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 [1]. 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 [2] or [3]). In this context the considered model is often what we call here "usual" logistic regression. It is given by

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

where one observes n couples $(z_1, Y_1),\ldots,(z_n, Y_n) \in \mathbb{R}^d \times \{0, 1\}$, and β_0 is the unknown parameter to estimate. Throughout the paper, we consider a fixed design setting (i.e $z_1, \ldots, z_n$ are consider deterministic).

*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

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

where the outputs $Y_i \in \{0, 1\}$, $i = 1, \ldots, n$ are independent and f_0 (not necessarily 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 $\mathbb{D} = \{\phi_1, \ldots, \phi_p\}$: $\hat{f}(.) := \sum_{j=1}^{p} \hat{\beta}_j \phi_j(.)$. 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 ℓ_0 penalization, *ie* penalyze by the number of non zero coefficients (see for instance AIC, BIC [4, 5]). 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 [6] proposes to use ℓ_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 ℓ_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 [7]), *predictor-corector* (see [8]) 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 ℓ_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 [9]).

Previously known results. Recently, a great deal of attention has been focused on ℓ_1-penalized based estimators. Most of this attention concerns regression models and ℓ_1-penalized least squares estimator of parameters in high dimensional linear and non linear additive regression. Among them one can cite [10, 11, 12, 13], who have studied the Lasso for linear model in nonparametric setting and proved sparsity oracle inequalities. Similar sparsity oracle inequalities are proved in [1], 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. [14] 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 [15]). The behavior of the Lasso and Group Lasso regarding their selection and estimation properties have been studied in: [16, 17, 18, 19, 20, 21] for Lasso in linear regression; [22, 23] for Group Lasso in linear regression; [24, 25, 26] 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). Zou [27] shows consistency in variable selection for adaptive Lasso in generalized linear models when the number of covariables p is fixed. Huang et al. [28] prove sign consistency and estimation consistency for high-dimensional logistic regression. Meir et al. [29] shown consistency for the Group Lasso in "usual" logistic model (1). To our knowledge there are only two non asymptotic results for the Lasso in logistic model: the first one is from Bach [30], 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 [31], who established non asymptotic oracle inequality for Lasso in high dimensional generalized linear models with Lipschitz loss functions. There is no non asymptotic result for the Group Lasso in 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 2.1 and Theorem 3.1) with no assumption on the Gram matrix, on the regressors nor on the margin. Secondly we provide "fast" oracle inequalities (see Theorem 2.2 and Theorem 3.2) under *restricted eigenvalues assumption* and some technical assumptions on the regressors. In each case, we give, as a consequence, the bounds for excess risk, $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ 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 ℓ_1-norm (for the Lasso) and group norm (for the Group Lasso). Simulation study is given to illustrate the numerical performance of Group Lasso and Lasso with such weights.

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 Section 2.3 and Section 3.3 we give as a consequence the bounds for excess risk, $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ and estimation errors for Lasso and Group Lasso in the "usual" logistic regression (1) . Section 4 is devoted to simulation study. The proofs are gathered in Section 6 and Appendix.

Definitions and notations

Consider the matrix $X = (\phi_j(z_i))_{1\leq i\leq n,\ 1\leq j\leq p}$ and $\{G_l, l = 1,\dots,g\}$ the partition of $\{1,\dots,p\}$. For any $\beta = (\beta_1,\dots,\beta_p)^T = (\beta^1,\dots,\beta^g)^T \in \mathbb{R}^p$, where $\beta^l = (\beta_j)_{j\in G_l}$ for $l = 1,\dots,g$. 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

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

We define the group norm of β as

$$\|\beta\|_{2,q} = \left(\sum_{l=1}^{g}\left(\sum_{j\in G_l}\beta_j^2\right)^{\frac{q}{2}}\right)^{\frac{1}{q}} = \left(\sum_{l=1}^{g}\|\beta^l\|_2^q\right)^{\frac{1}{q}},$$

for every $1\le q<\infty$. For $\beta\in\mathbb{R}^p$ $K(\beta)=\{j\in\{1,\dots,p\}:\beta_j\neq 0\}$ and $J(\beta)=\{l\in\{1,\dots,g\}:\beta^l\neq 0\}$, respectively the set of relevant coefficients (which characterizes the sparsity of the vector β) and the set of relevant groups. For all $\delta\in\mathbb{R}^p$ and a subset $I\subset\{1,\dots,p\}$, we denote by δ_I the vector in $\mathbb{R}^p$ that has the same coordinates as δ on I and zero coordinates on the complement I^c of I. Moreover $|I|$ denotes the cardinality of I. For all $h,f,g:\mathbb{R}^d\to R$, 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 $L_2(\frac{1}{n}\sum_{i=1}^{n}\delta_{z_i})$ norm (empirical norm). We consider empirical risk (logistic loss) for logistic model

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

We denote by R the expectation of $\hat{R}$ with respect to the distribution of $Y_1,\dots,Y_n$, *i.e*

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

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

$$R(f) = I\!E(\hat{R}(f)) = \hat{R}(f) + q_{f_0}(f). \tag{4}$$

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.

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 (2) by a linear combination of the functions of a dictionary

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

where $\phi_j : \mathbb{R}^d \to \mathbb{R}$ and p possibly $>> n$. The functions ϕ_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 β 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 [29, 32]), where predictors are included or excluded in groups. The logistic Group Lasso is the minimizer of the following optimization problem

$$f_{\hat{\beta}_{GL}} := \underset{f_\beta \in \Gamma_1}{\operatorname{argmin}} \left\{ \hat{R}(f_\beta) + r \sum_{l=1}^{g} \omega_l \|\beta^l\|_2 \right\}, \tag{5}$$

where

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

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 β 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 ℓ_1 and ℓ_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 $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ loss of Group Lasso estimator. Consider the following assumptions:

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

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

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

Assumptions $(\mathbf{B_1})$ and $(\mathbf{B_3})$ are technical assumptions useful to connect the excess risk and the $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ loss (see Lemma 6.1). An assumption similar to $(\mathbf{B_1})$ has been used in [11] to prove oracle inequality in gaussian regression model. The same kind of assumption as $(\mathbf{B_3})$ has been made in [33] to prove oracle inequality for support vector machine type with ℓ_1 complexity regularization.

Theorem 2.1. *Let $f_{\hat{\beta}_{GL}}$ be the Group Lasso solution defined in (5) with $r \geq 1$ and*

$$\omega_l = \frac{2|G_l|}{n}\sqrt{\frac{1}{2}\max_{j\in G_l}\sum_{i=1}^n \phi_j^2(z_i)\,(x+\log p) + \frac{2c_2|G_l|}{3n}\,(x+\log p)}\,, \tag{6}$$

where $x > 0$. Under Assumption $(\mathbf{B_2})$, with probability at least $1 - 2\exp(-x)$ we have

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq \inf_{\beta\in\mathbb{R}^p}\left\{R(f_\beta) - R(f_0) + 2r\|\beta\|_{2,1}\max_{1\leq l\leq g}\omega_l\right\}. \tag{7}$$

The first part of the right hand of Inequality (7) 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 (7) 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 β. Moreover this is a sharp oracle inequality in the sense that there is a constant 1 in front of term $\inf_{\beta\in\mathbb{R}^p}\{R(f_\beta) - R(f_0)\}$. This result is obtained without any assumption on the Gram matrix $(\Phi_n = X^TX/n)$. In order to obtain oracle inequality 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 $1 \leq s \leq g$ and a positive number a_0, the following condition holds $(\mathbf{RE_1})$

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

This is a natural extension to the Group Lasso of *restricted eigenvalues assumption* introduced in [1] (or Assumption $(\mathbf{RE_3})$ 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 $\{\Delta \neq 0 : \|\Delta_{K^c}\|_1 \leq a_0\|\Delta_K\|_1\}$ whereas for the Group Lasso the minimum is over $\{\Delta \neq 0 : \|\Delta_{K^c}\|_{2,1} \leq a_0\|\Delta_K\|_{2,1}\}$. This assumption has already been used in [15, 14] to prove oracle inequality for linear gaussian noise model under Group sparsity and for multi-task learning. 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 2.2. *Let $f_{\hat{\beta}_{GL}}$ be the Group Lasso solution defined in (5) with ω_l defined as in (6). Fix $\eta > 0$ and $1 \leq s \leq g$, assume that ($\mathbf{B_1}$), ($\mathbf{B_2}$), ($\mathbf{B_3}$) and ($\mathbf{RE_1}$) are satisfied, with $a_0 = 3 + 4/\eta$. Thus with probability at least $1 - 2\exp(-x)$ we have*

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta) \inf_{f_\beta \in \Gamma_1} \left\{ R(f_\beta) - R(f_0) + \frac{c(\eta)|J(\beta)|r^2 \left(\max_{1\leq l\leq g} \omega_l\right)^2}{c_0\epsilon_0\mu_1(s,a_0)^2} \right\}, \tag{8}$$

and

$$\|f_{\hat{\beta}_{GL}} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta) \inf_{f_\beta \in \Gamma_1} \left\{ \|f_\beta - f_0\|_n^2 + \frac{4c(\eta)|J(\beta)|r^2 \left(\max_{1\leq l\leq g} \omega_l\right)^2}{c_0'c_0\epsilon_0^2\mu_1(s,a)^2} \right\}. \tag{9}$$

Where $c(\eta)$ is a constant depending only on η; $c_0 = c_0(C_0, c_1)$ and $c_0' = c_0'(C_0, c_1)$ are constants depending on C_0 and c_1; $\epsilon_0 = \epsilon_0(c_1)$ 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". This rate is of same kind of the one obtain by [23] in Group Lasso for linear regression model with gaussian noise.

Remark 2.1. *Our results remain true if we assume that we are in the "neighborhood" of the target function. If we suppose that there exists ζ such that $\max_{1\leq i\leq n} |f_\beta(z_i) - f_0(z_i)| \leq \zeta$, then Lemma 6.1 is still true.*

Remark 2.2. *The choice of the weights ω_l comes from Bernstein's inequality. We could also use the following weights*

$$\omega_l' = \frac{2|G_l|}{n}\sqrt{2\max_{j\in G_l}\sum_{i=1}^n \mathbb{E}[\phi_j^2(z_i)\epsilon_i^2]\,(x+\log p)} + \frac{2|G_l|\max_{1\leq i\leq n}\max_{j\in G_l}|\phi_j(z_i)|}{3n}(x+\log p)\,.$$

Theorems 2.1 and 2.2 still hold true with such weights ω_l'. But these weights depend on the unknown function f_0 to be estimated through $\mathbb{E}(\epsilon_i^2) = \pi(f_0(z_i)(1-\pi(f_0(z_i))$. This is the reason for using weights ω_l slightly greater than ω_l'. We will show in simulation study (Section 4) how to use the weights ω_l' to improve the Group Lasso defined in [29] which used $\sqrt{|G_l|}$ as weight for the group l.

For the best of our knowledge, Inequalities (7), (8) and (9) 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.

2.3 Special case: f_0 linear

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 $X = (z_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$, the design matrix. Let $z_i = (z_{i1}, \dots, z_{ip})^T$ be the ith row of the matrix X and $z^{(j)} = (z_{1j}, \dots, z_{nj})^T$ is jth column. For $i = 1, \dots, n$

$$\mathbb{P}(Y_i = 1) = \frac{\exp(z_i^T \beta_0)}{1 + \exp(z_i^T \beta_0)}. \tag{10}$$

This corresponds to the "usual" logistic regression (1) *ie* logistic model that allows linear dependency between z_i and the distribution of Y_i. In this context, the Group Lasso estimator of β_0 is defined by

$$\hat{\beta}_{GL} := \underset{\beta:\ f_\beta \in \Gamma_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + \exp(z_i^T \beta)) - Y_i z_i^T \beta \right\} + r \sum_{l=1}^{g} \omega_l \|\beta^l\|_2. \tag{11}$$

Corollary 2.1. *Let assumption* $\mathbf{RE_1}(s,3)$ *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 (11) with*

$$\omega_l = \frac{2|G_l|}{n} \sqrt{\frac{1}{2} \max_{j \in G_l} \sum_{i=1}^{n} z_{ij}^2 \, (x + \log p) + \frac{2c_2|G_l|}{3n} (x + \log p)} \tag{12}$$

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

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \frac{9sr^2 \left(\max\limits_{1 \leq l \leq g} \omega_l \right)^2}{\mu^2(s,3) c_0 \epsilon_0} \tag{13}$$

$$\|f_{\hat{\beta}_{GL}} - f_{\beta_0}\|_n^2 \leq \frac{9sr^2 \left(\max\limits_{1 \leq l \leq g} \omega_l \right)^2}{\mu^2(s,3) c_0^2 \epsilon_0^2} \tag{14}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,1} \leq \frac{12rs \left(\max\limits_{1 \leq l \leq g} \omega_l \right)^2}{\mu^2(s,3) c_0 \epsilon_0 (\min\limits_{1 \leq l \leq g} \omega_l)} \tag{15}$$

$$\|\hat{\beta}_{GL} - \beta_0\|_{2,q}^q \leq \left(\frac{12rs \left(\max\limits_{1 \leq l \leq g} \omega_l \right)^2}{\mu^2(s,3) c_0 \epsilon_0 (\min\limits_{1 \leq l \leq g} \omega_l)} \right)^q \quad \text{for all} \ \ 1 < q \leq 2. \tag{16}$$

Remark 2.3. *In logistic regression model (27), if vector* β_0 *is sparse, ie* $|J(\beta_0)| \leq s$, *then Assumption* ($\mathbf{RE_1}$) *implies that* β_0 *is uniquely defined. Indeed, if there exists* β^* *such that for* $i = 1, \dots, n$, $\pi(z_i^T \beta_0) = \pi(z_i^T \beta^*)$, *it follows that* $X\beta_0 = X\beta^*$ *and* $|J(\beta^*)| \leq s$. *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* $\min\{\|X\beta\|_2 : |J(\beta)| \leq 2s, \beta \neq 0\} > 0$.

Remark 2.4. *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 [23]. Since then, some of these inqualities have been improved in Lounici [14]. Concerning in particular the gain on order rate, these improvement seem mainly based on the assumption that the noise is Gaussian. In our case (see proof of Theorem 2.1, formula (35)) the empirical process involve non gaussian variables and thus their method should not apply in our context.*

2.4 Non bounded functions

The results of Corollary 2.1 are obtained (as the consequence of Theorem 2.2) with the assumptions that f_{β_0} and all $f_\beta \in \Gamma_1$ are bounded. In some situations these assumptions could not be verified. In this section we will establish the same results without assuming ($\mathbf{B_1}$) or ($\mathbf{B_3}$) *ie* neither f_{β_0} nor f_β is bounded. We consider the Group Lasso estimator defined in (11) 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 ($\mathbf{RE_2}$)

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

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

This is an extension of the Assumption $\mathbf{RE_1}$ to the weighted Gram matrix $X^T DX/n$.

Theorem 2.3. *Consider the Group Lasso estimator $f_{\hat\beta_{GL}}$ defined by (11) with w_l defined as in (12) where $x>0$. Set $v = \max_{1\leq i\leq n}\max_{1\leq l\leq g}\|z_i^l\|_2$. Let Assumptions ($\mathbf{B_2}$) and ($\mathbf{RE_2}$) be satisfied with*

$$a_0 = \frac{3\max_{1\leq l\leq g}\omega_l}{\min_{1\leq l\leq g}\omega_l}.$$

If $r(1+a_0)^2\max_{1\leq l\leq g}\omega_l \leq \frac{\mu_2^2}{3v|J|}$, with probability at least $1-2\exp(-x)$ we have

$$R(f_{\hat\beta_{GL}}) - R(f_{\beta_0}) \leq \frac{9(1+a_0)^2 J(\beta_0)|r^2\left(\max_{1\leq l\leq g}\omega_l\right)^2}{\mu_2^2(s,3)} \tag{17}$$

$$\|\hat\beta_{GL}-\beta_0\|_{2,1} \leq \frac{6(1+a_0)^2|J(\beta_0)|r\left(\max_{1\leq l\leq g}\omega_l\right)}{\mu_2^2(s,3)} \tag{18}$$

$$\|\hat\beta_{GL}-\beta_0\|_{2,q}^q \leq \left(\frac{6(1+a_0)^2|J(\beta_0)|r\left(\max_{1\leq l\leq g}\omega_l\right)}{\mu_2^2(s,3)}\right)^q \quad \text{for all } 1<q\leq 2. \tag{19}$$

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

$$\epsilon_0 \leq \pi(f_{\beta_0}(z_i))[1 - \pi(f_{\beta_0}(z_i))] \qquad \textit{for all} \quad i = 1, \ldots, n$$

then,

$$\|X\hat{\beta}_{GL} - X\beta_0\|_n^2 \leq \frac{36(1+a_0)^2|J(\beta_0)|r^2\left(\max\limits_{1\leq l\leq g}\omega_l\right)^2}{\mu^2(s,3)\epsilon_0}. \tag{20}$$

Inequalities (18) and (19) are the extensions of the results in [30] 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 ℓ_1-penalized empirical risk

$$f_{\hat{\beta}_L} := \underset{f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \hat{R}(f_\beta) + r\sum_{j=1}^{p} \omega_j |\beta_j| \right\}, \tag{21}$$

where the minimum is taken over the set

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

and ω_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 ω_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* ℓ_1*-penalized minimum defined in (21). Let Assumption (*$\mathbf{B_2}$*) be satisfied.*

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

$$\omega_j = \frac{2}{n}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\phi_j^2(z_i)(x + \log p)} + \frac{2c_2(x + \log p)}{3n}. \tag{22}$$

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

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

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 $1 - 2p^{1-A^2/4c_2}$ we have

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

As previously, the variance terms are of order $\|\beta\|_1\sqrt{\log p/n}$ for any β. 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 eigenvalues condition.

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

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

This assumption has been introduced in [1], 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 [10, 11, 12] are more restrictive than *restricted eigenvalues assumption*. In those papers either Φ_n is positive definite, or mutual coherence condition is imposed. We refer to [34] 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.

Theorem 3.2. *Let $f_{\hat{\beta}_L}$ be the ℓ_1-penalized minimum defined in (21). Fix $\eta > 0$ and $1 \leq s \leq p$. Assume that ($\mathbf{B_1}$), ($\mathbf{B_2}$), ($\mathbf{B_3}$) and ($\mathbf{RE_3}$) are satisfied, with $a_0 = 3 + 4/\eta$.*

A-) Let $x > 0$ be fixed and $r \geq 1$. For $j = \{1, \ldots, p\}$, ω_j defined as in (22). Thus with probability at least $1 - 2\exp(-x)$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta)\inf_{f_\beta\in\Gamma}\left\{R(f_\beta) - R(f_0) + \frac{c(\eta)|K(\beta)|r^2\left(\max_{1\leq j\leq p}\omega_j\right)^2}{c_0\epsilon_0\mu^2(s, 3+4/\eta)}\right\}, \tag{23}$$

and

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta)\inf_{f_\beta \in \Gamma}\left\{\|f_\beta - f_0\|_n^2 + \frac{4c(\eta)|K(\beta)|r^2\left(\max_{1\leq j\leq p}\omega_j\right)^2}{c_0'c_0\epsilon_0^2\mu^2(s,3+4/\eta)}\right\}. \tag{24}$$

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 $1 - 2p^{1-A^2/4c_2}$ we have

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1+\eta)\inf_{f_\beta\in\Gamma}\left\{R(f_\beta) - R(f_0) + \frac{A^2c(\eta)}{c_0\epsilon_0\mu^2(s,3+4/\eta)}\frac{|K(\beta)|r^2\log p}{n}\right\}, \tag{25}$$

and

$$\|f_{\hat{\beta}_L} - f_0\|_n^2 \leq \frac{c_0'}{4c_0\epsilon_0}(1+\eta)\inf_{f_\beta\in\Gamma}\left\{\|f_\beta - f_0\|_n^2 + \frac{4c(\eta)A^2}{c_0'c_0\epsilon_0^2\mu^2(s,3+4/\eta)}\frac{|K(\beta)|r^2\log p}{n}\right\}. \tag{26}$$

In both cases $c(\eta)$ is a constant depending only on η; $c_0 = c_0(C_0, c_1)$ 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 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 [1] for linear regression model. For the best of our knowledge, (24) and (26) are the first non asymptotic oracle inequalities for the $L_2(\frac{1}{n}\sum_i^n \delta_{z_i})$ norm in logistic model. Some non asymptotic oracle inequalities for excess risk like (23) or (25) have been established in [31] 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 *RE* condition, which can be seen as empirical version of Assumption C in [31]. The confidence (probability that result holds true) of Inequality (23) 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 exhibit better performance, at least in the specific cases studied in the simulation part (see Section 4).

3.3 Special case: f_0 linear

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 $z_i = (z_{i1},\ldots,z_{ip})^T$. Denote $X = (z_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ the design matrix. Thus for $i = 1,\ldots,n$

$$\mathbb{P}(Y_i = 1) = \pi(z_i^T\beta_0) = \frac{\exp(z_i^T\beta_0)}{1+\exp(z_i^T\beta_0)}. \tag{27}$$

The Lasso estimator of β_0 is thus defined as

$$\hat{\beta}_L := \underset{\beta:\ f_\beta \in \Gamma}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + \exp(z_i^T \beta)) - Y_i z_i^T \beta \right\} + r \sum_{j=1}^{p} \omega_j |\beta_j| \right\}. \tag{28}$$

This infimum is achieved and might not be unique. When the design matrix X has full rank, the solution of optimization Problem (28) is usually unique.

Corollary 3.1. *Let assumption RE(s,3) be satisfied and $|K(\beta_0)| \leq s$, where $1 \leq s \leq p$. Consider the Lasso estimator $f_{\hat{\beta}_L}$ defined by (28) 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

$$R(f_{\hat{\beta}_L}) - R(f_{\beta_0}) \leq \frac{9sr^2 \left(\max\limits_{1 \leq j \leq p} \omega_j \right)^2}{\mu^2(s,3) c_0 \epsilon_0} \tag{29}$$

$$\|f_{\hat{\beta}_L} - f_{\beta_0}\|_n^2 \leq \frac{9s^2 r^2 \left(\max\limits_{1 \leq j \leq p} \omega_j \right)^2}{\mu^2(s,3) c_0^2 \epsilon_0^2} \tag{30}$$

$$\|\hat{\beta}_L - \beta_0\|_1 \leq \frac{12sr \left(\max\limits_{1 \leq j \leq p} \omega_j \right)^2}{\mu^2(s,3) c_0 \epsilon_0 \left(\min\limits_{1 \leq j \leq p} \omega_j \right)} \tag{31}$$

$$\|\hat{\beta}_L - \beta_0\|_q^q \leq \left(\frac{12sr \left(\max\limits_{1 \leq j \leq p} \omega_j \right)^2}{\mu^2(s,3) c_0 \epsilon_0 \left(\min\limits_{1 \leq j \leq p} \omega_j \right)} \right)^q \quad \text{for all } \ 1 < q \leq 2. \tag{32}$$

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 $1 - 2p^{1 - A^2/4c_2}$.

Line (29) and Line (31) of the corollary are similar to those of Theorem 5 in [30] except that we assume that $f_{\hat{\beta}_L}$ is bounded whereas he assume that r is bounded, and use the same weights $\omega_j =, \ldots, = \omega_p = 1$. 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 [1] for linear model with an s-sparse vector. Remark 2.3 remains true in this section.

4 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 Group Lasso using the weights we proposed to the Group Lasso proposed in [29]. Recall that Group Lasso for logistic regression proposed in [29] used $\sqrt{|G_l|}$ as weight for the group l, to ensure that the penalty term is of the order of the number of parameters $|G_l|$. We consider the Group Lasso defined in (11), with the weights defined in (12), which we denote by weight.GL. We also consider the Group Lasso estimator defined in (11) with weights

$$\omega_l' = \frac{2|G_l|}{n}\sqrt{2\max_{j\in G_l}\sum_{i=1}^{n}\mathbb{E}[z_{ij}^2\epsilon_i^2]\,(2+\log p)} + \frac{2|G_l|\max_{1\leq i\leq n}\max_{j\in G_l}|z_{ij}|}{3n}\,(2+\log p)\,,$$

which we denote by weight.theoretical.GL. Note that these are the exact weights, and as mentioned in Remark 2.2, all our results remain true with these weights. But the only drawback is that these weights depend on the unknown β_0 (the parameter to be estimated) through $\mathbb{E}[\epsilon_i^2] = \pi(z_i^T\beta_0)[1-\pi(z_i^T\beta_0)]$. Later, we will show how to estimate $\mathbb{E}[\epsilon_i^2]$ in order to estimate ω_l.

4.1 Data generation

▶ *Data generation for the Group Lasso.* We simulated our covariate matrix X with different numbers of covariates, observations and groups. The columns of X were independent and identically distributed (i.i.d.) gaussian, and the response y was constructed from logistic model (10) with $\beta_0^1 = (1,\ldots,1)$, $\beta_0^2 = (-1.5,\ldots,-1.5)$, $\beta_0^3 = (2,\ldots,2)$ and $\beta_0^l = (0,\ldots,0)$ for $l \notin \{1,2,3\}$. This corresponds to the model with $|J(\beta_0)|$=3. We consider different values of $|J^c(\beta_0)|$ to change the amount of sparsity. Denote by $nk = |G_l|$, $l \in \{1,2,3\}$ and $nkc = |G_l|$, $l \notin \{1,2,3\}$. For each data set we calculate the prediction error, estimation error, the rate of true selection and the rate of false relevant or irrelevant coefficients. True selection corresponds to the situation where the procedure selects exactly the true relevant coefficients. The rate of relevant or irrelevant coefficients is the rate of bad selection in an estimation (the procedure declares relevant yet it is irrelevant or declares irrelevant yet it is relevant).

▶ *Data generation for the Lasso.* We simulated 500 datasets consisting of n observations from logistic model (27), with $\beta_0 = (1.5,-1,2,0,0,\ldots,0) \in \mathbb{R}^p$ where $|K(\beta_0)| = 3$ and $p = 3+k \in \{200,200,1000\}$. The columns of X were i.i.d. gaussian. We first consider the Lasso with the weights ω_j' which we denote weight.theoretical. As we can not compute these weights in practice, we propose to estimate them as follows. Since the only unknown term in ω_j' is $I\!E(\epsilon_i^2)$, 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 β_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 β_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.

The results for the four methods are presented in the Figure 4, Figure 5 and Figure 6. Lasso represents the "classical" Lasso (without weight); weight.Logit is the Lasso with weights estimated using procedure (2); weight.Lasso is the Lasso with weights estimated by the procedure (1); weight.theoretical is the Lasso with theoretical weights. For all the methods, r will be estimated by cross-validation.

4.2 Comments

Referring to Figure 1, Figure 2 and Figure 3, we see that the performance of all methods increases until some optimal performance, and then decreases. This means that when we reach the optimal model, which corresponds to the model with r_{opt}-value, nothing is gain by adding other variables. Moreover it is important to note that the optimal value (r_{opt}) in prediction or estimation is different to the optimal value in selection. In prediction, estimation or selection, the Group Lasso using our weights outperforms the Group Lasso defined in [29]. According to Figure 1 for instance, weight.theoretical.GL reaches 99% of true selection rate, while weight.GL peaks at 97% and the Group Lasso comes in last with 66% of true selection.

According to Figure 4, Figure 5 and Figure 6 we can see that 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 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.

5 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. 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 $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ loss. We also provided as a consequence of oracle inequalities the bounds for excess risk, $L_2(\frac{1}{n}\sum_{i=1}^n \delta_{z_i})$ error and estimation error in the case where the true function f_0 is linear ("usual" logistic regression (1)). We shown in simulation

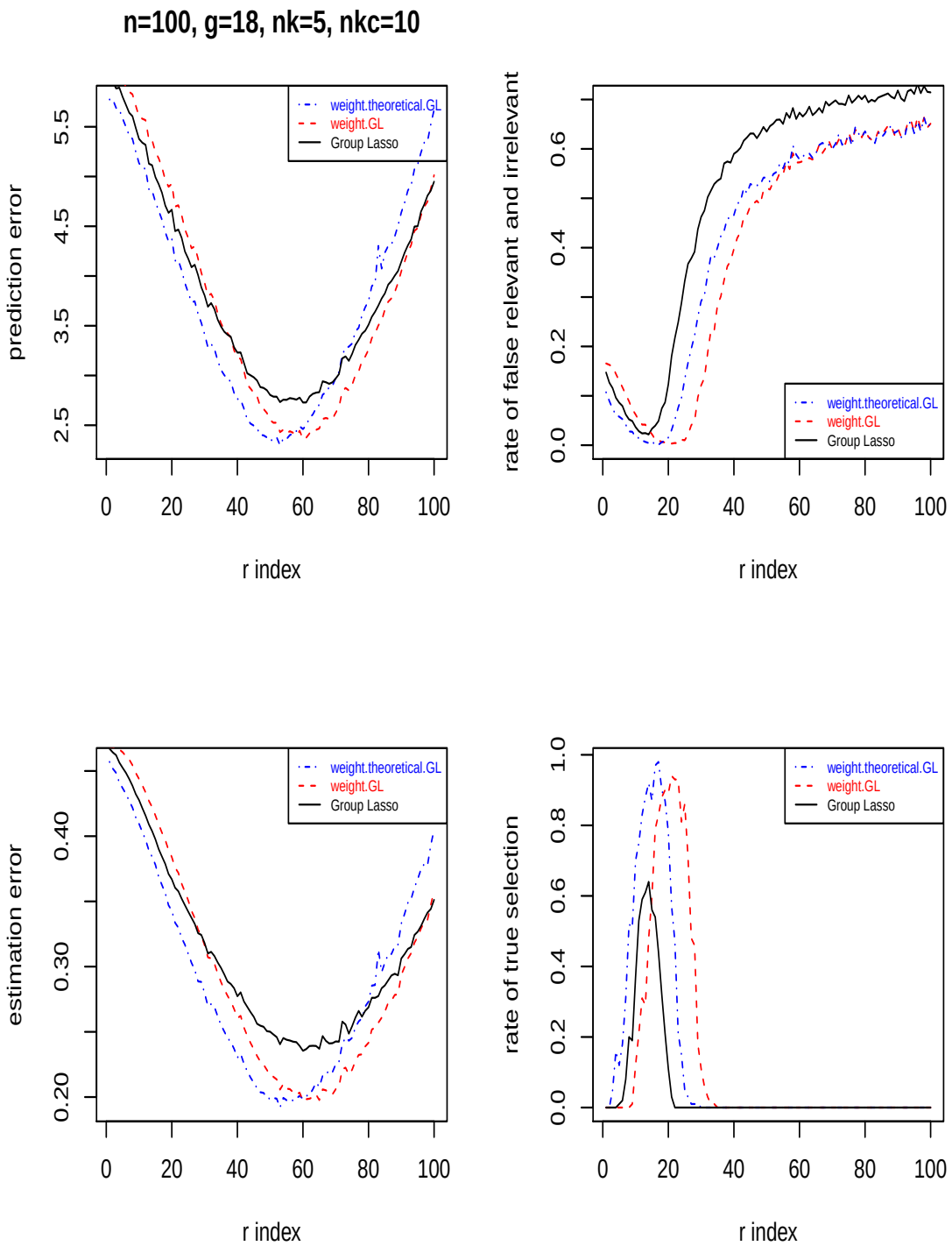

Figure 1: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients. All methods were fit from a path of 100 tuning parameters r from r_{max} to r_{min}. Each point corresponds to the average after 100 simulations from the setup described in Section 4.*

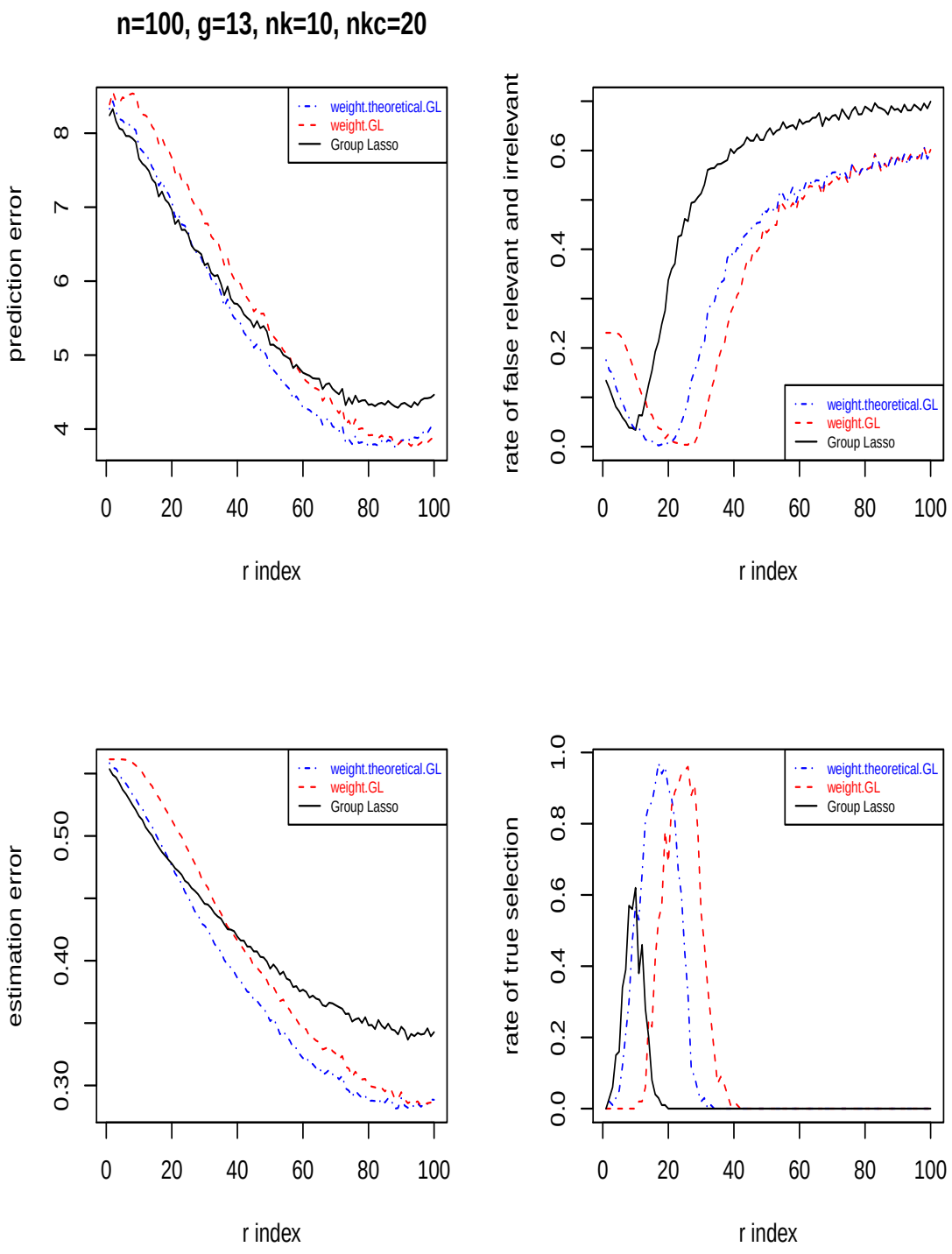

Figure 2: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients. All methods were fit from a path of 100 tuning parameters r from r_{max} to r_{min}. Each point corresponds to the average after 100 simulations from the setup described in Section 4.*

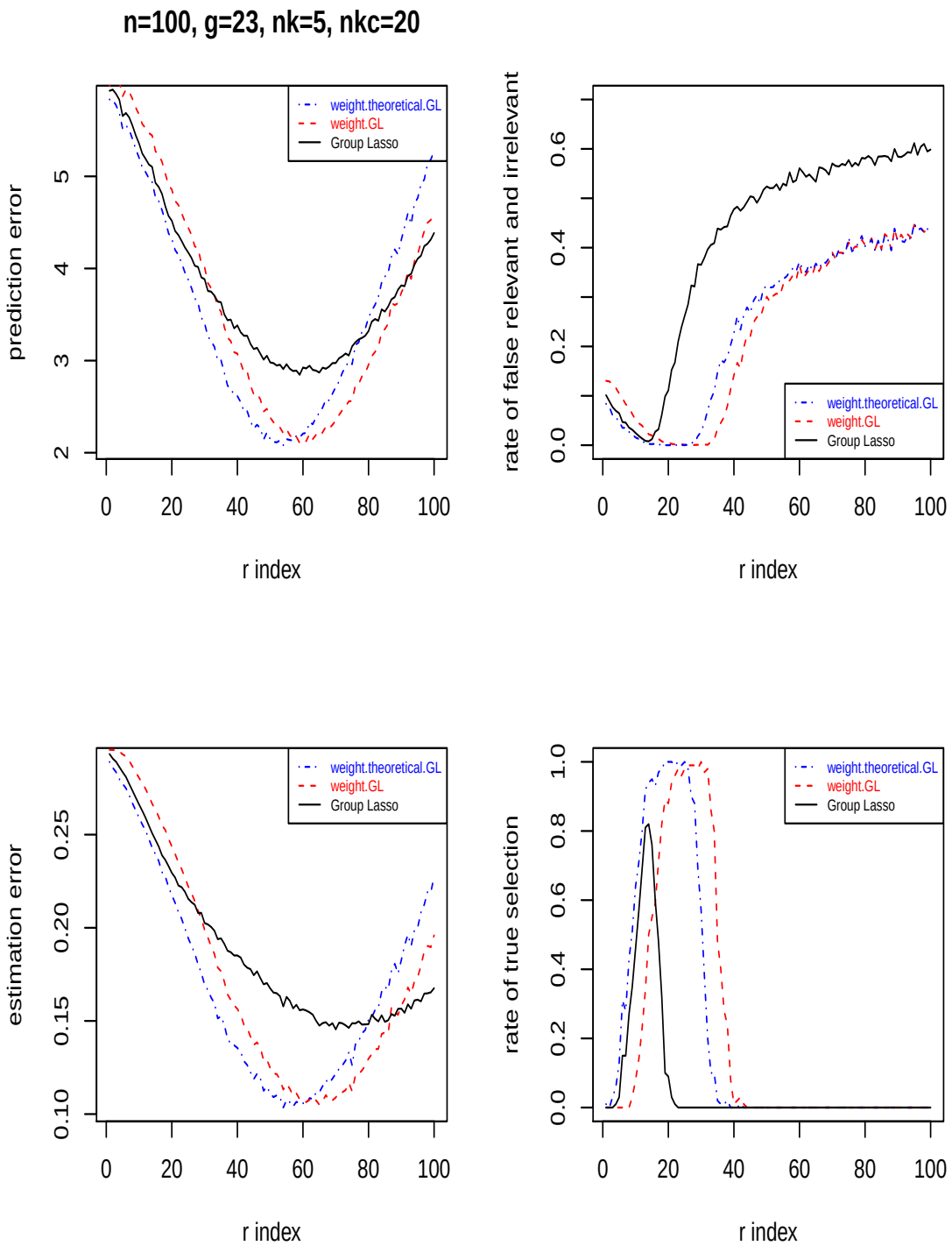

Figure 3: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients. All methods were fit from a path of 100 tuning parameters r from r_{max} to r_{min}. Each point corresponds to the average after 100 simulations from the setup described in Section 4.*

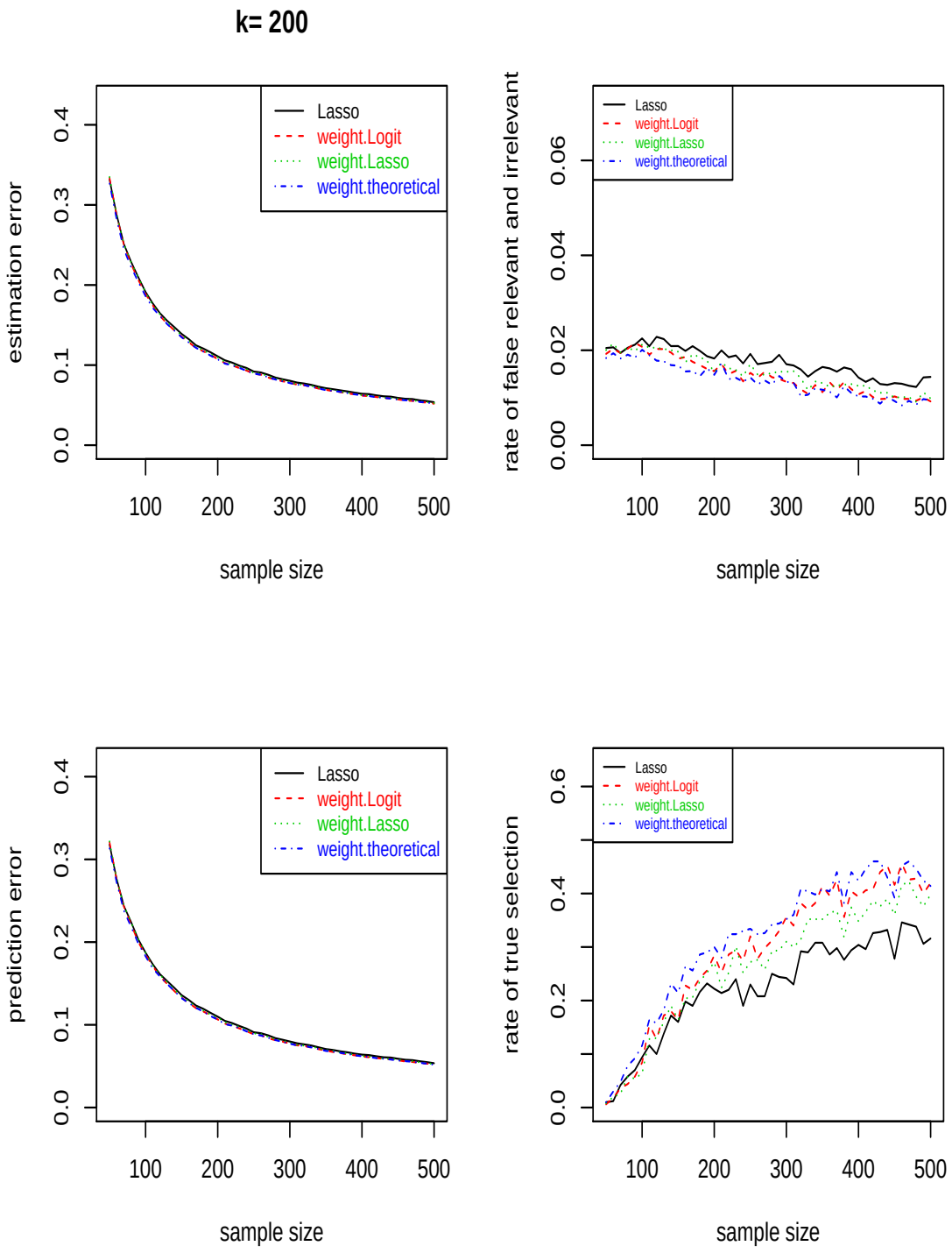

Figure 4: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients. k=200 from the setup described in Section 4*

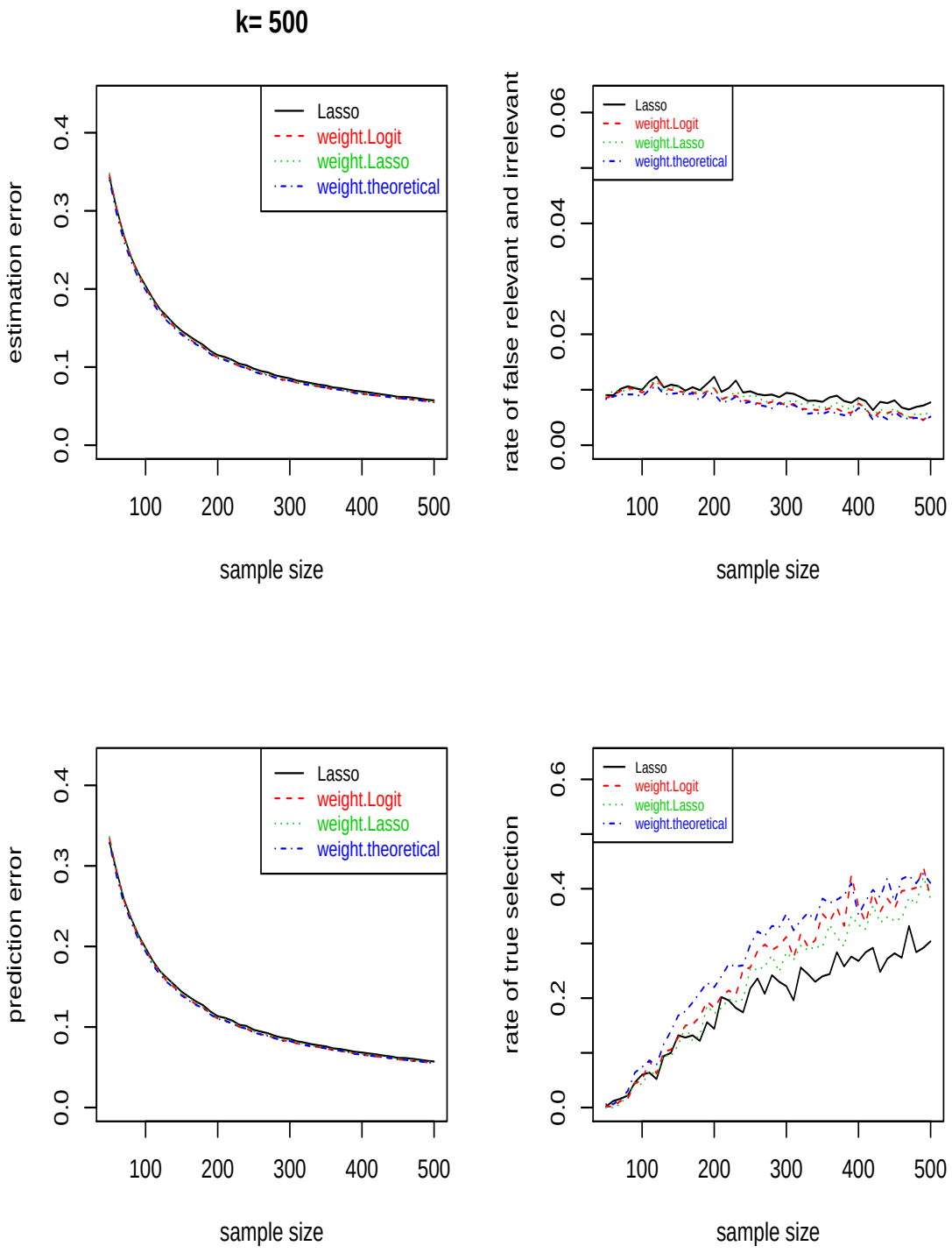

Figure 5: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficients. k=500 from the setup described in Section 4*

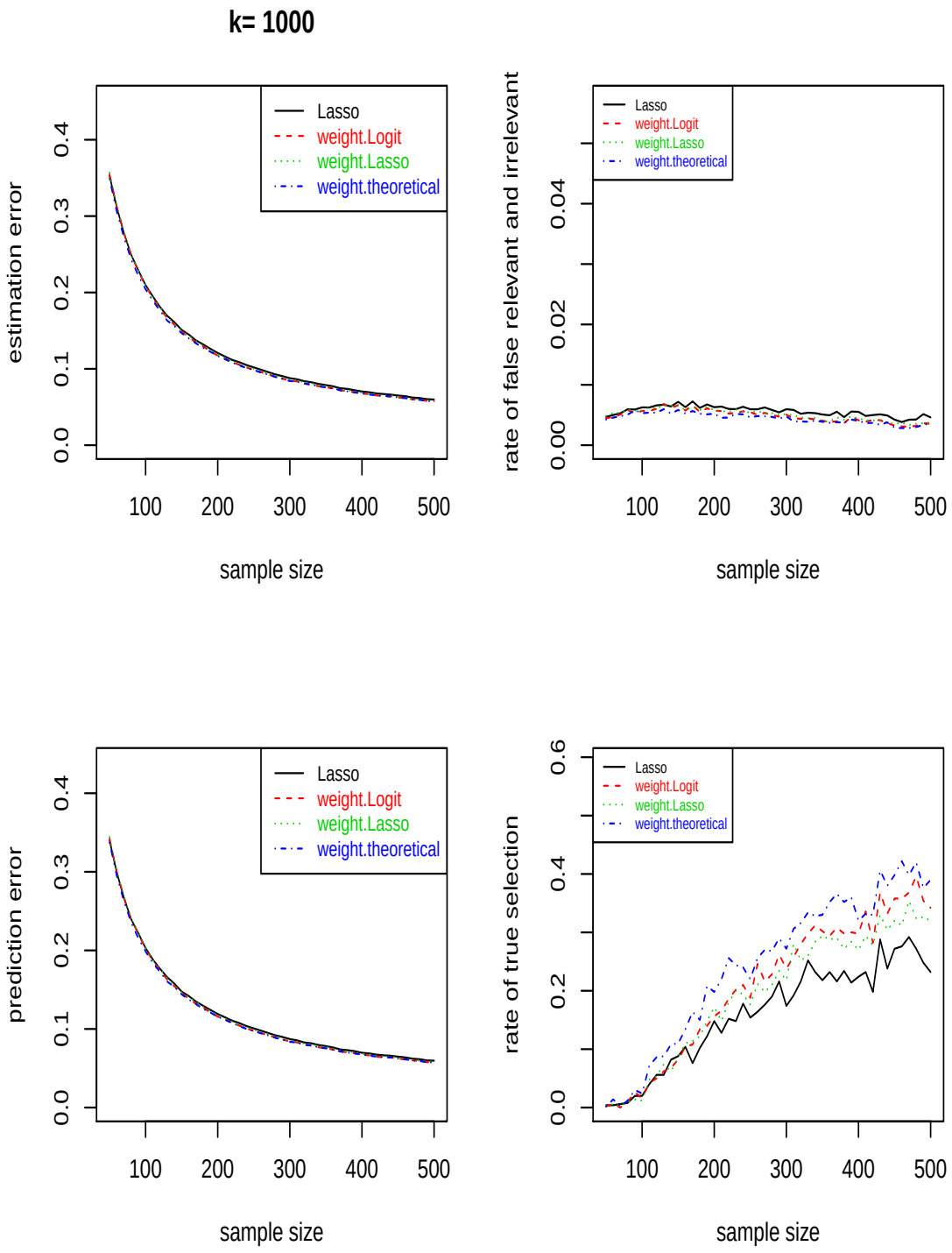

Figure 6: *Evolution of estimation error, prediction error, rate of true selection, and the rate of false relevant or irrelevant coefficient. k=1000 from the setup described in Section 4*

study that the weighted versions of Lasso and Group Lasso we proposed exhibit better properties than the canonical Lasso and Group 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.

6 Proofs of main results

6.1 Proof of Theorem 2.1

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

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

By applying Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) \quad &\leq \quad R(f_\beta) - R(f_0) + \sum_{l=1}^{g}\frac{1}{n}\sqrt{\sum_{j\in G_l}\left(\sum_{i=1}^{n}\phi_j(z_i)\epsilon_i\right)^2}\,\|(\hat{\beta}_{GL}-\beta)^l\|_2 \\ &\quad + r\sum_{l=1}^{g}\omega_l\|\beta^l\|_2 - r\sum_{l=1}^{g}\omega_l\|\hat{\beta}^l_{GL}\|_2. \end{aligned} \tag{33}$$

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 $l \in \{1,\ldots,g\}$ and the event

$$\mathcal{A} = \bigcap_{l=1}^{g}\left\{Z_l \leq r\omega_l/2\right\}. \tag{34}$$

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

On the event $\mathcal{A}$:

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

This implies that

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

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

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

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

$$\mathbb{P}(\mathcal{A}^c) = \mathbb{P}\left(\bigcup_{l=1}^{g}\{\sqrt{\sum_{j\in G_l}\left(\sum_{i=1}^{n}\phi_j(z_i)\epsilon_i\right)^2} > nr\omega_l/2\}\right) \tag{35}$$

$$\leq \sum_{l=1}^{g}\mathbb{P}\left(\sqrt{\sum_{j\in G_l}\left(\sum_{i=1}^{n}\phi_j(z_i)\epsilon_i\right)^2} > nr\omega_l/2\right). \tag{36}$$

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

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_{l=1}^{g}\mathbb{P}\left(\sqrt{\sum_{j\in G_l}(T_j^l)^2} > nr\omega_l/2\right)$$

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

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

$$\left\{\sum_{j\in G_l}|T_j^l| > nr\omega_l/2\right\} \subset \underset{j\in G_l}{\cup}\left\{|T_j^l| > \frac{nr\omega_l}{2|G_l|}\right\}, \tag{37}$$

it follows that

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

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

$$\mathbb{P}(|T_j^l| > \frac{nr\omega_l}{2|G_l|}) \leq \mathbb{P}\left(|T_j^l| > \sqrt{2v_j^l\,(x+\log p)} + \frac{c_2}{3}\,(x+\log p)\right), \quad r \geq 1.$$

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

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

It follows that

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_{l=1}^{g}\sum_{j\in G_l}\mathbb{P}\left(|T_j^l| > \frac{n\omega_l}{2|G_l|}\right) \leq \exp(-x). \tag{38}$$

This ends the proof of the Theorem 2.1. ■

6.2 Proof of Theorem 2.2

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

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

On the event $\mathcal{A}$ defined in (34), adding the term $\frac{r}{2}\sum_{l=1}^g \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2$ to both sides of Inequality (33) yields to

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

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

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) + \frac{r}{2}\sum_{l=1}^g \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \leq R(f_\beta) - R(f_0) + 2r\sum_{l\in J} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2. \tag{39}$$

we get from Equation (39) that

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq R(f_\beta) - R(f_0) + 2r\sum_{l\in J} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2 \tag{40}$$

Consider separately the two events:

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

and

$$\mathcal{A}_1^c = \{\eta(R(f_\beta) - R(f_0)) < 2r\sum_{l\in J} \omega_l \|(\hat{\beta}_{GL} - \beta)^l\|_2\}. \tag{41}$$

On the event $\mathcal{A} \cap \mathcal{A}_1$, we get from (40)

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)(R(f_\beta) - R(f_0)), \tag{42}$$

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 (40) that

$$\begin{aligned} R(f_{\hat{\beta}_{GL}}) - R(f_0) &\leq R(f_\beta) - R(f_0) + 2r\sqrt{|J(\beta)|}\sqrt{\sum_{l\in J} \omega_l^2 \|(\hat{\beta}_{GL} - \beta)^l\|_2^2} \\ &\leq R(f_\beta) - R(f_0) + 2r\sqrt{|J(\beta)|}\|\delta_J\|_2. \end{aligned} \tag{43}$$

On the other hand we get from Equation (39) that

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

and using (41) 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 ($\mathbf{RE_1}$) with $a_0 = 3+4/\eta$, and conclude that

$$\mu_1^2\|\delta_J\|_2^2 \leq \frac{\|X\delta\|_2^2}{n} = \frac{1}{n}(\hat{\beta}_{GL}-\beta)^T W X^T X W(\hat{\beta}_{GL}-\beta) \leq (\max_{1\leq l\leq g}\omega_l)^2\|f_{\hat{\beta}_{GL}}-f_\beta\|_n^2. \tag{44}$$

Gathering Equations (43) and (44) we get

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

We now use Lemma 6.1 which compares excess risk to empirical norm.

Lemma 6.1. *Under assumptions ($\mathbf{B_1}$) 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 ϵ_0 is a constant depending on c_1 and c_2.

(See the Appendix for the proof of Lemma 6.1).

Consequently

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

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

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

This implies that

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq \frac{b+1}{b-1}\left\{ R(f_\beta) - R(f_0) + \frac{2b^2 r^2 (\max_{1\leq l\leq g} \omega_l)^2 |J(\beta)|}{(b+1)\mu_1^2 c_0 \epsilon_0} \right\}. \tag{45}$$

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

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)\left\{ R(f_\beta) - R(f_0) + \frac{c(\eta) r^2 (\max_{1\leq l\leq g} \omega_l)^2 |J(\beta)|}{\mu_1^2 c_0 \epsilon_0} \right\}. \tag{46}$$

According to Inequalities (42) and (46) we conclude that on event $\mathcal{A}$,

$$R(f_{\hat{\beta}_{GL}}) - R(f_0) \leq (1+\eta)\left\{ R(f_\beta) - R(f_0) + \frac{c(\eta) r^2 (\max_{1\leq l\leq g} \omega_l)^2 |J(\beta)|}{\mu_1^2 c_0 \epsilon_0} \right\}, \tag{47}$$

where $c(\eta) = 2(1+2/\eta)^2/(2+2/\eta)$. Inequality (8) of the Theorem 2.2 follows. Inequality (9) follows from Lemma 6.1. This ends the proof of the Theorem 2.2 by considering (38). ■

6.3 Proof of Corollary 2.1

Set $\delta = W(\hat{\beta}_{GL} - \beta_0)$, Line (13) of Corollary 2.1 follows directly from Equation (47) with $\beta = \beta_0$ and $\eta = 1$. Note that on the event $\mathcal{A}$ defined in (34), we have

$$\|\delta_{J(\beta_0)^c}\|_{2,1} \leq 3\|\delta_{J(\beta_0)}\|_{2,1}. \tag{48}$$

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

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

which implies

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

On the event A we have

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

This yields to (48). Line (14) follows from Line (13) by applying Lemma 6.1. Line (15) follows from Line (14) by using Equation (44) and $\|\delta\|_{2,1}^2 \leq 16s\|\delta_{J(\beta_0)}\|_2^2$. Line (16) is the consequence of the Lemma A.3 with $a_l = \|(\hat{\beta}_{GL} - \beta_0)^l\|_2$ and

$$b_1 = \frac{12rs\left(\max_{1\leq l\leq g}\omega_l\right)^2}{\mu^2(s,3)c_0\epsilon_0(\min_{1\leq l\leq g}\omega_l)}. \quad ■$$

6.4 Proof of Theorem 2.3

On the event $\mathcal{A}$ defined in (34), using Inequality (33) with $\beta = \beta_0$ yields

$$R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \leq \sum_{l=1}^{g} \frac{3r\omega_l}{2} \|(\hat{\beta}_{GL} - \beta_0)^l\|_2. \tag{49}$$

By Lemma A.1 we have,

$$\frac{\langle h, h \rangle_{f_{\beta_0}}}{\|h\|_\infty^2} (\exp(-\|h\|_\infty) + \|h\|_\infty - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}) \tag{50}$$

where

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

One can easily verify that $\|h\|_\infty \leq v\|\delta'\|_{2,1}$ with $\delta' = \hat{\beta}_{GL} - \beta_0$. Equation (50) 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, Inequality (48) 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 ($\mathbf{RE_2}$) 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 $\|\delta'\|_{2,1}^2 \leq (1+a_0)^2 |J| \|\delta'_J\|_2^2$, with $J = J(\beta_0)$ to write

$$\frac{\mu_2^2}{(1+a_0)^2 |J| v^2} (\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1) \leq R(f_{\hat{\beta}_{GL}}) - R(f_{\beta_0}).$$

According to Equation (49) we have

$$\exp(-v\|\delta'\|_{2,1}) + v\|\delta'\|_{2,1} - 1 \leq \frac{3r(1+a_0)^2 \left(\max\limits_{1 \leq l \leq g} \omega_l\right) v^2 |J|}{2\mu_2^2} \|\delta'\|_{2,1}. \tag{51}$$

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

$$e^{\frac{-2a}{1-a}} + (1-a)\frac{2a}{1-a} - 1 \geqslant 0 \tag{52}$$

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

$$e^{-v\|\delta'\|_{2,1}} + v\|\delta'\|_{2,1} - 1 \geqslant \frac{v^2\|\delta'\|_{2,1}^2}{v\|\delta'\|_{2,1}+2}. \tag{53}$$

This implies using Equation (51) that

$$v\|\delta'\|_{2,1} \leq \frac{3r(1+a_0)^2\left(\max_{1\leq l\leq g}\omega_l\right)|J|v/\mu_2^2}{1-3r(1+a_0)^2\left(\max_{1\leq l\leq g}\omega_l\right)|J|v/2\mu_2^2}.$$

Now if $r(1+a_0)^2\max_{1\leq l\leq g}\omega_l \leq \frac{\mu_2^2}{3v|J|}$, we have $v\|\delta'\|_{2,1}\leq 2$ and consequently

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

Now, Inequality (51) implies

$$\|\delta'\|_{2,1} \leq \frac{6(1+a_0)^2|J|r\left(\max_{1\leq l\leq g}\omega_l\right)}{\mu_2^2}.$$

This proves the Line (18). Line (17) follows from (18) by using Inequality (49). Line (19) is the consequence of Lemma A.3 taking $a_l = \|(\hat{\beta}_{GL}-\beta_0)^l\|_2$ and $b_1 = 6(1+a_0)^2|J|r(\min_{1\leq l\leq g}\omega_l)/\mu_2^2(s,3)$. Line (20) follows from Line (17) and Inequality (50). ■

6.5 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 (33) that

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

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

$$E = \bigcap_{j=1}^{p}\left\{|S_j| \leq nr\omega_j/2\right\}. \tag{54}$$

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

On the event E:

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

We conclude that on the event E we have

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

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

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

For $j \in \{1, \ldots, p\}$, 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

$$\mathbb{P}(|S_j| > nr\omega_j/2) \leq \mathbb{P}\left(|S_j| > \sqrt{2v_j(x + \log p)} + \frac{c_2}{3}(x + \log p) \right), \quad r \geq 1.$$

By applying Bernstein's inequality (see [35, 36]) to the right hand side of the previous inequality we get

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

It follows that

$$\mathbb{P}(E^c) \leq \sum_{j=1}^{p} \mathbb{P}(|S_j| > r\omega_j n/2) \leq 2\exp(-x). \tag{55}$$

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 [35, 36]). This leads to

$$\begin{aligned}
\mathbb{P}(E^c) &= \mathbb{P}\left(\bigcup_{j=1}^{p} \{ |\sum_{i=1}^{n} \phi_j(z_i)(Y_i - I\!E(Y_i))| > rn/2 \} \right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(|\sum_{i=1}^{n} \phi_j(z_i)(Y_i - I\!E(Y_i))| > rn/2) \\
&\leq 2p\exp\left(-\frac{2(rn/2)^2}{\sum_{i=1}^{n} 2c_2} \right) = 2p\exp\left(-\frac{r^2 n}{4c_2} \right) = 2p^{1 - \frac{A^2}{4c_2}}.
\end{aligned} \tag{56}$$

This ends the proof of Theorem 3.1. ■

6.6 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 $W = \mathrm{Diag}(w_1, \ldots, w_p)$. It follows from Inequality (47) that

$$R(f_{\hat{\beta}_L}) - R(f_0) \leq (1 + \eta) \left\{ R(f_\beta) - R(f_0) + \frac{c(\eta) r^2 \left(\max_{1 \leq j \leq p} \omega_j \right)^2 |K(\beta)|}{\mu^2 c_0 \epsilon_0} \right\}, \tag{57}$$

where $c(\eta) = 2(1+2/\eta)^2/(2+2/\eta)$. This ends the proof of Inequality (23) of the Theorem 3.2. Inequality (24) follows from Lemma 6.1. To prove Inequalities (25) and (26) we just replace ω_j by $A\sqrt{\frac{\log p}{n}}$.
This ends the proof of the Theorem 3.2 by using (55) and (56). ■

6.7 Proof of Corollary 3.1

Set $\delta = W(\hat{\beta}_L - \beta_0)$. The result (29) directly comes by taking $\beta = \beta_0$ and $\eta = 2$ in (57). Note that, on the event E defined in (54), we have

$$\|\delta_{K(\beta_0)^c}\|_1 \leq 3\|\delta_{K(\beta_0)}\|_1. \tag{58}$$

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

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

which implies that

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

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 (58) follows. Line (30) follows from Line (29) by applying Lemma 6.1. Line (31) follows from Line(30) by using Inequality (44) and $\|\delta\|_1^2 \leq 16s\|\delta_{K(\beta_0)}\|_2^2$. The last line follows from Lemma A.3 in Appendix with $a_j = |\hat{\beta}_{L,j} - \beta_{0j}|$ and

$$b_1 = \frac{12sr\left(\max\limits_{1\leq j\leq p}\omega_j\right)^2}{\mu^2(s,3)c_0\epsilon_0\left(\min\limits_{1\leq j\leq p}\omega_j\right)}. \quad ■$$

A Appendix

The proof of Lemma 6.1 are based on property of self concordant function (see for instance [37]), *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 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 [30].

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 [30]. Precisely we extend Proposition 1 in [30] to the functions which are not necessarily linear (see Lemma A.1). This allows us to establish Lemma 6.1.

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

$$\frac{\langle h,h\rangle_f}{\|h\|_\infty^2}(\exp(-\|h\|_\infty)+\|h\|_\infty-1) \le R(f+h)-R(f)+(q_f-q_{f_0})(h), \tag{59}$$

$$R(f+h)-R(f)+(q_f-q_{f_0})(h) \le \frac{\langle h,h\rangle_f}{\|h\|_\infty^2}(\exp(\|h\|_\infty)-\|h\|_\infty-1), \tag{60}$$

and

$$\langle h,h\rangle_f e^{-\|h\|_\infty} \le \langle h,h\rangle_{f+h} \le \langle h,h\rangle_f e^{\|h\|_\infty}. \tag{61}$$

A.1 Proof of Lemma A.1

We use the following lemma (see [30] 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 $t \in \mathbb{R}$ $|g^{'''}(t)| \le S g^{''}(t)$, for some $S \ge 0$. Then , for all $t \ge 0$:*

$$\frac{g^{''}(0)}{S^2}(\exp(-St)+St-1) \le g(t)-g(0)-g'(0)t \le \frac{g^{''}(0)}{S^2}(\exp(St)-St-1). \tag{62}$$

We refer to Appendix A of [30] 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 $l^{'''}(u) \le l^{''}(u)$ we have,

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

We now apply Lemma A.2 to $g(t)$ with $S = \|h\|_\infty$, taking $t = 1$. Using Equation (4) 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 6.1

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 6.1 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 assumption ($\mathbf{B_3}$) 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 6.1.

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

We start by writing

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

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

$$\sum_{i=1}^{p} a_j^q \leq b_1^{2-q} b_1^{2q-2} = b_1^q. \tag{63}$$

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*

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

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

This lemma is obtain by gathering Proposition 2.9 and inequality (2.23) from [36].

Lemma A.5 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ 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*

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

This lemma is a consequence of Proposition 2.7 in [36].

References

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

[2] M. Garcia-Magariños, A. Antoniadis, R. Cao, 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) Art. 30, 30.

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

[4] 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, pp. 267–281.

[5] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (2) (1978) 461–464.

[6] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B 58 (1) (1996) 267–288.

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

[8] M. Y. Park, T. Hastie, L_1-regularization path algorithm for generalized linear models, Journal of the Royal Statistical Society Series B 69 (4) (2007) 659–677.

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

[10] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, Aggregation and sparsity via l_1 penalized least squares, in: Learning theory, Vol. 4005 of Lecture Notes in Comput. Sci., Springer, Berlin, 2006, pp. 379–391.

[11] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, Aggregation for Gaussian regression, Annals of Statistics 35 (4) (2007) 1674–1697.

[12] F. Bunea, A. Tsybakov, M. Wegkamp, Sparsity oracle inequalities for the Lasso, Electronic Journal of Statistics 1 (2007) 169–194.

[13] P. Massart, C. Meynet, The Lasso as an ℓ_1-ball model selection procedure, Electronic Journal of Statistics 5 (2011) 669–687.

[14] K. Lounici, M. Pontil, S. van de Geer, A. B. Tsybakov, Oracle inequalities and optimal inference under group sparsity, Annals of Statistics 39 (4) (2011) 2164–2204.

[15] K. Lounici, M. Pontil, A. Tsybakov, S. Van De Geer, Taking advantage of sparsity in multi-task learning, In COLT'09.

[16] K. Knight, W. Fu, Asymptotics for lasso-type estimators, Annals of Statistics 28 (5) (2000) 1356–1378.

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

[18] P. Zhao, B. Yu, On model selection consistency of Lasso, J. Mach. Learn. Res. 7 (2006) 2541–2563.

[19] M. R. Osborne, B. Presnell, B. A. Turlach, A new approach to variable selection in least squares problems, IMA J. Numer. Anal. 20 (3) (2000) 389–403.

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

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

[22] C. Chesneau, M. Hebiri, Some theoretical results on the grouped variables lasso, Mathematical Methods of Statistics 17 (4) (2008) 317–326.

[23] Y. Nardi, A. Rinaldo, On the asymptotic properties of the group lasso estimator for linear models, Electronic Journal of Statistics 2 (2008) 605–633.

[24] P. Ravikumar, J. Lafferty, H. Liu, L. Wasserman, Sparse additive models, Journal of the Royal Statistical Society Series B 71 (5) (2009) 1009–1030.

[25] L. Meier, S. Van De Geer, P. Bühlmann, High-dimensional additive modeling, Annals of Statistics 37 (6B) (2009) 3779–3821.

[26] J. Huang, J. Horowitz, F. Wei, Variable selection in nonparametric additive models, Annals of statistics 38 (4) (2010) 2282.

[27] H. Zou, The adaptive lasso and its oracle properties, J. Amer. Statist. Assoc. 101 (476) (2006) 1418–1429.

[28] J. Huang, S. Ma, C. Zhang, The iterated lasso for high–dimensional logistic regression, Technical Report 392.

[29] L. Meier, S. van de Geer, P. Bühlmann, The group Lasso for logistic regression, Journal of the Royal Statistical Society Series B 70 (1) (2008) 53–71.

[30] F. Bach, Self-concordant analysis for logistic regression, Electronic Journal of Statistics 4 (2010) 384–414.

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

[32] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society Series B 68 (1) (2006) 49–67.

[33] B. Tarigan, S. A. van de Geer, Classifiers of support vector machine type with l_1 complexity regularization, Bernoulli 12 (6) (2006) 1045–1076.

[34] S. A. van de Geer, P. Bühlmann, On the conditions used to prove oracle results for the Lasso, Electronic Journal of Statistics 3 (2009) 1360–1392.

[35] S. Boucheron, G. Lugosi, O. Bousquet, Concentration inequalities, Advanced Lectures on Machine Learning (2004) 208–240.

[36] P. Massart, Concentration inequalities and model selection, Vol. 1896 of Lecture Notes in Mathematics, Springer, Berlin, 2007, lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[37] Y. Nesterov, 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.