometry coefficient as in (Candes and Tao, 2005). So our definitions is a natural extension of theirs by incorporating the group structure. In general, we prefer δ and θ to be small. It is easy to see that when group structure is available, both δ_m and $\theta_{m,m'}$ will be smaller than their corresponding non-group versions. In other words, this illustrates that utilizing group structure can be helpful in weakening assumptions (Huang and Zhang, 2009).

Theorem 1 provides a sharp non-asymptotic bound for the estimation error $\|\widehat{\beta}^{gD} - \beta\|_2$. It also illustrates the striking advantage of the group Dantzig selector over the Lasso or the Dantzig selector when group structure is available. We start with an assumption which requires the largest eigenvalue of X_{G_j} to be bounded from above. This is not a restriction in many cases since we often have control over the choice of X_{G_j}.

Assumption 1. *For $\forall j \in \{1, \ldots, d\}$, X_{G_j} is full rank and the largest eigenvalues of all X_{G_j} are uniformly bounded from above by $1/\sqrt{\rho_{k_0}}$.*

Theorem 1. *(ℓ_2-bound) Let $\widehat{\beta}^{gD}$ be a solution to* (4) *with*

$$\lambda_g = \frac{\sigma}{\sqrt{\rho_{k_0}}} \left(\sqrt{k_0} + \sqrt{2}\sqrt{\log d + \log n} \right).$$

Let $|F| = s$, without loss of generality, we assume $s/4$ is an integer. Under Assumption 1 and

$$\delta_{1.25s} + \theta_{s,1.25s} < 1, \qquad (10)$$

then we have with probability larger than $1 - 1/n$

$$\|\widehat{\beta}^{gD} - \beta\|_2 \leq C \cdot (\sqrt{sk_0} + \sqrt{2}\sqrt{s\,(\log d + \log n)}), \qquad (11)$$

where

$$C = \frac{\sqrt{20}\sigma}{(1 - \delta_{1.25s} - \theta_{s,1.25s})\sqrt{\rho_{k_0}}}.$$

Proof. In the following, we suppress the superscript of $\widehat{\beta}^{gD}$ and simply denote it as $\widehat{\beta}$. We start with several simple technical lemmas. Lemma 1 purely depends on the objective function of the optimization problem, and this is the only place where the optimization formulation plays a role.

Lemma 1. *Let $\widehat{\beta}$ be a solution to* (4)*, then we have $\sum_{j \in F^c} \|\widehat{\beta}_{G_j}\|_2 \leq \sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2$.*

Proof. Since $\widehat{\beta}$ is the minimizer of the objective function in (4) and $\beta_{F^c} = \mathbf{0}$, we have

$$\begin{aligned} \sum_{j \in F} \|\beta_{G_j}\|_2 &\geq \sum_{j \in F} \|\widehat{\beta}_{G_j}\|_2 + \sum_{j \in F^c} \|\widehat{\beta}_{G_j}\|_2 \\ &\geq \sum_{j \in F} \|\beta_{G_j}\|_2 - \sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2 + \sum_{j \in F^c} \|\widehat{\beta}_{G_j}\|_2. \end{aligned}$$

We obtain the desired result after canceling the term $\sum_{j \in F} \|\beta_{G_j}\|_2$ on both sides. □

Lemma 2. *Let $M \subset \{1, \ldots, d\}$. With probability larger than $1 - 1/n$, we have*

$$\begin{aligned} &\max_{j \in M} \|X_{G_j}^T X(\widehat{\beta} - \beta)\|_2 \\ &\quad \leq \quad \frac{2\sigma}{\sqrt{\rho_{k_0}}} \left(\sqrt{k_0} + \sqrt{2}\sqrt{\log(|M|) + \log n} \right). \qquad (12) \end{aligned}$$

Proof. For all $j \in \{1, \ldots, d\}$ we have

$$\|X_{G_j}^T X(\widehat{\beta} - \beta)\|_2 \leq \|X_{G_j}^T (X\widehat{\beta} - \mathbf{y})\|_2 + \|X_{G_j}^T (\mathbf{y} - X\beta)\|_2.$$

Since $\|X_{G_j}^T (X\widehat{\beta} - \mathbf{y})\|_2 \leq \lambda_g$ and by Assumption 1, we have

$$\begin{aligned} &\forall j \in \{1, \ldots, d\}, \quad \|X_{G_j}^T (\mathbf{y} - X\beta)\|_2 \qquad (13) \\ &\quad \leq \quad \frac{1}{\sqrt{\rho_{k_0}}} \left\| \left(X_{G_j}^T X_{G_j} \right)^{-0.5} X_{G_j}^T (\mathbf{y} - X\beta) \right\|_2. \end{aligned}$$

The result follows from a direct application of Proposition 4.1 in (Huang and Zhang, 2009), coupled with a union bound over the group index set M. □

Without loss of generality, we assume $F = \{1, \ldots, s\}$. After some rearrangement, we also assume

$$\begin{aligned} &\|\widehat{\beta}_{G_{s+1}} - \beta_{G_{s+1}}\|_2 \\ &\quad \geq \quad \|\widehat{\beta}_{G_{s+2}} - \beta_{G_{s+2}}\|_2 \geq \cdots \geq \|\widehat{\beta}_{G_d} - \beta_{G_d}\|_2. \end{aligned}$$

Then, denote $F^* = \{s+1, \ldots, 1.25s\}$ and for all $i \geq 1$, $F_i = \{1.25s + (i-1)s + 1, \ldots, 1.25s + is\}$ with the last subset of size less than or equal to s. Also, for each F_i ($i \geq 1$), we denote F_{i1} be the subset containing the first $0.75s$ elements and F_{i2} be the subset contains the remaining.

The next technical lemma is obtained from (Cai et al., 2009): for a non-decreasing sequence of nonnegative numbers, the ℓ_2-norm of the last 4/5 proportion can be bounded by the ℓ_1-norm of the first 4/5 proportion over a large constant factor.

Lemma 3. *Given any descending chain of real numbers $a_1 \geq a_2 \geq \cdots \geq a_{0.25s} \geq b_1 \geq \cdots \geq b_{0.75s} \geq c_1 \geq \cdots \geq c_{0.25s} \geq 0$, we have*

$$\sqrt{\sum_{i=1}^{0.75s} b_i^2 + \sum_{i=1}^{0.25s} c_i^2} \leq \frac{\sum_{i=1}^{0.25s} a_i + \sum_{i=1}^{0.75s} b_i}{\sqrt{s}}. \quad (14)$$

Proof. Follows by a direct application of Lemma 2 in (Cai et al., 2009). □

Given Lemma 3, we obtain the following key lemma:

Lemma 4. *Given F, F^*, F_i, we have*

$$\sum_{i\geq 1} \sqrt{\sum_{j\in F_i} \|\widehat{\beta}_{G_j}\|_2^2} \leq \sqrt{\sum_{j\in F\cup F^*} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2^2}.$$

Proof. By (14) in Lemma 3, we directly obtain

$$\begin{aligned}
&\sqrt{s}\sum_{i\geq 1} \sqrt{\sum_{j\in F_i} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2^2} \\
&\leq \sum_{j\in \cup_{i\geq 1} F_i} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2 + \sum_{j\in F^*} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2 \\
&= \sum_{j\in F^c} \|\widehat{\beta}_{G_j}\|_2.
\end{aligned}$$

Lemma 1 further implies that $\sum_{j\in F^c} \|\widehat{\beta}_{G_j}\|_2 \leq \sum_{j\in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2$. The desired result follows from the fact $\sum_{j\in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2 \leq \sqrt{s}\sqrt{\sum_{j\in F\cup F^*} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2^2}$. □

To finalize Theorem 1, we consider a pivotal quantity $|\langle X(\widehat{\beta} - \beta), \sum_{j\in F\cup F^*} X_{G_j}(\widehat{\beta}_{G_j} - \beta_{G_j})\rangle|$. Let

$$K(\widehat{\beta}, F, F^*) = \sqrt{\sum_{j\in F\cup F^*} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2^2} \quad (15)$$

It is obvious that, with probability larger than $1 - \frac{1}{n}$,

$$\begin{aligned}
&|\langle X(\widehat{\beta} - \beta), \sum_{j\in F\cup F^*} X_{G_j}(\widehat{\beta}_{G_j} - \beta_{G_j})\rangle| \quad (16)\\
&\leq \sum_{j\in F\cup F^*} |\langle X_{G_j}^T X(\widehat{\beta} - \beta), \widehat{\beta}_{G_j} - \beta_{G_j}\rangle| \\
&\leq \sqrt{1.25s} \max_{1\leq j\leq d} \|X_{G_j}^T X(\widehat{\beta} - \beta)\|_2 K(\widehat{\beta}, F, F^*) \\
&\leq \sigma\sqrt{\frac{5s}{\rho_{k_0}}}(\sqrt{k_0} + \sqrt{2}\sqrt{\log d + \log n}) K(\widehat{\beta}, F, F^*).
\end{aligned}$$

where we apply the Cauchy-Schwartz inequality twice in the second inequality and the third inequality follows from Lemma 2.

On the other hand, we have

$$\begin{aligned}
&|\langle X(\widehat{\beta} - \beta), \sum_{j\in F\cup F^*} X_{G_j}(\widehat{\beta}_{G_j} - \beta_{G_j})\rangle| \\
&\geq \| \sum_{j\in F\cup F^*} X_{G_j}(\widehat{\beta}_{G_j} - \beta_{G_j})\|_2^2 \\
&- \sum_{i\geq 1} |\langle \sum_{j\in F_i} X_{G_j}(\widehat{\beta}_{G_j} - \beta_{G_j}), \sum_{j\in F\cup F^*} X_{G_j}(\widehat{\beta}_{G_j} - \beta_{G_j})\rangle| \\
&\geq (1 - \delta_{1.25s}) \sum_{j\in F\cup F^*} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2^2 \\
&- \theta_{s,1.25s} K(\widehat{\beta}, F, F^*) \Big(\sum_{i\geq 1} \sqrt{\sum_{j\in F_i} \|\widehat{\beta}_{G_j}\|_2^2} \Big) \\
&\geq (1 - \delta_{1.25s} - \theta_{s,1.25s}) \sum_{j\in F\cup F^*} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2^2,
\end{aligned}$$

where in the last inequality we apply Lemma 4. Combining this lower bound with the upper bound in (16), we have

$$\begin{aligned}
&\sqrt{\sum_{j\in F\cup F^*} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2^2} \\
&\leq c_1 \left(\sqrt{sk_0} + \sqrt{2}\sqrt{s\left(\log d + \log n\right)} \right).
\end{aligned}$$

where

$$c_1 = \sigma\sqrt{\frac{5}{(1 - \delta_{1.25s} - \theta_{s,1.25s})^2 \rho_{k_0}}}. \quad (17)$$

The result of Theorem 1 then follows by applying Lemma 4 again. □

We can identify the benefit of our group Dantzig selector by comparing the result of Theorem 2 to the ℓ_2-norm of the Lasso and the Dantzig selector estimation error given in Cai et al. (2009). In particular, there are two advantages of utilizing the group structure. First, our condition in equation (10) is much more relaxed since δ_m and $\theta_{m,m'}$ defined in definition 2 and 3 are smaller than the non-group version because they only need to be satisfied for group compatible vectors γ and γ'. Second, the upper bound we obtained for the group Dantzig selector is $O(\sqrt{sk_0} + \sqrt{s(\log d + \log n)})$. Recall that for the Lasso or the Dantzig selector, such a bound is $O(\sqrt{sk_0(\log d + \log k_0)})$. If there exists some $c > 0$, $k_0 \geq c \log d$, the bound for the group Dantzig selector becomes $O(\sqrt{sk_0})$. In contrast, the bound for the Lasso becomes $O(\sqrt{sk_0 \log d})$. When $d \gg n$, this $\log d$ factor difference can be striking in applications. This observation is also confirmed by our numerical experiments.

5 ℓ_1-THEORY

In this section, we study the theoretical property of the group Dantzig selector using block ℓ_1-norm, which is complementary to the ℓ_2-analysis in Section 4. For sparse learning problems, sometimes ℓ_1-norm can be a more suitable error metric than the usual ℓ_2-norm. For example, we assume that the true signal $\beta \in \mathbf{R}^p$ is a zero vector and consider two estimators β_1 and β_2 given by $\beta_1 = (1, 0, \ldots, 0)$ and $\beta_2 = (1/\sqrt{p}, \ldots, 1/\sqrt{p})$. We see $\|\beta_1 - \beta\|_2 = \|\beta_2 - \beta\|_2 = 1$, in contrast, $\|\beta_1 - \beta\|_1 = 1 \ll \sqrt{p} = \|\beta_2 - \beta\|_1$. In this case, ℓ_1-norm is more consistent with the intuition since when considering the sparseness issue, β_1 is a better estimator than β_2 for the true signal β. By a similar argument, the block ℓ_1-norm could be more suitable than the ℓ_2-norm for certain sparse learning problems when group structure is available.

In the following, we introduce a remarkably simple analytical framework for block ℓ_1-norm bound based on a new condition defined by *group relevant isometry coefficient* and *group mutual orthogonality coefficient*, which is easily accessible to a broad audience in the machine learning community;

Definition 4. *Let $|F| = s$, the group relevant isometry coefficient $\bar{\delta}_F$ of X is defined as*

$$\begin{aligned} \bar{\delta}_F = \inf_{\delta \geq 0} & \ \delta \\ \text{s.t.} \quad & \sqrt{1-\delta} \sum_{j \in F} \|\gamma_{G_j}\|_2 \leq \sqrt{s} \|X_{G_F} \gamma_{G_F}\|_2 \\ \text{for all} \quad & \gamma_{G_F} \in \mathbf{R}^{|G_F|}. \end{aligned}$$

Definition 5. *Let $|F| = s$ and $1 \leq m \leq |F^c|$, the group mutual orthogonality coefficient $\bar{\theta}_{F,m}$ of X is defined as the smallest number that satisfies, for all $M \subset F^c$,*

$$\begin{aligned} & |{\gamma_{G_F}}^T X_{G_F}^T X_{G_M} {\gamma'}_{G_M}| \\ \leq \ & \frac{\bar{\theta}_{F,m}}{\sqrt{s|M|}} (\sum_{k \in M} \|\gamma'_{G_k}\|_2)(\sum_{j \in F} \|\gamma_{G_j}\|_2), \end{aligned}$$

where $\gamma_{G_F} \in \mathbf{R}^{|G_F|}$, $\gamma'_{G_M} \in \mathbf{R}^{|G_M|}$ are arbitrary and $|\mathrm{gsupp}(\gamma'_{G_M})| \leq m$, $\mathrm{gsupp}(\gamma_{G_F}) \subset F$.

Comparing the definitions of $\bar{\delta}_F$ and $\bar{\theta}_{F,m}$ with δ_k and $\theta_{k,k'}$ as in Definitions 2 and 3, a significant difference is that $\bar{\delta}_F$ and $\bar{\theta}_{F,m}$ are defined based on F, which is the unknown sparse index set for nonzero groups. To derive a sharp block ℓ_1-bound, we need a condition called *sparsity oracle condition*, which requires $\bar{\delta}_F + \bar{\theta}_{F,|F|} < 1$. The following is the main theorem.

Theorem 2. *(block ℓ_1-bound) Let $\widehat{\beta}^{gD}$ defined in* (4) *with*

$$\lambda_g = \frac{\sigma}{\sqrt{\rho_{k_0}}} \left(\sqrt{k_0} + \sqrt{2} \sqrt{\log d + \log n} \right).$$

Let $|F| = s$, under Assumption 1 and

$$\bar{\delta}_F + \bar{\theta}_{F,s} < 1,$$

we have, with probability larger than $1 - 1/n$

$$\sum_{j=1}^{d} \|\widehat{\beta}_{G_j}^{gD} - \beta_{G_j}\|_2 \leq C_2 \cdot \left(\sqrt{k_0} + \sqrt{2} \sqrt{\log d + \log n} \right),$$

where

$$C_2 = \frac{4\sigma s}{(1 - \bar{\delta}_F - \bar{\theta}_{F,s}) \sqrt{\rho_{k_0}}}.$$

Proof. Again we suppress the superscript of $\widehat{\beta}^{gD}$ and simply denote it as $\widehat{\beta}$. Since $d \gg n \geq s$, follow by a simple argument, we can assume that the smallest s elements of $\widehat{\beta}$ are all zeros, therefore, without loss of generality, we can assume $|F^c|/s$ is an integer. Consider the pivotal quantity

$$|\langle X(\widehat{\beta} - \beta), X_{G_F}(\widehat{\beta}_{G_F} - \beta_{G_F}) \rangle|.$$

Similar to the proof in Theorem 1, with probability larger than $1 - 1/n$,

$$\begin{aligned} & |\langle X(\widehat{\beta} - \beta), X_{G_F}(\widehat{\beta}_{G_F} - \beta_{G_F}) \rangle| \\ & \quad \leq \sum_{j \in F} |\langle X_{G_j}^T X(\widehat{\beta} - \beta), \widehat{\beta}_{G_j} - \beta_{G_j} \rangle| \\ & \quad \leq \max_{1 \leq j \leq d} \|X_{G_j}^T X(\widehat{\beta} - \beta)\|_2 \sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2 \\ & \quad \leq 2\lambda_g \sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2, \end{aligned} \tag{18}$$

where the last inequality follows from Lemma 2. On the other hand, by the definitions of $\bar{\delta}_F$, $\bar{\theta}_{F,s}$,

$$\begin{aligned} & |\langle X(\widehat{\beta} - \beta), X_{G_F}(\widehat{\beta}_{G_F} - \beta_{G_F}) \rangle| \\ \geq \ & \|X_{G_F}(\widehat{\beta}_{G_F} - \beta_{G_F})\|_2^2 \\ & - |\langle X_{G_{F^c}} \widehat{\beta}_{G_{F^c}}, X_{G_F}(\widehat{\beta}_{G_F} - \beta_{G_F}) \rangle| \\ \geq \ & \frac{1 - \bar{\delta}_F}{s} (\sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2)^2 \\ & - \frac{\bar{\theta}_{F,s}}{s} (\sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2)(\sum_{j \in F^c} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2) \\ \geq \ & \frac{1 - \bar{\delta}_F - \bar{\theta}_{F,s}}{s} (\sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2)^2. \end{aligned} \tag{19}$$

Where the last inequality utilizes Lemma 1. Combining (19) and (18), we obtain

$$\begin{aligned} \sum_{j=1}^{d} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2 & \leq 2 \sum_{j \in F} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2 \\ & \leq \frac{4s\lambda_g}{1 - \bar{\delta}_F - \bar{\theta}_{F,s}}, \end{aligned}$$

where we apply Lemma 1 again. By plugging in λ_g, we obtain the desired result. □

It would be interesting to compare the group sparsity oracle condition used in Theorem 2 with the group restricted isometry condition used in Theorem 1. By the Cauchy-Schwartz inequality, it's easy to see that $\bar{\delta}_F \leq \delta_{1.25s}$. To compare $\bar{\theta}_{F,s}$ with $\delta_{s,1.25s}$, it's worthy to point out that $\bar{\theta}_{F,s}$ is a much more refined quantity, since it only involves calculating the " structured mutual correlation" between the relevant variables and irrelevant variables. For example, the correlation between irrelevant variables can be arbitrary large, in this case, it's obvious that the group restricted isometry condition in Theorem 1 will be violated, however, the group sparsity oracle condition in Theorem 2 may still holds.

6 NUMERICAL RESULTS

In this section, we report numerical results on both simulated and real datasets. They provide empirical evidences on why the group Dantzig selector is superior to the Lasso when the correct group structure is available. More experiments on uneven sized groups will be reported elsewhere. In summary, our results are consistent with those obtained in (Huang and Zhang, 2009).

6.1 The benefit of group sparsity

We compare the group Dantzig selector and the Lasso in high dimensional problems. We use a similar setting as in (Zhao and Yu, 2007) by taking different (n, p, k_0, s) combinations with s denotes the number of nonzero groups. For each (n, p, k_0, s) combination, we sample 1000 times the covariance matrix Σ from a Wishart distribution $\mathsf{Wishart}(p, I_p)$ and the true parameter vector β_{G_j} for the j-th nonzero group is $(8 \cdot (0.5)^{j-1}, \ldots, 8 \cdot (0.5)^{j-1})$. For each Σ we sample a design matrix X from the multivariate normal distribution $N(\mathbf{0}, \Sigma)$. The response vector $\mathbf{y} = X\beta + \epsilon$ is then calculated using

$$\epsilon \sim N(\mathbf{0}, (\sqrt{0.6})^2 I_n).$$

The noise level σ^2 is set to 0.6 to manifest the asymptotic characterizations. For both methods, the tuning parameters η and λ_g are chosen optimally over the full solution paths to optimize the true loss $\|\widehat{\beta} - \beta\|_2$ or $\sum_{j=1}^{d} \|\widehat{\beta}_{G_j} - \beta_{G_j}\|_2$. Since β is unknown in real applications, this is not a practical model selection method. But for our purpose, the advantage of using such "oracle scores" is that the simulation results will only depend on the methods themselves, but not on the model selection procedures. The results are reported in Table 1 and Table 2. We see that the performance of the group Dantzig selector (GDS) is much better than the Lasso for both ℓ_2-norm and block ℓ_1-norm losses. The only case the Lasso is better is the no group structure case, in this case, our method becomes the Dantzig selector, but since we only use 16 tuning parameters to build the path, the performance is slightly worse.

Table 1: Comparison of the ℓ_2-performance of the Lasso and the group Dantzig selector

(n, p, k_0, s)	Lasso(ℓ_2–loss)	GDS(ℓ_2–loss))
$(100, 480, 16, 3)$	26.558 (3.2045)	4.8352 (1.4258)
$(100, 200, 1, 20)$	0.0499 (0.0100)	0.3461 (0.0705)
$(100, 480, 12, 5)$	19.229 (3.1682)	12.820 (2.9661)
$(100, 480, 40, 1)$	43.493 (2.7088)	1.7806 (0.0048)
$(100, 480, 4, 12)$	1.2234 (0.6394)	1.5318 (0.2392)

Table 2: Comparison of the block ℓ_1-performance of the Lasso and the group Dantzig selector

(n, p, k_0, s)	Lasso(ℓ_1–loss)	GDS(ℓ_1–loss))
$(100, 480, 16, 3)$	54.817 (2.1804)	9.3408 (3.5994)
$(100, 200, 1, 20)$	0.2132 (0.0482)	1.0909 (0.1661)
$(100, 480, 12, 5)$	51.546 (2.2146)	5.1714 (0.9811)
$(100, 480, 40, 1)$	43.493 (2.7088)	1.7806 (0.0048)
$(100, 480, 4, 12)$	5.9667 (2.8910)	4.6063 (0.5976)

6.2 Boston Housing Data

We apply the group Dantzig selector to the corrected Boston Housing data in (Ravikumar et al., 2007). The dataset contains 506 records about housing prices in suburbs of Boston. Each record has 10 continuous features which might be useful in describing housing price, and the response variable is the median house price. We consider a sparse additive model and use exactly the same experimental protocol as in (Ravikumar et al., 2007) , but replace their method with the group Dantzig selector and the group Lasso. For this, we expand each variable using 5 polynomial basis. Thus cast the problem to be a parametric regression with equally-sized group structure. The regularization paths of the two methods are shown in figure 1. Though these two paths look similar, there are subtle difference. The top 6 variables selected by the group Lasso are: lstat, rm, dis, nox, crim and ptratio; while for

the group Dantzig selector, the order of nox and ptratio are exchanged. Interestingly, nox is the one treated as a borderline variable in (Ravikumar et al., 2007).

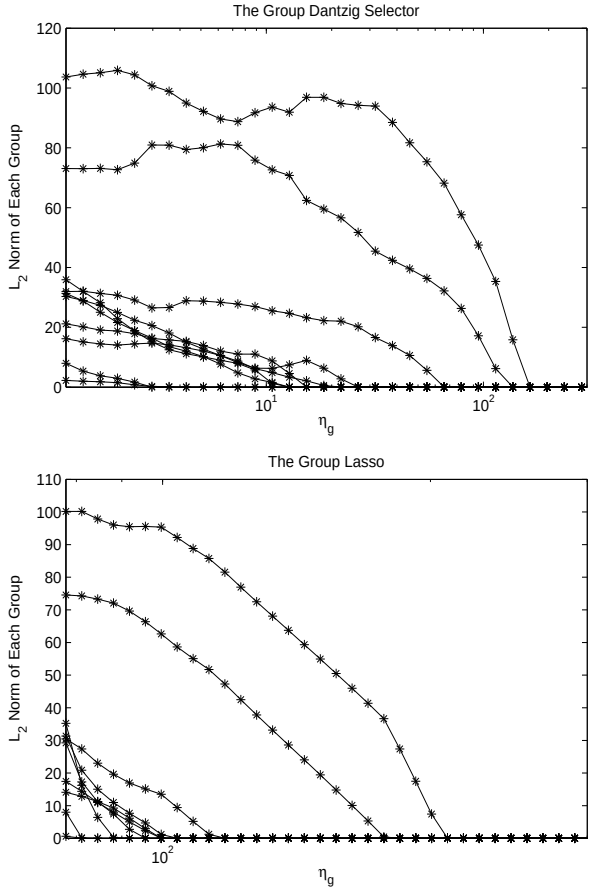

Figure 1: The regularization paths for (Upper) the group Dantzig selector and (Lower) the Group Lasso.

7 Conclusions

We present a novel method called the group Dantzig selector for sparse learning problems with group structures. The method has a convex formulation and good theoretical properties. Under the classical ℓ_2-framework, we provide a convincing theory to show that the group Dantzig selector is superior to the Lasso or the Dantzig selector for learning problems with grouped variables. To gain more insights and to make the results more accessible to a wide audience in the machine learning community, we also provide a new and remarkably simple ℓ_1-framework. We believe that the group Dantizg selector can be another useful tool for high dimensional sparse learning with group structure.

References

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 8: 1179–1225, 2008.

T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *Technical report, University of Pennsylvania*, 2009.

E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

E. J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51 (12):4203–4215, 2005.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific and Statistical Computing*, 20:33–61, 1998.

C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables lasso. *Mathematical Methods of Statistics*, 27(4):317–326, 2008.

J. Huang and T. Zhang. The benefit of group sparsity,. *Technical report, Rutgers University*, 2009.

G. M. James, P. Radchenko, and J. Lv. Dasso: connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society Series B*, 71(1):127–142, 2009.

H. Liu and J. Zhang. On the estimation consistency of the group lasso and its applications. *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *Proceedings of the Annual Conference on Learning Theory (COLT 2009)*, 2009.

L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B, Methodological*, 70:53–71, 2007.

N. Meinshausen, G. Rocha, and B. Yu. Discussion: A tale of three cousins: Lasso, l2boosting and dantzig. *The Annals of Statistics*, 35:2373–2384, 2007.

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

G. Obozinski and M. Wainwright. Union support recovery in high-dimensional multivariate regression. In *Advances in Neural Information Processing Systems 21*, 2008.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In *Advances in Neural Information Processing Systems 20*, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological*, 58:267–288, 1996.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B, Methodological*, 68:49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. of Mach. Learn. Res.*, 7:2541–2567, 2007.