Group Lasso for generalized linear models in high dimension

Mélanie Blazère, Jean-Michel Loubes, Fabrice Gamboa

Group Lasso for generalized linear models in high dimension

Mélanie Blazère, Jean-Michel Loubes and Fabrice Gamboa*

Abstract

We present a Group Lasso procedure for generalized linear models and we study the properties of this estimator applied to sparse high-dimensional generalized linear models. Under general conditions on the joint distribution of the pair observable covariates, we provide oracle inequalities promoting group sparsity of the covariables and we get convergence rates for the prediction and estimation error. We show the ability of this estimator to recover good sparse approximation of the true model. At last we extend this procedure to the case of an Elastic net penalty and we apply these results to the so-called Poisson regression case where the output is modeled as a Poisson process whose intensity relies on a linear combination of the covariables. The Group Lasso method enables to select few groups of meaningful variables among the set of inputs.

Keywords: Generalized linear model, ℓ_1-regularization, Group Lasso, sparse model, oracle inequalities.

1 Introduction

Handling high dimensional data is nowadays required for many practical applications, ranging from astronomy, economics, industrial problems to biomedicine. Being able to extract information from these large data sets has been at the heart of statistical studies over the last decades, and many papers have extensively studied this setting in a lot of fields ranging from statistical inference to machine learning. We refer for instance to references therein.

For high-dimensional data, classical methods based on a direct minimization of the empirical risk can lead to over fitting. Actually adding a complexity penalty enables to avoid it by selecting fewer coefficients. Using an ℓ_0-penalty leads to sparse solutions but the usely non-convex minimization problem turns out to be extremely difficult to handle when the number of parameters becomes large. Hence the ℓ_1 type penalty has been introduced to overcome this issue. On the one hand this penalty achieves sparsity of an estimated parameter vector and on the other hand it requires only convex optimization type calculations, which are computationally feasible even for high dimensional data. The use of a ℓ_1 type penalty, first proposed in [21] by Tibshirani, is now a well established procedure which has been studied in a large variety of models. We refer for example to [2], [5], [22], [8], [20] and [4].

Group sparsity can be promoted by imposing a ℓ_2 penalty to individual group of variables and then a ℓ_1 penalty to the resulting block norms. Yuan and Lin [27] proposed an extension of the Lasso in the case of linear regression and presented an algorithm when the model matrices in each group are orthonormal. This extension, called the Group Lasso, encourages blocks sparsity. Wei and Huang [25] studied the properties of the Group Lasso for linear regression and Lounici, Pontil, van de Geer and Tsybakov [13] stated oracle inequalities in linear Gaussian noise under group sparsity. Meir, van de Geer and Bühlmann [15] considered the Group Lasso in the case of logistic regression.

In this paper, we focus on the Group Lasso penalty to select and estimate parameters in the generalized linear model. One of the application is the Poisson regression model. More precisely, we consider the generalized linear model introduced by McCullagh and Nelder [14]. Let F be a

*M. Blazère, J-M. Loubes and F. Gamboa are with the Institute of Mathematics of Toulouse, Université Paul Sabatier, Toulouse, France. E-mail: (melanie.blazere, jean-michel.loubes, fabrice.gamboa)@math.univ-toulouse.fr.

distribution on $\mathbb{R}$ and let (X, Y) be a pair of random variables with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. The conditional law of $Y|X = x$ is modeled by a distribution from the exponential family and the canonical parameter is the linear predictor. Thus the conditional distribution of the observations given $X = x$ is $P(Y; \beta|x) = \exp(y\beta^T x - \psi(\beta^T x))$, where $\beta^T x$ satisfies $\int \exp(y\beta^T x)F(dy) < \infty$ and ψ is the normalized function. Notice that $\mathbb{E}(Y|X) \overset{\text{a.s}}{=} \psi'(\beta^T X)$ in other words $\beta^T X \overset{\text{a.s}}{=} h(\mathbb{E}(Y|X))$ where $h = \psi'^{-1}$ is the so-called link function. Some common examples of generalized linear models are the Poisson regression for count data, logistic and probit regression for binary data or multinomial regression for categorical data. The quantities of interest that we would like to estimate are the component $(\beta_j)_{1 \leqslant j \leqslant p}$ of β and for a given x we may also wish to predict the response $Y|X = x$. A natural field of applications for such models is given by genomics and Poisson regression type models. In particular thousands of variables such as expressions of genes and bacterias can be measured for each animal (mices) in a (pre-)clinical study thanks to the developpement of micro arrays (see, for example, [3] and [7]). A typical goal is to classify their health status, e.g. healthy or diseased, based on their bio-molecular profile, i.e. the thousands of bio-molecular variables measured for each individual (see for instance [9], [18] and [17]).

The paper falls into the following parts. In Section 2 we describe the model and the Group Lasso estimator for generalized linear models. In Section 3 we present the main results on coefficients estimation and prediction error and in Section 4 we consider the model in the particular case of Poisson regression. We study here the general model in the case of a mixture of ℓ_1 and ℓ_2 penalty. The Group Lasso estimator is then used on simulated data sets in Section 5 and its performances are compared to those of the Lasso estimator. The Appendix is devoted to the proof of the main theorems.

2 Sparse Variable selection for generalized linear models

2.1 The model

The Exponential family on the real line is a unified family of distributions parametrized by a one dimensional parameter widely used for practical modelling. Let F be a probability distribution on $\mathbb{R}$ not concentrated on a point and

$$\Theta := \left\{ \theta \in \mathbb{R} : \int \exp(\theta x) F(dx) < \infty \right\}.$$

Define

$$M(\theta) := \int \exp(\theta x) F(dx) \quad (\theta \in \Theta)$$

and

$$\psi(\theta) := \log(M(\theta)) \quad (\theta \in \Theta).$$

Let $P(y; \theta) = \exp(\theta y - \psi(\theta))$ with $\theta \in \Theta$. The densities of probability (related to a mesure adapted to the continuous or discrete case) $\{P(.; \theta) : \theta \in \Theta\}$ is called the exponential family. Θ is the natural parameter space and θ is called the canonical parameter. The exponential family includes most of the commonly used distributions like normal, gamma, poisson or binomial distributions [14].

We consider a pair of random variables (X, Y) such that the conditional distribution $Y|X = x$ is $P(Y; \beta^*, x) = \exp(y\beta^{*T} x - \psi(\beta^{*T} x))$, with $\beta^{*T} x \in \Theta$. Our aim is to estimate the components $(\beta_j)_{1 \leqslant j \leqslant p}$ of β in order to predict the response $Y|X = x$ conditionally on a given value of x. We assume that

- (H.1): there exists a constant $L > 0$ such that $|X_{ij}| \leqslant L$ for all i and all j a.s.
- (H.2): for all $x \in [-L, L]^p$, $\beta^{*T} x \in \overset{\circ}{\Theta}$

and we consider

$$\Lambda = \left\{ \beta \in \mathbb{R}^p : \forall x \in [-L, L]^p, \beta^T x \in \Theta \right\}.$$

Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d copies of (X, Y). We consider the case of high-dimensional regression i.e $p \gg n$. The log-likelihood for a generalized linear model is given by

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \left[Y_i \beta^T X_i - \psi(\beta^T X_i) \right].$$

We denote the generalized linear model loss function by

$$l(\beta) =: l(\beta; x, y) =: -y f_\beta(x) + \psi(f_\beta(x)), \tag{1}$$

where $f_\beta(x) := \beta^T x$. Notice that this function is convex in β (as ψ is convex). The associated risk is defined by $\mathbb{P}l(\beta) =: \mathbb{E}l(\beta; Y, X)$ and the empirical risk by $\mathbb{P}_n l(\beta)$ where

$$\mathbb{P}_n l(\beta) := \frac{1}{n} \sum_{i=1}^{n} \left[-Y_i \beta^T X_i + \psi(\beta^T X_i) \right].$$

Obviously

$$\beta^* = \underset{\beta \in \Lambda}{\operatorname{argmin}} \, \mathbb{P}l(\beta).$$

2.2 Group Lasso for generalized linear model

Assume that X is structured into G_n groups each of size d_g for $g \in \{1, ..., G_n\}$. For $i = 1, ..., n$ we set

$$X_i = (X_i^1, ..., X_i^g, ..., X_i^{G_n})$$

where

$$X_i^g = (X_{i,1}^g, ..., X_{i,d_g}^g)$$

and $\sum_{g=1}^{G_n} d_g = p$. This decomposition is often natural in biology and micro arrays data when the covariates are genes expresssion (see, for example, [19] and [26]). We allow the number of groups to increase with the sample size n, so we can consider the case where $G_n \gg n$. Define $d_{\max} := \max_{g \in \{1,...,G_n\}} d_g$ and $d_{\min} := \min_{g \in \{1,...,G_n\}} d_g$. For $\beta \in \mathbb{R}^p$ we denote by β^g the sub-vector of β whose indexes correspond to the index set of the g^{th} group of X.

Let us consider the Group Lasso estimator which achieves group sparsity and is obtained as the solution of the convex optimization problem

$$\hat{\beta}_n = \underset{\beta \in \Lambda}{\operatorname{argmin}} \left\{ \mathbb{P}_n l(\beta) + 2r_n \sum_{g=1}^{G_n} s(d_g) \|\beta^g\|_2 \right\} \tag{2}$$

where r_n is the tuning parameter.

Here $\|.\|_2$ refers to the Euclidian norm and s is a given function. An increase in r_n leads to a diminution of the β^g to zero, this means that some blocks become simultaneously zero and groups of predictors drop out of the model. Typically we choose $s(d_g) := \sqrt{d_g}$ to penalize more heavily groups of large size. Notice that if all the groups are of size one then we recover the Lasso estimator. The Group Lasso achieves variables selection and estimation simultaneously as the Lasso does. The penalty function is the sum of the ℓ_2 norm of the groups of variables. In fact the Group Lasso estimator acts like the Lasso at the group level [16], [13], [27].

We study estimation and prediction properties of the Group Lasso in high dimensional settings when the number of groups exceeds the sample size: $G_n \gg n$. Define $H^* = \{g : \beta^{*g} \neq 0\}$ the index set of the groups for which the correponding sub vectors of β^* are non-zero and $s^* := |H^*|$. Such a set characterizes the sparsity of the model. In the following H^{*c} denotes the index set of the groups which are not in H^*. We can notice that H^* and s^* depends on n but for simplicity we do not specify this dependency. In general, it will be hopeless to estimate all unknown parameters from data except if we make the assumption that the true parameter is group sparse. We assume that β^* is partitioned into a number of groups, in correspondance with the partition of X, only few of which are relevant. In other words we require that many subgroups of β^* are equal to zero i.e. $s^* \ll G_n$. We also assume (H.3): there exists a constant $B > 0$ such that $\sum_{g=1}^{G_n} \sqrt{d_g} \|\beta^{*g}\|_2 \leqslant B$.

3 Main Results

3.1 Bounds for estimation and prediction error

Under some assumptions on the value of the parameter r_n and on the Gram matrix X^TX we are going to show the ability of this estimator to recover good sparse approximation of the true model. To obtain oracle inequality on the Group Lasso for generalized linear model we need a concentration inequality on the empirical process $\mathbb{P}_n\left(l(\beta)\right)$ for $\beta \in \Lambda$. First we break down the empirical process into a linear part and a part which depends on the normalized parameter ψ

$$\left(\mathbb{P}_n - \mathbb{P}\right)\left(l(\beta)\right) = \left(\mathbb{P}_n - \mathbb{P}\right)\left(l_l(\beta)\right) + \left(\mathbb{P}_n - \mathbb{P}\right)\left(l_\psi(\beta)\right)$$

where $l_l(\beta) := l_l(\beta, x, y) = -y\beta^T x$ and $l_\psi(\beta) := l_\psi(\beta, x) = \psi(\beta^T x)$. Define

$$\mathcal{A} = \bigcap_{g=1}^{G_n} \{L_g \leqslant r_n/2\}$$

where

$$L_g := \left\| \frac{1}{\sqrt{d_g} n} \sum_{i=1}^{n} \left(Y_i X_i^g - \mathbb{E}(Y X^g)\right) \right\|_2$$

for all $g \in \{1, ..., G_n\}$ and

$$\mathcal{B} = \left\{ \sup_{\beta : \sum_{g=1}^{G_n} \sqrt{d_g} \|\beta^g - \beta^{*g}\|_2 \leqslant M} |\nu_n(\beta, \beta^*)| \leqslant \frac{r_n}{2} \right\}$$

where

$$\nu_n(\beta, \beta^*) := \frac{\left(\mathbb{P}_n - \mathbb{P}\right)\left(l_\psi(\beta^*) - l_\psi(\beta)\right)}{\sum_{g=1}^{G_n} \sqrt{d_g} \|\beta^g - \beta^{*g}\|_2 + \varepsilon_n}$$

with $M = 8B + \varepsilon_n$ and $\varepsilon_n = \frac{1}{n}$. We assume that G_n and n are such that $\frac{\log(2G_n)}{n} \leqslant 1$. The next proposition shows that the event $\mathcal{A} \cap \mathcal{B}$ occurs with high probability when the number of groups increases for some suitable values of the tuning parameter.

Proposition 3.1. *Let*

$$r_n \geqslant AKL \left\{ C_{L,B} \vee \max_{\{|x| \leqslant L\kappa_n\} \cap \Theta} |\psi'(x)| \right\} \sqrt{\frac{2\log(2G_n)}{n}}$$

where K is a universal constant, $A > \sqrt{2}$, $\kappa_n := 17B + \frac{2}{n}$ and $C_{L,B}$ is defined in Lemma 3.2. We have

$$\mathbb{P}\left(\mathcal{A} \cap \mathcal{B}\right) \geqslant 1 - (C + 2d_{max})(2G_n)^{-A^2/2}.$$

where C is a universal constant.

The proof rests on concentration inequalities. Indeed infering a bound for the probability of the events $\mathcal{A}$ and $\mathcal{B}$ is equivalent to prove concentration inequalities for the linear and non linear part of the empirical process. A concentration inequality for the linear part is derived from Bernstein inequality once the following lemma, which provides moment bounds for Y, has been proved.

Lemma 3.2. *Let (X, Y) a pair of random variables whose conditional distribution is $P(Y; \beta^*|x) = \exp(y{\beta^*}^T x - \psi({\beta^*}^T x))$ and assume assumptions (H.1-3) are fulfilled. For all $k \in \mathbb{N}^*$ there exists a constant $C_{L,B}$ such that $\mathbb{E}(|Y|^k) \leqslant k!(C_{L,B})^k$.*

On the other hand, since ψ is lipchitzian on the compact set $\left\{\beta : \sum_{g=1}^{G_n} \sqrt{d_g} \|\beta^g - \beta^{*g}\|_2 \leqslant M\right\}$ we use concentration results for lipchitzian loss functions [11] to bound the probability of the event $\mathcal{B}$.

Then, on the event $\mathcal{A}$ we have an upper bound for the linear part of the empirical process $\left(\mathbb{P}_n - \mathbb{P}\right)\left(l_l(\beta^*) - l_l(\hat{\beta}_n)\right)$.

Proposition 3.3. *On the event* $\mathcal{A}$

$$(\mathbb{P}_n - \mathbb{P})\left(l_l(\beta^*) - l_l(\hat{\beta}_n)\right) \leqslant \frac{r_n}{2}\sum_{g=1}^{G_n}\sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2.$$

Proof. We have

$$\begin{aligned}
&(\mathbb{P}_n - \mathbb{P})\left(l_l(\beta^*) - l_l(\hat{\beta}_n)\right) \\
&= \sum_{g=1}^{G_n}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^g - \mathbb{E}(YX^g)\right](\hat{\beta}_n^g - \beta^{*g}) \\
&\leqslant \sum_{g=1}^{G_n}\sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 \left\|\frac{1}{\sqrt{d_g}n}\sum_{i=1}^{n}(Y_i X_i^g - \mathbb{E}(YX^g))\right\|_2
\end{aligned}$$

The last bound is obtained by using Cauchy-Schwarz inequality. Therefore, on the event $\mathcal{A}$, the proposition follows. □

Thus the difference between the linear part of the empirical process and its expectation is upper bounded by the tuning parameter multiplied by the norm (associated to the penalty of the Group Lasso) of the difference between the Group Lasso estimator and the target parameter β^*. We can state a similar result for the non linear part of the empirical process, the key of the proof is based on the fact that the estimator $\hat{\beta}_n$ is in a neighborhood of the target parameter β^* on the event $\mathcal{A}\bigcap\mathcal{B}$.

Lemma 3.4. *On the event* $\mathcal{A}\bigcap\mathcal{B}$ *we have* $\sum_{g=1}^{G_n}\sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 \leqslant M$, *where we recall that* $M = 8B + \varepsilon_n$ *and* $\varepsilon_n = \frac{1}{n}$.

The next proposition provides an upper bound for $(\mathbb{P}_n - \mathbb{P})\left(l_\psi(\beta^*) - l_\psi(\hat{\beta}_n)\right)$ and is directly involved by the definition of $\mathcal{B}$ and Lemma 3.4.

Proposition 3.5. *On the event* $\mathcal{A}\cap\mathcal{B}$

$$\begin{aligned}
&(\mathbb{P}_n - \mathbb{P})\left(l_\psi(\beta^*) - l_\psi(\hat{\beta}_n)\right) \\
&\leqslant \frac{r_n}{2}\left(\sum_{g=1}^{G_n}\sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 + \varepsilon_n\right).
\end{aligned}$$

Then the key condition to derive oracle inequalities rests on the correlation between the covariates i.e. on the behavior of the Gram matrix X^TX which is necessarily singular when $p > n$. Meier, van de Geer and Bühlmann [15] proved that the group Lasso is consistent in the particular case of logistic regression and gave bounds for the prediction error under the assumption that $\mathbb{E}(XX^T)$ is non singular. In this paper we give sharp bounds for estimation and prediction errors for generalized linear models using a weaker condition similar to the restricted eigenvalue condition (RE) of Bickel, Ritov and Tsybakov [2]. This condition is quite weaker than the one of Bunea, Tsybakov and Wegkamp [6]. In their article Bickel, Ritov and Tsybakov also give several sufficient conditions for RE (which are easier to check). Here we use a condition which is a group version of the Stabil Condition first introduced by Bunea [5] for logistic regression in the case of an ℓ_1 penalty. This condition is similar (within a constant ε) to the condition used by Lounici, Pontil, van de Geer and Tsybakov [13] to state oracle inequalities for linear regression. For c_0, $\varepsilon > 0$, we define the restricted set as

$$S(c_0, \varepsilon) = \left\{\delta : \sum_{g\in H^{*c}}\sqrt{d_g}\|\delta^g\|_2 \leqslant c_0\sum_{g\in H^*}\sqrt{d_g}\|\delta^g\|_2 + \varepsilon\right\}.$$

Let $\hat{\Sigma}_n$ be the $p\times p$ empirical covariance matrix i.e. $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i {X_i}^T$.

Definition : Group Stabil Condition
Let $c_0, \varepsilon > 0$ be given. $\hat{\Sigma}_n$ satisfies the Group Stabil condition $GS(c_0, \varepsilon)$ if there exists $0 < k < 1$ such that

$$\delta^T \hat{\Sigma}_n \delta \geqslant k \sum_{g \in H^*} \|\delta^g\|_2^2 - \varepsilon \quad a.s$$

for any $\delta \in S(c_0, \varepsilon)$.

The following theorem provides meaningful bounds when the true model is sparse (i.e $s^* \ll G_n$) and $\log(G_n)$ is small as compared to n. Let $\gamma^* := \sqrt{\sum_{g \in H^*} d_g}$ and we recall that $\kappa_n := 17B + \frac{2}{n}$.

Theorem 3.6. *Assume condition $GS(3, \frac{1}{2n})$ is fulfilled. Let*

$$r_n \geqslant AKL \left\{ C_{L,B} \vee \max_{\{|x| \leqslant L\kappa_n\} \cap \Theta} |\psi^{'}(x)| \right\} \sqrt{\frac{\log(2G_n)}{n}}$$

where $A > \sqrt{2}$, K is a universal constant and $C_{L,B}$ is defined in Lemma 3.2. Then, with probability at least $1 - (C + 2d_{max})(2G_n)^{-A^2/2}$ (where C is given in Proposition 3.1), we have

$$\|\hat{\beta}_n - \beta^*\|_1 \leqslant \frac{4}{c_n k \sqrt{d_{min}}} r_n \gamma^{*2} + \left(1 + \frac{1}{r_n}\right) \frac{1}{2n\sqrt{d_{min}}}.$$

and

$$\mathbb{E}\left(\hat{\beta}_n^T X - \beta^{*T} X\right)^2 \leqslant \frac{16}{c_n^2 k} r_n^2 \gamma^{*2} + \frac{2r_n + 1}{c_n 2n}.$$

where

$$c_n := \min_{\left\{|x| \leqslant L(9B + \frac{1}{n})\right\} \cap \Theta} \left\{ \psi^{''}(x) \right\}.$$

Notice that $c_n > 0$ since the measure associated to the distribution F is not concentrated on a point. These results are similar to those of Nardi and Rinaldino [16] who proved asymptotic properties of the Group Lasso estimator for linear models. We can notice that if $\gamma^* = O(1)$ then the bound on the estimation error is of the order $O(\sqrt{\frac{\log G_n}{n}})$ and then the Group Lasso estimator still remains consistent, for ℓ_2 coefficients estimation error and for ℓ_2 prediction under the Group Stabil condition, if the number of groups increases almost as fast as $O(\exp(n))$. The term $\sqrt{\log G_n}$ is the price to pay for having a large number of factors and not knowing where are the non zero factors.

3.2 Lasso for generalized linear models

When each group is of size one we recover the Lasso estimator

$$\hat{\beta}_n = \underset{\beta \in \Lambda}{\operatorname{argmin}} \{\mathbb{P}_n l(\beta) + 2r_n \|\beta\|_1\} \tag{3}$$

where $\|\beta\|_1 = \sum_{j=1}^n |\beta_j|$. Thus following step by step the proof of Theorem 3.6 we can easily deduce bounds for estimation and prediction error for the Lasso estimator in the case of generalized linear models. The Lasso is a special case of the Group Lasso where $\gamma^* = \sqrt{s^*}$, $s^* := |I^*| = \left|\left\{j : \beta_j^* \neq 0\right\}\right|$ and $G_n = p$. We still consider high-dimensional data i.e. $n \ll p$ and sparsity assumption on the target β i.e. $s^* \ll p$ and we assume (H.1-3) except that for (H.3) we consider the ℓ_1 norm i.e. $\|\beta\|_1 \leqslant B$. The condition GS in this case requires the existence of $0 < k < 1$ such that $\delta^T \hat{\Sigma}_n \delta \geqslant k \sum_{j \in I^*} \delta_j^2 - \varepsilon \quad a.s$ for any $\delta \in S(c_0, \varepsilon) = \left\{\delta \in \mathbb{R}^p : \sum_{j \in I^{*c}} |\delta_j| \leqslant c_0 \sum_{j \in I^*} |\delta_j| + \varepsilon\right\}$. We recover the condition $St(c_0, \varepsilon)$ of [5].

Theorem 3.7. *Assume condition $St(3, \frac{1}{2n})$ is fulfilled. Let*

$$r_n \geqslant AKL \left\{ C_{L,B} \vee \max_{\{|x| \leqslant L\kappa_n\} \cap \Theta} |\psi^{'}(x)| \right\} \sqrt{\frac{\log(2p)}{n}}$$

where $A > \sqrt{2}$ and $C_{L,B}$ depends only on L and B. We have, with probability at least $1-C(2p)^{-A^2/2}$ (where C is a universal constant),

$$\|\hat{\beta}_n - \beta^*\|_1 \leqslant \frac{4r_n s^*}{c_n k} + \left(1 + \frac{1}{r_n}\right)\frac{1}{2n}$$

and

$$\mathbb{E}\left(\hat{\beta}_n^T X - \beta^{*T} X\right)^2 \leqslant \frac{16}{c_n^2 k} r_n^2 s^* + \frac{2r_n + 1}{2nc_n}$$

with $c_n := \min_{\left\{|x| \leqslant L(9B + \frac{1}{n})\right\} \cap \Theta} \left\{\psi''(x)\right\}$.

Proof. The proof of Theorem 3.7 follows the same guidelines as the one for the Group Lasso. The main difference comes from concentration inequalities for the linear and log Laplace transform part of the loss function, leading to simpler bounds. □

This result extends the one of Bunea [5] for logistic regression. The bound on the estimation error is of the order $O(s^*\sqrt{\frac{\log p}{n}})$ and the bounds are meaningful if r_n is small in particular if $n \gg \log(p)$ and s^* is small.

4 Applications and extensions

4.1 Group Lasso for Poisson regression

Sparse logistic regression has been widely studied in the literature (see, for example, [5] and [15]) but not the Poisson one in a sparse model. This last model is also very useful for many practical applications. For instance it is used to model count data and contingency tables. Poisson models are a special case of generalized linear models where the conditionnal law of Y given $X = x$ has a Poisson distribution with parameter $\lambda^*(x) := \exp(\beta^{*T} x)$. Therefore the conditional mean for Poisson regression is modeled by $\mathbb{E}(Y|X = x) = \exp(\beta^{*T} x)$. Thus the normalized function ψ is the exponential function and is defined on $\mathbb{R}$. For this special link function we are going to specify the constants which appear in Theorem 3.6. The log-likelihood based on the observations is given by

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \left[Y_i \beta^T X_i - \exp(\beta^T X_i) - \log(Y_i!)\right]$$

and thus the Poisson loss function is defined by

$$l(\beta) =: l(\beta; x, y) =: -y\beta^T x + \exp(\beta^T x).$$

It is formula (1) with $\psi = \exp$. In the particular case of Poisson regression the conditional law is defined by $Y|X \sim \mathcal{P}(\lambda^*(X))$ and the higher moments for a Poisson distribution are given by

$$\mathbb{E}\left(Y^k|X\right) = \sum_{l=1}^{k} \left(\lambda^*(X)\right)^l S_{l:k}$$

where $S_{l:k} = \frac{1}{l!}\sum_{i=0}^{l}(-1)^{l-i}\binom{l}{i} i^k \geqslant 0$ is the number of partitions of a set with l members into k undersets. The number $\sum_{l=1}^{k} S_{l:k} := B_k$ is called the k^{th} Bell number and this number satisfies the relation $B_{n+1} = \sum_{k=0}^{n}\binom{k}{n} B_k$ (see for example [10]). So we can easily prove by induction that $B_k \leqslant k!$ for all $k \geqslant 1$. Then we have on the event $\{0 \leqslant \lambda^*(X) < 1\}$

$$\mathbb{E}\left(Y^k\right) = \mathbb{E}\left(\sum_{i=1}^{k} \left(\lambda^*(X)\right)^i S_{i:k}\right) \leqslant \sum_{i=1}^{k} S_{i:k} \leqslant k!$$

and on the event $\{1 \leqslant \lambda(X)\}$ using (H.1) combined with (H.3) we find

$$\mathbb{E}\left(Y^k\right) = \sum_{i=1}^{k} S_{i:k}\mathbb{E}\left(\left(\lambda^*(X)\right)^i\right) \leqslant k!(e^{LB})^k$$

Because $e^{LB} \geqslant 1$ we deduce that for all $k \geqslant 1$

$$\mathbb{E}|Y|^k \leqslant k!(e^{LB})^k.$$

Therefore, for Poisson regression, we have $C_{L,B} = e^{LB}$. Besides $\max_{|x|\leqslant L\kappa_n}\left\{|\psi^{'}(x)|\right\} = e^{L\kappa_n}$ and $\min_{|x|\leqslant L\kappa_n}\left\{\psi^{''}(x)\right\} = e^{-L\kappa_n}$ where we recall that $\kappa_n := 17B + \frac{2}{n}$. Therefore, in the case of Poisson regression, Theorem 3.6 becomes

Corollary 4.1. *Assume that condition $GS(3, \frac{1}{2n})$ holds. If*

$$r_n \geqslant AKLe^{L(17B+\frac{2}{n})}\sqrt{\frac{\log(2G_n)}{n}}$$

with $A > \sqrt{2}$ then, with probability at least $1 - (C + 2d_{max})(2G_n)^{-A^2/2}$, we have

$$\|\hat{\beta}_n - \beta^*\|_1 \leqslant \frac{4}{c_n k\sqrt{d_{min}}} r_n {\gamma^*}^2 + \left(1 + \frac{1}{r_n}\right)\frac{1}{2n\sqrt{d_{min}}}$$

and

$$\mathbb{E}\left(\hat{\beta}_n^T X - {\beta^*}^T X\right)^2 \leqslant \frac{16}{(c_n)^2 k} r_n^2 {\gamma^*}^2 + \frac{2r_n + 1}{2nc_n}$$

with $c_n := e^{-L(9B+\frac{1}{n})}$.

4.2 Elastic net for generalized linear models

The most difficult part of the proof of Theorem 3.6 is to prove Proposition 3.1 (see above). Once this proposition has been proved it becomes easy to generalize the results presented above to any standard penalization using a modified version of the condition GS (depending on the norm we use). For example we can replace the ℓ_1 norm by a combination of ℓ_1 and ℓ_2 norms. It is the so-called Elastic net introduced by Zou, Trevor and Hastie [28] in the frame of linear regression. They showed that this estimator outperforms the Lasso in many situations for real world data and simulations. It is an alternative to the Group Lasso (the Elastic net has a behaviour similar to the one of the Group Lasso estimator). As the Lasso does, the Elastic net encourages sparsity and group selection but contrary to the Lasso when the sample size n is smaller than p the Elastic net can select more than n significant variables. This estimator is the solution of a convex optimization problem and Zou, Trevor and Hastie in their paper [28] proposed an algorithm to solve this problem. The Elastic net estimator for the generalized linear model is defined by

$$\hat{\beta}_n = \underset{\beta\in\Lambda}{\operatorname{argmin}}\left\{\mathbb{P}_n l(\beta) + 2r_n\|\beta\|_1 + t_n\|\beta\|_2^2\right\} \tag{4}$$

where r_n and t_n are the penalty parameters. Theorem 4.2 is an extension of the results first proved by Bunea [5] in the special case of logistic regression. Let $2t_nB = r_n$. We have the following theorem,

Theorem 4.2. *Assume condition $St(4, \frac{1}{2n})$ holds. Let*

$$r_n \geqslant AKL\left\{C_{L,B} \vee \underset{\{|x|\leqslant L(17B+\frac{2}{n})\}\cap\Theta}{max}|\psi^{'}(x)|\right\}\sqrt{\frac{\log(2p)}{n}}$$

where $A > \sqrt{2}$ and $C_{L,B}$ depends only on L and B. Then, with probability at least $1 - C(2p)^{-A^2/2}$ (where C is a universal constant), we have

$$\|\hat{\beta}_n - \beta^*\|_1 \leqslant \frac{(2.5)^2 r_n s^*}{t_n + c_n k} + \left(1 + \frac{1}{r_n}\right)\frac{1}{2n}$$

and

$$\mathbb{E}\left(\hat{\beta}_n^T X - \beta^{*T} X\right)^2 \leqslant \frac{2(2.5)^2}{c_n k(t_n + c_n k)} r_n^2 s^* + \frac{2r_n + 3}{2nc_n}.$$

where $c_n := \min_{\{|x| \leqslant L(9B+\frac{1}{n})\} \cap \Theta} \left\{\psi''(x)\right\}$.

We can notice that thanks to the ℓ_2 penalty the bound for the ℓ_1 and ℓ_2 errors are less sensitive to small value of k and small value of c_n (which can appears when L or B are large).

5 Simulations

We are going to compare the performances of the Lasso and Group Lasso for Poisson Regression on simulated data sets. Computations have been performed using `R`. We use the package `grplasso` developped by Meir, van de Geer and Buhlmann [15] for the Group Lasso and the package `glmnet` developped by Friedman, Hastie and Tibshirani [8] for the Lasso. The function `glmnet` fits the entire lasso regularization path for some generalized linear model via penalized maximum likelihood. We use this function in the particular case of Poisson regression. The function `grplasso` fits the solution of a Group Lasso problem for a model of type `grpl.model` which generates models to be used for Group Lasso algorithm and identify the exponential family of the response and the link function which is used. Here we consider the function `PoissReg()` which generates a Poisson model.

We simulate 100 data sets for each simulation and we ran the Lasso and Group Lasso on these data sets. Each data set X is cut into three separate subsets: a training data set, a validation data set and a test data set. We simulate responses via the model $Y \sim \mathcal{P}(\exp(X\beta))$ where $\beta = (\beta^1, ..., \beta^G)$ with g groups with non zero coefficients among the G groups. The training data set (X_{train} of size n_{train}) is used to fit the model (we estimate the target β for a sequence of tuning parameter λ and denote by β_λ the estimate of β obtained for such a parameter for the Lasso and the Group Lasso estimator). Then, we use the validation data (X_{valid} of size n_{valid}) to evaluate the performance of the fitted model according to a specific loss function. We define the optimal tuning parameter as the one for which the deviation from the fitted mean to the response is minimal i.e. $\lambda_{\text{opt}} \in \underset{\lambda}{\text{argmin}} \left\{\frac{1}{n_{\text{valid}}} \|Y - \exp(X_{\text{valid}}\beta_\lambda)\|_2^2\right\}$. From that, we determine the model with the parameter vector $\beta_{\lambda_{\text{opt}}}$. Then we compute the hits i.e the number of correctly identified relevant variables, the false positives i.e the number of non significant variables choosen as relevant and the degreee of freedom i.e the total number of variables selected in the model. Finally, we estimate the performance of the selected model by computing the coefficients estimation error $\|\beta - \beta_{\lambda_{\text{opt}}}\|_1$ and the prediction error $\|X_{\text{test}}\beta - X_{\text{test}}\beta_{\lambda_{\text{opt}}}\|_2$ on the test data (X_{test} of size n_{test}). We ran the Lasso and Group Lasso on these data sets.

Eight models are considered in the simulations. For each simulation $n_{train} = 50$, $n_{valid} = 50$, $n_{test} = 100$. To compare the Lasso and Group Lasso we use a random design matrix where the predictors are simulated as followed according to a uniform distribution to have bounded predictors.

1.
   - $X_i = U_1 + \varepsilon_i$ for $1 \leqslant i \leqslant 10$ with $U_1 \sim U([0, 1])$
   - $X_i = U_2 + \varepsilon_i$ for $11 \leqslant i \leqslant 20$ with $U_2 \sim U([0, 1])$
   - $X_i = U_i$ for the last 100 variables $U_i \sim U([-0.1, 0.1])$

   with ε_i i.i.d $\sim U([0, 0.01])$.
   The covariates within the first two blocks are highly correlated (~ 0.8) and there are small correlations between the blocks. The target is

$$\beta = (\underbrace{0.3, ..., 0.3}_{10}, \underbrace{0.2, ..., 0.2}_{10}, \underbrace{\underbrace{0, ..., 0}_{10}, ..., \underbrace{0, ..., 0}_{10}}_{10}).$$

2. The simulation is the same as the first one except that ε_i i.i.d $\sim U([0, 1])$. Thus there are small correlations within and between groups (~ 0.5).

3. The simulation is the same as the first one except that ε_i i.i.d $\sim U([0, 1.2])$. Thus there are very small correlations within and between groups (~ 0.2).

   For all the following simulations the non zero groups are generated in the same way as in the second example. For $j = 1, ..., G^*$ and $i = 1, ..., d_j$, $X_i = U_j + \varepsilon_i$ where $U_j \sim U([0, 1])$ and ε_i i.i.d $\sim U([0, 1])$. The non influencial groups are simulated according to $X_i = U_i$ with $U_i \sim U([-0.1, 0.1])$. In the two following simulations we increase the size of the non zero groups

4. $$\beta = (\underbrace{0.3, ..., 0.3}_{20}, \underbrace{0.2, ..., 0.2}_{20}, \underbrace{\underbrace{0, ..., 0}_{20}, ..., \underbrace{0, ..., 0}_{20}}_{10}).$$

5. $$\beta = (\underbrace{0.3, ..., 0.3}_{5}, \underbrace{0.2, ..., 0.2}_{5}, \underbrace{\underbrace{0, ..., 0}_{5}, ..., \underbrace{0, ..., 0}_{5}}_{10}).$$

   In the last three simulations it is the number of the non zero groups which is increased.

6. $$\beta = (\underbrace{0.2, ..., 0.2}_{10}, \underbrace{0.2, ..., 0.2}_{10}, \underbrace{\underbrace{0, ..., 0}_{10}, ..., \underbrace{0, ..., 0}_{10}}_{10}).$$

7. $$\beta = (\underbrace{\underbrace{0.2, ..., 0.2}_{10}, ..., \underbrace{0.2, ..., 0.2}_{10}}_{4}, \underbrace{\underbrace{0, ..., 0}_{10}, ..., \underbrace{0, ..., 0}_{10}}_{10}).$$

8. $$\beta = (\underbrace{\underbrace{0.2, ..., 0.2}_{10}, ..., \underbrace{0.2, ..., 0.2}_{10}}_{6}, \underbrace{\underbrace{0, ..., 0}_{10}, ..., \underbrace{0, ..., 0}_{10}}_{10}).$$

The results of the eight simulations are reported in Table I hereafter where p is the number of covariates, s the number of significant covariates, G the number of groups, G^* the number of non zero groups and v is the size of the non zero groups. The Group Lasso seems to perform better than the Lasso to include the relevant predictors into the model particularly when there are high within group correlations. The Lasso tends to select fewer variables among the influencial ones (the Lasso selects only some variables from the groups of highly correlated predictors) than the Group Lasso does in the case of highly correlated covariates and when the size or the number of the nonzero groups is large. The Group Lasso succeeds in including the true significant groups in most of the cases. On the contrary the Group Lasso estimator tends to add more irrelevant covariates than the Lasso does in particular when the number or the size of the non influencial groups is large. We can expect such a result because the Group Lasso estimator includes not single variable one after anoter but groups of variables (when one of the covariates is included in the model all the others which belongs to the same group are also included in the model). Thus the Group Lasso selects models that are larger than the true model. However when the Group Lasso selects a number of covariates which is the same than the number of significants covariates it is, with high probability, the correct groups which are included. We have also measured the performances of the Lasso and Group Lasso in terms of estimation error and prediction error. The Group Lasso seems to perform better than the Lasso in term of estimation error and of prediction error in most of the cases and the improvement is particularly meaningful for prediction error. We can also notice that the performances of the two estimators decreases with the increase of G and G^*. To conclude, the Group Lasso estimator seems to perform better than the Lasso to include the hits in the model and in terms of prediction and estimation error when the covariates are structured into groups and in particular in the case of high correlations within groups.

Simulation	1	2	3	4	5	6	7	8
s/p	20/120	20/120	20/120	40/240	10/60	20/120	40/140	60/160
G	12	12	12	12	12	12	14	16
G^*	2	2	2	2	2	2	4	6
v	10	10	10	20	5	10	10	10
Mean Hit lasso (%)	86,9	99,4	99,95	66,27	87,6	95,9	78,79	54,16
Mean Hit group lasso (%)	100	100	100	98,5	100	100	100	98,8
Mean False positive lasso (%)	15,02	10,61	9,25	2,39	47,38	18,96	9,33	8,87
Mean False positive group lasso (%)	28,7	36,1	33,9	74	34,2	29,6	61,9	37,9
Mean Nonzero lasso	32,4	30,49	29,24	31,29	32,45	38,14	40,85	41,37
Mean Nonzero gp lasso	48,7	56,1	53,9	187,4	27,1	49,6	101,9	97,2
Mean Prediction error lasso	0,19	0,15	0,18	6,73	0,15	0,17	3,22	17,97
Mean Prediction error gp lasso	0,021	0,007	0,004	1,98	0,03	0,016	0,012	1,32
Mean Estimation error lasso	6,36	2,84	2,08	9,79	15,31	6,37	8,45	16,28
Mean Estimation error gp lasso	5,54	2,66	1,64	18,65	4,22	3,40	2,77	12,76

Table 1: Simulation results of the eight models

6 Conclusion

We consider the generalized linear model in high dimensional settings and use the Group Lasso estimator to estimate the regression parameter β when the covariates are naturally structured into groups of variables and the true parameter is group sparse (just a few variables are relevant to explain the response). Under some assumptions on the sparsity of β, on the correlations between the groups of covariates and on the tuning parameter of the estimator we state general oracle inequalities for estimation and error prediction for the Group Lasso estimator applied to generalized linear models. In the particular case of groups of size one, we provide original inequalities for the Lasso in the case of generalized linear models extending the results of Bunea [5] for logistic regression. Furthermore, we extend these results to other penalties such as the Elastic net. Then we compare the performances of the Lasso to the ones of the Group Lasso on simulated data. We show the improvement in terms of variables selection and prediction errror of the Group Lasso compared to the Lasso when the covariates are structured into groups. Moreover we illustrate on these simulated data the impact of the total number of groups, of the number of non zero groups and of the size of the groups on the performances of the Group Lasso. The conclusion is that the Group Lasso estimator behave well in a high dimensional setting under sparsity and group correlations assumptions. However the main drawback of the Group Lasso is that we need an a priori knowledge on the groups and it is not always possible.

Appendices

A Proof of Theorem 3.6

A.1 The main steps of the proof

The proof follows the guidelines in [4] or [12]. Using the mere definition of $\hat{\beta}_n$, we have

$$\mathbb{P}_n l(\hat{\beta}_n) + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g} \|\hat{\beta}_n^g\|_2 \leqslant \mathbb{P}_n l(\beta^*) + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g} \|\beta^{*g}\|_2. \tag{5}$$

Hence we get

$$\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right) + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g} \|\hat{\beta}_n^g\|_2$$

$$\leqslant (\mathbb{P}_n - \mathbb{P})\left(l(\beta^*) - l(\hat{\beta}_n)\right) + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\beta^{*g}\|_2. \tag{6}$$

We decompose the empirical process into a linear part and a part which depends on the normalized parameter ψ.

$$(\mathbb{P}_n - \mathbb{P})\left(l(\beta^*) - l(\hat{\beta}_n)\right)$$
$$= (\mathbb{P}_n - \mathbb{P})\left(l_l(\beta^*) - l_l(\hat{\beta}_n)\right) + (\mathbb{P}_n - \mathbb{P})\left(l_\psi(\beta^*) - l_\psi(\hat{\beta}_n)\right)$$

where

$$l_l(\beta) := l_l(\beta, x, y) = -y\beta^T x$$

and

$$l_\psi(\beta) := l_\psi(\beta, x) = \psi(\beta^T x).$$

From Proposition 3.3 and Proposition 3.5 and by adding $r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2$ to both sides of the inequality (6) we find, on $\mathcal{A} \cap \mathcal{B}$, that

$$r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 + \mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right)$$

$$\leqslant 2r_n \sum_{g=1}^{G_n} \sqrt{d_g}\left(\|\hat{\beta}_n^g - \beta^{*g}\|_2 + \|\beta^{*g}\|_2 - \|\hat{\beta}_n^g\|_2\right) + \frac{r_n}{2}\varepsilon_n.$$

If $g \notin H^*$ then $\|\hat{\beta}_n^g - \beta^{*g}\|_2 + \|\beta^{*g}\|_2 - \|\hat{\beta}_n^g\| = 0$ and otherwise $\|\beta^{*g}\|_2 - \|\hat{\beta}_n^g\|_2 \leqslant \|\hat{\beta}_n^g - \beta^{*g}\|_2$. So the last inequality can be bounded by

$$4r_n \sum_{g \in H^*} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 + \frac{r_n}{2}\varepsilon_n. \tag{7}$$

By the definition of β^* we have $\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right) > 0$ and therefore

$$\sum_{g \notin H^*} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 \leqslant 3 \sum_{g \in H^*} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 + \frac{\varepsilon_n}{2}$$

i.e $\hat{\beta}_n - \beta^* \in GS(3, \frac{\varepsilon_n}{2})$. The next proposition provides a lower bound for $\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right)$.

Proposition A.1. *On the event $\mathcal{A} \cap \mathcal{B}$ we have*

$$\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right) \geqslant c_n \mathbb{E}\left[\left(f_{\hat{\beta}_n}(X) - f_{\beta^*}(X)\right)^2\right]$$

with $c_n := \min_{|x| \leqslant L(9B + \frac{1}{n})} \left\{\psi''(x)\right\}$.

Proof.

$$\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right) = -\mathbb{E}\left[\mathbb{E}(Y|X)\left(f_{\hat{\beta}_n}(X) - f_{\beta^*}(X)\right)\right]$$
$$+\mathbb{E}\left[\psi'(f_{\beta^*}(X))\left(f_{\hat{\beta}_n}(X) - f_{\beta^*}(X)\right)\right]$$
$$+\mathbb{E}\left[\psi''(f_{\tilde{\beta}}(X))\left(f_{\hat{\beta}_n}(X) - f_{\beta^*}(X)\right)^2\right]$$

where $\tilde{\beta}^T X$ is an intermediate point between $\hat{\beta}_n^T X$ and $\tilde{\beta}^T X$ given by a second order Taylor expansion of ψ. Since $\psi'(f_{\beta^*}(X)) = \mathbb{E}(Y|X)$ we find

$$\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right) = \mathbb{E}\left[\psi''(f_{\tilde{\beta}}(X))\left(f_{\hat{\beta}_n}(X) - f_{\beta^*}(X)\right)^2\right].$$

Besides we have

$$|\tilde{\beta}^T X| \leqslant |\tilde{\beta}^T X - \beta^{*T} X| + |\beta^{*T} X|$$
$$\leqslant \sum_{g=1}^{G_n} \|\beta^{*g} - \beta^g\|_2 \|X^g\|_2 + \sum_{g=1}^{G_n} \|\beta^{*g}\|_2 \|X^g\|_2.$$

Applying (H.1), we find

$$\|X^g\|_2 \leqslant L\sqrt{d_g}$$

Then using Lemma 3.4 and (H.3) we find

$$|\tilde{\beta}^T X| \leqslant LM + LB \quad \text{a.s.} \tag{8}$$

Furthermore, β^* and $\hat{\beta}_n$ belongs to Λ which is a convex set. Therefore $\tilde{\beta} \in \Lambda$ and $\tilde{\beta}^T X \in \Theta$ a.s. Thus we conclude

$$\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right) \geqslant c_n \mathbb{E}\left[\left(f_{\hat{\beta}_n}(X) - f_{\beta^*}(X)\right)^2\right]$$

where $c_n := \min_{\{|x| \leqslant L(M+B)\} \cap \Theta} \psi''(x)$. □

From Propositon A.1 and (7) we deduce that

$$r_n \sum_{g=1}^{G_n} \sqrt{d_g} \|\hat{\beta}_n^g - \beta^{*g}\|_2 + c_n \mathbb{E}\left(\hat{\beta}_n^T X - \beta^{*T} X\right)^2$$
$$\leqslant 4 r_n \sum_{g \in H^*} \sqrt{d_g} \|\hat{\beta}_n^g - \beta^{*g}\|_2 + \frac{r_n}{2} \varepsilon_n. \tag{9}$$

Then the end of the proof is similar to the end of the proof of Theorem 2.4 in [5]. Let Σ be the $p \times p$ matrix whose entries are $\mathbb{E}(X_k X_j)$. We have

$$\mathbb{E}\left(\hat{\beta}_n^T X - \beta^{*T} X\right)^2 = (\hat{\beta}_n - \beta^*)^T \Sigma (\hat{\beta}_n - \beta^*).$$

Because condition $GS(3, \frac{\varepsilon_n}{2})$ is satisfied we have

$$c_n (\hat{\beta}_n - \beta^*)^T \Sigma (\hat{\beta}_n - \beta^*) \geqslant c_n k \sum_{g \in H^*} \|\hat{\beta}_n^g - \beta^{*g}\|_2^2 - \frac{\varepsilon_n}{2}.$$

Then by using Cauchy-Schwarz inequality in (9) we find

$$r_n \sum_{g=1}^{G_n} \sqrt{d_g} \|\hat{\beta}_n^g - \beta^{*g}\|_2 + c_n k \sum_{g \in H^*} \|\hat{\beta}_n^g - \beta^{*g}\|_2^2$$
$$\leqslant 4 r_n \sqrt{\sum_{g \in H^*} d_g} \sqrt{\sum_{g \in H^*} \|\hat{\beta}_n^g - \beta^{*g}\|_2^2} + (r_n + 1) \frac{\varepsilon_n}{2}. \tag{10}$$

Now the fact that $2xy \leqslant tx^2 + y^2/t$ for all $t > 0$ leads to the following inequality

$$r_n \sum_{g=1}^{G_n} \sqrt{d_g} \|\hat{\beta}_n^g - \beta^{*g}\|_2 + c_n k \sum_{g \in H^*} \|\hat{\beta}_n^g - \beta^{*g}\|_2^2$$
$$\leqslant 4 t r_n^2 \gamma^{*2} + \frac{1}{t} \sum_{g \in H^*} \|\hat{\beta}_n^g - \beta^{*g}\|_2^2 + (r_n + 1) \frac{\varepsilon_n}{2}. \tag{11}$$

Replacing t by $\dfrac{1}{c_n k}$ in (11) we obtain

$$\sum_{g=1}^{G_n} \sqrt{d_g} \|\hat{\beta}_n^g - \beta^{*g}\|_2 \leqslant \frac{4}{c_n k} r_n \gamma^{*2} + (1 + \frac{1}{r_n}) \frac{\varepsilon_n}{2}.$$

What is more

$$\sqrt{d_{\min}}\|\hat{\beta}_n - \beta^*\|_2 \leqslant \sum_{g=1}^{G_n} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2.$$

Thus

$$\|\hat{\beta}_n - \beta^*\|_2 \leqslant \frac{4}{c_n k\sqrt{d_{\min}}} r_n {\gamma^*}^2 + (1 + \frac{1}{r_n})\frac{\varepsilon_n}{2\sqrt{d_{\min}}}. \tag{12}$$

A similar argument could be made to prove

$$\mathbb{E}\left(\hat{\beta}_n^T X - {\beta^*}^T X\right)^2 \leqslant \frac{16}{(c_n)^2 k} r_n^2 {\gamma^*}^2 + \frac{2r_n + 1}{c_n}\frac{\varepsilon_n}{2}. \tag{13}$$

Finally we conclude the proof using Proposition 3.1.

A.2 Proof of Proposition 3.1

Proof. Let $A > \sqrt{2}$. We recall that we have made the assumption $\frac{\log(2G_n)}{n} \leqslant 1$. Then we deduce Proposition 3.1 from the two following lemmas.

Lemma A.2. *Let*

$$r_n \geqslant \left(A8\sqrt{2}LC_{L,B}\sqrt{\frac{\log(2G_n)}{n}}\right) \vee \left(A^2 16LC_{L,B}\frac{\log(2G_n)}{n}\right)$$

with $A > 1$. *Then*

$$\mathbb{P}\{\mathcal{A}\} \geqslant 1 - 2d_{max}(2G_n)^{1-A^2}.$$

We can notice that $\mathbb{P}(\mathcal{A}) \underset{n\to\infty}{\longrightarrow} 1$.

Lemma A.3. *Let*

$$r_n \geqslant A20L\left(\max_{\{|x|\leqslant L\kappa_n\}\cap\Theta} |\psi'(x)|\right)\sqrt{\frac{2\log 2G_n}{n}}$$

where $A \geqslant 1$. *Then*

$$\mathbb{P}(\mathcal{B}) \geqslant 1 - C(2G_n)^{-A^2/2}$$

where we recall $\kappa_n := 17B + \frac{2}{n}$. *We can notice that* $\mathbb{P}(\mathcal{B}) \underset{n\to\infty}{\longrightarrow} 1$.

Thus if

$$r_n \geqslant AKL\left\{C_{L,B} \vee \max_{\{|x|\leqslant L\kappa_n\}\cap\Theta} |\psi'(x)|)\right\}\sqrt{\frac{2\log(2G_n)}{n}}$$

with K choosen such that

$$r_n \geqslant \max(C_1, C_2, C_3)$$

where

$$C_1 := A8\sqrt{2}LC_{L,B}\sqrt{\frac{\log(2G_n)}{n}}$$

$$C_2 := A^2 16LC_{L,B}\frac{\log(2G_n)}{n}$$

and

$$C_3 := A20L\left(\max_{\{|x|\leqslant L\kappa_n\}\cap\Theta} |\psi'(x)|\right)\sqrt{\frac{2\log 2G_n}{n}}$$

then $\mathbb{P}(\mathcal{A}\cap\mathcal{B}) \geqslant 1 - (2d_{\max} + C)(2G_n)^{-A^2/2}$. ∎

A.3 Proofs of the technical lemmata

Proof of Lemma 3.2

Proof. Let $\theta \in \overset{\circ}{\Theta}$ and Y_θ be a real random variable with density $\exp(\theta y - \psi(\theta))F(dy)$. First we prove that for all $k \in \mathbb{N}$, there exists a constant $C_\theta > 0$ which depends on θ such that

$$\mathbb{E}|Y_\theta|^k \leqslant k! C_\theta^k.$$

The k^{th} absolute moment of Y_θ is the k^{th} derivate of $H_\theta := \mathbb{E}(e^{s|Y_\theta|})$ at 0. Let $s \in \mathbb{C}$ be given. We have

$$H_\theta(s) = \frac{M_+(s+\theta) + M_-(\theta - s)}{M(\theta)}$$

where we define M_+ and M_- as

$$M_+ : \left| \begin{array}{ccc} \mathbb{C} & \longrightarrow & \mathbb{C} \\ z & \longmapsto & \int_{y \geqslant 0} e^{yz} F(dy) \end{array} \right.$$

and

$$M_- : \left| \begin{array}{ccc} \mathbb{C} & \longrightarrow & \mathbb{C} \\ z & \longmapsto & \int_{y < 0} e^{yz} F(dy). \end{array} \right.$$

M is analytic on $\Omega_\Theta := \left\{ z \in \mathbb{C} : \mathrm{Re}(z) \in \overset{\circ}{\Theta} \right\}$ and so does M_+ and M_-. Therefore $H_\theta : s \longmapsto \mathbb{E}(e^{s|Y_\theta|})$ is analytic on $U_\theta := \left\{ s \in \mathbb{C} : \mathrm{Re}(s+\theta) \in \overset{\circ}{\Theta} \quad \text{and} \quad \mathrm{Re}(\theta - s) \in \overset{\circ}{\Theta} \right\}$. Since $\theta \in \overset{\circ}{\Theta}$, H_θ is analytic at the point 0 and hence the function is also holomorphic in a neighborhood of 0. We recall the following result for analytic functions (see [24]).
Theorem: *If f is holomorphic on a domain Ω of $\mathbb{C}$ then $f \in \mathcal{C}^\infty(\Omega)$ and if in addition Ω is simply connected then for all contour γ around $z \in \Omega$ we have*

$$f^{(n)}(z) = \frac{n!}{2i\pi} \int_\gamma \frac{f(v)}{(v-z)^{n+1}} dv.$$

Using the previous theorem on H_θ at 0 and taking γ as a circle with radius R centered in 0 (such that H_θ is holomorphic on $D(0,R)$, of course R depends on θ) we obtain that for all $k \in \mathbb{N}$

$$|H_\theta^{(k)}(0)| \leqslant \frac{k!}{2\pi} \left| \int_\gamma \frac{H_\theta(v)}{v^{k+1}} dv \right| \leqslant \frac{k!}{R^k} \sup_{|z| \leqslant R} |H_\theta(z)|. \tag{14}$$

and the result follows with $C_\theta := \max\left(1, \frac{1}{R} \sup_{|z| \leqslant R} |H_\theta(z)|\right)$. Thanks to assumption (H.2) we can apply (14) with $\theta = {\beta^*}^T X$ and find that for all $k \in \mathbb{N}$,

$$\mathbb{E}\left(|Y|^k | X\right) \leqslant k! \left(C_{{\beta^*}^T X}\right)^k.$$

Finally (H.1) combined with (H.3) leads to

$$\mathbb{E}\left(|Y|^k\right) \leqslant k! \left(\sup_{|\theta| \leqslant LB} C_\theta \right)^k$$

and the result follows with $C_{L,B} := \sup_{|\theta| \leqslant LB} C_\theta$. □

Proof of Lemma 3.4

Proof. The proof is based on the convexity of the loss function and of the penalty, the main idea of the proof is similar to the one used by Bühlmann and van de Geer [4] for the Lasso to show consistency of the excess of risk. Define $t := \dfrac{M}{M + \sum_{g=1}^{G_n} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2}$ and $\tilde{\beta} := t\hat{\beta}_n + (1-t)\beta^*$. Notice $\sum_{g=1}^{G_n} \sqrt{d_g}\|\tilde{\beta}^g - \beta^{*g}\|_2 \leqslant M$. By convexity of $\beta \to l_\psi(\beta)$ and $\beta \to \|\beta\|_2$ combined with the fact that $\hat{\beta}_n$ satisifies (5) we find

$$\mathbb{P}\left(l(\tilde{\beta}) - l(\beta^*)\right) + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\tilde{\beta}^g\|_2$$

$$\leqslant (\mathbb{P}_n - \mathbb{P})\left(l(\beta^*) - l(\tilde{\beta})\right) + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\beta^{*g}\|_2.$$

On the event $\mathcal{A} \bigcap \mathcal{B}$ we have

$$\mathbb{P}\left(l(\tilde{\beta}) - l(\beta^*)\right) + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\tilde{\beta}^g\|_2$$

$$\leqslant r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\tilde{\beta}^g - \beta^{*g}\|_2 + r_n\frac{\varepsilon_n}{2} + 2r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\beta^{*g}\|_2.$$

Because $\mathbb{P}\left(l(\tilde{\beta}) - l(\beta^*)\right) \geqslant 0$, by adding to both sides of the inequality $2r_n \sum_{g=1}^{G_n} \sqrt{d_g}\|\beta^{*g}\|_2$ and by using the triangular inequality we have

$$\sum_{g=1}^{G_n} \sqrt{d_g}\|\tilde{\beta}^g - \beta^{*g}\|_2 \leqslant \frac{\varepsilon_n}{2} + 4\sum_{g=1}^{G_n} \sqrt{d_g}\|\beta^{*g}\|_2.$$

Therefore, using (H.3), we have

$$\sum_{g=1}^{G_n} \sqrt{d_g}\|\tilde{\beta}^g - \beta^{*g}\|_2 \leqslant \frac{\varepsilon_n}{2} + 4B = \frac{M}{2}.$$

i.e.

$$t\sum_{g=1}^{G_n} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 \leqslant \frac{M}{2}$$

and then the definition of t leads to

$$\sum_{g=1}^{G_n} \sqrt{d_g}\|\hat{\beta}_n^g - \beta^{*g}\|_2 \leqslant M.$$

□

Proof of Lemma A.2

Proof. We have

$$\mathbb{P}(\mathcal{A}^c) \leqslant \sum_{g=1}^{G_n} \mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\left(Y_i X_i^g - \mathbb{E}(Y X^g)\right)\right\|_2^2 > \frac{r_n^2}{4}d_g\right\}$$

$$\leqslant \sum_{g=1}^{G_n}\sum_{j=1}^{d_g} \mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\left(Y_i X_{i,j}^g - \mathbb{E}\left(Y X_j^g\right)\right)\right| > \frac{r_n}{2}\right\}. \tag{15}$$

For $j = 1, ..., d_g$ and $i = 1, ..., n$, let

$$W_{ij}^g := Y_i X_{ij}^g - \mathbb{E}\left(Y X_j^g\right).$$

The random variables $\{W_{ij}\}_{i=1,\dots,n}$ are independant, identically distributed and centered and for all $m \geqslant 2$

$$\mathbb{E}|W_{ij}^g|^m \leqslant \sum_{k=0}^{m} \binom{k}{m} \mathbb{E}|YX_j|^k \left(\mathbb{E}|YX_j|\right)^{m-k}$$

By using Jensen inequality we obtain

$$\mathbb{E}|W_{ij}^g|^m \leqslant 2^m \max_{k=1,\dots,m} \left\{\mathbb{E}|YX_j|^k \mathbb{E}|YX_j|^{m-k}\right\}.$$

For all $k \in \mathbb{N}$, by (H.1) and Lemma 3.2 we have

$$\mathbb{E}|YX_j|^k \leqslant L^k k! \left(C_{L,B}\right)^k.$$

Therefore $\mathbb{E}|W_{ij}^g|^m \leqslant m!(2LC_{L,B})^m$. Hence the conditions are satisfied to apply Bernstein concentration inequality [1] with $K = 2LC_{L,B}$ and $\sigma^2 = 8(LC_{L,B})^2$. Thus we obtain

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} W_{ij}^g\right| > r_n/2\right)$$

$$\leqslant 2\left(\exp\left(\frac{-nr_n}{16LC_{L,B}}\right) + \exp\left(\frac{-nr_n^2}{32(2LC_{L,B})^2}\right)\right). \tag{16}$$

Finally, from (15) and (16), we deduce that $\mathbb{P}(\mathcal{A}^c)$ is bounded by

$$2d_{\max}G_n\left(\exp\left(\frac{-nr_n}{16LC_{L,B}}\right) + \exp\left(\frac{-nr_n^2}{32(LC_{L,B})^2}\right)\right).$$

Therefore if

$$r_n \geqslant A^2 16LC_{L,B}\frac{\log(2G_n)}{n} \vee A8\sqrt{2}LC_{L,B}\sqrt{\frac{\log(2G_n)}{n}}$$

with $A > 1$ then $\mathbb{P}\left\{\mathcal{A}^C\right\} \leqslant 2d_{\max}(2G_n)^{1-A^2}$. □

Proof of Lemma A.3

Proof. The proof rests on the following Lemma

Lemma A.4. *Let $R > 0$ be given. Define*

$$Z_R := \sup_{\sum_{g=1}^{G_n}\sqrt{d_g}\|\beta^g - \beta^{*g}\|_2 \leqslant R} \left\{\left|\left(\mathbb{P}_n - \mathbb{P}\right)\left(l_\psi(\beta^*) - l_\psi(\beta)\right)\right|\right\}.$$

If $A \geqslant 1$ then

$$\mathbb{P}\left(Z_R \geqslant A5DLR\sqrt{\frac{2\log 2G_n}{n}}\right) \leqslant (2G_n)^{-A^2} \tag{17}$$

where $D := \max_{\{|x| \leqslant L(R+B)\} \cap \Theta} \left\{\left|\psi'(x)\right|\right\}$.

Proof. Let β satisfy $\sum_{g=1}^{G_n}\sqrt{d_g}\|\beta^g - \beta^{*g}\|_2 \leqslant R$. Notice that if we change X_i by X_i' while keeping the others fixed then Z_R is modified of at most $\frac{2}{n}LR\exp\left(L(R+B)\right)$. To see this let

$$\mathbb{P}_n = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j, Y_j}$$

and

$$\mathbb{P}_n' = \frac{1}{n}\sum_{j=1, j\neq i}^{n} 1_{X_j, Y_j} + 1_{X_i', Y_i'}$$

then we have

$$(\mathbb{P}_n - \mathbb{P})(l_\psi(\beta^*) - l_\psi(\beta)) - (\mathbb{P}_n^{'} - \mathbb{P})(l_\psi(\beta^*) - l_\psi(\beta))$$
$$= \frac{1}{n}\left(l_\psi(\beta^*, X_i) - l_\psi(\beta, X_i) - l_\psi(\beta^*, X_i^{'}) + l_\psi(\beta, X_i^{'})\right)$$
$$\leqslant \frac{1}{n}|\psi^{'}(\tilde{\beta}^T X_i)||\beta^{*T} X_i - \beta^T X_i| + \frac{1}{n}|\psi^{'}(\tilde{\beta}^T X_i^{'})||\beta^{*T} X_i^{'} - \beta^T X_i^{'}|$$

with $\tilde{\beta}^T X_i$ which is an intermediate point between $\beta^T X_i$ and $\beta^{*T} X_i$ (using a first order Taylor expansion of the exponential function). Then, using the same argument as for (8), we have

$$|\tilde{\beta}^T X_i| \leqslant LR + LB.$$

Therefore

$$(\mathbb{P}_n - \mathbb{P})(l_\psi(\beta^*) - l_\psi(\beta)) - (\mathbb{P}_n^{'} - \mathbb{P})(l_\psi(\beta^*) - l_\psi(\beta))$$
$$\leqslant \frac{2}{n} LR \max_{\{|x| \leqslant L(R+B)\} \cap \Theta} |\psi^{'}(x)| = \frac{2}{n} LRD. \tag{18}$$

We can apply McDiarmid's inequality also called the bounded difference inequality.

Theorem. *Let A a set. Assume $g : A^N \to \mathbb{R}$ is a function that satisfies the bounded difference inequality*

$$\sup_{x_1,...,x_n,x_i^{'} \in A} |g(x_1, ..., x_n) - g(x_1, ..., x_{i-1}, x_i^{'}, x_{i+1}, ..., x_n)| \leqslant c_i.$$

Let $X_1, .., X_n$ be independent random variables all taking values in the set A. Then for all $t > 0$,

$$\mathbb{P}\left\{g(X_1, ..., X_n) - \mathbb{E}g(X1, ..., X_n) \geqslant t\right\} \leqslant e^{-2t^2 / \sum_{i=1}^n c_i^2}.$$

We can apply McDiarmid's inequality to Z_R and obtain

$$\mathbb{P}(Z_R - \mathbb{E}Z_R \geqslant u) \leqslant \exp\left(-\frac{nu^2}{2R^2L^2\left(D\right)^2}\right).$$

Therefore if $r_n \geqslant ADLR\sqrt{\frac{2\log 2G_n}{n}}$ with $A > 0$ then

$$\mathbb{P}(Z_R - \mathbb{E}Z_R \geqslant r_n) \leqslant (2G_n)^{-A^2}. \tag{19}$$

Now we have to bound the mean $\mathbb{E}Z_R$.

Lemma A.5.

$$\mathbb{E}Z_R \leqslant 4RLD\sqrt{\frac{2\log(2G_n)}{n}}.$$

Proof. First let us introduce two theorems. Let $X_1, ..., X_n$ independent random variables with values in some space $\mathcal{X}$ and $\mathcal{F}$ a class of real-valued functions on $\mathcal{X}$.

Theorem: Symmetrization theorem [23]. *Let $\epsilon_1, ..., \epsilon_n$ be Rademacher sequence independent of $X_1, ..., X_n$. Then*

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}|\sum_{i=1}^n \{f(X_i) - \mathbb{E}(f(X_i))\}|\right)$$
$$\leqslant 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}|\sum_{i=1}^n \epsilon_i f(X_i)|\right).$$

Theorem: Contraction principle [11]. *Let $x_1, ..., x_n$ elements of $\mathcal{X}$ and $\varepsilon_1, ..., \varepsilon_n$ be Rademacher sequence. Consider Lipschitz functions g_i. Then for any function $h : \mathcal{X} \to \mathbb{R}$, we have*

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}|\sum_{i=1}^n \epsilon_i \{g_i(f(x_i)) - g_i(h(x_i))\}|\right)$$

$$\leqslant 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}|\sum_{i=1}^{n}\epsilon_i(f(x_i)-h(x_i))|\right)$$

Let $\epsilon_1,...,\epsilon_n$ a Rademacher sequence independant of $X_1,...,X_n$ and let $\mathcal{S}_R := \left\{\beta\in\mathbb{R}^p : \sum_{g=1}^{G_n}\sqrt{d_g}\|\beta^g-\beta^{*g}\|_2\leqslant R\right\}$. Then by the Symmetrization theorem and the Contraction theorem (ψ is D-lipschitz on the compact set $\mathcal{S}_R$) we have

$$\mathbb{E}Z_R \leqslant 4D\mathbb{E}\left(\sup_{\beta\in\mathcal{S}_R}\frac{1}{n}\sum_{i=1}^{n}\left|\epsilon_i(\beta^{*T}X_i-\beta^TX_i)\right|\right)$$

$$\leqslant 4DR\mathbb{E}\left(\max_{g\in\{1,...,G_n\}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{\|X_i^g\|_2}{\sqrt{d_g}}\right|\right),$$

by Holder inequality for the last bound. Now we are going to use the following theorem which is a consequence of Hoeffding inequality.

Theorem. *Let $X_1,...,X_n$ be independent random variables on $\mathcal{X}$ and $f_1,...,f_n$ real-valued functions on $\mathcal{X}$ which satisfies for all $j=1,...,p$ and all $i=1,...,n$*

$$\mathbb{E}f_j(X_i)=0,\quad |f_j(X_i)|\leqslant a_{ij}.$$

Then

$$\mathbb{E}\max_{1\leqslant j\leqslant p}\left|\sum_{i=1}^{n}f_j(X_i)\right|\leqslant\sqrt{2\log(2p)}\max_{1\leqslant j\leqslant p}\sqrt{\sum_{i=1}^{n}a_{ij}^2}.$$

Therefore we obtain

$$\mathbb{E}\left(\max_{g\in\{1,...,G_n\}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\frac{\|X_i^g\|_2}{\sqrt{d_g}}\right|\right)\leqslant L\sqrt{\frac{2\log(2G_n)}{n}}.$$

Thus

$$\mathbb{E}Z_R\leqslant 4RLD\sqrt{\frac{2\log(2G_n)}{n}}. \tag{20}$$

□

So we can conclude from (19) and (20) that if $A\geqslant 1$ then

$$\mathbb{P}\left(Z_R\geqslant A5DLR\sqrt{\frac{2\log 2G_n}{n}}\right)\leqslant(2G_n)^{-A^2} \tag{21}$$

for all $R>0$. □

Split up

$$\left\{\beta\in\mathbb{R}^p:\sum_{g=1}^{G_n}\sqrt{d_g}\|\beta^g-\beta^{*g}\|_2\leqslant M\right\},$$

where $M=8B+\varepsilon_n$, into two sets which are

$$E_1=\left\{\beta:\sum_{g=1}^{G_n}\sqrt{d_g}\|\beta^g-\beta^{*g}\|_2\leqslant\varepsilon_n\right\},$$

$$E_2=\left\{\beta:\varepsilon_n\leqslant\sum_{g=1}^{G_n}\sqrt{d_g}\|\beta^g-\beta^{*g}\|_2\leqslant M\right\}$$

$$\subseteq\bigcup_{j=1}^{j_n}\left\{\beta:2^{j-1}\varepsilon_n<\sum_{g=1}^{G_n}\sqrt{d_g}\|\beta^g-\beta^{*g}\|_2\leqslant 2^j\varepsilon_n\right\}$$

where $j_n := \lfloor \log_2(nM) \rfloor + 1$ is the smaller integer such that $2^{j_n}\varepsilon_n \geqslant M$. To simplify notations let

$$\nu_n(\beta, \beta^*) := \frac{(\mathbb{P}_n - \mathbb{P})\left(l_\psi(\beta^*) - l_\psi(\beta)\right)}{\sum_{g=1}^{G_n} \sqrt{d_g}\|\beta^g - \beta^{*g}\|_2 + \varepsilon_n},$$

$$\alpha_n(\beta, \beta^*) := (\mathbb{P}_n - \mathbb{P})\left(l_\psi(\beta^*) - l_\psi(\beta)\right)$$

and

$$\Phi(t) := \max_{\{|x| \leqslant t\} \cap \Theta} |\psi^{'}(x)|.$$

Let $A \geqslant 1$. We recall that $\kappa_n := 17B + \frac{2}{n} = 2M + B$. On the event E_1,

$$\mathbb{P}\left(\sup_{\beta \in E_1} |\nu_n(\beta, \beta^*)| \geqslant A10L\Phi(L\kappa_n)\sqrt{\frac{2\log(2G_n)}{n}}\right)$$

$$\leqslant \mathbb{P}\left(\sup_{\beta \in E_1} |\nu_n(\beta)| \geqslant A10L\Phi(L\kappa_n)\varepsilon_n\sqrt{\frac{2\log(2G_n)}{n}}\right)$$

$$\leqslant \mathbb{P}\left(\sup_{\beta \in E_1} |\nu_n(\beta)| \geqslant A5L\Phi(L(\varepsilon_n + B))\varepsilon_n\sqrt{\frac{2\log(2G_n)}{n}}\right)$$

given that $2M \geqslant \varepsilon_n$. From Lemma A.4 with $R = \varepsilon_n$ we deduce

$$\mathbb{P}\left(\sup_{\beta \in E_1} |\nu_n(\beta, \beta^*)| \geqslant A10L\Phi(L\kappa_n)\sqrt{\frac{2\log(2G_n)}{n}}\right)$$

$$\leqslant (2G_n)^{-A^2}. \tag{22}$$

On the event E_2, using the same type of argument as for (22) with $R = 2^j\varepsilon_n$ (given that $2M \geqslant 2^j\varepsilon_n$) for all $j = 1, ..., j_n$, we find

$$\mathbb{P}\left(\sup_{\beta \in E_2} |\nu_n(\beta, \beta^*)| \geqslant A10L\Phi(L\kappa_n)\sqrt{\frac{2\log(2G_n)}{n}}\right)$$

$$\leqslant j_n(2G_n)^{-A^2}.$$

Finally we have

$$\leqslant C^{'}(2G_n)^{-\frac{A^2}{2}} \tag{23}$$

where $C^{'}$ is a constant (because $j_n = \lfloor \log_2(nM) \rfloor + 1$ and $n \ll G_n$) and the result of Lemma A.3 follows from (23) and (22) with $C = 1 + C^{'}$. □

B Proof of Theorem 4.2

The main step of the proof are the same as for the Lasso.

Proof. By the same arguments as the ones used to prove (6) we have

$$\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right) + 2r_n\|\hat{\beta}_n\|_1 + t_n\|\hat{\beta}_n\|_2^2$$

$$\leqslant (\mathbb{P}_n - \mathbb{P})\left(l(\beta^*) - l(\hat{\beta}_n)\right) + 2r_n\|\beta\|_1 + t_n\|\beta^*\|_2^2. \tag{24}$$

The upper bound of $(\mathbb{P}_n - \mathbb{P})\left(l_l(\beta^*) - l_l(\hat{\beta}_n)\right)$, of $(\mathbb{P}_n - \mathbb{P})\left(l_\psi(\beta^*) - l_\psi(\hat{\beta}_n)\right)$ and the lower bound of $\mathbb{P}\left(l(\hat{\beta}_n) - l(\beta^*)\right)$ remains the same as those presented in the proof of Theorem 3.7 (see Proposition 3.3, Proposition 3.5 and Proposition A.1). Once these three propositions are proved, the rest of the

proof is similar to the one for logistic regression presented in [5]. On the event $\mathcal{A}\cap\mathcal{B}$ (which occurs with probability at least $1-2(2p)^{1-A^2}-C^{'}(2p)^{-A^2/2}\geqslant 1-C(2p)^{-A^2/2}$) by adding $r_n\|\hat{\beta}_n-\beta^*\|_1$ and $t_n\sum_{j\in I^*}(\beta_j^*-\hat{\beta}_j)^2$ to both sides of the inequality (24), we have

$$\begin{aligned}
&r_n\|\hat{\beta}_n-\beta^*\|_1+\mathbb{P}\left(l(\hat{\beta}_n)-l(\beta^*)\right)+t_n\sum_{j\in I^*}(\beta_j^*-\hat{\beta}_j)^2\\
&\leqslant 2r_n\|\hat{\beta}_n-\beta^*\|_1+2r_n\|\beta^*\|_1-2r_n\|\hat{\beta}_n\|_1+t_n\sum_{j\in I^*}(\beta_j^*-\hat{\beta}_j)^2\\
&\quad -t_n\|\hat{\beta}_n\|_2^2+t_n\|\beta^*\|_2^2+\frac{r_n}{2}\varepsilon_n \qquad (25)
\end{aligned}$$

with $\varepsilon_n=\dfrac{1}{n}$. On one hand, by the same argument as for (7) we have

$$2r_n\|\hat{\beta}_n-\beta^*\|_1+2r_n\|\beta^*\|_1-2r_n\|\hat{\beta}_n\|_1\leqslant 4r_n\sum_{j\in I^*}|\beta_j^*-\hat{\beta}_j|$$

and on the other hand

$$\begin{aligned}
&t_n\sum_{j\in I^*}(\beta_j^*-\hat{\beta}_j)^2-t_n\|\hat{\beta}_n\|_2^2+t_n\|\beta^*\|_2^2\\
&\leqslant 2t_n\sum_{j\in I^*}\beta_j^{*2}-2t_n\sum_{j\in I^*}\beta_j^*\hat{\beta}_j\\
&\leqslant r_n\sum_{j\in I^*}|\beta_j^*-\hat{\beta}_j|.
\end{aligned}$$

Therefore inequality (25) can be bounded by

$$\begin{aligned}
&r_n\|\hat{\beta}-\beta^*\|_1+\mathbb{P}\left(l(\hat{\beta})-l(\beta^*)\right)+t_n\sum_{j\in I^*}(\beta_j^*-\hat{\beta}_j)^2\\
&\leqslant 4r_n\sum_{j\in I^*}|\beta_j^*-\hat{\beta}_j|+r_n\sum_{j\in I^*}|\beta_j^*-\hat{\beta}_j|+\frac{r_n}{2}\varepsilon_n.
\end{aligned}$$

Since $\mathbb{P}\left(l(\hat{\beta})-l(\beta^*)\right)\geqslant 0$ we have

$$r_n\|\hat{\beta}-\beta^*\|_1\leqslant 5r_n\sum_{j\in I^*}|\beta_j^*-\hat{\beta}_j|+\frac{r_n}{2}\varepsilon_n$$

and then

$$\sum_{j\in I^{*C}}|\beta_j^*-\hat{\beta}_j|\leqslant 4\sum_{j\in I^*}|\beta_j^*-\hat{\beta}_j|+\frac{\varepsilon_n}{2}.$$

Thus $(\hat{\beta}-\beta^*)\in St(4,\frac{\varepsilon_n}{2})$. Using Proposition(A.1) in the case of groups of size one we find

$$\begin{aligned}
&r_n\|\hat{\beta}-\beta^*\|_1+t_n\sum_{j\in I^*}(\beta_j^*-\hat{\beta}_j)^2+c_n\mathbb{E}\left(\hat{\beta}^TX-\beta^{*T}X\right)^2\\
&\leqslant 5r_n\sum_{j\in I^*}|\beta_j^*-\hat{\beta}_j|+\frac{r_n}{2}\varepsilon_n,
\end{aligned}$$

with $c_n:=\min_{\{|x|\leqslant L(9B+\frac{1}{n})\}\cap\Theta}\left\{\psi^{''}(x)\right\}$. The rest of the proof follows the guidelines of the proof of (12) and (13) and leads to

$$\|\hat{\beta}-\beta^*\|_1\leqslant\frac{(2.5)^2r_ns^*}{t_n+c_nk}+(1+\frac{1}{r_n})\frac{\varepsilon_n}{2}.$$

and

$$\mathbb{E}\left(\hat{\beta}^TX-\beta^{*T}X\right)^2\leqslant\frac{2(2.5)^2}{c_nk(t_n+c_nk)}r_n^2s^*+\frac{2r_n+3}{2nc_n}.$$

□

References

[1] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[2] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[3] P.O. Brown, D. Botstein, et al. Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21:33–37, 1999.

[4] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

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

[6] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[7] D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer, J.M. Trent, et al. Expression profiling using cdna microarrays. *Nature genetics*, 21:10–14, 1999.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[9] R.A. Heller, M. Schena, A. Chai, D. Shalon, T. Bedilion, J. Gilmore, D.E. Woolley, and R.W. Davis. Discovery and analysis of inflammatory disease-related genes using cdna microarrays. *Proceedings of the National Academy of Sciences*, 94(6):2150–2155, 1997.

[10] J. Katriel. On a generalized recurrence for bell numbers. *Journal of Integer Sequences*, 11(2):3, 2008.

[11] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[12] J.M. Loubes and S. Van De Geer. Adaptive estimation with soft thresholding penalties. *Statistica Neerlandica*, 56(4):453–478, 2002.

[13] K. Lounici, M. Pontil, S. Van De Geer, and A.B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[14] P. McCullagh and J.A. Nelder. Generalized linear models. monographs on statistics and applied probability 37. *Chapman Hall, London*, 1989.

[15] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[16] 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.

[17] P.J. Park, L. Tian, and I.S. Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18:S120–S127, 2002.

[18] J. Quackenbush. Microarray analysis and tumor classification. *New England Journal of Medicine*, 354(23):2463–2472, 2006.

[19] M.R. Segal, K.D. Dahlquist, and B.R. Conklin. Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980, 2003.

[20] B. Tarigan and S.A. Van De Geer. Classifiers of support vector machine type with\ ell1 complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006.

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

[22] S.A. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[23] A. Van der Vaart and J. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

[24] C. Wagschal. *Fonctions holomorphes, équations différentielles: exercices corrigés*. Hermann, 2003.

[25] F. Wei and J. Huang. Consistent group selection in high-dimensional linear regression. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(4):1369, 2010.

[26] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, J.R. Marks, and J.R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98(20):11462–11467, 2001.

[27] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2005.

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