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 \(\ell_2\)-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 \(\ell_1\)-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 \(\ell_1\)-norm regularization with a sum of \(\ell_2\)-norm (or, block \(\ell_1\)-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 \(\ell_2\)-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 \(\ell_2\)-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 available. Some related work also appear in the multi-task learning literature, 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-
The Group Dantzig Selector

In this paper, we introduce 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 ℓ1-theory. Section 5 introduces necessary notations and problem formulations. Section 3 presents in detail the group Dantzig selector. Section 4 studies the ℓ1-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 \( X = (x_1, \ldots, x_p) \) where \( x_j \in \mathbb{R}^n \) for all \( j \in \{1, \ldots, p\} \). The response vector \( y = (y_1, \ldots, y_n) \in \mathbb{R}^n \) is assumed to be generated from the linear model \( y = X\beta + \epsilon \) with \( \epsilon \sim N(0, \sigma^2 I) \). Although we assume \( \epsilon \) 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 \( \beta \) to be \( \text{supp}(\beta) = \{ j : \beta_j \neq 0, j = 1, \ldots, p \} \).

Given \( M \subset \{1, \ldots, p\} \), we denote \( \beta_M \) to be the sub-vector of \( \beta \) with elements indexed by \( M \). Similarly, \( X_M \) is the \( n \times |M| \) submatrix of \( X \) with columns indexed by elements in \( M \). Given a vector \( v \in \mathbb{R}^n \), let \( 1 < q < \infty \), we denote \( \|v\|_q = (v_1^q + \ldots + v_n^q)^{1/q} \), and we use \( \|v\|_\infty = \max_{1 \leq j \leq n} |v_j| \) and \( \|v\|_0 = |\text{supp}(v)| \). Without loss of generality, we assume \( \|x_j\|_2 = 1 \) 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:

\[
\hat{\beta}^L = \arg \min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad \|y - X\beta\|_2 \leq \eta.
\]

\[
\hat{\beta}^D = \arg \min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad \|X^T(y - X\beta)\|_\infty \leq \lambda.
\]

where \( \eta \) and \( \lambda \) are positive regularization parameters.

There are cases where we know a group structure on \( \beta \) 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 \( \{1, \ldots, p\} = \bigcup_{j=1}^d G_j \) 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 \( G_M = \bigcup_{j \in M} G_j \). We also denote \( \text{gsupp}(\beta) = \{ j : \|\beta_{G_j}\|_2 \neq 0, j = 1, \ldots , d \} \).

In this paper, we always denote \( \text{gsupp}(\beta) = F \). The Group Lasso can be applied to solve the group sparse learning problem:

\[
\text{(group Lasso)} \quad \hat{\beta}^L = \arg \min_{\beta} \sum_{j=1}^d \|\beta_{G_j}\|_2 \quad \text{s.t.} \quad \|y - X\beta\|_2 \leq \eta_g,
\]

where \( \eta_g \) 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: \( |G_j| = k_0 \). 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 \( \hat{\beta}^D \) is a solution to the following convex optimization problem

\[
\hat{\beta}^D = \arg \min_{\beta} \sum_{j=1}^d \|\beta_{G_j}\|_2 \quad \text{s.t.} \quad \max_{1 \leq j \leq d} \|X_{G_j}^T(y - X\beta)\|_2 \leq \lambda_g,
\]

where \( \lambda_g \geq 0 \) is a regularizer parameter.

The problem in (4) is convex, thus can be easily solved by any off-the-shelf convex programming solver. 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 \( \max_{1 \leq j \leq d} \|X_{G_j}^T(y - X\beta)\|_2 \leq \lambda_g \) as the constraint term instead of \( \|X^T(y - X\beta)\|_\infty \), since the
it’s obvious that

\[ \text{chain} \]

\[ \text{ment. Define} \]

\[ \text{This follows by a standard variational dual-} \]

\[ \text{arity argument is also true for the group Dantzig selector.} \]

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

\[
\hat{\beta}^{gL} = \arg \min_{\beta} \|X\beta\|_2 \quad \text{s.t.} \quad \max_{1 \leq j \leq d} \|X^T_{G_j} (y - X\beta)\|_2 \leq \lambda_g.
\]

**Proof.** From the Lagrangian duality, for each \( \eta_g \) there exists a nonnegative \( \lambda_g \) such that each \( \hat{\beta}^{gL} \) solves (3) also solves the unconstrained optimization problem:

\[
\min_{\beta} F(\beta)
\]

where

\[
F(\beta) = \frac{1}{2} \|y - X\beta\|_2^2 + \lambda_g \sum_{j=1}^d \|\beta_{G_j}\|_2.
\]

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

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

it’s obvious that

\[
F(\beta) = \max_{\forall \beta, \|v\|_{G_j} \leq 1} H(\beta, v).
\]

By strong duality, the desired result follows from the chain

\[
\min_{\beta} F(\beta) = \min_{\beta} \sup_{\forall \beta, \|v\|_{G_j} \leq 1} H(\beta, v) = \sup_{\forall \beta, \|v\|_{G_j} \leq 1} \min_{\beta} H(\beta, v) = \sup_{\forall \beta, \|v\|_{G_j} \leq 1} \min_{\beta} \frac{1}{2} \|X\beta\|_2^2,
\]

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 \( \hat{\beta}^L \) is a solution to the Lasso problem in (1) with the regularization parameter \( \eta \), if and only if there exists a nonnegative number \( \lambda \), such that

\[ \hat{\beta}^L = \arg \min_{\beta} \|X\beta\|_2 \quad \text{s.t.} \quad \|X^T (y - X\beta)\|_\infty \leq \lambda. \quad (7) \]

Comparing (7) with (2), the Dantzig selector simply modifies an equivalent form of the Lasso by replacing \( \|\beta\|_1 \). 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 \( \ell_1 \)-norm. Therefore the definition in (4) naturally comes out.

Under certain conditions, the solution \( \hat{\beta}^{D} \) to the dantzig selector in (2) can be shown to be identical to the Lasso solution \( \hat{\beta}^{L} \) defined in (7) using the same regularization parameter \( \lambda \) (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 orthogonal 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 \( \hat{\beta}^{gD} \) be the unique group Dantzig selector solution in (4) and \( \hat{\beta}^{gL} \) be the unique group Lasso solution in (5), both using the same regularization parameter \( \lambda_g \). If for all \( j, k \in \{1, \ldots, d\}, j \neq k \),

\[
X^T_{G_j} X_{G_j} = I_{k_0} \quad \text{and} \quad X^T_{G_j} X_{G_k} = 0,
\]

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

4 \( \ell_2 \)-theory

In this section, we study the theoretical property of the group Dantzig selector using the \( \ell_2 \)-norm criteria \( \|\hat{\beta}^{gD} - \beta\|_2 \). 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 \( 1 \leq m \leq d \), the m-group restricted isome-
try coefficient $\delta_m$ for $X$ is defined as
\[
\delta_m = \inf_{\delta > 0} \delta \quad \text{s.t.} \quad \sqrt{1 - \delta} \|\gamma\|_2 \leq \|X\gamma\|_2 \leq \sqrt{1 + \delta} \|\gamma\|_2
\]
for all $\gamma \in \mathbb{R}^p$, $|\text{gsupp}(\gamma)| \leq m$.

**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
\[
|\gamma^T X^T X \gamma'| \leq \theta_{m,m'} \|\gamma\|_2 \|\gamma'|_2
\]
where $\gamma, \gamma' \in \mathbb{R}^p$ 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 $\delta_n$ 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 $\delta$ and $\theta$ to be small. It is easy to see that when group structure is available, both $\delta_n$ and $\theta_{m,m'}$ will be smaller than their corresponding non-group versions. In other words, this implies 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 $\|\hat{\beta}^D - \beta\|_2$. 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 guarantees the estimation error is bounded from above by $\sqrt{k_0}$.

**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 $1/\sqrt{k_0}$.

**Theorem 1.** ($l_2$-bound) Let $\hat{\beta}^D$ be a solution to (4) with
\[
\lambda_g = \frac{\sigma}{\sqrt{k_0}} \left( \sqrt{k_0} + \sqrt{2} \sqrt{\log d + \log n} \right).
\]
Let $|F| = s$, without loss of generality, we assume $s/n$ is an integer. Under Assumption 1 and
\[
\delta_{1.25s} + \theta_{s,1.25s} < 1,
\]
then we have with probability larger than $1 - 1/n$
\[
\|\hat{\beta}^D - \beta\|_2 \leq C \cdot (\sqrt{s k_0} + \sqrt{2 s (\log d + \log n)}),
\]
where
\[
C = \frac{\sqrt{20 \sigma}}{(1 - \delta_{1.25s} - \theta_{s,1.25s}) \sqrt{k_0}}.
\]

**Proof.** In the following, we suppress the superscript of $\hat{\beta}^D$ and simply denote it as $\hat{\beta}$. 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 $\hat{\beta}$ be a solution to (4), then we have
\[
\sum_{j \in F^c} \|\hat{\beta}_{G_j}\|_2 \leq \sum_{j \in F} \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2.
\]

**Proof.** Since $\hat{\beta}$ is the minimizer of the objective function in (4) and $\beta_{F^c} = 0$, we have
\[
\sum_{j \in F} \|\hat{\beta}_{G_j}\|_2 \geq \sum_{j \in F} \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2
\]
\[
\geq \sum_{j \in F} \|\beta_{G_j} - \beta_{G_j}\|_2 + \sum_{j \in F^c} \|\hat{\beta}_{G_j}\|_2.
\]

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
\[
\max_{j \in M} \|X_{G_j}^T X (\hat{\beta} - \beta)\|_2
\]
\[
\leq \frac{2 \sigma}{\sqrt{k_0}} \left( \sqrt{k_0} + \sqrt{2} \sqrt{\log(|M|) + \log n} \right).
\]

**Proof.** For all $j \in \{1, \ldots, d\}$ we have
\[
\|X_{G_j}^T X (\hat{\beta} - \beta)\|_2 \leq \|X_{G_j}^T (X\hat{\beta} - y)\|_2 + \|X_{G_j}^T (y - X\beta)\|_2.
\]

Since $\|X_{G_j}^T (X\hat{\beta} - y)\|_2 \leq \lambda_g$ and by Assumption 1, we have
\[
\forall j \in \{1, \ldots, d\}, \quad \|X_{G_j}^T (y - X\beta)\|_2
\]
\[
\leq \frac{1}{\sqrt{k_0}} \left( \left\|X_{G_j}^T X_{G_j}\right\|^{-0.5} \left\|X_{G_j}^T (y - X\beta)\right\|_2 \right).
\]

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
\[
\|\hat{\beta}_{G_{s+1}} - \beta_{G_{s+1}}\|_2
\]
\[
\geq \|\hat{\beta}_{G_{s+2}} - \beta_{G_{s+2}}\|_2 \geq \cdots \geq \|\hat{\beta}_{G_{d}} - \beta_{G_{d}}\|_2.
\]

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 $\ell_2$-norm of the last 4/5 proportion can be bounded by the $\ell_1$-norm of the first 4/5 proportion over a large constant factor.

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

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

(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

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

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

$$
\sqrt{s} \sum_{i \geq 1} \sqrt{\sum_{j \in F_i} \|\hat{\beta}_{j} - \beta_{j}\|_2^2} \\
\leq \sum_{j \in c_{i} \geq 1 \in F} \|\hat{\beta}_{j} - \beta_{j}\|_2 + \sum_{j \in F^*} \|\hat{\beta}_{j} - \beta_{j}\|_2 \\
= \sum_{j \in F^*} \|\hat{\beta}_{j} - \beta_{j}\|_2.
$$

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

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

$$
K(\bar{\beta}, F, F^*) = \sqrt{\sum_{j \in F \cup F^*} \|\hat{\beta}_{j} - \beta_{j}\|_2^2}.
$$

(15)

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

$$
|(X(\bar{\beta} - \beta), \sum_{j \in F \cup F^*} X_{G_j}(\hat{\beta}_{G_j} - \beta_{G_j}))| \\
\leq \sum_{j \in F \cup F^*} |X_{G_j}^T X(\bar{\beta} - \beta), \hat{\beta}_{G_j} - \beta_{G_j})| \\
\leq \sqrt{1.25s} \max_{1 \leq j \leq d} \|X_{G_j}^T X(\bar{\beta} - \beta)\|_2 K(\bar{\beta}, F, F^*) \\
\leq \sigma \sqrt{\frac{5s}{\rho_{k_0}}} (\sqrt{k_0} + \sqrt{2\sqrt{\log d + \log n}}) K(\bar{\beta}, F, F^*).
$$

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

$$
|(X(\bar{\beta} - \beta), \sum_{j \in F \cup F^*} X_{G_j}(\hat{\beta}_{G_j} - \beta_{G_j}))| \\
\geq \| \sum_{j \in F \cup F^*} X_{G_j}(\hat{\beta}_{G_j} - \beta_{G_j}) \|_2^2 - \sum_{i \geq 1} \| \sum_{j \in F} X_{G_j}(\hat{\beta}_{G_j} - \beta_{G_j}) \|_2 \| X_{G_j}(\hat{\beta}_{G_j} - \beta_{G_j}) \|_2 \\
\geq (1 - \delta_{1.25s}) \| \sum_{j \in F \cup F^*} X_{G_j}(\hat{\beta}_{G_j} - \beta_{G_j}) \|_2^2 - \theta_{s,1.25s} K(\bar{\beta}, F, F^*) \left( \sum_{i \geq 1} \| \sum_{j \in F} \| \hat{\beta}_{G_j} \|_2^2 \right) \\
\geq (1 - \delta_{1.25s} - \theta_{s,1.25s}) \| \sum_{j \in F \cup F^*} X_{G_j}(\hat{\beta}_{G_j} - \beta_{G_j}) \|_2^2,
$$

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

$$
\sqrt{\sum_{j \in F \cup F^*} \|\hat{\beta}_{j} - \beta_{j}\|_2^2} \\
\leq c_1 \left( \sqrt{5s} \sigma \sqrt{k_0} + \sqrt{2\sqrt{\log d + \log n}} \right).
$$

where

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

(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 $\ell_2$-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 $\delta_{m}$ and $\delta_{m', m''}$ defined in definition 2 and 3 are smaller than the non-group version because they only need to be satisfied for group compatible vectors $\gamma$ and $\gamma'$. Second, the upper bound we obtained for the group Dantzig selector is $O(\sqrt{s}k_0 + \sqrt{s(\log d + \log n)})$. Recall that for the Lasso or the Dantzig selector, such a bound is $O(\sqrt{s}k_0 + \sqrt{s(\log d + \log n)})$. If there exists some $c > 0$, $k_0 \geq c \log d$, the bound for the group Dantzig selector becomes $O(\sqrt{s}k_0)$. In contrast, the bound for the Lasso becomes $O(\sqrt{s}k_0 \log d)$. When $d \gg n$, this log $d$ factor difference can be striking in applications. This observation is also confirmed by our numerical experiments.
5 $\ell_1$-THEORY

In this section, we study the theoretical property of the group Dantzig selector using block $\ell_1$-norm, which is complementary to the $\ell_2$-analysis in Section 4. For sparse learning problems, sometimes $\ell_1$-norm can be a more suitable error metric than the usual $\ell_2$-norm. For example, we assume that the true signal $\beta \in \mathbb{R}^p$ is a zero vector and consider two estimators $\beta_1$ and $\beta_2$ 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, $\ell_1$-norm is more consistent with the intuition since when considering the sparseness issue, $\beta_1$ is a better estimator than $\beta_2$ for the true signal $\beta$. By a similar argument, the block $\ell_1$-norm could be more suitable than the $\ell_2$-norm for certain sparse learning problems when group structure is available.

In the following, we introduce a remarkably simple analytical framework for block $\ell_1$-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 $\delta_F$ of $X$ is defined as
\[
\delta_F = \inf_{\delta \geq 0} \sqrt{1 - \delta} \sum_{j \in F} \|\gamma_{G_j}\|_2 \leq \sqrt{s} \|X \gamma_{G_F}\|_2
\]
for all $\gamma_{G_F} \in \mathbb{R}^{[G_F]}$.

**Definition 5.** Let $|F| = s$ and $1 \leq m \leq |F^c|$, the group mutual orthogonality coefficient $\theta_{F,m}$ of $X$ is defined as the smallest number that satisfies, for all $M \subset F^c$,
\[
|\gamma_{G_F}^T X_{F,m}^\gamma_{G_M}^\gamma'_{G_M}| \leq \frac{\theta_{F,m}}{\sqrt{|s| |M|}} \left( \sum_{k \in M} \|\gamma_k\|_2 \right) \left( \sum_{j \in F} \|\gamma_j\|_2 \right),
\]
where $\gamma_{G_F} \in \mathbb{R}^{[G_F]}$, $\gamma'_{G_M} \in \mathbb{R}^{[G_M]}$ are arbitrary and $|\text{gsupp}(\gamma_{G_M}^\gamma)| \leq m$, $\text{gsupp}(\gamma_{G_F}^\gamma) \subset F$.

Comparing the definitions of $\delta_F$ and $\theta_{F,m}$ with $\delta_k$ and $\theta_{k,k'}$ as in Definitions 2 and 3, a significant difference is that $\delta_F$ and $\theta_{F,m}$ are defined based on $F$, which is the unknown sparse index set for nonzero groups. To derive a sharp block $\ell_1$-bound, we need a condition called sparsity oracle condition, which requires $\delta_F + \theta_{F,|F|} < 1$. The following is the main theorem.

**Theorem 2.** (block $\ell_1$-bound) Let $\hat{\beta}^{\text{GD}}$ defined in (4) with
\[
\lambda_g = \frac{\sigma}{\sqrt{p_k}}, \left( \sqrt{k_0} + \sqrt{2 \log d + \log n} \right).
\]
Let $|F| = s$, under Assumption 1 and
\[
\delta_F + \theta_{F,s} < 1,
\]
we have, with probability larger than $1 - 1/n$
\[
\sum_{j=1}^d \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2 \leq C_2 \left( \sqrt{k_0} + \sqrt{2 \log d + \log n} \right),
\]
where
\[
C_2 = \frac{4\sigma s}{(1 - \delta_F - \theta_{F,s}) \sqrt{p_k}}.
\]

**Proof.** Again we suppress the superscript of $\hat{\beta}^{\text{GD}}$ and simply denote it as $\hat{\beta}$. Since $d \gg n \geq s$, follow by a simple argument, we can assume that the smallest $s$ elements of $\hat{\beta}$ are all zeros, therefore, without loss of generality, we can assume $|F^c|/s$ is an integer. Consider the pivotal quantity
\[
\|X(\hat{\beta} - \beta), X_{G_F}(\hat{\beta}_{G_F} - \beta_{G_F})\|.
\]
Similar to the proof in Theorem 1, with probability larger than $1 - 1/n$
\[
\|X(\hat{\beta} - \beta), X_{G_F}(\hat{\beta}_{G_F} - \beta_{G_F})\| \leq \sum_{j \in F} \|X_{G_j}^T X(\hat{\beta} - \beta), \hat{\beta}_{G_j} - \beta_{G_j}\| \leq \|X_{G_j}^T X(\hat{\beta} - \beta), \hat{\beta}_{G_j} - \beta_{G_j}\| \\
\leq 2\lambda_g \sum_{j \in F} \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2,
\]
where the last inequality follows from Lemma 2. On the other hand, by the definitions of $\delta_F, \theta_{F,s}$,
\[
\|X(\hat{\beta} - \beta), X_{G_F}(\hat{\beta}_{G_F} - \beta_{G_F})\| \geq \|X_{G_F}(\hat{\beta}_{G_F} - \beta_{G_F})\|^2 \\
- \|X_{G_F}(\hat{\beta}_{G_F} - \beta_{G_F}), X_{G_F}(\hat{\beta}_{G_F} - \beta_{G_F})\|^2 \\
\geq 1 - \frac{\delta_F}{s} \sum_{j \in F} \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2^2 \\
- \frac{\theta_{F,s}}{s} \sum_{j \in F} \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2^2 \\
\geq 1 - \frac{\delta_F - \theta_{F,s}}{s} \sum_{j \in F} \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2^2.
\]
Where the last inequality utilizes Lemma 1. Combining (19) and (18), we obtain
\[
\sum_{j=1}^d \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2 \leq 2 \sum_{j \in F} \|\hat{\beta}_{G_j} - \beta_{G_j}\|_2 \\
\leq \frac{4\sigma s \lambda_g}{1 - \delta_F - \theta_{F,s}}.
\]
where we apply Lemma 1 again. By plugging in \( \lambda_g \), 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 \( \delta_F \leq \delta_{1.25^2} \). To compare \( \vartheta_{\epsilon,s} \) with \( \delta_{s,1.25^2} \), it’s worthy to point out that \( \vartheta_{\epsilon,s} \) 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 \(\Sigma\) from a Wishart distribution \(\text{Wishart}(p, I_p)\) and the true parameter vector \(\beta_{G_j}\) for the \(j\)-th nonzero group is \((8 \cdot (0.5)^{j-1}, \ldots, 8 \cdot (0.5)^{j-1})\). For each \(\Sigma\) we sample a design matrix \(X\) from the multivariate normal distribution \(N(0, \Sigma)\). The response vector \(y = X \beta + \epsilon\) is then calculated using

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

The noise level \(\sigma^2\) is set to 0.6 to manifest the asymptotic characterizations. For both methods, the tuning parameters \(\eta\) and \(\lambda_g\) are chosen optimally over the full solution paths to optimize the true loss \(\|\hat{\beta} - \beta\|_2\) or \(\sum_{j=1}^d \|\beta_{G_j} - \beta_{G_j}\|_2\). Since \(\beta\) 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 \(\ell_2\)-norm and block \(\ell_1\)-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 \(\ell_2\)-performance of the Lasso and the group Dantzig selector

<table>
<thead>
<tr>
<th>((n, p, k_0, s))</th>
<th>Lasso(\ell_2)-loss</th>
<th>GDS(\ell_2)-loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>(100, 480, 16, 3)</td>
<td>26.558 (3.2045)</td>
<td>4.8352 (1.4258)</td>
</tr>
<tr>
<td>(100, 200, 1, 20)</td>
<td>0.0499 (0.0100)</td>
<td>0.3461 (0.0705)</td>
</tr>
<tr>
<td>(100, 480, 12, 5)</td>
<td>19.229 (3.1682)</td>
<td>12.820 (2.9661)</td>
</tr>
<tr>
<td>(100, 480, 40, 1)</td>
<td>43.493 (2.7088)</td>
<td>1.7806 (0.0048)</td>
</tr>
<tr>
<td>(100, 480, 4, 12)</td>
<td>1.2234 (0.6394)</td>
<td>1.5318 (0.2392)</td>
</tr>
</tbody>
</table>

### Table 2: Comparison of the block \(\ell_1\)-performance of the Lasso and the group Dantzig selector

<table>
<thead>
<tr>
<th>((n, p, k_0, s))</th>
<th>Lasso(\ell_1)-loss</th>
<th>GDS(\ell_1)-loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>(100, 480, 16, 3)</td>
<td>54.817 (2.1804)</td>
<td>9.3408 (3.5994)</td>
</tr>
<tr>
<td>(100, 200, 1, 20)</td>
<td>0.2132 (0.0482)</td>
<td>1.0909 (0.1661)</td>
</tr>
<tr>
<td>(100, 480, 12, 5)</td>
<td>51.546 (2.2146)</td>
<td>5.1714 (0.9811)</td>
</tr>
<tr>
<td>(100, 480, 40, 1)</td>
<td>43.403 (2.7088)</td>
<td>1.7806 (0.0048)</td>
</tr>
<tr>
<td>(100, 480, 4, 12)</td>
<td>5.9667 (2.8910)</td>
<td>4.6063 (0.5976)</td>
</tr>
</tbody>
</table>

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: \text{Istat, rm, dis, nox, crim} and \text{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 $\ell_2$-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 $\ell_1$-framework. We believe that the group Dantzig selector can be another useful tool for high dimensional sparse learning with group structure.

References


