# Non-asymptotic oracle inequalities for the Lasso and Group Lasso in high dimensional logistic model 

Marius Kwemou

## - To cite this version:

Marius Kwemou. Non-asymptotic oracle inequalities for the Lasso and Group Lasso in high dimensional logistic model. ESAIM: Probability and Statistics, 2016, 20, pp.309-331. 10.1051/ps/2015020 . hal-01601407

HAL Id: hal-01601407
https://hal.science/hal-01601407
Submitted on 27 May 2020

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

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

# NON-ASYMPTOTIC ORACLE INEQUALITIES FOR THE LASSO AND GROUP LASSO IN HIGH DIMENSIONAL LOGISTIC MODEL 

Marius Kwemou ${ }^{1,2}$


#### 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 eigenvalue assumption as introduced in [P.J. Bickel, Y. Ritov and A.B. Tsybakov, Ann. Statist. 37 (2009) 1705-1732].


Mathematics Subject Classification. 62H12, 62J12, 62J07, 62G20.
Received July 1st, 2014. Revised November 25, 2015. Accepted December 4, 2015.

## 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 [11] or [37]). In this context the considered model is often what we call here "usual" logistic regression. It is given by

$$
\begin{equation*}
\mathbb{P}\left(Y_{i}=1\right)=\pi\left(z_{i}^{T} \beta_{0}\right)=\frac{\exp \left(z_{i}^{T} \beta_{0}\right)}{1+\exp \left(z_{i}^{T} \beta_{0}\right)}, \tag{1.1}
\end{equation*}
$$

where one observes $n$ couples $\left(z_{1}, Y_{1}\right), \ldots,\left(z_{n}, Y_{n}\right) \in \mathbb{R}^{d} \times\{0,1\}$, and $\beta_{0}$ is the unknown parameter to estimate. Throughout the paper, we consider a fixed design setting (i.e. $z_{1}, \ldots, z_{n}$ are considered deterministic).

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

$$
\begin{equation*}
\mathbb{P}\left(Y_{i}=1\right)=\frac{\exp \left(f_{0}\left(z_{i}\right)\right)}{1+\exp \left(f_{0}\left(z_{i}\right)\right)}, \tag{1.2}
\end{equation*}
$$

where the outputs $Y_{i} \in\{0,1\}, i=1, \ldots, n$ are independent and $f_{0}$ (not necessarily linear) is an unknown function (see [12]). We aim at estimating $f_{0}$ by constructing a suitable approximation. More precisely we

[^0]estimate $f_{0}$ by a sparse approximation of linear combination of elements of a given dictionary of functions $\mathbb{D}=\left\{\phi_{1}, \ldots, \phi_{p}\right\}: \hat{f}():.=\sum_{j=1}^{p} \hat{\beta}_{j} \phi_{j}($.$) . Our purpose expresses the belief that, in many instances, even if p$ 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, i.e. penalized by the number of non zero coefficients (see for instance AIC, BIC $[1,34]$ ). 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 [32] 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 a 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 exist some algorithms to solve this convex problem, glmnet (see [10]), predictor-corector (see [30]) 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 [21]).

### 1.1. 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 $[3,6-8,15,20]$, who have studied the Lasso for linear model in nonparametric setting and proved sparsity oracle inequalities. Similar sparsity oracle inequalities are proved in [4], and those results hold under the so-called restricted eigenvalue assumption on the Gram matrix. Those kind of results have been recently stated for the variants of the Lasso. For instance Lounici et al. [18] under a group version of restricted eigenvalue 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 [17]). The behavior of the Lasso and Group Lasso regarding their selection and estimation properties have been studied in: $[16,24,25,29,39,40]$ for Lasso in linear regression; $[9,26]$ for Group Lasso in linear regression; $[14,23,33]$ for additive models. 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 (1.1). Zou [41] shows consistency in variable selection for adaptive Lasso in generalized linear models when the number of covariables p is fixed. Huang et al. [13] prove sign consistency and estimation consistency for high-dimensional logistic regression. Meir et al. [22] shown consistency for the Group Lasso in "usual" logistic regression. Meir et al. [22] shown consistency for the Group Lasso in "usual" logistic model (1.1). 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 eigenvalue assumption on the weighted Gram matrix. The second one is from van de Geer [35], who established non asymptotic oracle inequality for Lasso in high dimensional generalized linear models with Lipschitz loss functions. Non asymptotic results concerning Group Lasso for logistic regression model have been established by Negahban et al. [27], with the assumption that $f_{0}$ is linear.

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 do not assume that $f_{0}$ is linear. We first state "slow" oracle inequalities (see Thms. 2.1 and 3.1 ) with no assumption on the Gram matrix, on the regressors nor on the margin. Secondly we provide "fast" oracle inequalities (see Thms. 2.2 and 3.2 ) under restricted eigenvalue assumption and some technical assumptions on the regressors. In each case, we give, as a consequence, the bounds for excess risk, $L_{2}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_{i}}\right)$ 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 (for the Lasso) and group norm (for the Group Lasso).

This paper is organized as follows. In Section 2, we describe our weighted Group Lasso estimation procedure and state non asymptotic oracle inequalities for the Group Lasso estimator. In Section 3 we describe our weighted Lasso estimation procedure and state non asymptotic oracle inequalities for the Lasso estimator. In Sections 2.3 and 3.3 we give as a consequence the bounds for excess risk, $L_{2}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_{i}}\right)$ and estimation errors for Lasso and Group Lasso in the "usual" logistic regression (1.1). The proofs are gathered in Section 5 and Appendix.

### 1.2. Definitions and notations

Consider the matrix $X=\left(\phi_{j}\left(z_{i}\right)\right)_{1 \leq i \leq n, 1 \leq j \leq p}$ and $\left\{G_{l}, l=1, \ldots, g\right\}$ the partition of $\{1, \ldots, p\}$. For any $\beta=\left(\beta_{1}, \ldots, \beta_{p}\right)^{T}=\left(\beta^{1}, \ldots, \beta^{g}\right)^{T} \in \mathbb{R}^{p}$, where $\beta^{l}=\left(\beta_{j}\right)_{j \in G_{l}}$ for $l=1, \ldots, g$. Let $f_{\beta}()=.\sum_{j=1}^{p} \beta_{j} \phi_{j}()=$. $\sum_{l=1}^{g} \sum_{j \in G_{l}} \beta_{j} \phi_{j}($.$) . With our notations$

$$
\left(f_{\beta}\left(z_{1}\right), \ldots, f_{\beta}\left(z_{n}\right)\right)^{T}=X \beta
$$

We define the group norm of $\beta$ as

$$
\|\beta\|_{2, q}=\left(\sum_{l=1}^{g}\left(\sum_{j \in G_{l}} \beta_{j}^{2}\right)^{\frac{q}{2}}\right)^{\frac{1}{q}}=\left(\sum_{l=1}^{g}\left\|\beta^{l}\right\|_{2}^{q}\right)^{\frac{1}{q}}
$$

for every $1 \leq q<\infty$. For $\beta \in \mathbb{R}^{p} K(\beta)=\left\{j \in\{1, \ldots, p\}: \beta_{j} \neq 0\right\}$ and $J(\beta)=\left\{l \in\{1, \ldots, g\}: \beta^{l} \neq 0\right\}$, respectively the set of relevant coefficients (which characterizes the sparsity of the vector $\beta$ ) and the set of relevant groups. For all $\delta \in \mathbb{R}^{p}$ and a subset $I \subset\{1, \ldots, p\}$, 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 $I$. Moreover $|I|$ denotes the cardinality of $I$. For all $h, f, g: \mathbb{R}^{d} \rightarrow R$, we define the scalar products

$$
\langle f, h\rangle_{n}=\frac{1}{n} \sum_{i=1}^{n} h\left(z_{i}\right) f\left(z_{i}\right)
$$

and

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

We use the notation

$$
q_{f}(h)=\frac{1}{n} \sum_{i=1}^{n} h\left(z_{i}\right)\left(Y_{i}-\pi\left(f\left(z_{i}\right)\right)\right)
$$

$\|h\|_{\infty}=\max _{i}\left|h\left(z_{i}\right)\right|$ and $\|h\|_{n}=\sqrt{\langle h, h\rangle_{n}}=\sqrt{\frac{1}{n} \sum_{i=1}^{n} h^{2}\left(z_{i}\right)}$ which denote the $L_{2}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_{i}}\right)$ norm (empirical norm). We consider empirical risk (logistic loss) for logistic model

$$
\begin{equation*}
\hat{R}(f)=\frac{1}{n} \sum_{i=1}^{n} \log \left(1+\exp \left(f\left(z_{i}\right)\right)\right)-Y_{i} f\left(z_{i}\right) \tag{1.3}
\end{equation*}
$$

We denote by $R$ the expectation of $\hat{R}$ with respect to the distribution of $Y_{1}, \ldots, Y_{n}$, i.e.

$$
R(f)=\mathbb{E}(\hat{R}(f))=\frac{1}{n} \sum_{i=1}^{n} \log \left(1+\exp \left(f\left(z_{i}\right)\right)\right)-\mathbb{E}\left(Y_{i}\right) f\left(z_{i}\right)
$$

It is clear that $R($.$) is a convex function and f_{0}$ is a minimum of $R($.$) when the model is well-specified (i.e.$ when (1.2) is satisfied). Note that with our notations

$$
\begin{equation*}
R(f)=\mathbb{E}(\hat{R}(f))=\hat{R}(f)+q_{f_{0}}(f) \tag{1.4}
\end{equation*}
$$

We shall use both the excess risk of $f_{\hat{\beta}}, R\left(f_{\hat{\beta}}\right)-R\left(f_{0}\right)$ and the prediction loss $\left\|f_{\hat{\beta}}-f_{0}\right\|_{n}^{2}$ to evaluate the quality of the estimator. Note that $R\left(f_{\hat{\beta}}\right)$ 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.

## 2. Group Lasso for LOGistic Regression model

### 2.1. Estimation procedure

The goal is not to estimate the parameters of the "true" model (since there is no true parameter) but rather to construct an estimator that mimics the performance of the best model in a given class, whether this model is true or not. Our aim is then to estimate $f_{0}$ in Model (1.2) by a linear combination of the functions of a dictionary

$$
\mathbb{D}=\left\{\phi_{1}, \ldots, \phi_{p}\right\}
$$

where $\phi_{j}: \mathbb{R}^{d} \rightarrow \mathbb{R}$ and $p$ possibly $\gg n$. 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

$$
f_{\beta}(.)=\sum_{j=1}^{p} \beta_{j} \phi_{j}(.),
$$

where $\beta$ has to be estimated.
In this section we assume that the set of relevant predictors have known group structure, for example in gene expression data these groups may be gene pathways, or factor level indicators in categorical data. 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 [22,38]), where predictors are included or excluded in groups. The logistic Group Lasso is the minimizer of the following optimization problem

$$
\begin{equation*}
f_{\hat{\beta}_{G L}}:=\underset{f_{\beta} \in \Gamma}{\operatorname{argmin}}\left\{\hat{R}\left(f_{\beta}\right)+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2}\right\} \tag{2.1}
\end{equation*}
$$

where

$$
\Gamma \subseteq\left\{f_{\beta}(.)=\sum_{l=1}^{g} \sum_{j \in G_{l}} \beta_{j} \phi_{j}(.), \beta \in \mathbb{R}^{p}\right\}
$$

The tuning parameter $r>0$ is used to adjust the trade-off between minimizing the loss and finding a solution which is sparse at the group level, i.e., to a vector $\beta$ such that $\beta^{l}=0$ for some of the groups $l \in$ $\{1, \ldots, g\}$. Sparsity is the consequence of the effect of non-differentiable penalty. This penalty can be viewed as an intermediate between $\ell_{1}$ and $\ell_{2}$ type penalty, which has the attractive property that it does variables selection at the group level. The weights $\omega_{l}>0$, which we will define later, are used to control the amount of penalization per group.

### 2.2. Oracle inequalities

In this section we state non asymptotic oracle inequalities for excess risk and $L_{2}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_{i}}\right)$ loss of Group Lasso estimator. Consider the following assumptions:

There exists a constant $0<c_{1}<\infty$ such that $\max _{1 \leq i \leq n}\left|f_{0}\left(z_{i}\right)\right| \leq c_{1}$.
There exists a constant $0<c_{2}<\infty$ such that $\max _{1 \leq i \leq n} \max _{1 \leq j \leq p}\left|\phi_{j}\left(z_{i}\right)\right| \leq c_{2}$.
There exists a constant $C_{0}$ such that the set $\Gamma=\Gamma\left(C_{0}\right)=\left\{f_{\beta}, \max _{1 \leq i \leq n}\left|f_{\beta}\left(z_{i}\right)\right| \leq C_{0}\right\}$ is non-empty. $\quad\left(\mathbf{B}_{\mathbf{3}}\right)$
Assumptions $\left(\mathbf{B}_{\mathbf{1}}\right)$ and $\left(\mathbf{B}_{\mathbf{3}}\right)$ are technical assumptions useful to connect the excess risk and the $L_{2}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_{i}}\right)$ loss (see Lem. 5.1). An assumption similar to $\left(\mathbf{B}_{1}\right)$ has been used in $[7]$ to prove oracle inequality in gaussian regression model. The same kind of assumption as $\left(\mathbf{B}_{3}\right)$ has been made in [31] to prove oracle inequality for support vector machine type with $\ell_{1}$ complexity regularization.

Theorem 2.1. Let $f_{\hat{\beta}_{G L}}$ be the Group Lasso solution defined in (2.1) with $r \geq 1$ and

$$
\begin{equation*}
\omega_{l}=\frac{2 \sqrt{\left|G_{l}\right|}}{n} \sqrt{\frac{1}{2} \max _{j \in G_{l}} \sum_{i=1}^{n} \phi_{j}^{2}\left(z_{i}\right)(x+\log p)}+\frac{2 c_{2} \sqrt{\left|G_{l}\right|}}{3 n}(x+\log p), \tag{2.2}
\end{equation*}
$$

where $x>0$. Under assumption $\left(\mathbf{B}_{\mathbf{2}}\right)$, with probability at least $1-2 \exp (-x)$ we have

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq \inf _{\beta \in \mathbb{R}^{p}}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r\|\beta\|_{2,1} \max _{1 \leq l \leq g} \omega_{l}\right\} . \tag{2.3}
\end{equation*}
$$

The first part of the right hand of Inequality (2.3) 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 $\mathbb{D}$ such that $f_{0}$ could well be approximated by a linear combination of the functions of $\mathbb{D}$. The second part of the right hand of Inequality (2.3) is the variance term and is usually referred as the rate of the oracle inequality. In Theorem 2.1, we speak about "slow" oracle inequality, with the rate at the order $\|\beta\|_{2,1} \sqrt{\log p / n}$ for any $\beta$. Moreover this is a sharp oracle inequality in the sense that there is a constant 1 in front of term $\inf _{\beta \in \mathbb{R}^{p}}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)\right\}$. This result is obtained without any assumption on the Gram matrix $\left(\Phi_{n}=X^{T} X / n\right)$. In order to obtain oracle inequality with a "fast rate" of order $\log p / n$ we need additional assumption on the restricted eigenvalue of the Gram matrix, namely the restricted eigenvalue assumption.

For some integer $s$ such that $1 \leq s \leq g$ and a positive number $a_{0}$, the following condition holds
$\left(\mathbf{R E}_{1}\right)$

$$
\mu_{1}\left(s, a_{0}\right):=\min _{K \subseteq\{1, \ldots p\}:|K| \leq s} \min _{\Delta \neq 0:\left\|\Delta_{K^{c}}\right\|_{2,1} \leq a_{0}\left\|\Delta_{K}\right\|_{2,1}} \frac{\|X \Delta\|_{2}}{\sqrt{n}\left\|\Delta_{K}\right\|_{2}}>0 .
$$

This is a natural extension to the Group Lasso of restricted eigenvalue assumption introduced in [4] (or Assumption ( $\mathbf{R E}_{\mathbf{3}}$ ) used below) for the usual Lasso. The only difference lies on the set where the minimum is taken: for the Lasso the minimum is taken over $\left\{\Delta \neq 0:\left\|\Delta_{K^{c}}\right\|_{1} \leq a_{0}\left\|\Delta_{K}\right\|_{1}\right\}$ whereas for the Group Lasso the minimum is over $\left\{\Delta \neq 0:\left\|\Delta_{K^{c}}\right\|_{2,1} \leq a_{0}\left\|\Delta_{K}\right\|_{2,1}\right\}$. This assumption has already been used in [17,18] to prove oracle inequality for linear gaussian noise model under Group sparsity and for multi-task learning. To emphasize the dependency of Assumption $\left(\mathbf{R E}_{\mathbf{1}}\right)$ on $s$ and $a_{0}$ we will sometimes refer to it as $R E\left(s, a_{0}\right)$.

Theorem 2.2. Let $f_{\hat{\beta}_{G L}}$ be the Group Lasso solution defined in (2.1) with $\omega_{l}$ defined as in (2.2). Fix $\eta>0$ and $1 \leq s \leq g$, assume that $\left(\mathbf{B}_{\mathbf{1}}\right),\left(\mathbf{B}_{\mathbf{2}}\right),\left(\mathbf{B}_{\mathbf{3}}\right)$ and $\left(\mathbf{R E}_{\mathbf{1}}\right)$ are satisfied, with $a_{0}=3+4 / \eta$. Thus with probability at
least $1-2 \exp (-x)$ we have

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq(1+\eta) \inf _{f_{\beta} \in \Gamma}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{c(\eta)|J(\beta)| r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{c_{0} \epsilon_{0} \mu_{1}\left(s, a_{0}\right)^{2}}\right\} \tag{2.4}
\end{equation*}
$$

and

$$
\begin{equation*}
\left\|f_{\hat{\beta}_{G L}}-f_{0}\right\|_{n}^{2} \leq \frac{c_{0}^{\prime}}{4 c_{0} \epsilon_{0}}(1+\eta) \inf _{f_{\beta} \in \Gamma}\left\{\left\|f_{\beta}-f_{0}\right\|_{n}^{2}+\frac{4 c(\eta)|J(\beta)| r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{c_{0}^{\prime} c_{0} \epsilon_{0}^{2} \mu_{1}(s, a)^{2}}\right\}, \tag{2.5}
\end{equation*}
$$

where $c(\eta)$ is a constant depending only on $\eta ; c_{0}=c_{0}\left(C_{0}, c_{1}\right)$ and $c_{0}^{\prime}=c_{0}^{\prime}\left(C_{0}, c_{1}\right)$ are constants depending on $C_{0}$ and $c_{1} ; \epsilon_{0}=\epsilon_{0}\left(c_{1}\right)$ is a constant depending on $c_{1}$; and $r \geq 1$.

In Theorem 2.2, the variance terms are of order $\log p / n$. Hence we say that the corresponding non asymptotic oracle inequalities have "fast rates". For the best of our knowledge, Inequalities (2.3), (2.4) and (2.5) are the first non asymptotic oracle inequalities for the Group Lasso in logistic regression model. These inequalities allow us to bound the prediction errors of Group Lasso by the best sparse approximation and a variance term. The major difference with existing results concerning Group Lasso for logistic regression model (see [22,27]) is that $f_{0}$ is not necessarily linear.

Remark 2.3. Our results remain true if we assume that we are in the "neighborhood" of the target function. If we suppose that there exists $\zeta$ such that $\max _{1 \leq i \leq n}\left|f_{\beta}\left(z_{i}\right)-f_{0}\left(z_{i}\right)\right| \leq \zeta$, then Lemma 5.1 is still true.

Remark 2.4. The choice of the weights $\omega_{\ell}$ comes from Bernstein's inequality. We could also use the following weights

$$
\omega_{l}^{\prime}=\frac{2 \sqrt{\left|G_{l}\right|}}{n} \sqrt{2 \max _{j \in G_{l}} \sum_{i=1}^{n} \mathbb{E}\left[\phi_{j}^{2}\left(z_{i}\right) \epsilon_{i}^{2}\right](x+\log p)}+\frac{2 \sqrt{\left|G_{l}\right|} \max _{1 \leq i \leq n j \in G_{l}} \max \left|\phi_{j}\left(z_{i}\right)\right|}{3 n}(x+\log p),
$$

where $\epsilon_{i}=Y_{i}-\mathbb{E}\left[Y_{i}\right], i=1, \ldots, n$. Theorems 2.1 and 2.2 still hold true with such weights $\omega_{l}^{\prime}$. But these weights depend on the unknown function $f_{0}$ to be estimated through $\mathbb{E}\left(\epsilon_{i}^{2}\right)=\pi\left(f_{0}\left(z_{i}\right)\left(1-\pi\left(f_{0}\left(z_{i}\right)\right)\right.\right.$. This is the reason for using weights $\omega_{l}$ slightly greater than $\omega_{l}^{\prime}$. We also note that our weights are proportional to the square root of groups sizes, which is in acordance with the weights previously proposed for grouping strategies (see [22]).

### 2.3. Special case: $f_{0}$ linear

In this section we assume that $f_{0}$ is a linear function i.e. $f_{0}\left(z_{i}\right)=f_{\beta_{0}}\left(z_{i}\right)=\sum_{l=1}^{g} \sum_{j \in G_{l}} \beta_{j} z_{i j}$. Denote by $X=\left(z_{i j}\right)_{1 \leq i \leq n, 1 \leq j \leq p}$, the design matrix. Let $z_{i}=\left(z_{i 1}, \ldots, z_{i p}\right)^{T}$ be the ith row of the matrix $X$ and $z^{(j)}=\left(z_{1 j}, \ldots, z_{n j}\right)^{T}$ is $j$ th column. For $i=1, \ldots, n$

$$
\begin{equation*}
\mathbb{P}\left(Y_{i}=1\right)=\frac{\exp \left(z_{i}^{T} \beta_{0}\right)}{1+\exp \left(z_{i}^{T} \beta_{0}\right)} \tag{2.6}
\end{equation*}
$$

This corresponds to the "usual" logistic regression (1.1) i.e. logistic model that allows linear dependency between $z_{i}$ and the distribution of $Y_{i}$. In this context, the Group Lasso estimator of $\beta_{0}$ is defined by

$$
\begin{equation*}
\hat{\beta}_{G L}:=\underset{\beta: f_{\beta} \in \Gamma}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n}\left\{\log \left(1+\exp \left(z_{i}^{T} \beta\right)\right)-Y_{i} z_{i}^{T} \beta\right\}+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2} . \tag{2.7}
\end{equation*}
$$

Corollary 2.5. Let assumption $\mathbf{R E}_{1}(s, 3)$ be satisfied and $\left|J\left(\beta_{0}\right)\right| \leq s$, where $1 \leq s \leq g$. Consider the Group Lasso estimator $f_{\widehat{\beta}_{G L}}$ defined by (2.7) with

$$
\begin{equation*}
\omega_{l}=\frac{2 \sqrt{\left|G_{l}\right|}}{n} \sqrt{\frac{1}{2} \max _{j \in G G_{l}} \sum_{i=1}^{n} z_{i j}^{2}(x+\log p)}+\frac{2 c_{2} \sqrt{\left|G_{l}\right|}}{3 n}(x+\log p) \tag{2.8}
\end{equation*}
$$

where $x>0$. Under the assumptions of Theorem 2.2, with probability at least $1-2 \exp (-x)$ we have

$$
\begin{align*}
& R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right) \leq \frac{9 s r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}}  \tag{2.9}\\
&\left\|f_{\hat{\beta}_{G L}}-f_{\beta_{0}}\right\|_{n}^{2} \leq \frac{9 s r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{\mu^{2}(s, 3) c_{0}^{2} \epsilon_{0}^{2}}  \tag{2.10}\\
&\left\|\hat{\beta}_{G L}-\beta_{0}\right\|_{2,1} \leq \frac{12 r s\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}\left(\min _{1 \leq l \leq g} \omega_{l}\right)}  \tag{2.11}\\
&\left\|\hat{\beta}_{G L}-\beta_{0}\right\|_{2, q}^{q} \leq\left(\frac{12 r s\left(\max _{1 \leq \leq \leq g} \omega_{l}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}\left(\min _{1 \leq l \leq g} \omega_{l}\right)}\right)^{q} \quad \text { for all } 1<q \leq 2 . \tag{2.12}
\end{align*}
$$

Remark 2.6. In logistic regression model (3.7), if vector $\beta_{0}$ is sparse, i.e. $\left|J\left(\beta_{0}\right)\right| \leq s$, then assumption $\left(\mathbf{R E}_{\mathbf{1}}\right)$ implies that $\beta_{0}$ is uniquely defined. Indeed, if there exists $\beta^{*}$ such that for $i=1, \ldots, n, \pi\left(z_{i}^{T} \beta_{0}\right)=\pi\left(z_{i}^{T} \beta^{*}\right)$, it follows that $X \beta_{0}=X \beta^{*}$ and $\left|J\left(\beta^{*}\right)\right| \leq s$. Then according to assumption $R E\left(s, a_{0}\right)$ with $a_{0} \geqslant 1$, we necessarily have $\beta_{0}=\beta^{*}$. Indeed if $R E\left(s, a_{0}\right)$ is satisfied with $a_{0} \geqslant 1$, then $\min \left\{\|X \beta\|_{2}:|J(\beta)| \leq 2 s, \beta \neq 0\right\}>0$.

Remark 2.7 (Theoretical advantage of Group Lasso over the Lasso). Concerning results on oracle inequality for the Group Lasso few results exist. The first oracle inequality for the Group Lasso in the additive regression model is due to [26]. Since then, some of these inequalities have been improved in Lounici et al. [18], concerning in particular the gain on order rate. More precisely, Lounici et al. [18] have found a rate of order $\log g / n$ for Group Lasso in gaussian linear model, which is better than is corresponding rate for the Lasso, $\log p / n$ (since $g \leq p$ ). This improvement seems mainly based on the assumption that the noise is gaussian. In our case (see proof of Thm. 2.1, formula (5.3)) the empirical process involves non gaussian variables and thus their method should not apply in our context. However the probability that their results are true depends on $g$ whereas the probability that our results hold does not depend on $g$.

We can find the rate of order $\log g / n$ by choosing this constant $x$ in the weights in a certain manner. Indeed, let us assume (without loss of generality) that the groups are all of equal size $\left|G_{1}\right|=\cdots=\left|G_{g}\right|=m$, so that $p=m . g$. Since the weights in (2.2) are defined for all $x>0$, if we take $x=q \log g-\log m>0$ where $q$ is a positive constant such that $g^{q}>m$. Then the weights in (2.2) become

$$
\omega_{l}=\frac{2 \sqrt{\left|G_{l}\right|}}{n} \sqrt{\frac{1}{2} \max _{j \in G_{l}} \sum_{i=1}^{n} \phi_{j}^{2}\left(z_{i}\right)[(1+q) \log g]}+\frac{2 c_{2} \sqrt{\left|G_{l}\right|}}{3 n}[(1+q) \log g],
$$

thus

$$
\omega_{l}^{2} \sim \frac{\log g}{n},
$$

and the results in Theorems 2.1 and 2.2 hold with probability at least

$$
1-2 \frac{m}{g^{q}}
$$

In the special case where the $g>2 m$ these results are true for all $q>0$.

### 2.4. Non bounded functions

The results of Corollary 2.5 are obtained (as the consequence of Thm. 2.2) with the assumptions that $f_{\beta_{0}}$ and all $f_{\beta} \in \Gamma$ are bounded. In some situations these assumptions could not be verified. In this section we will establish the same results without assuming $\left(\mathbf{B}_{\mathbf{1}}\right)$ or $\left(\mathbf{B}_{\mathbf{3}}\right)$ i.e. neither $f_{\beta_{0}}$ nor $f_{\beta}$ is bounded. We consider the Group Lasso estimator defined in (2.7) and the following assumption:

For some integer $s$ such that $1 \leq s \leq g$ and a positive number $a_{0}$, the following condition holds
$\left(\mathbf{R E}_{2}\right)$

$$
\mu_{2}\left(s, a_{0}\right):=\min _{K \subseteq\{1, \ldots p\}:|K| \leq s} \min _{\Delta \neq 0:\left\|\Delta_{K^{c}}\right\|_{2,1} \leq a_{0}\left\|\Delta_{K}\right\|_{2,1}} \frac{\Delta^{T} X^{T} D X \Delta}{n\left\|\Delta_{K}\right\|_{2}^{2}}>0
$$

where $D=\operatorname{Diag}\left(\operatorname{var}\left(Y_{i}\right)\right)$.
This is an extension of the assumption $\mathbf{R E}_{\boldsymbol{1}}$ to the weighted Gram matrix $X^{T} D X / n$.
Theorem 2.8. Consider the Group Lasso estimator $f_{\hat{\beta}_{G L}}$ defined by (2.7) with $w_{l}$ defined as in (2.8) where $x>0$. Set $v=\max _{1 \leq i \leq n} \max _{1 \leq l \leq g}\left\|z_{i}^{l}\right\|_{2}$. Let assumptions $\left(\mathbf{B}_{\mathbf{2}}\right)$ and $\left(\mathbf{R E}_{\mathbf{2}}\right)$ be satisfied with

$$
a_{0}=\frac{3 \max _{1 \leq l \leq g} \omega_{l}}{\min _{1 \leq l \leq g} \omega_{l}}
$$

If $r\left(1+a_{0}\right)^{2} \max _{1 \leq l \leq g} \omega_{l} \leq \frac{\mu_{2}^{2}}{3 v|J|}$, with probability at least $1-2 \exp (-x)$ we have

$$
\begin{align*}
& R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right) \leq \frac{9\left(1+a_{0}\right)^{2} J\left(\beta_{0}\right) \mid r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{\mu_{2}^{2}(s, 3)}  \tag{2.13}\\
&\left\|\hat{\beta}_{G L}-\beta_{0}\right\|_{2,1} \leq \frac{6\left(1+a_{0}\right)^{2}\left|J\left(\beta_{0}\right)\right| r\left(\max _{1 \leq l \leq g} \omega_{l}\right)}{\mu_{2}^{2}(s, 3)}  \tag{2.14}\\
&\left\|\hat{\beta}_{G L}-\beta_{0}\right\|_{2, q}^{q} \leq\left(\frac{6\left(1+a_{0}\right)^{2}\left|J\left(\beta_{0}\right)\right| r\left(\max _{1 \leq l \leq g} \omega_{l}\right)}{\mu_{2}^{2}(s, 3)}\right)^{q} \quad \text { for all } 1<q \leq 2 \tag{2.15}
\end{align*}
$$

Moreover if we assume that there exists $0<\epsilon_{0} \leq 1 / 2$ such that

$$
\epsilon_{0} \leq \pi\left(f_{\beta_{0}}\left(z_{i}\right)\right)\left[1-\pi\left(f_{\beta_{0}}\left(z_{i}\right)\right)\right] \quad \text { for all } i=1, \ldots, n
$$

then,

$$
\begin{equation*}
\left\|X \hat{\beta}_{G L}-X \beta_{0}\right\|_{n}^{2} \leq \frac{36\left(1+a_{0}\right)^{2}\left|J\left(\beta_{0}\right)\right| r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{\mu^{2}(s, 3) \epsilon_{0}} \tag{2.16}
\end{equation*}
$$

Inequalities (2.14) and (2.15) are the extensions of the results in [2] for the Lasso to Group Lasso in logistic regression model.

In this section we studied some properties of the Group Lasso. However the Group Lasso is based on prior knowledge that the set of relevant predictors have known group structure. If this group sparsity condition is not satisfied, the sparsity can be achieve by simply using the Lasso. We will show in the next section how to adapt the results of this section to the Lasso.

## 3. LASSO FOR LOGISTIC REGRESSION

### 3.1. Estimation procedure

The Lasso estimator $f_{\hat{\beta}_{L}}$ is defined as a minimizer of the following $\ell_{1}$-penalized empirical risk

$$
\begin{equation*}
f_{\hat{\beta}_{L}}:=\underset{f_{\beta} \in \Gamma}{\operatorname{argmin}}\left\{\hat{R}\left(f_{\beta}\right)+r \sum_{j=1}^{p} \omega_{j}\left|\beta_{j}\right|\right\} \tag{3.1}
\end{equation*}
$$

where the minimum is taken over the set

$$
\Gamma \subseteq\left\{f_{\beta}(.)=\sum_{j=1}^{p} \beta_{j} \phi_{j}(.), \beta=\left(\beta_{1}, \ldots, \beta_{p}\right) \in \mathbb{R}^{p}\right\}
$$

and $\omega_{j}$ are positive weights to be specified later. The "classical" Lasso penalization corresponds to $\omega_{j}=1$, where $r$ 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.

### 3.2. Oracle inequalities

In this section we provide non asymptotic oracle inequalities for the Lasso in logistic regression model.
Theorem 3.1. Let $f_{\hat{\beta}_{L}}$ be the $\ell_{1}$-penalized minimum defined in (3.1). Let assumption $\left(\mathbf{B}_{\mathbf{2}}\right)$ be satisfied.
(A) Let $x>0$ be fixed and $r \geq 1$. For $j=\{1, \ldots, p\}$, let

$$
\begin{equation*}
\omega_{j}=\frac{2}{n} \sqrt{\frac{1}{2} \sum_{i=1}^{n} \phi_{j}^{2}\left(z_{i}\right)(x+\log p)}+\frac{2 c_{2}(x+\log p)}{3 n} \tag{3.2}
\end{equation*}
$$

Thus with probability at least $1-2 \exp (-x)$ we have

$$
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) \leq \inf _{\beta \in \mathbb{R}^{p}}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+2\|\beta\|_{1} r \max _{1 \leq j \leq p} \omega_{j}\right\}
$$

(B) Let $A>2 \sqrt{c_{2}}$. For $j=\{1, \ldots, p\}$, let $\omega_{j}=1$, and

$$
r=A \sqrt{\frac{\log p}{n}}
$$

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

$$
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) \leq \inf _{\beta \in \mathbb{R}^{p}}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 A\|\beta\|_{1} r \sqrt{\frac{\log p}{n}}\right\}
$$

As previously, the variance terms are of order $\|\beta\|_{1} \sqrt{\log p / n}$ for any $\beta$. Hence these are sharp oracle inequalities with "slow" rates. These results are obtained without any assumption on the Gram matrix. To obtain oracle inequalities with a "fast rate", of order $\log p / n$, we need the restricted eigenvalue condition.

For some integer $s$ such that $1 \leq s \leq p$ and a positive number $a_{0}$, the following condition holds
$\left(\mathbf{R E}_{3}\right)$

$$
\mu\left(s, a_{0}\right):=\min _{K \subseteq\{1, \ldots p\}:|K| \leq s} \min _{\Delta \neq 0:\left\|\Delta_{K^{c}}\right\|_{1} \leq a_{0}\left\|\Delta_{K}\right\|_{1}} \frac{\|X \Delta\|_{2}}{\sqrt{n}\left\|\Delta_{K}\right\|_{2}}>0
$$

This assumption has been introduced in [4], 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 [6-8] are more restrictive than restricted eigenvalue assumption. In those papers either $\Phi_{n}$ is positive definite, or mutual coherence condition is imposed. We refer to [36] for a complete comparison of the assumptions used to prove oracle inequality for the Lasso. Especially it is proved that restricted eigenvalue assumption is weaker than the neighborhood stability or irrepresentable condition.

Theorem 3.2. Let $f_{\hat{\beta}_{L}}$ be the $\ell_{1}$-penalized minimum defined in (3.1). Fix $\eta>0$ and $1 \leq s \leq p$. Assume that $\left(\mathbf{B}_{\mathbf{1}}\right),\left(\mathbf{B}_{\mathbf{2}}\right),\left(\mathbf{B}_{\mathbf{3}}\right)$ and $\left(\mathbf{R E}_{\mathbf{3}}\right)$ are satisfied, with $a_{0}=3+4 / \eta$.
(A) Let $x>0$ be fixed and $r \geq 1$. For $j=\{1, \ldots, p\}$, $\omega_{j}$ defined as in (3.2). Thus with probability at least $1-2 \exp (-x)$ we have

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) \leq(1+\eta) \inf _{f_{\beta} \in \Gamma}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{c(\eta)|K(\beta)| r^{2}\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}}{c_{0} \epsilon_{0} \mu^{2}(s, 3+4 / \eta)}\right\} \tag{3.3}
\end{equation*}
$$

and

$$
\begin{equation*}
\left\|f_{\hat{\beta}_{L}}-f_{0}\right\|_{n}^{2} \leq \frac{c_{0}^{\prime}}{4 c_{0} \epsilon_{0}}(1+\eta) \inf _{f_{\beta} \in \Gamma}\left\{\left\|f_{\beta}-f_{0}\right\|_{n}^{2}+\frac{4 c(\eta)|K(\beta)| r^{2}\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}}{c_{0}^{\prime} c_{0} \epsilon_{0}^{2} \mu^{2}(s, 3+4 / \eta)}\right\} \tag{3.4}
\end{equation*}
$$

(B) Let $A>2 \sqrt{c_{2}}$. For $j=\{1, \ldots, p\}$, let $\omega_{j}=1$, and

$$
r=A \sqrt{\frac{\log p}{n}}
$$

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

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) \leq(1+\eta) \inf _{f_{\beta} \in \Gamma}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{A^{2} c(\eta)}{c_{0} \epsilon_{0} \mu^{2}(s, 3+4 / \eta)} \frac{|K(\beta)| r^{2} \log p}{n}\right\} \tag{3.5}
\end{equation*}
$$

and

$$
\begin{equation*}
\left\|f_{\hat{\beta}_{L}}-f_{0}\right\|_{n}^{2} \leq \frac{c_{0}^{\prime}}{4 c_{0} \epsilon_{0}}(1+\eta) \inf _{f_{\beta} \in \Gamma}\left\{\left\|f_{\beta}-f_{0}\right\|_{n}^{2}+\frac{4 c(\eta) A^{2}}{c_{0}^{\prime} c_{0} \epsilon_{0}^{2} \mu^{2}(s, 3+4 / \eta)} \frac{|K(\beta)| r^{2} \log p}{n}\right\} \tag{3.6}
\end{equation*}
$$

In both cases $c(\eta)$ is a constant depending only on $\eta ; c_{0}=c_{0}\left(C_{0}, c_{1}\right)$ and $c_{0}^{\prime}=c_{0}^{\prime}\left(C_{0}, c_{1}\right)$ are constants depending on $C_{0}$ and $c_{1}$; and $\epsilon_{0}=\epsilon_{0}\left(c_{1}\right)$ is a constant depending on $c_{1}$.

In this theorem the variance terms are 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 [4] for linear regression model. For the best
of our knowledge, (3.4) and (3.6) are the first non asymptotic oracle inequalities for the $L_{2}\left(\frac{1}{n} \sum_{i}^{n} \delta_{z_{i}}\right)$ norm in logistic model. Some non asymptotic oracle inequalities for excess risk like (3.3) or (3.5) have been established in [35] 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 results hold under $R E$ condition, which can be seen as empirical version of assumption C in [35]. The confidence (probability that result holds true) of Inequality (3.3) does not depend on $n$ or $p$ while the confidence of her results depends on $n$ and $p$. Moreover, the weights we proposed from Bernstein's inequality are different and easy to interpret.

### 3.3. Special case: $f_{0}$ linear

In this section we assume that $f_{0}$ is a linear function that is $f_{0}\left(z_{i}\right)=f_{\beta_{0}}\left(z_{i}\right)=\sum_{j=1}^{p} \beta_{0 j} z_{i j}=z_{i}^{T} \beta_{0}$, where $z_{i}=\left(z_{i 1}, \ldots, z_{i p}\right)^{T}$. Denote $X=\left(z_{i j}\right)_{1 \leq i \leq n, 1 \leq j \leq p}$ the design matrix. Thus for $i=1, \ldots, n$

$$
\begin{equation*}
\mathbb{P}\left(Y_{i}=1\right)=\pi\left(z_{i}^{T} \beta_{0}\right)=\frac{\exp \left(z_{i}^{T} \beta_{0}\right)}{1+\exp \left(z_{i}^{T} \beta_{0}\right)} . \tag{3.7}
\end{equation*}
$$

The Lasso estimator of $\beta_{0}$ is thus defined as

$$
\begin{equation*}
\hat{\beta}_{L}:=\underset{\beta: f_{\beta} \in \Gamma}{\operatorname{argmin}}\left\{\frac{1}{n} \sum_{i=1}^{n}\left\{\log \left(1+\exp \left(z_{i}^{T} \beta\right)\right)-Y_{i} z_{i}^{T} \beta\right\}+r \sum_{j=1}^{p} \omega_{j}\left|\beta_{j}\right|\right\} . \tag{3.8}
\end{equation*}
$$

When the design matrix $X$ has full rank, the solution of optimization problem (3.8) is usually unique. When $p \gg n$ this infimum might not be unique.

Corollary 3.3. Let assumption $R E(s, 3)$ be satisfied and $\left|K\left(\beta_{0}\right)\right| \leq s$, where $1 \leq s \leq p$. Consider the Lasso estimator $f_{\hat{\beta}_{L}}$ defined by (3.8) with

$$
\omega_{j}=\frac{2}{n} \sqrt{\frac{1}{2} \sum_{i=1}^{n} z_{i j}^{2}(x+\log p)}+\frac{2 c_{2}(x+\log p)}{3 n} .
$$

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

$$
\begin{align*}
& R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{\beta_{0}}\right) \leq \frac{9 s r^{2}\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}}  \tag{3.9}\\
&\left\|f_{\hat{\beta}_{L}}-f_{\beta_{0}}\right\|_{n}^{2} \leq \frac{9 s^{2} r^{2}\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}}{\mu^{2}(s, 3) c_{0}^{2} \epsilon_{0}^{2}}  \tag{3.10}\\
&\left\|\hat{\beta}_{L}-\beta_{0}\right\|_{1} \leq \frac{12 s r\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}\left(\min _{1 \leq j \leq p} \omega_{j}\right)}  \tag{3.11}\\
&\left\|\hat{\beta}_{L}-\beta_{0}\right\|_{q}^{q} \leq\left(\frac{12 s r\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}\left(\min _{1 \leq j \leq p} \omega_{j}\right)}\right)^{q} \quad \text { for all } 1<q \leq 2 . \tag{3.12}
\end{align*}
$$

If $r=A \sqrt{\log p / n}$ and $\omega_{j}=1$ for all $j \in\{1, \ldots, p\}$ we have the same results with probability at least $1-$ $2 p^{1-A^{2} / 4 c_{2}}$.

Line (3.9) and Line (3.11) of the corollary are similar to those of Theorem 5 in [2]. Note that, up to differences in constant factors, the rates obtained in this corollary are the same as those obtained in Theorem 7.2 in [4] for linear model with an s-sparse vector. Remark 2.6 remains true in this section.

## 4. Conclusion

In this paper we stated non asymptotic oracle inequalities for the Lasso and Group Lasso. Our results are non asymptotic: the number $n$ of observations is fixed while the number $p$ of covariates can grow with respect to $n$ and can be much larger than $n$. The major difference with existing results concerning Group Lasso or Lasso for logistic regression model is that we do not assume that $f_{0}$ is linear. First we provided sharp oracle inequalities for excess risk, with "slow" rates, with no assumption on the Gram matrix, on the regressors nor on the margin. Secondly, under RE condition we provided "fast" oracle inequalities for excess risk and $L_{2}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_{i}}\right)$ loss. We also provided as a consequence of oracle inequalities the bounds for excess risk, $L_{2}\left(\frac{1}{n} \sum_{i=1}^{n} \delta_{z_{i}}\right)$ error and estimation error in the case where the true function $f_{0}$ is linear ("usual" logistic regression (1.1)).

## 5. Proofs of main results

### 5.1. Proof of Theorem 2.1

Since $\hat{\beta}_{G L}$ is the minimizer of $\hat{R}\left(f_{\beta}\right)+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2}$, we get

$$
R\left(f_{\hat{\beta}_{G L}}\right)-\frac{1}{n} \varepsilon^{T} X \hat{\beta}_{G L}+r \sum_{l=1}^{g} \omega_{l}\left\|\hat{\beta}_{G L}^{l}\right\|_{2} \leq R\left(f_{\beta}\right)-\frac{1}{n} \varepsilon^{T} X \beta+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2},
$$

where $\epsilon=\left(\epsilon_{1}, \ldots, \epsilon_{n}\right)^{T}$ with $\epsilon_{i}=Y_{i}-\mathbb{E}\left[Y_{i}\right]$ for $i=1, \ldots, n$. By applying Cauchy-Schwarz inequality, we obtain

$$
\begin{align*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq & R\left(f_{\beta}\right)-R\left(f_{0}\right)+\sum_{l=1}^{g} \frac{1}{n} \sqrt{\sum_{j \in G_{l}}\left(\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \epsilon_{i}\right)^{2}}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} \\
& +r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2}-r \sum_{l=1}^{g} \omega_{l}\left\|\hat{\beta}_{G L}^{l}\right\|_{2} . \tag{5.1}
\end{align*}
$$

Set $Z_{l}=n^{-1} \sqrt{\sum_{j \in G_{l}}\left(\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \epsilon_{i}\right)^{2}}$, for $l \in\{1, \ldots, g\}$ and the event

$$
\begin{equation*}
\mathcal{A}=\bigcap_{l=1}^{g}\left\{Z_{l} \leq r \omega_{l} / 2\right\} . \tag{5.2}
\end{equation*}
$$

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

$$
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+r \sum_{l=1}^{g} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2}-r \sum_{l=1}^{g} \omega_{l}\left\|\hat{\beta}_{G L}^{l}\right\|_{2} .
$$

This implies that

$$
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2} .
$$

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

$$
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq \inf _{\beta \in \mathbb{R}^{p}}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r\|\beta\|_{2,1} \max _{1 \leq l \leq g} \omega_{l}\right\}
$$

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

$$
\begin{align*}
\mathbb{P}\left(\mathcal{A}^{c}\right) & =\mathbb{P}\left(\bigcup_{l=1}^{g}\left\{\sqrt{\sum_{j \in G_{l}}\left(\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \epsilon_{i}\right)^{2}}>n r \omega_{l} / 2\right\}\right)  \tag{5.3}\\
& \leq \sum_{l=1}^{g} \mathbb{P}\left(\sqrt{\sum_{j \in G_{l}}\left(\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \epsilon_{i}\right)^{2}}>n r \omega_{l} / 2\right) . \tag{5.4}
\end{align*}
$$

For $j \in G_{l}$ set $T_{j}^{l}=\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \epsilon_{i}$, we have

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{A}^{c}\right) & \leq \sum_{l=1}^{g} \mathbb{P}\left(\sqrt{\sum_{j \in G_{l}}\left(T_{j}^{l}\right)^{2}}>n r \omega_{l} / 2\right) \\
& =\sum_{l=1}^{g} \mathbb{P}\left(\sum_{j \in G_{l}}\left(T_{j}^{l}\right)^{2}>\left(n r \omega_{l}\right)^{2} / 4\right) .
\end{aligned}
$$

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

$$
\begin{equation*}
\left\{\sum_{j \in G_{l}}\left(T_{j}^{l}\right)^{2}>\left(n r \omega_{l}\right)^{2} / 4\right\} \subset \cup_{j \in G_{l}}\left\{\left(T_{j}^{l}\right)^{2}>\frac{\left(n r \omega_{l}\right)^{2}}{4\left|G_{l}\right|}\right\} \tag{5.5}
\end{equation*}
$$

it follows that

$$
\mathbb{P}\left(\mathcal{A}^{c}\right) \leq \sum_{l=1}^{g} \sum_{j \in G_{l}} \mathbb{P}\left(\left|T_{j}^{l}\right|>\frac{n r \omega_{l}}{2 \sqrt{\left|G_{l}\right|}}\right)
$$

For $j \in G_{l}$, set $v_{j}^{l}=\sum_{i=1}^{n} \mathbb{E}\left(\phi_{j}^{2} \epsilon_{i}^{2}\right)$. Since $\sum_{i=1}^{n} \phi_{j}^{2}\left(z_{i}\right) \geqslant 4 v_{j}^{l}$, we have

$$
\mathbb{P}\left(\left|T_{j}^{l}\right|>\frac{n r \omega_{l}}{2 \sqrt{\left|G_{l}\right|}}\right) \leq \mathbb{P}\left(\left|T_{j}^{l}\right|>\sqrt{2 v_{j}^{l}(x+\log p)}+\frac{c_{2}}{3}(x+\log p)\right), \quad r \geq 1
$$

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

$$
\mathbb{P}\left(\left|T_{j}^{l}\right|>\frac{n \omega_{l}}{2 \sqrt{\left|G_{l}\right|}}\right) \leq 2 \exp (-x-\log p)
$$

It follows that

$$
\begin{equation*}
\mathbb{P}\left(\mathcal{A}^{c}\right) \leq \sum_{l=1}^{g} \sum_{j \in G_{l}} \mathbb{P}\left(\left|T_{j}^{l}\right|>\frac{n \omega_{l}}{2 \sqrt{\left|G_{l}\right|}}\right) \leq 2 \exp (-x) \tag{5.6}
\end{equation*}
$$

This ends the proof of the Theorem 2.1.

### 5.2. Proof of Theorem 2.2

Fix an arbitrary $\beta \in \mathbb{R}^{p}$ such that $f_{\beta} \in \Gamma$. Set $\delta=W\left(\hat{\beta}_{G L}-\beta\right)$ where $W=\operatorname{Diag}\left(W_{1}, \ldots, W_{p}\right)$ is a block diagonal matrix, with $W_{l}=\operatorname{Diag}\left(\omega_{l}, \ldots, \omega_{l}\right)$. Since $\hat{\beta}_{G L}$ is the minimizer of $\hat{R}\left(f_{\beta}\right)+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2}$, we get

$$
R\left(f_{\hat{\beta}_{G L}}\right)-\frac{1}{n} \varepsilon^{T} X \hat{\beta}_{G L}+r \sum_{l=1}^{g} \omega_{l}\left\|\hat{\beta}_{G L}^{l}\right\|_{2} \leq R\left(f_{\beta}\right)-\frac{1}{n} \varepsilon^{T} X \beta+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2} .
$$

On the event $\mathcal{A}$ defined in (5.2), adding the term $\frac{r}{2} \sum_{l=1}^{g} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}$ to both sides of inequality (5.1) yields to

$$
R\left(f_{\hat{\beta}_{G L}}\right)+\frac{r}{2} \sum_{l=1}^{g} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} \leq R\left(f_{\beta}\right)+r \sum_{l=1}^{g} \omega_{l}\left(\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}-\left\|\hat{\beta}_{G L}^{l}\right\|_{2}+\left\|\beta^{l}\right\|_{2}\right) .
$$

Since $\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}-\left\|\hat{\beta}_{G L}^{l}\right\|_{2}+\left\|\beta^{l}\right\|_{2}=0$ for for $l \notin J(\beta)=J$, we have

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right)+\frac{r}{2} \sum_{l=1}^{g} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} \tag{5.7}
\end{equation*}
$$

we get from equation (5.7) that

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} . \tag{5.8}
\end{equation*}
$$

Consider separately the two events:

$$
\mathcal{A}_{1}=\left\{2 r \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} \leq \eta\left(R\left(f_{\beta}\right)-R\left(f_{0}\right)\right)\right\},
$$

and

$$
\begin{equation*}
\mathcal{A}_{1}^{c}=\left\{\eta\left(R\left(f_{\beta}\right)-R\left(f_{0}\right)\right)<2 r \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}\right\} . \tag{5.9}
\end{equation*}
$$

On the event $\mathcal{A} \cap \mathcal{A}_{1}$, we get from (5.8)

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq(1+\eta)\left(R\left(f_{\beta}\right)-R\left(f_{0}\right)\right), \tag{5.10}
\end{equation*}
$$

and the result follows. On the event $\mathcal{A} \cap \mathcal{A}_{1}^{c}$, all the following inequalities are valid. On one hand, by applying Cauchy-Schwarz inequality, we get from (5.8) that

$$
\begin{align*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) & \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r \sqrt{|J(\beta)|} \sqrt{\sum_{l \in J} \omega_{l}^{2}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}^{2}} \\
& \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r \sqrt{|J(\beta)|}\left\|\delta_{J}\right\|_{2} . \tag{5.11}
\end{align*}
$$

On the other hand we get from equation (5.7) that

$$
\frac{1}{2} \sum_{l=1}^{g} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2},
$$

and using (5.9) we obtain

$$
\frac{1}{2} \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}+\frac{1}{2} \sum_{l \in J^{c}} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2} \leq \frac{2}{\eta} \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}+2 \sum_{l \in J} \omega_{l}\left\|\left(\hat{\beta}_{G L}-\beta\right)^{l}\right\|_{2}
$$

which implies

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

We can therefore apply assumption $\left(\mathbf{R E}_{\mathbf{1}}\right)$ with $a_{0}=3+4 / \eta$, and conclude that

$$
\begin{equation*}
\mu_{1}^{2}\left\|\delta_{J}\right\|_{2}^{2} \leq \frac{\|X \delta\|_{2}^{2}}{n}=\frac{1}{n}\left(\hat{\beta}_{G L}-\beta\right)^{T} W X^{T} X W\left(\hat{\beta}_{G L}-\beta\right) \leq\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}\left\|f_{\hat{\beta}_{G L}}-f_{\beta}\right\|_{n}^{2} \tag{5.12}
\end{equation*}
$$

Gathering equations (5.11) and (5.12) we get

$$
\begin{aligned}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) & \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r\left(\max _{1 \leq l \leq g} \omega_{l}\right) \sqrt{|J(\beta)|} \mu_{1}^{-1}\left\|f_{\hat{\beta}_{G L}}-f_{\beta}\right\|_{n} \\
& \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r\left(\max _{1 \leq l \leq g} \omega_{l}\right) \sqrt{|J(\beta)|} \mu_{1}^{-1}\left(\left\|f_{\hat{\beta}_{G L}}-f_{0}\right\|_{n}+\left\|f_{\beta}-f_{0}\right\|_{n}\right)
\end{aligned}
$$

We now use Lemma 5.1 which compares excess risk to empirical norm.
Lemma 5.1. Under assumptions $\left(\mathbf{B}_{\mathbf{1}}\right)$ and $\left(\mathbf{B}_{\mathbf{3}}\right)$ we have

$$
c_{0} \epsilon_{0}\left\|f_{\beta}-f_{0}\right\|_{n}^{2} \leq R\left(f_{\beta}\right)-R\left(f_{0}\right) \leq \frac{1}{4} c_{0}^{\prime}\left\|f_{\beta}-f_{0}\right\|_{n}^{2}
$$

where $c_{0}$ and $c_{0}^{\prime}$ are constants depending on $C_{0}$; and $\epsilon_{0}$ is a constant depending on $c_{1}$ and $c_{2}$.
(See the Appendix for the proof of Lem. 5.1).
Consequently

$$
\begin{aligned}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq & R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{2 r\left(\max _{1 \leq l \leq g} \omega_{l}\right) \sqrt{|J(\beta)|} \mu_{1}^{-1}}{\sqrt{c_{0} \epsilon_{0}}} \sqrt{R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right)} \\
& +\frac{2 r\left(\max _{1 \leq l \leq g} \omega_{l}\right) \sqrt{|J(\beta)|} \mu_{1}^{-1}}{\sqrt{c_{0} \epsilon_{0}}} \sqrt{R\left(f_{\beta}\right)-R\left(f_{0}\right)} .
\end{aligned}
$$

Using inequality $2 u v<u^{2} / b+b v^{2}$ for all $b>1$, with $u=r\left(\max _{1 \leq l \leq g} \omega_{l}\right) \frac{\sqrt{|J(\beta)|} \mu_{1}^{-1}}{\sqrt{c_{0} \epsilon_{0}}}$ and $v$ being either $\sqrt{R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right)}$ or $\sqrt{R\left(f_{\beta}\right)-R\left(f_{0}\right)}$ we have

$$
\begin{aligned}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq & R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 b\left(\frac{r\left(\max _{1 \leq l \leq g} \omega_{l}\right) \sqrt{|J(\beta)|} \mu_{1}^{-1}}{\sqrt{c_{0} \epsilon_{0}}}\right)^{2} \\
& +\frac{R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right)}{b}+\frac{R\left(f_{\beta}\right)-R\left(f_{0}\right)}{b}
\end{aligned}
$$

This implies that

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq \frac{b+1}{b-1}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{2 b^{2} r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}|J(\beta)|}{(b+1) \mu_{1}^{2} c_{0} \epsilon_{0}}\right\} \tag{5.13}
\end{equation*}
$$

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

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq(1+\eta)\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{c(\eta) r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}|J(\beta)|}{\mu_{1}^{2} c_{0} \epsilon_{0}}\right\} \tag{5.14}
\end{equation*}
$$

According to inequalities (5.10) and (5.14) we conclude that on event $\mathcal{A}$,

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{0}\right) \leq(1+\eta)\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{c(\eta) r^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}|J(\beta)|}{\mu_{1}^{2} c_{0} \epsilon_{0}}\right\} \tag{5.15}
\end{equation*}
$$

where $c(\eta)=2(1+2 / \eta)^{2} /(2+2 / \eta)$. Inequality (2.4) of the Theorem 2.2 follows. Inequality (2.5) follows from Lemma 5.1. This ends the proof of the Theorem 2.2 by considering (5.6).

### 5.3. Proof of Corollary 2.5

Set $\delta=W\left(\hat{\beta}_{G L}-\beta_{0}\right)$, Line (2.9) of Corollary 2.5 follows directly from equation (5.15) with $\beta=\beta_{0}$ and $\eta=1$. Note that on the event $\mathcal{A}$ defined in (5.2), we have

$$
\begin{equation*}
\left\|\delta_{J\left(\beta_{0}\right)^{c}}\right\|_{2,1} \leq 3\left\|\delta_{J\left(\beta_{0}\right)}\right\|_{2,1} \tag{5.16}
\end{equation*}
$$

Indeed, since $\hat{\beta}_{G L}$ is the minimizer of $\hat{R}\left(f_{\beta}\right)+r \sum_{l=1}^{g} \omega_{l}\left\|\beta^{l}\right\|_{2}$,

$$
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right)+r \sum_{l=1}^{g} \omega_{l}\left\|\hat{\beta}_{G L}^{l}\right\|_{2} \leq \frac{1}{n} \varepsilon^{T} X\left(\hat{\beta}_{G L}-\beta_{0}\right)+r \sum_{l=1}^{g} \omega_{l}\left\|\beta_{0}^{l}\right\|_{2}
$$

which implies

$$
r\left\|W \hat{\beta}_{G L}\right\|_{2,1} \leq \sum_{l=1}^{g} \frac{1}{n} \sqrt{\sum_{j \in G_{l}}\left(\sum_{i=1}^{n}\left(z_{i j}\right) \epsilon_{i}\right)^{2}}\left\|\left(\hat{\beta}_{G L}-\beta_{0}\right)^{l}\right\|_{2}+r\left\|W \beta_{0}\right\|_{2,1}
$$

On the event $A$ we have

$$
\begin{aligned}
\left\|W\left(\hat{\beta}_{G L}\right)_{J\left(\beta_{0}\right)}\right\|_{2,1}+\left\|W\left(\hat{\beta}_{G L}\right)_{J^{c}\left(\beta_{0}\right)}\right\|_{2,1} \leq & \frac{1}{2}\left(\left\|W\left(\hat{\beta}_{G L}-\beta_{0}\right)_{J\left(\beta_{0}\right)}\right\|_{2,1}+\left\|W\left(\hat{\beta}_{G L}\right)_{J^{c}\left(\beta_{0}\right)}\right\|_{2,1}\right) \\
& +\left\|W\left(\beta_{0}\right)_{J\left(\beta_{0}\right)}\right\|_{2,1}
\end{aligned}
$$

This yields to (5.16). Line (2.10) follows from line (2.9) by applying Lemma 5.1. Line (2.11) follows from line (2.10) by using Equation (5.12) and $\|\delta\|_{2,1}^{2} \leq 16 s\left\|\delta_{J\left(\beta_{0}\right)}\right\|_{2}^{2}$. Line (2.12) is the consequence of the Lemma A. 3 with $a_{l}=\left\|\left(\hat{\beta}_{G L}-\beta_{0}\right)^{l}\right\|_{2}$ and

$$
b_{1}=\frac{12 r s\left(\max _{1 \leq l \leq g} \omega_{l}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}\left(\min _{1 \leq l \leq g} \omega_{l}\right)}
$$

### 5.4. Proof of Theorem 2.8

On the event $\mathcal{A}$ defined in (5.2), using inequality (5.1) with $\beta=\beta_{0}$ yields

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right) \leq \sum_{l=1}^{g} \frac{3 r \omega_{l}}{2}\left\|\left(\hat{\beta}_{G L}-\beta_{0}\right)^{l}\right\|_{2} . \tag{5.17}
\end{equation*}
$$

By Lemma A. 1 we have,

$$
\begin{equation*}
\frac{\langle h, h\rangle_{f_{\beta_{0}}}}{\|h\|_{\infty}^{2}}\left(\exp \left(-\|h\|_{\infty}\right)+\|h\|_{\infty}-1\right) \leq R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right) \tag{5.18}
\end{equation*}
$$

where

$$
h\left(z_{i}\right)=\left(f_{\hat{\beta}_{G L}}-f_{\beta_{0}}\right)\left(z_{i}\right)=\sum_{l=1}^{g} \sum_{j \in G_{l}}\left(\hat{\beta}_{G L, j}-\beta_{0 j}\right) z_{i j}
$$

One can easily verify that $\|h\|_{\infty} \leq v\left\|\delta^{\prime}\right\|_{2,1}$ with $\delta^{\prime}=\hat{\beta}_{G L}-\beta_{0}$. Equation (5.18) and the decreasing of $t \mapsto$ $\frac{\exp (-t)+t-1}{t^{2}}$ lead to

$$
\frac{\delta^{\prime} T X^{T} D X \delta^{\prime}}{n\left(v\left\|\delta^{\prime}\right\|_{2,1}\right)^{2}}\left(\exp \left(-v\left\|\delta^{\prime}\right\|_{2,1}\right)+v\left\|\delta^{\prime}\right\|_{2,1}-1\right) \leq R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right)
$$

Now, inequality (5.16) implies

$$
\left\|\delta_{J\left(\beta_{0}\right)^{c}}^{\prime}\right\|_{2,1} \leq 3 \frac{\left(\max _{1 \leq l \leq g} \omega_{l}\right)}{\min _{1 \leq l \leq g} \omega_{l}}\left\|\delta_{J\left(\beta_{0}\right)}^{\prime}\right\|_{2,1}
$$

We can therefore apply assumption $\left(\mathbf{R E}_{\mathbf{2}}\right)$ with $a_{0}=3\left(\max _{1 \leq l \leq g} \omega_{l}\right) / \min _{1 \leq l \leq g} \omega_{l}$ and get that

$$
\frac{\mu_{2}^{2}\left\|\delta_{J}^{\prime}\right\|_{2}^{2}}{v^{2}\left\|\delta^{\prime}\right\|_{2,1}^{2}}\left(\exp \left(-v\left\|\delta^{\prime}\right\|_{2,1}\right)+v\left\|\delta^{\prime}\right\|_{2,1}-1\right) \leq R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right)
$$

We can use that $\left\|\delta^{\prime}\right\|_{2,1}^{2} \leq\left(1+a_{0}\right)^{2}|J|\left\|\delta_{J}^{\prime}\right\|_{2}^{2}$, with $J=J\left(\beta_{0}\right)$ to write

$$
\frac{\mu_{2}^{2}}{\left(1+a_{0}\right)^{2}|J| v^{2}}\left(\exp \left(-v\left\|\delta^{\prime}\right\|_{2,1}\right)+v\left\|\delta^{\prime}\right\|_{2,1}-1\right) \leq R\left(f_{\hat{\beta}_{G L}}\right)-R\left(f_{\beta_{0}}\right)
$$

According to equation (5.17) we have

$$
\begin{equation*}
\exp \left(-v\left\|\delta^{\prime}\right\|_{2,1}\right)+v\left\|\delta^{\prime}\right\|_{2,1}-1 \leq \frac{3 r\left(1+a_{0}\right)^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right) v^{2}|J|}{2 \mu_{2}^{2}}\left\|\delta^{\prime}\right\|_{2,1} \tag{5.19}
\end{equation*}
$$

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

$$
\begin{equation*}
\mathrm{e}^{\frac{-2 a}{1-a}}+(1-a) \frac{2 a}{1-a}-1 \geqslant 0 \tag{5.20}
\end{equation*}
$$

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

$$
\begin{equation*}
\mathrm{e}^{-v\left\|\delta^{\prime}\right\|_{2,1}}+v\left\|\delta^{\prime}\right\|_{2,1}-1 \geqslant \frac{v^{2}\left\|\delta^{\prime}\right\|_{2,1}^{2}}{v\left\|\delta^{\prime}\right\|_{2,1}+2} \tag{5.21}
\end{equation*}
$$

This implies using equation (5.19) that

$$
v\left\|\delta^{\prime}\right\|_{2,1} \leq \frac{3 r\left(1+a_{0}\right)^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)|J| v / \mu_{2}^{2}}{1-3 r\left(1+a_{0}\right)^{2}\left(\max _{1 \leq l \leq g} \omega_{l}\right)|J| v / 2 \mu_{2}^{2}}
$$

Now if $r\left(1+a_{0}\right)^{2} \max _{1 \leq l \leq g} \omega_{l} \leq \frac{\mu_{2}^{2}}{3 v \mid J}$, we have $v\left\|\delta^{\prime}\right\|_{2,1} \leq 2$ and consequently

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

Now, inequality (5.19) implies

$$
\left\|\delta^{\prime}\right\|_{2,1} \leq \frac{6\left(1+a_{0}\right)^{2}|J| r\left(\max _{1 \leq l \leq g} \omega_{l}\right)}{\mu_{2}^{2}} .
$$

This proves the line (2.14). Line (2.13) follows from (2.14) by using inequality (5.17). Line (2.15) is the consequence of Lemma A. 3 taking $a_{l}=\left\|\left(\hat{\beta}_{G L}-\beta_{0}\right)^{l}\right\|_{2}$ and $b_{1}=6\left(1+a_{0}\right)^{2}|J| r\left(\min _{1 \leq l \leq g} \omega_{l}\right) / \mu_{2}^{2}(s, 3)$. Line (2.16) follows from line (2.13) and inequality (5.18).

### 5.5. Proof of Theorem 3.1

Note that Lasso can be derived by Group Lasso by taking one predictor per group i.e. $p=g$ and $G_{j}=\{j\}$ for $j \in\{1, \ldots, p\}$. This implies, using (5.1) that

$$
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+\sum_{j=1}^{p}\left|\frac{1}{n} \sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \varepsilon_{i}\right|\left|\hat{\beta}_{L, j}-\beta_{j}\right|+r \sum_{j=1}^{p} \omega_{j}\left|\beta_{j}\right|-r \sum_{j=1}^{p} \omega_{j}\left|\hat{\beta}_{L, j}\right| .
$$

For $1 \leq j \leq p$, set $S_{j}=\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \varepsilon_{i}$ and let us denote by $E$, the event

$$
\begin{equation*}
E=\bigcap_{j=1}^{p}\left\{\left|S_{j}\right| \leq n r \omega_{j} / 2\right\} . \tag{5.22}
\end{equation*}
$$

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

$$
\begin{aligned}
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) & \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+r \sum_{j=1}^{p} \omega_{j}\left|\hat{\beta}_{L, j}-\beta_{j}\right|+r \sum_{j=1}^{p} \omega_{j}\left|\beta_{j}\right|-r \sum_{j=1}^{p} \omega_{j}\left|\hat{\beta}_{L, j}\right| \\
& \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r \sum_{j=1}^{p} \omega_{j}\left|\beta_{j}\right| .
\end{aligned}
$$

We conclude that on the event $E$ we have

$$
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) \leq \inf _{\beta \in \mathbb{R}^{p}}\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+2 r\|\beta\|_{1} \max _{1 \leq j \leq p} \omega_{j}\right\}
$$

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

$$
\begin{aligned}
\mathbb{P}\left(\mathrm{e}^{c}\right) & \leq \mathbb{P}\left(\bigcup_{j=1}^{p}\left\{\left|\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right)\left(Y_{i}-\mathbb{E}\left(Y_{i}\right)\right)\right|>r \omega_{j} n / 2\right\}\right) \\
& \leq \sum_{j=1}^{p} \mathbb{P}\left(\left|\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right)\left(Y_{i}-\mathbb{E}\left(Y_{i}\right)\right)\right|>r \omega_{j} n / 2\right) .
\end{aligned}
$$

For $j \in\{1, \ldots, p\}$, set $v_{j}=\sum_{i=1}^{n} \mathbb{E}\left(\phi_{j}^{2} \epsilon_{i}^{2}\right)$. Since $\sum_{i=1}^{n} \phi_{j}^{2}\left(z_{i}\right) \geqslant 4 v_{j}$, we have

$$
\mathbb{P}\left(\left|S_{j}\right|>n r \omega_{j} / 2\right) \leq \mathbb{P}\left(\left|S_{j}\right|>\sqrt{2 v_{j}(x+\log p)}+\frac{c_{2}}{3}(x+\log p)\right), \quad r \geq 1
$$

By applying Bernstein's inequality (see $[5,19]$ ) to the right hand side of the previous inequality we get

$$
\mathbb{P}\left(\left|S_{j}\right|>n r \omega_{j} / 2\right) \leq 2 \exp (-x-\log p)
$$

It follows that

$$
\begin{equation*}
\mathbb{P}\left(\mathrm{e}^{c}\right) \leq \sum_{j=1}^{p} \mathbb{P}\left(\left|S_{j}\right|>r \omega_{j} n / 2\right) \leq 2 \exp (-x) \tag{5.23}
\end{equation*}
$$

When $\omega_{j}=1$, for all $j \in\{1, \ldots, p\}$ and $r=A \sqrt{\frac{\log p}{n}}$, we apply Hoeffding's inequality (see [5, 19]). This leads to

$$
\begin{align*}
\mathbb{P}\left(\mathrm{e}^{c}\right) & =\mathbb{P}\left(\bigcup_{j=1}^{p}\left\{\left|\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right)\left(Y_{i}-\mathbb{E}\left(Y_{i}\right)\right)\right|>r n / 2\right\}\right) \\
& \leq \sum_{j=1}^{p} \mathbb{P}\left(\left|\sum_{i=1}^{n} \phi_{j}\left(z_{i}\right)\left(Y_{i}-\mathbb{E}\left(Y_{i}\right)\right)\right|>r n / 2\right) \\
& \leq 2 p \exp \left(-\frac{2(r n / 2)^{2}}{\sum_{i=1}^{n} 2 c_{2}}\right)=2 p \exp \left(-\frac{r^{2} n}{4 c_{2}}\right)=2 p^{1-\frac{A^{2}}{4 c_{2}}} \tag{5.24}
\end{align*}
$$

This ends the proof of Theorem 3.1.

### 5.6. Proof of Theorem 3.2

Fix an arbitrary $\beta \in \mathbb{R}^{p}$ such that $f_{\beta} \in \Gamma$, and set $\delta=W\left(\hat{\beta}_{L}-\beta\right)$, where $W=\operatorname{Diag}\left(w_{1}, \ldots, w_{p}\right)$. It follows from inequality (5.15) that

$$
\begin{equation*}
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{0}\right) \leq(1+\eta)\left\{R\left(f_{\beta}\right)-R\left(f_{0}\right)+\frac{c(\eta) r^{2}\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}|K(\beta)|}{\mu^{2} c_{0} \epsilon_{0}}\right\} \tag{5.25}
\end{equation*}
$$

where $c(\eta)=2(1+2 / \eta)^{2} /(2+2 / \eta)$. This ends the proof of inequality (3.3) of the Theorem 3.2. Inequality (3.4) follows from Lemma 5.1. To prove inequalities (3.5) and (3.6) we just replace $\omega_{j}$ by $A \sqrt{\frac{\log p}{n}}$.

This ends the proof of the Theorem 3.2 by using (5.23) and (5.24).

### 5.7. Proof of Corollary 3.3

Set $\delta=W\left(\hat{\beta}_{L}-\beta_{0}\right)$. The result (3.9) directly comes by taking $\beta=\beta_{0}$ and $\eta=2$ in (5.25). Note that, on the event $E$ defined in (5.22), we have

$$
\begin{equation*}
\left\|\delta_{K\left(\beta_{0}\right)^{c}}\right\|_{1} \leq 3\left\|\delta_{K\left(\beta_{0}\right)}\right\|_{1} \tag{5.26}
\end{equation*}
$$

Indeed, since $\hat{\beta}_{L}$ is the minimizer of $\hat{R}\left(f_{\beta}\right)+r \sum_{j=1}^{p} \omega_{j}\left|\beta_{j}\right|$, then

$$
R\left(f_{\hat{\beta}_{L}}\right)-R\left(f_{\beta_{0}}\right)+r \sum_{j=1}^{p} \omega_{j}\left|\hat{\beta}_{L, j}\right| \leq \frac{1}{n} \varepsilon^{T} X\left(\hat{\beta}_{L}-\beta_{0}\right)+r \sum_{j=1}^{p} \omega_{j}\left|\beta_{0 j}\right|
$$

which implies that

$$
r\left\|W \hat{\beta}_{L}\right\|_{1} \leq \sum_{j=1}^{p}\left|\frac{1}{n} \sum_{i=1}^{n} \phi_{j}\left(z_{i}\right) \varepsilon_{i}\right|\left|\hat{\beta}_{L, j}-\beta_{j}\right|+r\left\|W \beta_{0}\right\|_{1} .
$$

On the event $E$ we have

$$
\begin{aligned}
\left\|W\left(\hat{\beta}_{L}\right)_{K\left(\beta_{0}\right)}\right\|_{1}+\left\|W\left(\hat{\beta}_{L}\right)_{K^{c}\left(\beta_{0}\right)}\right\|_{1} \leq & \frac{1}{2}\left(\left\|W\left(\hat{\beta}_{L}-\beta_{0}\right)_{K\left(\beta_{0}\right)}\right\|_{1}+\left\|W\left(\hat{\beta}_{L}\right)_{K^{c}\left(\beta_{0}\right)}\right\|_{1}\right) \\
& +\left\|W\left(\beta_{0}\right)_{K\left(\beta_{0}\right)}\right\|_{1}
\end{aligned}
$$

Thus (5.26) follows. Line (3.10) follows from line (3.9) by applying Lemma 5.1. Line (3.11) follows from line (3.10) by using inequality (5.12) and $\|\delta\|_{1}^{2} \leq 16 s\left\|\delta_{K\left(\beta_{0}\right)}\right\|_{2}^{2}$. The last line follows from Lemma A. 3 in Appendix with $a_{j}=\left|\hat{\beta}_{L, j}-\beta_{0 j}\right|$ and

$$
b_{1}=\frac{12 s r\left(\max _{1 \leq j \leq p} \omega_{j}\right)^{2}}{\mu^{2}(s, 3) c_{0} \epsilon_{0}\left(\min _{1 \leq j \leq p} \omega_{j}\right)}
$$

## A. Appendix

The proof of Lemma 5.1 are based on property of self concordant function (see for instance [28]), i.e., the functions whose third derivatives are controlled by their second derivatives. A one-dimensional, convex function $g$ is called self concordant if

$$
\left|g^{\prime \prime \prime}(x)\right| \leq C g^{\prime \prime}(x)^{3 / 2}
$$

The function we use $(g(t)=\hat{R}(g+t h))$ is not really 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 necessarily linear as assumed in [2]. Precisely we extend Proposition 1 in [2] to the functions which are not necessarily linear (see Lem. A.1). This allows us to establish Lemma 5.1.

Lemma A.1. For all $h, f: \mathbb{R}^{d} \rightarrow \mathbb{R}$, we have

$$
\begin{gather*}
\frac{\langle h, h\rangle_{f}}{\|h\|_{\infty}^{2}}\left(\exp \left(-\|h\|_{\infty}\right)+\|h\|_{\infty}-1\right) \leq R(f+h)-R(f)+\left(q_{f}-q_{f_{0}}\right)(h)  \tag{A.1}\\
R(f+h)-R(f)+\left(q_{f}-q_{f_{0}}\right)(h) \leq \frac{\langle h, h\rangle_{f}}{\|h\|_{\infty}^{2}}\left(\exp \left(\|h\|_{\infty}\right)-\|h\|_{\infty}-1\right) \tag{A.2}
\end{gather*}
$$

and

$$
\begin{equation*}
\langle h, h\rangle_{f} \mathrm{e}^{-\|h\|_{\infty}} \leq\langle h, h\rangle_{f+h} \leq\langle h, h\rangle_{f} \mathrm{e}^{\|h\|_{\infty}} . \tag{A.3}
\end{equation*}
$$

Proof of Lemma A.1. We use the following lemma (see [2], Lem. 1) that we recall here:
Lemma A.2. Let $g$ be a convex three times differentiable function $g: \mathbb{R} \rightarrow \mathbb{R}$ such that for all $t \in \mathbb{R}\left|g^{\prime \prime \prime}(t)\right| \leq$ $S g^{\prime \prime}(t)$, for some $S \geq 0$. Then, for all $t \geq 0$ :

$$
\begin{equation*}
\frac{g^{\prime \prime}(0)}{S^{2}}(\exp (-S t)+S t-1) \leq g(t)-g(0)-g^{\prime}(0) t \leq \frac{g^{\prime \prime}(0)}{S^{2}}(\exp (S t)-S t-1) \tag{A.4}
\end{equation*}
$$

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

$$
g(t)=\hat{R}(f+t h)=\frac{1}{n} \sum_{i=1}^{n} l\left((f+t h)\left(z_{i}\right)\right)-Y_{i}(f+t h)\left(z_{i}\right), \quad f, h \in H,
$$

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

$$
g^{\prime \prime}(t)=\frac{1}{n} \sum_{i=1}^{n} h^{2}\left(z_{i}\right) l^{\prime \prime}\left((f+t h)\left(z_{i}\right)\right)=\langle h, h\rangle_{f+t h},
$$

and

$$
g^{\prime \prime \prime}(t)=\frac{1}{n} \sum_{i=1}^{n} h^{3}\left(z_{i}\right) l^{\prime \prime \prime}\left((f+t h)\left(z_{i}\right)\right) .
$$

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

$$
\begin{aligned}
\left|g^{\prime \prime \prime}(t)\right| & =\left|\frac{1}{n} \sum_{i=1}^{n} h^{3}\left(z_{i}\right) l^{\prime \prime \prime}\left((f+t h)\left(z_{i}\right)\right)\right| \\
& \leq \frac{1}{n} \sum_{i=1}^{n} h^{2}\left(z_{i}\right) l^{\prime \prime}\left((f+t h)\left(z_{i}\right)\right)\|h\|_{\infty}=\|h\|_{\infty} g^{\prime \prime}(t)
\end{aligned}
$$

We now apply Lemma A. 2 to $g(t)$ with $S=\|h\|_{\infty}$, taking $t=1$. Using equation (1.4) we get the first and second inequality of Lemma A.1. Now by considering $g(t)=\langle h, h\rangle_{f+t h}$, a short calculation leads to $\left|g^{\prime}(t)\right| \leq\|h\|_{\infty} g(t)$ which implies $g(0) \mathrm{e}^{-\|h\|_{\infty} t} \leq g(t) \leq g(0) \mathrm{e}^{\|h\|_{\infty} t}$. By applying the last inequality to $g(t)$, and taking $t=1$ we get the third inequality of Lemma A.1.

## A.1. Proof of Lemma 5.1

Set $h_{0}=f_{\beta}-f_{0}$ from Lemma A. 1 below,

$$
\frac{\left\langle h_{0}, h_{0}\right\rangle_{f_{0}}}{\left\|h_{0}\right\|_{\infty}^{2}}\left(\exp \left(-\left\|h_{0}\right\|_{\infty}\right)+\left\|h_{0}\right\|_{\infty}-1\right) \leq R\left(f_{\beta}\right)-R\left(f_{0}\right)
$$

Using assumptions $\left(\mathbf{B}_{3}\right),\left(\mathbf{B}_{1}\right)$ and the decreasing of $t \mapsto \frac{\exp (-t)+t-1}{t^{2}}$, we claim that there exists $c_{0}=$ $c_{0}\left(C_{0}, c_{1}\right)>0$ such that

$$
c_{0} \leq \frac{\left.\exp \left(-\left\|h_{0}\right\|_{\infty}\right)+\left\|h_{0}\right\|_{\infty}-1\right)}{\left\|h_{0}\right\|_{\infty}^{2}}
$$

According to assumption $\left(\mathbf{B}_{1}\right)$, there exists $0 \leq \epsilon_{0} \leq 1 / 2$ such that for $1 \leq i \leq n$

$$
\epsilon_{0} \leq \pi\left(f_{0}\left(z_{i}\right)\right)\left(1-\pi\left(f_{0}\left(z_{i}\right)\right)\right) \leq 1-\epsilon_{0} .
$$

The proof of the left hand side of Lemma 5.1 follows from the fact that $\epsilon_{0}\left\|h_{0}\right\|_{n}^{2} \leq\left\langle h_{0}, h_{0}\right\rangle_{f_{0}}$. From the second line of Lemma A. 1 we have

$$
R\left(f_{\beta}\right)-R\left(f_{0}\right) \leq \frac{\left\langle h_{0}, h_{0}\right\rangle_{f_{0}}}{\left\|h_{0}\right\|_{\infty}^{2}}\left(\exp \left(\left\|h_{0}\right\|_{\infty}\right)-\left\|h_{0}\right\|_{\infty}-1\right) .
$$

Using assumption $\left(\mathbf{B}_{\mathbf{3}}\right)$ and increasing of $t \mapsto \frac{\exp (t)-t-1}{t^{2}}$ thus there exists $c_{0}^{\prime}=c_{0}^{\prime}\left(C_{0}, c_{1}\right)>0$ such that

$$
\begin{aligned}
R\left(f_{\beta}\right)-R\left(f_{0}\right) & \leq c_{0}^{\prime}\left\langle h_{0}, h_{0}\right\rangle_{f_{0}} \\
& \leq c_{0}^{\prime} \frac{1}{4}\left\|h_{0}\right\|_{n}^{2} .
\end{aligned}
$$

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.2. Proof of Lemma A. 3

We start by writing

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

Since $\sum_{i=1}^{p} a_{j}^{2} \leq\left(\sum_{i=1}^{p} a_{j}\right)^{2} \leq b_{1}^{2}$, thus

$$
\begin{equation*}
\sum_{i=1}^{p} a_{j}^{q} \leq b_{1}^{2-q} b_{1}^{2 q-2}=b_{1}^{q} \tag{A.5}
\end{equation*}
$$

This ends the proof.
Lemma A. 4 (Bernstein's inequality). Let $X_{1}, \ldots, X_{n}$ be independent real valued random variables such that for all $i \leq n, X_{i} \leq b$ almost surely, then we have

$$
\mathbb{P}\left[\left|\sum_{i=1}^{n} X_{i}-\mathbb{E}\left(X_{i}\right)\right| \geqslant \sqrt{2 v x}+b x / 3\right] \leq 2 \exp (-x)
$$

where $v=\sum_{i=1}^{n} \mathbb{E}\left(X_{i}^{2}\right)$.
This lemma is obtain by gathering Proposition 2.9 and inequality (2.23) from [19].
Lemma A. 5 (Hoeffding's inequality). Let $X_{1}, \ldots, X_{n}$ be independent random variables such that $X_{i}$ takes its values in $\left[a_{i}, b_{i}\right]$ almost surely for all $i \leq n$. Then for any positive $x$, we have

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

This lemma is a consequence of Proposition 2.7 in [19].

Acknowledgements. We are grateful to two anonymous reviewers, the Associate Editor and Editor for their helpful comments which led to considerable improvements in the paper. 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.

## References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory (Tsahkadsor, 1971). Akadémiai Kiadó, Budapest (1973) 267-281.
[2] F. Bach, Self-concordant analysis for logistic regression. Electron. J. Statist. 4 (2010) 384-414.
[3] P.L. Bartlett, S. Mendelson and J. Neeman, 1-regularized linear regression: persistence and oracle inequalities. Probab. Theory Relat. Fields 154 (2012) 193-224.
[4] P.J. Bickel, Y. Ritov and A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37 (2009) 17051732.
[5] S. Boucheron, G. Lugosi and O. Bousquet, Concentration inequalities. Adv. Lect. Machine Learn. (2004) 208-240.
[6] F. Bunea, A.B. Tsybakov and M.H. Wegkamp, Aggregation and sparsity via $l_{1}$ penalized least squares. In Learning theory, vol. 4005 of Lect. Notes Comput. Sci. Springer, Berlin (2006) 379-391.
[7] F. Bunea, A.B. Tsybakov and M.H. Wegkamp, Aggregation for Gaussian regression. Ann. Statist. 35 (2007) $1674-1697$.
[8] F. Bunea, A. Tsybakov and M. Wegkamp, Sparsity oracle inequalities for the Lasso. Electron. J. Statist. 1 (2007) 169-194.
[9] C. Chesneau and M. Hebiri, Some theoretical results on the grouped variables lasso. Math. Methods Statist. 17 (2008) 317-326.
[10] J. Friedman, T. Hastie and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. J. Statist. Software 33 (2010) 1.
[11] M. Garcia-Magariños, A. Antoniadis, R. Cao and W. González-Manteiga, Lasso logistic regression, GSoft and the cyclic coordinate descent algorithm: application to gene expression data. Stat. Appl. Genet. Mol. Biol. 9 (2010) 30.
[12] T. Hastie, Non-parametric logistic regression. SLAC PUB-3160 (1983).
[13] J. Huang, S. Ma and CH Zhang, The iterated lasso for high-dimensional logistic regression. Technical Report 392 (2008).
[14] J. Huang, J.L. Horowitz and F. Wei, Variable selection in nonparametric additive models. Ann. Statist. 38 (2010) 2282.
[15] G.M James, P. Radchenko and J. Lv, Dasso: connections between the dantzig selector and lasso. J. Roy. Statist. Soc. Ser. B 71 (2009) 127-142.
[16] K. Knight and W. Fu, Asymptotics for lasso-type estimators. Ann. Statist. 28 (2000) 1356-1378.
[17] K. Lounici, M. Pontil, A.B. Tsybakov and S. Van De Geer, Taking advantage of sparsity in multi-task learning. In COLT'09 (2009).
[18] K. Lounici, M. Pontil, S. van de Geer and A.B. Tsybakov, Oracle inequalities and optimal inference under group sparsity. Ann. Statist. 39 (2011) 2164-2204.
[19] P. Massart, Concentration inequalities and model selection. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6-23, 2003. With a foreword by Jean Picard. Vol. 1896 of Lect. Notes Math. Springer, Berlin (2007).
[20] P. Massart and C. Meynet, The Lasso as an $\ell_{1}$-ball model selection procedure. Electron. J. Statist. 5 (2011) $669-687$.
[21] J. McAuley, J. Ming, D. Stewart and P. Hanna, Subband correlation and robust speech recognition. IEEE Trans. Speech Audio Process. 13 (2005) 956-964.
[22] L. Meier, S. van de Geer and P. Bühlmann, The group Lasso for logistic regression. J. Roy. Statist. Soc. Ser. B 70 (2008) 53-71.
[23] L. Meier, S. Van De Geer and P. Bühlmann, High-dimensional additive modeling. Ann. Statist. 37 (2009) 3779-3821.
[24] N. Meinshausen and P. Bühlmann, High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34 (2006) 1436-1462.
[25] N. Meinshausen and B. Yu, Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist. 37 (2009) 246-270.
[26] Y. Nardi and A. Rinaldo, On the asymptotic properties of the group lasso estimator for linear models. Electron. J. Statist. $\mathbf{2}$ (2008) 605-633.
[27] S.N. Negahban, P. Ravikumar, M.J. Wainwright and B. Yu, A unified framework for high-dimensional analysis of $M$-estimators with decomposable regularizers. Statist. Sci. 27 (2012) 538-557.
[28] Y. Nesterov and A. Nemirovskii, Interior-point polynomial algorithms in convex programming. Vol. 13 of SIAM Studies in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1994).
[29] M.R. Osborne, B. Presnell and B.A. Turlach, A new approach to variable selection in least squares problems. IMA J. Numer. Anal. 20 (2000) 389-403.
[30] M.Y. Park and T. Hastie, $L_{1}$-regularization path algorithm for generalized linear models. J. Roy. Statist. Soc. Ser. B 69 (2007) 659-677.
[31] B. Tarigan and S.A. van de Geer, Classifiers of support vector machine type with $l_{1}$ complexity regularization. Bernoulli 12 (2006) 1045-1076.
[32] R. Tibshirani, Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58 (1996) $267-288$.
[33] P. Ravikumar, J. Lafferty, H. Liu and L. Wasserman, Sparse additive models. J. Roy. Statist. Soc. Ser. B 71 (2009) 1009-1030.
[34] G. Schwarz, Estimating the dimension of a model. Ann. Statist. 6 (1978) 461-464.
[35] S.A. van de Geer, High-dimensional generalized linear models and the lasso. Ann. Statist. 36 (2008) 614-645.
[36] S.A. van de Geer and P. Bühlmann, On the conditions used to prove oracle results for the Lasso. Electron. J. Statist. 3 (2009) 1360-1392.
[37] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel and K. Lange, Genome-wide association analysis by lasso penalized logistic regression. Bioinform. 25 (2009) 714-721.
[38] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables. J. Roy. Statist. Soc. Ser. B 68 (2006) 49-67.
[39] C.-H. Zhang and J. Huang, The sparsity and bias of the LASSO selection in high-dimensional linear regression. Ann. Statist. 36 (2008) 1567-1594.
[40] P. Zhao and B. Yu, On model selection consistency of Lasso. J. Mach. Learn. Res. 7 (2006) 2541-2563.
[41] H. Zou, The adaptive lasso and its oracle properties. J. Am. Statist. Assoc. 101 (2006) 1418-1429.


[^0]:    Keywords and phrases. Logistic model, Lasso, Group Lasso, high-dimensional, oracle inequality.
    ${ }^{1}$ Laboratoire de Mathématique et modélisation d'Evry UMR CNRS 8071- USC INRA, Université d'Évry Val d'Essonne, Evry, France. marius.kwemou@gmail.com
    ${ }^{2}$ LERSTAD, Université Gaston Berger de Saint-Louis, Sénégal.

