# Non-asymptotic Oracle Inequalities for the Lasso and Group Lasso in high dimensional logistic model

Marius Kwemou

# Non-asymptotic Oracle Inequalities for the Lasso and Group Lasso in high dimensional logistic model

Marius Kwemou
Laboratoire Statistique et Génome UMR CNRS 8071- USC INRA, Université d'Évry Val d'Essonne, France
LERSTAD, Université Gaston Berger de Saint-Louis, Sénégal
Institut de Recherche pour le Développement, IRD* UMR 216, Paris
*e-mail*: marius.kwemou@genopole.cnrs.fr

**Abstract**

We consider the problem of estimating a function $f_0$ in logistic regression model. We propose to estimate this function $f_0$ by a sparse approximation build as a linear combination of elements of a given dictionary of $p$ functions. This sparse approximation is selected by the Lasso or Group Lasso procedure. In this context, we state non asymptotic oracle inequalities for Lasso and Group Lasso under restricted eigenvalues assumption as introduced in [3]. Those theoretical results are illustrated through a simulation study.

**keywords** Logistic model, Lasso, Group Lasso, High-dimensional.

## 1 Introduction

During the last few years, logistic regression problems with more and more high-dimensional data occur in a wide variety of scientific fields, especially in studies that attempt to find risk factors for disease and clinical outcomes. For example in gene expression data analysis or in genome wide association analysis the number $p$ of predictors may be of the same order or largely higher than the sample size $n$ (thousands $p$ of predictors for only a few dozens of individuals $n$, see for instance [9] or [28]). In this context the considered model is often what we call here "usual" logistic regression. It is given by

$$\mathbb{P}(Y_i = 1|Z = z_i) = \pi(z_i^T \beta_0) = \frac{\exp(z_i^T \beta_0)}{1 + \exp(z_i^T \beta_0)}, \tag{1}$$

where one observes $n$ independent and identically distributed (i.i.d) random couples $(Z_1, Y_1)$,..,$(Z_n, Y_n)$ from the joint distribution of $(Z, Y) \in \mathbb{R}^d \times \{0, 1\}$ and $\beta_0$ is the parameter to estimate.

*Research partially supported by IRD/DSF Département Soutien et Formation des communautés scientifiques du Sud.

In this paper, we consider a more general logistic model described by

$$\mathbb{P}(Y_i = 1|Z_i = z_i) = \frac{\exp(f_0(z_i))}{1 + \exp(f_0(z_i))}, \tag{2}$$

where $f_0$ (not necessarly linear) is an unknown function. We aim at estimating $f_0$ by constructing a suitable approximation. More precisely we estimate $f_0$ by a sparse approximation of linear combination of elements of a given dictionary of functions $\mathbb{D} = \{\phi_1, ..., \phi_p\}$: $\hat{f}(.) := \sum_{j=1}^{p} \hat{\beta}_j \phi_j(.)$. Our purpose expresses the belief that, in many instances, even if $p$ is large, only a subset of $\mathbb{D}$ may be needed to approximate $f_0$ well. This construction can be done by minimizing the empirical risk. However, it is well-known that with a large number of parameters in high dimensional data situations, direct minimization of empirical risk can lead to *Overfitting*: the classifier can only behave well in training set, and can be bad in test set. The procedure would also be unstable: since empirical risk is data dependent, hence random, small change in the data can lead to very different estimators. Penalization is used to overcome those drawbacks. One could use $\ell_0$ penalization, *ie* penalyze by the number of non zero coefficients (see for instance AIC, BIC [1, 23]). Such a penalization would produce interpretable models, but leads to non convex optimization and there is not efficient algorithm to solve this problem in high dimensional framework. Tibshirani [25] proposes to use $\ell_1$ penalization, which is a regularization technique for simultaneous estimation and selection. This penalization leads to convex optimization and is important from computational point of view (as well as from theoretical point of view). As the consequence of the optimality conditions, regularization by the $\ell_1$ penalty tends to produce some coefficients that are exactly zero and shrink others, thus the name of Lasso (Least Absolute Shrinkage and Selection Operator). There exit some algorithms to solve this convex problem, *glmnet* (see [8]), *predictor-corector* (see [22]) among the others.

A related Lasso-type procedure is the Group Lasso, where the covariates are assumed to be clustered in groups, and instead of $\ell_1$-penalty (summing the absolute values of each individual loading) the sum of Euclidean norms of the loadings in each group is used. It shares the same kind of properties as the Lasso, but encourages predictors to be selected in groups. This is useful when the set of predictors is partitioned into prescribed groups, only few being relevant in the estimation process. Group Lasso has numerous applications: when categorical predictors (factors) are present, the Lasso solution is not adequate since it only selects individual dummy variables instead of whole factors. In this case, categorical variables are usually represented as groups of dummy variables. In speech and signal processing for example, the groups may represent different frequency bands (see [16]).

**Previously known results.** Recently, a great deal of attention has been focused on $\ell_1$-penalized based estimators. Most of this attention concerns regression models and $\ell_1$-penalized least squares estimator of parameters in high dimensional linear and non linear additive regression. Among them one can cite [4, 6, 5, 15], who have studied the Lasso for linear model in nonparametric setting and proved sparsity oracle inequalities. Similar sparsity oracle inequalities are proved in [3], and those results hold under the so-called *restricted eigenvalues assumption* on the Gram matrix. Those kind of results have been recently

stated for the variants of the Lasso. For instance Lounici et al. [13] under a group version of *restricted eigenvalues assumption* stated oracle inequalities in linear gaussian noise model under Group sparsity. Those results lead to the refinements of their previous results for multi-task learning (see [12]). The behavior of the Lasso for linear regression regarding its selection and estimation properties have been studied in [11, 18, 31, 21, 30, 19] among other. Few results on the Lasso and Group Lasso concern logistic regression model. Most of them are asymptotic results and concern the "usual" logistic regression model defined by (1). Zou [32] shows consistency in variable selection for adaptive Lasso in generalized linear models when the number of covariables p is fixed. Huang et al. [10] prove sign consistency and estimation consistency for high-dimensional logistic regression. To our knowledge there are only two non asymptotic results for the Lasso in logistic model: the first one is from Bach [2], who provided bounds for excess risk (generalization performance) and estimation error in the case of "usual" logistic regression model under *restricted eigenvalues assumption* on the weighted Gram matrix. The second one is from van de Geer [26], who established non asymptotic oracle inequality for Lasso in high dimensional generalized linear models with Lipschitz loss functions. Meir et al. [17] shown consistency for the Group Lasso in "usual" logistic model.

In this paper, we state general non asymptotic oracle inequalities for the Lasso and Group Lasso in logistic model within the framework of high-dimensional statistics. We first state "slow" oracle inequalities (see Theorem 3.1 and Theorem 4.1) without any assumption on the Gram matrix. Secondly we provide "fast" oracle inequalities (see Theorem 3.2 and Theorem 4.2) under *restricted eigenvalues assumption* . In each case, we give, as a consequence, the bounds for excess risk, prediction and estimation errors for Lasso and Group Lasso in the "usual" logistic regression. Our non asymptotic results lead to an adaptive data-driven weighting of the $\ell_1$-norm. Simulation study is given to illustrate the numerical performance of Lasso with such weights.

This paper is organized as follows. In Section 2, we describe our estimation procedure base $\ell_1$-penalized logistic loss. In Section 3 and 4 we state our main results on oracle inequalities for Lasso and Group Lasso. In each case, we give as a consequence the bounds for excess risk, prediction and estimation errors for Lasso and Group Lasso in the "usual" logistic regression. Section 4.3 is devoted to simulation study. The proofs are gathered in Section 5 and Appendix.

## 2 $L_1$-penalized logistic regression

Our aim is to estimate $f_0$ in Model (2) by a linear combination of the function of a dictionary

$$\mathbb{D} = \{\phi_1, ..., \phi_p\},$$

where $\phi_j : \mathbb{R}^d \to \mathbb{R}$ and $p >> n$. The functions $\phi_j$ can be viewed as estimators of $f_0$ constructed from independent training sample, or estimators computed using $p$ different values of the tuning parameter of the same method. They can also be a collection of basis functions, that can approximate $f_0$, like wavelets,

splines, kernels, etc... We implicitly assume that $f_0$ can be well approximated by a linear combination

$$f_\beta(.) = \sum_{j=}^{p} \beta_j \phi_j(.),$$

where $\beta$ has to be estimated.

We consider empirical risk (logistic loss) for logistic model

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(f(z_i))) - Y_i f(z_i). \tag{3}$$

The $\ell_1$-penalized logistic regression estimator $f_{\hat{\beta}_L}$ is defined as a minimizer of the following $\ell_1$-penalized empirical risk

$$f_{\hat{\beta}_L} := \underset{f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \hat{R}(f_\beta) + \sum_{j=1}^{p} \omega_j |\beta_j| \right\}, \tag{4}$$

where the minimun is taken over the set

$$\Gamma \subseteq \left\{ f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.),\ \beta = (\beta_1, ..., \beta_p) \in \mathbb{R}^p \right\}$$

and $\omega_j$ are positive weights to be precised later. The "classical" Lasso penalization corresponds to $\omega_j = \lambda$, where $\lambda$ is the tuning parameter which makes balance between goodness-of-fit and sparsity. The Lasso estimator has the property that it does predictors selection and estimation at the same time. Indeed for large values of $\omega_j$, the related components $\hat{\beta}_j$ are set exactly to 0 and the other are shrunken toward zero.

## 2.1 Definitions and notation

Throughout the paper, we consider a fixed design setting (i.e $z_1, ..., z_n$ are consider determistic). Consider the matrix $X = (\phi_j(z_i))_{1\le i\le n,\ 1\le j\le p}$. For any $\beta = (\beta_1, ..., \beta_p)^T \in \mathbb{R}^p$, define $f_\beta(.) = \sum_{j=1}^{p} \beta_j \phi_j(.)$. Using this notation

$$(f_\beta(z_1), ..., f_\beta(z_n))^T = X\beta.$$

Let $K(\beta) = \{j \in \{1, ..., p\} : \beta_j \neq 0\}$ and $|K(\beta)|$ the cardinality of $K(\beta)$, which characterizes the sparsity of the vector $\beta$. For all $\delta \in \mathbb{R}^p$ and a subset $I \subset \{1, ..., p\}$, we denote by $\delta_I$ the vector in $\mathbb{R}^p$ that has the same coordinates as $\delta$ on $I$ and zero coordinates on the complement $I^c$ of $I$. For all $h, f, g : \mathbb{R}^d \to R$, we define the scalar products

$$\langle f, h\rangle_n = \frac{1}{n}\sum_{i=1}^{n} h(z_i) f(z_i),$$

and

$$\langle f, h\rangle_g = \frac{1}{n}\sum_{i=1}^{n} h(z_i) f(z_i) \pi(g(z_i))(1-\pi(g(z_i))), \quad \text{where} \quad \pi(t) = \frac{\exp(t)}{1+\exp(t)}.$$

We use the notation

$$q_f(h) = \frac{1}{n}\sum_{i=1}^{n} h(z_i)(Y_i - \pi(f(z_i))),$$

$\|h\|_\infty = \max_i |h(z_i)|$ and $\|h\|_n = \sqrt{\langle h, h\rangle_n} = \sqrt{\frac{1}{n}\sum_{i=1}^n h^2(z_i)}$ which denote the empirical norm. We denote by $R$ the conditionnal expectation of $\hat{R}$ given $z_1, ..., z_n$, *ie*

$$R(f) = I\!E(\hat{R}(f)) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(f(z_i))) - I\!E(Y_i) f(z_i).$$

It is clear that $R(.)$ is a convex function and $f_0$ is a minimum of $R(.)$ when the model is well-specified (*ie* when (2) is satisfied). Note that with our notations

$$R(f) = I\!E(\hat{R}(f)) = \hat{R}(f) + q_{f_0}(f). \tag{5}$$

We shall use both the excess risk of $f_{\hat{\beta}}$, $R(f_{\hat{\beta}}) - R(f_0)$ and the prediction loss $\|f_{\hat{\beta}} - f_0\|_n^2$ to evaluate the quality of the estimator. Note that $R(f_{\hat{\beta}})$ corresponds to the average Kullback-Leibler divergence to the best model when the model is well-specified, and is common for the study of logistic regression.

# 3 Oracle inequalities for the Lasso

Consider the following assumptions:

There exists a constant $0 < c_1 < \infty$ such that $\max_{1\le i\le n} |f_0(z_i)| \le c_1$. **($\mathbf{B_1}$)**

There exists a constant $c_2 > 0$ such that $\max_{1\le i\le n} \max_{1\le j\le p} |\phi_j(z_i)| < c_2$. **($\mathbf{B_2}$)**

For all $f_\beta \in \Gamma$, there is some universal constant $C_0$ such that $\max_{1\le i\le n} |f_\beta(z_i)| \le C_0$. **($\mathbf{B_3}$)**

Assumptions ($\mathbf{B_1}$) and ($\mathbf{B_3}$) are technical Assumptions useful to connect the excess risk and the prediction loss (see Lemma 5.1 and 5.2). An assumption similar to ($\mathbf{B_1}$) have been used in [7] to prove oracle inequality in gaussian regression model. The same kind of assumption as ($\mathbf{B_3}$) have been made in [24] to prove oracle inequality for support vector machine type with $\ell_1$ complexity regularization.

**Theorem 3.1.** *Let $f_{\hat{\beta}_L}$ be the $\ell_1$-penalized minimum defined in (4).*

*A-) Let $x > 0$ be fixed. For $j = \{1, ..., p\}$, let*

$$\omega_j = \frac{2}{n}\sqrt{\frac{1}{2}\sum_{i=1}^{n} \phi_j^2(z_i)(x + \log p)} + \frac{2c_2(x + \log p)}{3n}. \tag{6}$$

*Thus with probability at least $1 - \exp(-x)$ we have*

$$R(f_{\hat{\beta}_L}) - R(f_0) \le \inf_{\beta\in\mathbb{R}^p} \left\{ R(f_\beta) - R(f_0) + 2\|\beta\|_1 \max_{1\le j\le p} \omega_j \right\}.$$

*B-) Let $A > 2\sqrt{c_2}$. For $j = \{1, ..., p\}$, let*

$$w_j = \lambda = A\sqrt{\frac{\log p}{n}}.$$

*Thus with probability at least $1 - 2p^{1-A^2/4c_2}$ we have*

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^p} \left\{ R(f_\beta) - R(f_0) + 2A\|\beta\|_1 \sqrt{\frac{\log p}{n}} \right\}.$$

In each inequality the first part of the right hand corresponds to the approximation error (bias). The selection of the dictionary can be very important to minimize this approximation error. It is recommended to choose a dictionary $\mathbb{D}$ as $f_0$ could well be approximated by a linear combination of the functions of $\mathbb{D}$. The second part of the right hand in each inequality is the variance term and is usually refered as the rate of the oracle inequality. In Theorem 3.1, we speak about "slow" oracle inequalities, with the rate at the order $\|\beta\|_1\sqrt{\log p/n}$ for any $\beta$. Those results are obtained without any assumption on the Gram matrix ($\Phi_n = X^T X/n$). In order to obtain oracle inequalities with a fast rate of order $\log p/n$ we need additional assumption on the restricted eigenvalues of the Gram matrix, namely the *restricted eigenvalues assumption*.

For some integer $s$ such that $1 \leq s \leq p$ and a positive number $a_0$, the following condition holds ($\mathbf{RE_1}$)

$$\mu(s, a_0) := \min_{K \subseteq \{1,...p\}: |K| \leq s} \quad \min_{\Delta \neq 0: \|\Delta_{K^c}\|_1 \leq a_0 \|\Delta_K\|_1} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0.$$

This assumption has been introduced in [3], where several sufficient conditions for this assumption are described. This condition is known to be one of the weakest to derive fast rates for the Lasso. For instance conditions on the Gram matrix used to prove oracle inequality in [4, 6, 5] are more restrictive than *restricted eigenvalues assumption*. In those papers either $\Phi_n$ is positive definite, or mutual coherence condition is imposed. We refer to [27] for a complete comparison of the assumptions used to prove oracle inequality for the Lasso. Especially it is proved that *restricted eigenvalues assumption* is weaker than the neighborhood stability or irrepresentable condition. To emphasize the dependency of Assumption ($\mathbf{RE_1}$) on $s$ and $a_0$ we will sometimes refer to it as $RE(s, a_0)$.

**Theorem 3.2.** *Let $f_{\hat{\beta}_L}$ be the $\ell_1$-penalized minimum defined in (4). Fix $\eta > 0$ and $1 \leq s \leq p$. Assume that Assumptions ($\mathbf{B_1}$), ($\mathbf{B_2}$), ($\mathbf{B_3}$) and ($\mathbf{RE_1}$) are satisfied, with $a_0 = 3 + 4/\eta$.*

*A-) Let $x > 0$ be fixed. For $j = \{1, ..., p\}$, $\omega_j$ defined as in (6). Thus with probability at least $1 - \exp(-x)$ we have*

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma} \left\{ R(f_\beta) - R(f_0) + \frac{c(\eta)|K(\beta)| \left( \max_{1 \leq j \leq p} \omega_j \right)^2}{c_0 \epsilon_0 \mu^2(s, 3 + 4/\eta)} \right\}, \tag{7}$$

*and*

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)|K(\beta)| \left( \max_{1\leq j\leq p} \omega_j \right)^2}{c_0' c_0 \epsilon_0^2 \mu^2(s, 3+4/\eta)} \right\}. \tag{8}$$

*B-) Let $A > 2\sqrt{c_2}$. For $j = \{1, ..., p\}$, let*

$$w_j = \lambda = A\sqrt{\frac{\log p}{n}}.$$

*Thus with probability at least $1 - 2p^{1-A^2/4c_2}$ we have*

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma} \left\{ R(f_\beta) - R(f_0) + \frac{A^2 c(\eta)}{c_0 \epsilon_0 \mu^2(s, 3+4/\eta)} \frac{|K(\beta)| \log p}{n} \right\}, \tag{9}$$

*and*

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)A^2}{c_0' c_0 \epsilon_0^2 \mu^2(s, 3+4/\eta)} \frac{|K(\beta)| \log p}{n} \right\}. \tag{10}$$

*In both cases $c(\eta)$ is a constant depending only on $\eta$; $c_0 = c_0(C_0, c_1)$ and $c_0' = c_0'(C_0, c_1)$ are constants depending on $C_0$ and $c_1$; and $\epsilon_0 = \epsilon_0(c_1)$ is a constant depending on $c_1$.*

In this theorem the variance term is always of order $|K(\beta)| \log p/n$. Such order in sparse oracle inequalities usually refer to "fast rate". This rate is of same kind of the one obtain in [3] for linear regression model. For the best of our knowledge this is a first non asymptotic oracle inequality for prediction loss in logistic model. Some non asymptotic oracle inequality for excess risk has been established in [26] under different assumptions. Indeed, she stated oracle inequality for high dimensional generalized linear model with Lipschitz loss function, where logistic regression is a particular case. Her result assumes to be hold in the "neighborhood" of the target function, while our result is true for all bounded functions. Note also that our result holds under *RE* condition and we use more explicit weight (or parameter $\lambda$) on the penalty term.

**Remark 3.1.** *It is important to note that our results remain true if we assume that we are in the "neighborhood" of the target function. It suffices to note that if we suppose that there exists $\zeta$ such that $\max_{1\leq i\leq n} |f_\beta(z_i) - f_0(z_i)| \leq \zeta$, then Lemma 5.1 is still true.*

**Remark 3.2.** *The choice of the weights $\omega_j$ in part A of the Theorem 3.2 comes from Bernstein's inequality. It is important to note that we could have used the following weights*

$$\omega_j' = \frac{2}{n}\sqrt{2\sum_{i=1}^n I\!E(\phi_j^2(z_i)\epsilon_i^2)(x + \log p)} + \frac{2\max_{1\leq i\leq n} |\phi_j(z_i)|(x + \log p)}{3n}.$$

*Our results still hold with such weights $\omega_j'$. But these weights depend on the unknown function $f_0$ to be estimated. Indeed $I\!E(\epsilon_i^2) = \pi(f_0(z_i)(1 - \pi(f_0(z_i))$, that is the reason for using weights $\omega_j$ slightly greater than $\omega_j'$.*

### 3.1 Special case: variable selection in logistic regression model

In this section we assume that $f_0$ is a linear function that is $f_0(z_i) = f_{\beta_0}(z_i) = \sum_{j=1}^{p} \beta_{0j} z_{ij} = z_i^T \beta_0$, where $z_i = (z_{i1}, ..., z_{ip})^T$. Denote $X = (z_{ij})_{1 \le i \le n, 1 \le j \le p}$ the design matrix. Thus for $i = 1, ..., n$

$$\mathbb{P}(Y_i = 1/Z = z_i) = \pi(z_i^T \beta_0) = \frac{\exp(z_i^T \beta_0)}{1 + \exp(z_i^T \beta_0)}. \tag{11}$$

This corresponds to the usual logistic regression *ie* logistic model that allows linear dependency between $Z_i$ and the distribution of $Y_i$. The Lasso estimator of $\beta_0$ is thus defined as

$$\hat{\beta}_L := \underset{\beta:\ f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + \exp(z_i^T \beta)) - Y_i z_i^T \beta \right\} + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\}. \tag{12}$$

This infimum is achieved and might not be unique. When the design matrix $X$ has full rank, the solution of optimization problem 12 is usually unique.

**Corollary 3.1.** *Let assumption RE(s,3) be satisfied and $|K(\beta_0)| \le s$, where $1 \le s \le p$. Consider the Lasso estimator $f_{\hat{\beta}_L}$ defined by (12) with*

$$\omega_j = \frac{2}{n} \sqrt{\frac{1}{2} \sum_{i=1}^{n} z_{ij}^2 (x + \log p)} + \frac{2c_2(x + \log p)}{3n}$$

*Under the assumptions of Theorem 3.2 with probability at least $1 - \exp(-x)$ we have*

$$R(f_{\hat{\beta}_L}) - R(f_{\beta_0}) \le \frac{8s \left( \max_{1 \le j \le p} \omega_j \right)^2}{\mu^2(s,3) c_0 \epsilon_0} \tag{13}$$

$$\|f_{\hat{\beta}_L} - f_{\beta_0}\|_n^2 \le \frac{8s \left( \max_{1 \le j \le p} \omega_j \right)^2}{\mu^2(s,3) c_0^2 \epsilon_0^2} \tag{14}$$

$$\|\hat{\beta}_L - \beta_0\|_1 \le \frac{8\sqrt{2}s \left( \max_{1 \le j \le p} \omega_j \right)}{\mu^2(s,3) c_0 \epsilon_0} \tag{15}$$

$$\|\hat{\beta}_L - \beta_0\|_q^q \le \left( \frac{8\sqrt{2}s \left( \max_{1 \le j \le p} \omega_j \right)}{\mu^2(s,3) c_0 \epsilon_0} \right)^q \quad \text{for all} \quad 1 < q \le 2. \tag{16}$$

*If $\omega_j = \lambda = A\sqrt{\log p / n}$ we have the same results with probability at least $1 - 2p^{1 - A^2/4c_2}$.*

The first and third line of the corollary are similar to those of Theorem 5 in [2] except that we assume that $f_{\hat{\beta}_L}$ is bounded whereas he assume that $\lambda$ is bounded. Note that, up to differences in constant factors,

the rates obtained in this corollarly are the same as those obtained in Theorem 7.2 in [3] for linear model with an s-sparse vector.

**Remark 3.3.** *In logistic regression model (11), if vector $\beta_0$ is sparse, ie $|K(\beta_0)| \leq s$, then Assumption ($\mathbf{RE_1}$) implies that $\beta_0$ is uniquely defined. Indeed, if there exists $\beta^*$ such that for $i = 1, ..., n$, $\pi(z_i^T\beta_0) = \pi(z_i^T\beta^*)$, it follows that $X\beta_0 = X\beta^*$ and $|K(\beta^*)| \leq s$. Then according to assumption $RE(s, a_0)$ with $a_0 \geqslant 1$, we necessarily have $\beta_0 = \beta^*$. Indeed if $RE(s, a_0)$ is satisfied with $a_0 \geqslant 1$, then* $\min\{\|X\beta\|_2 : |K(\beta)| \leq 2s, \beta \neq 0\} > 0$.

# 4 Oracle inequalities for the Group Lasso

In this section we assume that prior knowledge is available. More precisely the set of relevents preditors have known group structure, and we wish to achieves sparsity at the level of groups. This group sparsity assumption suggests us to use the Group Lasso method. We consider the Group Lasso for logistic regression (see [17, 29]), where predictors are included or excluded in groups. The logistic Group Lasso is the minimizer of the following optimization problem, defined as

$$f_{\hat{\beta}_{GL}} := \underset{f_\beta \in \Gamma_1}{\operatorname{argmin}} \left\{ \hat{R}(f_\beta) + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 \right\}, \tag{17}$$

where $\omega_l > 0$ are the weights used to control the amount of penalization, and where

$$\Gamma_1 \subseteq \left\{ f_\beta(.) = \sum_{l=1}^{g} \sum_{j \in G_l} \beta_{lj}\phi_j(.),\ \beta \in \mathbb{R}^{\sum_{l=1}^{g}|G_l|} \right\}.$$

This penalty can be viewed as an intermediate betwen $\ell_1$ and $\ell_2$ type penalty. It has the attractive property that it does variables selection at the group level. Denote $X = (\phi_j(z_i))_{1\leq i\leq n, 1\leq j\in G_l,\ l=1,...,g}$ the design matrix and $\beta = (\beta^1, ..., \beta^g)^T$, where $\beta^l = (\beta_{lj})_{j\in G_l}$ for $l = 1, ..., g$. Let $f_\beta(.) = \sum_{l=1}^{g}\sum_{j\in G_l}\beta_{lj}\phi_j(.)$. With our notations

$$(f_\beta(z_1), ..., f_\beta(z_n))^T = X\beta.$$

We define the group norm of $\beta$ as

$$\|\beta\|_{2,1} = \sum_{l=1}^{g} \sqrt{\sum_{j\in G_l} \beta_{lj}^2} = \sum_{l=1}^{g} \|\beta^l\|_2.$$

Finally we set $J(\beta) = \{l : \beta^l \neq 0\}$, the set of relevant groups and $|J(\beta)|$ the number of such groups. Consider the following assumptions:

There exists a constant $a_2 > 0$ such that, $\max\limits_{1\leq i\leq n} \max\limits_{1\leq l\leq g} \max\limits_{j\in G_l} |\phi_j(z_i)| < a_2$. ($\mathbf{B_4}$)

For all $f_\beta \in \Gamma_1$, there is some universal constant $C_0$ such that, $\max_{1\le i\le n} |f_\beta(z_i)| \le C_0$. ($\mathbf{B_5}$)

Assumption ($\mathbf{B_4}$) could be viewed as a natural extension of Assumption $\mathbf{B_2}$ for the Lasso.

**Theorem 4.1.** *Let $f_{\hat\beta_{GL}}$ be the Group Lasso solution defined in (17) with*

$$\omega_l = \frac{2|G_l|}{n}\sqrt{\frac{1}{2}\max_{j\in G_l}\sum_{i=1}^n \phi_j^2(z_i)\left(x+\log\sum_{l=1}^g |G_l|\right)} + \frac{2a_2|G_l|}{3n}\left(x+\log\sum_{l=1}^g |G_l|\right) \tag{18}$$

*where $x>0$. Thus with probability at least $1-\exp(-x)$ we have*

$$R(f_{\hat\beta_{GL}}) - R(f_0) \le \inf_{\beta\in\mathbb{R}^{\sum_{l=1}^g |G_l|}} \left\{ R(f_\beta) - R(f_0) + 2\|\beta\|_{2,1}\max_{1\le l\le g}\omega_l \right\}.$$

As previously the weight being of order $\|\beta\|_{2,1}\sqrt{\log g/n}$ for any $\beta$, this is a "slow" oracle inequality. Those results are obtained without any assumption on the Gram matrix. To obtain oracle inequalities with a fast rate, of order $\log g/n$, we need some group version of the restricted eigenvalues of the Gram matrix.

For some integer $s$ such that $1\le s\le g$ and a positive number $a_0$, the following condition holds ($\mathbf{RE_2}$)

$$\mu_1(s,a_0) := \min_{K\subseteq\{1,\dots p\}:|K|\le s} \ \min_{\Delta\neq 0:\|\Delta_{K^c}\|_{2,1}\le a_0\|\Delta_K\|_{2,1}} \frac{\|X\Delta\|_2}{\sqrt{n}\|\Delta_K\|_2} > 0.$$

This is a natural extension to the Group Lasso of Assumption ($\mathbf{RE_1}$) for the usual Lasso. The only main difference lies on the set where the minimum is taken: for the Lasso the minimum is taken over $\{\Delta\neq 0 : \|\Delta_{K^c}\|_1 \le a_0\|\Delta_K\|_1\}$ whereas for the Group Lasso the minimun is over $\{\Delta\neq 0 : \|\Delta_{K^c}\|_{2,1} \le a_0\|\Delta_K\|_{2,1}\}$. This assumption has already been used in [12, 13] to prove oracle inequality for linear gaussian noise model under Group sparsity and for multi-task learning.

**Theorem 4.2.** *Let $f_{\hat\beta_{GL}}$ be the Group Lasso solution defined in (17) with $\omega_l$ defined as in (18). Fix $\eta>0$ and $1\le s\le g$, assume that Assumptions ($\mathbf{B_1}$), ($\mathbf{B_4}$), ($\mathbf{B_5}$) and ($\mathbf{RE_2}$) are satisfied, with $a_0 = 3+4/\eta$. Thus with probability at least $1-\exp(-x)$ we have*

$$R(f_{\hat\beta_{GL}}) - R(f_0) \le (1+\eta)\inf_{f_\beta\in\Gamma_1}\left\{ R(f_\beta) - R(f_0) + \frac{c(\eta)|J(\beta)|\left(\max_{1\le l\le g}\omega_l\right)^2}{c_0\epsilon_0\mu_1(s,3+4/\eta)^2}\right\},$$

*and*

$$\|f_{\hat\beta_{GL}} - f_0\|_n^2 \le \frac{c_0'}{4c_0\epsilon_0}(1+\eta)\inf_{f_\beta\in\Gamma_1}\left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)|J(\beta)|\left(\max_{1\le l\le g}\omega_l\right)^2}{c_0'c_0\epsilon_0^2\mu_1(s,3+4/\eta)^2}\right\},$$

*Where $c(\eta)$ is a constant depending only on $\eta$; $c_0 = c_0(C_0,c_1)$ and $c_0' = c_0'(C_0,c_1)$ are constants depending on $C_0$ and $c_1$; and $\epsilon_0 = \epsilon_0(c_1)$ is a constant depending on $c_1$.*

Once again the variance term are of order $\log g/n$, which refers to a"fast rate".

## 4.1 Special case: variable selection in the usual logistic model

In this section we assume that $f_0$ is a linear function *ie* $f_0(z_i) = f_{\beta_0}(z_i) = \sum_{l=1}^{g} \sum_{j \in G_l} \beta_j z_{ij}$. Denote by $X = (z_{ij})_{1 \leq i \leq n, 1 \leq j \in G_l,\ l=1,...,g}$, the design matrix. Let $z_i = (z_{ij})_{j \in G_l, l=1,...,g}$ be the ith row of the matrix $X$ and $z^{(j)} = (z_{1j}, ..., z_{nj})^T$ is jth column. For $i = 1, ..., n$

$$\mathbb{P}(Y_i = 1/Z = z_i) = \frac{\exp(z_i^T \beta_0)}{1 + \exp(z_i^T \beta_0)}. \tag{19}$$

The Group Lasso estimator of $\beta_0$ is defined by

$$\hat{\beta}_{GL} := \underset{\beta:\ f_\beta \in \Gamma_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + \exp(z_i^T \beta)) - Y_i z_i^T \beta \right\} + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2. \tag{20}$$

**Corollary 4.1.** *Let assumption RE(s,3) be satisfied and $|J(\beta_0)| \leq s$, where $1 \leq s \leq g$. Consider the Group Lasso estimator $f_{\hat{\beta}_{GL}}$ defined by (20) with*

$$\omega_l = \frac{2|G_l|}{n} \sqrt{\frac{1}{2} \max_{j \in G_l} \sum_{i=1}^{n} z_{ij}^2 \left( x + \log \sum_{l=1}^{g} |G_l| \right)} + \frac{2a_2|G_l|}{3n} \left( x + \log \sum_{l=1}^{g} |G_l| \right)$$

*where $x > 0$. Under the assumptions of Theorem 4.2, with probability at least $1 - \exp(-x)$ we have*

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \frac{8s \left( \max_{1 \leq l \leq g} \omega_l \right)^2}{\mu^2(s,3) c_0 \epsilon_0} \tag{21}$$

$$\|f_{\hat{\beta}_{GL}} - f_{\beta_0}\|_n^2 \leq \frac{8s \left( \max_{1 \leq l \leq g} \omega_l \right)^2}{\mu^2(s,3) c_0^2 \epsilon_0^2} \tag{22}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,1} \leq \frac{8\sqrt{2}s \left( \max_{1 \leq l \leq g} \omega_l \right)}{\sqrt{3}\mu^2(s,3) c_0 \epsilon_0} \tag{23}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,q}^q \leq \left( \frac{8\sqrt{2}s \left( \max_{1 \leq l \leq g} \omega_l \right)}{\sqrt{3}\mu^2(s,3) c_0 \epsilon_0} \right)^q \quad \text{for all} \ \ 1 < q \leq 2 \tag{24}$$

Remark 3.3 is still true in this case.

## 4.2 Non bounded functions

In this section we consider the Group Lasso estimator (20). We wish to establish the same result as Corollary 4.1 without assuming ($\mathbf{B_1}$) or ($\mathbf{B_5}$) *ie* neither $f_{\beta_0}$ nor $f_\beta$ is bounded. Let us consider the following

assumption.

For some integer $s$ such that $1 \leq s \leq g$ and a positive number $a_0$, the following condition holds **($\mathbf{RE_3}$)**

$$\mu_2(s, a_0) := \min_{K \subseteq \{1,\dots p\}: |K| \leq s} \quad \min_{\Delta \neq 0: \|\Delta_{K^c}\|_{2,1} \leq a_0 \|\Delta_K\|_{2,1}} \frac{\Delta^T X^T D X \Delta}{n \|\Delta_K\|_2^2} > 0,$$

where $D = \mathrm{Diag}\left(\mathrm{var}(Y_i)\right)$.

This is an extension of the Assumption **$\mathbf{RE_2}$** to the weighted Gram matrix $X^T D X / n$.

**Theorem 4.3.** *Consider the Group Lasso estimator $f_{\hat{\beta}_{GL}}$ defined by (20) with*

$$\omega_l = \frac{2|G_l|}{n} \sqrt{\frac{1}{2} \max_{j \in G_l} \sum_{i=1}^n z_{ij}^2 \left( x + \log \sum_{l=1}^g |G_l| \right)} + \frac{2 a_2 |G_l|}{3n} \left( x + \log \sum_{l=1}^g |G_l| \right),$$

*where $x > 0$. Set $v = \max_{1 \leq i \leq n} \max_{1 \leq l \leq g} \|z_i^l\|_2$. Let Assumptions (**$\mathbf{B_4}$**) and (**$\mathbf{RE_3}$**) be satisfied with*

$$a_0 = \frac{3 \max\limits_{1 \leq l \leq g} \omega_l}{\min\limits_{1 \leq l \leq g} \omega_l}.$$

*If $\max\limits_{1 \leq l \leq g} \omega_l \leq \mu_2 / 48 v |J(\beta_0)|$, with probability at least $1 - \exp(-x)$ we have*

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \frac{144 |J(\beta_0)| \left( \max\limits_{1 \leq l \leq g} \omega_l \right)^2}{\mu_2^2(s, 3)} \tag{25}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,1} \leq \frac{96 |J(\beta_0)| \left( \max\limits_{1 \leq l \leq g} \omega_l \right)}{\mu_2^2(s, 3)} \tag{26}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,q}^q \leq \left( \frac{96 |J(\beta_0)| \left( \max\limits_{1 \leq l \leq g} \omega_l \right)}{\mu_2^2(s, 3)} \right)^q \quad \text{for all} \quad 1 < q \leq 2. \tag{27}$$

*Moreover if we assume that there exists $0 < \epsilon_0 \leq 1/2$ such that*

$$\epsilon_0 \leq \pi(f_{\beta_0}(z_i))(1 - \pi(f_{\beta_0}(z_i))) \qquad \text{for all} \quad i = 1, \dots, n$$

*then,*

$$\|X \hat{\beta}_{GL} - X \beta_0\|_n^2 \leq \frac{144 |J(\beta_0)| \left( \max\limits_{1 \leq l \leq g} \omega_l \right)^2}{\mu^2(s, 3) \epsilon_0}. \tag{28}$$

### 4.3 Simulation study

To illustrate the theoretical part of this paper we provide in this section some experimental results on simulated data. Our aim is to compare the Lasso using the weights we proposed to the "classical" Lasso (without weights). We consider the Lasso estimator defined as

$$\hat{\beta}_L := \underset{\beta:\ f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1+\exp(z_i^T\beta)) - Y_i z_i^T \beta \right\} + \lambda \sum_{j=1}^{p} \omega_j |\beta_j| \right\}, \tag{29}$$

where $\lambda$ is a tuning parameter. Note that this is equivalent to the Lasso estimator defined in (12). For all the simulations $\lambda$ will be estimated by cross-validation. According to Remark 3.2 we will use the weights $\omega_j'$ which seem more realistic than $\omega_j$. We will proceed in two steps: (i) estimate the variance $I\!E(\epsilon_i^2)$, $\hat{\sigma}_i^2$; (ii) for each $j \in \{1, ..., p\}$, calculate the estimator $\hat{\omega}_j'$ of the weight $\omega_j'$, by replacing in $\omega_j'$ the variance $I\!E(\epsilon_i^2)$ by is estimator calculated in step (i). The great challenge is to find an estimator of $I\!E(\epsilon_i^2)$. We propose two estimators of $I\!E(\epsilon_i^2) = \pi(z_i^T\beta_0)(1-\pi(z_i^T\beta_0)$:

1. estimate by $\hat{\sigma}_i^2 = \pi(z_i^T\hat{\beta}_L)(1-\pi(z_i^T\hat{\beta}_L)$ where $\hat{\beta}_L$ is the "classical" Lasso estimator of $\beta_0$ (without weight);

2. the second estimator is $\hat{\sigma}_i^2 = \pi(z_i^T\hat{\beta}_{Logit})(1-\pi(z_i^T\hat{\beta}_{Logit})$, where $\hat{\beta}_{Logit}$ is an estimator of $\beta_0$ obtained after successively using the Lasso to screen coefficients and a logistic model which take into account coefficients different to zero in the Lasso.

*Data generation.* We simulate 500 datasets consisting of $n$ observations from logistic model (11) with $\beta_0 = (1.5, -1, 2, 0, 0, ..., 0) \in \mathbb{R}^p$ where $p = |K(\beta_0)| + k$, $(|K(\beta_0)| = 3)$. This corresponds to the model with $k$ null coefficients, where $z_i$ are standard normal. For each data set we calculate the prediction error, estimation error, the rate of true selection and the rate of false relevant or irrelevant coefficient. The true selection is when the procedure select exactly the true relevent coefficients. The rate of relevant or irrelevant coefficient is the rate of bad selection in the estimation (the procedure declares relevant yet it is irrelevant or declares irrelevant yet it is relevant ). The results for the four methods are presented in the Figure 2, Figure 3 and Figure 4. Lasso represents the "classical" Lasso (without weight); weight.Logit is the method (29) with weights estimated using procedure (2); weight.Lasso is the method (29) with weights estimated by the procedure (1); weight.theoretical is method (29) with theoretical weights.

### 4.4 Comments

For estimation or prediction error the performance of all the methods are almost the same. When the number of sample $n$ increases, the performance of all the methods also increases. The strength of the methods decreases with the number $k$ of null coefficients. The real difference is in rate of true selection

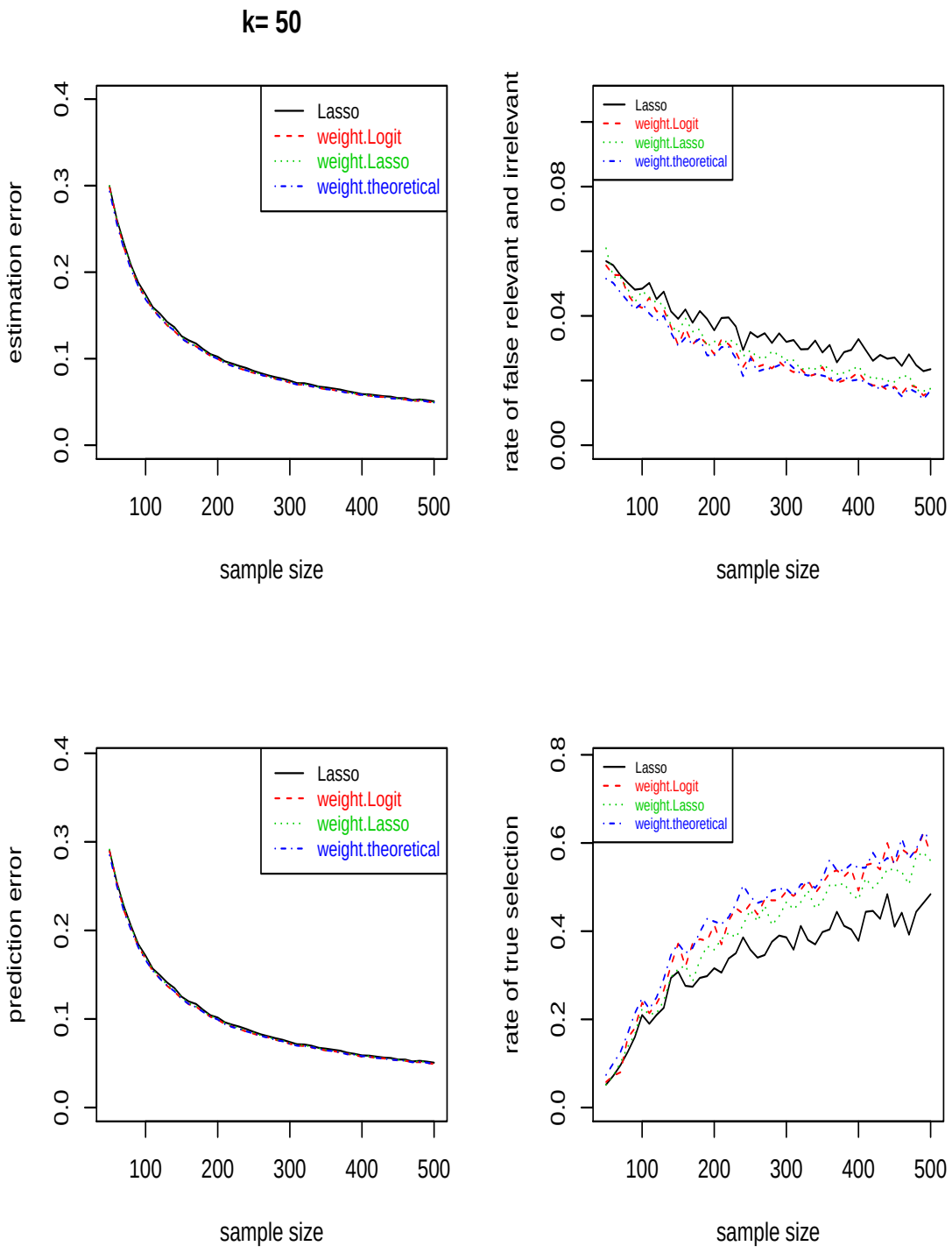




Figure 1: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficient. k=50 from the setup described in Section 4.3*

and the rate of false relevant and irrelevant where the weight.theoretical, weight.Logit and weight.Lasso outperform the Lasso. And weight.Logit seem to be better than weight.Lasso.


## Acknowledgements

We would like to thank Marie-Luce Taupin for the careful reading of the manuscript and for her helpful comments . We also thank Sarah Lemler for helpful discussions.


# 5 Proofs of main results

## 5.1 Proof of Theorem 3.1

Fix an arbitrary $\beta \in \mathbb{R}^p$ and set $\delta = W(\hat{\beta}_L - \beta)$, where $W = \mathrm{Diag}(w_1, ..., w_p)$. Short calculation shows that

$$R(f_\beta) = \hat{R}(f_\beta) + \frac{1}{n}\varepsilon^T X\beta$$

where $\varepsilon_i = Y_i - I\!E(Y_i)$, for $i \in 1, ..., n$. Since $\hat{\beta}_L$ is the minimizer of $\hat{R}(f_\beta) + \sum_{j=1}^p \omega_j|\beta_j|$,

$$R(f_{\hat{\beta}_L}) - \frac{1}{n}\varepsilon^T X\hat{\beta}_L + \sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j}| \leq R(f_\beta) - \frac{1}{n}\varepsilon^T X\beta + \sum_{j=1}^p \omega_j|\beta_j|,$$

thus

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq R(f_\beta) - R(f_0) + \frac{1}{n}\varepsilon^T X(\hat{\beta}_L - \beta) + \sum_{j=1}^p \omega_j|\beta_j| - \sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j}|.$$

This implies that

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq R(f_\beta) - R(f_0) + \sum_{j=1}^p \left|\frac{1}{n}\sum_{i=1}^n \phi_j(z_i)\varepsilon_i\right| |\hat{\beta}_{\lambda,j} - \beta_j| + \sum_{j=1}^p \omega_j|\beta_j| - \sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j}|.$$

For $1 \leq j \leq p$, set $S_j = \sum_{i=1}^n \phi_j(z_i)\varepsilon_i$ and let us denote by $E$, the event

$$E = \bigcap_{j=1}^p \{|S_j| \leq n\omega_j/2\}.$$

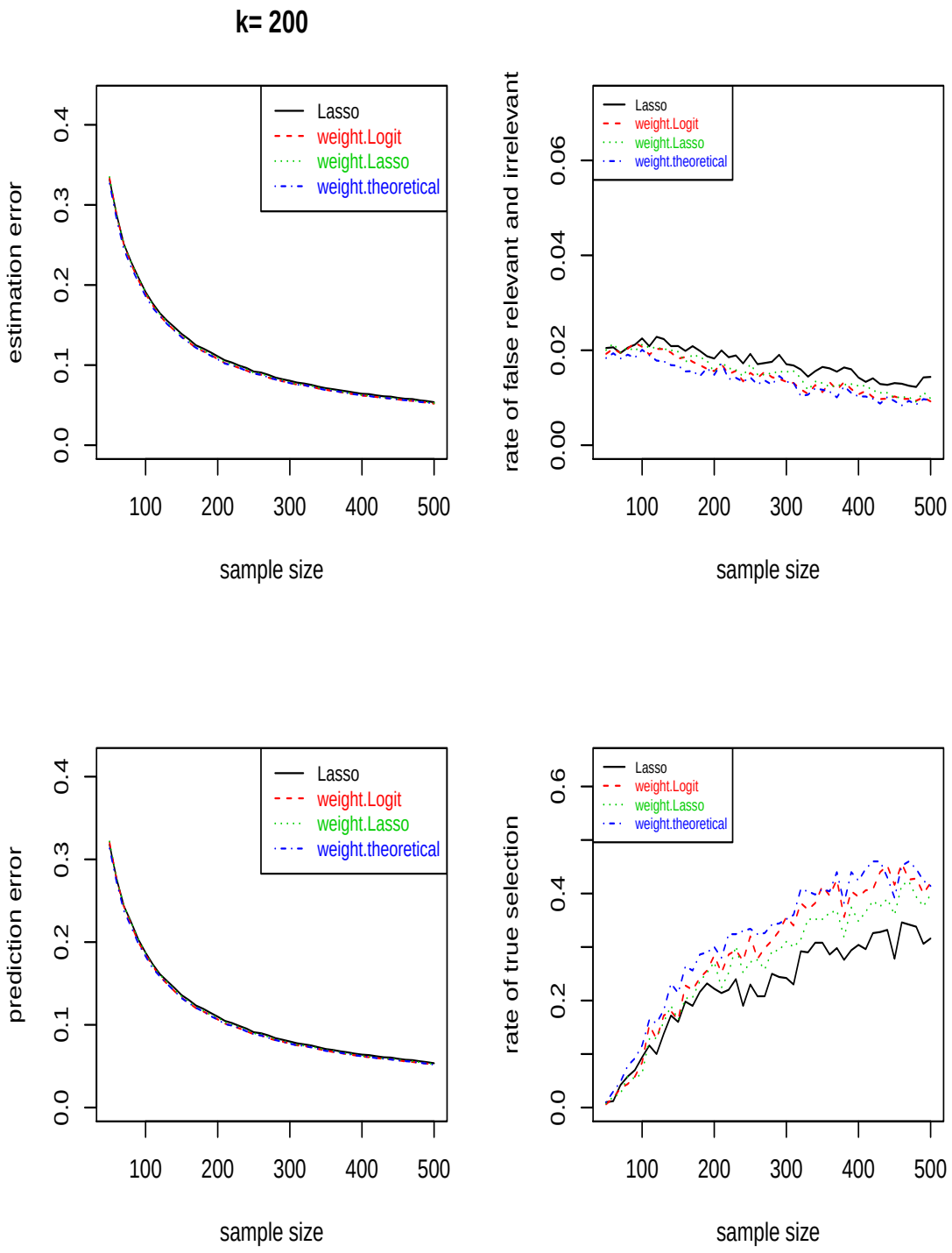




Figure 2: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficient. k=200 from the setup described in Section 4.3*

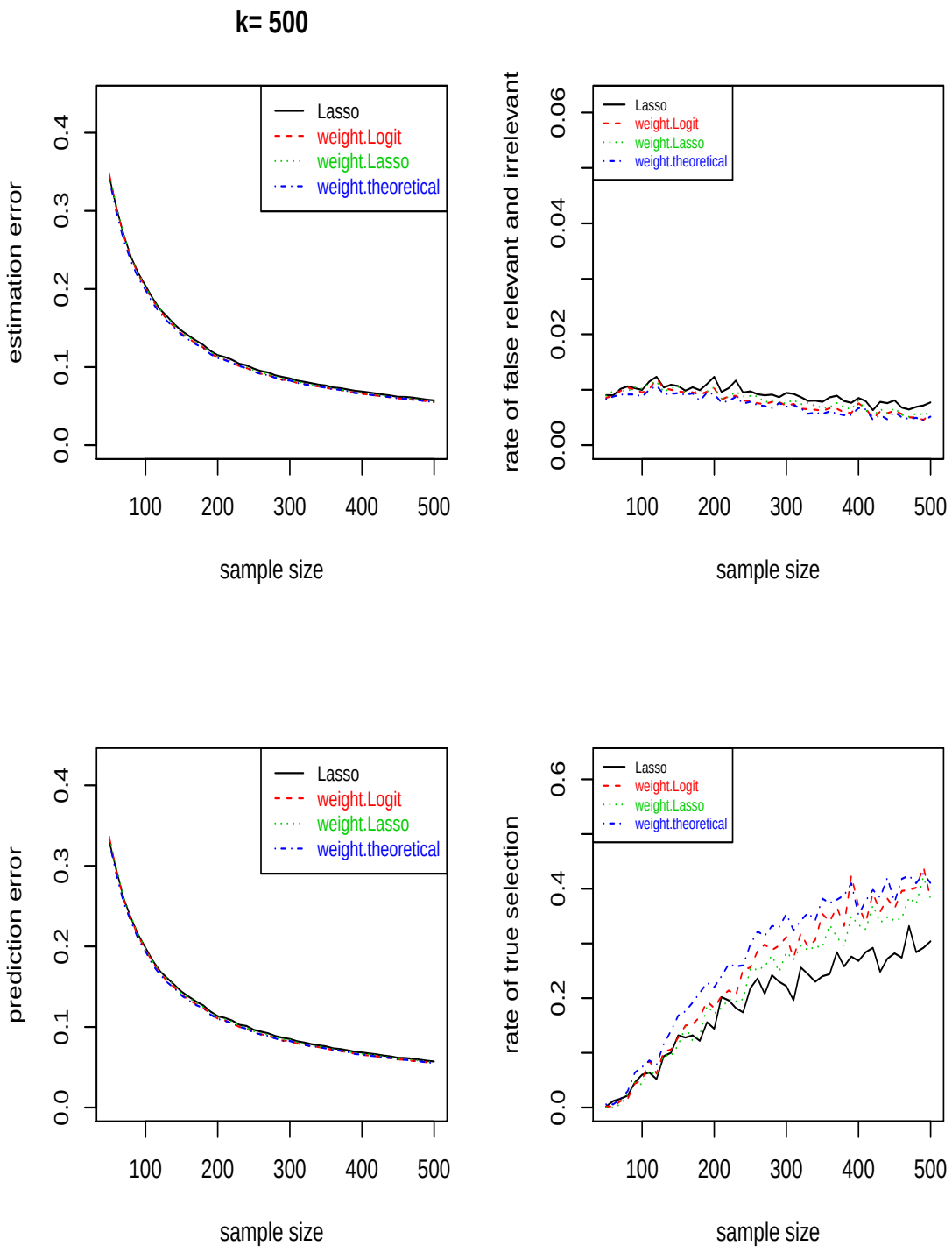




Figure 3: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficient. k=500 from the setup described in Section 4.3*

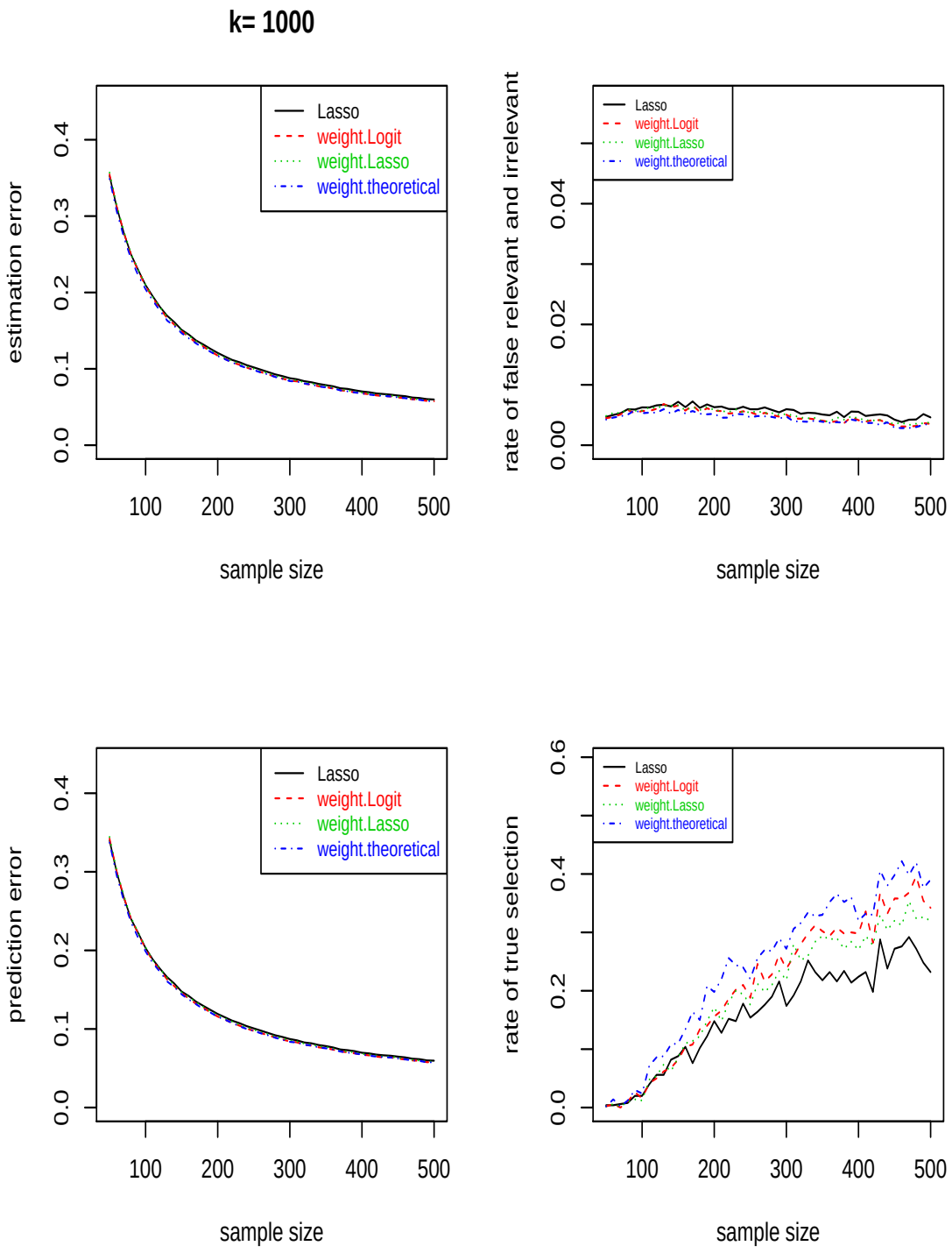




Figure 4: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficient. k=1000 from the setup described in Section 4.3*

We state the results on the event $E$ and then find an upper bound of $\mathbb{P}(E^c)$.

**On the event** $E$:

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq R(f_\beta) - R(f_0) + \sum_{j=1}^{p} \omega_j |\hat{\beta}_{\lambda,j} - \beta_j| + \sum_{j=1}^{p} \omega_j |\beta_j| - \sum_{j=1}^{p} \omega_j |\hat{\beta}_{\lambda,j}|$$
$$\leq R(f_\beta) - R(f_0) + 2\sum_{j=1}^{p} \omega_j |\beta_j|.$$

We conclude that on the event $E$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^p} \left\{ R(f_\beta) - R(f_0) + 2\|\beta\|_1 \max_{1 \leq j \leq p} \omega_j \right\}.$$

Now we are going to find an upper bound of $\mathbb{P}(E^c)$ :

$$\mathbb{P}(E^c) \leq \mathbb{P}\left( \bigcup_{j=1}^{p} \{ |\sum_{i=1}^{n} \phi_j(z_i)(Y_i - I\!E(Y_i))| > \omega_j n/2 \} \right)$$
$$\leq \sum_{j=1}^{p} \mathbb{P}(|\sum_{i=1}^{n} \phi_j(z_i)(Y_i - I\!E(Y_i))| > \omega_j n/2).$$

For $j \in \{1, ..., p\}$, set $v_j = \sum_{i=1}^{n} I\!E(\phi_j^2 \epsilon_i^2)$. Since $\sum_{i=1}^{n} \phi_j^2(z_i) \geqslant 4v_j$, we have

$$\mathbb{P}(|S_j| > n\omega_j/2) \leq \mathbb{P}\left( |S_j| > \sqrt{2v_j(x + \log p)} + \frac{c_2}{3}(x + \log p) \right).$$

By applying Bernstein's inequality (see Lemma A.4) to the right hand side of the previous inequality we get

$$\mathbb{P}(|S_j| > n\omega_j/2) \leq 2\exp(-x - \log p).$$

It follows that

$$\mathbb{P}(E^c) \leq \sum_{j=1}^{p} \mathbb{P}(|S_j| > \omega_j n/2) \leq \exp(-x). \tag{30}$$

When $\omega_j = \lambda$, for all $j \in \{1, ..., p, \}$ we apply Hoeffding's inequality (see Lemma A.5). This leads to

$$\mathbb{P}(E^c) = \mathbb{P}\left( \bigcup_{j=1}^{p} \{ |\sum_{i=1}^{n} \phi_j(z_i)(Y_i - I\!E(Y_i))| > \lambda n/2 \} \right)$$
$$\leq \sum_{j=1}^{p} \mathbb{P}(|\sum_{i=1}^{n} \phi_j(z_i)(Y_i - I\!E(Y_i))| > \lambda n/2)$$
$$\leq 2p \exp\left( -\frac{2(\lambda n/2)^2}{\sum_{i=1}^{n} 2c_2} \right) = 2p \exp\left( -\frac{\lambda^2 n}{4c_2} \right) = 2p^{1 - \frac{A^2}{4c_2}}. \tag{31}$$

This ends the proof of Theorem 3.1. ■

## 5.2 Proof of Theorem 3.2

Fix an arbitrary $\beta \in \mathbb{R}^p$ such that $f_\beta \in \mathit{\Gamma}$, and set $\delta = W(\hat{\beta}_L - \beta)$, where $W = \mathrm{Diag}(w_1, ..., w_p)$. Short calculation shows that

$$R(f_\beta) = \hat{R}(f_\beta) + \frac{1}{n}\varepsilon^T X\beta$$

where $\varepsilon_i = Y_i - I\!E(Y_i)$, for $i \in 1, ..., n$. Since $\hat{\beta}_L$ is the minimizer of $\hat{R}(f_\beta) + \sum_{j=1}^p \omega_j|\beta_j|$,

$$R(f_{\hat{\beta}_L}) - \frac{1}{n}\varepsilon^T X\hat{\beta}_L + \sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j}| \leq R(f_\beta) - \frac{1}{n}\varepsilon^T X\beta + \sum_{j=1}^p \omega_j|\beta_j|,$$

then

$$R(f_{\hat{\beta}_L}) \leq R(f_\beta) + \sum_{j=1}^p \left|\frac{1}{n}\sum_{i=1}^n \phi_j(z_i)\varepsilon_i\right| |\hat{\beta}_{\lambda,j} - \beta_j| + \sum_{j=1}^p \omega_j|\beta_j| - \sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j}|. \tag{32}$$

For $1 \leq j \leq p$, set $S_j = \sum_{i=1}^n \phi_j(z_i)\varepsilon_i$ and let us denote by $E$, the event

$$E = \bigcap_{j=1}^p \{|S_j| \leq n\omega_j/2\}.$$

We state the results on the event $E$ and find an upper bound of $\mathbb{P}(E^c)$.
**On the event** $E$: adding the term $\frac{1}{2}\sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j} - \beta_j|$ to both sides of Inequality (32) yields to

$$R(f_{\hat{\beta}_L}) + \frac{1}{2}\sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j} - \beta_j| \leq R(f_\beta) + \sum_{j=1}^p \omega_j(|\hat{\beta}_{\lambda,j} - \beta_j| + |\beta_j| - |\hat{\beta}_{\lambda,j}|).$$

Since $|\hat{\beta}_{\lambda,j} - \beta_j| + |\beta_j| - |\hat{\beta}_{\lambda,j}| = 0$ for $j \notin K(\beta)$, we have

$$R(f_{\hat{\beta}_L}) - R(f_0) + \frac{1}{2}\sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda j} - \beta_j| \leq R(f_\beta) - R(f_0) + 2\|\delta_K\|_1. \tag{33}$$

We get from Equation (33)

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq R(f_\beta) - R(f_0) + 2\|\delta_K\|_1. \tag{34}$$

Now, we consider separatly two events:

$$E_1 = \{2\|\delta_K\|_1 \leq \eta(R(f_\beta) - R(f_0))\},$$

and

$$E_2 = \{\eta(R(f_\beta) - R(f_0)) < 2\|\delta_K\|_1\}.$$

On the event $E \cap E_1$ we get from (34)

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta)(R(f_\beta) - R(f_0)), \tag{35}$$

and hence the result follows. On the event $E \cap E_2$ all subsequent inequalities are valid. On one hand, by applying Cauchy-Schwarz inequality, we get from (34) that

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq R(f_\beta) - R(f_0) + 2\sqrt{|K(\beta)|}\|\delta_K\|_2. \tag{36}$$

On the other hand, we get from Equation (33) that

$$\frac{1}{2}\sum_{j=1}^{p} \omega_j |\hat{\beta}_{\lambda j} - \beta_j| \leq R(f_\beta) - R(f_0) + 2\|\delta_K\|_1,$$

and using (5.2) we obtain

$$\|\delta\|_1 \leq (4 + 4/\eta)\|\delta_K\|_1,$$

which implies $\|\delta_{K^c}\|_1 \leq (3 + 4/\eta)\|\delta_K\|_1$. We can therefore apply Assumption ($\mathbf{RE_1}$). with $a_0 = 3 + 4/\eta$. This yields,

$$\mu^2\|\delta_K\|_2^2 \leq \frac{\|X\delta\|_2^2}{n} = \frac{1}{n}(\hat{\beta}_L - \beta)^T W X^T X W (\hat{\beta}_L - \beta) \leq \left(\max_{1\leq j\leq p} \omega_j\right)^2 \|f_{\hat{\beta}_L} - f_\beta\|_n^2. \tag{37}$$

Gathering Equations (36) and (37) we obtain

$$\begin{aligned} R(f_{\hat{\beta}_L}) - R(f_0) &\leq R(f_\beta) - R(f_0) + 2\left(\max_{1\leq j\leq p} \omega_j\right)\sqrt{|K(\beta)|}\mu^{-1}\|f_{\hat{\beta}_L} - f_\beta\|_n \\ &\leq R(f_\beta) - R(f_0) + 2\left(\max_{1\leq j\leq p} \omega_j\right)\sqrt{|K(\beta)|}\mu^{-1}(\|f_{\hat{\beta}_L} - f_0\|_n + \|f_\beta - f_0\|_n). \end{aligned}$$

We now apply Lemma 5.1

**Lemma 5.1.** *Under Asumptions ($\mathbf{B_1}$) and ($\mathbf{B_3}$) we have*

$$c_0\epsilon_0\|f_\beta - f_0\|_n^2 \leq R(f_\beta) - R(f_0) \leq \frac{1}{4}c_0'\|f_\beta - f_0\|_n^2,$$

*where $c_0$ and $c_0'$ are constants depending on $C_0$ and $c_1$; $\epsilon_0$ is a constant depending on $c_1$ and $c_2$.*

Consequently

$$\begin{aligned} R(f_{\hat{\beta}_L}) - R(f_0) \quad \leq \quad & R(f_\beta) - R(f_0) + \frac{2\left(\max_{1\leq j\leq p} \omega_j\right)\sqrt{|K(\beta)|}\mu^{-1}}{\sqrt{c_0\epsilon_0}}\sqrt{R(f_{\hat{\beta}_L}) - R(f_0)} \\ & + \frac{2\left(\max_{1\leq j\leq p} \omega_j\right)\sqrt{|K(\beta)|}\mu^{-1}}{\sqrt{c_0\epsilon_0}}\sqrt{R(f_\beta) - R(f_0)}. \end{aligned}$$

Using inequality $2uv < u^2/b + bv^2$ for all $b > 1$, with $u = \left(\max_{1\leq j\leq p} \omega_j\right) \sqrt{|K(\beta)|}\mu^{-1}/\sqrt{c_0\epsilon_0}$ and $v$ being either $\sqrt{R(f_{\hat{\beta}_L}) - R(f_0)}$ or $\sqrt{R(f_\beta) - R(f_0)}$ we have

$$\begin{aligned} R(f_{\hat{\beta}_L}) - R(f_0) \quad &\leq \quad R(f_\beta) - R(f_0) + 2b\left(\frac{\omega_{\max}\sqrt{|K(\beta)|}\mu^{-1}}{\sqrt{c_0\epsilon_0}}\right)^2 \\ &\quad + \frac{R(f_{\hat{\beta}_L}) - R(f_0)}{b} + \frac{R(f_\beta) - R(f_0)}{b}. \end{aligned}$$

It follows that

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq \frac{b+1}{b-1}\left\{R(f_\beta) - R(f_0) + \frac{2b^2\left(\max_{1\leq j\leq p} \omega_j\right)^2 |K(\beta)|}{(b+1)(b-1)\mu^2 c_0\epsilon_0}\right\}.$$

Now taking $b = 1 + 2/\eta$ leads to

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta)\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)\left(\max_{1\leq j\leq p} \omega_j\right)^2 |K(\beta)|}{\mu^2 c_0\epsilon_0}\right\}. \tag{38}$$

According to Inequalities (35) and (38), we conclude that on the event $E$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta)\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)\left(\max_{1\leq j\leq p} \omega_j\right)^2 |K(\beta)|}{\mu^2 c_0\epsilon_0}\right\}, \tag{39}$$

where $c(\eta) = 2(1+2/\eta)^2/[(2+2/\eta)(2/\eta)]$. This ends the proof of (7) of the Theorem 3.2. Equation (8) follows from Lemma 5.1: To prove Inequalities (9) and (10) we just replace $\omega_j$ by $\lambda$.
This ends the proof of the Theorem 3.2 by using (30) and (31). ■

## 5.3 Proof of Corollary 3.1

Set $\delta = W(\hat{\beta}_L - \beta_0)$. The result (13) directly comes by taking $\beta = \beta_0$ and $\eta = 2$ in (39). Note that, on the event $E$ we have

$$\|\delta_{K(\beta_0)^c}\|_1 \leq 3\|\delta_{K(\beta_0)}\|_1.$$

Indeed, since $\hat{\beta}_L$ is the minimizer of $\hat{R}(f_\beta) + \sum_{j=1}^p \omega_j|\beta_j|$, then

$$R(f_{\hat{\beta}_L}) - R(f_{\beta_0}) + \sum_{j=1}^p \omega_j|\hat{\beta}_{\lambda,j}| \leq \frac{1}{n}\varepsilon^T X(\hat{\beta}_L - \beta_0) + \sum_{j=1}^p \omega_j|\beta_{0j}|,$$

which implies that

$$\|W\hat{\beta}_L\|_1 \leq \sum_{j=1}^{p} \left| \frac{1}{n} \sum_{i=1}^{n} \phi_j(z_i)\varepsilon_i \right| |\hat{\beta}_{\lambda,j} - \beta_j| + \|W\beta_0\|_1.$$

On the event $E$ we have

$$\begin{aligned} \|W(\hat{\beta}_L)_{K(\beta_0)}\|_1 + \|W(\hat{\beta}_L)_{K^c(\beta_0)}\|_1 \quad &\leq \quad \frac{1}{2}(\|W(\hat{\beta}_L - \beta_0)_{K(\beta_0)}\|_1 + \|W(\hat{\beta}_L)_{K^c(\beta_0)}\|_1) \\ &\quad + \|W(\beta_0)_{K(\beta_0)}\|_1, \end{aligned}$$

thus $\|\delta_{K(\beta_0)^c}\|_1 \leq 3\|\delta_{K(\beta_0)}\|_1$. Result (14) follows from Equation (13) by applying Lemma 5.1. Equation (15) follows from (14) by applying Equation (37) and $\|\delta\|_1^2 \leq 16s\|\delta_{K(\beta_0)}\|_2^2$. The last line follows from Lemma A.3 in Appendix with $a_j = |\hat{\beta}_{\lambda j} - \beta_{0j}|$ and $b_1 = 8\sqrt{2}s(\max_{1\leq j\leq p} \omega_j)/\mu^2(s,3)c_0\epsilon_0$. ■

## 5.4 Proof of Theorem 4.1

Fix an arbitrary $\beta \in \mathbb{R}^{\sum_{l=1}^{g}|G_l|}$. Set $\delta = W(\hat{\beta}_{GL} - \beta)$ where $W = \text{Diag}(W_1, ..., W_p)$ is a block diagonal matrix, with $W_l = \text{Diag}(\omega_l, ..., \omega_l)$. Since $\hat{\beta}_{GL}$ is the minimizer of $\hat{R}(f_\beta) + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2$, we get

$$R(f_{\hat{\beta}_{GL}}) - \frac{1}{n}\varepsilon^T X\hat{\beta}_{GL} + \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l\|_2 \leq R(f_\beta) - \frac{1}{n}\varepsilon^T X\beta + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2.$$

By applying Cauchy-Schwarz inequality, this implies

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) \quad &\leq \quad R(f_\beta) - R(f_0) + \sum_{l=1}^{g} \frac{1}{n} \sqrt{\sum_{j\in G_l} \left( \sum_{i=1}^{n} \phi_j(z_i)\epsilon_i \right)^2} \|(\hat{\beta}_{GL} - \beta)^l\|_2 \\ &\quad + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 - \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l\|_2. \end{aligned}$$

Set $Z_l = n^{-1}\sqrt{\sum_{j\in G_l} \left(\sum_{i=1}^{n} \phi_j(z_i)\epsilon_i\right)^2}$, for $l \in \{1, ..., g\}$ and the event

$$\mathcal{A} = \bigcap_{l=1}^{g} \{Z_l \leq \omega_l/2\}.$$

We state the result on event $\mathcal{A}$ and find an upper bound of $\mathbb{P}(\mathcal{A}^c)$.

**On the event $\mathcal{A}$:**

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + \sum_{l=1}^{g} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 - \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l\|_2.$$

This implies that

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + 2\sum_{l=1}^{g} \omega_l \|\beta^l\|_2.$$

We conclude that on the event $\mathcal{A}$ we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq \inf_{\beta \in \mathbb{R}^{\sum_{l=1}^g |G_l|}} \left\{ R(f_\beta) - R(f_0) + 2\|\beta\|_{2,1} \max_{1 \leq l \leq g} \omega_l \right\}.$$

We now come to the bound on $\mathbb{P}(A^c)$ and write

$$\begin{aligned}
\mathbb{P}(A^c) &= \mathbb{P}\left( \bigcup_{l=1}^g \{ \sqrt{\sum_{j \in G_l} \left( \sum_{i=1}^n \phi_j(z_i)\epsilon_i \right)^2} > n\omega_l/2 \} \right) \\
&\leq \sum_{l=1}^g \mathbb{P}\left( \sqrt{\sum_{j \in G_l} \left( \sum_{i=1}^n \phi_j(z_i)\epsilon_i \right)^2} > n\omega_l/2 \right).
\end{aligned}$$

For $j \in G_l$ set $T_j^l = \sum_{i=1}^n \phi_j(z_i)\epsilon_i$, we have

$$\begin{aligned}
\mathbb{P}(A^c) &\leq \sum_{l=1}^g \mathbb{P}\left( \sqrt{\sum_{j \in G_l} (T_j^l)^2} > n\omega_l/2 \right) \\
&\leq \sum_{l=1}^g \mathbb{P}\left( \sum_{j \in G_l} |T_j^l| > n\omega_l/2 \right).
\end{aligned}$$

Using the fact that, for all $l \in \{1, ..., g\}$

$$\left\{ \sum_{j \in G_l} |T_j^l| > n\omega_l/2 \right\} \subset \bigcup_{j \in G_l} \left\{ |T_j^l| > \frac{n\omega_l}{2|G_l|} \right\}, \tag{40}$$

it follows that

$$\mathbb{P}(A^c) \leq \sum_{l=1}^g \sum_{j \in G_l} \mathbb{P}\left( |T_j^l| > \frac{n\omega_l}{2|G_l|} \right).$$

For $j \in G_l$, set $v_j^l = \sum_{i=1}^n I\!E(\phi_j^2 \epsilon_i^2)$. Since $\sum_{i=1}^n \phi_j^2(z_i) \geqslant 4v_j^l$, we have

$$\mathbb{P}(|T_j^l| > \frac{n\omega_l}{2|G_l|}) \leq \mathbb{P}\left( |T_j^l| > \sqrt{2v_j^l \left( x + \log \sum_{l=1}^g |G_l| \right)} + \frac{c_l}{3} \left( x + \log \sum_{l=1}^g |G_l| \right) \right).$$

By applying Bernstein's inequality (see Lemma A.4) to the right hand side of the previous inequality we get

$$\mathbb{P}(|T_j^l| > \frac{n\omega_l}{2|G_l|}) \leq 2 \exp\left( -x - \log \sum_{l=1}^g |G_l| \right).$$

It follows that

$$\mathbb{P}(A^c) \leq \sum_{l=1}^{g} \sum_{j \in G_l} \mathbb{P}\left( |T_j^l| > \frac{n\omega_l}{2|G_l|} \right) \leq \exp(-x). \tag{41}$$

this ends the proof of the Theorem 4.1. ■

## 5.5 Proof of Theorem 4.2

The proof starts as in the proof of Theorem 3.2. Fix an arbitrary $\beta \in \mathbb{R}^{\sum_{l=1}^{g} |G_l|}$ such that $f_\beta \in \Gamma_1$. Set $\delta = W(\hat{\beta}_{GL} - \beta)$ where $W = \text{Diag}(W_1, ..., W_p)$ is a block diagonal matrix, with $W_l = \text{Diag}(\omega_l, ..., \omega_l)$. Since $\hat{\beta}_{GL}$ is the minimizer of $\hat{R}(f_\beta) + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2$, we get

$$R(f_{\hat{\beta}_{GL}}) - \frac{1}{n}\varepsilon^T X \hat{\beta}_{GL} + \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l\|_2 \leq R(f_\beta) - \frac{1}{n}\varepsilon^T X \beta + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2$$

which implies:

$$R(f_{\hat{\beta}_{GL}}) \leq R(f_\beta) + \frac{1}{n}\varepsilon^T X(\hat{\beta}_{GL} - \beta) + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 - \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l\|_2 \tag{42}$$

By applying Cauchy-Schwarz inequality, we obtain

$$R(f_{\hat{\beta}_{GL}}) \leq R(f_\beta) + \sum_{l=1}^{g} \frac{1}{n} \sqrt{\sum_{j \in G_l} \left( \sum_{i=1}^{n} \phi_j(z_i)\epsilon_i \right)^2} \|(\hat{\beta}_{GL} - \beta)^l\|_2 + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 - \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l\|_2.$$

Set $Z_l = n^{-1} \sqrt{\sum_{j \in G_l} \left( \sum_{i=1}^{n} \phi_j(z_i)\epsilon_i \right)^2}$, for $l \in \{1, ..., g\}$ and the event

$$\mathcal{A} = \bigcap_{l=1}^{g} \{ Z_l \leq \omega_l / 2 \}.$$

We state the result on event $A$ and find an upper bound of $\mathbb{P}(A^c)$.

**On the event $\mathcal{A}$**: adding the term $\frac{1}{2}\sum_{l=1}^{g} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2$ to both sides of Inequality (42) yields to

$$R(f_{\hat{\beta}_{GL}}) + \frac{1}{2}\sum_{l=1}^{g} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq R(f_\beta) + \sum_{l=1}^{g} \omega_l (\|(\hat{\beta}_{GL} - \beta)^l\|_2 - \|\hat{\beta}_{GL}^l\|_2 + \|\beta^l\|_2).$$

Since $\|(\hat{\beta}_{GL} - \beta)^l\|_2 - \|\hat{\beta}_{GL}^l\|_2 + \|\beta^l\|_2 = 0$ for for $l \notin J(\beta) = J$, we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) + \frac{1}{2}\sum_{l=1}^{g} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq R(f_\beta) - R(f_0) + 2\sum_{l \in J} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2. \tag{43}$$

we get from Equation (43) that

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + 2\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 \tag{44}$$

Consider separatly the two events:

$$\mathcal{A}_1 = \{2\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq \eta(R(f_\beta) - R(f_0))\},$$

and

$$\mathcal{A}_2 = \{\eta(R(f_\beta) - R(f_0)) < 2\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2\}.$$

On the event $\mathcal{A}\cap\mathcal{A}_1$, we get from (44)

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)(R(f_\beta) - R(f_0)), \tag{45}$$

and the result fillows. On the event $\mathcal{A}\cap\mathcal{A}_1$, all the following inequalities are valid. On one hand, by applying Cauchy Schwarz inequality, we get from (44) that

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) &\leq R(f_\beta) - R(f_0) + 2\sqrt{|J(\beta)|}\sqrt{\sum_{l\in J}\omega_l^2\|(\hat{\beta}_{GL} - \beta)^l\|_2^2} \\ &\leq R(f_\beta) - R(f_0) + 2\sqrt{|J(\beta)|}\|\delta_J\|_2. \end{aligned} \tag{46}$$

On the other hand we get from Equation (43) that

$$\frac{1}{2}\sum_{l=1}^{g}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq R(f_\beta) - R(f_0) + 2\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2,$$

and using (5.5) we obtain

$$\frac{1}{2}\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 + \frac{1}{2}\sum_{l\in J^c}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq \frac{2}{\eta}\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2 + 2\sum_{l\in J}\omega_l\|(\hat{\beta}_{GL} - \beta)^l\|_2,$$

which implies

$$\|\delta_{J^c}\|_{2,1} \leq (3 + 4/\eta)\|\delta_J\|_{2,1}.$$

We can therefore apply Assumption (**RE$_2$**) with $a_0 = 3 + 4/\eta$. It follows that,

$$\mu_1^2\|\delta_J\|_2^2 \leq \frac{\|X\delta\|_2^2}{n} = \frac{1}{n}(\hat{\beta}_{GL} - \beta)^T W X^T X W(\hat{\beta}_{GL} - \beta) \leq (\max_{1\leq l\leq g}\omega_l)^2\|f_{\hat{\beta}_{GL}} - f_\beta\|_n^2. \tag{47}$$

Gathering Equation (46) and (47) we get

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) &\leq R(f_\beta) - R(f_0) + 2(\max_{1\leq l\leq g}\omega_l)\sqrt{|K(\beta)|}\mu_1^{-1}\|f_{\hat{\beta}_{GL}} - f_\beta\|_n \\ &\leq R(f_\beta) - R(f_0) + 2(\max_{1\leq l\leq g}\omega_l)\sqrt{|K(\beta)|}\mu_1^{-1}(\|f_{\hat{\beta}_{GL}} - f_0\|_n + \|f_\beta - f_0\|_n). \end{aligned}$$

We now apply Lemma 5.2

**Lemma 5.2.** *Under asumptions (**B**$_1$) and (**B**$_5$) we have*

$$c_0\epsilon_0\|f_\beta - f_0\|_n^2 \leq R(f_\beta) - R(f_0) \leq \frac{1}{4}c_0'\|f_\beta - f_0\|_n^2.$$

*where $c_0$ and $c_0'$ are constants depending on $C_0$ and $\epsilon_0$ is a constant depending on $c_1$ and $c_2$.*

Consequently

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) \quad &\leq \quad R(f_\beta) - R(f_0) + \frac{2(\max_{1\leq l\leq g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}\sqrt{R(f_{\hat{\beta}_{GL}}) - R(f_0)} \\ &\quad + \frac{2(\max_{1\leq l\leq g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}\sqrt{R(f_\beta) - R(f_0)}. \end{aligned}$$

Using inequality $2uv < u^2/b + bv^2$ for all $b > 1$, with $u = (\max_{1\leq l\leq g}\omega_l)\frac{\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}$ and $v$ being either $\sqrt{R(f_{\hat{\beta}_{GL}}) - R(f_0)}$ or $\sqrt{R(f_\beta) - R(f_0)}$ we have

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) \quad &\leq \quad R(f_\beta) - R(f_0) + 2b\left(\frac{(\max_{1\leq l\leq g}\omega_l)\sqrt{|J(\beta)|}\mu_1^{-1}}{\sqrt{c_0\epsilon_0}}\right)^2 \\ &\quad + \frac{R(f_{\hat{\beta}_{GL}}) - R(f_0)}{b} + \frac{R(f_\beta) - R(f_0)}{b}. \end{aligned}$$

This implies that

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq \frac{b+1}{b-1}\left\{R(f_\beta) - R(f_0) + \frac{2b^2(\max_{1\leq l\leq g}\omega_l)^2|J(\beta)|}{(b+1)(b-1)\mu_1^2c_0\epsilon_0}\right\}. \tag{48}$$

Now taking $b = 1 + 2/\eta$ leads to

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)(\max_{1\leq l\leq g}\omega_l)^2|J(\beta)|}{\mu_1^2c_0\epsilon_0}\right\}. \tag{49}$$

Accordind to Inequalities (45) and (49) we conclude that on event $\mathcal{A}$,

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)(\max_{1\leq l\leq g}\omega_l)^2|J(\beta)|}{\mu_1^2c_0\epsilon_0}\right\}, \tag{50}$$

where $c(\eta) = 2(1+2/\eta)^2/(2+2/\eta)(2/\eta)$. This ends the proof of (4.2) of the theorem 4.2. Equation 4.2 of the theorem follows from Lemma 5.2.

This ends the proof of the Theorem 4.2 by considering (41). ■

## 5.6 Proof of Corollary 4.1

Set $\delta = W(\hat{\beta}_{GL} - \beta_0)$, Line (21) of corollary 4.1 follows directly from Equation (50) with $\beta = \beta_0$ and $\eta = 2$. Note that on the event $A$, we have

$$\|\delta_{J(\beta_0)^c}\|_{2,1} \leq 3\|\delta_{J(\beta_0)}\|_{2,1}. \tag{51}$$

Indeed, since $\hat{\beta}_{GL}$ is the minimizer of $\hat{R}(f_\beta) + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2$,

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) + \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l_{GL}\|_2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta}_{GL} - \beta_0) + \sum_{l=1}^{g} \omega_l \|\beta^l_0\|_2$$

which implies

$$\|W\hat{\beta}_{GL}\|_{2,1} \leq \sum_{l=1}^{g} \frac{1}{n} \sqrt{\sum_{j \in G_l} \left( \sum_{i=1}^{n} (z_{ij})\epsilon_i \right)^2} \|(\hat{\beta}_{GL} - \beta_0)^l\|_2 + \|W\beta_0\|_{2,1}$$

On the event $A$ we have

$$\begin{aligned} \|W(\hat{\beta}_{GL})_{J(\beta_0)}\|_{2,1} + \|W(\hat{\beta}_{GL})_{J^c(\beta_0)}\|_{2,1} \quad &\leq \quad \frac{1}{2}(\|W(\hat{\beta}_{GL} - \beta_0)_{J(\beta_0)}\|_{2,1} + \|W(\hat{\beta}_{GL})_{J^c(\beta_0)}\|_{2,1}) \\ &\quad + \|W(\beta_0)_{J(\beta_0)}\|_{2,1}, \end{aligned}$$

thus $\|\delta_{J(\beta_0)^c}\|_{2,1} \leq 3\|\delta_{J(\beta_0)}\|_{2,1}$. Line (22) follows from Line 21 by applying Lemma (5.2). Line (23) follows from the second by using Equation (47) and $\|\delta\|^2_{2,1} \leq 16s\|\delta_{J(\beta_0)}\|^2_2$. Line (24) is the consequence of the Lemma A.3 with $a_l = \|(\hat{\beta}_{\lambda j} - \beta_0)^l\|_2$ and $b_1 = 8\sqrt{2}s(\max_{1 \leq l \leq g} \omega_l)/\mu^2(s,3)c_0\epsilon_0$. ■

## 5.7 Proof of Theorem 4.3

Set $\delta' = \hat{\beta}_{GL} - \beta_0$ and $J = J(\beta_0)$. By definition, $\hat{\beta}_{GL}$ is the minimizer of $\hat{R}(f_\beta) + \sum_{l=1}^{g} \omega_l \|\beta^l\|_2$, and hence

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) + \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l_{GL}\|_2 \leq \frac{1}{n}\varepsilon^T X(\hat{\beta}_{GL} - \beta_0) + \sum_{l=1}^{g} \omega_l \|\beta^l_0\|_2.$$

This implies that

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \frac{1}{n}\varepsilon^T X(\hat{\beta}_{GL} - \beta_0) - \sum_{l=1}^{g} \omega_l \|\hat{\beta}^l_{GL}\|_2 + \sum_{l=1}^{g} \omega_l \|\beta^l_0\|_2.$$

By applying Cauchy-Schwartz inequality we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \sum_{l=1}^{g} \frac{1}{n} \sqrt{\sum_{j \in G_l} \left( \sum_{i=1}^{n} (z_{ij})\epsilon_i \right)^2} \|(\hat{\beta}_{GL} - \beta_0)^l\|_2 + \sum_{l=1}^{g} \omega_l \|(\beta_0 - \hat{\beta}_{GL})^l\|_2,$$

Define the event

$$A = \bigcap_{l=1}^{g} \left\{ \frac{1}{n} \sqrt{\sum_{j \in G_l} \left( \sum_{i=1}^{n} (z_{ij})\epsilon_i \right)^2} \leq \omega_l/2 \right\}.$$

On the event $A$ we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \sum_{l=1}^{g} \frac{3\omega_l}{2} \|(\hat{\beta}_{GL} - \beta_0)^l\|_2. \tag{52}$$

By Lemma A.1 we have,

$$\frac{\langle h, h \rangle_{f_{\beta_0}}}{\|h\|_\infty^2} (\exp(-\|h\|_\infty) + \|h\|_\infty - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \tag{53}$$

where

$$h(z_i) = (f_{\hat{\beta}_{GL}} - f_{\beta_0})(z_i) = \sum_{l=1}^{g} \sum_{j \in G_l} (\hat{\beta}_{\lambda j} - \beta_{0j}) z_{ij}$$

and $\|h\|_\infty \leq v\|\delta'\|_{2,1}$, with $v = \max_{1\leq i\leq n} \max_{1\leq l\leq g} \|z_i^l\|_2$. Equation (53) and the decreasing of $t \mapsto \frac{\exp(-t)+t-1}{t^2}$ lead to

$$\frac{\delta'^T X^T D X \delta'}{n(v\|\delta'\|_{2,1})^2} (\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}).$$

Now, Inequalitiy (51) implies

$$\|\delta'_{J(\beta_0)^c}\|_{2,1} \leq 3 \frac{\left(\max\limits_{1\leq l\leq g} \omega_l\right)}{\min\limits_{1\leq l\leq g} \omega_l} \|\delta'_{J(\beta_0)}\|_{2,1}.$$

We can therefore apply Assumption ($\mathbf{RE_3}$) with $a_0 = 3(\max\limits_{1\leq l\leq g} \omega_l)/\min\limits_{1\leq l\leq g} \omega_l$ and get that

$$\frac{\mu_2^2 \|\delta'_J\|_2^2}{v^2 \|\delta'\|_{2,1}^2} (\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}).$$

We can use that $\|\delta'\|_{2,1}^2 \leq 16|J|\|\delta'_J\|_2^2$ to write

$$\frac{\mu_2^2}{16|J|v^2} (\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}).$$

According to Equation (52) we have

$$\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1 \leq \frac{24|J|v^2 \left(\max\limits_{1\leq l\leq g} \omega_l\right)}{\mu_2^2} \|\delta'\|_{2,1}. \tag{54}$$

Now, a short calculation shows that for all $a \in (0, 1]$,

$$e^{\frac{-2a}{1-a}} + (1-a)\frac{2a}{1-a} - 1 \geqslant 0 \tag{55}$$

Set $a = v\|\delta'\|_{2,1}/(v\|\delta'\|_{2,1} + 2)$. Thus $v\|\delta'\|_{2,1} = 2a/(1-a)$ and we have

$$e^{-v\|\delta'\|_{2,1}} + v\|\delta'\|_{2,1} - 1 \geqslant \frac{v^2\|\delta'\|_{2,1}^2}{v\|\delta'\|_{2,1} + 2}. \tag{56}$$

This implies using Equation (54) that

$$v\|\delta'\|_{2,1} \leq \frac{48\left(\max_{1\leq l\leq g}\omega_l\right)|J|v/\mu_2^2}{1-24\left(\max_{1\leq l\leq g}\omega_l\right)|J|v/\mu_2^2}.$$

Now if $\max_{1\leq l\leq g}\omega_l \leq \frac{\mu_2}{48v|J|}$, we have $v\|\delta'\|_{2,1} \leq 2$ and consequently

$$\frac{\exp(-v\|\delta'\|_{2,1})+v\|\delta'\|_{2,1}-1}{v^2\|\delta'\|_{2,1}^2} \geqslant 1/4.$$

Using Equation (54) we have that

$$\|\delta'\|_{2,1} \leq \frac{96|J|\left(\max_{1\leq l\leq g}\omega_l\right)}{\mu_2^2},$$

which prove the Line (26). Line (25) follows from (26) by using Equation (52). Line (27) is the consequence of Lemma A.3 Taking $a_l = \|(\hat{\beta}_{\lambda j}-\beta_0)^l\|_2$ and $b_1 = 144|J|(\min_{1\leq l\leq g}\omega_l)/\mu_2^2(s,3)$. Line (28) follows from (25) and Equation (53). ■

# A Appendix

The proof of Lemma 5.1 and 5.2 are based on property of self concordant function (see for instance [20]), *ie*, the functions whose third derivatives are controlled by their second derivatives. A one-dimensional, convex function $g$ is called self concordant if

$$|g^{'''}(x)| \leq Cg^{''}(x)^{3/2}.$$

The function we use ($g(t)=\hat{R}(g+th)$) is not realy self concordant but we can bound his third derivative by the second derivative times a constant. Our results on self-concordant functions are based on the ones of [2]. He has used and extended tools from convex optimization and self-concordance to provide simple extensions of theoretical results for the square loss to logistic loss. We use the same kind of arguments and state some relations between excess risk and prediction loss in the context of nonparametric logistic model, where $f_0$ is not necessarly linear as assumed in [2]. Precisely we extend Proposition 1 in [2] to the functions which are not necessarly linear (see Lemma A.1). This allows us to establish Lemma 5.1 and 5.2.

**Lemma A.1.** *For all $h, f : \mathbb{R}^d \to \mathbb{R}$,*

$$\frac{\langle h,h\rangle_f}{\|h\|_\infty^2}(\exp(-\|h\|_\infty)+\|h\|_\infty-1) \leq R(f+h)-R(f)+(q_f-q_{f_0})(h) \tag{57}$$

$$R(f+h)-R(f)+(q_f-q_{f_0})(h) \leq \frac{\langle h,h\rangle_f}{\|h\|_\infty^2}(\exp(\|h\|_\infty)-\|h\|_\infty-1) \tag{58}$$

$$\langle h,h\rangle_f e^{-\|h\|_\infty} \leq \langle h,h\rangle_{f+g} \leq \langle h,h\rangle_f e^{\|h\|_\infty}. \tag{59}$$

## A.1 Proof of Lemma A.1

We use the following lemma (see [2] Lemma 1) that we recall here:

**Lemma A.2.** *Let $g$ be a convex three times differentiable function $g : \mathbb{R} \to \mathbb{R}$ such that for all $t \in \mathbb{R}$ $|g^{'''}(t)| \leq S g^{''}(t)$, for some $S \geq 0$. Then , for all $t \geq 0$:*

$$\frac{g^{''}(0)}{S^2}(\exp(-St) + St - 1) \leq g(t) - g(0) - g'(0)t \leq \frac{g^{''}(0)}{S^2}(\exp(St) - St - 1). \tag{60}$$

We refer to Appendix A of [2] for the proof of this lemma.

Set

$$g(t) = \hat{R}(f + th) = \frac{1}{n}\sum_{i=1}^{n} l((f + th)(z_i)) - Y_i(f + th)(z_i), \quad f, h \in H,$$

where $l(u) = \log(1 + \exp(u))$. A short calculation leads to $l'(u) = \pi(u)$, $l^{''}(u) = \pi(u)(1 - \pi(u))$, $l^{'''}(u) = \pi(u)[1 - \pi(u)][1 - 2\pi(u)]$. It follows that

$$g^{''}(t) = \frac{1}{n}\sum_{i=1}^{n} h^2(z_i) l^{''}((f + th)(z_i)) = \langle h, h\rangle_{f+th},$$

and

$$g^{'''}(t) = \frac{1}{n}\sum_{i=1}^{n} h^3(z_i) l^{'''}((f + th)(z_i)).$$

Since $l^{'''}(u) \leq l^{''}(u)$ we have,

$$\begin{aligned} |g^{'''}(t)| &= \left|\frac{1}{n}\sum_{i=1}^{n} h^3(z_i) l^{'''}((f + th)(z_i))\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} h^2(z_i) l^{''}((f + th)(z_i))\|h\|_\infty = \|h\|_\infty g^{''}(t). \end{aligned}$$

We now apply Lemma A.2 to $g(t)$ with $S = \|h\|_\infty$, taking $t = 1$. Using Equation (5) we get the first and second inequality of Lemma A.1. Now by considering $g(t) = \langle h, h\rangle_{f+th}$, a short calculation leads to $|g'(t)| \leq \|h\|_\infty g(t)$ which implies $g(0)e^{-\|h\|_\infty t} \leq g(t) \leq g(0)e^{\|h\|_\infty t}$. By applying the last inequality to $g(t)$, and taking $t = 1$ we get the third inequality of Lemma A.1.

## A.2 Proof of Lemma 5.1

Set $h_0 = f_\beta - f_0$ from Lemma A.1 below,

$$\frac{\langle h_0, h_0\rangle_{f_0}}{\|h_0\|_\infty^2}(\exp(-\|h_0\|_\infty) + \|h_0\|_\infty - 1) \leq R(f_\beta) - R(f_0).$$

Using Assumptions ($\mathbf{B_3}$), ($\mathbf{B_1}$) and the decreasing of $t \mapsto \frac{\exp(-t)+t-1}{t^2}$, we claim that there exists $c_0 = c_0(C_0, c_1) > 0$ such that

$$c_0 \leq \frac{\exp(-\|h_0\|_\infty) + \|h_0\|_\infty - 1)}{\|h_0\|_\infty^2}.$$

According to Assumption ($\mathbf{B_1}$), there exists $0 \leq \epsilon_0 \leq 1/2$ such that for $1 \leq i \leq n$

$$\epsilon_0 \leq \pi(f_0(z_i))(1 - \pi(f_0(z_i))) \leq 1 - \epsilon_0.$$

The proof of the left hand side of Lemma 5.1 follows from the fact that $\epsilon_0 \|h_0\|_n^2 \leq \langle h_0, h_0 \rangle_{f_0}$. From the second line of Lemma A.1 we have

$$R(f_\beta) - R(f_0) \leq \frac{\langle h_0, h_0 \rangle_{f_0}}{\|h_0\|_\infty^2} (\exp(\|h_0\|_\infty) - \|h_0\|_\infty - 1).$$

Using asumption ($\mathbf{B_3}$) and increasing of $t \mapsto \frac{\exp(t)-t-1}{t^2}$ thus there exists $c_0' = c_0'(C_0, c_1) > 0$ such that

$$\begin{aligned} R(f_\beta) - R(f_0) &\leq c_0' \langle h_0, h_0 \rangle_{f_0} \\ &\leq c_0' \frac{1}{4} \|h_0\|_n^2. \end{aligned}$$

This end the proof of the right hand side of the Lemma 5.1.

**Lemma A.3.** *If we assume that* $\sum_{i=1}^p a_j \leq b_1$ *with* $a_j > 0$, *this implies that* $\sum_{i=1}^p a_j^q \leq b_1^q$, *with* $1 \leq q \leq 2$

## A.3 Proof of Lemma A.3

$$\begin{aligned} \sum_{i=1}^p a_j^q &= \sum_{i=1}^p a_j^{2-q} a_j^{2q-2} \\ &\leq \left( \sum_{i=1}^p a_j \right)^{2-q} \left( \sum_{i=1}^p a_j^2 \right)^{q-1} \end{aligned}$$

Since $\sum_{i=1}^p a_j^2 \leq \left(\sum_{i=1}^p a_j\right)^2 \leq b_1^2$, thus

$$\sum_{i=1}^p a_j^q \leq b_1^{2-q} b_1^{2q-2} = b_1^q \tag{61}$$

## A.4 Proof of Lemma 5.2

We act similarly to the proof of Lemma 5.1.

**Lemma A.4** (Bernsten's inequality). *Let $X_1, ..., X_n$ be independent real valued random variables such that for all $i \leq n$, $X_i \leq b$ almost surely, then we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}(E_i)\right| \geqslant \sqrt{2vx} + bx/3\right] \leq 2\exp(-x),$$

*where $v = \sum_{i=1}^{n} \mathbb{E}(X_i^2)$.*

This lemma is obtain by gattering Proposition 2.9 and inequality (2.23) from [14].

**Lemma A.5** (Hoeffding's inequality). *Let $X_1, ..., X_n$ be independent random variables such that $X_i$ takes its values in $[a_i, b_i]$ almost surely for all $i \leq n$. Then for any positive $x$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} X_i - \mathbb{E}(E_i)\right| \geqslant x\right] \leq 2\exp(-\frac{2x^2}{\sum_{i=1}^{n}(b_i - a_i)^2}).$$

This lemma is a consequence of Proposition 2.7 in [14].

# References


[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag, 1973.

[2] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[3] 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.

[4] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation and sparsity via $\ell_1$ penalized least squares. *Learning theory*, pages 379–391, 2006.

[5] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

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

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

[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] M. Garcia-Magariños, A. Antoniadis, R. Cao, and W. Gonzàlez-Manteiga. Lasso logistic regression, gsoft and the cyclic coordinate descent algorithm: Application to gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 2010.

[10] J. Huang, S. Ma, and CH Zhang. The iterated lasso for high-dimensional logistic regression. Technical report, 2008.

[11] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, pages 1356–1378, 2000.

[12] K. Lounici, M. Pontil, A.B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *Arxiv preprint arXiv:0903.1468*, 2009.

[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. Massart. Concentration inequalities and model selection. 2007.

[15] P. Massart and C. Meynet. An l1-oracle inequality for the lasso. *Arxiv preprint arXiv:1007.4791*, 2010.

[16] J. McAuley, J. Ming, D. Stewart, and P. Hanna. Subband correlation and robust speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 13(5):956–964, 2005.

[17] L. Meier, S. van de Geer, and P. Bühlmann. The group lasso for logistic regression. *group*, 70(Part 1):53–71, 2008.

[18] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[19] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[20] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Society for Industrial Mathematics, 1987.

[21] M.R. Osborne, B. Presnell, and B.A. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[22] M.Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal-Royal statistical society. SERIES B statistical methodology*, 69(4):659, 2007.

[23] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[24] B. Tarigan and S.A. van de Geer. Classifiers of support vector machine type with l1 complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006.

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

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

[27] S.A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[28] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[29] 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, 2006.

[30] C.H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

[31] P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[32] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.