Model selection and minimax estimation in generalized linear models

Felix Abramovich
Department of Statistics and Operations Research
Tel Aviv University

Vadim Grinshtein
Department of Mathematics and Computer Science
The Open University of Israel

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 expected Kullback-Leibler divergence between the true distribution of the data and that generated by a selected model, and establish the corresponding minimax lower bounds for 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 discuss also possible extensions of the proposed approach to model selection in GLM under additional structural constraints and aggregation.

1 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 of Akaike, 1973; Mallows, 1973 and Schwarz, 1978), 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 George & Foster (1994), Birgé & Massart (2001, 2007), Chen & Chen (2008), Bickel et al. (2009), Abramovich & Grinshtein (2010), Raskutti et al. (2011), Rigollet & Tsybakov (2011) among many others. A review on model selection in Gaussian regression for “p larger than n” setup can be found in Verzelen (2012).

Generalized linear models (GLM) is a generalization of Gaussian linear regression, where the distribution of response is not necessarily normal but belongs to the 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 McCullogh & Nelder (1989).

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, 1973 and BIC of Schwarz, 1978), 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., Fan & Song, 2010; Chen & Chen, 2012), but a rigorous theory of model selection for estimation and prediction in GLM remains essentially terra incognita. We can mention a recent paper by Rigollet (2012) that considered an 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 Birgé & Massart (2001, 2007) for Gaussian regression. We derive a general nonasymptotic upper bound for the expected Kullback-Leibler divergence between the unknown true distribution of the data and that generated by a selected model. Furthermore, for the properly chosen penalty we establish asymptotic minimaxity of the resulting estimator and its 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 2. Its main theoretical properties are presented in Section 3. In particular, we derive the upper bound for its expected Kullback-Leibler divergence from the true distribution and corresponding minimax lower bounds, and establish its asymptotic minimaxity over various sparse settings. Extensions to model selection under structural constraints and aggregation are discussed in Section 4. All the proofs are given in the Appendix.
2 Model selection procedure for GLM

2.1 Setup and notation

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

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

The function $b(\cdot)$ is assumed to be twice-differentiable. In this case $\mathbb{E}(Y_i) = b'(\theta_i)$ and $\text{Var}(Y_i) = ab''(\theta_i)$ (see McCullagh & Nelder, 1989). To complete GLM we assume the canonical link $\theta_i = \beta^t x_i$ or, equivalently, in the matrix form, $\theta = X\beta$, where $X_{n \times p}$ is the design matrix and $\beta \in \mathbb{R}^p$ is a vector of the unknown regression coefficients.

In what follows we assume the following assumption on the parameter space $\Theta$ 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 $0 < L \leq U < \infty$ such that the function $b''(\cdot)$ satisfies the following conditions:
   
   (a) $\sup_{t \in \mathbb{R}} b''(t) \leq U$

   (b) $\inf_{t \in \Theta} b''(t) \geq L$

Similar assumption was imposed in Rigollet (2012). Conditions on $b''(\cdot)$ in Assumption (A) are intended to exclude two degenerate cases, where the variance $\text{Var}(Y)$ is infinitely large or small. They also ensure strong convexity of $b(\cdot)$ over $\Theta$. For Gaussian distribution, $b''(\theta) = 1$ and, therefore, $L = U = 1$ for any $\Theta$. For the binomial distribution, $b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2}$, $U = \frac{1}{4}$, while $L$ depends on $\Theta$.

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

$$KL(\theta, \zeta) = \mathbb{E}_\theta \left\{ \ln \left( \frac{f_{\theta}(Y)}{f_{\zeta}(Y)} \right) \right\} = \frac{1}{a} \mathbb{E}_\theta \left\{ \sum_{i=1}^n Y_i(\theta_i - \zeta_i) - b(\theta_i) + b(\zeta_i) \right\}$$

$$= \frac{1}{a} \sum_{i=1}^n \left( b'(\theta_i)(\theta_i - \zeta_i) - b(\theta_i) + b(\zeta_i) \right) = \frac{1}{a} (b'(\theta)^t(\theta - \zeta) - (b(\theta) - b(\zeta))^t 1)$$
The goodness of an estimator \( \widehat{\theta} \) for the unknown \( \theta \) is measured by the expected Kullback-Leibler divergence between the true distribution \( f_{\theta} \) of the data and the resulting empirical distribution \( f_{\widehat{\theta}} \) generated by \( \widehat{\theta} \):

\[
\mathbb{E} KL(\theta, \widehat{\theta}) = \frac{1}{a} \left( \mathbf{b}'(\theta)^\dagger (\theta - \mathbb{E}(\theta)) - (\mathbf{b}(\theta) - \mathbb{E}b(\widehat{\theta}))' \mathbf{1} \right)
\]

In particular, for the Gaussian case, where \( \mathbf{b}(\theta) = \theta^2 / 2 \) and \( a = \sigma^2 \), \( \mathbb{E} KL(\theta, \widehat{\theta}) \) is the mean squared error \( E|\widehat{\theta} - \theta|^2 \) divided by the constant \( 2\sigma^2 \).

### 2.2 Penalized maximum likelihood model selection

Consider a GLM [1] with a vector of natural parameters \( \theta \) and the canonical link \( \theta = X\beta \).

For a given model (subset of predictors) \( M \) consider the corresponding maximum likelihood estimator (MLE) \( \widehat{\beta}_M \) of \( \beta \):

\[
\widehat{\beta}_M = \arg \max_{\beta \in \mathcal{B}_M} \ell(\beta) = \arg \max_{\beta \in \mathcal{B}_M} \left\{ \sum_{i=1}^{n} (Y_i (X \beta)_i - b((X \beta)_i)) \right\} = \arg \max_{\beta \in \mathcal{B}_M} \left\{ \mathbf{Y}'X \beta - b(X \beta)' \mathbf{1} \right\},
\]

where \( \mathcal{B}_M = \{ \beta \in \mathbb{R}^p : \beta_j = 0 \text{ if } x_j \text{ does not belong to } M, \text{ and } \beta^t x_i \in \Theta \text{ for all } i = 1, \ldots, n \} \). Note that generally \( \mathcal{B}_M \) depends on a given design matrix \( X \).

The MLE for \( \theta \) is \( \widehat{\theta}_M = X\widehat{\beta}_M \), and the ideally selected model (oracle choice) is then the one that minimizes \( \mathbb{E} KL(\theta, \widehat{\theta}_M) = \frac{1}{a} \left( \mathbf{b}'(\theta)^\dagger (\theta - \mathbb{E}(\theta)) - (\mathbf{b}(\theta) - \mathbb{E}b(\widehat{\theta}_M))' \mathbf{1} \right) \) or, equivalently, \(-\mathbf{b}'(\theta)^\dagger \mathbb{E}(\theta) + \mathbb{E}b(\widehat{\theta}_M)' \mathbf{1} \) over all \( M \). An oracle chosen model depends evidently on the unknown \( \theta \) and can only be used as a benchmark for any available model selection procedure.

Consider instead an empirical analog \( KL(\mathbb{E}[\mathbf{b}]^{-1}(\mathbf{Y}), \widehat{\theta}_M) \) of \( \mathbb{E} KL(\theta, \widehat{\theta}_M) \), where the true \( \mathbb{E} \mathbf{Y} = \mathbf{b}'(\theta) \) is replaced by \( \mathbf{Y} \). A naive approach of minimizing \( KL(\mathbb{E}[\mathbf{b}]^{-1}(\mathbf{Y}), \widehat{\theta}_M) \) yields maximizing \( \mathbf{Y}'\widehat{\theta}_M - b(\widehat{\theta}_M)' \mathbf{1} \) (or, equivalently, maximizing \( \ell(\widehat{\theta}_M) \)) over \( M \) 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 and consider the corresponding penalized maximum likelihood model selection criterion of the form

\[
\widehat{M} = \arg \max_{M} \left\{ \ell(\widehat{\theta}_M) - Pen(|M|) \right\} = \arg \min_{M} \left\{ \frac{1}{a} \left( b(X \widehat{\beta}_M) - \mathbf{Y}'X \widehat{\beta}_M \right) + Pen(|M|) \right\},
\]

where the MLE \( \widehat{\beta}_M \) 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 \( AIC = -2\ell(\widehat{\theta}_M) + 2|M| \) of Akaike (1973), \( BIC = -2\ell(\widehat{\theta}_M) + |M| \ln n \) of Schwarz (1973) and its extended version \( EBIC = -2\ell(\widehat{\theta}_M) + |M| \ln n + 2\gamma|M| \ln p, \ 0 \leq \gamma \leq 1 \) of Chen & Chen (2012) 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 \( RIC = -2\ell(\widehat{\theta}_M) + 2|M| \ln p \) of George & Foster (1994) yields \( Pen(|M|) = |M| \ln p \). Note that all the above penalties increase linearly with a model size \( |M| \).
3 Main results

In this section we investigate theoretical properties of the penalized maximum likelihood model selection procedure proposed in Section 2.2 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 divergence of the resulting estimator for a given $Pen(|M|)$ and then establish its asymptotic minimaxity for a properly chosen penalty.

3.1 General upper bound for the expected Kullback-Leibler divergence

Consider a GLM (1) with the canonical link $\theta = X\beta$ and the natural parameters $\theta_i \in \Theta$ satisfying Assumption (A). Let $r = \text{rank}(X)$. 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 $\beta \in B_M$, where $|M| > r$, there necessarily exists another $\beta^*$ with at most $r$ nonzero entries such that $X\beta = X\beta^*$.

We now present an upper bound for the expected Kullback-Leibler divergence from the true distribution of the data for 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 $\theta = X\beta$ 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

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

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

Fix any $0 < \alpha < 1$ and assume that the complexity penalty $Pen(\cdot)$ in (4) is such that

$$Pen(k) \geq \left(4 + \frac{1}{\alpha}\right) \frac{U}{L} k(A + \sqrt{2L_k + L_k}), \quad k = 1, \ldots, r$$

for some $A > 1$.

Let $\hat{M}$ be a model selected in (4) with $Pen(\cdot)$ satisfying (7) and $\hat{\beta}_{\hat{M}}$ be the corresponding MLE estimator (3) of $\beta$. Then,

$$(1 - \alpha)EKL(\theta, X\hat{\beta}_{\hat{M}}) \leq \inf_{M} \left\{ \inf_{\hat{\beta} \in B_M} KL(\theta, X\hat{\beta}) + Pen(|M|) \right\} + \left(4 + \frac{1}{\alpha}\right) \frac{U}{L} \frac{2A - 1}{2A - 2} S$$

The term $\inf_{\hat{\beta} \in B_M} KL(\theta, X\hat{\beta})$ in (7) can be interpreted as a Kullback-Leibler divergence between a true distribution $f_\theta$ of the data and the family of distributions $\{f_{X\hat{\beta}}, \hat{\beta} \in B_M\}$ 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 $1 \leq k < r$ are the numbers of all possible models of size $k$. 

5
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 $X\beta_M$ and can be still associated with a single (saturated) model.

For Gaussian regression, $\mathbb{E}KL(X\beta, X\tilde{\beta}_M) = \frac{1}{2\sigma^2}E||X\beta - X\tilde{\beta}_M||^2$, min$_{\beta \in B_M} KL(X\beta, X\tilde{\beta}) = \frac{1}{2\sigma^2}||X\beta - X\beta_M||^2$, where $X\beta_M$ is the projection of $X\beta$ 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 Birgé & Massart (2001, 2007). Thus, Theorem 1 essentially extends their results for GLM.

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

1. Constant weights. In this case $L_k = L$ for all $k = 1, \ldots, r$. The condition (5) implies

$$\sum_{k=1}^{r-1} \left( \frac{p}{k} \right) e^{-kL} + e^{-rL} \leq \sum_{k=1}^{p} \left( \frac{p}{k} \right) e^{-kL} = (1 + e^{-L})^p - 1$$

The above sum is bounded by an absolute constant for $L = \ln p.$ It can be easily verified that for $L = \ln p$ and $p \geq 5$, there exists $A > 1$ such that $A + \sqrt{2L} + L \leq 3L$. Consider any $C > 15$. Then, there exists $0 < \alpha < 1$ such that $C = 3(4 + \frac{1}{\alpha})$ and we have

$$(4 + \frac{1}{\alpha}) \frac{U}{L} k(A + \sqrt{2L} + L) \leq 3 \left( 4 + \frac{1}{\alpha} \right) \frac{U}{L} kL = C \frac{U}{L} k \ln p$$

that implies the RIC-type linear penalty

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

(8) in Theorem 1 with $C > 15$.

Note that the AIC criterion corresponding to $\text{Pen}(k) = k$ (see Section 2.2) does not satisfy (6).

2. Variable weights. Note that

$$\sum_{k=1}^{r-1} \left( \frac{p}{k} \right) e^{-kL_k} + e^{-rL_r} \leq \sum_{k=1}^{r-1} \left( \frac{ep}{k} \right) e^{-kL_k} + e^{-rL_r} = \sum_{k=1}^{r-1} e^{-kL_k} + e^{-rL_r}$$

(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 $\text{Pen}(k) \sim C \frac{U}{L} k \ln \left( \frac{pe}{k} \right)$ for $k = 1, \ldots, r - 1$ and $\text{Pen}(r) \sim C \frac{U}{L} r$ for some constant $C$.

More precisely, for any $C > 20$ there exist constants $\bar{C}, A > 1$ and $0 < \alpha < 1$ such that $C = 4A \bar{C}$ $(4 + \frac{1}{\alpha})$. Define $L_k = \bar{C} \ln \left( \frac{pe}{k} \right)$, $k = 1, \ldots, r - 1$ and $L_r = \bar{C}$. From (9) one can easily verify the condition (5) for such weights $L_k$. Furthermore, for $1 \leq k \leq r - 1$ we have

$$\left( 4 + \frac{1}{\alpha} \right) \frac{U}{L} k(A + \sqrt{2L_k} + L_k) \leq \left( 4 + \frac{1}{\alpha} \right) \frac{U}{L} kA(1 + \sqrt{L_k})^2 \leq \left( 4 + \frac{1}{\alpha} \right) \frac{U}{L} kA \frac{4L_k}{c}$$

$$= C \frac{U}{L} \ln \left( \frac{pe}{k} \right)$$

6
and similarly, for \( k = r \),

\[
\left( 4 + \frac{1}{\alpha} \right) \frac{U}{L} r \left( A + \sqrt{2Lr} + Lr \right) \leq C \frac{U}{L} r
\]

The corresponding (nonlinear) penalty in (9) is therefore

\[
Pen(k) = C \frac{U}{L} k \ln \left( \frac{p \epsilon}{k} \right), \quad k = 1, \ldots, r - 1 \quad \text{and} \quad Pen(r) = C \frac{U}{L} r,
\]

where \( C > 20 \). For Gaussian regression such \( k \ln \frac{p \epsilon}{k} \)-type penalties were considered in Birgé & Massart (2001, 2007), Bunea et al. (2007), Abramovich & Grinshtein (2010) and Rigollet & Tsybakov (2011).

The choice of \( C > 15 \) in (8) and \( C > 20 \) in (10) was mostly motivated by simplicity of calculus and these lower bounds may possibly be reduced by more accurate analysis.

### 3.2 Risk bounds for sparse models

Theorem 1 established a general upper bound for the expected Kullback-Leibler divergence of the selected model from a true model without any assumption on its size. 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 \( 1 \leq p_0 \leq r \), consider a set of models of size at most \( p_0 \). Obviously, \(|M| \leq p_0\) iff the \( l_0 \) (quasi)-norm of regression coefficients \( ||\beta||_0 \leq p_0 \), where \( ||\beta||_0 \) is the number of nonzero entries.

Define \( B(p_0) = \{ \beta \in \mathbb{R}^p : \beta' x_i \in \Theta \text{ for all } i = 1, \ldots, n, \text{ and } ||\beta||_0 \leq p_0 \} \).

Consider a GLM with the canonical link \( \theta = X\beta \) under Assumption (A), where \( \beta \in B(p_0) \). We refine the general upper bound (7) for a penalized maximum likelihood estimator (4) with a RIC-type linear penalty (8) and a nonlinear \( k \ln \frac{p \epsilon}{k} \)-type penalty (10) considered in Section 3.1 for models within \( B(p_0) \).

Apply the general upper bound (7) with \( \alpha \) and \( A \) corresponding to the chosen constant \( C \) in (8) and (10) (see Section 3.1), and the true \( \tilde{\theta} = X\tilde{\beta}, \tilde{\beta} \in B(p_0) \) in the RHS. For both penalties, we then have

\[
\sup_{\beta \in B(p_0)} \mathbb{E} KL(X\beta, X\tilde{\beta}_{\tilde{M}}) \leq \frac{1}{1 - \alpha} \left\{ Pen(p_0) + \left( 4 + \frac{1}{\alpha} \right) \frac{U}{L} \frac{2A - 1}{2A - 2} S \right\} \leq C_1 Pen(p_0)
\]

for some constant \( C_1 > 1 \) not depending on \( p_0, p \) and \( n \).

Thus, for the RIC-type penalty (8), (11) yields \( \sup_{\beta \in B(p_0)} \mathbb{E} KL(X\beta, X\tilde{\beta}_{\tilde{M}}) = O(p_0 \ln p) \), while for the nonlinear \( k \ln \frac{p \epsilon}{k} \)-type penalty (10) the expected Kullback-Leibler divergence is of a smaller order \( O \left( p_0 \ln \left( \frac{p \epsilon}{p_0} \right) \right) \). 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

\[
\sup_{\beta \in B(p_0)} \mathbb{E} KL(X\beta, X\tilde{\beta}_{\tilde{M}}) \leq \sup_{\beta \in B(r)} \mathbb{E} KL(X\beta, X\tilde{\beta}_{\tilde{M}}) \leq C_1 Pen(r) = O(r)
\]

for some constant \( C_1 > 1 \).
and the final upper bound for an estimator with the penalty \( U \) is, therefore,

\[
C_1 \frac{U}{L} \min \left( p_0 \ln \frac{pe}{p_0}, r \right)
\]  

with \( C_1 > 1 \).

To assess the accuracy of the upper bound \( \text{(13)} \), we establish the corresponding lower bound for the minimax expected Kullback-Leibler divergence over \( B(p_0) \).

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

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

\[
\phi_{\max}[k] = \max_{\beta : 1 \leq ||\beta||_0 \leq k} \frac{||X\beta||^2}{||\beta||^2}.
\]

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 \( X^tX \) 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 \( \theta = X\beta \) under Assumptions (A).

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

\[
\inf_{\hat{\theta}} \sup_{\beta \in B(p_0)} EKL(X\beta, \hat{\theta}) \geq \begin{cases} 
C_2 \frac{\tau[2p_0]}{\tau[p_0]} p_0 \ln \left( \frac{pe}{p_0} \right), & 1 \leq p_0 \leq r/2 \\
C_2 \frac{\tau[p_0]}{\tau[p_0]} r, & r/2 \leq p_0 \leq r 
\end{cases}
\]  

where the infimum is taken over all estimators \( \hat{\theta} \) of \( \theta \).

### 3.3 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 \( r_n = \text{rank}(X_{n,p_n}) \). 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 \( \beta_p \) and sets \( B_p(p_0) \). 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), \( \theta_i \in \Theta, \theta = X_p\beta_p \) and \( \text{rank}(X_p) = r \to \infty \).

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, Abramovich & Grinshtein (2010, Remark 1) showed 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 expected Kullback-Leibler divergence from the true model over \( \mathcal{B}_p(p_0) \) is of the order \( p_0 \ln \left( \frac{pe}{p_0} \right) \) 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 \( \frac{U}{L} \) such that for all sufficiently large \( r \),

\[
C_1 p_0 \ln \left( \frac{pe}{p_0} \right) \leq \inf_{\hat{\theta}} \sup_{\beta_p \in \mathcal{B}_p(p_0)} \mathbb{E}KL(X_p \beta_p, \hat{\theta}) \leq C_2 p_0 \ln \left( \frac{pe}{p_0} \right)
\]

for all \( 1 \leq p_0 \leq r \).

2. Consider penalized maximum likelihood model selection rule \( \text{Pen}(k) = C \frac{U}{L} k \ln \left( \frac{pe}{k} \right) \), \( k = 1, \ldots, r - 1 \) and \( \text{Pen}(r) = C \frac{U}{L} r \), where \( C > 20 \). Then, the resulting penalized MLE estimator \( X_p \hat{\beta}_{pM} \) attains the minimax convergence rates (in terms of \( \mathbb{E}KL(X_p \beta_p, X_p \hat{\beta}_{pM}) \)) simultaneously over all \( \mathcal{B}_p(p_0) \), \( 1 \leq p_0 \leq r \).

Corollary [1] is a generalization of the corresponding results of Abramovich & Grinshtein (2010) for Gaussian regression. It also shows that model selection criteria with RIC-type (linear) penalties \( \text{Pen}(k) \) of the form \( \text{Pen}(k) = C k \ln p \) are of the minimax order for sparse models with \( p_0 \ll p \) but only suboptimal otherwise.

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

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

**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 \( \beta \), may be a “blessing” for estimating \( \theta = X \beta \). For Gaussian regression Abramovich & Grinshtein (2010) 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 Abramovich & Grinshtein, 2010), and we do not discuss its extension for GLM in this paper.

4 Possible extensions

In this section we discuss some possible extensions of the results obtained in Section \(3\).

4.1 Model selection in GLM under structural constraints

So far we considered the complete variable selection, where the set of admissible models contains all \(2^p\) possible subsets of predictors \(x_1, \ldots, x_p\). 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 \(x_1, \ldots, x_{j-1}\) (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 Abramovich & Grinshtein (2013). 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 \(1 \leq p_0 \leq r\), where \(m(r) = 1\) if there are admissible models of size \(r\). Obviously, \(0 \leq m(p_0) \leq \binom{p}{p_0}\), where the two extreme cases \(m(p_0) = 1\) and \(m(p_0) = \binom{p}{p_0}\) for all \(p_0 = 1, \ldots, r - 1\), correspond respectively to the ordered and complete variable selection.

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

\[
\sum_{k=1}^{r-1} m(k)e^{-kL_k} + e^{-rL_k} \leq S
\]

and for any fixed \(0 < \alpha < 1\)

\[
Pen(k) \geq \left(4 + \frac{1}{\alpha}\right) \frac{U}{L} k(A + \sqrt{2L_k + L_k}), \quad k = 1, \ldots, r; \quad m(k) \geq 1
\]
for some $A > 1$. Then, under Assumption (A)

$$(1 - \alpha)\mathbb{E} KL(\theta, X\hat{\beta}_M) \leq \inf_{M \in \mathfrak{M}} \left\{ \inf_{\beta \in \mathcal{B}_M} KL(\theta, X\beta) + Pen(|M|) \right\} + \left( 4 + \frac{1}{\alpha} \right) \frac{U}{L} \frac{2A - 1}{2A - 2} S.$$  \hspace{1cm} (16)

See Birgé & Massart (2001, 2007), Abramovich & Grinshtein (2013) for similar results for Gaussian regression under structural constraints.

In particular, (15) holds for $L_k = c_1 k \max(\ln m(k), k)$, $k = 1, \ldots, r$; $m(k) \geq 1$ for some $c > 1$ leading to the penalty of the form

$$Pen(k) \sim \frac{U}{L} \max(\ln m(k), k)$$ \hspace{1cm} (17)

for all $1 \leq k \leq r$ such that $m(k) \geq 1$. For the complete variable selection, the penalty (17) is evidently the $k \ln k$-type penalty (10) from Section 3, while for the ordered variable selection it implies the AIC-type penalty of the form $Pen(k) = C U k$ 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 3.2 for the complexity penalty (17), under Assumption (A), the general upper bound (16) yields

$$\sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} KL(X\beta, X\hat{\beta}_M) = O(\max(\ln m(p_0), p_0))$$ \hspace{1cm} (18)

with a constant depending on the ratio $U/L$.

The upper bound (18) 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

$$\sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} KL(X\beta, X\hat{\beta}_M) = O(Pen(r)) = O(r)$$

that combining with (18) yields

$$\sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} KL(X\beta, X\hat{\beta}_M) = O(\min(\max(\ln m(p_0), p_0), r))$$  \hspace{1cm} (19)

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

$$\inf_{\hat{\theta}} \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} KL(X\beta, \tilde{\theta}) \geq \begin{cases} C_2 \frac{1}{r} \max \left\{ \tau \cdot \frac{\ln m(p_0)}{\ln p_0}, \tau \cdot [p_0]\right\}, & 1 \leq p_0 \leq r/2 \\ C_2 \frac{1}{r} \tau, & r/2 \leq p_0 \leq r \end{cases}$$ \hspace{1cm} (20)

for some $C_2 > 0$.

Thus, comparing the upper bounds (18)–(19) with the lower bound (20) one realizes that for weakly collinear design the proposed penalized maximum likelihood estimator with the complexity penalty of type (17) 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 3 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 (20) for a general case as well. See also Abramovich & Grinshtein (2013) for similar results for Gaussian regression.

4.2 Aggregation in GLM

An interesting statistical problem related to model selection is aggregation. Originated by Nemirovski (2000), it has been intensively studied in the literature during the last decade. See, for example, Tsybakov (2003), Young (2004), Leung & Barron (2006), Bunea et al. (2007) and Rigollet & Tsybakov (2011) for aggregation in Gaussian regression. Aggregation in GLM was considered in Rigollet (2012) and can be described as follows.

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

$$\inf_{\beta \in \mathcal{B}} KL(\theta, \theta_\beta)$$

Depending on the specific choice of $\mathcal{B} \subseteq \mathbb{R}^p$ there are different aggregation strategies. Following the terminology of Bunea et al. (2007) there are linear aggregation ($\mathcal{B} = \mathcal{B}_L = \mathbb{R}^p$), convex aggregation ($\mathcal{B} = \mathcal{B}_C = \{\beta \in \mathbb{R}^p: \beta_j \geq 0, \sum_{j=1}^{p} \beta_j = 1\}$), 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 ($\mathcal{B} = \mathcal{B}_{SS}(p_0) = \{\beta \in \mathbb{R}^p: ||\beta||_0 \leq p_0\}$ 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 $\theta$ is unknown, the solution of (21) is unavailable. The goal then is to construct an estimator (linear aggregator) $\hat{\theta}_\beta$ that mimics the ideal (oracle) solution $\theta_\beta$ of (21) as close as possible. More precisely, we would like to find $\hat{\theta}_\beta$ such that

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

(22)

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

For weakly collinear design, Rigollet (2012, Theorem 4.1) established the minimal possible asymptotic rates for $\Delta_\mathcal{B}(\theta, \theta_\beta)$ for linear, convex and model selection aggregation under Assump-
tion (A) and assumptions similar to those of Theorem 2:

\[
\inf_{\bar{\theta}} \sup_{\theta} \Delta_{B}(\theta, \bar{\theta}) = \begin{cases} 
O(r) & B = B_L \text{ (linear aggregation)} \\
O(\min(\sqrt{n\ln p}, r)) & B = B_C \text{ (convex aggregation)} \\
O(\min(\ln p, r)) & B = B_{MS} \text{ (model selection aggregation)}
\end{cases}
\]  

(23)

He also proposed an estimator \( \hat{\theta} \beta \) that achieves these optimal aggregation rates even with \( C = 1 \) in (22).

Using the results of Section 3 we can complete the case of subset selection aggregation in GLM, where under the assumptions of Theorem 4.1 of Rigollet (2012), \( B_{SS}(p_0) \) is essentially \( B(p_0) \) 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 \( B(p_0) \) there exists \( C_2 > 0 \) such that

\[
\inf_{\bar{\theta}} \sup_{\theta} \Delta_{B(p_0)}(\theta, \bar{\theta}) \geq C_2 L \min \left( \frac{\ln(p_0)}{p_0}, r \right)
\]  

(24)

In particular, (24) also yields the lower bounds (23) for excess-KL for linear (\( p_0 = r \)) and model selection (\( p_0 = 1 \)) aggregation. Furthermore, similar to model selection in GLM within \( B(p_0) \) considered in Section 3.2 from Theorem 1, it follows that for weakly collinear design, the penalized maximum likelihood estimator \( \hat{\theta} \beta \) with the complexity penalty (10) achieves the optimal rate (24) for subset selection aggregation over \( B(p_0) \) for all \( 1 \leq p_0 \leq r \) (and, therefore, for linear and model selection aggregation in particular) though with some \( C > 1 \) in (22).

Acknowledgement

The work was supported by the Israel Science Foundation (ISF), grant ISF-820/13. We are grateful to Alexander Goldenshluger and Ya’acov Ritov for valuable remarks.

Appendix

We first prove the following lemma establishing the equivalence of the Kullback-Leibler divergence \( KL(\theta_1, \theta_2) \) and the squared quadratic norm \( ||\theta_1 - \theta_2||^2 \) under Assumption (A) that will be used further in the proofs:

**Lemma 1.** Let Assumption (A) hold. Then, for any \( \theta_1, \theta_2 \in \mathbb{R}^n \) such that \( \theta_{1i}, \theta_{2i} \in \Theta, i = 1, \ldots, n, \)

\[
\frac{L}{2a} ||\theta_1 - \theta_2||^2 \leq KL(\theta_1, \theta_2) \leq \frac{U}{2a} ||\theta_1 - \theta_2||^2
\]

**Proof.** Recall that for a GLM

\[
KL(\theta_1, \theta_2) = \frac{1}{a} \sum_{i=1}^{n} \{ u'(\theta_{1i})(\theta_{1i} - \theta_{2i}) - b(\theta_{1i}) + b(\theta_{2i}) \}
\]  

(25)
A Taylor expansion of \( b(\theta_{2i}) \) around \( \theta_{1i} \) 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 \( \theta_{1i} \) and \( \theta_{2i} \), and substituting into (25) we have

\[
KL(\theta_1, \theta_2) = \frac{1}{2a} \sum_{i=1}^n b''(c_i)(\theta_{2i} - \theta_{1i})^2
\]

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

**Proof of Theorem 1**

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

\[
\beta_M = \arg \inf_{\beta \in \mathcal{B}_M} KL(\theta, X\beta),
\]

where \( \mathcal{B}_M \) is given in (3), and let \( \theta_M = X\beta_M \). As we have mentioned in Section 3.1, \( \theta_M \) can be interpreted as the closest vector to \( \theta \) within the span generated by a subset of columns of \( X \) corresponding to \( M \) w.r.t. a Kullback-Leibler divergence. Let also \( P_M \theta \) be an orthogonal projection of \( \theta \) on this span (evidently, \( \theta_M = P_M \theta \) for Gaussian case) and recall that \( \hat{\theta}_M = X\hat{\beta}_M \) is the MLE of \( \theta \) for the model \( M \). In particular, \( \hat{\theta}_{\hat{M}} = X\hat{\beta}_{\hat{M}} \). Finally, for any random variable \( \eta \) let \( \varphi_\eta(\cdot) \) be its moment generating function.

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

**Step 1.** Since \( \hat{M} \) is the minimizer defined in (1), for any given model \( M \)

\[
- \ell(\hat{\beta}_M) + Pen(|\hat{M}|) \leq -\ell(\beta_M) + Pen(|M|) \tag{26}
\]

By a straightforward calculus, one can easily verify that

\[
KL(\theta, \hat{\beta}_\hat{M}) - KL(\theta, \theta_M) = -\ell(\beta_M) - \ell(\hat{\beta}_\hat{M}) + \frac{1}{a}(Y - b'(\theta))^t(\hat{\beta}_\hat{M} - \theta_M) \tag{27}
\]

and, hence, (26) yields

\[
KL(\theta, \hat{\beta}_\hat{M}) + Pen(|\hat{M}|) \leq KL(\theta, \theta_M) + Pen(|M|) + \frac{1}{a}(Y - b'(\theta))^t(\hat{\beta}_\hat{M} - \theta_M) \tag{28}
\]

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

\[
\mathbb{E}\left((Y - b'(\theta))^t(\hat{\beta}_\hat{M} - \theta_M)\right) = \mathbb{E}\left((Y - b'(\theta))^t(\hat{\beta}_\hat{M} - \theta)\right)
\]

Furthermore, by the definition of \( \hat{\beta}_\hat{M} \), \( KL(\theta, \hat{\beta}_\hat{M}) \geq KL(\theta, \theta_{\hat{M}}) \), and since (28) holds for any model \( M \) in the RHS, we have

\[
(1 - \alpha)\mathbb{E}KL(\theta, \hat{\beta}_\hat{M}) \leq \inf_M \{KL(\theta, \theta_M) + Pen(|M|)\}
\]

\[
+ \mathbb{E}\left(\frac{1}{a}(Y - b'(\theta))^t(\hat{\beta}_\hat{M} - \theta) - Pen(|\hat{M}|) - \alpha KL(\theta, \theta_{\hat{M}})\right) \tag{29}
\]
for any $0 < \alpha < 1$.

**Step 2.** Consider now the term $\frac{1}{a}(Y - b'(\theta))^t(\hat{\theta}_M - \theta)$ in the RHS of (29). The selected model $\hat{M}$ in (13) can, in principle, be any model $M$ and we want, therefore, to control it uniformly over $M$. For any $M$ we have

$$\frac{1}{a}(Y - b'(\theta))^t(\hat{\theta}_M - \theta) = \frac{1}{a}(Y - b'(\theta))^t(\hat{\theta}_M - \theta_M) + \frac{1}{a}(Y - b'(\theta))^t(\theta_M - P_M\theta)$$

$$+ \frac{1}{a}(Y - b'(\theta))^t(P_M\theta - \theta)$$

(30)

Let $\xi_M$ be the projection of $Y - b'(\theta)$ on the span of columns of $X$ corresponding to the model $M$. Then, by the Cauchy-Schwarz inequality

$$(Y - b'(\theta))^t(\hat{\theta}_M - \theta_M) = \xi_M^t(\hat{\theta}_M - \theta_M) \leq ||\xi_M|| \cdot ||\hat{\theta}_M - \theta_M||$$

(31)

Since $\hat{\theta}_M$ is the MLE for a given $M$, $\ell(\hat{\theta}_M) \geq \ell(\theta_M)$ and, therefore, (27) implies

$$KL(\theta, \hat{\theta}_M) \leq KL(\theta, \theta_M) + \frac{1}{a}(Y - b'(\theta))^t(\hat{\theta}_M - \theta_M)$$

(32)

Similar to the proof of Lemma 6.3 of Rigollet (2012), using a Taylor expansion it follows that under Assumption (A), $KL(\theta, \hat{\theta}_M) - KL(\theta, \theta_M) \geq \frac{c_L}{aL}||\hat{\theta}_M - \theta_M||^2$ that together with (31) and (32) yields

$$\frac{1}{a}(Y - b'(\theta))^t(\hat{\theta}_M - \theta_M) \leq \frac{2}{aL}||\xi_M||^2$$

(33)

Similarly to (31),

$$(Y - b'(\theta))^t(\theta_M - P_M\theta) \leq ||\xi_M|| \cdot ||\theta_M - P_M\theta||$$

(34)

and using the inequality $2c_1c_2 \leq \delta c_1^2 + \frac{1}{\delta}c_2^2$ for all $c_1, c_2$ and $\delta > 0$, from (34) we have

$$\frac{1}{a}(Y - b'(\theta))^t(\theta_M - P_M\theta) \leq \frac{\delta}{2a}||\xi_M||^2 + \frac{1}{2a\delta}||\theta_M - P_M\theta||^2$$

(35)

for all $\delta > 0$.

Hence, combining (30), (33) and (35) implies

$$\frac{1}{a}(Y - b'(\theta))^t(\hat{\theta}_M - \theta) \leq \frac{c}{2a}||\xi_M||^2 + \frac{1}{a}(Y - b'(\theta))^t(P_M\theta - \theta) + \frac{1}{2a\delta}||\theta_M - P_M\theta||^2,$$

(36)

where $c = (\frac{1}{2} + \delta)$.

Define

$$Q(M) = Pen(||M||) + \alpha KL(\theta, \theta_M) - \frac{1}{2a\delta}||\theta_M - P_M\theta||^2$$

and

$$R(M) = \frac{c}{2a}||\xi_M||^2 + \frac{1}{a}(Y - b'(\theta))^t(P_M\theta - \theta) - Q(M)$$
Then, from (29) and (30),

\[
(1 - \alpha)\mathbb{E} KL(\theta, \tilde{\theta}_M) \leq \inf_M \{ KL(\theta, \theta_M) + Pen(|M|) \} + \mathbb{E} R(\tilde{M})
\]

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

\textbf{Step 3.} Consider \( \varphi||\xi_M||^2(\cdot) \). Let \( \zeta_j, \ j = 1, \ldots, |M| \) be any orthonormal basis of the span of columns of \( X \) corresponding to the model \( M \) and \( \xi_j = (Y - b'(\theta))^t \zeta_j \) be the coordinates of \( \xi_M \) in this basis. We have

\[
\varphi(\xi_j^2(s)) = \mathbb{E} e^{s \xi_j^2} = \sum_{h=0}^{\infty} \frac{\mathbb{E} \xi_j^{2h}}{h!} s^h
\]

Applying Chernoff bound for \( h > 0 \) yields

\[
\mathbb{E} \xi_j^{2h} = \int_0^\infty P(\xi_j^{2h} > t) dt \leq \int_0^\infty \exp \left\{ - \sup_{u > 0} (ut^{1/(2h)} - \ln \varphi(\xi_j(u))) \right\} dt,
\]

where by (6.3) of Rigollet (2012)

\[
\ln \varphi(\xi_j(u)) = \ln \mathbb{E} e^{u(\gamma - b'(\theta))^t \zeta_j} \leq \frac{1}{2a} u^2 u^2
\]

Thus, after straightforward calculus

\[
\mathbb{E} \xi_j^{2h} \leq \int_0^\infty \exp \left\{ - \frac{t^{1/h}}{2aU} \right\} dt = h!(2aU)^h
\]

Hence, (38) implies

\[
\varphi(\xi_j^2(s)) \leq 1 + \sum_{h=1}^{\infty} (2aU)^h = \frac{1}{1 - 2aUs}
\]

and, therefore,

\[
\varphi||\xi_M||^2(s) \leq \frac{1}{(1 - 2aUs)^{|M|}}
\]

for all \( 0 < s < \frac{1}{2aU} \).

Consider the random variable \( \eta_M = (Y - b'(\theta))^t (P_M \theta - \theta) \). Applying (6.3) in Lemma 6.1 of Rigollet (2012) yields

\[
\varphi(\eta_M(s)) \leq \exp \left\{ \frac{1}{2} s^2 Ua||P_M \theta - \theta||^2 \right\}
\]

In addition, \( ||\xi_M||^2 = \sum_{j=1}^{|M|} (\gamma - b'(\theta))^t \zeta_j^2 \) and \( \eta_M = (Y - b'(\theta))^t (P_M \theta - \theta) \) are uncorrelated since \( \zeta_j \)'s belong to the span of columns of \( X \) corresponding to \( M \), while \( P_M \theta - \theta \) is orthogonal to it. Hence, defining \( Z = \frac{1}{2\pi} ||\xi_M||^2 - 2a|M|U + \frac{1}{2} \eta_M = R(M) + Q(M) - c|M|U \) we have

\[
\varphi_Z(s) = e^{-|M|cUs} \cdot \varphi||\xi_M||^2 \left( \frac{cs}{2a} \right) \cdot \varphi(\eta_M) \left( \frac{s}{a} \right) \leq \frac{1}{(1 - cUs)^{|M|}} \cdot \exp \left\{ -|M|cUs + \frac{Us^2}{2a} ||P_M \theta - \theta||^2 \right\}
\]

for all \( 0 < s < \frac{1}{cU} \).
Let \( x = c \mathcal{U} s \) \((0 < x < 1)\) and \( \rho = \frac{||P_M \theta - \theta||^2}{2ac\mathcal{U}} \). Then, from (39)

\[
\ln \varphi_Z(s) \leq |M|(-\ln(1-x) - x) + \rho x^2 \leq \frac{|M|x^2}{2(1-x)} + \rho x^2 \leq \frac{|M|x^2}{2(1-x)} + \rho x
\]

and, therefore,

\[
\ln \frac{\varphi_{\mathcal{U}}}{\mathcal{U} - \rho}(x) \leq \frac{|M|}{2} \frac{x^2}{1-x}
\]

for all \( 0 < x < 1 \).

We can now apply Lemma 2 of Birgé & Massart (2007) to get \( P(\frac{Z}{\mathcal{U}} - \rho \geq \sqrt{2|M|t + t}) \leq e^{-t} \) for all \( t > 0 \), that is,

\[
P\left\{ \frac{1}{c \mathcal{U}} \left( R(M) + Q(M) - \frac{||P_M \theta - \theta||^2}{2ac} \right) \geq |M| + \sqrt{2|M|t + t} \right\} \leq e^{-t} \tag{40}
\]

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

Consider the term \( Q(M) - \frac{||P_M \theta - \theta||^2}{2ac} \) in the LHS of (40), where recall that \( c = \frac{4}{L^2} + \delta > \delta \). We have

\[
Q(M) - \frac{1}{2ac} ||P_M \theta - \theta||^2 = Pen(|M|) + \alpha K L (\theta, \theta_M) - \frac{1}{2a\delta} ||\theta_M - P_M \theta||^2 - \frac{1}{2ac} ||P_M \theta - \theta||^2
\]

\[
> Pen(|M|) + \alpha K L (\theta, \theta_M) - \frac{1}{2a\delta} (||\theta_M - P_M \theta||^2 + ||P_M \theta - \theta||^2)
\]

\[
= Pen(|M|) + \alpha K L (\theta, \theta_M) - \frac{1}{2a\delta} ||\theta_M - \theta||^2
\]

\[
\geq Pen(|M|) + \frac{1}{2a} (\alpha L - \frac{1}{\delta}) ||\theta_M - \theta||^2,
\]

where the last inequality follows from Lemma 11. Hence, for any \( \delta \geq \frac{1}{a2} \), \( Q(M) - \frac{1}{2ac} ||P_M \theta - \theta||^2 \geq Pen(|M|) \) and from (40),

\[
P\left\{ \frac{1}{c \mathcal{U}} \left( R(M) + Pen(|M|) \right) \geq |M| + \sqrt{2|M|t + t} \right\} \leq e^{-t}
\]

Let \( k = |M| \) and take \( t = kL_k + \omega \) for any \( \omega > 0 \), where \( L_k > 0 \) are the weights from Theorem 14.

Using inequalities \( \sqrt{c_1 + c_2} \leq \sqrt{c_1} + \sqrt{c_2} \) and \( c_1 c_2 \leq \frac{1}{2} (c_1 + c_2) \) for any positive \( c_1, c_2 \) and \( \epsilon \), we have

\[
\sqrt{2kt} \leq k \sqrt{2L_k} + \sqrt{2k\omega} \leq k \sqrt{2L_k} + k\epsilon + \frac{\omega}{2\epsilon}
\]

and, therefore,

\[
P\left\{ \frac{1}{c \mathcal{U}} \left( R(M) + Pen(k) \right) \geq k \left( 1 + \epsilon + \sqrt{2L_k} + L_k \right) + \omega \left( 1 + \frac{1}{2\epsilon} \right) \right\} \leq e^{-(kL_k + \omega)}
\]

Take the smallest possible \( \delta = \frac{1}{a2} \) that yields \( c = (4 + \frac{1}{a2}) \frac{1}{L^2} \). For the penalty \( Pen(k) \) satisfying (9) with some \( A > 1 \) and \( \epsilon = A - 1 \), we then have

\[
P\left\{ \frac{1}{c \mathcal{U}} R(M) \geq \omega \left( 1 + \frac{1}{2\epsilon} \right) \right\} \leq e^{-(kL_k + \omega)} \tag{41}
\]
for all $M$.

Finally, under the condition (5) on the weights $L_k$, (41) implies
\[
P \left\{ R(\hat{M}) \geq c \mathcal{U} \omega \left( 1 + \frac{1}{2\epsilon} \right) \right\} \leq \sum_M P \left\{ R(M) \geq c \mathcal{U} \omega \left( 1 + \frac{1}{2\epsilon} \right) \right\} \leq \sum_M e^{- (KL_1 + \omega)} \leq \mathcal{S}e^{-\omega}
\]
and, hence,
\[
\mathbb{E} R(\hat{M}) \leq \int_0^\infty P(R(\hat{M}) > t)dt \leq c \mathcal{U} \left( 1 + \frac{1}{2\epsilon} \right) S = \left( 4 + \frac{1}{\alpha} \right) \frac{\mathcal{U}}{c} \frac{2A - 1}{2A - 2} S
\]
that together with (37) completes the proof. \hfill \Box

**Proof of Theorem 2**

Due to Lemma 1 the minimax lower bound for the Kullback-Leibler divergence can be reduced to the lower bound for the corresponding quadratic risk:
\[
\inf_{\theta} \sup_{\beta \in \mathcal{B}(\theta)} \mathbb{E} KL(X\beta, \bar{\theta}) \geq \frac{C}{2\alpha} \inf_{\theta} \sup_{\beta \in \mathcal{B}(\theta)} \mathbb{E}\|X\beta - \bar{\theta}\|^2
\]
(42)

The idea of the proof now is to find a subset $\mathcal{B}^*(p_0) \subseteq \mathcal{B}(p_0)$ of vectors $\beta$ and the corresponding subset $\Theta^*(p_0) = \{ \theta \in \mathbb{R}^n : \theta = X\beta, \beta \in \mathcal{B}^*(p_0) \}$ such that for any $\theta_1, \theta_2 \in \Theta^*(p_0)$, $||\theta_1 - \theta_2||^2 \geq 4s^2(p_0)$ and $KL(\theta_1, \theta_2) \leq (1/16) \ln \text{card}(\Theta^*(p_0))$. It will follow then from Lemma A.1 of Bunea et al. (2007) that $s^2(p_0)$ is the minimax lower bound for the quadratic risk over $\mathcal{B}(p_0)$.

To construct such subsets of vectors we can exploit the techniques similar to that used in the corresponding proofs for the quadratic risk in linear regression (e.g., Abramovich & Grinshtein, 2010; Rigollet & Tsybakov, 2011). Consider three cases.

**Case 1.** $p_0 \leq r/2$

Define the subset $\tilde{\mathcal{B}}(p_0)$ of all vectors $\beta \in \mathbb{R}^p$ that have $p_0$ entries equal to $C_{p_0}$, where $C_{p_0}$ will be defined below and others are zeros: $\tilde{\mathcal{B}}(p_0) = \{ \beta \in \mathbb{R}^p : \beta \in \{0, C_{p_0}\}^p, ||\beta||_0 = p_0 \}$.

From Lemma A.3 of Rigollet & Tsybakov (2011), there exists a subset $\mathcal{B}^*(p_0) \subseteq \tilde{\mathcal{B}}(p_0)$ such that $\ln \text{card}(\mathcal{B}^*(p_0)) \geq \tilde{c}_{p_0} \ln \left( \frac{p_0}{p_{\gamma}} \right)$ for some constant $0 < \tilde{c} < 1$, and for any pair $\theta_1, \theta_2 \in \mathcal{B}^*(p_0)$, the Hamming distance $\rho(\theta_1, \theta_2) = \sum_{j=1}^p I\{ \theta_{1j} \neq \theta_{2j} \} \geq \tilde{c}_{p_0}.$

Take $C_{p_0}^2 = \frac{1}{16} \tilde{c}_{p_0}^{-1} \phi_{\min}[2p_0] \ln \left( \frac{p_0}{p_{\gamma}} \right)$. By the assumptions of the theorem, $\mathcal{B}^*(p_0) \subseteq \tilde{\mathcal{B}}(p_0) \subseteq \mathcal{B}(p_0)$. Consider the corresponding subset $\Theta^*(p_0)$. Evidently, $\text{card}(\Theta^*(p_0)) = \text{card}(\mathcal{B}^*(p_0))$, and for any $\theta_1, \theta_2 \in \Theta^*(p_0)$ associated with $\beta_1, \beta_2 \in \mathcal{B}^*(p_0)$ we then have
\[
||\theta_1 - \theta_2||^2 = ||X(\beta_1 - \beta_2)||^2 \geq \phi_{\min}[2p_0] ||\beta_1 - \beta_2||^2 \geq \tilde{c}_{p_0} \phi_{\min}[2p_0] C_{p_0}^2 p_0 = 4s^2(p_0),
\]
(43)

where $s^2(p_0) = \frac{1}{6} \phi_{\max}[2p_0] \ln \left( \frac{p_0}{p_{\gamma}} \right)$.

On the other hand,
\[
K(\theta_1, \theta_2) \leq \frac{\mathcal{U}}{2a} ||\theta_1 - \theta_2||^2 \leq \frac{\mathcal{U}}{2a} \phi_{\max}[2p_0] C_{p_0}^2 \rho(\beta_1, \beta_2) \leq \frac{\mathcal{U}}{a} \phi_{\max}[2p_0] C_{p_0}^2 p_0 \leq \frac{1}{16} \ln \text{card}(\Theta^*(p_0)),
\]
(44)
where the first inequality follows from Lemma A.1 of Bunea et al. (2007) and (42) complete then the proof for this case.

Case 2. \( \frac{r}{2} \leq p_0 \leq r, \quad p_0 \geq 8 \)

In this case consider the subset \( \tilde{B}(p_0) = \{ \beta \in \mathbb{R}^p : \beta \in \{(0, C_{p_0})^{p_0}, 0, \ldots, 0\} \} \), where \( C_{p_0}^2 = \frac{\ln 2}{\mathcal{U} \phi^{-1}} \phi_{\max}[p_0] \). From the assumptions of the theorem \( \tilde{B}(p_0) \subseteq B(p_0) \). Varshamov-Gilbert bound (see, e.g., Tsybakov, 2009, Lemma 2.9) guarantees the existence of a subset \( B^*(p_0) \subset \tilde{B}(p_0) \) such that \( \ln \text{card}(B^*(p_0)) \geq \frac{p_0}{8} \ln 2 \) and the Hamming distance \( \rho(\beta_1, \beta_2) \geq \frac{p_0}{8} \) for any pair \( \beta_1, \beta_2 \in B^*(p_0) \).

Note that for any \( \beta_1, \beta_2 \in B^*(p_0) \), \( \beta_1 - \beta_2 \) has at most \( p_0 \) nonzero components and repeating the arguments for the Case 1, one obtains the minimax lower bound \( s^2(p_0) = C_a \mathcal{U} \tau[p_0] \geq \frac{C}{2} a \mathcal{U} \tau[p_0] r \) for the quadratic risk. Applying (42) completes the proof.

Case 3. \( \frac{r}{2} \leq p_0 \leq r, \quad 2 \leq p_0 < 8 \)

For this case, obviously, \( 2 \leq r < 16 \). Consider a trivial subset \( B^*_{p_0} \) containing just two vectors \( \beta_1 \equiv 0 \) and \( \beta_2 \) that has first \( p_0 \) nonzero entries equal to \( C_{p_0} \), where \( C_{p_0}^2 = \frac{\ln 2}{\mathcal{U} \phi_{\max}[p_0]} \). Under the assumptions of the theorem \( B^*_{p_0} \subset B(p_0) \). For the corresponding \( \theta_1 = X\beta_1 \) and \( \theta_2 = X\beta_2 \), (43) and (44) yield

\[
KL(\theta_1, \theta_2) \leq \frac{\mathcal{U}}{2a} \phi_{\max}[p_0] |C_{p_0}^2| \frac{1}{16} \ln \text{card}(\Theta^*_{p_0})
\]

and

\[
\| \theta_1 - \theta_2 \|^2 \geq \phi_{\min}[p_0] p_0 C_{p_0}^2 = C_a \mathcal{U} \tau[p_0] \geq \frac{C}{2} a \mathcal{U} \tau[p_0] r
\]

and the proof follows from Lemma A.1 of Bunea et al. (2007).

References

[1] Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. *Electr. J. Statist.* 4, 932–949.

[2] Abramovich, F. and Grinshtein, V. (2013). Model selection in regression under structural constraints. *Electr. J. Statist.* 7, 480–498.

[3] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. in *Second International Symposium on Information Theory*. (eds. B.N. Petrov and F. Czákı). Akademiai Kiadó, Budapest, 267–281.

[4] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* 37, 1705–1732.

[5] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* 3, 203–268.

[6] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* 138, 33–73.
[7] Bunea, F., Tsybakov, A. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**, 1674–1697.

[8] Chen, J. and Chen Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika* **94**, 759–771.

[9] Chen, J. and Chen Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica* **22**, 555-574.

[10] Fan, J. and Son, R. (2010). SURE independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–3604.

[11] Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.

[12] Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regression. *IEEE Trans. Inform. Theory* **52**, 3596–3410.

[13] Mallows, C.L. (1973). Some comments on $C_p$. *Technometrics* **15**, 661–675.

[14] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London.

[15] Nemirovski, A. (2000). *Topics in Non-Parametric Statistics*. In Lectures on Probability Theory and Statistics (Saint-Flour, 1998). Lecture Notes in Math. 1738, 85277. Springer, Berlin.

[16] Raskutti, G., Wainwright, M.J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional regression over $l_q$ balls. *IEEE Trans. Inform. Theory* **57**, 6976–6994.

[17] Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40**, 639-665.

[18] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731–771.

[19] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

[20] Tsybakov, A. (2003). Optimal rates of aggregation. *Proceeding of COLT-2003*, Lecture Notes in Artif. Intellig. **2777**, 303–313.

[21] Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

[22] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensionals phenomenon. *Electr. J. Statist.* **6**, 38–90.
[23] YANG, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* 10, 25–47.
Supplementary material for
“Model selection and minimax estimation in generalized linear models”

**Theorem (2’).** Consider a GLM with the canonical link \( \theta = X\beta \) under Assumption (A).

Let \( 1 \leq p_0 \leq r \), \( m(p_0) \geq 1 \). Assume that \( B(p_0) \subseteq B(p_0) \), where \( B(p_0) \) was defined in Section 4.1 and the subsets \( \tilde{B}(p_0) \) will be defined later in the proof.

Then, there exists a constant \( C_2 > 0 \) such that

\[
\inf_{\theta} \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} KL(X\beta, \tilde{\theta}) \geq \begin{cases} 
C_2 \frac{\tau}{\alpha} \max \left\{ \tau[p_0] \ln m(p_0), \tau[p_0] p_0 \right\}, & 1 \leq p_0 \leq r/2 \\
C_2 \frac{\tau}{\alpha} \tau[p_0] r, & r/2 \leq p_0 \leq r 
\end{cases}
\]

\[
(1)
\]

Proof. Throughout the proof we use \( C \) to denote a generic positive constant not necessarily the same each time it is used.

The general scheme of the proof is similar to that of Theorem 2 for complete variable selection. The minimax lower bound for the Kullback-Leibler divergence is reduced to the lower bound for the corresponding quadratic risk by Lemma 1:

\[
\inf_{\theta} \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} KL(X\beta, \tilde{\theta}) \geq \frac{C}{2a} \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} ||X\beta - \tilde{\theta}||^2
\]

The goal then is to find the subsets of regression coefficients \( B^*(p_0) \subseteq B(p_0) \) and the corresponding subsets \( \Theta^*(p_0) \) suitable for applying Lemma A.1 of Bunea et al. (2007).

We first show that

\[
\inf_{\theta} \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} ||X\beta - \tilde{\theta}||^2 \geq C \frac{\alpha}{U} \tau[p_0] p_0
\]

\[
(2)
\]

for all \( 1 \leq p_0 \leq r \).

Consider any particular admissible model \( M(p_0) \) of size \( p_0 \). Evidently,

\[
\inf_{\theta} \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} ||X\beta - \tilde{\theta}||^2 \geq \inf_{\theta} \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E} ||X\beta - \tilde{\theta}||^2
\]

which is essentially a minimax lower bound for the quadratic risk for a fixed GLM of size \( p_0 \).

For \( p_0 \geq 8 \), define the subset \( \mathcal{B}(p_0) \) that includes \( 2^{p_0} \) vectors \( \beta \in \mathcal{B}_M(p_0) \) with the corresponding \( p_0 \) nonzero entries equal to \( \pm C_{p_0} \), where \( C_{p_0}^2 = \frac{\ln 2}{256 \alpha} \phi^{-1}_{\max} [p_0] \). By Varshamov-Gilbert bound (see, e.g., Lemma 2.9 of Tsybakov, 2009) there exists a subset \( B^*(p_0) \subseteq \mathcal{B}(p_0) \) such that \( \ln \text{card}(B^*_p) \geq \frac{L2}{2\alpha} \ln 2 \) and the Hamming distance \( |\rho(\beta_1, \beta_2) - \frac{p_0}{8} | \) for any pair \( \beta_1, \beta_2 \in B^*_p \). Let \( \Theta^*(p_0) \) be the subset \( \{ \theta \in \mathbb{R}^n : \theta = X\beta, \beta \in B^*_p \} \). Then, along the lines of the proof of Case 2 of Theorem 2, for any pair \( \beta_1, \beta_2 \in B^*(p_0) \) and the corresponding \( \theta_1, \theta_2 \in \Theta^*(p_0) \) we have

\[
KL(\theta_1, \theta_2) \leq \frac{U}{2\alpha} \phi_{\max}[p_0] 4C_{p_0}^2 \rho(\beta_1, \beta_2) \leq \frac{U}{2\alpha} \phi_{\max}[p_0] 4C_{p_0}^2 p_0 \leq \frac{1}{16} \ln \text{card}(\Theta^*(p_0)),
\]
We now show that for $1 \leq p_0 \leq r/2$, in addition to (2),

$$\inf_\theta \sup_{\beta \in \mathcal{B}(p_0)} \mathbb{E}||X \beta - \tilde{\theta}||^2 \geq C \frac{a}{U} \tau[p_0] \frac{\ln m(p_0)}{\ln p_0} \tag{3}$$

The main difficulty in proving (3) w.r.t the proof of Theorem 2 is that we cannot use Varshamov-Gilbert bound and its generalizations (e.g., Lemma A.3 of Rigollet & Tsybakov, 2011) any longer due to the structural constraints on the set of admissible models. Instead, following Abramovich & Grinshtein (2013) we exploit the recent combinatorial results of Gutin & Jones (2012).

For $m(p_0) = 1$ the lower bound (3) holds trivially. Consider $m(p_0) > 1$. For any admissible model $M \in \mathcal{M}$ of size $p_0$ define a vector $\beta_M \in \mathbb{R}^p$ that has the corresponding $p_0$ nonzero entries equal to $C_{p_0}$ and $-C_{p_0}$ respectively (obviously, $\beta_2 = -\beta_1$), where $C_{p_0}^2 = \frac{\ln 2}{2m(\mathcal{U})^2} \phi_\max^{-1}[p_0]$, and then repeat the arguments from the previous proof for $p_0 \geq 8$.

We now show that for $1 \leq p_0 \leq r/2$, $\mu_{\theta_1 - \theta_2} \geq \phi_{\min}[p_0] \mu_{\beta_1 - \beta_2} \geq \phi_{\min}[2p_0] \mu_{\beta_1 - \beta_2} \geq \frac{a}{8} \frac{\tau[p_0]}{\ln m(p_0)} \ln p_0$, and (2) follows from Lemma A.1 of Bunea et al. (2007).

For $p_0 < 8$, define a trivial subset $\mathcal{B}^*(p_0)$ of two vectors $\beta_1$ and $\beta_2$ with the corresponding $p_0$ nonzero entries equal to $C_{p_0}$ and $-C_{p_0}$ respectively (obviously, $\beta_2 = -\beta_1$), where $C_{p_0}^2 = \frac{\ln 2}{2m(\mathcal{U})^2} \phi_\max^{-1}[p_0]$, and then repeat the arguments from the previous proof for $p_0 \geq 8$.

Theorem 2 of Gutin & Jones (2012) implies that for any constant $\tilde{c} > 2$ there exists a subset $\mathcal{B}^*(p_0) \subseteq \mathcal{B}(p_0)$ such that

$$\frac{\rho_{\max}}{\rho_{\min}} \leq \tilde{c} \quad \text{and} \quad \ln \text{card}(\mathcal{B}^*(p_0)) \geq \gamma \ln m(p_0),$$

where $\rho_{\max} = \max_{\beta_1, \beta_2 \in \mathcal{B}^*(p_0)} \rho(\beta_1, \beta_2)$ and $\rho_{\min} = \min_{\beta_1 \neq \beta_2 \in \mathcal{B}^*(p_0)} \rho(\beta_1, \beta_2)$ are respectively the maximal and the minimal Hamming distances among all pairs of vectors from $\mathcal{B}^*(p_0)$, and

$$\gamma = \left[ \frac{\ln(p_0/2)}{\ln(\tilde{c}/2)} \right]^{-1}$$

Set now

$$C_{p_0}^2 = \frac{a}{8} \frac{\gamma \ln m(p_0)}{\rho_{\max} \phi_{\max}[2p_0]}$$

Under the assumptions of the theorem $\mathcal{B}^*(p_0) \subseteq \mathcal{B}(p_0)$. Consider the corresponding subset $\Theta^*(p_0)$, where $\text{card}(\mathcal{B}^*(p_0)) = \text{card}(\Theta^*(p_0))$. For any $\theta_1, \theta_2 \in \Theta^*(p_0)$ associated with $\beta_1, \beta_2 \in \mathcal{B}^*(p_0)$ we then have

$$||\theta_1 - \theta_2||^2 = ||X(\beta_1 - \beta_2)||^2 \geq \phi_{\min}[2p_0] ||\beta_1 - \beta_2||^2 \geq \phi_{\min}[2p_0] C_{p_0}^2 \rho_{\min} \geq \frac{a}{8} \frac{\rho_{\max} \phi_{\max}[2p_0]}{\tau[p_0]} \gamma \ln m(p_0)$$

On the other hand,

$$K(\theta_1, \theta_2) \leq \frac{U}{2a} ||\theta_1 - \theta_2||^2 \leq \frac{U}{2a} \phi_{\max}[2p_0] C_{p_0}^2 \rho(\beta_1, \beta_2) \leq \frac{U}{2a} \phi_{\max}[2p_0] C_{p_0}^2 \rho_{\max} \leq \frac{1}{16} \text{card}(\Theta^*(p_0))$$

and Lemma A.1 of Bunea et al. (2007) yields (3).

Combining (1), (2) and (3) completes the proof.
References

[1] Gutin, G. and Jones, M. (2012). Note on large subsets of binary vectors with similar distances. *SIAM J. Discrete Math.* 26, 1108-1111.