



HAL
open science

Generalization of l1 constraints for high dimensional regression problems

Pierre Alquier, Mohamed Hebiri

► **To cite this version:**

Pierre Alquier, Mohamed Hebiri. Generalization of l1 constraints for high dimensional regression problems. 2008. hal-00336101v3

HAL Id: hal-00336101

<https://hal.science/hal-00336101v3>

Preprint submitted on 27 May 2010 (v3), last revised 4 Jul 2011 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Generalization of ℓ_1 constraints for high dimensional regression problems

Pierre ALQUIER⁽¹⁾ and Mohamed HEBIRI⁽²⁾

(1, 2) LPMA, CNRS-UMR 7599,
Université Paris 7 - Diderot, UFR de Mathématiques,
175 rue de Chevaleret F-75013 Paris, France.

(1) CREST-LS,
3, avenue Pierre Larousse
92240 Malakoff, France.

(2) ETH-Zürich

Abstract

We focus on the high dimensional linear regression $Y \sim \mathcal{N}(X\beta^*, \sigma^2 I_n)$, where $\beta^* \in \mathbb{R}^p$ is the parameter of interest. In this setting, several estimators such as the LASSO [Tib96] and the Dantzig Selector [CT07] are known to satisfy interesting properties whenever the vector β^* is sparse. Interestingly both of the LASSO and the Dantzig Selector can be seen as orthogonal projections of 0 into $\mathcal{DC}(s) = \{\beta \in \mathbb{R}^p, \|X'(Y - X\beta)\|_\infty \leq s\}$ - using an ℓ_1 distance for the Dantzig Selector and ℓ_2 for the LASSO. For a well chosen $s > 0$, this set is actually a confidence region for β^* . In this paper, we investigate the properties of estimators defined as projections on $\mathcal{DC}(s)$ using general distances. We prove that the obtained estimators satisfy oracle properties close to the one of the LASSO and Dantzig Selector. On top of that, it turns out that these estimators can be tuned to exploit a different sparsity or/and slightly different estimation objectives.

Keywords: High-dimensional data, LASSO, Restricted eigenvalue assumption, Sparsity, Variable selection.

AMS 2000 subject classifications: Primary 62J05, 62J07; Secondary 62F25.

1 Introduction

In many modern applications, one has to deal with very large datasets. Regression problems may involve a large number of covariates, possibly larger than the sample size. In this situation, a major issue lies in dimension reduction which can be performed through the selection of a small amount of relevant covariates. For this purpose, numerous regression methods have been proposed in the literature, ranging from the classical information criteria such as C_p , AIC and BIC to the more recent regularization-based techniques such as the ℓ_1 penalized least square estimator, known as the LASSO [Tib96], and the Dantzig selector [CT07]. These ℓ_1 -regularized regression methods have recently witnessed several developments due to the attractive feature of computational feasibility, even for high dimensional data when the number of covariates p is large.

Consider the linear regression model

$$Y = X\beta^* + \varepsilon, \quad (1)$$

where Y is a vector in \mathbb{R}^n , $\beta^* \in \mathbb{R}^p$ is the parameter vector, X is an $n \times p$ real-valued matrix with possibly much fewer rows than columns, $n \ll p$, and ε is a random noise vector in \mathbb{R}^n . Here, for the sake of simplicity, we will assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Let \mathbb{P} denote the probability distribution of Y in this setting. Moreover, we assume that the matrix X is normalized in such a way that $X'X/n$ has only 1 on its diagonal. The analysis of regularized regression methods for high dimensional data usually involves a sparsity assumption on β^* through the *sparsity index* $\|\beta^*\|_0 = \sum_{j=1, \dots, p} \mathbb{I}(\beta_j^* \neq 0)$ where $\mathbb{I}(\cdot)$ is the indicator function. For any $q \geq 1$, $d \geq 0$ and $a \in \mathbb{R}^d$, denote by $\|a\|_q^q = \sum_{i=1}^d |a_i|^q$ and $\|a\|_\infty = \max_{1 \leq i \leq d} |a_i|$, the ℓ_q and the ℓ_∞ norms respectively. When the design matrix X is normalized, the LASSO and the Dantzig selector minimize respectively $\|X\beta\|_2^2$ and $\|\beta\|_1$ under the constraint $\|X'(Y - X\beta)\|_\infty \leq s$ where s is a positive tuning parameter (e.g. [OPT00, Alq08] for the dual form of the LASSO). This geometric constraint is central in the approach developed in the present paper and we shall use it in a general perspective. Let us mention that several objectives may be considered by the statistician when we deal with the model given by Equation (1). Usually, we consider three specific objectives in the high-dimensional setting (i.e., $p \geq n$):

Goal 1 - Prediction: The reconstruction of the signal $X\beta^*$ with the best possible accuracy is first considered. The quality of the reconstruction with an estimator $\hat{\beta}$ is often measured with the squared error $\|X\hat{\beta} - X\beta^*\|_2^2$. In the standard form, results are stated as follows: under assumptions on the matrix X and with high probability, the prediction error is bounded by $C \log(p) \|\beta^*\|_0$ where C is a positive constant. Such results for the prediction issue have been obtained in

[BRT09, Bun08, BTW07b] for the LASSO and in [BRT09] for the Dantzig selector. We also refer to [Kol09a, Kol09b, MVdGB09, vdG08, DT07, CH08] for related works with different estimators (non-quadratic loss, penalties slightly different from ℓ_1 and/or random design). The results obtained in the works above-mentioned are optimal up to a logarithmic factor as it has been proved in [BTW07a]. See also [vdGB09] for a very nice survey paper on the various conditions used to prove these results.

Goal 2 - Estimation: Another wishful thinking is that the estimator $\hat{\beta}$ is close to β^* in terms of the ℓ_q distance for $q \geq 1$. The estimation bound is of the form $C \|\beta^*\|_0 (\log(p)/n)^{q/2}$ where C is a positive constant. Such results are stated for the LASSO in [BTW07a, BTW07b] when $q = 1$, for the Dantzig selector in [CT07] when $q = 2$ and have been generalized in [BRT09] with $1 \leq q \leq 2$ for both the LASSO and the Dantzig selector.

Goal 3 - Selection: Since we consider variable selection methods, the identification of the true support $\{j : \beta_j^* \neq 0\}$ of the vector β^* is to be considered. One expects that the estimator $\hat{\beta}$ and the true vector β^* share the same support at least when n grows to infinity. This is known as the variable selection consistency problem and it has been considered for the LASSO and the Dantzig Selector in several works [Bun08, Lou08, MB06, MY09, Wai06, ZY06].

In this paper, we focus on variants of **Goal 1** and **Goal 2**, using estimators $\hat{\beta}$ that also satisfy the constraint $\|X'(Y - X\hat{\beta})\|_\infty \leq s$. It is organized as follows. In Section 2 we give some general geometrical considerations on the LASSO and the Dantzig Selector that motivates the introduction of the general form of estimator:

$$\underset{\beta \in \|X'(Y - X\beta)\|_\infty \leq s}{\text{Argmin}} \quad \|\beta\|$$

for any semi-norm $\|\cdot\|$. In Section 3, we focus on two particular cases of interest in this family, and give some sparsity inequalities in the spirit of the ones in [BRT09]. We show that under the hypothesis that $P\beta^*$ is sparse for a known matrix P , we are able to estimate properly β^* . Finally, Section 4 is dedicated to proofs.

2 Some geometrical considerations

Definition 2.1. *Let us put, for any $s > 0$, $\mathcal{DC}(s) = \{\beta \in \mathbb{R}^p : \|X'(Y - X\beta)\|_\infty \leq s\}$.*

Lemma 1. *For any $s > 0$, $\mathbb{P}(\beta^* \in \mathcal{DC}(s)) > 1 - p \exp(-s^2/(2n\sigma^2))$.*

This means that $\mathcal{DC}(s)$ is a confidence region for β^* . Moreover, note that $\mathcal{DC}(s)$ is convex and closed. Let $\|\cdot\|$ be any semi-norm in \mathbb{R}^p . Let $\Pi_{\|\cdot\|}^s$ denote an orthogonal projection on $\mathcal{DC}(s)$ with respect to $\|\cdot\|$:

$$\Pi_{\|\cdot\|}^s(b) \in \underset{\beta \in \mathcal{DC}(s)}{\text{Argmin}} \|\beta - b\|.$$

From properties of projections, we know that

$$\beta^* \in \mathcal{DC}(s) \Rightarrow \forall b \in \mathbb{R}^p, \|\Pi_{\|\cdot\|}^s(b) - \beta^*\| \leq \|b - \beta^*\|.$$

There is a very simple interpretation to this inequality: if b is any estimator of β^* , then, with probability at least $1 - p \exp(-s^2/(2n\sigma^2))$, $\Pi_{\|\cdot\|}^s(b)$ is a better estimator. In order to perform shrinkage it seems natural to take $b = 0$.

Definition 2.2. *We define our general estimator by*

$$\hat{\beta}_s^{\|\cdot\|} = \Pi_{\|\cdot\|}^s(0) \in \underset{\beta \in \mathcal{DC}(s)}{\text{Argmin}} \|\beta\|.$$

We have the following examples:

1. for $\|\cdot\| = \|\cdot\|_1$, we obtain the definition of the Dantzig Selector given in [CT07].
2. for $\|\beta\| = \|X\beta\|_2$, we obtain the program $\underset{\beta \in \mathcal{DC}(s)}{\text{Argmin}} \|X\beta\|_2$. It was proved in [OPT00] for example that a particular solution of this program is Tibshirani's LASSO estimator [Tib96] known as

$$\tilde{\beta}_s = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} [\|Y - X\beta\|_2^2 + 2s\|\beta\|_1].$$

3. for $\|\beta\| = \|X'X\beta\|_q$ with $q > 0$, it is proved in [Alq08] that the solution coincides with the "Correlation Selector" and it does not depend on q .

In the next Section, we exhibit other cases of interest and provide some theoretical results on the performances of the estimators.

3 Generalized LASSO and Dantzig Selector and Sparsity Inequalities

Let A be a $p \times p$ symmetric positive matrix and P be a $p \times p$ invertible matrix satisfying both $(X'X)P = A$ and $\text{Ker}A = \text{Ker}X$. The idea is that, for a well chosen $\|\cdot\|$, we will build estimators that will be useful to estimate β^* when $P\beta^*$

is sparse, which means that they will be close to β^* in the sense of the semi-norm induced by A .

Let \tilde{A}^{-1} be any pseudo-inverse of A . Let $\Omega = (X'X)\tilde{A}^{-1}(X'X)/n$ (note that Ω is uniquely defined, even if \tilde{A}^{-1} is not).

Definition 3.1. We define the "Generalized Dantzig Selector", $\hat{\beta}_s^{GDS}$, as $\hat{\beta}_s^{\|\cdot\|}$ for $\|b\| = \|Pb\|_1$, and the "Generalized LASSO", $\hat{\beta}_s^{GL}$, as $\hat{\beta}_s^{\|\cdot\|}$ for $\|b\| = (b'Ab)^{1/2}$.

Remark 1. In the case where the program $\min_{\beta \in \mathcal{DC}(s)} \beta' A \beta$ has multiple solutions we define $\hat{\beta}_s^{GL}$ as one of the solutions that minimizes $\|P\beta\|_1$ among all the solutions β . The case where the program $\min_{\beta \in \mathcal{DC}(s)} \|P\beta\|_1$ has multiple solution does not cause any trouble: we can take $\hat{\beta}_s^{GDS}$ as any of these solution without any effect on its statistical properties.

We now present the assumptions we need to state the Sparsity Inequalities. Note that they essentially involve the matrix Ω , and then, the matrices X and A .

Assumption A(c) for $c > 0$: for any $\alpha \in \mathbb{R}^p$ such that

$$\sum_{j:(P\beta^*)_j=0} |\alpha_j| \leq 3 \sum_{j:(P\beta^*)_j \neq 0} |\alpha_j|,$$

we have

$$\sum_{j:(P\beta^*)_j \neq 0} \alpha_j^2 \leq c\alpha' \Omega \alpha.$$

This assumption can be seen as a modification of assumptions that can be found in [BRT09]: in [BRT09], the same assumption is made on the matrix $X'X$. So here, in the case where $A = X'X$ and $P = I_p$, we will obtain exactly the same assumption that in [BRT09].

Theorem 1. Let us take $\varepsilon \in]0, 1[$ and $s = 2\sigma(2n \log(p/\varepsilon))^{1/2}$. Assume that Assumption A(c) is satisfied for some $c > 0$. With probability at least $1 - \varepsilon$ we have simultaneously:

$$(\hat{\beta}_s^{GDS} - \beta^*)' A (\hat{\beta}_s^{GDS} - \beta^*) \leq 72\sigma^2 c \|P\beta^*\|_0 \log\left(\frac{p}{\varepsilon}\right),$$

$$\|P(\hat{\beta}_s^{GDS} - \beta^*)\|_1 \leq 18\sqrt{2}\sigma \|P\beta^*\|_0 \sqrt{\frac{c \log(p/\varepsilon)}{n}},$$

$$(\hat{\beta}_s^{GL} - \beta^*)' A (\hat{\beta}_s^{GL} - \beta^*) \leq 128\sigma^2 c \|P\beta^*\|_0 \log\left(\frac{p}{\varepsilon}\right),$$

and finally

$$\|P(\hat{\beta}_s^{GL} - \beta^*)\|_1 \leq 32\sqrt{2}\sigma \|P\beta^*\|_0 \sqrt{\frac{c \log(p/\varepsilon)}{n}}.$$

In the case $A = X'X$ and $P = I_p$, we obtain the same result as in [BRT09]. However, it is worth noting that the use of $\hat{\beta}_s^{GL}$ is particularly useful when $P\beta^*$ is sparse and β^* is not. In this case the errors of the LASSO and the Dantzig Selector are not controlled anymore. This generalization is also of some interests especially when Assumption $A(c)$ is satisfied for Ω , but not satisfied if we replace Ω by $X'X$.

4 Proofs

4.1 Proof of Lemma 1

We have $Y \sim \mathcal{N}(X\beta^*, \sigma^2 I_n)$ and so $Y - X\beta^* \sim \mathcal{N}(0, \sigma^2 I_n)$ and finally $X'(Y - X\beta^*) \sim \mathcal{N}(0, \sigma^2 X'X)$. Let us put $V = X'(Y - X\beta^*)$ and let V_j denote the j -th coordinate of V . Note that $X'X$ is normalized such that for any j , $V_j \sim \mathcal{N}(0, \sigma^2 n)$, so: $\mathbb{P}(|V_j| > s) \leq \exp(-s^2/(2n\sigma^2))$. Then $\mathbb{P}(\|V\|_\infty > s) \leq p \exp(-s^2/(2n\sigma^2))$. \square

4.2 Proof of Theorem 1

We use arguments from [BRT09]. From now, we assume that the event $\{\beta^* \in \mathcal{DC}(s/2)\} = \{\|X'(Y - X\beta^*)\|_\infty < s/2\}$ is satisfied. According to Lemma 1, the probability of this event is at least $1 - p \exp(-s^2/(8n\sigma^2)) = 1 - \varepsilon$ as $s = 2(2n \log(p/\varepsilon))^{1/2}$.

Proof of the results on the Generalized Dantzig Selector.

We have

$$\begin{aligned} (\hat{\beta}_s^{GDS} - \beta^*)' A (\hat{\beta}_s^{GDS} - \beta^*) &= (\hat{\beta}_s^{GDS} - \beta^*)' X'X P (\hat{\beta}_s^{GDS} - \beta^*) \\ &\leq \|X'X (\hat{\beta}_s^{GDS} - \beta^*)\|_\infty \|P (\hat{\beta}_s^{GDS} - \beta^*)\|_1 \\ &\leq \left(\|X'(Y - X\beta^*)\|_\infty + \|X'(Y - X\hat{\beta}_s^{GDS})\|_\infty \right) \|P (\hat{\beta}_s^{GDS} - \beta^*)\|_1 \\ &\leq (s/2 + s) \|P (\hat{\beta}_s^{GDS} - \beta^*)\|_1 \end{aligned}$$

since $\hat{\beta}_s^{GDS} \in \mathcal{DC}(s)$, and $\{\beta^* \in \mathcal{DC}(s/2)\}$ is satisfied. By definition of $\hat{\beta}_s^{GDS}$,

$$\begin{aligned} 0 &\leq \|P\beta^*\|_1 - \|P\hat{\beta}_s^{GDS}\|_1 \\ &= \sum_{(P\beta^*)_j \neq 0} |(P\beta^*)_j| - \sum_{(P\hat{\beta}_s^{GDS})_j \neq 0} |(P\hat{\beta}_s^{GDS})_j| - \sum_{(P\hat{\beta}_s^{GDS})_j = 0} |(P\hat{\beta}_s^{GDS})_j| \\ &\leq \sum_{(P\beta^*)_j \neq 0} |(P\beta^*)_j - (P\hat{\beta}_s^{GDS})_j| - \sum_{(P\hat{\beta}_s^{GDS})_j = 0} |(P\beta^*)_j - (P\hat{\beta}_s^{GDS})_j|. \end{aligned}$$

This means that

$$\|P(\hat{\beta}_s^{GDS} - \beta^*)\|_1 \leq 2 \sum_{(P\beta^*)_j \neq 0} |(P\beta^*)_j - (P\hat{\beta}_s^{GDS})_j|.$$

We can summarize all that we have now:

$$\begin{aligned} (\hat{\beta}_s^{GDS} - \beta^*)' A(\hat{\beta}_s^{GDS} - \beta^*) &\leq \frac{3s}{2} \|P(\hat{\beta}_s^{GDS} - \beta^*)\|_1 \\ &\leq 3s \sum_{(P\beta^*)_j \neq 0} |(P\beta^*)_j - (P\hat{\beta}_s^{GDS})_j|. \end{aligned} \quad (2)$$

Let us remark that Inequality (2) implies that the vector $\alpha = P(\hat{\beta}_s^{GDS} - \beta^*)$ may be used in Assumption $A(c)$. This leads to

$$\begin{aligned} (\hat{\beta}_s^{GDS} - \beta^*)' A(\hat{\beta}_s^{GDS} - \beta^*) &\leq 3s \sum_{(P\beta^*)_j \neq 0} |(P\beta^*)_j - (P\hat{\beta}_s^{GDS})_j| \\ &\leq 3s \left(\|P\beta^*\|_0 \sum_{(P\beta^*)_j \neq 0} [(P\beta^*)_j - (P\hat{\beta}_s^{GDS})_j]^2 \right)^{\frac{1}{2}} \\ &\leq 3s \left(\|P\beta^*\|_0 c (P\hat{\beta}_s^{GDS} - P\beta^*)' \Omega (P\hat{\beta}_s^{GDS} - P\beta^*) \right)^{\frac{1}{2}} \\ &= 3s \left(\|P\beta^*\|_0 \frac{c}{n} (\hat{\beta}_s^{GDS} - \beta^*)' A(\hat{\beta}_s^{GDS} - \beta^*) \right)^{\frac{1}{2}}. \end{aligned} \quad (3)$$

As a consequence,

$$(\hat{\beta}_s^{GDS} - \beta^*)' A(\hat{\beta}_s^{GDS} - \beta^*) \leq 9s^2 \|P\beta^*\|_0 c = 72\sigma^2 c \|P\beta^*\|_0 \log\left(\frac{p}{\varepsilon}\right).$$

Plugging this result into Inequality (3) and using Inequality (2) again, we obtain:

$$\|P(\hat{\beta}_s^{GDS} - \beta^*)\|_1 \leq 18\sqrt{2}\sigma \|P\beta^*\|_0 \sqrt{\frac{c \log(p/\varepsilon)}{n}}.$$

Proof of the results on the Generalized LASSO.

First, we establish an important property of the Generalized LASSO estimator. We prove that

$$\begin{aligned} \forall \beta \in \mathbb{R}^p, \quad &\|Y - X P \hat{\beta}_s^{GL}\|_2^2 + 2s \|P \hat{\beta}_s^{GL}\|_1 + (\hat{\beta}_s^{GL})' P' (n\Omega - X'X) P \hat{\beta}_s^{GL} \\ &\leq \|Y - X P \beta\|_2^2 + 2s \|P \beta\|_1 + \beta' P' (n\Omega - X'X) P \beta. \end{aligned} \quad (4)$$

To prove Inequality (4), we write the Lagrangian of the program that defines $\hat{\beta}_s^{GL}$:
Let us write the Lagrangian of this program:

$$\mathcal{L}(\beta, \lambda, \mu) = \beta' A \beta + \lambda' [X'(X\beta - Y) - sE] + \mu' [X'(Y - X\beta) - sE],$$

where $E = (1, \dots, 1)'$, λ and μ are vectors in \mathbb{R}^p . Moreover, for any j , $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j \mu_j = 0$. Any solution $\underline{\beta} = \underline{\beta}(\lambda, \mu)$ must satisfy

$$0 = \frac{\partial \mathcal{L}}{\partial \beta}(\underline{\beta}, \lambda, \mu) = 2A\underline{\beta} + X'X(\lambda - \mu),$$

and then $A\underline{\beta} = (X'X)(\mu - \lambda)/2$. Note that $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j \mu_j = 0$ imply that there is a $\gamma_j \in \mathbb{R}$ such that $\gamma_j = (\mu_j - \lambda_j)/2$, $|\gamma_j| = (\lambda_j + \mu_j)/2$. Hence $\lambda_j = 2(\gamma_j)_-$ and $\mu_j = 2(\gamma_j)_+$, where for any a , $(a)_+ = \max(a; 0)$ and $(a)_- = \max(-a; 0)$. Let also γ denote the vector which j -th component is exactly γ_j , we obtain:

$$A\underline{\beta} = (X'X)\gamma. \quad (5)$$

Note that this also implies that:

$$\underline{\beta}' A \underline{\beta} = \underline{\beta}' (X'X)\gamma = \underline{\beta}' A \tilde{A}^{-1} (X'X)\gamma = \gamma' (X'X) \tilde{A}^{-1} (X'X)\gamma = n\gamma' \Omega \gamma.$$

Using these relations, the Lagrangian may be written:

$$\begin{aligned} \mathcal{L}(\underline{\beta}, \lambda, \mu) &= n\gamma' \Omega \gamma + 2\gamma' X'Y - 2\gamma' (X'X)\underline{\beta} - 2s \sum_{j=1}^p |\gamma_j| \\ &= 2\gamma' X'Y - n\gamma' \Omega \gamma - 2s \|\gamma\|_1 \end{aligned}$$

Note that λ and β , and so γ , should maximize this value. Hence, γ is to minimize

$$-2\gamma' X'Y + n\gamma' \Omega \gamma + 2s \|\gamma\|_1 + Y'Y$$

Now, note that

$$Y'Y - 2\gamma' X'Y = \|Y - X\gamma\|_2^2 - \gamma' (X'X)\gamma$$

and then γ also minimizes

$$\|Y - X\gamma\|_2^2 + 2s \|\gamma\|_1 + \gamma' [n\Omega - (X'X)] \gamma.$$

Let us put $b = P^{-1}\gamma$, then b is to minimize

$$\|Y - XPb\|_2^2 + 2s \|Pb\|_1 + (Pb)' [n\Omega - (X'X)] (Pb). \quad (6)$$

Now, we know that $\hat{\beta}_s^{GL}$ must satisfy the relation

$$A\hat{\beta}_s^{GL} = X'XPb = Ab,$$

where b is any minimizer of (6). But we can check that this equality implies that

$$\|Y - XPb\|_2^2 = \|Y - XP\hat{\beta}_s^{GL}\|_2^2,$$

and

$$(Pb)' [n\Omega - (X'X)] (Pb) = (P\hat{\beta}_s^{GL})' [n\Omega - (X'X)] (P\hat{\beta}_s^{GL}).$$

So, we necessarily have $\|Pb\|_1 = \|P\hat{\beta}_s^{GL}\|_1$ (otherwise we would have a contradiction with Remark 1). Then $\hat{\beta}_s^{GL}$ is also a minimizer of (6). This proves Equation (4).

The next step is to apply Equation (4) with $\beta = \beta^*$ to obtain

$$\begin{aligned} \|Y - XP\hat{\beta}_s^{GL}\|_2^2 + 2s\|P\hat{\beta}_s^{GL}\|_1 + (\hat{\beta}_s^{GL})' P'(n\Omega - X'X) P\hat{\beta}_s^{GL} \\ \leq \|Y - XP\beta^*\|_2^2 + 2s\|P\beta^*\|_1 + (P\beta^*)'(n\Omega - X'X) P\beta^*. \end{aligned}$$

For the sake of simplicity, we can define $\hat{\gamma} = P\hat{\beta}_s^{GL}$ and $\gamma^* = P\beta^*$ and we obtain

$$\begin{aligned} \|Y - X\hat{\gamma}\|_2^2 + 2s\|\hat{\gamma}\|_1 + \hat{\gamma}'(n\Omega - X'X)\gamma \\ \leq \|Y - X\gamma^*\|_2^2 + 2s\|\gamma^*\|_1 + (\gamma^*)'(n\Omega - X'X)\gamma^*. \end{aligned}$$

Computations lead to

$$\begin{aligned} \|X(\hat{\gamma} - \gamma^*)\|_2^2 + 2s\|\hat{\gamma}\|_1 + \hat{\gamma}'(n\Omega - X'X)\hat{\gamma} - 2(Y - X\gamma^*)' X\hat{\gamma} \\ + 2(\gamma^*)'(n\Omega - X'X)(\gamma^* - \gamma) \leq 2s\|\gamma^*\|_1 + (\gamma^*)'(n\Omega - X'X)\hat{\gamma} \\ - 2(Y - X\gamma^*)' X\gamma^*, \end{aligned}$$

and then

$$\begin{aligned} \|X(\hat{\gamma} - \gamma^*)\|_2^2 \\ \leq 2s(\|\gamma^*\|_1 - \|\hat{\gamma}\|_1) + 2(Y - X\gamma^*)' X(\hat{\gamma} - \gamma^*) - (\gamma^* - \hat{\gamma})'(n\Omega - X'X)(\gamma^* - \hat{\gamma}). \end{aligned}$$

As a consequence

$$\begin{aligned} (\gamma^* - \hat{\gamma})' n\Omega (\gamma^* - \hat{\gamma}) &\leq 2s(\|\gamma^*\|_1 - \|\hat{\gamma}\|_1) + 2(Y - X\gamma^*)' X(\hat{\gamma} - \gamma^*) \\ &\leq 2s \sum_{j=1}^p (|\gamma_j^*| - |\hat{\gamma}_j|) + 2\|X'(Y - X\beta^*)\|_\infty \sum_{j=1}^p |\hat{\gamma}_j - \gamma_j^*| \\ &\leq 2s \sum_{j=1}^p (|\gamma_j^*| - |\hat{\gamma}_j|) + s \sum_{j=1}^p |\hat{\gamma}_j - \gamma_j^*|. \end{aligned}$$

So we obtain

$$\begin{aligned}
(\gamma^* - \hat{\gamma})' n \Omega (\gamma^* - \hat{\gamma}) + s \sum_{j=1}^p |\hat{\gamma}_j - \gamma_j^*| &\leq 2s \sum_{j=1}^p (|\hat{\gamma}_j| - |\gamma_j^*|) + 2s \sum_{j=1}^p |\hat{\gamma}_j - \gamma_j^*| \\
&= 2s \sum_{j:\gamma_j^* \neq 0} (|\hat{\gamma}_j| - |\gamma_j^*|) + 2s \sum_{j:\gamma_j^* \neq 0} |\hat{\gamma}_j - \gamma_j^*| = 4s \sum_{j:\gamma_j^* \neq 0} |\hat{\gamma}_j - \gamma_j^*|. \quad (7)
\end{aligned}$$

In particular, Equation (7) implies that

$$\sum_{j:\gamma_j^* = 0} |\hat{\gamma}_j - \gamma_j^*| \leq 3 \sum_{j:\gamma_j^* \neq 0} |\hat{\gamma}_j - \gamma_j^*|,$$

and so $\alpha = \hat{\gamma}_j - \gamma_j^*$ may be used in Assumption $A(c)$. Then Inequality (7) becomes

$$\begin{aligned}
(\gamma^* - \hat{\gamma})' n \Omega (\gamma^* - \hat{\gamma}) &\leq 4s \sum_{j:\gamma_j^* \neq 0} |\hat{\gamma}_j - \gamma_j^*| \leq 4s \left(\|\gamma^*\|_0 \sum_{j:\gamma_j^* \neq 0} (\hat{\gamma}_j - \gamma_j^*)^2 \right)^{\frac{1}{2}} \\
&\leq 4s (\|\gamma^*\|_0 c (\gamma^* - \hat{\gamma})' \Omega (\gamma^* - \hat{\gamma}))^{\frac{1}{2}}.
\end{aligned}$$

That leads to

$$(\hat{\beta}_s^{GL} - \beta^*)' A(\hat{\beta}_s^{GL} - \beta^*) = (\gamma^* - \hat{\gamma})' n \Omega (\gamma^* - \hat{\gamma}) \leq 128 \sigma^2 c \|P\beta^*\|_0 \log \left(\frac{p}{\varepsilon} \right).$$

We plug this result into Inequality (7) again to obtain

$$\|\hat{\gamma} - \gamma^*\|_1 \leq 32 \sqrt{2} \sigma \|P\beta^*\|_0 \sqrt{\frac{c \log(p/\varepsilon)}{n}}.$$

This ends the proof. □

- [Alq08] P. Alquier. Lasso, iterative feature selection and the correlation selector: Oracle inequalities and numerical performances. *Electron. J. Stat.*, pages 1129–1152, 2008.
- [BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.
- [BTW07a] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

- [BTW07b] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electron. J. Stat.*, 1:169–194, 2007.
- [Bun08] F. Bunea. *Consistent selection via the Lasso for high dimensional approximating regression models*, volume 3. IMS Collections, 2008.
- [CH08] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables lasso. *Mathematical Methods of Statistics*, 17(4):317–326, 2008.
- [CT07] E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, 35, 2007.
- [DT07] A. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. *COLT 2007 Proceedings. Lecture Notes in Computer Science 4539 Springer*, pages 97–111, 2007.
- [Kol09a] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.
- [Kol09b] V. Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359, 2009.
- [Lou08] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.
- [MB06] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.
- [MVdGB09] L. Meier, S. Van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 2009.
- [MY09] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.
- [OPT00] M. Osborne, B. Presnell, and B. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000.
- [Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
- [vdG08] S. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

- [vdGB09] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Elect. Journ. Statist.*, 3:1360–1392, 2009.
- [Wai06] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming. Technical report n. 709, Department of Statistics, UC Berkeley, 2006.
- [ZY06] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.