Article

Prediction and Variable Selection in High-Dimensional Misspecified Binary Classification

Konrad Furmańczyk 1,*,† and Wojciech Rejchel 2,†

1 Institute of Information Technology, Warsaw University of Life Sciences (SGGW), Nowoursynowska 159, 02-776 Warszawa, Poland
2 Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Chopina 12/18, 87-100 Toruń, Poland; wrejchel@gmail.com
* Correspondence: konrad_furmanczyk@sggw.pl
† These authors contributed equally to this work.

Received: 20 April 2020; Accepted: 11 May 2020; Published: 13 May 2020

Abstract: In this paper, we consider prediction and variable selection in the misspecified binary classification models under the high-dimensional scenario. We focus on two approaches to classification, which are computationally efficient, but lead to model misspecification. The first one is to apply penalized logistic regression to the classification data, which possibly do not follow the logistic model. The second method is even more radical: we just treat class labels of objects as they were numbers and apply penalized linear regression. In this paper, we investigate thoroughly these two approaches and provide conditions, which guarantee that they are successful in prediction and variable selection. Our results hold even if the number of predictors is much larger than the sample size. The paper is completed by the experimental results.

Keywords: misclassification risk; model misspecification; penalized estimation; supervised classification; variable selection consistency

1. Introduction

Large-scale data sets, where the number of predictors significantly exceeds the number of observations, become common in many practical problems from, among others, biology or genetics. Currently, the analysis of such data sets is a fundamental challenge in statistics and machine learning. High-dimensional prediction and variable selection are arguably the most popular and intensively studied topics in this field. There are many methods trying to solve these problems such as those based on penalized estimation \([1,2]\). The main representative of them is Lasso \([3]\), that relates to \(l_1\)-norm penalization. Its properties in model selection, estimation and prediction are deeply investigated, among others, in \([2,4–10]\). The results obtained in the above papers can be applied only if some specific assumptions are satisfied. For instance, these conditions concern the relation between the response variable and predictors. However, it is quite common that a complex data set does not satisfy these model assumptions or they are difficult to verify, which leads to the fact that the considered model is specified incorrectly. The model misspecification problem is the core of the current paper. We investigate this topic in the context of high-dimensional binary classification (binary regression).

In the classification problem we are to predict or to guess the class label of the object on the basis of its observed predictors. The object is described by the random vector \((X, Y)\), where \(X \in \mathbb{R}^p\) is a vector of predictors and \(Y \in \{-1, 1\}\) is the class label of the object. A classifier is defined as a measurable function \(f : \mathbb{R}^p \rightarrow \mathbb{R}\), which determines the label of an object in the following way:

\[
\text{if } f(x) \geq 0, \text{ then we predict that } y = 1.
\]
Otherwise, we guess that $y = -1$.

The most natural approach is to look for a classifier $f$, which minimizes the misclassification risk (probability of incorrect classification)

$$ R(f) = P(Y = 1, f(X) < 0) + P(Y = -1, f(X) \geq 0). \tag{1} $$

Let $\eta(x) = P(Y = 1|X = x)$. It is clear that $f_B(x) = \text{sign}(2\eta(x) - 1)$ minimizes the risk (1) in the family of all classifiers. It is called the Bayes classifier and we denote its risk as $R_B = R(f_B)$. Obviously, in practice we do not know the function $\eta$, so we cannot find the Bayes classifier. However, if we possess a training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ containing independent copies of $(X, Y)$, then we can consider a sample analog of (1), namely the empirical misclassification risk

$$ \frac{1}{n} \sum_{i=1}^{n} \left[ \mathbb{I}(Y_i = 1, f(X_i) < 0) + \mathbb{I}(Y_i = -1, f(X_i) \geq 0) \right], \tag{2} $$

where $\mathbb{I}$ is the indicator function. Then a minimizer of (2) could be used as our estimator.

The main difficulty in this approach lies in discontinuity of the function (2). It entails that finding its minimizer is computationally difficult and not effective. To overcome this problem, one usually replaces the discontinuous loss function by its convex analog $\phi : \mathbb{R} \to [0, \infty]$, for instance the logistic loss, the hinge loss or the exponential loss. Then we obtain the convex empirical risk

$$ Q(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_if(X_i)). \tag{3} $$

In the high-dimensional case one usually obtains an estimator by minimizing the penalized version of (3). Those tricks have been successfully used in the classification theory and have allowed to invent boosting algorithms [11], support vector machines [12] or Lasso estimators [3]. In this paper we are mainly interested in Lasso estimators, because they are able to solve both variable selection and prediction problems simultaneously, while the first two algorithms are developed mainly for prediction.

Thus, we consider linear classifiers

$$ f_B(x) = b_0 + \sum_{j=1}^{p} b_j x_j, \tag{4} $$

where $b = (b_0, b_1, \ldots, b_p) \in \mathbb{R}^{p+1}$. For a fixed loss function $\phi$ we define the Lasso estimator as

$$ \hat{b} = \arg\min_{b \in \mathbb{R}^{p+1}} \hat{Q}(f_b) + \lambda \sum_{j=1}^{p} |b_j|, \tag{5} $$

where $\lambda$ is a positive tuning parameter, which provides a balance between minimizing the empirical risk and the penalty. The form of the penalty is crucial, because its singularity at the origin implies that some coordinates of the minimizer $\hat{b}$ are exactly equal to zero, if $\lambda$ is sufficiently large. Thus, calculating (5) we simultaneously select significant predictors in the model and we estimate their coefficients, so we are also able to predict the class of new objects. The function $\hat{Q}(f_b)$ and the penalty are convex, so (5) is a convex minimization problem, which is an important fact from both practical and theoretical points of view. Notice that the intercept $b_0$ is not penalized in (5).

The random vector (5) is an estimator of

$$ b_\ast = \arg\min_{b \in \mathbb{R}^{p+1}} Q(f_b), \tag{6} $$

where $Q(f_b) = \mathbb{E}\phi(Yf_b(X))$. In this paper we are mainly interested in minimizers (6) corresponding to quadratic and logistic loss functions. The latter has a nice information-theoretic interpretation.
Namely, it can be viewed as the Kullback–Leibler projection of unknown $\eta$ on logistic models [13]. The Kullback–Leibler divergence [14] plays an important role in the information theory and statistics, for instance it is involved in information criteria in model selection [15] or in detecting influential observations [16].

In general, the classifier corresponding to (6) need not coincide with the Bayes classifier. Obviously, we want to have a “good” estimator, which means that its misclassification risk should be as close to the risk of the Bayes classifier as possible. In other words, its excess risk

$$\mathcal{E}(\hat{b}, f_b) = E_D R(\hat{b}) - R_B$$

should be small, where $E_D$ is the expectation with respect to the data $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and we write simply $R(\hat{b})$ instead of $R(f_b)$. Our goal is to study the excess risk (7) for the estimator (5) with different loss functions $\phi$. We do it by looking for the upper bounds of (7).

In the excess risk (7) we compare two misclassification risks defined in (1). In the literature one can also find a different approach, which replaces the misclassification risks $R(\cdot)$ in (7) by the convex risks $Q(\cdot)$. In that case the excess risk depends on the loss function $\phi$. To deal with this fact one uses the results from [17,18], which state the relation between the excess risk (7) and its analog based on the convex risk $Q(\cdot)$. In this paper we do not follow this way and work, right from the beginning, with the excess risk independent of $\phi$. Only the estimator (5) depends on the loss $\phi$.

In this paper we are also interested in variable selection. We investigate this problem in the following semiparametric model

$$\eta(x) = g(\beta_0 + \sum_{j=1}^{p} \beta_j x_j),$$

where $\eta(x) = P(Y = 1|X = x)$, $\beta \in \mathbb{R}^{p+1}$ is the true parameter and $g$ is unknown function. Thus, we suppose that predictors influence class probability through the function $g$ of the linear combination $\beta_0 + \sum_{j=1}^{p} \beta_j x_j$. The goal of variable selection is the identification of the set of significant predictors

$$T = \{1 \leq j \leq p : \beta_j \neq 0\}.$$  

Obviously, in the model (8) we cannot estimate an intercept $\beta_0$ and we can identify the vector $(\beta_1, \ldots, \beta_p)$ only up to a multiplicative constant, because any shift or scale change in $\beta_0 + \sum_{j=1}^{p} \beta_j X_j$ can be absorbed by $g$. However, we show in Section 5 that in many situations the Lasso estimator (5) can properly identify the set (9).

The literature on the classification problem is comprehensive. We just mention a few references: [12,19–21]. The predictive quality of classifiers is often investigated by obtaining upper bounds for their excess risks. It is an important problem and was studied thoroughly, among others in [17,18,22–24]. The variable selection and predictive properties of estimators in the high-dimensional scenario were studied, for instance, in [2,10,13,25,26]. In the current paper we investigate the behaviour of classifiers in possibly misspecified high-dimensional classification, which appears frequently in practice. For instance, while working with binary regression one often assumes incorrectly that the data follow the logistic regression model. Then the problem is solved using the Lasso penalized maximum likelihood method. Another approach to binary regression, which is widely used due to its computational simplicity, is just treating labels $Y_i$ as they were numbers and applying standard Lasso. For instance, such method is used in [1], [Subsections 4.2 and 4.3]) or [2], Subsection 2.4.1). These two approaches to classification sometimes give unexpectedly good results in variable selection and prediction, but the reason of this phenomenon has not been deeply studied in the literature. Among the above mentioned papers only [2,13,25] take up this issue. However, [25] focuses mainly on the predictive properties of Lasso classifiers with the hinge loss. Bühlmann and van de Geer [2] and Kubkowski and Mielniczuk [13] study general Lipschitz loss functions. The latter paper considers
only the variable selection problem. In [2] one also investigates prediction, but they do not study classification with the quadratic loss.

In this paper we are interested in both variable selection and predictive properties of classifiers with convex (but not necessarily Lipschitz) loss functions. The prominent example is classification with the quadratic loss function, which has not been investigated so far in the context of the high-dimensional misspecified model. In this case the estimator (5) can be calculated efficiently using the existing algorithms, for instance [27] or [28], even if the number of predictors is much larger than the sample size. It makes this estimator very attractive, while working with large data sets. In [28] one provides also the efficient algorithm for Lasso estimators with the logistic loss in the high-dimensional scenario. Therefore, misspecified classification with the logistic loss plays an important role in this paper as well. Our goal is to study thoroughly such estimators and provide conditions, which guarantee that they are successful in prediction and variable selection.

The paper is organized as follows: in the next section we provide basic notations and assumptions, which are used in this paper. In Section 3 we study predictive properties of Lasso estimators with different loss functions. We will see that these properties depend strongly on the estimation quality of estimators, which is studied in Section 4. In Section 5 we consider variable selection. In Section 6 we show numerical experiments, which describe the quality of estimators in practice. The proofs and auxiliary results are relegated to Appendix A.

2. Assumptions and Notation

In this paper we work in the high-dimensional scenario \( p \gg n \). As usual we assume that the number of predictors \( p \) can vary with the sample size \( n \), which could be denoted as \( p(n) = p_n \). However, to make notation simpler we omit the lower index and write \( p \) instead of \( p_n \). The same refers to the other objects appearing in this paper.

In the further sections we will need the following notation:

- \( X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top \);
- \( X = (X_1, X_2, \ldots, X_n)^\top \) is the \((n \times p)\)-matrix of predictors;
- \( A \subset \{1, \ldots, p\} \). Then \( A^c = \{1, \ldots, p\} \setminus A \) is a complement of \( A \);
- \( X_A \) is a submatrix of \( X \), with columns whose indices belong to \( A \);
- \( b_A \) is a restriction of a vector \( b \in \mathbb{R}^p \) to the indices from \( A \);
- \( |A| \) is the number of elements in \( A \);
- \( \tilde{A} = A \cup \{0\} \), so the set \( \tilde{A} \) contains indices from \( A \) and the intercept;
- The \( l_q \)-norm of a vector is defined as \( |b|_q = \left( \sum_{j=1}^p |b_j|^q \right)^{1/q} \) for \( q \in [1, \infty] \);
- For \( x \in \mathbb{R}^p \) we denote \( \tilde{x} = (1, x)^\top \);
- \( \tilde{X} \) is the matrix \( X \) with the column of ones binded from the left side;
- \( b_{\text{quad}} \), \( b_{\text{quad}}^\wedge \) are minimizers in (5), (6), respectively, with the quadratic loss function;
- \( b_{\text{log}} \), \( b_{\text{log}}^\wedge \) are minimizers in (5), (6), respectively, with the logistic loss function;
- The Kullback–Leibler (KL) distance [14] between two binary distributions with success probabilities \( \pi_1 \) and \( \pi_2 \) is defined as
  \[
  KL(\pi_1, \pi_2) = \pi_1 \log \left( \frac{\pi_1}{\pi_2} \right) + (1 - \pi_1) \log \left( \frac{1 - \pi_1}{1 - \pi_2} \right),
  \]

  Obviously, we have \( KL(\pi_1, \pi_2) \geq 0 \) and \( KL(\pi_1, \pi_2) = 0 \) if only if \( \pi_1 = \pi_2 \). Moreover, the KL distance need not be symmetric;
- the set of nonzero coefficients of \( b_{\text{quad}}^\wedge \) is denoted as
  \[
  T = \{1 \leq j \leq p : (b_{\text{quad}}^\wedge)_j \neq 0 \}. \]
Notice that the intercept is not contained in (11) even if it is nonzero. We also specify assumptions, which are used in this paper.

**Assumption 1.** We assume that \((X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)\) are i.i.d. random vectors. Moreover, predictors are univariate subgaussian, i.e., for each \(a \in \mathbb{R}\) and \(j \in \{1, \ldots, p\}\) we have \(E \exp(aX_j) \leq \exp(\sigma_j^2a^2/2)\) for positive numbers \(\sigma_j\). We also denote \(\sigma = \max_{1 \leq j \leq p} \sigma_j\). Finally, we suppose that the matrix \(H = E[XX^T]\) is positive definite and \(H_{jj} = 1\) for \(j = 1, \ldots, p\).

In Sections 4 and 5 we need stronger version of Assumption 1.

**Assumption 2.** We suppose that the subvector of predictors \(X_T\) is subgaussian with the coefficient \(\sigma_0 > 0\), i.e., for each \(u \in \mathbb{R}^{|T|}\) we have \(E \exp(u^T X_T) \leq \exp(\sigma_0^2 u^T H_T u/2)\), where \(H_T = (E[X_j X_k])_{j,k \in T}\). The remaining conditions are as in Assumption 1. We also denote \(\sigma = \max(\sigma_0, \sigma_j, j \notin T)\).

Subgaussianity of predictors is a standard assumption while working with random predictors in high-dimensional models, cf. [13]. In particular, Assumption 1 implies that \(E[X] = 0\) and \(\sigma \geq 1\) [29].

### 3. Predictive Properties of Classifiers

In this part of the paper we study prediction properties of classifiers with convex loss functions. To do it we look for upper bounds of the excess risk (7) of estimators.

As usual the excess risk in (7) can be decomposed as

\[
E_D R(\hat{b}) - R(b_*) + R(b_*) - R_B.
\]

The second term in (12) is the approximation risk and compares the predictive ability of the “best” linear classifier (6) to the Bayes classifier. The first term in (12) is called the estimation risk and describes how the estimation process influences the predictive property of classifiers.

In the next theorem we bound from above the estimation risk of classifiers. To make the result more transparent we use notations \(P_D\) and \(P_X\) in (13), which indicate explicitly which probability we consider, i.e., \(P_D\) is probability with respect to the data \(D\) and \(P_X\) is with respect to the new object \(X\). In further results we omit these lower indexes and believe that it does not lead to confusion.

**Theorem 1.** For \(c > 0\) we consider an event \(\Omega = \{||\hat{b} - b_*||_1 \leq c\}\). We have

\[
E_D R(\hat{b}) - R(b_*) \leq 2P_D(\Omega^C) + P_X(||b_*^T \hat{X}|| \leq c||X||_\infty).
\]  

In Theorem 1 we obtain the upper bound for the estimation risk. This risk becomes small, if we establish that probability of the event \(\Omega^C\) is small and the sequence \(c\), which is involved in \(\Omega\) and in the second term on the right-hand side of (13), decreases sufficiently fast to zero. Therefore, Theorem 1 shows that to have a small estimation risk it is enough to prove that for each \(\varepsilon \in (0, 1)\) there exists \(c\) such that

\[
P(||\hat{b} - b_*||_1 \leq c) \geq 1 - \varepsilon.
\]

Moreover, numbers \(\varepsilon\) and \(c\) should be sufficiently small. This property will be studied thoroughly in the next section. Notice that the first term on the right-hand side of (13) relates to the fact, how well (5) estimates (6). Moreover, the second expression on the right-hand side of (13) can be bounded from above, if predictors are sufficiently regular, for instance subgaussian.

So far, we have been interested in the estimation risk of estimators. In the next result we establish the upper bound for the approximation risk as well. This bound combined with (13) enables us to bound from above the excess risk of estimators. We prove this fact for the quadratic loss \(\phi(t) = (1 - t)^2\) and the logistic loss \(\phi(t) = \log(1 + e^{-t})\), which play prominent roles in this paper.
Theorem 2. Suppose that Assumption 1 is fulfilled. Moreover, a random variable $b^\top \tilde{X}$ has a density $h$, which is continuous on the interval $U = [-2c \sqrt{\log p}, 2c \sqrt{\log p}]$ and $\tilde{h} = \sup_{u \in U} h(u)$.

(a) We have
\[
\mathcal{E}(\hat{b}^{\text{quad}}, f_B) \leq 2P(\Omega^c) + 4\sigma \tilde{h}\log c \sqrt{\log p} + 2/p + \sqrt{E \left[ 2\eta(X) - 1 - (b^\text{quad})^\top \tilde{X} \right]^2},
\]
where $\tilde{h}$ refers to the density $h$ of $(b^\text{quad})^\top \tilde{X}$.

(b) Let $\eta_\log(u) = 1/(1 + \exp(-u))$. Then we obtain
\[
\mathcal{E}(\hat{b}^{\text{log}}, f_B) \leq 2P(\Omega^c) + 4\sigma \tilde{h}\log c \sqrt{\log p} + 2/p + \sqrt{2E \left[ KL \left( \eta(X), \eta_\log((b^\text{log})^\top \tilde{X}) \right) \right]},
\]
where $KL(\cdot, \cdot)$ is the Kullback–Leibler distance defined in (10) and $\tilde{h}$ refers to the density $h$ of $(b^\text{log})^\top \tilde{X}$. Additionally, assuming that there exists $\delta \in (0, 1)$ such that $\delta \leq \eta(X) \leq 1 - \delta$ and $\delta \leq \eta_\log((b^\text{log})^\top \tilde{X}) \leq 1 - \delta$, we have
\[
E \left[ KL \left( \eta(X), \eta_\log((b^\text{log})^\top \tilde{X}) \right) \right] \leq (2\delta(1-\delta))^{-1}E \left[ \eta(X) - \eta_\log((b^\text{log})^\top \tilde{X}) \right]^2.
\]

In Theorem 2 we establish upper bounds on the excess risks for Lasso estimators (5). They describe predictive properties of these classifiers. In this paper we consider linear classifiers, so the misclassification risk of an estimator is close to the Bayes risk, if the “truth” can be approximated linearly in a satisfactory way. For the classifier with the logistic loss this fact is described by (18) and (19), which measure the distance between true success probability and the one in logistic regression. In particular, when the true model is logistic, then (18) and (19) vanish. The expression (16) relates to the approximation error in the case of the quadratic loss. It measures how well the conditional expectation $E[Y|X]$ can be described by the “best” (with respect to the loss $\phi$) linear function $(b^\text{quad})^\top \tilde{X}$.

The right-hand sides of (15) and (17) relate to estimation risk. They have been already discussed after Theorem 1. Using subgaussianity of predictors we have made them more explicit. The main ingredient of bounds in Theorem 2, namely $P(\Omega^c)$, is studied in the next section.

Results in Theorem 2 refer to Lasso estimators with quadratic and logistic loss functions. Similar results are given in [2, Theorem 6.4]. They refer to the case that the convex excess risk is considered, i.e., the misclassification risks $R(\cdot)$ are replaced by the convex risks $Q(\cdot)$ in (7). Moreover, these results do not consider Lasso estimators with the quadratic loss applied to classification, which is an approach playing a key role in the current paper. Furthermore, in [2, Theorem 6.4] the estimation error $\hat{b} - b_*$ is measured in the $l_1$-norm, which is enough for prediction. However, for variable selection the $l_\infty$-norm gives better results. Such results will be established in Sections 4 and 5. Finally, results of [2] need more restrictive assumptions than ours. For instance, predictors should be bounded and a function $f_B$ should be sufficiently close to $f_B$ in the supremum norm.

Analogous bounds to those in Theorem 2 can be obtained for other loss functions, if we combine Theorem 1 with results of [17]. Finally, we should stress that the estimator $\hat{b}$ need not rely on the Lasso method. All we require is that the bound (14) can be established for this estimator.

4. On the Event $\Omega$

In this section we show that probability of the event $\Omega$ can be close to one. Such results for classification models with Lipschitz loss functions were established in [2,13]. Therefore, we focus on
the quadratic loss function, which is obviously non-Lipschitz. This loss function is important from the practical point of view, but was not considered in these papers. Moreover, in our results the estimation error in \( \Omega \) can be measured in the \( l_q \)-norms, \( q \geq 1 \), not only in the \( l_1 \)-norm as in [2,13]. Bounds in the \( l_\infty \)-norm lead to better results in variable selection, which are given in Section 5.

We start with introducing the cone invertibility factor (CIF), which plays a significant role in investigating properties of estimators based on the Lasso penalty [9]. In the case \( n > p \) one usually uses the minimal eigenvalue of the matrix \( \tilde{X}^\top \tilde{X} / n \) to express the strength of correlations between predictors. Obviously, in the high-dimensional scenario this value is equal to zero and the minimal eigenvalue needs to be replaced by some other measure of predictors interdependency, which would describe the potential of consistent estimation of model parameters.

For \( \xi > 1 \) we define a cone

\[
C(\xi) = \{ b \in \mathbb{R}^{p+1} : |b_T|_1 \leq \xi |b_{\tilde{T}}|_1 \},
\]

where we recall that \( \tilde{T} = T \cup \{0\} \). In the case when \( p \gg n \) three different characteristics measuring the potential for consistent estimation of the model parameters have been introduced:

- The restricted eigenvalue [8]:

\[
RE(\xi) = \inf_{0 \neq b \in C(\xi)} \frac{b^\top \tilde{X}^\top \tilde{X} b / n}{|b|^2},
\]

- The compatibility factor [7]:

\[
K(\xi) = \inf_{0 \neq b \in C(\xi)} \frac{|T| b^\top \tilde{X}^\top \tilde{X} b / n}{|b_T|^2},
\]

- The cone invertibility factor (CIF, [9]): for \( q \geq 1 \)

\[
\bar{F}_q(\xi) = \inf_{0 \neq b \in C(\xi)} \frac{|T|^{1/q} |\tilde{X}^\top \tilde{X} b / n|_\infty}{|b_q|}.
\]

In this article we will use CIF, because this factor allows for a sharp formulation of convergency results for all \( l_q \) norms with \( q \geq 1 \), see ([9], Section 3.2). The population (non-random) version of CIF is given by

\[
F_q(\xi) = \inf_{0 \neq b \in C(\xi)} \frac{|T|^{1/q} |\tilde{X}^\top \tilde{X} b / n|_\infty}{|b_q|},
\]

where \( \tilde{H} = E [\tilde{X} \tilde{X}^\top] \). The key property of the random and the population versions of CIF, \( \bar{F}_q(\xi) \) and \( F_q(\xi) \), is that, in contrast to the smallest eigenvalues of matrices \( \tilde{X}^\top \tilde{X} / n \) and \( \tilde{H} \), they can be close to each other in the high-dimensional setting, see ([30], Lemma 4.1) or ([31], Corollary 10.1). This fact is used in the proof of Theorem 3 (given below).

Next, we state the main results of this section.

**Theorem 3.** Let \( a \in (0,1) \), \( q \geq 1 \) and \( \xi > 1 \) be arbitrary. Suppose that Assumption 2 is satisfied and

\[
n \geq K_1 |T|^2 \sigma^4 (1 + \xi)^2 \log(p/a) \frac{F_q^2(\xi)}{\bar{F}_q^2(\xi)},
\]

and

\[
\lambda \geq K_2 \frac{\xi + 1}{\xi - 1} \sigma^2 \sqrt{\log(p/a) / n},
\]
where $K_1, K_2$ are universal constants. Then there exists a universal constant $K_3 > 0$ such that with probability at least $1 - K_3 a$ we have

$$|b_{\text{quad}}^* - b_{\text{quad}}^a|_q \leq \frac{2a^q |T|^{1/q} \lambda}{(\xi + 1) F_q(\xi)}. \quad (22)$$

In Theorem 3 we provide the upper bound for the estimation error of the Lasso estimator with the quadratic loss function. This result gives the conditions for estimation consistency of $b_{\text{quad}}^*$ in the high-dimensional scenario, i.e., the number of predictors can be significantly greater than the sample size. Indeed, consistency in the $l_\infty$-norm holds e.g., when $p = \exp(n^{a_1}), |T| = n^{a_2}, a = \exp(-n^{a_2})$, where $a_1 + 2a_2 < 1$. Moreover, $\lambda$ is taken as the right-hand side of the inequality (21) and finally $F_\infty(\xi)$ is bounded from below (or slowly converging to 0) and $\sigma$ is bounded from above (or slowly diverging to $\infty$).

The choice of the $\lambda$ parameter is difficult in practice, which is a common drawback of Lasso estimators. However, Theorem 3 gives us a hint how to choose $\lambda$. The “safe” choice of $\lambda$ is the right-hand side of the inequality (21), so, roughly speaking, $\lambda$ should be proportional to $\sqrt{\log(p)/n}$. In the experimental part of the paper the parameter $\lambda$ is chosen using the cross-validation method. As we will observe it gives satisfactory results for the Lasso estimators in both prediction and variable selection.

Theorem 3 is a crucial fact, which gives the upper bound for (15) in Theorem 2. Namely, taking $q = 1, a = 1/p$ and $\lambda$ equal to the right-hand side of the inequality (21), we obtain the following consequence of Theorem 3.

**Corollary 1.** Suppose that Assumptions 2 is satisfied. Moreover, assume that there exist $\xi_0 > 1$ and constants $C_1 > 0$ and $C_2 < \infty$ such that $F_1(\xi_0) \geq C_1$ and $\sigma \leq C_2$. If $n \geq K_1 |T|^2 \log p$, then

$$P\left( |b_{\text{quad}}^* - b_{\text{quad}}^a|_1 \leq K_2 |T| \sqrt{\frac{\log p}{n}} \right) \geq 1 - K_3 / p, \quad (23)$$

where the constants $K_1$ and $K_2$ depend only on $\xi_0, C_1, C_2$ and $K_3$ is a universal constant provided in Theorem 3.

The above result works for Lasso estimators with the quadratic loss. In the case of the logistic loss analogous result is obtained in ([13], Theorem 1). In fact, their results relate to the case of quite general Lipschitz loss functions, which can be useful in extending Theorem 2 to such cases.

5. Variable Selection Properties of Estimators

In Section 3 we are interested in predictive properties of estimators. In this part of the paper we focus on variable selection, which is another important problem in high-dimensional statistics. As we have already noticed upper bounds for probability of the event $\Omega$ are crucial in proving results concerning prediction. It also plays a key role in establishing results relating to variable selection. In this section we again focus on the Lasso estimators with the quadratic loss functions. The analogous results for Lipschitz loss functions were considered in ([13], Corollary 1).

In the variable selection problem we want to find significant predictors, which, roughly speaking, give us some information on the observed phenomenon. We consider this problem in the semiparametric model, which is defined in (8). In this case the set of significant predictors is given by (9). As we have already mentioned vectors $\beta$ and $b_{\text{quad}}^*$ need not be the same. However, in [32] one proved that for a real number $\gamma$ the following relation

$$(b_{\text{quad}}^*)_j = \gamma \beta_j, \quad j = 1, \ldots, p \quad (24)$$

holds under Assumption 3, which is now stated.
Assumption 3. Let \( \hat{\beta} = (\beta_1, \ldots, \beta_p) \). We assume that for each \( \theta \in \mathbb{R}^p \) the conditional expectation \( \mathbb{E}[\theta^\top X | \hat{\beta}^\top X] \) exists and
\[
\mathbb{E}[\theta^\top X | \hat{\beta}^\top X] = d_\theta \hat{\beta}^\top X
\]
for a real number \( d_\theta \in \mathbb{R} \).

The coefficient \( \gamma \) in (24) can be easily calculated. Namely, we have
\[
\gamma = \frac{\mathbb{E}[Y \hat{\beta}^\top X]}{\hat{\beta}^\top H \hat{\beta}} = \frac{2\mathbb{E}[g(\hat{\beta}^\top X) \hat{\beta}^\top X]}{\hat{\beta}^\top H \hat{\beta}}.
\]
Standard arguments [33] show that \( \gamma \) is nonzero, if \( g \) is monotonic. In this case we have that the set \( \mathbb{T} \) defined in (9) equals to \( T \) defined in (11).

Assumption 3 is a well-known condition in the literature, see e.g., [13,32,34–36]. It is always satisfied in the simple regression model (i.e., when \( X_1 \in \mathbb{R} \)), which is often used for initial screening of explanatory variables, see, e.g., [37]. It is also satisfied when \( X \) comes from the elliptical distribution, like the multivariate normal distribution or multivariate t-distribution. In the interesting paper [38] one advocates that Assumption 3 is a nonrestrictive condition, when the number of predictors is large, which is the case that we focus on in this paper.

Now, we state the results of this part of the paper. We will use the notation \( b_{\text{quad}}^{\text{min}} = \min_{j \in T} |(b_{\text{quad}}^{\text{min}})_j| \).

Corollary 2. Suppose that conditions of Theorem 3 are satisfied for \( q = \infty \). If \( b_{\text{quad}}^{\text{min}} \geq \frac{4\xi \lambda}{(\xi + 1) F_{\text{so}}(\xi)} \), then
\[
P\left( \forall j \in T \setminus K \in T \left| |(b_{\text{quad}}^{\text{min}})_j| > (b_{\text{quad}}^{\text{min}})_j \right. \right) \geq 1 - K_3 a,
\]
where \( K_3 \) is the universal constant from Theorem 3.

In Corollary 2 we show that the Lasso estimator with the quadratic loss is able to separate predictors, if the nonzero coefficients of \( b_{\text{quad}}^{\text{min}} \) are large enough in absolute values. In the case that \( T \) equals (9) (i.e., \( T \) is the set of significant predictors) we can prove that the thresholded Lasso estimator is able to find the true model with high-probability. This fact is stated in the next result. The thresholded Lasso estimator is denoted by \( \hat{\beta}_{\text{th}}^{\text{quad}} \) and defined as
\[
(b_{\text{th}}^{\text{quad}})_j = \hat{b}_{j}^{\text{quad}} \mathbb{I}(\hat{b}_{j}^{\text{quad}} \geq \delta), \quad j = 1, \ldots, p,
\]
where \( \delta > 0 \) is a threshold. We set \( (b_{\text{th}}^{\text{quad}})_0 = b_{\text{quad}}^{\text{quad}} \) and denote \( \hat{T}_{\text{th}} = \{ 1 \leq j \leq p : (b_{\text{th}}^{\text{quad}})_j \neq 0 \} \).

Corollary 3. Let \( g \) in (8) be monotonic. We suppose that Assumption 3 and conditions of Theorem 3 are satisfied for \( q = \infty \). If
\[
\frac{2\xi \lambda}{(\xi + 1) F_{\text{so}}(\xi)} < \delta \leq b_{\text{quad}}^{\text{min}} / 2,
\]
then
\[
P(\hat{T}_{\text{th}} = T) \geq 1 - K_3 a,
\]
where \( K_3 \) is the universal constant from Theorem 3.

Corollary 3 states that the Lasso estimator after thresholding is able to find the true model with high probability, if the threshold is appropriately chosen. However, Corollary 3 does not give a constructive way of choosing the threshold, because both endpoints of the interval \( \left[ \frac{2\xi \lambda}{(\xi + 1) F_{\text{so}}(\xi)}, \frac{b_{\text{quad}}^{\text{min}}}{2} \right] \)
are unknown. It is not a surprising fact and has been already observed, for instance, in linear models ([9], Theorem 8). In the literature we can find methods, which help to choose a threshold in practice, for instance the approach relying on information criteria developed in [39,40].

Finally, we discuss the condition of Corollary 3 that \( b_{\min}^{\text{quad}} \) cannot be too small, i.e., \( b_{\min}^{\text{quad}} \geq \frac{4\xi \lambda}{(\xi + 1) F_\infty(\xi)} \). We know that \( (b_{\star}^{\text{quad}})_j = \gamma \beta_j \) for \( j = 1, \ldots, p \), so the considered condition requires that

\[
\min_{j \in T} |\beta_j| \geq \frac{4\xi \lambda}{|\gamma| (\xi + 1) F_\infty(\xi)} . \tag{26}
\]

Compared to the similar condition for the Lasso estimators in the well-specified models, we observe that the denominator in (26) contains an additional factor \(|\gamma|\). This number is usually smaller than one, which means that in the misspecified models the Lasso estimator needs larger sample size to work well. This phenomenon is typical for misspecified models and the similar restrictions hold for competitors [13].

6. Numerical Experiments

In this section we present simulation study, where we compare the accuracy of considered estimators in prediction and variable selection.

We consider the model (8) with predictors generated from the \( p \)-dimensional normal distribution \( N(0, H) \), where \( H_{jj} = 1 \) and \( H_{jk} = 0.5 \) for \( j \neq k \). The true parameter is

\[
\beta = (1, \pm 1, \pm 1, \ldots, \pm 1, 0, 0, \ldots, 0), \tag{27}
\]

where signs are chosen at random. The first coordinate in (27) corresponds to the intercept and the next ten coefficients relate to significant predictors in the model. We study two cases:

- Scenario 1: \( g(x) = \exp(x) / (1 + \exp(x)) \);
- Scenario 2: \( g(x) = \arctan(x) / \pi + 0.5 \).

In each scenario we generate the data \((X_1, Y_1), \ldots, (X_n, Y_n)\) for \( n \in \{100, 350, 600\}\). The corresponding numbers of predictors are \( p \in \{100, 1225, 3600\} \), so the number of predictors exceeds significantly the sample size in the experiments. For every model we consider two Lasso estimators with unpenalized intercepts (5): the first one with the logistic loss and the second one with the quadratic loss. They are denoted by “logistic” and “quadratic”, respectively. To calculate them we use the “glmnet” package [28] in the “R” software [41]. The tuning parameters \( \lambda \) are chosen on the basis of 10-fold cross-validation.

Observe that applying the Lasso estimator with the logistic loss function to Scenario 1 leads to a well-specified model, while using the quadratic loss implies misspecification. In Scenario 2 both estimators work in misspecified models.

Simulations for each scenario are repeated 300 times.

To describe the quality of estimators in variable selection we calculate two values:

- \( \text{TD} \)—the number of correctly selected relevant predictors;
- \( \text{sep} \)—the number of relevant predictors, whose Lasso coefficients are greater in absolute value than the largest in absolute value Lasso coefficient corresponding to irrelevant predictors.

So, we want to confirm that the considered estimators are able to separate predictors, which we establish in Section 5. Using TD we also study “screening” properties of estimators, which are easier than separability.

The classification accuracy of estimators is measured in the following way: we generate a test sample containing 1000 objects. On this set we calculate
- **pred**—the fraction of correctly predicted classes of objects for each estimator.

The results of experiments are collected in Tables 1 and 2. By the “oracle” we mean the classifier, which works only with significant predictors and uses the function $g$ from the true model (8) in the estimation process.

| Table 1. Results for Scenario 1. |
|---------------------------------|
| $n = 100$ | Quadratic | Logistic | Oracle |
| TD       | 6.3       | 6.1      |        |
| sep      | 2.2       | 2.3      |        |
| pred     | 0.734     | 0.736    | 0.810  |
| $n = 350$ |           |          |        |
| TD       | 9.3       | 9.5      |        |
| sep      | 6.0       | 6.3      |        |
| pred     | 0.774     | 0.779    | 0.831  |
| $n = 600$ |           |          |        |
| TD       | 9.8       | 9.9      |        |
| sep      | 8.6       | 8.9      |        |
| pred     | 0.791     | 0.795    | 0.832  |

| Table 2. Results for Scenario 2. |
|---------------------------------|
| $n = 100$ | Quadratic | Logistic | Oracle |
| TD       | 4.8       | 4.6      |        |
| sep      | 1.4       | 1.4      |        |
| pred     | 0.697     | 0.698    | 0.768  |
| $n = 350$ |           |          |        |
| TD       | 8.1       | 8.2      |        |
| sep      | 3.9       | 3.9      |        |
| pred     | 0.730     | 0.731    | 0.805  |
| $n = 600$ |           |          |        |
| TD       | 9.4       | 9.4      |        |
| sep      | 6.8       | 6.9      |        |
| pred     | 0.750     | 0.752    | 0.809  |

Finally, we also compare execution time of both algorithms. In Table 3 we show the averaged relative time difference

$$
\frac{t^{log} - t^{quad}}{t^{quad}},
$$

where $t^{quad}$ and $t^{log}$ is time of calculating Lasso with quadratic and logistic loss functions, respectively.

| Table 3. Relative time difference (28) of algorithms. |
|-----------------------------------------------------|
| Scenario 1 | Scenario 2 |
| $n = 350$  | 0.02 | 0.06 |
| $n = 600$  | 0.11 | 0.13 |

Looking at the results of experiments we observe that both estimators perform in a satisfactory way. Their predictive accuracy is relatively close to the oracle, especially when the sample size is larger. In variable selection we see that both estimators are able to find significant predictors and separate predictors in both scenarios. Again we can notice that properties of estimators become better, when $n$ increases.
In Scenario 2 the quality of both estimators in prediction and variable selection is comparable. In Scenario 1, which is well-specified for Lasso with the logistic loss, we observe its dominance over Lasso with the quadratic loss. However, this dominance is not large. Therefore, using Lasso with the quadratic loss we obtain slightly worse accuracy of the procedure, but this algorithm is computationally faster. The computational efficiency is especially important, when we study large data sets. As we can see in Table 3 execution time of estimators is almost the same for \( n = 350 \), but for \( n = 600 \) the relative time difference becomes greater than 10%.

Author Contributions: Both authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: The research of K.F. was partially supported by Warsaw University of Life Sciences (SGGW).

Acknowledgments: We would like to thank J. Mielniczuk and the reviewers for their valuable comments, which have improved the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs and Auxiliary Results

This section contains proofs of results from the paper. Additional lemmas are also provided.

Appendix A.1. Results from Section 3

Proof of Theorem 1. For arbitrary \( b \in \mathbb{R}^{p+1} \) the averaged misclassification risk of \( f_b \) can be expressed as

\[
E_D R(b) = E_D E_{(X,Y)} \left[ I(Y = 1)I(b^T \hat{X} < 0) + I(Y = -1)I(b^T \hat{X} \geq 0) \right].
\] (A1)

Moreover, we have

\[
I(Y = -1)I(b^T \hat{X} \geq 0) = I(Y = -1) \left[ 1 - I(b^T \hat{X} < 0) \right].
\] (A2)

Applying (A1) and (A2) for \( \hat{b} \) and \( b^* \), we obtain

\[
\begin{align*}
|E_D R(\hat{b}) - R(b^*)| & = \left| E_D E_{(X,Y)} \left[ I(Y = 1) - I(Y = -1) \right] I(b^T \hat{X} < 0) - I(b^*_T \hat{X} < 0) \right| \\
& \leq E_D E_{(X,Y)} \left| I(b^T \hat{X} < 0) - I(b^*_T \hat{X} < 0) \right| \\
& = P(b^T \hat{X} < 0, b^*_T \hat{X} \geq 0) + P(b^T \hat{X} \geq 0, b^*_T \hat{X} < 0),
\end{align*}
\]

where \( P \) is probability with respect to both the data \( D \) and the new object \( X \). Observe that on the event \( \Omega \), we have that

\[
\hat{b}^T \hat{X} \leq c|\hat{X}|_\infty + b^*_T \hat{X},
\]

so

\[
\begin{align*}
P(\hat{b}^T \hat{X} \geq 0, b^*_T \hat{X} < 0) &= P(\hat{b}^T \hat{X} \geq 0, b^*_T \hat{X} < 0, \Omega) \\
& + P(\hat{b}^T \hat{X} \geq 0, b^*_T \hat{X} < 0, \Omega^c) \\
& \leq P_X(-c|\hat{X}|_\infty \leq b^*_T \hat{X} < 0) + P_D(\Omega^c).
\end{align*}
\]

Analogously, we obtain

\[
P(b^T \hat{X} < 0, b^*_T \hat{X} \geq 0) \leq P_X(0 \leq b^*_T \hat{X} \leq c|\hat{X}|_\infty) + P_D(\Omega^c)
\]
from
\[ b^\top \bar{X} \geq -c|\bar{X}|_\infty + b_1^\top \bar{X}, \]
which finishes the proof.
\[ \square \]

**Lemma A1.** Suppose that Assumption 1 is fulfilled. Moreover, a random variable \( b^\top \bar{X} \) has a density \( h \), which is continuous on the interval \( U = [-2\sigma c \sqrt{\log p}, 2\sigma c \sqrt{\log p}] \) and \( \bar{h} = \sup_{u \in U} h(u) \). Then
\[ P_X(|b^\top \bar{X}| \leq c|\bar{X}|_\infty) \leq 4\sigma \bar{h} c \sqrt{\log p} + 2/p. \]  

**Proof.** For simplicity, we omit the lower index \( X \) in probability \( P_X \) in this proof. We take \( a > 1 \) and obtain inequalities
\[ P(|b^\top \bar{X}| \leq c|\bar{X}|_\infty) = P(|b^\top \bar{X}| \leq c|\bar{X}|_\infty, |\bar{X}|_\infty \leq a) \]
\[ + P(|b^\top \bar{X}| \leq c|\bar{X}|_\infty, |\bar{X}|_\infty > a) \]
\[ \leq P(|b^\top \bar{X}| \leq ca) + P(|\bar{X}|_\infty > a). \]  

The second expression in (A4) equals \( P(|X|_\infty > a) \), because \( a > 1 \). It can be handled using subgaussianity of \( X \) as follows: take \( z > 0 \) and notice that by the Markov inequality and the fact that \( \exp(|u|) \leq \exp(u) + \exp(-u) \) for each \( u \in \mathbb{R} \), we obtain
\[ P(|X|_\infty > a) \leq e^{-za} E \exp(z|X|_\infty) \leq e^{-za} \sum_{j=1}^{p} E \exp(z|X_j|) \]
\[ \leq 2p \exp(\sigma z^2/2 - az). \]

Taking \( z = a/\sigma^2 \), we obtain
\[ P(|X|_\infty > a) \leq 2p \exp(-a^2/(2\sigma^2)). \]

Then we choose \( a = 2\sigma \sqrt{\log p} \), which is not smaller than one, because \( \sigma \geq 1 \) from Assumption 1.

Finally, the first term in (A4) can be bounded from above by \( 2ca\bar{h} = 4\sigma \bar{h} c \sqrt{\log p} \) by the mean value theorem.
\[ \square \]

**Proof of Theorem 2.** The right-hand side of (15) and (17) are upper bounds on the estimation risk. They are obtained using Theorem 1 and Lemma A1. The expressions (16) and (18) are upper bounds for the approximation risk in the case of estimators with the quadratic and logistic loss functions, respectively. In particular, (16) follows from ([17], [Theorem 2.1]) applied for \( f^{\log}_{b_i} \) and Example 3.1. Establishing (18) is similar: we just use ([17], [Theorem 2.1]) applied for \( f^{\log}_{b_i} \) and Example 3.5 to show that
\[ R(b^{\log}_{b_i}) - R_B \leq \sqrt{2E \left[ KL \left( \eta(X), \eta^{\log}_{b_i}(b^{\log}_{b_i})^\top \bar{X} \right) \right]}, \]  

where the Kullback–Leibler distance \( KL(\cdot, \cdot) \) is defined in (10).

Next, we define the function \( h(a) = a \log a + (1 - a) \log(1 - a) \) for \( a \in (0, 1) \). Clearly, we have \( KL(a, b) = h(a) - h(b) - h'(b)(a - b) \) and \( h''(a) = (a(1 - a))^{-1} \). Therefore, from the mean value theorem
\[ KL(a, b) = \frac{(a - b)^2}{2c(1 - c)}, \]  

for some \( c \) between \( a \) and \( b \). To finish the proof we apply (A6) to the right-hand side of (A5) with \( \delta < c < 1 - \delta \).
Appendix A.2. Results from Section 4

To simplify notation we write \( \hat{b}, \hat{b}_s \) for \( \hat{b}^{\text{quad}}, \hat{b}^{\text{quad}}_s \), respectively, in this section. Moreover, we also denote \( \hat{b}_s = ((b_s)_1, \ldots, (b_s)_p) \).

We start with establishing results, which help us to prove Theorem 3.

**Lemma A2.** For \( \hat{b}_s = H^{-1}E[XY] \) we have \( \hat{b}^\top_s HH\hat{b}_s \leq 1 \).

**Proof.** The proof is elementary and based on the inequality

\[
0 \leq E \left[ E[Y|X] - \hat{b}^\top_s X \right]^2. \tag{A7}
\]

The right-hand side of (A7) can be expressed as

\[
E \left[ E[Y|X]^2 - 2\hat{b}^\top_s E[XY] + \hat{b}^\top_s X X^\top \hat{b}_s \right] = E[E[Y|X]^2 - 2\hat{b}^\top_s E[XY] + \hat{b}^\top_s H\hat{b}_s]. \tag{A8}
\]

Using \( \hat{b}_s = H^{-1}E[XY] \), we have \( \hat{b}^\top_s E[XY] = \hat{b}^\top_s HH^{-1}E[XY] = \hat{b}^\top_s H\hat{b}_s \) and we can bound from above the right-hand side of (A8) by

\[
E \left[ Y^2 \right] - \hat{b}^\top_s H\hat{b}_s,
\]

which finishes the proof. \( \square \)

The next result is given in ([42], Corollary 8.2).

**Lemma A3.** Suppose that \( Z_1, \ldots, Z_n \) are i.i.d. random variables and there exists \( L > 0 \) such that \( C^2 = E \exp(|Z_1|/L) \) is finite. Then for arbitrary \( u > 0 \)

\[
P \left( \frac{1}{n} \sum_{i=1}^{n} (Z_i - E[Z_i]) > 2L \left( C \sqrt{\frac{2u}{n}} + \frac{u}{n} \right) \right) \leq \exp(-u). \tag{A9}
\]

**Lemma A4.** For arbitrary \( j = 1, \ldots, p \) and \( u > 0 \) we have

\[
P \left( \frac{2}{n} \sum_{i=1}^{n} X_{ij}(X^\top b_s + E[Y] - Y_i) > 16.4\sigma^2 \left( 3 \sqrt{\frac{2u}{n}} + \frac{u}{n} \right) \right) \leq \exp(-u). \tag{A9}
\]

**Proof.** Fix \( j = 1, \ldots, p \) and \( u > 0 \). Recall that \( H\hat{b}_s = E[XY] \) and \( E[X] = 0 \). Thus, we work with an average of i.i.d. centred random variables, so we can use Lemma A3. We only have to find \( L, C > 0 \) such that

\[
E \exp \left( |X_j(X^\top b_s + E[Y])|/L \right) \leq C^2, \tag{A10}
\]

where \( X_i \) is the \( j \)-th coordinate of \( X \). For each positive number \( a, b \) we have the inequality \( ab \leq \frac{a^2}{2} + \frac{b^2}{2} \).

Therefore, we have

\[
|X_j(X^\top b_s + E[Y])| \leq \frac{X_j^2}{2} + (X^\top b_s)^2 + 4.
\]

Applying this fact and the Schwarz inequality we obtain

\[
E \exp \left( |X_j(X^\top b_s - Y)|/L \right) \leq \exp \left( \frac{4}{L} \right) \sqrt{E \exp \left( \frac{X_j^2}{2L} \right)} \sqrt{E \exp \left( \frac{2(X^\top b_s)^2}{L} \right)}. \tag{A11}
\]
The variable $X_j$ is subgaussian, so using ([43], Lemma 7.4) we can bound the first expectation in (A11) by \( \left( 1 - \frac{2\sigma^2}{L} \right)^{-1/2} \) provided that \( L > 2\sigma^2 \). The second expectation in (A11) can be bounded using subgaussianity of the vector $X_T$, ([43], Lemma 7.4) and Lemma A2 in the following way

\[
E \exp \left( \frac{2(X^\top b_\ast)^2}{L} \right) \leq \left( 1 - \frac{4\sigma^2}{L} \right)^{-1/2},
\]

provided that \( 4\sigma^2 < L \). Taking \( L = 4.1\sigma^2 \) we can bound \( \exp(4/L) \leq 2.7 \), because \( H_{jj} = 1 \) implies that \( \sigma \geq 1 \). Thus, we obtain $C \leq 3$, where $C$ is the upper bound in (A10). It finishes the proof.

\[\square\]

**Lemma A5.** Suppose that assumptions of Theorem 3 are satisfied. Then for arbitrary \( a \in (0, 1) \), \( q \geq 1 \), \( \xi > 1 \) with probability at least \( 1 - Ka \) we have \( \bar{F}_q(\xi) \geq F_q(\xi)/2 \), where $K$ is an universal constant.

**Proof.** Fix \( a \in (0, 1) \), \( q \geq 1 \), \( \xi > 1 \). We start with considering the $l_\infty$-norm of the matrix

\[
\frac{1}{n} \frac{X \tilde{X}^\top}{X \tilde{X}^\top} - E \left[ \frac{X \tilde{X}^\top}{X \tilde{X}^\top} \right]_{\infty} = \max \left( \max_{j=1, \ldots, p} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij}X_{jk} - E [X_jX_k] \right| \right),
\]

(A12)

\[
\max_{j=1, \ldots, p} \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right| \right),
\]

(A13)

We focus only on the right-hand side of (A12), because (A13) can be done similarly. Thus, fix \( j, k \in \{1, \ldots, p\} \). Using subgaussianity of predictors, Lemma A3 and argumentation similar to the proof of Lemma A4 we have

\[
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij}X_{ik} - E [X_jX_k] \right| > K_2\sigma^2 \sqrt{\log(p^2/a) \over n} \right) \leq 2a \over p^2,
\]

where $K_2$ is an universal constant. The values of constants $K_i$ that appear in this proof can change from line to line.

Therefore, using union bounds we obtain

\[
P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_{ij}X_{ik} - E [X_jX_k] \right| > K_2\sigma^2 \sqrt{\log(p^2/a) \over n} \right) \leq K_3a.
\]

Proceeding similarly to the proof of [30, Lemma 4.1] we have the following probabilistic inequality

\[
\bar{F}_q(\xi) \geq F_q(\xi) - K_2(1 + \xi) |T| \sigma^2 \sqrt{\log(p^2/a) \over n}.
\]

To finish the proof we use (20) with $K_1$ being sufficiently large. \[\square\]

**Proof of Theorem 3.** Let \( a \in (0, 1) \), \( q \geq 1 \), \( \xi > 1 \) be arbitrary. The main part of the proof is to show that with high probability

\[
|\hat{b} - b_\ast|_q \leq \frac{\xi |T|^{1/q} \lambda}{(\xi + 1) F_q(\xi)}.
\]

(A14)

Then we apply Lemma A5 to obtain (22).
Thus, we focus on showing that (A14) holds with high probability. Denote \( A = \{ |\nabla \hat{Q}(b_*)|_\infty \leq \frac{1}{\sqrt{n}} \lambda \} \). We start with bounding from below probability of \( A \). Recall that \( b_* \) is the minimizer of \( Q(b) = E(1 - Yb^\top \hat{X})^2 \), which can be easily calculated, namely
\[
\hat{b}_* = H^{-1} E[YX] \quad \text{and} \quad (b_*)_0 = E[Y].
\]

For every \( j = 1, \ldots, p \) the \( j \)-th partial derivative of \( \hat{Q}(b_*) \) at \( b_* \) is
\[
\nabla_j \hat{Q}(b_*) = \frac{2}{n} \sum_{i=1}^{n} X_i j^\top b_* + E[Y] - Y_i.
\]

The derivative with respect to the \( b_0 \) is
\[
\nabla_0 \hat{Q}(b_*) = \frac{2}{n} \sum_{i=1}^{n} (X_i^\top \hat{b}_* + E[Y] - Y_i).
\]

Taking \( \lambda \), which satisfies (21), and using union bounds, we obtain that
\[
P(A^c) \leq \sum_{j=0}^{p} P \left( |\nabla_j \hat{Q}(b_*)| > K_2 e^2 \sqrt{\frac{\log p/a}{n}} \right).
\]

Consider a summand on the right-hand side of (A17), which corresponds to \( j \in \{1, \ldots, p\} \). From (A15) we can handle it using Lemma A4. We just take \( u = \log(p/a) \) and sufficiently large \( K_2 \). Probability of the first term on the right-hand side of (A17), which corresponds to \( j = 0 \), can be bounded from above analogously as in the proof of Lemma A4. Proceeding is even easier, so we omit it.

In further argumentation we consider only the event \( A \). Besides, we denote \( \hat{\theta} = \hat{b} - b_* \), where \( \hat{b} \) is a minimizer of a convex function (5), that is equivalent to
\[
\begin{align*}
\nabla_j \hat{Q}(\hat{b}) &= -\lambda \text{sign}(\hat{b}_j) \quad \text{for} \quad \hat{b}_j \neq 0; \\
|\nabla_j \hat{Q}(\hat{b})| &\leq \lambda \quad \text{for} \quad \hat{b}_j = 0; \\
\nabla_0 \hat{Q}(\hat{b}) &= 0,
\end{align*}
\]

where \( j = 1, \ldots, p \).

First, we prove that \( \hat{\theta} \in C(\xi) \). Here our argumentation is standard [9]. From (A18) and the fact that \( |\theta|_1 = |\theta_T|_1 + |\theta_{T^c}|_1 + |\theta_0| \) we can calculate
\[
0 \leq 2 \theta^\top \mathbb{E}^\top \hat{X} \hat{\theta} / n = \theta^\top \left[ \nabla \hat{Q}(\hat{b}) - \nabla \hat{Q}(b_*) \right] = \sum_{j \in T} \theta_j \nabla_j \hat{Q}(\hat{b}) + \sum_{j \in T^c} \hat{b}_j \nabla_j \hat{Q}(\hat{b}) - \theta^\top \nabla \hat{Q}(b_*) \\
\leq \lambda \sum_{j \in T} |\theta_j| - \lambda \sum_{j \in T^c} |\hat{b}_j| + |\theta_0| \nabla \hat{Q}(b_*) |_\infty \\
= [\lambda + |\nabla \hat{Q}(b_*)|_\infty |\theta_T|_1 + ||\nabla \hat{Q}(b_*)|_\infty - \lambda] |\theta_{T^c}|_1 + |\theta_0||\nabla \hat{Q}(b_*)|_\infty .
\]

Thus, using the fact that we consider the event \( A \) we get
\[
|\theta_T|_1 \leq \frac{\lambda + |\nabla \hat{Q}(b_*)|_\infty |\theta_T|_1 + |\nabla \hat{Q}(b_*)|_\infty - \lambda |\theta_0|}{\lambda - |\nabla \hat{Q}(b_*)|_\infty} \leq \xi |\theta_T|_1.
\]

Therefore, from the definition of \( F_q(\xi) \) we have
\[
|\hat{b} - b_*|_q \leq |T|^{1/q} |\xi^\top \hat{X}(\hat{b} - b_*)|_n |_{\infty} \leq |T|^{1/q} |\nabla \hat{Q}(\hat{b})|_\infty / 2 + |\nabla \hat{Q}(b_*)|_\infty / 2 \leq F_q(\xi) .
\]
Using (A18) and the fact, that we are on \(A\), we obtain (A14).

\[\]

**Appendix A.3. Results from Section 5**

**Proof of Corollary 2.** The proof is a simple consequence of the bound (22) with \(q = \infty\) obtained in Theorem 3. Indeed, for arbitrary predictors \(j \in T\) and \(k \not\in T\) we obtain

\[
\left| \hat{b}_j^{\text{quad}} \right| \geq \left| (b_j^{\text{quad}}) - (b_k^{\text{quad}}) \right| - \left| \hat{b}_k^{\text{quad}} - (b_k^{\text{quad}}) \right| \geq \left| \hat{b}_k^{\text{quad}} - (b_k^{\text{quad}}) \right| \geq \left| \hat{b}_k^{\text{quad}} - (b_k^{\text{quad}}) \right| = \left| \hat{b}_k^{\text{quad}} \right|
\]

\[\]

**Proof of Corollary 3.** The proof is almost the same as the proof of Corollary 2, so it is omitted.

**References**

1. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning; Data Mining, Inference and Prediction*; Springer: New York, NY, USA, 2001.
2. Bühlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*; Springer: New York, NY, USA, 2011.
3. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* 1996, 58, 267–288.
4. Meinshausen, N.; Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 2006, 34, 1436–1462.
5. Zhao, P.; Yu, B. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.* 2006, 7, 2541–2563.
6. Zou, H. The adaptive Lasso and its oracle properties. *J. Am. Stat. Assoc.* 2006, 101, 1418–1429.
7. van de Geer, S. High-dimensional generalized linear models and the Lasso. *Ann. Stat.* 2008, 36, 614–645.
8. Bickel, P.J.; Ritov, Y.; Tsybakov, A.B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* 2009, 37, 1705–1732.
9. Ye, F.; Zhang, C.H. Rate minimaxity of the Lasso and Dantzig selector for the \(l_q\) loss in \(l_r\) balls. *J. Mach. Learn. Res.* 2010, 11, 3519–3540.
10. Huang, J.; Zhang, C.H. Estimation and Selection via Absolute Penalized Convex Minimization and Its Multistage Adaptive Applications. *J. Mach. Learn. Res.* 2012, 13, 1839–1864.
11. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comp. Syst. Sci.* 1997, 55, 119–139.
12. Vapnik, VN. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
13. Kubkowski, M.; Mielenzczuk, J. Selection Consistency of Lasso-Based Procedures for Misspecified High-Dimensional Binary Model and Random Regressors. *Entropy* 2020, 22, 153.
14. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Statist.* 1951, 22, 79–86.
15. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* 1978, 6, 461–464.
16. Quintans, F.; Contreras-Reyes, J.E.; Wiff, R.; Arellano-Valle, R.B. Flexible Bayesian analysis of the von Bertalanffy growth function with the use of a log-skew-t distribution. *Fish. Bull.* 2017, 115, 12—26.
17. Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat.* 2004, 32, 56–85.
18. Bartlett, P.L.; Jordan, M.I.; McAuliffe, J.D. Convexity, classification and risk bounds. *J. Am. Stat. Assoc.* 2006, 101, 138–156.
19. Devroye, L.; Györfi, L.; Lugosi, G. *A Probabilistic Theory of Pattern Recognition*; Springer-Verlag: New York, NY, USA, 1996.
20. Boucheron, S.; Bousquet, O.; Lugosi, G. Introduction to statistical learning theory. *Adv. Lect. Mach. Learn.* 2004, 36, 169–207.
21. Boucheron, S.; Bousquet, O.; Lugosi, G. Theory of classification: A survey of some recent advances. *ESAIM P&S* 2005, 9, 323–375.
22. Bartlett, P.L.; Bousquet, O.; Mendelson, S. Local Rademacher complexities. *Ann. Stat.* 2005, 33, 1497–1537.
23. Audibert, J.Y.; Tsybakov, A.B. Fast learning rates for plug-in classifiers. *Ann. Stat.* **2007**, *35*, 608–633.
24. Blanchard, G.; Bousquet, O.; Massart, P. Statistical performance of support vector machines. *Ann. Stat.* **2008**, *36*, 489–531.
25. Tarigan, B.; van de Geer, S. Classifiers of support vector machine type with $l_1$ complexity regularization. *Bernoulli* **2006**, *12*, 1045–1076.
26. Abramovich, F.; Grinshtein, V. High-Dimensional Classification by Sparse Logistic Regression. *IEEE Trans. Inf. Theory* **2019**, *65*, 3068–3079.
27. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499.
28. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
29. Buldygin, V.; Kozachenko, Y. *Metric Characterization of Random Variables and Random Processes*; American Mathematical Society: Providence, RI, USA, 2000.
30. van de Geer, S.; Bühlmann, P. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **2009**, *3*, 1360–1392.
31. Li, K.C.; Duan, N. Regression analysis under link violation. *Ann. Stat.* **1989**, *17*, 1009–1052.
32. Thorisson, H. Coupling methods in probability theory. *Scand. J. Stat.* **1995**, *22*, 159–182.
33. Brillinger, D.R. *A Generalized Linear Model with Gaussian Regressor Variables*; A Festschrift for Erich Lehmann; Bickel, P.J.; Doksum, K.; Hodges, J.L., Eds.; Wadsworth: Belmont, CA, USA, 1983; pp. 97–114.
34. Ruud, P.A. Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models. *Econometrica* **1983**, *51*, 225–228.
35. Zhong, W.; Zhu, L.; Li, R.; Cui, H. Regularized quantile regression and robust feature screening for single index models. *Stat. Sin.* **2016**, *26*, 69–95.
36. Fan, J.; Lv, J. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **2008**, *70*, 849–911.
37. Hall, P.; Li, K.C. On almost Linearity of Low Dimensional Projections from High Dimensional Data. *Ann. Stat.* **1993**, *21*, 867–889.
38. Pokarowski, P.; Mielniczuk, J. Combined $l_1$ and Greedy $l_0$ Penalized Least Squares for Linear Model Selection. *J. Mach. Learn. Res.* **2015**, *16*, 961–992.
39. Pokarowski, P.; Rejchel, W.; Solty, A.; Frej, M.; Mielniczuk, J. Improving Lasso for model selection and prediction. *arXiv* 2019, arXiv:1907.03025.
40. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
41. van de Geer, S. *Estimation and Testing under Sparsity*; Springer: Berlin, Germany, 2016.
42. Baraniuk, R.; Davenport, M.A.; Duarte, M.F.; Hegde, C. *An Introduction to Compressive Sensing*; Connexions, Rice University: Houston, TX, USA, 2011.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).