

Quasi-likelihood and/or robust estimation in high dimensions*

Sara van de Geer and Patric Müller

ETH Zürich

Abstract. We consider the theory for the high-dimensional generalized linear model with the Lasso. After a short review on theoretical results in literature, we present an extension of the oracle results to the case of quasi-likelihood loss. We prove bounds for the prediction error and ℓ_1 -error. The results are derived under fourth moment conditions on the error distribution. The case of robust loss is also given. We moreover show that under an irrepresentable condition, the ℓ_1 -penalized quasi-likelihood estimator has no false positives.

Key words and phrases: high-dimensional model, quasi-likelihood estimation, robust estimation, sparsity, variable selection.

1. A REVIEW OF THE THEORY IN LITERATURE

Consider n independent observations $\{(x_i^T, Y_i)\}_{i=1}^n$, where $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is a random response variable, and x_i is a fixed p -dimensional vector of co-variables, $i = 1, \dots, n$. In a high-dimensional model, the number of co-variables p is much larger than the number of observations n . There has been much literature on the linear model for this situation. In that case, one assumes that

$$Y_i = x_i^T \beta^0 + \epsilon_i, \quad i = 1, \dots, n,$$

where $\beta^0 \in \mathbb{R}^p$ is an unknown vector of coefficients, and $\epsilon_1, \dots, \epsilon_n$ are independent noise variables. The Lasso estimator (Tibshirani [1996]) is

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^n |Y_i - x_i^T \beta|^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}.$$

The parameter $\lambda > 0$ is a regularization parameter, and $\|\beta\|_1 := \sum_{j=1}^p |\beta_j|$ is the ℓ_1 -norm of β . For the case of orthogonal design, that is, the case where the

Seminar for Statistics, ETH Zürich, Rämistrasse 101, 8092 Zürich (e-mail: geer@stat.math.ethz.ch; mueller@stat.math.ethz.ch).

*Research supported by SNF 20PA21-120050.

columns of the $n \times p$ design matrix

$$\mathbf{X} := \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$$

are orthogonal, the Lasso estimator is the soft-thresholding estimator (Donoho [1995]). We study in this paper the extension of the theoretical results for the Lasso estimator, to the case of generalized linear models.

The theory for the Lasso with least squares loss is well established. We refer to Bunea et al. [2006], Bunea et al. [2007a], Bunea et al. [2007c], van de Geer [2007], Lounici [2008], Bickel et al. [2009]. See also Bühlmann and van de Geer [2011] and the references therein. The main results concern oracle inequalities for the prediction error $\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2$ and variable selection properties of the Lasso. Oracle results say that the prediction error of the Lasso estimator is up to log-factors as good as that of an oracle that uses the least squares “estimator” with only the co-variables in the unknown active set $S_0 := \{j : \beta_j^0 \neq 0\}$. Variable selection results roughly state that with large probability the estimated active set $\hat{S} := \{j : \hat{\beta}_j \neq 0\}$ is with large probability equal to the true active set S_0 . Both results depend on appropriate conditions: for prediction one assumes restricted eigenvalue condition (Koltchinskii [2009a], Koltchinskii [2009b], Bickel et al. [2009]) or compatibility conditions (van de Geer [2007]), and for variable selection, one assumes the neighborhood stability (Meinshausen and Bühlmann [2006]) or equivalent irrepresentable condition (Zhao and Yu [2006]). Clearly, variable selection is a harder problem than prediction, so that one expects conditions for the former to be stronger than those for the latter. Indeed, van de Geer and Bühlmann [2009] show that the irrepresentable condition implies the compatibility condition.

Concerning work on oracle inequalities for general loss, an earlier paper which uses ℓ_1 -regularization in this context is Loubes and van de Geer [2002]. Here, the case of orthogonal design is considered (thus, it has $p \leq n$). The technique of proof is however very much along the lines of the later proofs for non-orthogonal design (with possibly $p > n$), as developed by van de Geer [2007] and others. Some remarks on the proof technique can be found in van de Geer [2001], highlighting that with an ℓ_1 -penalty one can derive oracle inequalities with rates faster than $1/\sqrt{n}$, despite the fact that the penalty-term $\lambda\|\beta^0\|_1$ itself is generally of larger order than $1/\sqrt{n}$. The case of quantile regression was studied in van de Geer [2003], again only for the case of orthonormal design. In Targin and van de Geer [2006], hinge loss with ℓ_1 -penalty is studied. Here the design is not assumed to be orthogonal, and is in fact random. This paper does not use restricted eigenvalue or compatibility conditions, but rather a weighted eigenvalue condition. It shows that the ℓ_1 -penalty leads to estimators which are both adaptive to the “smoothness” or “complexity” of the underlying regression function, as well as to the “margin behavior” of the problem. The margin behavior expresses the amount of curvature of the theoretical risk near its minimum. The paper Bunea et al. [2007b] considers the density estimation problem.

In van de Geer [2007], results are derived for generalized linear models with ℓ_1 -penalty and p possibly larger than n , assuming the compatibility condition. It covers the case of quadratic loss and of general Lipschitz loss, and it allows for random design. Similar results are in van de Geer [2008], although there the compatibility condition is replaced by one somewhat in the spirit conditions in Juditsky and Nemirovski [2011]. In Bühlmann and van de Geer [2011], one can find further details concerning sparsity oracle inequalities for high-dimensional generalized linear models.

There is a large body of literature extending the oracle results for the linear model to matrix versions. It is beyond the scope of this paper to review this work, and we only point to the generalization to robust loss, as given in Candès et al. [2009].

Within this volume, the paper Negahban et al. [2011] gives a general account of oracle results for high-dimensional M-estimators. After our Theorem 5.2, we briefly discuss its relation with Negahban et al. [2011].

Concerning variable selection, the fact that the irrepresentable condition is rather strong has led to considering modifications of the Lasso, such as two step procedures, and the SCAD introduced by Fan [1997], see e.g. Wu and Liu [2009] for the case of quantile regression.

Our paper focusses only on the theoretical aspects. There is much literature on applications of the Lasso in generalized linear models, see Wu et al. [2009] for example. The computational aspects are well-studied: see Friedman et al. [2010]. The paper Lambert-Lacroix and Zwald [2011] contains apart from theory also software descriptions and a real data example for the case of Huber loss. In Wang et al. [2007], ℓ_1 -regularization with least absolute deviations loss is studied and compared numerically with the least squares Lasso.

We present new results for prediction and variable selection for the case of quasi-likelihood estimation. The findings for prediction are along the lines as those in van de Geer [2008], but this time completed with the compatibility condition. The paper details and extends the findings in Bühlmann and van de Geer [2011]. We also show that a weighted form of the irrepresentable condition implies consistent variable selection.

2. QUASI-LIKELIHOOD AND ROBUST LOSS

We model the dependence of the distribution of Y_i on x_i via a linear function $f_{\beta^0}(x_i) := x_i^T \beta^0$, where β^0 is a vector of unknown coefficients. The problem is to estimate β^0 or the linear predictor vector $f_{\beta^0} := \mathbf{X}\beta^0$, where $\mathbf{X}^T := (x_1, \dots, x_n)$. We study a high-dimensional situation, where the number of variables p can be much larger than the sample size n . (For technical reasons, we assume that p is at least 2.) The vector β^0 is assumed to be sparse, that is, its number of non-zero coefficients is assumed to be small. See Subsection 2.2 for more details on sparsity.

We consider two models. The first one is a generalized linear model, with a given inverse link function G , that is

$$\mathbf{E}(Y_i|x_i) := \mu_0(x_i) = G(x_i^T \beta^0), \quad i = 1, \dots, n,$$

with $\beta^0 \in \mathbb{R}^p$ a vector of unknown coefficients. The quasi-(log)likelihood function is

$$Q(y, \mu) := \int_y^\mu \frac{y - u}{V(u)} du, \quad y, \mu \in \mathcal{Y},$$

where $V : \mathbb{R} \rightarrow (0, \infty)$ is a given variance function, see also McCullagh and Nelder [1989]. Together, quasi-likelihood and link function define quasi-likelihood loss, as follows:

DEFINITION 2.1. *The quasi-likelihood loss function is*

$$\rho(y, z) := -Q(y, G(z)), \quad y \in \mathcal{Y}, \quad z \in \mathbb{R}.$$

In our second model, the dependence of the distribution of Y_i on x_i may be described through quantiles or other aspects of the distribution. In particular, one can define this dependence via a loss function $\{\rho(y, z) : y \in \mathcal{Y}, z \in \mathbb{R}\}$, and

$$f_i^0 := \arg \min_{z \in \mathbb{R}} \mathbf{E} \left(\rho(Y_i, z) \middle| x_i \right).$$

The generalized linear model assumes that $f_i^0 = x_i^T \beta^0$ for some $\beta^0 \in \mathbb{R}^p$.

The robust case is the one where, for all $y \in \mathcal{Y}$, the loss function $\rho(y, z)$ is Lipschitz in z , with Lipschitz constant not depending on y . Without loss of generality one can then assume the Lipschitz constant to be equal to one. This leads to the following definition:

DEFINITION 2.2. *The loss function ρ is robust if for all $y \in \mathcal{Y}$,*

$$|\rho(y, z) - \rho(y, \tilde{z})| \leq |z - \tilde{z}|, \quad \forall z, \tilde{z}.$$

.

Quasi-likelihood loss is sometimes robust, but there are also many examples where it is not. Moreover, there are many (robust) loss functions which do not correspond to minus quasi-likelihoods. See Section 3 for some examples.

To handle the large p situation, one needs a regularized estimation method. Let us write a linear function with coefficients β as

$$f_\beta(x) = x^T \beta.$$

In what follows, we sometimes, with some abuse of notation, let f_β be the n -dimensional vector $\mathbf{X}\beta = (f_\beta(x_1), \dots, f_\beta(x_n))^T \in \mathbb{R}^n$ as well.

The ℓ_1 -norm of a vector $\beta \in \mathbb{R}^p$ is

$$\|\beta\|_1 := \sum_{j=1}^p |\beta_j|.$$

We examine the ℓ_1 -penalized estimator $\hat{\beta}$ of β^0 , defined as

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n \rho(Y_i, f_\beta(x_i)) + \lambda \|\beta\|_1 \right\}.$$

Here, $\lambda > 0$ is a tuning parameter. Large values correspond to more regularization, which means more shrinkage of the estimator $\hat{\beta}$. The expression

$$\frac{1}{n} \sum_{i=1}^n \rho(Y_i, f_\beta(x_i))$$

is called the empirical risk (at β). For least squares loss (i.e., $\rho(y, u) = (y - u)^2$), the empirical risk is the usual sum of squares (normalized by $1/n$). The above estimator is then called the Lasso estimator (Tibshirani [1996]).

We will study loss functions ρ that are either minus quasi-likelihoods or robust (or both). The normalized Euclidean norm on \mathbb{R}^n is

$$\|f\|_n := \sqrt{f^T f / n}, \quad f \in \mathbb{R}^n.$$

We will establish bounds for the “prediction error” $\|f_{\hat{\beta}} - f_{\beta^0}\|_n^2$, the ℓ_1 -error $\|\hat{\beta} - \beta^0\|_1$, and (for the case of quasi-likelihood loss) present sufficient conditions for variable selection using $\hat{\beta}$.

2.1 Convex loss

We require throughout this paper, both for quasi-likelihood loss as well as for robust loss, that the map

$$z \mapsto \rho(y, z)$$

is convex for all $y \in \mathcal{Y}$. This assumption is important from a computational point of view. It also plays a crucial role in our theory, as it allows us to prove that the estimator $\hat{\beta}$ is in an ℓ_1 -neighborhood of β^0 . This in turn will be invoked to establish sup-norm bounds for $f_{\hat{\beta}}$.

2.2 Sparsity

The indices of the set of non-zero coefficients of β^0 is called the (true) active set. It is denoted by

$$S_0 := \{j : \beta_j^0 \neq 0\}.$$

Its cardinality $s_0 := |S_0|$ is called the sparsity index of β^0 . It is assumed that s_0 is relatively small, at least smaller than $\sqrt{n/\log p}$ in order of magnitude (see (5.3), (6.1), (7.2), (7.3) and (8.1)). The vector β^0 is sparse if s_0 is small.

More generally, one can call a vector β^0 sparse if it can in some sense be approximated by a vector with only a few non-zero entries. To avoid too many digressions, we will not elaborate on this issue, but only present a brief outline after the formulation of the main oracle result (see Remark 5.5).

2.3 Results in this paper

As β^0 is unknown, its active set S_0 and its sparsity index s_0 are unknown as well. We will show in Theorems 5.2 and 6.1 that the prediction error of the ℓ_1 -penalized estimator $\hat{\beta}$ is, up to a $\log p$ -term, the same as that of minimizer of the empirical risk without penalty but with all coefficients not in S_0 restricted to be zero. The latter is not an estimator, as it depends on the unknown S_0 . It is often referred to as the oracle. We moreover show that a version of the irrepresentable condition, appropriate for quasi-likelihood loss, is sufficient for variable selection (see Theorem 7.3). All our results are stated in a non-asymptotic form, but to facilitate the interpretation, we also give asymptotic formulations.

2.4 Organization of the paper

The next section provides some examples of quasi-likelihood and robust loss. Section 4 gives the definition of the so-called *compatibility constant*, which will occur in the oracle results. Section 5 gives oracle inequalities for the prediction and ℓ_1 -error for quasi-likelihood loss, and Section 6 does the same for robust loss. In Section 7 we address the variable selection problem in the quasi-likelihood context. Similar arguments can be used in the robust context, but this is omitted here. Section 8 briefly discusses the case of random design, and Section 9 concludes. The proofs are in the supplemental article van de Geer and Müller [2012]. Lemmas ?? and ?? there are based on a concentration inequality (see Massart [2000]) and a contraction inequality (see Ledoux and Talagrand [1991]). These lemmas use only fourth moment assumptions, and are perhaps of interest in themselves.

3. EXAMPLES OF LOSS FUNCTIONS

3.1 Least squares loss

The least squares criterion has $\mathcal{Y} = \mathbb{R}$. It corresponds to a quasi-likelihood loss with variance function $V(u) = 1$ for all $u \in \mathbb{R}$. The link function is then the identity, which is the canonical link function for this case. The loss function is convex, but not robust.

3.2 Logistic loss

When the response Y_i is binary, say $Y_i \in \{0, 1\}$, $i = 1, \dots, n$, we have

$$\mathbf{E}(Y_i|x_i) = \mathbb{P}(Y_i = 1|x_i).$$

In logistic regression, one takes the quasi-likelihood with variance function $V(u) = u(1 - u)$, $u \in (0, 1)$, and the canonical link function

$$\gamma(\mu) := \log\left(\frac{\mu}{1 - \mu}\right), \quad \mu \in (0, 1),$$

that is

$$G(z) = \gamma^{-1}(z) = \frac{e^z}{1 + e^z}, \quad z \in \mathbb{R}.$$

Hence, in this case

$$\rho(y, z) = yz - \log(1 + e^z), \quad z \in \mathbb{R}.$$

Because $\mathcal{Y} = \{0, 1\}$, one sees that this leads to a robust loss function, i.e., $z \mapsto \rho(y, z)$ is Lipschitz in z for all $y \in \mathcal{Y}$. We acknowledge that logistic regression is not robust in the sense of having a bounded influence function (but we will in fact assume in Condition A1 that the covariables are bounded). As in all cases of quasi-likelihood with canonical link function, the loss also convex.

3.3 Binary response with other link functions

Consider binary response $Y_i \in \{0, 1\}$ as in Subsection 3.2, but now with more general inverse link function G :

$$\mathbb{P}(Y_i = 1|x_i) = G(x_i^T \beta^0), \quad i = 1, \dots, n.$$

If $G : \mathbb{R} \rightarrow [0, 1]$ is a strictly increasing symmetric distribution function, then quasi-likelihood loss is convex. This is because the hazard $g(u)/(1 - G(u))$ (g being the derivate of G) is a decreasing function of u . When the hazard is uniformly bounded, quasi-likelihood loss is also robust.

3.4 Quantile regression

If the dependence of the distribution of $Y_i \in \mathbb{R}$ on x_i is via its α -quantile ($0 < \alpha < 1$), we take as loss function

$$\rho(y, z) = \rho(y - z),$$

where

$$\rho(z) = \alpha|z|1\{z > 0\} + (1 - \alpha)|z|1\{z \leq 0\}.$$

This is clearly a robust loss function, but it does not correspond to a quasi-likelihood.

4. THE COMPATIBILITY CONDITION

Let $S \subset \{1, \dots, p\}$ be an index set with cardinality s . We define for all $\beta \in \mathbb{R}^p$,

$$\beta_{S,j} := \beta_j 1\{j \in S\}, \quad j = 1, \dots, p, \quad \beta_{S^c} := \beta - \beta_S.$$

Below, we present for constants $L > 0$ the compatibility constant $\phi(L, S)$ introduced in van de Geer [2007]. For normalized design (i.e., $\|\mathbf{X}_j\|_n = 1$ for all j , where \mathbf{X}_j denotes the j -th column of \mathbf{X}), one can view $1 - \phi^2(1, S)/2$ as an ℓ_1 -version of the canonical correlation between the linear space spanned by the variables in S on the one hand, and the linear space of the variables in S^c on the other hand. Instead of all linear combinations with normalized ℓ_2 -norm, we now consider all linear combinations with normalized ℓ_1 -norm of the coefficients. For a geometric interpretation, we refer to van de Geer and Lederer [2012].

Definition *The compatibility constant is*

$$\phi^2(L, S) := \min\{s\|f_\beta\|_n^2 : \|\beta_S\|_1 = 1, \|\beta_{S^c}\|_1 \leq L\}.$$

The compatibility constant is closely related to (and never smaller than) the restricted eigenvalue as defined in Bickel et al. [2009], which is

$$\phi_{\text{RE}}^2(L, S) = \min\left\{\frac{\|f_\beta\|_n^2}{\|\beta_S\|_2^2} : \|\beta_{S^c}\|_1 \leq L\|\beta_S\|_1\right\}.$$

The calculation of the compatibility constant is a nonlinear eigenvalue problem (see e.g. Hein and Buehler [2010] for computational aspects of nonlinear eigenvalues). Lower bounds that hold with high-probability follow for example if \mathbf{X} is an i.i.d. sample from a p -dimensional vector with non-degenerate covariance matrix (see Section 8 for some details). See also Koltchinskii [2009a], and see van de Geer and Bühlmann [2009] for a discussion of the relation between restricted eigenvalues and compatibility.

For oracle results, we need $\phi(L, S_0)$ to be strictly positive for some $L > 1$ (depending on the tuning parameter λ). In this paper, we take $L = 3$ for definiteness, and we require throughout that $\phi(3, S_0) > 0$ (except when we consider sparse approximations of the truth, see Remark 5.5). If $\phi(3, S_0) = 0$, one sees that some conditions (e.g. condition (5.3)) become impossible.

As we will see, all bounds in this paper involve not so much the sparsity index s_0 itself, but rather the *effective sparsity*

$$\Gamma_{\text{effective}}(S_0) := \frac{s_0}{\phi^2(3, S_0)}.$$

EXAMPLE 4.1. *As a simple numerical example, let us suppose $n = 2$, $p = 3$, $S_0 = \{3\}$, and*

$$\mathbf{X} = \sqrt{n} \begin{pmatrix} 5/13 & 0 & 1 \\ 12/13 & 1 & 0 \end{pmatrix}.$$

Thus, the sparsity index is $s_0 = 1$. One can easily verify that there is no $\beta \in \mathbb{R}^p$ with $\mathbf{X}\beta = 0$ and $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$. Thus, the compatibility constant $\phi^2(3, S_0)$ is strictly positive. In fact, $\phi(3, S_0)$ is equal to the distance of \mathbf{X}_1 to line that connects $3\mathbf{X}_1$ and $-3\mathbf{X}_2$, that is $\phi(3, S_0) = \sqrt{2/13}$. The effective sparsity is

$\Gamma_{\text{effective}}(S_0) = 13/2$.

Alternatively, when

$$\mathbf{X} = \sqrt{n} \begin{pmatrix} 12/13 & 0 & 1 \\ 5/13 & 1 & 0 \end{pmatrix},$$

then $\phi(3, S) = 0$. This is due to the sharper angle between \mathbf{X}_1 and \mathbf{X}_3 .

5. ORACLE INEQUALITIES FOR QUASI-LIKELIHOOD LOSS

5.1 The case of least squares loss

To appreciate the results we will present for the general case, it may be useful to first reconsider the standard linear model and least squares loss. Let $Y = (Y_1, \dots, Y_n)^T$ and suppose

$$Y = \mathbf{X}\beta^0 + \epsilon.$$

Let $\hat{\beta}$ be the Lasso estimator

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - \mathbf{X}\beta\|_n^2 + \lambda \|\beta\|_1 \right\}.$$

Let \mathbf{X}_j denote the j -th column of the design matrix \mathbf{X} . If the errors $\epsilon; = (\epsilon_1, \dots, \epsilon_n)^T$ are independent with mean zero and the design is normalized (that is, $\|\mathbf{X}_j\|_n = 1$ for all j) one can prove that uniformly in j , the ‘‘correlations’’ $\epsilon^T \mathbf{X}_j / n$ are small in absolute value, generally as small as $O(\sqrt{\log p / n})$. The regularization parameter λ is to be chosen in such a way that it ‘‘overrules’’ these correlations. Indeed, this allows one to prove the following result (see Bühlmann and van de Geer [2011], Theorem 6.1) by rather elementary means (recall the notation $f_\beta := \mathbf{X}\beta$):

THEOREM 5.1. *Suppose that $\lambda \geq 4 \max_{1 \leq j \leq p} |\epsilon^T \mathbf{X}_j| / n$. Then*

$$\|f_{\hat{\beta}} - f_{\beta^0}\|_n^2 + \lambda \|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 \Gamma_{\text{effective}}(S_0).$$

This result says that if the effective sparsity $\Gamma_{\text{effective}}(S_0)$ is of the same order as the sparsity index $s_0 := |S_0|$ (i.e., if the compatibility constant stays away from zero), then for a large class of error distributions the Lasso estimator with $\lambda \asymp \sqrt{\log p / n}$ is up to constants and a $(\log p)$ -factor as good as the oracle least squares ‘‘estimator’’ which knows the active set S_0 . The performance of $\hat{\beta}$ is here measured in terms of its prediction error¹ $\|X(\hat{\beta} - \beta^0)\|_n^2$. Theorem 5.1 moreover says that the ℓ_1 error converges with rate $\lambda \Gamma_{\text{effective}}(S_0)$. Looking ahead at more general loss functions, ideas are based on quadratic approximations, which are generally only valid in a neighborhood of β^0 . This is why in our work, we will assume that $\lambda \Gamma_{\text{effective}}(S_0)$ is small, say $\lambda \Gamma_{\text{effective}}(S_0) \leq \gamma$, where γ is a sufficiently small constant. With $\lambda \asymp \sqrt{\log p / n}$, and a compatibility constant staying away from zero, it means we assume the sparsity index s_0 to be sufficiently smaller than $\sqrt{n / \log p}$.

¹The prediction error of the predictor $f_{\hat{\beta}}$ of an independent copy $Y_{\text{new}} := f_{\beta^0} + \epsilon_{\text{new}}$ of Y is rather $\|f_{\hat{\beta}} - f_{\beta^0}\|_n^2 + \sigma^2$, where $\sigma^2 = \mathbf{E}\|\epsilon_{\text{new}}\|_n^2$. We however do not include the additional variance σ^2 in our definition.

5.2 General quasi-likelihood loss

As in the situation of the standard linear model and least squares loss, we will study the error $\|f_{\hat{\beta}} - f_{\beta^0}\|_n^2$ and the ℓ_1 -error. For prediction, one will be interested in estimating the mean $\mu_0 = G(f_{\beta^0})$ of the response variable Y . Our Conditions A3 and A4 below will ensure that G has a bounded derivative on an appropriate domain. This means that bounds for $\|f_{\hat{\beta}} - f_{\beta^0}\|_n$ immediately lead to similar bounds for $\|G(f_{\hat{\beta}}) - G(f_{\beta^0})\|_n$. With some abuse of terminology, we refer to $\|f_{\hat{\beta}} - f_{\beta^0}\|_n^2$ as the prediction error.

The theoretical properties of the ℓ_1 -penalized quasi-likelihood estimator $\hat{\beta}$ depend on the tail-behavior of the error

$$\epsilon_i := Y_i - \mu_0(x_i), \quad i = 1, \dots, n.$$

We will need at least finite second moments of the errors. For definiteness, we assume the errors have finite fourth moments. With higher order moments, the confidence level in the oracle result of Theorem 5.2 will be larger, and when the errors have sup-exponential tails, one can derive exponential probability inequalities for prediction error and ℓ_1 -error.

Condition A $_{\epsilon}$ *There exist constants $\sigma > 0$ and $\kappa > 0$ such that*

$$\max_{1 \leq i \leq n} \mathbb{E} \epsilon_i^2 \leq \sigma^2,$$

and

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(\epsilon_i^2 - \mathbb{E} \epsilon_i^2 \right)^2 \leq \kappa^4.$$

The next conditions, Conditions A1-A4, allow us to use quadratic approximations in a neighborhood of β^0 . We assume throughout that the inverse link function G is increasing and that its derivative

$$g(z) := \frac{dG(z)}{dz}, \quad z \in \mathbb{R},$$

exists. We further define

$$(5.1) \quad \gamma(\mu) := \int_{y_0}^{\mu} \frac{1}{V(u)} du, \quad B(\mu, \mu_0) := \int_{\mu_0}^{\mu} \frac{u - \mu_0}{V(u)} du, \quad \mu \in \mathcal{Y},$$

where y_0 is an arbitrary but fixed constant. We let

$$(5.2) \quad H(z) := \gamma(G(z)), \quad z \in \mathbb{R},$$

that is, $H := \gamma \circ G$. Note that γ is (up to an additive constant) the canonical link function. When $G = \gamma^{-1}$, we get $H(z) = z$ for all z . The term $yH(z)$ in the quasi-likelihood $Q(y, G(z))$ containing the response y is then linear in z . In a sense, H measures the departure from linearity of this term. We let

$$h(z) := \frac{dH(z)}{dz} = \frac{g(z)}{V(G(z))}, \quad z \in \mathbb{R}.$$

The quantity $B(\mu, \mu_0)$ is the “regret” for choosing the expectation μ instead of the “true” μ_0 .

Condition A1 *There exists a constant K_X such that*

$$\max_{1 \leq j \leq p} \max_{1 \leq i \leq n} |x_{i,j}| \leq K_X.$$

We remark that Condition A1 serves as normalization of the design, albeit not in terms of the $\|\cdot\|_n$ norm but rather in supremum norm. As our results will be presented in non-asymptotic form, it is in principle possible to see the effect when, say, K_X grows with p and/or n .

Condition A2 *There exists a constant K_0 such that*

$$\max_{1 \leq i \leq n} |f_{\beta^0}(x_i)| \leq K_0.$$

Condition A3 *With K_X and K_0 given in Conditions A1 and A2 respectively, there exists a positive constant C_h such that for all $|z| \leq K_X + K_0$,*

$$1/C_h \leq h(z) \leq C_h.$$

Condition A4 *With K_X and K_0 given in Conditions A1 and A2 respectively, there exists a constant C_V , such that for all $|z| \leq K_X + K_0$,*

$$2/C_V \leq V \circ G(z) \leq C_V/2.$$

REMARK 5.1. *There is an interplay between Conditions A_ϵ , A1 and A2. For example, for quadratic loss, we do not need A1 and A2 when the errors are (sub)Gaussian. Conditions A1 and A2 are imposed so that we need the Conditions A3 and A4 only in the neighborhood $|z| \leq K_X + K_0$. As for Condition A3, when G is the inverse of the canonical link function γ , it holds with $C_h = 1$, as H is then the identity. For quadratic loss, and logistic loss for example (which have canonical link function), Condition A4 holds as well. We actually will only need the lower bound for $V \circ G$ in this section, and the upper bound will come into play in Section 7.*

To organize the constants appearing in our results, let use the short hand notation

$$\begin{aligned} C_{h,V} &:= C_V C_h^2, \\ C_{h,X} &:= 16C_h K_X, \\ \Gamma(S_0) &:= 16C_{h,V} \Gamma_{\text{effective}}(S_0). \end{aligned}$$

Thus, up to constants $\Gamma(S_0)$ is the effective sparsity. As in the case of least squares loss, we assume the regularization parameter λ to be of order at least $\sqrt{\log p/n}$. The larger λ , the larger the confidence level of our bounds will be (in Theorem 5.1 this the probability of $4 \max_{1 \leq j \leq p} |\epsilon^T \mathbf{X}_j|/n \leq \lambda$) but then these bounds themselves are also larger. We introduce a variable $t > 0$ to describe this effect, and define

$$\lambda_\epsilon(t) := C_{h,X} \sigma \sqrt{\frac{2(t + \log p)}{n}}.$$

If we choose the tuning parameter λ at least as large as $4\lambda_\epsilon(t)$, the confidence level will be at least $1 - \alpha(t)$, where

$$\alpha(t) := \alpha(t) := 3 \exp[-t] + 3\kappa^4/(n\sigma^4).$$

The variable t is in principle arbitrary, but it is however not allowed to be arbitrarily large. As we can only apply the quadratic approximations in a neighborhood of β^0 we will need to show that $\hat{\beta}$ is with large probability in such a neighborhood. For that reason, we cannot let the tuning parameter λ to be arbitrarily large (as a large λ will give slow rates): see condition (5.4) in 5.2 below. A reasonable choice for t is for example $t \asymp \log n$, in which case $\alpha(t) \asymp 1/n$.

THEOREM 5.2. *Let $\hat{\beta}$ be the ℓ_1 -penalized quasi-likelihood estimator. Assume Conditions A_ϵ and A1-A4. Suppose that*

$$(5.3) \quad \lambda_\epsilon(t) \Gamma(S_0) \leq \frac{1}{4}.$$

Take

$$(5.4) \quad 4\lambda_\epsilon(t) \leq \lambda \leq \frac{1}{\Gamma(S_0)}.$$

With probability at least $1 - \alpha(t)$, it holds that

$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{\lambda}{2} \Gamma(S_0),$$

and

$$\|f_{\hat{\beta}} - f_{\beta^0}\|_n^2 \leq \frac{3}{4} C_{h,V} \lambda^2 \Gamma(S_0).$$

REMARK 5.2. *Our result in Theorem 5.2 is comparable to Corollary 3 in Negahban et al. [2011], albeit that we do not assume bounded responses or canonical link function, and our compatibility condition is weaker than the there assumed restricted eigenvalue condition. On the other hand, we require (5.3), and only give bounds for the ℓ_1 -error and prediction error, not for the ℓ_2 -error.*

REMARK 5.3. *We have presented the result in a non-asymptotic form, but did not try to optimize the constants.*

REMARK 5.4. *Thus, up to the compatibility constant, and taking λ of order $\sqrt{\log p/n}$, the prediction error is of order $s_0 \log p/n$:*

$$\|\hat{f} - f^0\|_n^2 = \mathcal{O}\left(\frac{s_0 \log p}{n}\right).$$

An oracle that knows S_0 and does empirical risk minimization without penalty but with the restriction that all coefficients not in S_0 are set to zero, has a prediction error of order s_0/n . We see that for not knowing S_0 one pays a price of order $\log p$. We moreover have

$$\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}\left(s_0 \sqrt{\frac{\log p}{n}}\right).$$

REMARK 5.5. *We have presented the above oracle inequality involving the sparsity of the true β^0 . If the truth is not sparse, or if actually the generalized linear model is misspecified, one may replace the truth by a sparse linear approximation of the truth, and the oracle inequality involves a trade-off between the approximation error on the one hand, and the sparsity and compatibility constant on the other. This trade-off is of the following form. Let for an arbitrary index set $S \subset \{1, \dots, p\}$,*

$$f_S := \arg \min_{f=f_{\beta_S}} \bar{B}_n(G \circ f, \mu_0),$$

where $\bar{B}(G \circ f, \mu_0)$ is the average regret

$$\bar{B}(G \circ f, \mu_0) := \frac{1}{n} \sum_{i=1}^n B(G \circ f(x_i), \mu_0(x_i)).$$

Thus, f_S is the best approximation of f^0 using only the variables in S . Then under some regularity conditions the prediction error of $\bar{B}(G \circ f_{\hat{\beta}}, \mu_0)$ is with probability $(1 - \alpha)$ bounded by

$$\text{const.} \min_{\text{sets } S} \left\{ \bar{B}(G \circ f_S, \mu_0) + \frac{\lambda^2 |S|}{\phi^2(L, S)} \right\}.$$

The “const.” depends on the constants occurring in the regularity conditions, the constant L depends moreover on the choice of λ , and the confidence level α depends on all these. For more details on this extension, we refer to Bühlmann and van de Geer [2011] and the references therein.

REMARK 5.6. *Condition (5.3) assumes that the sparsity index s_0 is sufficiently smaller than $\sqrt{n/\log p}$, a condition we already announced in Subsection 5.1. This assumption plays its part in all our results: it will also be important for variable selection and simplifies the derivation of results for the case of random design. In the case of least squares loss, the assumption can be avoided, even in some cases with random design. It should however be noted that a large s_0 means a slow rate. In particular, when the sparsity is of larger order than $\sqrt{n/\log p}$, the bound for the prediction error is of larger order than $\sqrt{\log p/n}$, and this cannot*

be improved up to the $\log p$ -term. Thus, then the bounds are actually quite large in order of magnitude. Indeed, recall that the prediction error is $\|f_{\hat{\beta}} - f_{\beta^0}\|_n^2$, which is the squared distance between $f_{\hat{\beta}}$ and f_{β^0} . Assumption (5.3) allows to conclude that $\|\hat{\beta} - \beta^0\|_1 \leq 1$, and hence, that $|f_{\hat{\beta}}(x_i)| \leq K_X + K_0$ for all i . The latter was used because we only want to require Conditions A3 and A4 for bounded values of the argument z . When dealing with least squares loss, Conditions A3 and A4 hold for all $z \in \mathbb{R}$. This means that with least squares loss, Assumption (5.3) can be dropped in Theorem 5.2 (see Theorem 5.1).

REMARK 5.7. The lower bound in (5.4) for the tuning parameter λ depends on the noise level σ as well as other unknown constants. In practice, one may for instance apply cross-validation. The noise level σ can also be treated as additional parameter which can be estimated along with β^0 . See Städler and van de Geer [2010] for a discussion.

6. ORACLE INEQUALITIES FOR ROBUST LOSS

In this section, we assume throughout that ρ is robust loss, see Definition 2.2.

We define for $i = 1, \dots, n$,

$$l_i(z) = \mathbb{E}\rho(Y_i, z|x_i), \quad z \in \mathbb{R},$$

and assume that $\ddot{l}_i(z) := d^2l_i(z)/dz^2$ exists.

Condition B For K_X and K_0 given in Conditions A1 and A2 respectively, we have for some constant C_l and for all i ,

$$\inf_{|z| \leq K_X + K_0} \ddot{l}_i(z) \geq 2/C_l.$$

EXAMPLE 6.1. The least absolute deviations loss is $\rho(y, z) := |y - z|$. Let G_i be distribution function of Y_i given x_i ($i = 1, \dots, n$). Then f_i^0 is the median of G_i and Condition B requires that G_i has a strictly positive density g_i on $\{|z| \leq K_X + K_0\}$ for all i .

We now define

$$\Gamma(S_0) := 16C_l \left[\frac{s_0}{\phi^2(3, S_0)} \right].$$

Fix some $t > 0$ and define

$$\lambda_\epsilon(t) := 16K_X \sqrt{\frac{2(t + \log p)}{n}}.$$

The following theorem is a reformulation of results in van de Geer [2007], van de Geer [2007] or Bühlmann and van de Geer [2011].

THEOREM 6.1. *Let $\hat{\beta}$ be the ℓ_1 -penalized robust estimator. Assume Conditions A1, A2 and B. Suppose that*

$$(6.1) \quad \lambda_\epsilon(t)\Gamma_0(S_0) \leq \frac{1}{4}.$$

Take

$$4\lambda_\epsilon(t) \leq \lambda \leq \frac{1}{\Gamma(S_0)}.$$

With probability at least $1 - \alpha(t)$, where $\alpha(t) := 3 \exp[-t]$, it holds that

$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{\lambda}{2}\Gamma(S_0),$$

and

$$\|\hat{f}_\beta - f_{\beta^0}\|_n^2 \leq \frac{3}{4}C_l\lambda^2\Gamma(S_0).$$

REMARK 6.1. *Similar remarks can be made as for the ℓ_1 -penalized quasi-likelihood estimator. The new element in the result is that with robustness the tuning parameter λ does not depend on some noise level σ .*

7. VARIABLE SELECTION WITH QUASI-LIKELIHOOD LOSS

Note that the bounds for the ℓ_1 -error $\|\hat{\beta} - \beta^0\|_1$, given in Theorems 5.2 and 6.1, can be invoked to show that, with large probability, the ℓ_1 -regularized estimator will detect most of the non-zero coefficients β^0 which are large enough: for all $\eta > 0$,

$$\#\{\hat{\beta}_j \neq 0, |\beta_j^0| \geq \lambda/\eta\} \geq \#\{|\beta_j^0| \geq \lambda/\eta\} - \eta\|\hat{\beta} - \beta^0\|_1/\lambda.$$

In other words, if a large proportion of the non-zero coefficients is sufficiently far above the noise level in absolute value, then there will also be many true positives. By this argument, if all non-zero coefficients of β^0 are of larger order than $\lambda\Gamma(S_0)$, we will have $\hat{S} \supset S_0$, where

$$\hat{S} := \{j : \hat{\beta}_j \neq 0\}.$$

This section will study the false positives. We show that for the case of quasi-likelihood loss, an irrepresentable condition similar to Meinshausen and Bühlmann [2006] and Zhao and Yu [2006] implies that there are no false positives, i.e., that $\hat{S} \subset S_0$. Such result can also be obtained for robust loss, but is omitted here.

7.1 The case of least squares loss

Again, as preparation, let us first consider the standard linear model and the least squares Lasso estimator

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_n^2 + \lambda\|\beta\|_1 \right\}.$$

Let $\mathbf{X}(S) := (\mathbf{X}_j)_{j \in S}$ be the design matrix consisting of the variables in S , and let

$$\hat{\Sigma}_{1,1}(S) := \mathbf{X}^T(S)\mathbf{X}(S)/n, \quad \hat{\Sigma}_{1,2}(S) := \mathbf{X}^T(S^c)\mathbf{X}(S)/n.$$

In Bühlmann and van de Geer [2011] (Exercise 7.5) or van de Geer et al. [2011], one can find the following result.

THEOREM 7.1. *Suppose that $\lambda > \lambda_0$ where $\lambda_0 \geq 2 \max_{1 \leq j \leq p} |\epsilon^T \mathbf{X}_j|/n$. Assume moreover the irrerepresentable condition*

$$\sup_{\|\tau_{S_0}\|_\infty \leq 1} \|\hat{\Sigma}_{2,1}(S_0)\hat{\Sigma}_{1,1}^{-1}(S_0)\tau_{S_0}\|_\infty < \frac{\lambda - \lambda_0}{\lambda + \lambda_0}.$$

Then $\hat{S} \subset S_0$.

We remark that an irrerepresentable condition (see also below in Definition 7.1) is always rather strong. However, for exact variable selection, an irrerepresentable condition is essentially necessary, as shown in Meinshausen and Bühlmann [2006], Zhao and Yu [2006], Bühlmann and van de Geer [2011]. By thresholding the estimated coefficients and refitting, or by applying the adaptive Lasso, one can often improve on variable selection and yet maintain a good prediction and estimation error. The conditions for the latter are much less restrictive than the irrerepresentable condition. We refer to van de Geer et al. [2011] for details.

7.2 General quasi-likelihood loss

The results are based on the Karush-Kuhn-Tucker (or KKT-)conditions, see Bertsimas and Tsitsiklis [1997]. In our context, they read as follows:

KKT conditions *We have*

$$\frac{\partial}{\partial \beta} \frac{1}{n} \sum_{i=1}^n Q(Y_i, x_i^T \beta) \Big|_{\beta=\hat{\beta}} = -\lambda \hat{\tau}.$$

Here $\|\hat{\tau}\|_\infty \leq 1$, and moreover

$$\hat{\tau}_j \mathbf{1}\{\hat{\beta}_j \neq 0\} = \text{sign}(\hat{\beta}_j), \quad j = 1, \dots, p.$$

Let

$$\hat{\Sigma}_{j,k} := \frac{1}{n} \sum_{i=1}^n x_{i,j} x_{i,k} w_i^2,$$

where

$$w_i^2 := h^2(x_i^T \beta^0) V \circ G(x_i^T \beta^0), \quad i = 1, \dots, n.$$

Thus, $\hat{\Sigma}$ is the weighted Gram matrix

$$\hat{\Sigma} = \mathbf{X}^T W^2 \mathbf{X} / n, \quad W^2 := \text{diag}(w_1^2, \dots, w_n^2).$$

We write $\mathbf{X}_W := W\mathbf{X}$, so that $\hat{\Sigma} = \mathbf{X}_W^T \mathbf{X}_W / n$.

Let $\mathbf{X}_W(S)$ be the weighted design matrix consisting of the variables in S , and

$$\hat{\Sigma}_{1,1}(S) := \mathbf{X}_W^T(S) \mathbf{X}_W(S) / n, \quad \hat{\Sigma}_{2,1}(S) := \mathbf{X}_W^T(S^c) \mathbf{X}_W(S) / n.$$

DEFINITION 7.1. *Let $0 < \theta \leq 1$ be given. We say that the θ -irrepresentable condition is met for the set S if*

$$\max_{\|\tau_S\|_\infty \leq 1} \|\hat{\Sigma}_{2,1}(S) \hat{\Sigma}_{1,1}^{-1}(S) \tau_S\|_\infty \leq \theta.$$

Here is how the θ -irrepresentable condition can be linked with variable selection.

THEOREM 7.2. *Let $0 \leq \lambda_0 < \lambda$. Suppose that*

$$(7.1) \quad \hat{\Sigma}(\hat{\beta} - \beta^0) = -v,$$

where $|v_j| \leq \lambda + \lambda_0$, and $v_j \hat{\beta}_j \geq (\lambda - \lambda_0) |\hat{\beta}_j|$, $j = 1, \dots, p$. Suppose moreover the θ -irrepresentable condition is met for S_0 , with $\theta < (\lambda - \lambda_0) / (\lambda + \lambda_0)$. Then $\hat{S} \subset S_0$.

In the proof of Theorem 7.3 below, we show that the equation (7.1) in Theorem 7.2 holds for some v satisfying the conditions of this theorem. This allows us then to conclude that $\hat{S} \subset S_0$.

As one sees in the KKT conditions, the derivative at $\hat{\beta}$ of the loss function occurs. We will need to compare this by the derivative at β^0 . To bring this to an end we need, in addition to Conditions A3 and A4, certain Lipschitz conditions on h and g .

Condition A5 *For K_X and K_0 given in Conditions A1 and A2 respectively, we have for all $|z_0| \leq |z| \leq K_X + K_0$, and some constant L_h ,*

$$|h(z) - h(z_0)| \leq L_h |z - z_0|.$$

Condition A6 *For K_X and K_0 given in Conditions A1 and A2 respectively, we have for all $|z_0| \leq |z| \leq K_X + K_0$, and some constant L_g ,*

$$|g(z) - g(z_0)| \leq L_g |z - z_0| / 2.$$

REMARK 7.1. *Under the additional Conditions A5 and A6, one can improve the constants in Theorem 5.2. It is also clear that Conditions A5 and A6 hold for least squares and logistic loss.*

With these new constants, we define

$$L_{h,V} := (L_g + L_h C_V) C_h, \quad L_{h,X} + 16L_h K_X^2.$$

We moreover let

$$\Gamma_\epsilon := \Gamma(S_0) := 16C_{h,V} \Gamma_{\text{effective}}(S_0),$$

and

$$\Gamma_0 := \Gamma_0(S_0) := 6L_{h,V} C_{h,V}^2 \Gamma_{\text{effective}}(S_0).$$

Fix some $t > 0$ and define

$$\lambda_\epsilon(t) := C_{h,X} \sigma \sqrt{\frac{2(t + \log p)}{n}},$$

and

$$\lambda_0(t) := L_{h,X} \sigma \sqrt{\frac{2(t + 2 \log p)}{n}}.$$

Define

$$\alpha(t) := 9 \exp[-t] + 9\kappa^4 / (n\sigma^4).$$

Thus, up to constants, Γ_ϵ and Γ_0 are the effective sparsity. Moreover, for $t \asymp \log n$ (say), $\lambda_\epsilon(t) \asymp \lambda_0(t) \asymp \sqrt{\log(p \vee n)/n}$ and $\alpha(t) \asymp 1/n$.

We arrive at the main result of this section.

THEOREM 7.3. *Let $\hat{\beta}$ be the ℓ_1 -penalized quasi-likelihood estimator. Assume Conditions A_ϵ and A1-A6. Assume that (5.3) holds, i.e.,*

$$\lambda_\epsilon(t) \Gamma_\epsilon \leq \gamma_1 \leq \frac{1}{4}.$$

where γ_1 is given by

$$\gamma_1 := \frac{\lambda_\epsilon(t)}{\lambda}.$$

Assume now that

$$(7.2) \quad \lambda_\epsilon(t) \Gamma_0 \leq \gamma_1 \gamma_\epsilon \text{ for some } \gamma_\epsilon < 1 - \gamma_1,$$

as well as

$$(7.3) \quad \lambda_0(t) \Gamma_\epsilon \leq \gamma_0 \text{ for some } \gamma_0 < 1 - \gamma_\epsilon - \gamma_1.$$

Assume furthermore the θ -irrepresentable condition with

$$\theta < \frac{1 - \gamma}{1 + \gamma}, \quad \gamma := \gamma_\epsilon + \gamma_0 + \gamma_1.$$

With probability at least $1 - \alpha(t)$, it holds that $\hat{S} \subset S_0$.

REMARK 7.2. *Let us take $\lambda_\epsilon(t) \asymp \lambda_0(t) \asymp \lambda \asymp \sqrt{\log p/n}$. The constants γ_0 , γ_1 and γ_ϵ are small, depending on the constants appearing in Conditions A_ϵ and A1-A6. Fixing these, they can be kept away from zero, and hence also the θ -irrepresentable condition is assumed for a value of θ that stays away from zero. Conditions 7.2, and 7.2 again require that the effective sparsity is sufficiently smaller than $\sqrt{\log p/n}$. Formulated differently, the results of Theorems 7.3 and 5.2 imply that if the θ -irrepresentable condition holds and if $\Gamma_{\text{effective}}(S_0) \leq \gamma\sqrt{\log p/n}$ for sufficiently small values of θ and γ (depending only on the constants appearing in Conditions A_ϵ and A1-A6) then with an appropriate choice of $\lambda \asymp \sqrt{\log p/n}$ the Lasso estimator has with large probability prediction error $\Gamma_{\text{effective}}(S_0) \log p/n$, ℓ_1 -error $\Gamma_{\text{effective}}(S_0)\sqrt{\log p/n}$ and no false positives.*

8. RANDOM DESIGN

Consider quasi-likelihood loss. It is easy to see that under the conditions of Theorem 5.2, one has with large probability

$$(\hat{\beta} - \beta^0)^T \hat{\Sigma} (\hat{\beta} - \beta^0) \leq 6C_{h,V}^3 \lambda^2 \Gamma_{\text{effective}}(S_0).$$

This follows from $w_i^2 \leq C_{h,V}/2$, where as in Section 7, $w_i^2 = h^2(x_i^T \beta^0)V \circ G(x_i^T \beta^0)$, $i = 1, \dots, n$. Let Σ be some other $p \times p$ positive semi-definite matrix. Then

$$\|(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0)\|_\infty \leq \lambda_X \|\hat{\beta} - \beta^0\|_1,$$

where

$$\lambda_X := \max_{j,k} |\hat{\Sigma}_{j,k} - \Sigma_{j,k}|.$$

Thus, under the conditions of Theorem 5.2, one has that with large probability

$$\|(\hat{\Sigma} - \Sigma)(\hat{\beta} - \beta^0)\|_\infty \leq \lambda \lambda_X \Gamma(S_0)/2.$$

One can verify that if $\lambda_X \Gamma(S_0)$ is small enough, say for some γ_X sufficiently small

$$(8.1) \quad \lambda_X \Gamma(S_0) \leq \gamma_X,$$

then one may reformulate the compatibility condition replacing $\|f_\beta\|_n^2$ by $\beta^T \Sigma \beta$, and the theory for prediction and ℓ_1 -error goes through essentially without new arguments. One can then also establish bounds for $(\hat{\beta} - \beta^0)^T \hat{\Sigma} (\hat{\beta} - \beta^0)$. Similarly, one may reformulate the θ -irrepresentable condition with $\hat{\Sigma}$ replaced by Σ , and obtain variable selection without needing new arguments. In the case where Σ is the population version of $\hat{\Sigma}$, the latter built from an i.i.d. sample of covariables, one can show that with large probability λ_X is of order $\sqrt{\log p/n}$. In other words (and modulo the compatibility constant), then condition (8.1) is another instance where it is required that the sparsity s_0 is not of larger order than $\sqrt{n/\log p}$. We refer to Bühlmann and van de Geer [2011] for more precise statements.

9. CONCLUSION

The results of this paper show that the oracle and variable selection properties of the Lasso for the linear model also hold for the generalized linear model. We prove this under the assumption that the sparsity is sufficiently smaller than $\sqrt{n/\log p}$. We note that the results rely heavily on the convexity of the loss function. This allows one to work with an unbounded parameter space. If the estimators are a priori restricted to lie in a given bounded set, one can extend the results to non-convex loss (see Städler and van de Geer [2010] for the mixture model, and Schelldorfer et al. [2011] for the mixed effects model) and one can moreover prove oracle results for the almost linear in s_0 regime of sparsity.

Supplementary Material

Supplementary material for “Quasi-likelihood and/or robust estimation in high dimensions”

(<http://lib.stat.cmu.edu/sts/???/???>). Due to space constraints, the proofs and technical details have been given in the supplementary document van de Geer and Müller [2012].

REFERENCES

- D. Bertsimas and J.N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific Belmont, MA, 1997.
- P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.
- P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.
- F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation and sparsity via ℓ_1 -penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory, COLT 2006. Lecture Notes in Artificial Intelligence*, pages 379–391. Springer Verlag, 2006.
- F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674, 2007a.
- F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparse density Estimation with ℓ_1 penalties. In *Proceedings of 20th Annual Conference on Learning Theory, COLT 2007. Lecture Notes in Artificial Intelligence*, pages 530–544. Springer, 2007b.
- F. Bunea, A. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007c.
- E. Candès, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *Journal of the Association for Computing Machinery*, 58:1–37, 2009.
- D.L. Donoho. De-noising via soft-thresholding. *IEEE Transactions on Information Theory*, 41:613–627, 1995.
- J. Fan. Comments on Wavelets in statistics: A review, by A. Antoniadis. *Journal of the American Statistical Association*, 6:131–138, 1997.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 2010.
- M. Hein and T. Buehler. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In *Advances in Neural Information Processing Systems, NIPS 2010*, volume 23, pages 847–855, 2010.
- A. Juditsky and A. Nemirovski. Accuracy guarantees for ℓ_1 -recovery. *IEEE Transactions on Information Theory*, 2011. to appear.

- V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45:7–57, 2009a.
- V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15:799–828, 2009b.
- S. Lambert-Lacroix and L. Zwald. Robust regression through the Hubers criterion and adaptive Lasso penalty. *Electronic Journal of Statistics*, 5:1015–1053, 2011.
- M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Verlag, New York, 1991.
- J.-M. Loubes and S. van de Geer. Adaptive estimation in regression, using soft thresholding type penalties. *Statistica Neerlandica*, 56:453–478, 2002.
- K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.
- P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28:863–884, 2000.
- P. McCullagh and J.A. Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, 1989.
- N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.
- S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 2011. In this volume.
- J. Schelldorfer, P. Bühlmann, and S. van de Geer. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.
- P. Städler, N. P. Bühlmann and S. van de Geer. L1-penalization in mixture regression models (with discussion). *Test*, 19:209–285, 2010.
- B. Tarigan and S.A. van de Geer. Classifiers of support vector machine type with ℓ_1 complexity regularization. *Bernoulli*, 12:1045–1076, 2006.
- R. Tibshirani. Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.
- S. van de Geer. Least squares estimation with complexity penalties. *Mathematical Methods of Statistics*, 10:355–374, 2001.
- S. van de Geer. Adaptive quantile regression. In *Recent Trends in Nonparametric Statistics*, pages 235–250. Elsevier Science, 2003. Eds. M.G. Akritas and D.N. Politis.
- S. van de Geer and J. Lederer. The Lasso, correlated design, and improved oracle inequalities. In *IMS Collections: A Festschrift in Honor of Jon Wellner*. IMS, 2012. To appear.
- S. van de Geer and P. Müller. Supplementary material for “Quasi-likelihood and/or robust estimation in high dimensions”, 2012.
- S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5: 688–749, 2011.
- S.A. van de Geer. The deterministic Lasso. In *JSM proceedings, 2007, 140*. American Statistical Association, 2007.
- S.A. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645, 2008.
- S.A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
- H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics*, 25(3):347–355, 2007.
- T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, and K. Lange. Genomewide association analysis by Lasso penalized logistic regression. *Bioinformatics*, 25:714–721, 2009.
- Y. Wu and Y. Liu. Variable selection in quantile regression. *Statistica Sinica*, 19:801–817, 2009.
- P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.