Model Selection and Minimax Estimation in Generalized Linear Models

Felix Abramovich and Vadim Grinshtein

***Abstract*— We consider model selection in generalized linear models (GLM) for high-dimensional data and propose a wide class of model selection criteria based on penalized maximum likelihood with a complexity penalty on the model size. We derive a general nonasymptotic upper bound for the Kullback–Leibler risk of the resulting estimators and establish the corresponding minimax lower bounds for the sparse GLM. For the properly chosen (nonlinear) penalty, the resulting penalized maximum likelihood estimator is shown to be asymptotically minimax and adaptive to the unknown sparsity. We also discuss possible extensions of the proposed approach to model selection in the GLM under additional structural constraints and aggregation.**

***Index Terms*— Complexity penalty, generalized linear models, Kullback-Leibler risk, minimax estimator, model selection, sparsity.**

I. INTRODUCTION

REGRESSION analysis of high-dimensional data, where the number of potential explanatory variables (predictors) p might be even large relative to the sample size n faces a severe "curse of dimensionality" problem. Reducing the dimensionality of the model by selecting a sparse subset of "significant" predictors becomes therefore crucial. The interest to model selection in regression goes back to seventies (e.g., seminal papers [4], [15], [21]), where the considered "classical" setup assumed $p \ll n$. Its renaissance started in 2000s with the new challenges brought to the door of statistics by exploring data, where p is of the order of n or even larger. Analysing the "p larger than n" setup required novel approaches and techniques, and led to novel model selection procedures. The corresponding theory (risk bounds, oracle inequalities, minimax rates, variable selection consistency, etc.) for model selection in Gaussian linear regression has been intensively developed in the literature in the last decade. See [1], [5]–[7], [9], [12], [18], and [20] among many others. A review on model selection in Gaussian regression for "p larger than n" setup can be found in [25].

Manuscript received July 14, 2015; revised January 31, 2016; accepted March 29, 2016. Date of publication April 21, 2016; date of current version May 18, 2016. This work was supported by the Israel Science Foundation under Grant ISF-820/13.

F. Abramovich is with the Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel (e-mail: felix@post.tau.ac.il).

V. Grinshtein is with the Department of Mathematics and Computer Science, The Open University of Israel, Ra'anana 4353701, Israel (e-mail: vadimg@openu.ac.il).

Communicated by I. Nikiforov, Associate Editor for Detection and Estimation.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Digital Object Identifier 10.1109/TIT.2016.2555812

Generalized linear models (GLM) is a generalization of Gaussian linear regression, where the distribution of response is not necessarily normal but belongs to the natural exponential family of distributions. Important examples include binomial and Poisson data arising in a variety of statistical applications. Foundations of a general theory for GLM have been developed in [16].

Although most of the proposed model selection criteria for Gaussian regression have been extended and are nowadays widely used in GLM (e.g., AIC of Akaike [4] and BIC of Schwarz [21]), not much is known on their theoretical properties in the general GLM setup. There are some results on variable selection consistency of several model selection criteria (e.g., [10], [11]), but a rigorous theory of model selection for estimation and prediction in GLM remains essentially *terra incognita*. We can mention [26] that investigated the Lasso estimator in GLM and [19] that considered aggregation problem for GLM. The presented paper intends to contribute to fill this gap.

We introduce a wide class of model selection criteria for GLM based on the penalized maximum likelihood estimation with a complexity penalty on the model size. In particular, it includes AIC, BIC and some other well-known criteria. In a way, this approach can be viewed as an extension of that of [6] and [7] for Gaussian regression. We derive a general nonasymptotic upper bound for the Kullback-Leibler risk of the resulting estimator. Furthermore, for the properly chosen penalty we establish its asymptotic minimaxity and adaptiveness to the unknown sparsity. Possible extensions to model selection under additional structural constraints and aggregation are also discussed.

The paper is organized as follows. The penalized maximum likelihood model selection procedure for GLM is introduced in Section II. Its main theoretical properties are presented in Section III. In particular, we derive the upper bound for its Kullback-Leibler risk and corresponding minimax lower bounds, and establish its asymptotic minimaxity over various sparse settings. We illustrate the obtained general results on the example of logistic regression. Extensions to model selection under structural constraints and aggregation are discussed in Section IV. All the proofs are given in the Appendix.

II. MODEL SELECTION PROCEDURE FOR GLM

A. Notation and Preliminaries

Consider a GLM setup with a response variable Y and a set of p predictors $x_1, ..., x_p$. We observe a series of independent observations $(\mathbf{x}_i, Y_i)$, $i = 1, \dots, n$, where the design points $\mathbf{x}_i \in \mathbb{R}^p$ are deterministic, and the distribution $f_{\theta_i}(y)$ of Y_i belongs to a (one-parameter) natural exponential family with a natural parameter θ_i and a scaling parameter a:

$$f_{\theta_i}(y) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a} + c(y,a)\right\} \tag{1}$$

The function $b(\cdot)$ is assumed to be twice-differentiable. In this case $\mathbb{E}(Y_i) = b'(\theta_i)$ and $Var(Y_i) = ab''(\theta_i)$ (see [16]). To complete GLM we assume the canonical link $\theta_i = \boldsymbol{\beta}^t \mathbf{x}_i$ or, equivalently, in the matrix form, $\boldsymbol{\theta} = X\boldsymbol{\beta}$, where $X_{n\times p}$ is the design matrix and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of the unknown regression coefficients.

In what follows we assume the following assumption on the parameter space Θ and the second derivative $b''(\cdot)$: *Assumption (***A***):*

1) *Assume that $\theta_i \in \Theta$, where the parameter space $\Theta \subseteq \mathbb{R}$ is a closed (finite or infinite) interval.*
2) *Assume that there exist constants $0 < \mathcal{L} \le \mathcal{U} < \infty$ such that the function $b''(\cdot)$ satisfies the following conditions:*
   a) $\sup_{t\in\mathbb{R}} b''(t) \le \mathcal{U}$
   b) $\inf_{t\in\Theta} b''(t) \ge \mathcal{L}$

Similar assumptions were imposed in [19] and [26]. Conditions on $b''(\cdot)$ in Assumption (A) are intended to exclude two degenerate cases, where the variance $Var(Y)$ is infinitely large or small. They also ensure strong convexity of $b(\cdot)$ over Θ. For Gaussian distribution, $b''(\theta) = 1$ and, therefore, $\mathcal{L} = \mathcal{U} = 1$ for any Θ. For the binomial distribution, $b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2}$, $\mathcal{U} = \frac{1}{4}$, while the condition (b) is equivalent to the boundedness of $\theta : \Theta = \{\theta : |\theta| \le C_0\}$, where $\mathcal{L} = \frac{e^{C_0}}{(1+e^{C_0})^2}$.

Let $f_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\zeta}}$ be two possible joint distributions of the data from the natural exponential family with n-dimensional vectors of natural parameters $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ correspondingly. A Kullback-Leibler divergence $KL(\boldsymbol{\theta}, \boldsymbol{\zeta})$ between $f_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\zeta}}$ is then

$$\begin{aligned} KL(\boldsymbol{\theta}, \boldsymbol{\zeta}) &= \mathbb{E}_{\boldsymbol{\theta}}\left\{\ln\left(\frac{f_{\boldsymbol{\theta}}(\mathbf{Y})}{f_{\boldsymbol{\zeta}}(\mathbf{Y})}\right)\right\} \\ &= \frac{1}{a}\ \mathbb{E}_{\boldsymbol{\theta}}\left\{\sum_{i=1}^n Y_i(\theta_i - \zeta_i) - b(\theta_i) + b(\zeta_i)\right\} \\ &= \frac{1}{a}\sum_{i=1}^n \left\{b'(\theta_i)(\theta_i - \zeta_i) - b(\theta_i) + b(\zeta_i)\right\} \\ &= \frac{1}{a}\left(b'(\boldsymbol{\theta})^t(\boldsymbol{\theta} - \boldsymbol{\zeta}) - (b(\boldsymbol{\theta}) - b(\boldsymbol{\zeta}))^t \mathbf{1}\right), \end{aligned}$$

where $b(\boldsymbol{\theta}) = (b(\theta_1), \cdots, b(\theta_n))$ and $b(\boldsymbol{\zeta}) = (b(\zeta_1), \dots, b(\zeta_n))$.

For a given estimator $\widehat{\boldsymbol{\theta}}$ of the unknown $\boldsymbol{\theta}$ consider the Kullback-Leibler loss $KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$ – the Kullback-Leibler divergence between the true distribution $f_{\boldsymbol{\theta}}$ of the data and its empirical distribution $f_{\widehat{\boldsymbol{\theta}}}$ generated by $\widehat{\boldsymbol{\theta}}$. The goodness of $\widehat{\boldsymbol{\theta}}$ is measured by the corresponding Kullback-Leibler risk:

$$\mathbb{E}KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{1}{a}\left(b'(\boldsymbol{\theta})^t(\boldsymbol{\theta} - \mathbb{E}(\widehat{\boldsymbol{\theta}})) - (b(\boldsymbol{\theta}) - \mathbb{E}b(\widehat{\boldsymbol{\theta}}))^t\mathbf{1}\right) \tag{2}$$

where the expectation is taken w.r.t. the true distribution $f_{\boldsymbol{\theta}}$. In particular, for the Gaussian case, where $b(\theta) = \theta^2/2$ and $a = \sigma^2$, $\mathbb{E}KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$ is the mean squared error $E||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ divided by the constant $2\sigma^2$. The binomial distribution will be considered in more details in Section III-D below.

B. Penalized Maximum Likelihood Model Selection

Consider a GLM (1) with a vector of natural parameters $\boldsymbol{\theta}$ and the canonical link $\boldsymbol{\theta} = X\boldsymbol{\beta}$.

For a given model $M \subset \{1, \dots, p\}$ consider the corresponding maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_M &= \arg\max_{\widetilde{\boldsymbol{\beta}}\in\mathcal{B}_M} \ell(\widetilde{\boldsymbol{\beta}}) = \arg\max_{\widetilde{\boldsymbol{\beta}}\in\mathcal{B}_M}\left\{\sum_{i=1}^n (Y_i(X\widetilde{\beta})_i - b((X\widetilde{\beta})_i)\right\} \\ &= \arg\max_{\widetilde{\boldsymbol{\beta}}\in\mathcal{B}_M}\left\{\mathbf{Y}^t X\widetilde{\boldsymbol{\beta}} - b(X\widetilde{\boldsymbol{\beta}})^t\mathbf{1}\right\}, \end{aligned} \tag{3}$$

where $\mathcal{B}_M = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \notin M \text{ and } \boldsymbol{\beta}^t\mathbf{x}_i \in \Theta$ for all $i = 1, \dots, n\}$. Note that generally $\mathcal{B}_M$ depends on a given design matrix X. Except Gaussian regression, the MLE $\widehat{\boldsymbol{\beta}}_M$ in (3) is not available in the closed form but can be obtained numerically by the iteratively reweighted least squares algorithm (see [16, Sec. 2.5]).

The MLE for $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}}_M = X\widehat{\boldsymbol{\beta}}_M$, and the ideally selected model (oracle choice) is then the one that minimizes $\mathbb{E}KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_M) = \frac{1}{a}\left(b'(\boldsymbol{\theta})^t(\boldsymbol{\theta} - \mathbb{E}(\widehat{\boldsymbol{\theta}}_M)) - (b(\boldsymbol{\theta}) - \mathbb{E}b(\widehat{\boldsymbol{\theta}}_M))^t\mathbf{1}\right)$ or, equivalently, $-b'(\boldsymbol{\theta})^t\mathbb{E}(\widehat{\boldsymbol{\theta}}_M) + \mathbb{E}b(\widehat{\boldsymbol{\theta}}_M)^t\mathbf{1}$ over M from the set of all 2^p possible models $\mathfrak{M}$. An oracle chosen model depends evidently on the unknown $\boldsymbol{\theta}$ and can only be used as a benchmark for any available model selection procedure.

Consider instead an empirical analog $KL([b']^{-1}(\mathbf{Y}), \widehat{\boldsymbol{\theta}}_M)$ of $\mathbb{E}KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_M)$, where the true $\mathbb{E}\mathbf{Y} = b'(\boldsymbol{\theta})$ is replaced by $\mathbf{Y}$. A naive approach of minimizing $KL([b']^{-1}(\mathbf{Y}), \widehat{\boldsymbol{\theta}}_M)$ yields maximizing $\mathbf{Y}^t\widehat{\boldsymbol{\theta}}_M - b(\widehat{\boldsymbol{\theta}}_M)^t\mathbf{1}$ (or, equivalently, maximizing $\ell(\widehat{\boldsymbol{\beta}}_M)$) over $M \in \mathfrak{M}$ and obviously leads to the saturated model.

A common remedy to avoid such a trivial unsatisfactory choice is to add a complexity penalty $Pen(|M|)$ on the model size $|M|$ and consider the corresponding *penalized* maximum likelihood model selection criterion of the form

$$\begin{aligned} \widehat{M} &= \arg\max_{M\in\mathfrak{M}}\left\{\ell(\widehat{\boldsymbol{\beta}}_M) - Pen(|M|)\right\} \\ &= \arg\min_{M\in\mathfrak{M}}\left\{\frac{1}{a}\left(b(X\widehat{\boldsymbol{\beta}}_M)^t\mathbf{1} - \mathbf{Y}^tX\widehat{\boldsymbol{\beta}}_M\right) + Pen(|M|)\right\}, \end{aligned} \tag{4}$$

where the MLE $\widehat{\boldsymbol{\beta}}_M$ for a given model M are given in (3). The properties of the resulting model selection procedure depends crucially on the choice of the complexity penalty. The commonly used criteria for model selection in GLM are $AIC = -2\ell(\widehat{\boldsymbol{\beta}}_M) + 2|M|$ of [4], $BIC = -2\ell(\widehat{\boldsymbol{\beta}}_M) + |M|\ln n$ of [21] and its extended version $EBIC = -2\ell(\widehat{\boldsymbol{\beta}}_M) + |M|\ln n + 2\gamma|M|\ln p$, $0 \le \gamma \le 1$ of [10] correspond to $Pen(|M|) = |M|$, $Pen(|M|) = \frac{|M|}{2}\ln n$ and $Pen(|M|) = \frac{|M|}{2}\ln n + \gamma|M|\ln p$ in (4) respectively. A similar extension of RIC criterion $RIC = -2\ell(\widehat{\boldsymbol{\beta}}_M) + 2|M|\ln p$ of [12] yields

$Pen(|M|) = |M| \ln p$. Note that all the above penalties increase *linear*ly with a model size $|M|$.

III. Main Results

In this section we investigate theoretical properties of the penalized maximum likelihood model selection procedure proposed in Section II-B and discuss the optimal choice for the complexity penalty $Pen(|M|)$ in (4). We start from deriving a (nonasymptotic) upper bound for the expected Kullback-Leibler risk of the resulting estimator for a given $Pen(|M|)$ and then establish its asymptotic minimaxity for a properly chosen penalty. To illustrate the general results we consider the example of logistic regression.

A. General Upper Bound for the Kullback-Leibler Risk

Consider a GLM (1) with the canonical link $\boldsymbol{\theta} = X\boldsymbol{\beta}$ and the natural parameters $\theta_i \in \Theta$ satisfying Assumption (A). Let $r = rank(X)$. The number of possible predictors p might be larger than the sample size n. We assume that any r columns of X are linearly independent and consider only models of sizes at most r in (4) since otherwise, for any $\boldsymbol{\beta} \in \mathcal{B}_M$, where $|M| > r$, there necessarily exists another $\boldsymbol{\beta}^*$ with at most r nonzero entries such that $X\boldsymbol{\beta} = X\boldsymbol{\beta}^*$.

We now present an upper bound for the Kullback-Leibler risk of the proposed maximum penalized likelihood estimator valid for a wide class of penalties. Moreover, it does not require the GLM assumption on the canonical link $\boldsymbol{\theta} = X\boldsymbol{\beta}$ and can still be applied when a link function is misspecified.

Theorem 1: Consider a GLM (1), where $\theta_i \in \Theta$, $i = 1, \ldots, n$ and let Assumption (A) hold.

Let L_k, $k = 1, \ldots, r$ be a sequence of positive weights such that

$$\sum_{k=1}^{r-1} \binom{p}{k} e^{-kL_k} + e^{-rL_r} \le S \tag{5}$$

for some absolute constant S not depending on r, p and n.

Assume that the complexity penalty $Pen(\cdot)$ in (4) is such that

$$Pen(k) \ge 2\,\frac{\mathcal{U}}{\mathcal{L}}\,k(A + 2\sqrt{2L_k} + 4L_k), \quad k = 1, \ldots, r \tag{6}$$

for some $A > 1$.

Let $\widehat{M}$ be a model selected in (4) with $Pen(\cdot)$ satisfying (6) and $\widehat{\boldsymbol{\beta}}_{\widehat{M}}$ be the corresponding MLE estimator (3) of $\boldsymbol{\beta}$. Then,

$$\mathbb{E}KL(\boldsymbol{\theta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) \le \frac{4}{3} \inf_{M \in \mathfrak{M}} \left\{ \inf_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} KL(\boldsymbol{\theta}, X\widetilde{\boldsymbol{\beta}}) + Pen(|M|) \right\} + \frac{16}{3}\,\frac{\mathcal{U}}{\mathcal{L}}\,\frac{2A-1}{A-1}\,S \tag{7}$$

The term $\inf_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} KL(\boldsymbol{\theta}, X\widetilde{\boldsymbol{\beta}})$ in (7) can be interpreted as a Kullback-Leibler divergence between a true distribution $f_{\boldsymbol{\theta}}$ of the data and the family of distributions $\{f_{X\widetilde{\boldsymbol{\beta}}}, \widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M\}$ generated by the span of a subset of columns of X corresponding to the model M. The binomial coefficients $\binom{p}{k}$ appearing in the condition (5) for $1 \le k < r$ are the numbers of all possible models of size k. The case $k = r$ is treated slightly differently in (5). For $p = r$, there is evidently a single saturated model. For $p > r$, although there are $\binom{p}{r}$ various models of size r, all of them are nevertheless undistinguishable in terms of $X\boldsymbol{\beta}_M$ and can be still associated with a single (saturated) model.

For Gaussian regression, $\mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) = \frac{1}{2\sigma^2}\mathbb{E}||X\boldsymbol{\beta} - X\widehat{\boldsymbol{\beta}}_{\widehat{M}}||^2$, $\min_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} KL(X\boldsymbol{\beta}, X\widetilde{\boldsymbol{\beta}}) = \frac{1}{2\sigma^2}||X\boldsymbol{\beta} - X\boldsymbol{\beta}_M||^2$, where $X\boldsymbol{\beta}_M$ is the projection of $X\boldsymbol{\beta}$ on the span of columns of M, $\mathcal{L} = \mathcal{U} = 1$ and the upper bound (7) is similar (up to somewhat different constants) to those of [6] and [7]. Thus, Theorem 1 essentially extends their results for GLM.

Consider two possible choices of weights L_k and the corresponding penalties.

1. Constant weights. The simplest choice of the weights L_k's is to take them equal, i.e. $L_k = L$ for all $k = 1, \ldots, r$. The condition (5) implies then

$$\sum_{k=1}^{r-1} \binom{p}{k} e^{-kL} + e^{-rL} \le \sum_{k=1}^{p} \binom{p}{k} e^{-kL} = (1 + e^{-L})^p - 1$$

The above sum is bounded by an absolute constant for $L = \ln p$. It can be easily verified that for $L = \ln p$ and $p \ge 3$, there exists $A > 1$ such that $A + 2\sqrt{2L} + 4L \le 8L$. Thus, $2\,\frac{\mathcal{U}}{\mathcal{L}}\,k(A + 2\sqrt{2L} + 4L) \le 16\,\frac{\mathcal{U}}{\mathcal{L}}\,k \ln p$ that implies the RIC-type *linear* penalty

$$Pen(k) = C\,\frac{\mathcal{U}}{\mathcal{L}}\,k \ln p, \quad k = 1, \ldots, r \tag{8}$$

in Theorem 1 with $C \ge 16$.

Note that the AIC criterion corresponding to $Pen(k) = 2k$ (see Section II-B) does not satisfy (6).

2. Variable weights. Using the inequality $\binom{p}{k} \le \left(\frac{pe}{k}\right)^k$ (see, e.g., [3, Lemma A1]), one has

$$\sum_{k=1}^{r-1} \binom{p}{k} e^{-kL_k} + e^{-rL_r} \le \sum_{k=1}^{r-1} \left(\frac{pe}{k}\right)^k e^{-kL_k} + e^{-rL_r} = \sum_{k=1}^{r-1} e^{-k(L_k - \ln(pe/k))} + e^{-rL_r} \tag{9}$$

that suggests the choice of $L_k \sim c \ln\left(\frac{pe}{k}\right)$, $k = 1, \ldots, r-1$ and $L_r = c$ for some $c > 1$, and leads to the *nonlinear* penalty of the form $Pen(k) \sim C\,\frac{\mathcal{U}}{\mathcal{L}}\,k \ln\left(\frac{pe}{k}\right)$ for $k = 1, \ldots, r-1$ and $Pen(r) \sim C\,\frac{\mathcal{U}}{\mathcal{L}}\,r$ for some constant C.

More precisely, for any $C > 16$ there exist constants $\widetilde{C}, A > 1$ such that $C \ge 16A\widetilde{C}$. Define $L_k = \widetilde{C} \ln\left(\frac{pe}{k}\right)$, $k = 1, \ldots, r-1$ and $L_r = \widetilde{C}$. From (9) one can easily verify the condition (5) for such weights L_k. Furthermore, for $1 \le k \le r-1$ we have

$$\begin{aligned} 2\,\frac{\mathcal{U}}{\mathcal{L}}k(A + 2\sqrt{2L_k} + 4L_k) &< 2A\,\frac{\mathcal{U}}{\mathcal{L}}k\left((1 + \sqrt{2L_k})^2 + 2L_k\right) \\ &< 2A\,\frac{\mathcal{U}}{\mathcal{L}}k\left((1 + \sqrt{2})^2 L_k + 2L_k\right) \\ &\le 16A\,\frac{\mathcal{U}}{\mathcal{L}}kL_k \le C\,\frac{\mathcal{U}}{\mathcal{L}}k \ln\left(\frac{pe}{k}\right) \end{aligned}$$

and similarly, for $k = r$,

$$2\,\frac{\mathcal{U}}{\mathcal{L}}\,r(A + 2\sqrt{2L_r} + 4L_r) \le C\,\frac{\mathcal{U}}{\mathcal{L}}\,r$$

The corresponding (nonlinear) penalty in (6) is therefore

$$Pen(k) = C\ \frac{\mathcal{U}}{\mathcal{L}}\ k \ln\left(\frac{pe}{k}\right), \quad k = 1, \ldots, r-1 \tag{10}$$

and

$$Pen(r) = C\ \frac{\mathcal{U}}{\mathcal{L}}\ r,$$

where $C > 16$. For Gaussian regression such $k \ln \frac{p}{k}$-type penalties were considered in [1], [6]–[8], and [20].

The choice of $C > 16$ in (8) and (10) was mostly motivated by simplicity of calculus and it may possibly be reduced by more accurate analysis.

B. Risk Bounds for Sparse Models

Theorem 1 established a general upper bound for the Kullback-Leibler risk without any assumption on the size of a true underlying model. Analysing large data sets it is commonly assumed that only a subset of predictors has a real impact on the response. Such extra *sparsity* assumption becomes especially crucial for "p larger than n" setups. We now show that for sparse models the upper bound (7) can be improved.

For a given $1 \le p_0 \le r$, consider a set of models of size at most p_0. Obviously, $|M| \le p_0$ iff the l_0 (quasi)-norm of regression coefficients $||\boldsymbol{\beta}||_0 \le p_0$, where $||\boldsymbol{\beta}||_0$ is the number of nonzero entries. Define $\mathcal{B}(p_0) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}^t \mathbf{x}_i \in \Theta$ for all $i = 1, \ldots, n,$ and $||\boldsymbol{\beta}||_0 \le p_0\}$.

Consider a GLM with the canonical link $\boldsymbol{\theta} = X\boldsymbol{\beta}$ under Assumption (A), where $\boldsymbol{\beta} \in \mathcal{B}(p_0)$. We refine the general upper bound (7) for a penalized maximum likelihood estimator (4) with a RIC-type linear penalty (8) and a nonlinear $k \ln \frac{p}{k}$-type penalty (10) considered in Section III-A for sparse models with $\boldsymbol{\beta} \in \mathcal{B}(p_0)$.

Apply the general upper bound (7) with A corresponding to the chosen constant C in (8) and (10) (see Section III-A), and the true $\boldsymbol{\theta}^* = X\boldsymbol{\beta}^*$, $\boldsymbol{\beta}^* \in \mathcal{B}(p_0)$ in the RHS. For both penalties, we then have

$$\begin{aligned}\sup_{\boldsymbol{\beta} \in \mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) &\le \frac{4}{3}\ Pen(p_0) + \frac{16}{3}\ \frac{\mathcal{U}}{\mathcal{L}}\ \frac{2A-1}{A-1}\ S \\ &\le C_1 Pen(p_0)\end{aligned} \tag{11}$$

for some constant $C_1 > 4/3$ not depending on p_0, p and n.

Thus, for the RIC-type penalty (8), (11) yields $\sup_{\boldsymbol{\beta} \in \mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) = O(p_0 \ln p)$, while for the nonlinear $k \ln \frac{p}{k}$-type penalty (10) the Kullback-Leibler risk is of a smaller order $O\left(p_0 \ln(\frac{pe}{p_0})\right)$. Moreover, the latter can be improved further for dense models, where $p_0 \sim r$. Indeed, for a saturated model of size r in the RHS of (7), the penalty (10) yields

$$\begin{aligned}\sup_{\boldsymbol{\beta} \in \mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) &\le \sup_{\boldsymbol{\beta} \in \mathcal{B}(r)} \mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) \\ &\le C_1 Pen(r) = O(r)\end{aligned} \tag{12}$$

and the final upper bound for an estimator with the penalty (10) is, therefore,

$$C_1\ \frac{\mathcal{U}}{\mathcal{L}}\ \min\left(p_0 \ln \frac{pe}{p_0}, r\right) \tag{13}$$

with $C_1 > 4/3$.

To assess the accuracy of the upper bound (13) we establish the corresponding lower bound for the minimax Kullback-Leibler risk over $\mathcal{B}(p_0)$.

We introduce first some additional notation. For any given $k = 1, \ldots, r$, let $\phi_{min}[k]$ and $\phi_{max}[k]$ be the k-sparse minimal and maximal eigenvalues of the design defined as

$$\phi_{min}[k] = \min_{\boldsymbol{\beta}: 1 \le ||\boldsymbol{\beta}||_0 \le k} \frac{||X\boldsymbol{\beta}||^2}{||\boldsymbol{\beta}||^2},$$

$$\phi_{max}[k] = \max_{\boldsymbol{\beta}: 1 \le ||\boldsymbol{\beta}||_0 \le k} \frac{||X\boldsymbol{\beta}||^2}{||\boldsymbol{\beta}||^2}$$

In other words, $\phi_{min}[k]$ and $\phi_{max}[k]$ are respectively the minimal and maximal eigenvalues of all $k \times k$ submatrices of the matrix $X^t X$ generated by any k columns of X. Let $\tau[k] = \phi_{min}[k]/\phi_{max}[k]$, $k = 1, \ldots, r$.

Theorem 2: Consider a GLM with the canonical link $\boldsymbol{\theta} = X\boldsymbol{\beta}$ *under Assumptions (A).*

Let $1 \le p_0 \le r$ *and assume that* $\widetilde{\mathcal{B}}(p_0) \subseteq \mathcal{B}(p_0)$*, where the subsets* $\widetilde{\mathcal{B}}(p_0)$ *are defined in the proof. Then, there exists a constant* $C_2 > 0$ *such that*

$$\begin{aligned}&\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\beta} \in \mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, \widehat{\theta}) \\ &\ge \begin{cases} C_2\ \frac{\mathcal{L}}{\mathcal{U}} \tau[2p_0]\ p_0 \ln\left(\frac{pe}{p_0}\right), & p_0 \le \frac{r}{2} \\ C_2\ \frac{\mathcal{L}}{\mathcal{U}} \tau[p_0]\ r, & p_0 > \frac{r}{2} \end{cases}\end{aligned} \tag{14}$$

where the infimum is taken over all estimators $\hat{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}$*.*

C. Asymptotic Adaptive Minimaxity

We consider now the asymptotic properties of the proposed penalized MLE estimator as the sample size n increases. The number of predictors $p = p_n$ may increase with n as well, where we allow $p > n$ or even $p \gg n$. In such asymptotic setup there is essentially a *sequence* of design matrices X_{n,p_n}, where $r_n = rank(X_{n,p_n})$. For simplicity of notation, in what follows we omit the index n and denote X_{n,p_n} by X_p to highlight the dependence on p, and let r tend to infinity. Similarly, we define the corresponding sequences of regression coefficients $\boldsymbol{\beta}_p$ and sets $\mathcal{B}_p(p_0)$. The considered asymptotic GLM setup can now be viewed as a sequence of GLM models of the form $Y_i \sim f_{\theta_i}(y)$, $i = 1, \ldots, n$, where $f_{\theta_i}(y)$ are given in (1), $\theta_i \in \Theta$, $\boldsymbol{\theta} = X_p \boldsymbol{\beta}_p$ and $rank(X_p) = r \to \infty$.

As before, we assume that any r columns of X_p are linearly independent and, therefore, $\tau_p[r] > 0$. We distinguish between two possible cases: *weakly collinear* design, where the sequence $\tau_p[r]$ is bounded away from zero by some constant $c > 0$, and *multicollinear* design, where $\tau_p[r] \to 0$. Intuitively, it is clear that weak collinearity of the design cannot hold when p is "too large" relative to r. Indeed, [1, Remark 1] shows that for weakly collinear design, necessarily $p = O(r)$ and, thus, $p = O(n)$.

For weakly collinear design the following corollary is an immediate consequence of (13) and Theorem 2:

Corollary 1: Consider a GLM with the canonical link and weakly collinear design. Then, as r increases, under Assumption (A) and other assumptions of Theorem 2 the following statements hold:

1) *The asymptotic minimax Kullback-Leibler risk from the true model over $\mathcal{B}_p(p_0)$ is of the order $\min\left(p_0 \ln\left(\frac{pe}{p_0}\right), r\right)$ or essentially $p_0 \ln\left(\frac{pe}{p_0}\right)$ (since $p = O(r)$ – see comments above), that is, there exist two constants $0 < C_1 \leq C_2 < \infty$ depending possibly on the ratio $\frac{\mathcal{U}}{\mathcal{L}}$ such that for all sufficiently large r,*

$$C_1\ p_0 \ln\left(\frac{pe}{p_0}\right) \leq \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\beta}_p \in \mathcal{B}_p(p_0)} \mathbb{E}KL(X_p\boldsymbol{\beta}_p, \widehat{\boldsymbol{\theta}}) \leq C_2\ p_0 \ln\left(\frac{pe}{p_0}\right)$$

for all $1 \leq p_0 \leq r$.

2) *Consider penalized maximum likelihood model selection rule (4) with the complexity penalty $Pen(k) = C\,\frac{\mathcal{U}}{\mathcal{L}}\,k \ln\left(\frac{pe}{k}\right)$, $k = 1, \ldots, r-1$ and $Pen(r) = C\,\frac{\mathcal{U}}{\mathcal{L}}\,r$, where $C > 16$. Then, the resulting penalized MLE estimator $X_p\widehat{\boldsymbol{\beta}}_{p\widehat{M}}$ attains the minimax convergence rates (in terms of $\mathbb{E}KL(X_p\boldsymbol{\beta}_p, X_p\widehat{\boldsymbol{\beta}}_{p\widehat{M}})$) simultaneously over all $\mathcal{B}_p(p_0)$, $1 \leq p_0 \leq r$.*

Corollary 1 is a generalization of the corresponding results of [1] for Gaussian regression. It also shows that model selection criteria with RIC-type (linear) penalties (8) of the form $Pen(k) = Ck \ln p$ are of the minimax order for sparse models with $p_0 \ll p$ but only sub-optimal otherwise.

We would like to finish this section with several important remarks:

Remark 1: Under Assumption (A), $KL(\boldsymbol{\theta}, \boldsymbol{\zeta}) \asymp ||\boldsymbol{\theta} - \boldsymbol{\zeta}||^2$ (see Lemma 1 in the Appendix) and Corollary 1 implies then that $X_p\widehat{\boldsymbol{\beta}}_{p\widehat{M}}$ is also a minimax-rate estimator for natural parameters $\boldsymbol{\theta}_p = X_p\boldsymbol{\beta}_p$ in terms of the quadratic risk simultaneously over all $\mathcal{B}_p(p_0)$, $p_0 = 1, \ldots, r$. Furthermore, since $||X_p\widehat{\boldsymbol{\beta}}_{p\widehat{M}} - X_p\boldsymbol{\beta}_p||^2 \asymp ||\widehat{\boldsymbol{\beta}}_{p\widehat{M}} - \boldsymbol{\beta}_p||^2$ for weakly collinear design, the same is true for $\widehat{\boldsymbol{\beta}}_{p\widehat{M}}$ as an estimator of the regression coefficients $\boldsymbol{\beta}_p \in \mathcal{B}_p(p_0)$.

Remark 2: As we have mentioned, multicollinear design necessarily appears when $p \gg n$. For such type of design, $\tau_p[r]$ tends to zero, and there is a gap in the rates in the upper and lower bounds (13) and (14). Somewhat surprisingly, multicollinearity, being a "curse" for consistency of variable selection or estimation of regression coefficients $\boldsymbol{\beta}$, may be a "blessing" for estimating $\boldsymbol{\theta} = X\boldsymbol{\beta}$. For Gaussian regression [1] showed that strong correlations between predictors can be exploited to reduce the size of a model (thus, to decrease the variance) without paying much extra price in the bias and, therefore, to improve the upper bound (13). The analysis of multicollinear case is however much more delicate and technical even for the linear regression (see [1]), and we do not discuss its extension for GLM in this paper.

Remark 3: Like any model selection criteria based on complexity penalties, minimization in (4) is a nonconvex optimization problem that generally requires search over all possible models. To make computations practically feasible for high-dimensional data, common approaches are either various greedy algorithms (e.g., forward selection) that approximate the global solution of (4) by a stepwise sequence of local ones, or convex relaxation methods, where the original nonconvex problem is replaced by a related convex program. The most well-known and well-studied method is the celebrated Lasso ([22]). For *linear* complexity penalties of the form $Pen(|M|) = C|M|$ it replaces the original l_0-norm $|M| = ||\widehat{\boldsymbol{\beta}}_M||_0$ in (4) by the l_1-norm $||\widehat{\boldsymbol{\beta}}_M||_1$. Theoretical properties of Lasso for Gaussian regression have been intensively studied in the literature during the last decade (see, e.g., [5]). van de Geer [26] investigated Lasso in the GLM setup but with random design. In particular, she showed that under assumptions similar to Assumption (A) and some additional restrictions on the design, Lasso with a properly chosen tuning parameter C behaves similar to the RIC-type estimator and its Kullback-Leibler risk achieves the sub-optimal rate $O(p_0 \ln p)$.

D. Example: Logistic Regression

We now illustrate the obtained general results on logistic regression.

Consider the Bernoulli distribution $Bin(1, p)$. A simple calculus shows that it belongs to the natural exponential family with the natural parameter $\theta = \ln \frac{p}{1-p}$, $b(\theta) = \ln(1 + e^{\theta})$ and $a = 1$. Thus, $b''(\theta) = \frac{e^{\theta}}{(1+e^{\theta})^2} \leq 1/4$ and, as we have already mentioned in Section II-A, the condition (a) of Assumption (A) is satisfied with $\mathcal{U} = 1/4$ for any Θ, while the condition (b) is equivalent to the boundedness of θ: $\Theta = \{\theta : |\theta| \leq C_0\}$, where $\mathcal{L} = \frac{e^{C_0}}{(1+e^{C_0})^2}$. In terms of the original parameter of the binomial distribution $p = \frac{e^{\theta}}{1+e^{\theta}}$ it means that p is bounded away from zero and one: $\delta \leq p \leq 1 - \delta$ for some $0 < \delta < 1/2$ and $\mathcal{L} = \delta(1-\delta)$. The same restriction on p appears in [26].

Consider now a logistic regression, where a binary data $Y_i \sim Bin(1, p_i)$, $\mathbf{x}_i \in \mathbb{R}^p$ are deterministic and $\ln \frac{p_i}{1-p_i} = \boldsymbol{\beta}^t \mathbf{x}_i$, $i = 1, \ldots, n$. Following (3), for a given model M, the MLE of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_M = \arg \max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \sum_{i=1}^{n} \left\{\mathbf{x}_i^t \widetilde{\boldsymbol{\beta}}_M Y_i - \ln\left(1 + \exp(\mathbf{x}_i^t \widetilde{\boldsymbol{\beta}}_M)\right)\right\}, \quad (15)$$

where $\mathcal{B}_M$ was defined in (3). The MLE for the resulting probabilities p_{Mi}'s are $\widehat{p}_{Mi} = \frac{\exp(\widehat{\boldsymbol{\beta}}_M \mathbf{x}_i)}{1+\exp(\widehat{\boldsymbol{\beta}}_M \mathbf{x}_i)}$, $i = 1, \ldots, n$.

The model $\widehat{M}$ is selected w.r.t. the penalized maximum likelihood model selection criterion (4):

$$\widehat{M} = \arg \min_{M \in \mathfrak{M}} \left\{ \sum_{i=1}^{n} \left(\ln\left(1 + \exp(\mathbf{x}_i^t \widehat{\boldsymbol{\beta}}_M)\right) - \mathbf{x}_i^t \widehat{\boldsymbol{\beta}}_M Y_i\right) + Pen(|M|) \right\} \quad (16)$$

A straightforward calculus shows that the Kullback-Leibler divergence $KL(\mathbf{p}_1, \mathbf{p}_2)$ between two sample distributions from $Bin(1, p_{1i})$ and $Bin(1, p_{2i})$, $i = 1, \ldots, n$ is

$$KL(\mathbf{p}_1, \mathbf{p}_2) = \sum_{i=1}^{n} \left\{ p_{1i} \ln\left(\frac{p_{1i}}{p_{2i}}\right) + (1 - p_{1i}) \ln\left(\frac{1 - p_{1i}}{1 - p_{2i}}\right) \right\}$$

Assume that there exists a constant $C_0 < \infty$ such that $\max_{1\le i\le n} |\boldsymbol{\beta}^t \mathbf{x}_i| \le C_0$ or, equivalently, $\delta \le p_i \le 1-\delta$, $i = 1,\dots,n$ for some positive $\delta < 1/2$ (see above). Assumption (A) is, therefore, satisfied with $\mathcal{U} = 1/4$ and $\mathcal{L} = \delta(1-\delta)$.

Consider the $k \ln \frac{p}{k}$-type complexity penalty (10) $Pen(k) = Ck \ln \frac{pe}{k}$ for $k = 1,\dots,r-1$ and $Pen(r) = Cr$ in (16), where $C > \frac{4}{\delta(1-\delta)}$. From our general results from the previous sections it then follows that

$$EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) = O\left(\min\left(p_0 \ln \frac{pe}{p_0}, r\right)\right),$$

where $p_0 = ||\boldsymbol{\beta}||_0$ is the size of the true (unknown) underlying logistic regression model. For weakly collinear design, as r increases, it is the minimax rate of convergence.

Similarly, the RIC-type penalty $Pen(k) = Ck \ln p$, $k = 1,\dots,r$ with $C > \frac{4}{\delta(1-\delta)}$ in (16) yields the sub-optimal rate $O(p_0 \ln p)$.

IV. Possible Extensions

In this section we discuss some possible extensions of the results obtained in Section III.

A. Model Selection in GLM Under Structural Constraints

So far we considered the complete variable selection, where the set of admissible models $\mathfrak{M}$ contains all 2^p possible subsets of predictors $x_1,\dots,x_p$. However, in various GLM setups there may be additional structural constraints on the set of admissible models. Thus, for the ordered variable selection, where the predictors have some natural order, x_j can enter the model only after $x_1,\dots,x_{j-1}$ (e.g., polynomial regression). Models with interactions that cannot be selected without the corresponding main effects is an example of hierarchical constraints. Factor predictors associated with groups of indicator (dummy) variables, where either none or all of the group is selected, is an example of group structural constraints.

Model selection under structural constraints for Gaussian regression was considered in [2]. Its extension to GLM may be described as follows. Let $m(p_0)$ be the number of all admissible models of size p_0. As before we can consider only $1 \le p_0 \le r$, where $m(r) = 1$ if there are admissible models of size r. Obviously, $0 \le m(p_0) \le \binom{p}{p_0}$, where the two extreme cases $m(p_0) = 1$ and $m(p_0) = \binom{p}{p_0}$ for all $p_0 = 1,\dots,r-1$, correspond respectively to the ordered and complete variable selection.

Let $\mathfrak{M}$ be the set of all admissible models. We slightly change the original definition of $\mathcal{B}_M$ in (3) by the additional requirement that $\beta_j = 0$ iff $j \notin M$ to have $||\boldsymbol{\beta}||_0 = |M|$ for all $\boldsymbol{\beta} \in \mathcal{B}_M$. The model $\widehat{M}$ is selected w.r.t. (4) from all models in $\mathfrak{M}$ and the penalty $Pen(k)$ is relevant only for k with $m(k) \ge 1$. From the proof (see the Appendix) it follows that Theorem 1 can be immediately extended to a restricted set of models $\mathfrak{M}$ with an obviously modified condition (5) on the weights L_k. Namely, let

$$\sum_{k=1}^{r-1} m(k) e^{-kL_k} + e^{-rL_k} \le S \tag{17}$$

and

$$Pen(k) \ge 2\frac{\mathcal{U}}{\mathcal{L}}\, k(A + 2\sqrt{2L_k} + 4L_k), \quad k = 1,\dots,r;\ m(k) \ge 1$$

for some $A > 1$. Then, under Assumption (A)

$$\mathbb{E}KL(\boldsymbol{\theta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) \le \frac{4}{3} \inf_{M\in\mathfrak{M}} \left\{ \inf_{\widetilde{\boldsymbol{\beta}}\in\mathcal{B}_M} KL(\boldsymbol{\theta}, X\widetilde{\boldsymbol{\beta}}) + Pen(|M|) \right\} + \frac{16}{3}\,\frac{\mathcal{U}}{\mathcal{L}}\,\frac{2A-1}{A-1}\, S, \tag{18}$$

See [2], [6], and [7] for similar results for Gaussian regression under structural constraints.

In particular, (17) holds for $L_k = c\frac{1}{k}\max(\ln m(k), k)$, $k = 1,\dots,r;\ m(k) \ge 1$ for some $c > 1$ leading to the penalty of the form

$$Pen(k) \sim \frac{\mathcal{U}}{\mathcal{L}} \max(\ln m(k), k) \tag{19}$$

for all $1 \le k \le r$ such that $m(k) \ge 1$. For the complete variable selection, the penalty (19) is evidently the $k \ln \frac{p}{k}$-type penalty (10) from Section III, while for the ordered variable selection it implies the AIC-type penalty of the form $Pen(k) = C\frac{\mathcal{U}}{\mathcal{L}}\, k$ for some $C > 0$.

Consider now all admissible models of size p_0 and the corresponding set of regression coefficients $\mathcal{B}(p_0) = \bigcup_{M\in\mathfrak{M}:|M|=p_0} \mathcal{B}_M$. Repeating the arguments from Section III-B, for the complexity penalty (19), under Assumption (A), the general upper bound (18) yields

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) = O(Pen(p_0)) = O(\max(\ln m(p_0), p_0)) \tag{20}$$

with a constant depending on the ratio $\mathcal{U}/\mathcal{L}$.

The upper bound (20) can be improved further if there exist admissible models of size r. In this case $m(r) = 1$ and similar to (12) for complete variable selection, we have

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) = O(Pen(r)) = O(r)$$

that combining with (20) yields

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, X\widehat{\boldsymbol{\beta}}_{\widehat{M}}) = O(\min(\max(\ln m(p_0), p_0), r)) \tag{21}$$

In the supplementary material we show that if $m(p_0) \ge 1$, under Assumption (A) and correspondingly modified other assumptions of Theorem 2, the minimax lower bound over $\mathcal{B}(p_0)$ is

$$\inf_{\widetilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\beta}\in\mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta}, \widetilde{\theta}) \ge \begin{cases} C_2\,\frac{\mathcal{L}}{\mathcal{U}}\,\max\left\{\tau[2p_0]\frac{\ln m(p_0)}{\ln p_0}, \tau[p_0]p_0\right\}, & 1 \le p_0 \le \frac{r}{2} \\ C_2\,\frac{\mathcal{L}}{\mathcal{U}}\,\tau[p_0]r, & \frac{r}{2} \le p_0 \le r \end{cases} \tag{22}$$

for some $C_2 > 0$.

Thus, comparing the upper bounds (20)–(21) with the lower bound (22) one realizes that for weakly collinear design the proposed penalized maximum likelihood estimator with the complexity penalty of type (19) is asymptotically (as r increases) at least nearly-minimax (up to a possible $\ln p_0$-factor) simultaneously for all $1 \leq p_0 \leq r/2$ and for all $1 \leq p_0 \leq r$ if, in addition, $m(r) = 1$ (i.e., there exist admissible models of size r). In particular, for the ordered variable selection, both bounds are of the same order $O(p_0)$. In Section III we showed that it also achieves the exact minimax rate for complete variable selection. So far we can only conjecture that the $\ln p_0$-factor can be removed in (22) for a general case as well. See also [2] for similar results for Gaussian regression.

B. Aggregation in GLM

An interesting statistical problem related to model selection is aggregation. Originated by [17], it has been intensively studied in the literature during the last decade. See, for example, [8], [14], [20], [23], and [27] for aggregation in Gaussian regression. Aggregation in GLM was considered in [19] and can be described as follows.

We observe $(\mathbf{x}_i, Y_i),\ i = 1, \ldots, n$, where the distribution $f_{\theta_i}(\cdot)$ of Y_i belongs to the natural exponential family with a natural parameter θ_i (1). Unlike GLM regression with the canonical link, where we assume that $\theta_i = \boldsymbol{\beta}^t \mathbf{x}_i$, in aggregation setup we do not rely on such modeling assumption but simply seek the best linear approximation $\boldsymbol{\theta}_{\boldsymbol{\beta}} = \sum_{j=1}^p \beta_j \mathbf{x}_j$ of $\boldsymbol{\theta}$ w.r.t. Kullback-Leibler divergence, where $\boldsymbol{\beta} \in \mathcal{B} \subseteq \mathbb{R}^p$, by solving the following optimization problem:

$$\inf_{\boldsymbol{\beta} \in \mathcal{B}} KL(\boldsymbol{\theta}, \boldsymbol{\theta}_{\boldsymbol{\beta}}) \tag{23}$$

Depending on the specific choice of $\mathcal{B} \subseteq \mathbb{R}^p$ there are different aggregation strategies. Following the terminology of [8]) there are *linear* aggregation ($\mathcal{B} = \mathcal{B}_L = \mathbb{R}^p$), *convex* aggregation ($\mathcal{B} = \mathcal{B}_C = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta}_j \geq 0,\ \sum_{j=1}^p \beta_j = 1\}$), *model selection* aggregation ($\mathcal{B} = \mathcal{B}_{MS}$ is a subset of vectors with a single nonzero entry), and *subset selection* or p_0*-sparse* aggregation ($\mathcal{B} = \mathcal{B}_{SS}(p_0) = \{\boldsymbol{\beta} \in \mathbb{R}^p : ||\boldsymbol{\beta}||_0 \leq p_0\}$ for a given $1 \leq p_0 \leq r$). In fact, linear and model selection aggregation can be viewed as two extreme cases of subset selection aggregation, where $\mathcal{B}_L = \mathcal{B}_{SS}(r)$ and $\mathcal{B}_{MS} = \mathcal{B}_{SS}(1)$.

Since in practice $\boldsymbol{\theta}$ is unknown, the solution of (23) is unavailable. The goal then is to construct an estimator (linear aggregator) $\boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}$ that mimics the ideal (oracle) solution $\boldsymbol{\theta}_{\boldsymbol{\beta}}$ of (23) as close as possible. More precisely, we would like to find $\boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}$ such that

$$\mathbb{E}KL(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}) \leq C \inf_{\boldsymbol{\beta} \in \mathcal{B}} KL(\boldsymbol{\theta}, \boldsymbol{\theta}_{\boldsymbol{\beta}}) + \Delta_{\mathcal{B}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}), \quad C \geq 1 \tag{24}$$

with the minimal possible $\Delta_{\mathcal{B}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}})$ (called *excess-KL*) and C close to one.

For weakly collinear design, [19, Th. 4.1] established the minimal possible asymptotic rates for $\Delta_{\mathcal{B}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}})$ for linear, convex and model selection aggregation under Assumption (A) and assumptions similar to those of Theorem 2:

$$\inf_{\boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}} \sup_{\boldsymbol{\theta}} \Delta_{\mathcal{B}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}) = \begin{cases} O(r), & \mathcal{B} = \mathcal{B}_L \\ O\left(\min(\sqrt{n \ln p}, r)\right), & \mathcal{B} = \mathcal{B}_C \\ O\left(\min(\ln p, r)\right), & \mathcal{B} = \mathcal{B}_{MS} \end{cases} \tag{25}$$

He also proposed an estimator $\boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}$ that achieves these optimal aggregation rates even with $C = 1$ in (24).

Using the results of Section III we can complete the case of subset selection aggregation in GLM, where under the assumptions of [19, Th. 4.1], $\mathcal{B}_{SS}(p_0)$ is essentially $\mathcal{B}(p_0)$ considered in the context of GLM model selection in previous sections. Indeed, repeating the arguments in the proof of Theorem 2 (see Appendix) implies that for $\mathcal{B}(p_0)$ there exists $C_2 > 0$ such that

$$\inf_{\boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}} \sup_{\boldsymbol{\theta}} \Delta_{\mathcal{B}(p_0)}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}}) \geq C_2 \frac{\mathcal{L}}{\mathcal{U}} \min\left(p_0 \ln\left(\frac{pe}{p_0}\right), r\right) \tag{26}$$

In particular, (26) also yields the lower bounds (25) for excess-KL for linear ($p_0 = r$) and model selection ($p_0 = 1$) aggregation. Furthermore, similar to model selection in GLM within $\mathcal{B}(p_0)$ considered in Section III-B, from Theorem 1 it follows that for weakly collinear design, the penalized maximum likelihood estimator $\boldsymbol{\theta}_{\widehat{\boldsymbol{\beta}}_{\widehat{M}}}$ with the complexity penalty (10) achieves the optimal rate (26) for subset selection aggregation over $\mathcal{B}(p_0)$ for all $1 \leq p_0 \leq r$ (and, therefore, for linear and model selection aggregation in particular) though with some $C > 4/3$ in (24). Similar to the results of [20] for Gaussian regression, we may conjecture that to get $C = 1$ one should average estimators from all models with properly chosen weights rather than select a single one as in model selection.

APPENDIX A

We first prove the following lemma establishing the equivalence of the Kullback-Leibler divergence $KL(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and the squared quadratic norm $||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2$ under Assumption (A) that will be used further in the proofs:

Lemma 1: Let Assumption (A) hold. Then, for any $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in \mathbb{R}^n$ *such that* $\theta_{1i}, \theta_{2i} \in \Theta,\ i = 1, \ldots, n$,

$$\frac{\mathcal{L}}{2a}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2 \leq KL(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \leq \frac{\mathcal{U}}{2a}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||^2$$

Proof: Recall that for a GLM

$$KL(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{a} \sum_{i=1}^n \left\{ b'(\theta_{1i})(\theta_{1i} - \theta_{2i}) - b(\theta_{1i}) + b(\theta_{2i}) \right\} \tag{27}$$

A Taylor expansion of $b(\theta_{2i})$ around θ_{1i} yields $b(\theta_{2i}) = b(\theta_{1i}) + b'(\theta_{1i})(\theta_{2i} - \theta_{1i}) + \frac{b''(c_i)}{2}(\theta_{2i} - \theta_{1i})^2$, where c_i lies between θ_{1i} and θ_{2i}, and substituting into (27) we have

$$KL(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{2a} \sum_{i=1}^n b''(c_i)(\theta_{2i} - \theta_{1i})^2$$

Due to Assumption (A), Θ is an interval and, therefore, $c_i \in \Theta$. Hence, $\mathcal{L} \leq b''(c_i) \leq \mathcal{U}$ that completes the proof. $\square$

Appendix B
Proof of Theorem 1

We introduce first some notation. For a given model M, define

$$\boldsymbol{\beta}_M = \arg \inf_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} KL(\boldsymbol{\theta}, X\widetilde{\boldsymbol{\beta}}),$$

where $\mathcal{B}_M$ is given in (3), and let $\boldsymbol{\theta}_M = X\boldsymbol{\beta}_M$. As we have mentioned in Section III-A, $\boldsymbol{\theta}_M$ can be interpreted as the closest vector to $\boldsymbol{\theta}$ within the span generated by a subset of columns of X corresponding to M w.r.t. a Kullback-Leibler divergence. Recall also that $\widehat{\boldsymbol{\theta}}_M = X\widehat{\boldsymbol{\beta}}_M$ is the MLE of $\boldsymbol{\theta}$ for the model M and, in particular, $\widehat{\boldsymbol{\theta}}_{\widehat{M}} = X\widehat{\boldsymbol{\beta}}_{\widehat{M}}$. Finally, for any random variable η let $\varphi_\eta(\cdot)$ be its moment generating function.

For the clarity of exposition, we split the proof into several steps.

Step 1: Since $\widehat{M}$ is the minimizer defined in (4), for any given model M

$$-\ell(\widehat{\boldsymbol{\beta}}_{\widehat{M}}) + Pen(|\widehat{M}|) \leq -\ell(\boldsymbol{\beta}_M) + Pen(|M|) \tag{28}$$

By a straightforward calculus, one can easily verify that

$$\begin{aligned} KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\widehat{M}}) - KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) &= \ell(\boldsymbol{\beta}_M) - \ell(\widehat{\boldsymbol{\beta}}_{\widehat{M}}) \\ &\quad + \frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{\widehat{M}} - \boldsymbol{\theta}_M) \end{aligned} \tag{29}$$

and, hence, (28) yields

$$\begin{aligned} KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\widehat{M}}) + Pen(|\widehat{M}|) &\leq KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) + Pen(|M|) \\ &\quad + \frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{\widehat{M}} - \boldsymbol{\theta}_M) \end{aligned} \tag{30}$$

Note that $\mathbb{E}Y = b'(\boldsymbol{\theta})$, $\mathbb{E}\left\{(\mathbf{Y} - b'(\boldsymbol{\theta}))^t\boldsymbol{\zeta}\right\} = 0$ for any deterministic vector $\boldsymbol{\zeta} \in \mathbb{R}^n$ and, therefore,

$$\mathbb{E}\left((\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{\widehat{M}} - \boldsymbol{\theta}_M)\right) = \mathbb{E}\left((\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{\widehat{M}} - \boldsymbol{\theta})\right)$$

Furthermore, by the definition of $\boldsymbol{\theta}_{\widehat{M}}$, $KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\widehat{M}}) \geq KL(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{M}})$, and since (30) holds for any model M in the RHS, we have

$$\begin{aligned} \frac{3}{4}\,\mathbb{E}KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\widehat{M}}) &\leq \inf_M \{KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) + Pen(|M|)\} \\ &\quad + \mathbb{E}\left(\frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{\widehat{M}} - \boldsymbol{\theta}) - Pen(|\widehat{M}|) \right. \\ &\qquad \left. \frac{1}{4}KL(\boldsymbol{\theta}, \boldsymbol{\theta}_{\widehat{M}})\right) \end{aligned} \tag{31}$$

Step 2: Consider now the term $\frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{\widehat{M}} - \boldsymbol{\theta})$ in the RHS of (31). The selected model $\widehat{M}$ in (4) can, in principle, be any model M and we want, therefore, to control it uniformly over M. For any M we have

$$\begin{aligned} \frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}) &= \frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M) \\ &\quad + \frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\boldsymbol{\theta}_M - \boldsymbol{\theta}) \end{aligned} \tag{32}$$

Let Ξ_M be any orthonormal basis of the span of columns of X corresponding to the model M and $\boldsymbol{\xi}_M = \Xi_M \Xi_M^t(\mathbf{Y} - b'(\boldsymbol{\theta}))$ be the projection of $\mathbf{Y} - b'(\boldsymbol{\theta})$ on this span.

Then, by the Cauchy-Schwarz inequality

$$\begin{aligned} (\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M) &= \boldsymbol{\xi}_M^t(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M) \\ &\leq ||\boldsymbol{\xi}_M|| \cdot ||\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M|| \end{aligned} \tag{33}$$

Since $\widehat{\boldsymbol{\theta}}_M$ is the MLE for a given M, $\ell(\widehat{\boldsymbol{\theta}}_M) \geq \ell(\boldsymbol{\theta}_M)$ and, therefore, (29) implies

$$KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_M) \leq KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) + \frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M) \tag{34}$$

Similar to the proof of [19, Lemma 6.3], using a Taylor expansion it follows that under Assumption (A), $KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_M) - KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) \geq \frac{\mathcal{L}}{2a}||\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M||^2$ that together with (33) and (34) yields

$$\frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_M - \boldsymbol{\theta}_M) \leq \frac{2}{a\mathcal{L}}||\boldsymbol{\xi}_M||^2 \tag{35}$$

Define

$$\begin{aligned} R(M) &= \frac{2}{a\mathcal{L}}||\boldsymbol{\xi}_M||^2 + \frac{1}{a}(\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\boldsymbol{\theta}_M - \boldsymbol{\theta}) \\ &\quad - Pen(|M|) - \frac{1}{4}KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) \end{aligned}$$

Then, from (31),

$$\mathbb{E}KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{\widehat{M}}) \leq \frac{4}{3}\inf_M \{KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) + Pen(|M|)\} + \frac{4}{3}\,\mathbb{E}R(\widehat{M}) \tag{36}$$

and to complete the proof we need to find an upper bound for $\mathbb{E}R(\widehat{M})$.

Step 3: Consider $\varphi_{||\boldsymbol{\xi}_M||^2}(\cdot)$. By [19, eq. (6.3)],

$$\mathbb{E}e^{\mathbf{w}^t(\mathbf{Y} - b'(\boldsymbol{\theta}))} \leq e^{\frac{\mathcal{U}a||\mathbf{w}||^2}{2}}$$

for any $\mathbf{w} \in \mathbb{R}^n$. The projection matrix $\Xi_M \Xi_M^t$ is idempotent and $tr(\Xi_M \Xi_M^t) = |M|$. We can apply then [13, Remark 2.3] to have

$$\varphi_{||\boldsymbol{\xi}_M||^2}(s) \leq \exp\left\{a\mathcal{U}s|M| + \frac{a^2\mathcal{U}^2 s^2|M|}{1 - 2a\mathcal{U}s}\right\} \tag{37}$$

for all $0 < s < \frac{1}{2a\mathcal{U}}$.

Consider now the random variable $\eta_M = (\mathbf{Y} - b'(\boldsymbol{\theta}))^t(\boldsymbol{\theta}_M - \boldsymbol{\theta})$. Applying [19, eq. (6.3)] yields

$$\varphi_{\eta_M}(s) \leq \exp\left\{\frac{1}{2}s^2\mathcal{U}a||\boldsymbol{\theta}_M - \boldsymbol{\theta}||^2\right\} \tag{38}$$

Define $Z = \frac{2}{a\mathcal{L}}(||\boldsymbol{\xi}_M||^2 - a\mathcal{U}|M|) + \frac{1}{a}\eta_M = R(M) + Pen(|M|) + \frac{1}{4}KL(\boldsymbol{\theta}, \boldsymbol{\theta}_M) - 2\frac{\mathcal{U}}{\mathcal{L}}|M|$. Unlike Gaussian regression, $||\boldsymbol{\xi}||_M^2$ and η_M are not independent. However, by the Cauchy-Schwarz inequality

$$\varphi_Z(s) \leq e^{-2\frac{\mathcal{U}}{\mathcal{L}}|M|s} \cdot \sqrt{\varphi_{\frac{2}{a\mathcal{L}}||\boldsymbol{\xi}_M||^2}(2s)} \cdot \sqrt{\varphi_{\frac{1}{a}\eta_M}(2s)}$$

and from (37) and (38),

$$\varphi_Z(s) \leq \exp\left\{\frac{8\,\frac{\mathcal{U}^2}{\mathcal{L}^2}|M|s^2}{1 - 8\,\frac{\mathcal{U}}{\mathcal{L}}s} + \frac{\mathcal{U}s^2}{a}||\boldsymbol{\theta}_M - \boldsymbol{\theta}||^2\right\} \tag{39}$$

for all $0 < s < \frac{\mathcal{L}}{8\mathcal{U}}$.

Let $x = 8\frac{\mathcal{U}}{\mathcal{L}}s$ $(0 < x < 1)$ and $\rho = \frac{\mathcal{L}^2||\boldsymbol{\theta}_M-\boldsymbol{\theta}||^2}{64a\mathcal{U}}$. Then, using the obvious inequality $\rho x^2 < \rho x$ for $0 < x < 1$, after a straightforward calculus (39) yields

$$\ln \varphi_{\frac{\mathcal{L}}{8\mathcal{U}}Z-\rho}(x) \le \frac{|M|}{8}\frac{x^2}{1-x}$$

for all $0 < x < 1$.

We can now apply [7, Lemma 2] to get $P(\frac{\mathcal{L}}{8\mathcal{U}}Z - \rho \ge \sqrt{\frac{|M|}{2}\,t} + t) \le e^{-t}$ for all $t > 0$, that is,

$$P\left\{\frac{\mathcal{L}}{8\mathcal{U}}\left(R(M)+Pen(|M|)+\frac{1}{4}KL(\boldsymbol{\theta},\boldsymbol{\theta}_M) - \frac{\mathcal{L}||\boldsymbol{\theta}_M-\boldsymbol{\theta}||^2}{8a}\right) \ge \frac{|M|}{4}+\sqrt{\frac{|M|}{2}\,t}+t\right\} \le e^{-t}$$

Lemma 1 implies that $\frac{1}{4}KL(\boldsymbol{\theta},\boldsymbol{\theta}_M) - \frac{\mathcal{L}||\boldsymbol{\theta}_M-\boldsymbol{\theta}||^2}{8a} \ge 0$ and, therefore,

$$P\left\{\frac{\mathcal{L}}{8\mathcal{U}}\left(R(M)+Pen(|M|)\right) \ge \frac{|M|}{4}+\sqrt{\frac{|M|}{2}\,t}+t\right\} \le e^{-t} \tag{40}$$

Step 4: Based on (40) we can now find an upper bound for $\mathbb{E}R(\widehat{M})$.

Let $k = |M|$ and take $t = kL_k + \omega$ for any $\omega > 0$, where $L_k > 0$ are the weights from Theorem 1. Using inequalities $\sqrt{c_1+c_2} \le \sqrt{c_1}+\sqrt{c_2}$ and $\sqrt{c_1c_2} \le \frac{1}{2}(c_1\epsilon + \frac{c_2}{\epsilon})$ for any positive c_1, c_2 and ϵ, we have

$$\sqrt{kt} \le k\sqrt{L_k}+\sqrt{k\omega} \le k\sqrt{L_k}+\frac{1}{2}\left(k\epsilon+\frac{\omega}{\epsilon}\right)$$

and, therefore,

$$P\left\{\frac{\mathcal{L}}{8\mathcal{U}}\left(R(M)+Pen(k)\right) \ge \frac{k}{4}\left(1+\sqrt{2}\,\epsilon+2\sqrt{2L_k}+4L_k\right) + \omega\left(1+\frac{1}{2\sqrt{2}\,\epsilon}\right)\right\} \le e^{-(kL_k+\omega)}$$

For the penalty $Pen(k)$ satisfying (6) with some $A > 1$ and $\epsilon = (A-1)/\sqrt{2}$, we then have

$$P\left\{\frac{\mathcal{L}}{4\mathcal{U}}R(M) \ge \omega\frac{2A-1}{A-1}\right\} \le e^{-(kL_k+\omega)} \tag{41}$$

for all M.

Finally, under the condition (5) on the weights L_k, (41) implies

$$\begin{aligned} P\left\{R(\widehat{M}) \ge \frac{4\mathcal{U}}{\mathcal{L}}\,\omega\frac{2A-1}{A-1}\right\} &\le \sum_M P\left\{R(M) \ge \frac{4\mathcal{U}}{\mathcal{L}}\,\omega\frac{2A-1}{A-1}\right\} \le \sum_M e^{-(kL_k+\omega)} \\ &\le Se^{-\omega} \end{aligned}$$

and, hence,

$$\mathbb{E}R(\widehat{M}) \le \int_0^\infty P(R(\widehat{M}) > t)dt \le \frac{4\mathcal{U}}{\mathcal{L}}\;\frac{2A-1}{A-1}\;S$$

that together with (36) completes the proof. □

APPENDIX C
PROOF OF THEOREM 2

Due to Lemma 1, the minimax lower bound for the Kullback-Leibler risk can be reduced to the lower bound for the corresponding quadratic risk:

$$\inf_{\widetilde{\boldsymbol{\theta}}}\ \sup_{\boldsymbol{\beta}\in\mathcal{B}(p_0)} \mathbb{E}KL(X\boldsymbol{\beta},\widetilde{\boldsymbol{\theta}}) \ge \frac{\mathcal{L}}{2a}\inf_{\widetilde{\boldsymbol{\theta}}}\ \sup_{\boldsymbol{\beta}\in\mathcal{B}(p_0)} \mathbb{E}||X\boldsymbol{\beta}-\widetilde{\boldsymbol{\theta}}||^2 \tag{42}$$

Following a general reduction scheme for establishing the minimax risk lower bounds, the quadratic risk in (42) is first reduced to the probability of misclassification error among a properly chosen finite subset $\Theta^*(p_0) \subset \{\boldsymbol{\theta}\in\mathbb{R}^n : \boldsymbol{\theta} = X\boldsymbol{\beta},\ \boldsymbol{\beta}\in\mathcal{B}(p_0)\}$ such that for any $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \Theta^*(p_0)$, $||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||^2 \ge 4s^2(p_0)$:

$$\begin{aligned} \inf_{\widetilde{\boldsymbol{\theta}}}\ \sup_{\boldsymbol{\beta}\in\mathcal{B}(p_0)} \mathbb{E}||X\boldsymbol{\beta}-\widetilde{\boldsymbol{\theta}}||^2 &\ge \inf_{\widetilde{\boldsymbol{\theta}}}\ \max_{\boldsymbol{\theta}_j\in\Theta^*(p_0)} \mathbb{E}||\boldsymbol{\theta}_j-\widetilde{\boldsymbol{\theta}}||^2 \\ &\ge 4s^2(p_0)\inf_{\widetilde{\boldsymbol{\theta}}}\ \max_{\boldsymbol{\theta}_j\in\Theta^*(p_0)} P_{\boldsymbol{\theta}_j}(\widetilde{\boldsymbol{\theta}}\neq\boldsymbol{\theta}_j) \end{aligned}$$

and then bounding the latter from below (e.g., applying various versions of Fano' lemma). See [24, Sec. 2] for more details.

In particular, the idea of our proof is to find a finite subset $\mathcal{B}^*(p_0) \subseteq \mathcal{B}(p_0)$ of vectors $\boldsymbol{\beta}$ and the corresponding subset $\Theta^*(p_0) = \{\boldsymbol{\theta}\in\mathbb{R}^n : \boldsymbol{\theta} = X\boldsymbol{\beta},\ \boldsymbol{\beta}\in\mathcal{B}^*(p_0)\}$ such that for any $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \Theta^*(p_0)$, $||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||^2 \ge 4s^2(p_0)$ and $KL(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \le (1/16)\ln\text{card}(\Theta^*(p_0))$. It will follow then from [8, Lemma A.1] that $s^2(p_0)$ is the minimax lower bound for the quadratic risk over $\mathcal{B}(p_0)$.

To construct such subsets we can exploit the techniques similar to that used in the corresponding proofs for the quadratic risk in linear regression (e.g., [1], [20]). Consider three cases.

Case 1: $p_0 \le r/2$.

Define the subset $\widetilde{\mathcal{B}}(p_0)$ of all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ that have p_0 entries equal to C_{p_0}, where C_{p_0} will be defined below and others are zeros: $\widetilde{\mathcal{B}}(p_0) = \{\boldsymbol{\beta}\in\mathbb{R}^p : \boldsymbol{\beta}\in\{0,C_{p_0}\}^p,\ ||\boldsymbol{\beta}||_0 = p_0\}$. From [20, Lemma A.3], there exists a subset $\mathcal{B}^*(p_0) \subset \widetilde{\mathcal{B}}(p_0)$ such that $\ln\text{card}(\mathcal{B}^*(p_0)) \ge \tilde{c}p_0\ln\left(\frac{pe}{p_0}\right)$ for some constant $0 < \tilde{c} < 1$, and for any pair $\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2 \in \mathcal{B}^*(p_0)$, the Hamming distance $\rho(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = \sum_{j=1}^p \mathbb{I}\{\beta_{1j}\neq\beta_{2j}\} \ge \tilde{c}p_0$.

Take $C_{p_0}^2 = \frac{1}{16}\tilde{c}\frac{a}{\mathcal{U}}\phi_{\max}^{-1}[2p_0]\ln\left(\frac{pe}{p_0}\right)$. By the assumptions of the theorem, $\mathcal{B}^*(p_0) \subset \widetilde{\mathcal{B}}(p_0) \subseteq \mathcal{B}(p_0)$. Consider the corresponding subset $\Theta^*(p_0)$. Evidently, $\text{card}(\Theta^*(p_0)) = \text{card}(\mathcal{B}^*(p_0))$, and for any $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2 \in \Theta^*(p_0)$ associated with $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2 \in \mathcal{B}^*(p_0)$ we then have

$$\begin{aligned} ||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||^2 = ||X(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)||^2 &\ge \phi_{min}[2p_0]\;||\boldsymbol{\beta}_1-\boldsymbol{\beta}_2||^2 \\ &\ge \tilde{c}\phi_{min}[2p_0]\;C_{p_0}^2\;p_0 = 4s^2(p_0), \end{aligned} \tag{43}$$

where $s^2(p_0) = \frac{1}{64}\frac{a}{\mathcal{U}}\tilde{c}^2\tau[2p_0]p_0\ln\left(\frac{pe}{p_0}\right)$.

On the other hand,

$$K(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \le \frac{\mathcal{U}}{2a}||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||^2 \le \frac{\mathcal{U}}{2a}\ \phi_{max}[2p_0]C_{p_0}^2\rho(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \le \frac{\mathcal{U}}{a}\ \phi_{max}[2p_0]C_{p_0}^2 p_0 \le \frac{1}{16}\ln \mathrm{card}(\Theta^*(p_0)), \tag{44}$$

where the first inequality follows from Lemma 1. [8, Lemma A.1] and (42) complete then the proof for this case.

Case 2: $r/2 \le p_0 \le r,\ p_0 \ge 8$.

In this case consider the subset $\widetilde{\mathcal{B}}(p_0) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \boldsymbol{\beta} \in \{\{0, C_{p_0}\}^{p_0}, 0, \ldots, 0\}$, where $C_{p_0}^2 = \frac{\ln 2}{64}\frac{a}{\mathcal{U}}\phi_{\max}^{-1}[p_0]$. From the assumptions of the theorem $\widetilde{\mathcal{B}}(p_0) \subseteq \mathcal{B}(p_0)$. Varshamov-Gilbert bound (see, e.g., [24, Lemma 2.9]) guarantees the existence of a subset $\mathcal{B}^*(p_0) \subset \widetilde{\mathcal{B}}(p_0)$ such that $\ln \mathrm{card}(\mathcal{B}_{p_0}^*) \ge \frac{p_0}{8}\ln 2$ and the Hamming distance $\rho(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \ge \frac{p_0}{8}$ for any pair $\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2 \in \mathcal{B}_{p_0}^*$.

Note that for any $\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2 \in \mathcal{B}_{p_0}^*$, $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ has at most p_0 nonzero components and repeating the arguments for the Case 1, one obtains the minimax lower bound $s^2(p_0) = C\frac{a}{\mathcal{U}}\tau[p_0]p_0 \ge \frac{C}{2}\frac{a}{\mathcal{U}}\tau[p_0]r$ for the quadratic risk. Applying (42) completes the proof.

Case 3: $r/2 \le p_0 \le r,\ 2 \le p_0 < 8$.

For this case, obviously, $2 \le r < 16$. Consider a trivial subset $\mathcal{B}_{p_0}^*$ containing just two vectors $\boldsymbol{\beta}_1 \equiv 0$ and $\boldsymbol{\beta}_2$ that has first p_0 nonzero entries equal to C_{p_0}, where $C_{p_0}^2 = \frac{\ln 2}{64}\frac{a}{\mathcal{U}}\phi_{max}^{-1}[p_0]$. Under the assumptions of the theorem $\mathcal{B}_{p_0}^* \subset \mathcal{B}(p_0)$. For the corresponding $\boldsymbol{\theta}_1 = X\boldsymbol{\beta}_1$ and $\boldsymbol{\theta}_2 = X\boldsymbol{\beta}_2$, (43) and (44) yield

$$KL(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \le \frac{\mathcal{U}}{2a}\phi_{max}[p_0]8C_{p_0}^2 = \frac{1}{16}\ln \mathrm{card}(\Theta_{p_0}^*)$$

and

$$||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||^2 \ge \phi_{min}[p_0]p_0C_{p_0}^2 = C\frac{a}{\mathcal{U}}\tau[p_0]p_0 \ge \frac{C}{2}\frac{a}{\mathcal{U}}\tau[p_0]r$$

and the proof follows from [8, Lemma A.1]. □

Acknowledgement

The authors are grateful to Alexander Goldenshluger, Ya'acov Ritov and Ron Peled for valuable remarks.

References

[1] F. Abramovich and V. Grinshtein, "MAP model selection in Gaussian regression," *Electron. J. Statist.*, vol. 4, pp. 932–949, 2010.
[2] F. Abramovich and V. Grinshtein, "Model selection in regression under structural constraints," *Electron. J. Statist.*, vol. 7, pp. 480–498, 2013.
[3] F. Abramovich, V. Grinshtein, A. Petsa, and T. Sapatinas, "On Bayesian testimation and its application to wavelet thresholding," *Biometrika*, vol. 97, no. 1, pp. 181–198, 2010.
[4] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Inf. Theory*, Budapest, Hungary, 1973, pp. 267–281.
[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.
[6] L. Birgé and P. Massart, "Gaussian model selection," *J. Eur. Math. Soc.*, vol. 3, no. 3, pp. 203–268, 2001.
[7] L. Birgé and P. Massart, "Minimal penalties for Gaussian model selection," *Probab. Theory Related Fields*, vol. 138, no. 1, pp. 33–73, 2007.
[8] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Aggregation for Gaussian regression," *Ann. Statist.*, vol. 35, no. 4, pp. 1674–1697, 2007.
[9] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.
[10] J. Chen and Z. Chen, "Extended BIC for small-n-large-P sparse GLM," *Statist. Sin.*, vol. 22, no. 2, pp. 555–574, 2012.
[11] J. Fan and R. Son, "SURE independence screening in generalized linear models with NP-dimensionality," *Ann. Statist.*, vol. 38, no. 6, pp. 3567–3604, 2010.
[12] D. P. Foster and E. I. George, "The risk inflation criterion for multiple regression," *Ann. Statist.*, vol. 22, no. 4, pp. 1947–1975, 1994.
[13] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electron. Commun. Probab.*, vol. 17, pp. 1–6, Nov. 2012.
[14] G. Leung and A. R. Barron, "Information theory and mixing least-squares regressions," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3396–3410, Aug. 2006.
[15] C. L. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.
[16] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London, U.K.: Chapman & Hall, 1989.
[17] A. Nemirovski, "Topics in nonparametric statistics," in *Lectures on Probability Theory and Statistics: Ecole d'Ete de Probabilites de Saint-Flour XXVIII* (Lecture Notes in Mathematics). Berlin, Germany: Springer, 2000.
[18] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.
[19] P. Rigollet, "Kullback–Leibler aggregation and misspecified generalized linear models," *Ann. Statist.*, vol. 40, no. 2, pp. 639–665, 2012.
[20] P. Rigollet and A. Tsybakov, "Exponential Screening and optimal rates of sparse estimation," *Ann. Statist.*, vol. 39, no. 2, pp. 731–771, 2011.
[21] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[23] A. B. Tsybakov, "Optimal rates of aggregation," in *Proc. COLT*, vol. 2777. 2003, pp. 303–313.
[24] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York, NY, USA: Springer, 2009.
[25] N. Verzelen, "Minimax risks for sparse regressions: Ultra-high dimensional phenomenons," *Electron. J. Statist.*, vol. 6, pp. 38–90, 2012.
[26] S. A. van de Geer, "High-dimensional generalized linear models and the Lasso," *Ann. Statist.*, vol. 36, no. 2, pp. 614–645, 2008.
[27] Y. Yang, "Aggregating regression procedures to improve performance," *Bernoulli*, vol. 10, no. 1, pp. 25–47, 2004.

Felix Abramovich received M.Sc. (1986) in Applied Mathematics from Moscow Oil and Gaz Institute and Ph.D. (1993) in Statistics from Tel Aviv University. He is an Associate Professor at the Department of Statistics and Operations Research, Tel Aviv University. His research interests include high-dimensional estimation and inference.

Vadim Grinshtein received M.Sc. (1986) in Mathematics from Simferopol State University, USSR and Ph.D. (2001) in Mathematics from Tel Aviv University. He works at the Department of Mathematics and Computer Science, The Open University of Israel. His research interests include spectral theory of operators and mathematical statistics.