High-dimensional classification by sparse logistic regression

Felix Abramovich, Vadim Grinshtein

***Abstract*—We consider high-dimensional binary classification by sparse logistic regression. We propose a model/feature selection procedure based on penalized maximum likelihood with a complexity penalty on the model size and derive the non-asymptotic bounds for its misclassification excess risk. To assess its tightness we establish the corresponding minimax lower bounds. The bounds can be reduced under the additional low-noise condition. The proposed complexity penalty is remarkably related to the VC-dimension of a set of sparse linear classifiers. Implementation of any complexity penalty-based criterion, however, requires a combinatorial search over all possible models. To find a model selection procedure computationally feasible for high-dimensional data, we extend the Slope estimator for logistic regression and show that under an additional weighted restricted eigenvalue condition it is rate-optimal in the minimax sense.**

***Index Terms*—Complexity penalty; feature selection; high-dimensionality; misclassification excess risk; sparsity; VC-dimension.**

I. Introduction

Classification is one of the most important setups in statistical learning and has been studied in various contexts. Theoretical foundations of classification are presented in the books [13] and [24], while the surveys of the state-of-the-art can be found in [11] and [15] (Section 9).

Consider a general (binary) classification with a (high-dimensional) vector of features $\mathbf{x} \in \mathbb{R}^d$ and the outcome class label $Y|\mathbf{x} \sim Bin(1, p(\mathbf{x}))$. The accuracy of a classifier η is defined by a misclassification error $R(\eta) = P(Y \neq \eta(\mathbf{x}))$. It is well-known that $R(\eta)$ is minimized by the Bayes classifier $\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}$. However, the probability function $p(\mathbf{x})$ is unknown and the resulting classifier $\widehat{\eta}(\mathbf{x})$ should be designed from the data D: a random sample of n independent observations $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$. The design points $\mathbf{x}_i$ may be considered as fixed or random. The corresponding (conditional) misclassification error of $\widehat{\eta}$ is $R(\widehat{\eta}) = P(Y \neq \widehat{\eta}(\mathbf{x})|D)$ and the goodness of $\widehat{\eta}$ w.r.t. η^* is measured by the misclassification excess risk $\mathcal{E}(\widehat{\eta}, \eta^*) = ER(\widehat{\eta}) - R(\eta^*)$.

Felix Abramovich is with the Department of Statistics and Operations Research, Tel Aviv University, Israel (email: felix@tauex.tau.ac.il).

Vadim Grinshtein is with the Department of Mathematics and Computer Science, The Open University of Israel, Israel (email: vadimg@openu.ac.il).

A common general (nonparametric) approach for finding a classifier $\widehat{\eta}$ from the data is empirical risk minimization (ERM), where minimization of a true misclassification error $R(\eta)$ is replaced by minimization of the corresponding empirical risk $\widehat{R}_n(\eta) = \frac{1}{n}\sum_{i=1}^n I\{Y_i \neq \eta(\mathbf{x}_i)\}$ over a given class of classifiers. Misclassification excess risk of ERM classifiers has been intensively studied in the literature (see, e.g., [11] and [15] (Section 9) for surveys and references therein). However, ERM can be hardly used directly in practice due to its computational cost and is typically relaxed by some related convex minimization surrogate (e.g., SVM).

Another possibility to obtain $\widehat{\eta}$ is to estimate $p(\mathbf{x})$ from the data by some $\widehat{p}(\mathbf{x})$ and use a plug-in classifier of the form $\widehat{\eta}(\mathbf{x}) = I\{\widehat{p}(\mathbf{x}) \geq 1/2\}$. A standard approach is to assume some (parametric or nonparametric) model for $p(\mathbf{x})$. In this paper we consider one of the most commonly used models – logistic regression, where it is assumed that $p(\mathbf{x}) = \frac{\exp(\boldsymbol{\beta}^t\mathbf{x})}{1+\exp(\boldsymbol{\beta}^t\mathbf{x})}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector of unknown regression coefficients. The corresponding Bayes classifier is a linear classifier $\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\} = I\{\boldsymbol{\beta}^t\mathbf{x} \geq 0\}$. One then estimates $\boldsymbol{\beta}$ from the data by the maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}}$, plugs-in $\widehat{\boldsymbol{\beta}}$ (or, equivalently, $\widehat{p}(\mathbf{x})$) and the resulting (linear) classifier is $\widehat{\eta}(\mathbf{x}) = I\{\widehat{p}(\mathbf{x}) \geq 1/2\} = I\{\widehat{\boldsymbol{\beta}}^t\mathbf{x} \geq 0\}$. Unlike ERM, the MLE $\widehat{\boldsymbol{\beta}}$ though not available in the closed form, can be nevertheless obtained numerically by the fast iteratively reweighted least squares algorithm [19] (Section 2.5). Nonparamertric plug-in classifiers were considered in [26], [16], [3].

In the era of "Big Data", however, the number of features d describing the objects for classification might be very large and even larger than the sample size n (large d small n setups) that raises a severe "curse of dimensionality" problem.

Reducing the dimensionality of a feature space by selecting a sparse subset of "significant" features becomes essential. Thus, [7], [14] showed that even in simple cases, high-dimensional classification without feature selection might be as bad as just pure guessing.

Nevertheless, unlike model selection in high-dimensional Gaussian regression that has been intensively studied in 2000s (see [8], [9], [1], [21], [25] among many others), there are much less theoretical results on model/feature selection in classification. [13] (Chapter 18) and [24] (Chapter 4) considered selection from a sequence of classifiers within a sequence of classes by penalized ERM with the structural penalty depending on the Vapnik-Chervonenkis (VC) dimension of a class. They established the oracle inequalities and the upper bounds for the misclassification excess risk of the selected classifier but did not provide the lower bound to assess its optimality. See also [11] (Section 8) for related penalized ERM approaches and references therein. Recall, however, that a computational cost (even for a *given* model) is a serious drawback of any ERM-based procedure.

The main goal of the paper is to fill the gap. In particular, we investigate feature selection in sparse logistic regression classification. Although logistic regression is widely used in various classification problems, its rigorous theoretical ground has not been yet properly established. Model selection in a general framework of generalized linear models (GLM) and in logistic regression in particular was studied in [2]. The authors proposed model selection procedure based on penalized maximum likelihood with a complexity penalty on the model size and investigated the goodness-of-fit of the resulting estimator in terms of the Kullback-Leibler risk. They derived the nonasymptotic bounds for this risk and showed that the resulting estimator is asymptotically minimax and adaptive to the unknown sparsity. In this paper we utilize their approach for classification and consider the corresponding plug-in classifier. In particular, we show that the considered complexity penalty is remarkably related to the VC-dimension of a set of sparse linear classifiers. We establish the non-asymptotic upper bound for misclassification excess risk of the resulting classifier and construct explicitly the design for which it is sharp in the minimax sense. We also show that the excess risk bounds can be improved under the additional low-noise assumption.

Any model selection criterion based on a complexity penalty requires, however, a combinatorial search over all possible models that makes its usefulness problematic for high-dimensional data. A common remedy is to replace the original complexity penalty by a related convex surrogate. The probably most well-known techniques is Lasso. However, it can achieve only the sub-optimal rates under some extra conditions on the design matrix X ([23]). Recently, for Gaussian linear regression [10] proposed a Slope estimator. [6] showed that under certain additional conditions on X, Slope is rate-optimal for linear regression. We adapt it to the logistic regression (and, in fact, to a general GLM) setup and extend the results of [6]).

The rest of the paper is organized as follows. In Section II we present the model (feature) selection procedure for sparse logistic regression with fixed design based on a general procedure of [2] and provide the upper bounds for the resulting estimator in terms of Kullback-Leibler risk. In Section III we apply it for classification to establish the non-asymptotic upper bound for its misclassification excess risk and derive the corresponding minimax lower bounds. The improvement of the obtained risk bounds under the additional low-noise assumption is given in Section IV. In Section V we consider the logistic Slope classifier as a convex surrogate for the proposed feature selection procedure and show that its misclassification excess risk is still rate-optimal under an extra weighted restricted eigenvalue condition on the design matrix X. The random design case is considered in Section VI . Section VII provides a short real-data example. All the proofs are given in the Appendix.

II. Notation and preliminaries

Consider a sparse logistic regression model

$$Y_i \sim Bin(1, p_i), \qquad \ln \frac{p_i}{1-p_i} = \boldsymbol{\beta}^t \mathbf{x}_i \tag{1}$$

with deterministic design points $\mathbf{x}_i \in \mathbb{R}^d \quad i = 1, \ldots, n$, where we assume that the unknown vector of regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^d$ is sparse.

Let $d_0 = ||\boldsymbol{\beta}||_0$ be the size of true (unknown) model, where the l_0 (quasi)-norm of regression coefficients $||\boldsymbol{\beta}||_0$ is the number of nonzero entries. Let $X \in \mathbb{R}^{n \times d}$ be the design matrix of rows $\mathbf{x}_i$, $r = rank(X)$ and assume that any r columns of X are linearly independent.

For the model (1) the log-likelihood function is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \boldsymbol{\beta}^t \mathbf{x}_i \; Y_i - \ln(1 + \exp(\boldsymbol{\beta}^t \mathbf{x}_i) \right\}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

Let $\mathfrak{M}$ be the set of all 2^d possible models $M \subseteq \{1, \ldots, d\}$. For a given model M, define $\mathcal{B}_M = \{\boldsymbol{\beta} \in \mathbb{R}^d : \beta_j = 0 \text{ if } j \notin M\}$. The MLE $\widehat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ is then

$$\widehat{\boldsymbol{\beta}}_M = \arg \max_{\widetilde{\boldsymbol{\beta}} \in \mathcal{B}_M} \sum_{i=1}^n \left\{ \widetilde{\boldsymbol{\beta}}^t \mathbf{x}_i Y_i - \ln\left(1 + \exp(\widetilde{\boldsymbol{\beta}}^t \mathbf{x}_i)\right) \right\} \quad (2)$$

The $\widehat{\boldsymbol{\beta}}_M$ in (2) is not available in the closed form but can be obtained numerically by the iteratively reweighted least squares algorithm (see [19], Section 2.5). The corresponding MLE for probabilities p_i are $\widehat{p}_{Mi} = \frac{\exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}{1+\exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)}$.

Select the model $\widehat{M}$ by the penalized maximum likelihood model selection criterion of the form

$$\widehat{M} = \arg \min_{M \in \mathfrak{M}} \left\{ \sum_{i=1}^n \left(\ln\left(1 + \exp(\widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i)\right) - \widehat{\boldsymbol{\beta}}_M^t \mathbf{x}_i Y_i \right) + Pen(|M|) \right\}, \quad (3)$$

where $Pen(|M|)$ is a complexity penalty on the model size $|M|$. In fact, we can restrict $\mathfrak{M}$ in (3) to models with sizes at most r since for any $\boldsymbol{\beta}$ with $||\boldsymbol{\beta}||_0 > r$, there necessarily exists another $\boldsymbol{\beta}'$ with $||\boldsymbol{\beta}'||_0 \le r$ such that $X\boldsymbol{\beta} = X\boldsymbol{\beta}'$.

Within general GLM framework, [2] investigated the goodness-of-fit of the resulting estimator $\widehat{\mathbf{p}}_{\widehat{M}}$. They considered the Kullback-Leibler divergence $KL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}})$ between the data distribution with the true probabilities $\mathbf{p} = (p_1, \ldots, p_n)$ and its empirical distribution generated by $\widehat{\mathbf{p}}_{\widehat{M}}$ given by

$$KL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) = \frac{1}{n} \sum_{i=1}^n \left\{ p_i \ln\left(\frac{p_i}{\widehat{p}_{\widehat{M}i}}\right) + (1-p_i) \ln\left(\frac{1-p_i}{1-\widehat{p}_{\widehat{M}i}}\right) \right\}, \quad (4)$$

and measured the accuracy of $\widehat{\mathbf{p}}_{\widehat{M}}$ by the corresponding Kullback-Leibler risk $EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}})$ (in fact, the Kullback-Leibler divergence $KL(\cdot,\cdot)$ in [2] was defined as n times $KL(\cdot,\cdot)$ in this paper).

Assumption (A). *Assume that there exists* $0 < \delta < 1/2$ *such that* $\delta < p_i < 1 - \delta$ *or, equivalently, there exists* $C_0 > 0$ *such that* $|\boldsymbol{\beta}^t \mathbf{x}_i| < C_0$ *in (1) for all* $i = 1, \ldots, n$.

Assumption (A) prevents the variances $Var(Y_i) = p_i(1 - p_i)$ to be infinitely close to zero.

Consider a set of models of size at most d_0, where $1 \le d_0 \le r$. Obviously, $|M| \le d_0$ iff $||\boldsymbol{\beta}||_0 \le d_0$. [2] showed that for the complexity penalty

$$Pen(|M|) = c\,|M| \ln \frac{de}{|M|},\ |M| = 1, \ldots, r-1;\ Pen(r) = c \cdot r \quad (5)$$

in (3), where $c > \frac{4}{\delta(1-\delta)}$, under Assumption (A), the upper bound of the Kullnack-Leibler risk is given by

$$\sup_{\boldsymbol{\beta}: ||\boldsymbol{\beta}||_0 \le d_0} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) \le C \frac{1}{\delta(1-\delta)} \frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{n} \quad (6)$$

for some $C > 0$. They also derived the corresponding minimax lower bounds for the Kullback-Leibler risk and showed that for weakly-collinear design, the upper bound in (6) is of the optimal order (in the minimax sense).

The above results on the Kullback-Leibler risk can be extended to model selection under additional structural constraints on the set of admissible models $\mathfrak{M}$ (see Secition 4.1 of [2]).

In what follows we utilize (6) to derive the upper bounds for the misclassification excess risk of the corresponding plug-in classifier $\widehat{\eta}_{\widehat{M}}(\mathbf{x}) = I\{\widehat{\boldsymbol{\beta}}_{\widehat{M}}^t \mathbf{x} \ge 0\}$.

To gain more insight into the complexity penalty (5) within classification framework we present the following lemma on the Vapnik-Chervonenskis (VC) dimension of the set of all d_0-sparse linear classifiers :

Lemma 1. *Let* $\mathcal{C}(d_0) = \{\eta(\mathbf{x}) = I\{\boldsymbol{\beta}^t \mathbf{x} \ge 0\} : \boldsymbol{\beta} \in \mathbb{R}^d,\ ||\boldsymbol{\beta}||_0 \le d_0\}$ *be the set of all* d_0*-sparse linear classifiers and* $V(\mathcal{C}(d_0))$ *its VC-dimension. Then,*

$$d_0 \log_2\left(\frac{2d}{d_0}\right) \le V(\mathcal{C}(d_0)) \le 2 d_0 \log_2\left(\frac{de}{d_0}\right) \quad (7)$$

Thus, the complexity penalty $Pen(|M|)$ in (5) is essentially proportional to the VC-dimension of the corresponding class of $|M|$-sparse linear classifiers $\mathcal{C}(|M|)$.

III. Misclassification excess risk bounds

In this section we apply the selected model $\widehat{M}$ in (3) for classification and derive the bounds for the corresponding misclassification exceeess risk.

We consider first the fixed design. For a given design matrix X the misclassification error of a classifier η is $R_X(\eta) = \frac{1}{n}\sum_{i=1}^n P(Y_i \ne \eta(\mathbf{x}_i))$. Following our previous arguments define a (linear) plug-in classifier

$$\widehat{\eta}_{\widehat{M}}(\mathbf{x}) = I\{\widehat{\boldsymbol{\beta}}_{\widehat{M}}^t \mathbf{x} \ge 0\} \quad (8)$$

and consider its misclassification excess risk $\mathcal{E}_X(\widehat{\eta}_{\widehat{M}}, \eta^*) = ER_X(\widehat{\eta}_{\widehat{M}}) - R_X(\eta^*)$, where recall that the (ideal) Bayes classifier $\eta^*(\mathbf{x}) = I\{\boldsymbol{\beta}^t \mathbf{x} \ge 0\}$ with the true (unknown) $\boldsymbol{\beta}$ in (1). The remarkable results of [28] and [5] establish the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

relations between the Kullback-Leibler risk $EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}})$ and the misclassification excess risk $\mathcal{E}_X(\widehat{\eta}_{\widehat{M}}, \eta^*)$:

$$\mathcal{E}_X(\widehat{\eta}_{\widehat{M}}, \eta^*) \leq \sqrt{2EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}})} \tag{9}$$

Thus, (6) and (9) imply immediately the following upper bound for $\mathcal{E}_X(\widehat{\eta}_{\widehat{M}}, \eta^*)$:

Theorem 1. *Consider a sparse logistic regression model (1) with $||\boldsymbol{\beta}||_0 \leq d_0$. Let $\widehat{M}$ be a model selected in (3) with the complexity penalty (5) and consider the corresponding plug-in classifier $\widehat{\eta}_{\widehat{M}}(\mathbf{x})$ in (8). Under Assumption (A),*

$$\sup_{\eta^* \in \mathcal{C}(d_0)} \mathcal{E}_X(\widehat{\eta}_{\widehat{M}}, \eta^*) \leq C_1 \sqrt{\frac{1}{\delta(1-\delta)} \; \frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{n}} \tag{10}$$

for some $C_1 > 0$, simultaneously for all $1 \leq d_0 \leq r$.

The constant $\sqrt{\frac{1}{\delta(1-\delta)}}$ in (10) is a result of the direct application of (6) and (9). It can be improved by establishing the similar relations between misclassification excess risk and other losses rather than Kullback-Leibler in (9) and deriving the corresponding upper bounds for their risks. See, for example, the results and the proof of Theorem 6 below for the random design. In a way, the Kullback-Leibler loss implies the most conservative upper bound ([20]).

We now show that there exists a design matrix X_0 for which the upper bound for the misclassification excess risk (10) is essential sharp (up to a probably different constant).

Consider the set of all possible d_0-sparse linear classifiers $\mathcal{C}(d_0)$ defined in Lemma 1 and the case, where a Bayes classifier $\eta^*(\mathbf{x})$ is not perfect, that is, $R(\eta^*) > 0$ (aka an *agnostic* model). Then, the following result holds:

Theorem 2. *Consider a d_0-sparse agnostic logistic regression model (1), where $2 \leq d_0 \log_2\left(\frac{2d}{d_0}\right) \leq n$.*

Then, there exists a design matrix $X_0 \in \mathbb{R}^{n \times d}$ such that

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}(d_0)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \geq C_2 \sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}} \tag{11}$$

for some constant $C_2 > 0$, where the infimum is taken over all classifiers $\tilde{\eta}$ based on the data $(X_0, \mathbf{Y})$.

Theorem 2 is a particular case of Theorem 4 from Section IV below.

The upper and lower bounds established in Theorem 1 and Theorem 2 imply the asymptotic minimax rate for misclassification excess risk in sparse logistic regression model as n increases. We allow the number of features d to increase with n as well and even faster than n ($d \gg n$ setup). The following immediate Corollary 1 shows that the proposed classifier $\widehat{\eta}_{\widehat{M}}$ is asymptotically minimax in terms of "the worst case" design and adaptive to the unknown sparsity:

Corollary 1. *Consider a d_0-sparse logistic regression agnostic model (1), where d_0 satisfies $2 \leq d_0 \log_2\left(\frac{2d}{d_0}\right) \leq n$. Then, as n and d increase, for a fixed δ in Assumption (A),*

1) *The asymptotic minimax misclassification excess risk $\sup_X \inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}(d_0)} \mathcal{E}_X(\tilde{\eta}, \eta^*)$ is of the order*

$$\sqrt{\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n}} \sim \sqrt{\frac{V(\mathcal{C}(d_0))}{n}}$$

2) *The classifier $\widehat{\eta}_{\widehat{M}}$ defined in (8), where the model $\widehat{M}$ was selected by (3) with the complexity penalty (5), attains the minimax rates simultaneously for all $2 \leq d_0 \log_2\left(\frac{2d}{d_0}\right) \leq n$.*

Finally, we note that if the considered logistic regression model is misspecified and the Bayes classifier η^* is not linear, we still have the following risk decomposition

$$\begin{aligned} R_X(\widehat{\eta}_{\widehat{M}}) - R_X(\eta^*) = &\left(R_X(\widehat{\eta}_{\widehat{M}}) - R_X(\eta^*_L)\right) \\ &+ \left(R_X(\eta^*_L) - R_X(\eta^*)\right), \end{aligned} \tag{12}$$

where $\eta^*_L = \arg\min_{\eta \in \mathcal{C}(d)} R_X(\eta)$ is the best (ideal) linear classifier. Our previous arguments can then be applied to the first term in the RHS of (12) representing the estimation error, while the second term is an approximation error and measures the ability of linear classifiers to perform as good as η^*. Enriching the class of classifiers may improve the approximation error but will necessarily increase the estimation error in (12). In a way, it is similar to the variance/bias tradeoff in regression.

IV. Tighter risk bounds under low-noise condition

The main challenges for any classifier occur near the the boundary $\{\mathbf{x} : p(\mathbf{x}) = 1/2\}$ (equivalently, a hyperplane $\boldsymbol{\beta}^t\mathbf{x} = 0$ for the logistic regression model), where it is hard to predict the class label accurately. However, for regions, where $p(\mathbf{x})$ is bounded away from $1/2$ (margin or aka low-noise condition), the bounds for misclassification excess risk established in the previous Section III can be improved. Following [18] introduce the following low-noise assumption:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

Assumption (B). *Consider the logistic regression model (1) and assume that there exists $0 \leq h < 1/2$ such that*

$$|p_i - 1/2| \geq h \;\text{ or, equivalently, }\; |\boldsymbol{\beta}^t \mathbf{x}_i| \geq \ln\left(\frac{1+2h}{1-2h}\right) \tag{13}$$

for all $i = 1, \ldots, n$.

Assumption (B) essentially assumes the existence of the "corridor" of width $2\ln\left(\frac{1+2h}{1-2h}\right)$ that separates the two sets $\{\mathbf{x}_i : \boldsymbol{\beta}^t\mathbf{x}_i > 0,\; i = 1, \ldots, n\}$ and $\{\mathbf{x}_i : \boldsymbol{\beta}^t\mathbf{x}_i < 0,\; i = 1, \ldots, n\}$.

For a given design matrix X, define $\mathcal{C}_X(d_0, h) = \{\eta : \eta \in \mathcal{C}(d_0),\; |\boldsymbol{\beta}^t\mathbf{x}_i| \geq \ln\left(\frac{1+2h}{1-2h}\right),\; i = 1, \ldots, n\}$. Evidently, $\mathcal{C}_X(d_0, 0) = \mathcal{C}(d_0)$ for any X.

Theorem 3 below establishes the upper bound for the misclassification excess risk of the proposed classifier $\widehat{\eta}_{\widehat{M}}$ under the additional low noise Assumption (B):

Theorem 3. *Consider a sparse logistic regression model (1), where $||\boldsymbol{\beta}||_0 \leq d_0$. Assume that there exist $0 < h < \Delta < 1/2$ such that*

$$h \leq |p_i - 1/2| \leq \Delta \tag{14}$$

for all $i = 1, \ldots, n$.

Let $\widehat{M}$ be a model selected in (3) with the complexity penalty (5) and consider the corresponding classifier $\widehat{\eta}_{\widehat{M}}(\mathbf{x})$ in (8). Then, for all $1 \leq d_0 \leq r$,

$$\sup_{\eta^* \in \mathcal{C}_X(d_0,h)} \mathcal{E}_X(\widehat{\eta}_{\widehat{M}}, \eta^*) \leq C_1 \min\left(\sqrt{\frac{1-4h^2}{1-4\Delta^2}\;\frac{\min\left(d_0 \ln\frac{de}{d_0}, r\right)}{n}}\;,\; \frac{1-4h^2}{1-4\Delta^2}\;\frac{\min\left(d_0 \ln\frac{de}{d_0}, r\right)}{nh}\right) \tag{15}$$

for some $C_1 > 0$.

Thus, if the margin parameter h is large enough, namely, $h > \sqrt{\frac{d_0 \ln\frac{de}{d_0}}{n}}$, the misclassification excess risk bound (10) is reduced. The classifier $\widehat{\eta}_{\widehat{M}}(\mathbf{x})$ does not depend on h and the procedure is inherently adaptive to its value.

Similar to the previous Section III, one can construct a design matrix for which the upper bound (15) is sharp:

Theorem 4. *Consider a d_0-sparse agnostic logistic regression model (1) with $2 \leq d_0 \log_2\frac{2d}{d_0} \leq n$.*

There exists a design matrix $X_0 \in \mathbb{R}^{n \times d}$ such that under Assumption (B)

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{X_0}(d_0,h)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \geq C_2 \min\left(\sqrt{\frac{d_0 \ln\frac{de}{d_0}}{n}}\;,\; \frac{d_0 \ln\frac{de}{d_0}}{nh}\right) \tag{16}$$

for some $C_2 > 0$.

The design matrix X_0 is constructed explicitly in the proof of Theorem 4 in the Appendix. Note that Theorem 2 may be viewed as a particular case of Theorem 4 for $h = 0$.

V. Logistic Slope classifier

Solving for $\widehat{M}$ in (3) requires generally a combinatorial search over all possible models in $\mathfrak{M}$ that makes the use of complexity penalties to be computationally problematic when the number of features is large. Greedy algorithms (e.g., forward selection) approximate the global solution of (3) by a stepwise sequence of local ones. However, they require strong constraints on the design matrix X that can hardly hold for high-dimensional data. A more reasonable approach is convex relaxation, where the original combinatorial problem is replaced by a related convex surrogate. Thus, for linear-type complexity penalties of the form $Pen(|M|) = \lambda|M| = \lambda||\boldsymbol{\beta}||_0$, the celebrated Lasso replaces the l_0-(quasi) norm by l_1-norm:

$$\widehat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\widetilde{\boldsymbol{\beta}}} \left\{ \sum_{i=1}^n \left(\ln\left(1 + \exp(\widetilde{\boldsymbol{\beta}}^t\mathbf{x}_i)\right) - \widetilde{\boldsymbol{\beta}}^t\mathbf{x}_i Y_i \right) + \lambda||\widetilde{\boldsymbol{\beta}}||_1 \right\}$$

Assume that all columns of the design matrix X are normalized to have unit norms. From the results of [23] it follows that under an assumption similar to Assumption (A) and certain extra conditions on X, the logistic Lasso with a tuning parameter λ of the order $\sqrt{\ln d}$ results in sub-optimal Kullback-Leibler risk $O\left(\frac{d_0}{n}\ln d\right)$ and, therefore, sub-optimal misclassification excess risk $O\left(\sqrt{\frac{d_0}{n}\ln d}\right)$. For Gaussian regression, [6] showed that under certain conditions on X, Lasso can achieve the optimal rate with *adaptively chosen* λ by Lepski procedure.

Recently, for Gaussian regression, [10] suggested the Slope estimator – a penalized maximum likelihood estimator with a *sorted* l_1-norm penalty defined as follows:

$$\widehat{\boldsymbol{\beta}}_{Slope} = \arg\min_{\widetilde{\boldsymbol{\beta}}} \left\{ ||\mathbf{Y} - X\widetilde{\boldsymbol{\beta}}||_2^2 + \sum_{j=1}^d \lambda_j |\widetilde{\boldsymbol{\beta}}|_{(j)} \right\}, \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

where $||\cdot||_2$ denotes the Euclidean norm in $\mathbb{R}^n$, $|\widetilde{\boldsymbol\beta}|_{(1)} \geq \ldots \geq |\widetilde{\boldsymbol\beta}|_{(d)}$ are the descendingly ordered absolute values of $\widetilde{\boldsymbol\beta}_j$'s and $\lambda_1 \geq \ldots \geq \lambda_d > 0$ are the tuning parameters. It is a convex minimization problem. Unlike the constant λ in Lasso, there is a sequence of decreasing λ_j in Slope. [6] proved that under a weighted restricted eigenvalue condition on X with normalized columns, the quadratic risk of the Slope estimator (17) with $\lambda_j = A\sqrt{\ln(2d/j)}$ for a certain constant A is of the (rate-optimal) order $O\left(\frac{d_0}{n}\ln(\frac{de}{d_0})\right)$.

We will now extend the above results for Slope for logistic regression and, in fact, for a general GLM (see the Appendix C). Naturally modifying the definition of the Slope estimator for the considered logistic regression model (1), define

$$\widehat{\boldsymbol\beta}_{Slope} = \arg\min_{\widetilde{\boldsymbol\beta}} \left\{ \sum_{i=1}^n \left(\ln\left(1+\exp(\widetilde{\boldsymbol\beta}^t\mathbf{x}_i)\right) - \widetilde{\boldsymbol\beta}^t\mathbf{x}_i Y_i\right) + \sum_{j=1}^d \lambda_j |\widetilde{\boldsymbol\beta}|_{(j)} \right\}, \tag{18}$$

where $\lambda_1 \geq \ldots \geq \lambda_d > 0$. Note that (18) is also a convex program that makes the logistic Slope estimator computationally feasible for high-dimensional data. The corresponding estimated probabilities $\widehat{p}_{Slope,i} = \frac{\exp(\widehat{\boldsymbol\beta}^t_{Slope}\mathbf{x}_i)}{1+\exp(\widehat{\boldsymbol\beta}^t_{Slope}\mathbf{x}_i)},\ i = 1,\ldots,n$.

As usual, any convex relaxation requires certain extra conditions on the restricted minimal eigenvalue of the design matrix X over some set of vectors. In particular, similar to Gaussian regression considered in [6], we assume the following *Weighted Restricted Eigenvalue* (WRE) condition for Slope estimator (18) :

Assumption. *($WRE(d_0, c_0)$* **condition)** *Consider the sparse logistic regression model (1) with $||\boldsymbol\beta||_0 \leq d_0$, where the columns of the design matrix X are normalized to have unit norms. Consider the set $\mathcal{S}(d_0, c_0) = \{\mathbf{u} \in \mathbb{R}^d : \sum_{j=1}^d \sqrt{\ln(2d/j)}|u|_{(j)} \leq (1+c_0)||\mathbf{u}||_2 \sqrt{\sum_{j=1}^{d_0}\ln(2d/j)}\}$ and assume that $X\mathbf{u} \neq \mathbf{0}$ for any $\mathbf{u} \neq \mathbf{0} \in \mathcal{S}(d_0, c_0)$.*

An interesting discussion on the relations between the WRE condition and the restricted eigenvalue condition (RE) required for Lasso is given in Section 8 of [6].

Define a restricted minimal eigenvalue $\kappa(d_0, c_0)$ as follows :

$$\kappa(d_0, c_0) = \min_{\mathbf{u}\in\mathcal{S}(d_0,c_0);\mathbf{u}\neq\mathbf{0}} \frac{||X\mathbf{u}||_2}{||\mathbf{u}||_2} > 0$$

Theorem 5. *Consider a sparse logistic regression model (1), where $||\boldsymbol\beta||_0 \leq d_0$, the columns of the design matrix X are normalized to have unit norms and, in addition, X satisfies the $WRE(d_0, c_0)$ condition for some $c_0 > 1$. Assume that Assumption (A) holds.*

Let the tuning parameters

$$\lambda_j = A\,\frac{c_0+1}{c_0-1}\sqrt{\ln(2d/j)}, \quad j = 1,\ldots,d \tag{19}$$

with the constant $A \geq 20\sqrt{6}$.

Then,

$$\sup_{\boldsymbol\beta:||\boldsymbol\beta||_0\leq d_0} \mathbb{E}KL(\mathbf{p}, \widehat{\mathbf{p}}_{Slope}) \leq 8A^2\,\frac{c_0^2}{(c_0-1)^2}\,\frac{1}{\delta(1-\delta)} \times \left(\frac{2\pi+8}{\ln(2d)} + \frac{1}{\kappa^2(d_0,c_0)}\right)\frac{d_0}{n}\ln\left(\frac{2de}{d_0}\right) \tag{20}$$

for all $1 \leq d_0 \leq r$.

Note that λ_j's in (19) are of the same form as those in [6] for Gaussian regression but differ in a constant A.

Theorem 5 is a particular case of Theorem 8 for a general GLM (see Appendix).

Using (9) one immediately gets the corresponding result for the misclassification exceess risk of the logistic Slope classifier:

Corollary 2. *Assume all the conditions of Theorem 5 and choose λ_j according to (19). Consider the logistic Slope classifier $\widehat{\eta}_{Slope}(\mathbf{x}) = I\{\widehat{\boldsymbol\beta}^t_{Slope}\mathbf{x} \geq 0\}$. Then,*

$$\mathcal{E}_X(\widehat{\eta}_{Slope}, \eta^*) = O\left(\sqrt{\frac{d_0\ln\frac{de}{d_0}}{n}}\right) \tag{21}$$

Thus, the logistic Slope estimator is computationally feasible and yet achieves the optimal rates under the additional $WRE(d_0, c_0)$ condition on the design for all but very dense models for which $d_0\ln(\frac{de}{d_0}) > r$ (see Theorem 1).

Furthermore, following the arguments in the proof of Theorem 3, one can show that the bound (21) for $\mathcal{E}_X(\widehat{\eta}_{Slope}, \eta^*)$ may be reduced under the additional low noise Assumption (B).

VI. Random design

The results above have been obtained for the fixed design. In machine learning, it is more common to consider classification with random design. In this section we show that our main previous results for the fixed design can be extended for the random design.

Consider the following model:

$$Y|(\mathbf{X}=\mathbf{x}) \sim B(1, p(\mathbf{x})),\ p(\mathbf{x}) = \frac{\exp(\boldsymbol\beta^t\mathbf{x})}{1+\exp(\boldsymbol\beta^t\mathbf{x})} \quad \text{and}\ \ \mathbf{X} \sim q(\mathbf{x}), \tag{22}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

where $q(\cdot)$ is a marginal density of $\mathbf{X}$ with a bounded support $\mathcal{X} \subset \mathbb{R}^d$. By re-scaling we can assume without loss of generality that $||\mathbf{x}||_2 \leq 1$ for all $\mathbf{x} \in \mathcal{X}$, where recall that $||\cdot||_2$ is the Euclidean norm.

We assume that all X_j are linearly independent. Hence, the minimal eigenvalue $\lambda_{\min}(G)$ of the matrix $G = E(\mathbf{X}\mathbf{X}^t)$ is strictly positive.

Recall that the misclassification excess risk of a classifier $\widehat{\eta}$ designed from a random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from the joint distribution $(\mathbf{X}, Y)$ is $\mathcal{E}(\widehat{\eta}_{\widehat{M}}, \eta^*) = ER(\widehat{\eta}) - ER(\eta^*)$, where the Bayes classifier $\eta^*(\mathbf{x}) = I(\boldsymbol{\beta}^t\mathbf{x} \geq 0)$. Adapting a general result on the minimax lower bound for the misclassification excess risk for random design (see, e.g., [13], Chapter 14 and [11], Section 5.5) for a d_0-sparse agnostic logistic regression model (22), by Lemma 1 we have:

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}(d_0), q} \mathcal{E}(\tilde{\eta}, \eta^*) \geq C\sqrt{\frac{V(\mathcal{C}(d_0))}{n}} \geq \tilde{C}\sqrt{\frac{d_0 \ln \frac{de}{d_0}}{n}}$$

Similar to the fixed design setup, consider the penalized maximum likelihood model selection procedure (3) with the complexity penalty

$$Pen(|M|) = C|M| \ln \frac{de}{|M|}, \quad |M| = 1, \ldots, \min(d, n), \tag{23}$$

where the exact choice for the constant C will follow from the proof of Theorem 7 below.

The following Assumption (A1) is a direct analog of Assumption (A) for random design:

Assumption (A1). *Assume that there exists $0 < \delta < 1/2$ such that $\delta < p(\mathbf{x}) < 1 - \delta$ or, equivalently, there exists $C_0 > 0$ such that $|\boldsymbol{\beta}^t\mathbf{x}| < C_0$ in (22) for all $\mathbf{x} \in \mathcal{X}$.*

Theorem 6 extends the results of Theorem 1 for random design:

Theorem 6. *Consider a sparse logistic regression model (22), where $||\boldsymbol{\beta}||_0 \leq d_0$.*

Let $\widehat{M}$ be a model selected in (3) with the complexity penalty (23) and consider the corresponding plug-in classifier $\widehat{\eta}_{\widehat{M}}(\mathbf{x})$ in (8). Under Assumption (A1),

$$\sup_{\eta^* \in \mathcal{C}(d_0)} \mathcal{E}(\widehat{\eta}_{\widehat{M}}, \eta^*) \leq C\sqrt{\ln\left(\frac{1}{\delta\lambda_{\min}(G)}\right) \frac{d_0 \ln \frac{de}{d_0}}{n}}$$

for some positive $C > 0$, simultaneously for all $1 \leq d_0 \leq \min(d, n)$.

Thus, the classifier $\widehat{\eta}_{\widehat{M}}(\mathbf{x})$ is adaptively rate-optimal (in the minimax sense) for random design as well.

Theorem 6 is a particular case of Theorem 7 stated below.

We should note that similar upper bounds can be obtained for model selection by penalized ERM utilizing general results of [13], Chapter 18 and [24], Chapter 4 on structural penalties depending on a VC-dimension and applying Lemma 1 for their adaptation to sparse logistic regression. See also [11], Section 8 for related ERM approaches and references therein. Recall, however, that a computational cost is a crucial drawback of any ERM-based procedure.

The misclassification excess risk of $\widehat{\eta}_{\widehat{M}}(\mathbf{x})$ can be again improved under the low-noise condition which can even be formulated in a more general form for random design ([17] and [22]):

Assumption (B1). *Assume that there exist $C > 0$ and $\alpha \geq 0$ such that*

$$P\left(|p(\mathbf{X}) - 1/2| \leq h\right) \leq Ch^\alpha \tag{24}$$

for all $0 < h < h^*$, where $h^* < 1/2$.

Assumption (B) in Section IV for the fixed design can be viewed as a limiting case $\alpha = \infty$.

Theorem 7. *Consider a sparse logistic regression model (22), where $||\boldsymbol{\beta}||_0 \leq d_0$. Let $\widehat{M}$ be a model selected in (3) with the complexity penalty (23) and consider the corresponding plug-in classifier $\widehat{\eta}_{\widehat{M}}(\mathbf{x})$ in (8).*

Under Assumptions (A1) and (B1), there exists $C > 0$ such that

$$\sup_{\eta^* \in \mathcal{C}(d_0)} \mathcal{E}(\widehat{\eta}_{\widehat{M}}, \eta^*) \leq C\left(\ln\left(\frac{1}{\delta\lambda_{\min}(G)}\right) \frac{d_0 \ln \frac{de}{d_0}}{n}\right)^{\frac{\alpha+1}{\alpha+2}} \tag{25}$$

for all $1 \leq d_0 \leq \min(d, n)$.

Theorem 6 (no low-noise assumption) corresponds to the extreme case $\alpha = 0$. For another extreme case $\alpha = \infty$ (complete separation from 1/2), the upper bound (25) is $O\left(\frac{d_0 \ln \frac{de}{d_0}}{n}\right)$ similar to the results of Section IV for the fixed design.

The rates (25) can be reduced further under additional conditions on the support $\mathcal{X}$ and the density $q(\mathbf{x})$ using the arguments of [16] and [3] but this is beyond the scope of the paper.

To extend the results of Theorem 5 for Slope estimator for random design one needs the $WRE(d_0, c_0)$ condition to be held with high probability. It evidently depends on the marginal distribution $q(\mathbf{x})$. Thus, [6] (Theorem 8.3) showed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

that it is satisfied for multivariate Gaussian and even sub-Gaussian $\mathbf{X}$ when $d_0 \ln^2(de/d_0) \leq cn$ for some constant $c > 0$, and under mild conditions on the covariance matrix.

VII. Numerical example

We now demonstrate the performance of the proposed feature selection and classification procedures on a short numerical real-data study.

The online marketing data contains various information about 500 customers registered at the site: their personal data (e.g., gender, country of living, etc.) and purchase activities during the past period (e.g., frequencies and types of purchases, purchase amounts, currencies, etc.). Overall, there were 61 explanatory variables. Given such data, one of the main goals is to predict customers who are about to become inactive in order to incentivize them to remain by various discounts. The output then is a binary variable indicating whether a customer was still active during the next time period or not.

The data was randomly split into training ($n_1 = 400$) and test ($n_2 = 100$) sets. The number of possible features $d = 61$ was too large to perform a complete combinatorial search for penalized maximum likelihood model selection procedure (3) with the complexity penalty (5). Instead we used its forward selection version, logistic Lasso and logistic Slope classifiers. The corresponding tuning constants were chosen by 5-fold cross-validation and the resulting three classifiers were then applied to the test set.

The best misclassification rate was achieved by Slope (17%), followed by Lasso (19%) and forward selection (22%). In addition, we compared the sizes of the models selected by the three classifiers. The conservative forward selection procedure yielded a very sparse model with only 2 predictors – the time since last purchase and purchase amount. On the other hand, Lasso with the CV-chosen tuning parameter is known to tend to select too many variables (see, e.g., [12], Section 2) and resulted in the model of size 13 by adding 11 other variables. The Slope classifier with a decreasing sequence of tuning parameters commonly implies even larger models (32 in the considered example). Note, however, that prediction and model identification are two different problems and, in particular, the choices of tuning parameters for them should be different.

Appendix

Throughout the proofs we use various generic positive constants, not necessarily the same each time they are used even within a single equation.

Appendix A
Proof of Lemma 1

Denote for brevity $V = V(\mathcal{C}(d_0))$. For any fixed subset of d_0 β_j's the VC of the corresponding set of d_0-dimensional linear classifiers is known to be d_0 (e.g., [15], Exercise 9.5.2). Then, by Sauer's lemma the maximal number of different labelling of V points in $\mathbb{R}^{d_0}$ that such set of classifiers can produce is $\sum_{k=0}^{d_0} \binom{V}{k} \leq \left(\frac{Ve}{d_0}\right)^{d_0}$ (see, e.g., [15], Section 9.2.2). The overall number of different labelling is, therefore, $\binom{d}{d_0} \sum_{k=0}^{d_0} \binom{V}{k}$, and by the definition of $V(\mathcal{C}(d_0))$ we have

$$2^V \leq \binom{d}{d_0} \sum_{k=0}^{d_0} \binom{V}{k} \leq \left(\frac{de}{d_0}\right)^{d_0} \left(\frac{Ve}{d_0}\right)^{d_0} \leq \left(\frac{de}{d_0}\right)^{2d_0}$$

that implies an upper bound $V \leq 2\, d_0 \log_2 \left(\frac{de}{d_0}\right)$.

On the other hand, take $k = \log_2(2d/d_0)$ and let K be the $k \times 2^{k-1}$ matrix whose columns are all possible vectors with entries ± 1 and the first entry 1. Note that $d_0 2^{k-1} = d$. Let W be the $d_0 k \times d$ block-wise matrix consisting of $d_0 \times d_0$ blocks, each being a $k \times 2^{k-1}$ matrix, where the diagonal matrices are copies of K, while all others are zero matrices. Thus, W has $d_0 k = d_0 \log_2(2d/d_0)$ rows. It is easy to verify that these rows are shattered by half-spaces whose supporting vectors w have a single non-zero ± 1 entry in each of the d_0 blocks and, therefore, $V \geq d_0 \log_2(2d/d_0)$.

Appendix B
Tighter bounds for low-noise condition

A. *Proof of Theorem 3*

Assumption (14) obviously implies Assumption (A) with $\delta = 1/2 - \Delta$. In addition, under (14), $Var(Y_i) = p_i(1-p_i) \leq (1/2-h)(1/2+h) = (1-4h^2)/4$. Hence, adapting the results of [2] on Kullback-Leibler risk in general GLM framework for logistic regression, the upper bound (6) for $EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}})$ can be improved:

$$\sup_{\boldsymbol{\beta}:||\boldsymbol{\beta}||_0 \leq d_0} EKL(\mathbf{p}, \widehat{\mathbf{p}}_{\widehat{M}}) \leq C\, \frac{1-4h^2}{1-4\Delta^2}\, \frac{\min\left(d_0 \ln \frac{de}{d_0}, r\right)}{n} \tag{26}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

and, therefore, from (9) we have

$$\sup_{\eta^*\in\mathcal{C}_X(d_0,h)} \mathcal{E}_X(\widehat{\eta}_{\widehat{M}},\eta^*) \le C_1 \sqrt{\frac{1-4h^2}{1-4\Delta^2}\ \frac{\min\left(d_0\ln\frac{de}{d_0},r\right)}{n}}$$

On the other hand, we can adapt the general Theorem 3 of [5] for $\psi(f) = (1/2)\left((1-f)\ln(1-f)+(1+f)\ln(1+f)\right) \ge f^2/2$ corresponding to the Kullback-Leibler risk ([28], Section 3.5), $\alpha=1$ corresponding to (14) and $c=1/(2h)$ to get

$$\mathcal{E}_X(\widehat{\eta}_{\widehat{M}},\eta^*) \le \frac{4}{h}\ EKL(\mathbf{p},\widehat{\mathbf{p}}_{\widehat{M}})$$

Applying (26) implies then

$$\mathcal{E}_X(\widehat{\eta}_{\widehat{M}},\eta^*) \le C_1\ \frac{1-4h^2}{1-4\Delta^2}\ \frac{\min\left(d_0\ln\frac{de}{d_0},r\right)}{nh}$$

B. Proof of Theorem 4

For any $\tilde{\eta}$ and $\eta^*\in\mathcal{C}(d_0,h)$ we have

$$\begin{aligned}\mathcal{E}_X(\tilde{\eta},\eta^*) &= \frac{1}{n}\sum_{i=1}^n P(\tilde{\eta}_i\neq\eta_i^*)|2p_i-1| \\ &\ge \frac{2h}{n}\ E\left(\sum_{i=1}^n I\{\tilde{\eta}_i\neq\eta_i^*\}\right) = \frac{2h}{n}E||\tilde{\eta}-\eta^*||_1\end{aligned} \tag{27}$$

for any X.

As we have mentioned, the worst case scenario for classification is when $p_i=1/2\pm h$ or, equivalently, $|\boldsymbol{\beta}^t\mathbf{x}_i| = \ln\left(\frac{1+2h}{1-2h}\right)$. Let $V=d_0\log_2(2d/d_0)$. In the proof of Lemma 1 we constructed explicitly the matrix $W_{V\times d}$ whose rows $\mathbf{w}_1,\ldots,\mathbf{w}_V$ are shattered by $\mathcal{C}(d_0)$. Then, for any $\mathbf{p} = \{\frac{1}{2}\pm h\}^V$ there exists $\boldsymbol{\beta}\in\mathbb{R}^d$ such that $||\boldsymbol{\beta}||_0\le d_0$ and $\boldsymbol{\beta}^t\mathbf{w}_i = \ln\frac{p_i}{1-p_i} = \pm\ln\frac{1+2h}{1-2h}$ for all $i=1,\ldots,V$. Define also the corresponding binary vector $\mathbf{b}$ with $b_i = I\{\boldsymbol{\beta}^t\mathbf{w}_i\ge 0\}$, that is, $b_i=1$ if $p_i=\frac{1}{2}+h$ and $b_i=0$ if $p_i=\frac{1}{2}-h$. Obviously, the set of all $\mathbf{b}$'s is a hypercube $H^V=\{0,1\}^V$.

Define now a $n\times d$ design matrix X_0 with $\varkappa$ rows of $\mathbf{w}_1$, $\varkappa$ rows of $\mathbf{w}_2$, ..., $\varkappa$ rows of $\mathbf{w}_{V-1}$ and the remaining $n-(V-1)\varkappa$ rows of $\mathbf{w}_V$, where an integer $1\le\varkappa\le\lfloor\frac{n}{V-1}\rfloor$ will be defined later.

The proof will now follow the general scheme of the proof of Theorem 4 of [18] but with necessary modifications for the fixed design.

For any $\mathbf{p}\in\{\frac{1}{2}\pm h\}^V$ and the corresponding $\mathbf{b}\in H^V$ define an n-dimensional indicator vector $\eta_{\mathbf{b}} = (\underbrace{b_1,\ldots,b_1}_{\varkappa},\ldots,\underbrace{b_{V-1},\ldots,b_{V-1}}_{\varkappa},\underbrace{b_V,\ldots,b_V}_{n-(V-1)\varkappa})$ and let $\tilde{\mathcal{C}}_{X_0}(d_0,h) = \{\eta_{\mathbf{b}},\ \mathbf{b}\in H^V\}$. By its design, $\tilde{\mathcal{C}}_{X_0}(d_0,h)\subseteq\{\eta:\eta\in\mathcal{C}(d_0),\ |\boldsymbol{\beta}^t\mathbf{x}_{0i}| = \ln\frac{1+2h}{1-2h},\ i=1,\ldots,n\}\subseteq \mathcal{C}_{X_0}(d_0,h)$.

Hence, we can reduce the minimax risk over the entire $\mathcal{C}_{X_0}(d_0,h)$ to $\tilde{\mathcal{C}}_{X_0}(d_0,h)$:

$$\inf_{\tilde{\eta}}\sup_{\eta^*\in\mathcal{C}_{X_0}(d_0,h)}\mathcal{E}_{X_0}(\tilde{\eta},\eta^*) \ge \inf_{\tilde{\eta}}\sup_{\eta^*\in\tilde{\mathcal{C}}_{X_0}(d_0,h)}\mathcal{E}_{X_0}(\tilde{\eta},\eta^*) \tag{28}$$

Furthermore, for a given $\tilde{\eta}$, define $\tilde{\eta}^* = \arg\min_{\eta\in\tilde{\mathcal{C}}(d_0,h)}||\tilde{\eta}-\eta||_1$. Then, for any $\eta^*\in\tilde{\mathcal{C}}_{X_0}(d_0,h)$ we have

$$||\tilde{\eta}^*-\eta^*||_1 \le ||\tilde{\eta}^*-\tilde{\eta}||_1 + ||\tilde{\eta}-\eta^*||_1 \le 2||\tilde{\eta}-\eta^*||_1 \tag{29}$$

and, therefore, from (27)-(29)

$$\begin{aligned}&\inf_{\tilde{\eta}}\sup_{\eta^*\in\mathcal{C}_{X_0}(d_0,h)}\mathcal{E}_{X_0}(\tilde{\eta},\eta^*) \\ &\quad\ge \frac{h}{n}\inf_{\tilde{\eta}^*\in\tilde{\mathcal{C}}_{X_0}(d_0,h)}\ \sup_{\eta^*\in\tilde{\mathcal{C}}_{X_0}(d_0,h)} E||\tilde{\eta}^*-\eta^*||_1 \\ &\quad\ge \frac{h}{n}\ \varkappa\inf_{\tilde{\mathbf{b}}\in H^V}\sup_{\mathbf{b}^*\in H^V} E\left(\sum_{i=1}^{V-1} I\{\tilde{b}_i\neq b_i^*\}\right),\end{aligned} \tag{30}$$

where $\tilde{\mathbf{b}},\mathbf{b}^*\in H^V$ are the binary vectors corresponding to $\tilde{\eta}^*$ and η^* respectively (see above).

By a simple calculus one can verify that the square Hellinger distance $H^2\left(Bin(1,\frac{1}{2}+h),Bin(1,\frac{1}{2}-h)\right)$ between two Bernoulli distributions $Bin(1,\frac{1}{2}+h)$ and $Bin(1,\frac{1}{2}-h)$ is $1-\sqrt{1-4h^2}$. For any $\mathbf{b}\in H^V$ and the corresponding $\eta_{\mathbf{b}}$ define $\mathbf{p}_{\mathbf{b}}\in\mathbb{R}^n$ as follows: $p_{\mathbf{b}i}=\frac{1}{2}+h$ if $\eta_{\mathbf{b}i}=1$ and $p_{\mathbf{b}i}=\frac{1}{2}-h$ if $\eta_{\mathbf{b}i}=0,\ i=1,\ldots,\varkappa(V-1)$, and $p_{\mathbf{b}i}=0,\ i=\varkappa(V-1)+1,\ldots,n$. Then, for any $\mathbf{b}_1,\mathbf{b}_2\in H^V$ and the corresponding $\mathbf{p}_{\mathbf{b}_1}$ and $\mathbf{p}_{\mathbf{b}_2}$ we have

$$\begin{aligned}H^2(\mathbf{p}_{\mathbf{b}_1},\mathbf{p}_{\mathbf{b}_2}) &= \frac{1}{n}\sum_{i=1}^n H^2\left(Bin(1,p_{b_{1i}}),Bin(1,p_{b_{2i}})\right) \\ &= \frac{\varkappa}{n}(1-\sqrt{1-4h^2})\sum_{i=1}^{V-1} I\{b_{1i}\neq b_{2i}\}\end{aligned}$$

Hence, applying the version of Assouad's lemma given in Lemma 7 of [4] yields

$$\begin{aligned}&\inf_{\tilde{\mathbf{b}}\in H^V}\sup_{\mathbf{b}^*\in H^V} E\left(\sum_{i=1}^{V-1} I\{\tilde{b}_i\neq b_i^*\}\right) \\ &\quad\ge \frac{V-1}{2}\left(1-\sqrt{2\varkappa(1-\sqrt{1-4h^2}}\right) \\ &\quad\ge \frac{V-1}{2}\left(1-\sqrt{8\varkappa h^2}\right)\end{aligned}$$

that together with (30) implies

$$\inf_{\tilde{\eta}}\sup_{\eta^*\in\mathcal{C}_{X_0}(d_0,h)}\mathcal{E}_{X_0}(\tilde{\eta},\eta^*) \ge \varkappa\frac{h}{n}\frac{V-1}{2}\left(1-\sqrt{8\varkappa h^2}\right) \tag{31}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

Consider two cases.

Case 1. $h \leq \frac{1}{6}$.

For $h \geq \sqrt{\frac{V-1}{18n}}$, apply (31) for $\varkappa = \lfloor \frac{1}{18h^2} \rfloor$ (note that $2 \leq \varkappa \leq \lfloor \frac{n}{V-1} \rfloor$), to get

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{X_0}(d_0,h)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \geq \frac{V-1}{216nh} \geq C_2 \frac{d_0 \ln(\frac{de}{d_0})}{nh}$$

For $h < \sqrt{\frac{V-1}{18n}}$, one can follow all the above arguments for $\tilde{h} = \sqrt{\frac{V-1}{18n}}$ and the corresponding $\varkappa = \lfloor \frac{n}{V-1} \rfloor$ to have

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{X_0}(d_0,h)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \geq \inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{X_0}(d_0,\tilde{h})} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \geq C_2 \sqrt{\frac{d_0 \ln(\frac{de}{d_0})}{n}}$$

Case 2. $h > \frac{1}{6}$.

Set $\varkappa = 1$ and note that $C_{X_0}(d_0, \frac{1}{2}) \subseteq C_{X_0}(d_0, h)$ for any $0 \leq h \leq \frac{1}{2}$. Hence, (30) implies

$$\begin{aligned} \inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{X_0}(d_0,h)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) &\geq \inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{X_0}(d_0,\frac{1}{2})} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \\ &\geq \frac{1}{2n} \inf_{\tilde{\mathbf{b}} \in H^V} \sup_{\mathbf{b}^* \in H^V} E\left(\sum_{i=1}^{V-1} I\{\tilde{b}_i \neq b_i^*\}\right) \\ &\geq \frac{1}{2n} \inf_{\tilde{\mathbf{b}} \in H^V} \frac{1}{2^V} \sum_{\mathbf{b}_j \in H^V} E\left(\sum_{i=1}^{V-1} I\{\tilde{b}_i \neq b_{ji}\}\right) \\ &= \frac{1}{2n} \inf_{\tilde{\mathbf{b}} \in H^V} \sum_{i=1}^{V-1} \frac{1}{2^V} \sum_{j=1}^{2^V} P(\tilde{b}_i \neq b_{ji}) \end{aligned}$$

By obvious combinatoric calculus, for any (binary) vector $\tilde{\mathbf{b}}$, $\frac{1}{2^V} \sum_{j=1}^{2^V} P(\tilde{b}_i \neq b_{ji}) = \frac{1}{2}$ for any i and, therefore,

$$\inf_{\tilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{X_0}(d_0,h)} \mathcal{E}_{X_0}(\tilde{\eta}, \eta^*) \geq \frac{V-1}{4n} \geq C_2 \frac{d_0 \ln \frac{de}{d_0}}{nh}$$

for large $h > \frac{1}{6}$ (in fact, larger than any fixed h_0).

Appendix C
Slope estimator for a general GLM

Consider a GLM setup with a response variable Y and a set of d predictors $x_1, ..., x_d$. We observe a series of independent observations $(\mathbf{x}_i, Y_i)$, $i = 1, \ldots, n$, where the design points $\mathbf{x}_i \in \mathbb{R}^p$ are deterministic. 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:

$$f_{\theta_i}(y) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a} + c(y, a)\right\} \tag{32}$$

The function $b(\cdot)$ is assumed to be twice-differentiable. In this case $\mathbb{E}(Y_i) = b'(\theta_i)$ and $Var(Y_i) = ab''(\theta_i)$. 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. The logistic regression (1) is a particular case of a general GLM (32) for the Bernoulli distribution $Bin(1, p_i)$, where the natural parameter is $\theta = \ln \frac{p}{1-p}$, $b(\theta_0) = \ln(1+e^\theta)$ and $a = 1$.

Following [2] assume the extended version of Assumption (A) for GLM :

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 $0 < \mathcal{L} \leq \mathcal{U} < \infty$ such that the function $b''(\cdot)$ satisfies the following conditions:*
   a) $\sup_{t \in \mathbb{R}} b''(t) \leq \mathcal{U}$
   b) $\inf_{t \in \Theta} b''(t) \geq \mathcal{L}$

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 the binomial distribution, $\mathcal{U} = 1/4$ and Assumption (A') reduces to Assumption (A) with $\mathcal{L} = \delta(1-\delta)$.

Recall that the Slope estimator is a penalized maximum likelihood with an ordered l_1-norm penalty and, therefore, defined for a GLM as follows:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{Slope} &= \arg\min_{\widetilde{\boldsymbol{\beta}}} \left\{-\ell(\widetilde{\boldsymbol{\beta}}) + \sum_{j=1}^d \lambda_j |\widetilde{\boldsymbol{\beta}}|_{(j)}\right\} \\ &= \arg\min_{\widetilde{\boldsymbol{\beta}}} \left\{b(X\widetilde{\boldsymbol{\beta}})^t \mathbf{1} - \mathbf{Y}^t X\widetilde{\boldsymbol{\beta}} + \sum_{j=1}^d \lambda_j |\widetilde{\boldsymbol{\beta}}|_{(j)}\right\} \end{aligned} \tag{33}$$

for $\lambda_1 \geq \cdots \geq \lambda_d > 0$. The corresponding Kullback-Leibler risk

$$\begin{aligned} \mathbb{E}KL(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_{Slope}) = &\frac{1}{n}\frac{1}{a}\Big(b'(\boldsymbol{\theta})^t(\boldsymbol{\theta} - \mathbb{E}(\widehat{\boldsymbol{\theta}}_{Slope})) \\ &-(b(\boldsymbol{\theta}) - \mathbb{E}b(\widehat{\boldsymbol{\theta}}_{Slope}))^t \mathbf{1}\Big) \end{aligned} \tag{34}$$

where $\boldsymbol{\theta} = X\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\theta}}_{Slope} = X\widehat{\boldsymbol{\beta}}_{Slope}$ (see [2]).

Theorem 8. *Consider a GLM (32), where $||\boldsymbol{\beta}||_0 \leq d_0$, the columns of the design matrix X are normalized to have unit norms and X satisfies the $WRE(d_0, c_0)$ condition for some $c_0 > 1$. Assume that Assumption (A') holds.*

Let

$$\lambda_j = A\, \frac{c_0 + 1}{c_0 - 1} \sqrt{\frac{\mathcal{U}}{a}}\, \sqrt{\ln(2d/j)}, \quad j = 1, \ldots, d, \tag{35}$$

in (33) with the constant $A \geq 40\sqrt{6}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

Then, simultaneously for all $\boldsymbol{\beta} \in \mathbb{R}^d$ such that $||\boldsymbol{\beta}||_0 \le d_0$,

1)

$$P\left(KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \le \frac{8A^2}{n}\frac{c_0^2}{(c_0-1)^2}\frac{\mathcal{U}}{\mathcal{L}} \times \max\left\{\left(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}}\right)^2, \frac{d_0}{\kappa^2(d_0,c_0)}\ln\left(\frac{2de}{d_0}\right)\right\}\right) \ge 1-\Delta \tag{36}$$

for any $0 < \Delta < 1$.

2)

$$\mathbb{E}KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \le 8A^2\frac{c_0^2}{(c_0-1)^2}\frac{\mathcal{U}}{\mathcal{L}}\left(\frac{2\pi+8}{\ln(2d)}+\frac{1}{\kappa^2(d_0,c_0)}\right)\frac{d_0}{n}\ln\left(\frac{2de}{d_0}\right) \tag{37}$$

Proof. Since $\widehat{\boldsymbol{\beta}}_{Slope}$ is the minimizer of (33),

$$-\ell(\widehat{\boldsymbol{\beta}}_{Slope})+\sum_{j=1}^d \lambda_j|\widehat{\boldsymbol{\beta}}_{Slope}|_{(j)} \le -\ell(\boldsymbol{\beta})+\sum_{j=1}^d \lambda_j|\boldsymbol{\beta}|_{(j)}$$

From (29) of [2] one has

$$n\,KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) = \ell(\boldsymbol{\beta})-\ell(\widehat{\boldsymbol{\beta}}_{Slope})+\frac{1}{a}(\mathbf{Y}-b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta})$$

(recall that the Kullback-Leibler divergence $KL(\cdot,\cdot)$ in [2] was defined as n times $KL(\cdot,\cdot)$ in this paper). Thus,

$$KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \le \frac{1}{n\,a}(\mathbf{Y}-b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}) + \frac{1}{n}\left(\sum_{j=1}^d \lambda_j|\boldsymbol{\beta}|_{(j)}-\sum_{j=1}^d \lambda_j|\widehat{\boldsymbol{\beta}}_{Slope}|_{(j)}\right) \tag{38}$$

Let $\mathbf{u}=\widehat{\boldsymbol{\beta}}_{Slope}-\boldsymbol{\beta}$. Applying Lemma A.1 of [6] with $\tau=0$ implies

$$\sum_{j=1}^d \lambda_j|\boldsymbol{\beta}|_{(j)}-\sum_{j=1}^d \lambda_j|\widehat{\boldsymbol{\beta}}_{Slope}|_{(j)} \le \sqrt{\sum_{j=1}^{d_0}\lambda_j^2}\;||\mathbf{u}||_2 - \sum_{j=d_0+1}^d \lambda_j|u|_{(j)} \tag{39}$$

Consider now the first term of the RHS in (38). Since the distribution of Y belongs to the exponential family with the bounded variance $ab''(\theta) \le a\mathcal{U}$ (Assumption (A')), a centered zero mean random variable $Y-b'(\theta)$ is sub-Gaussian with the scale factor $\sqrt{a\mathcal{U}}$, that is, $\mathbb{E}e^{t(Y-b'(\theta))} \le e^{a\mathcal{U}t^2/2}$ and, therefore, $\mathbb{E}e^{(Y-b'(\theta))^2/(6\mathcal{U}a)} \le e$. Applying Theorem 9.1 of [6] (adapted to our normalization conditions on the columns of X) yields

$$\frac{1}{na}(\mathbf{Y}-b'(\boldsymbol{\theta}))^t(\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}) \le \frac{40\sqrt{6\mathcal{U}}}{n\sqrt{a}} \times \max\left(\sum_{j=1}^d |u|_{(j)}\sqrt{\ln(2d/j)}\;,\; ||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}})\right) \tag{40}$$

with probability at least $1-\Delta$.

Set

$$H(\mathbf{u}) = \sum_{j=1}^d |u|_{(j)}\sqrt{\ln(2d/j)} \le ||\mathbf{u}||_2\sqrt{\sum_{j=1}^{d_0}\ln(2d/j)}+\sum_{j=d_0+1}^d |u|_{(j)}\sqrt{\ln(2d/j)} = \tilde{H}(\mathbf{u}) \tag{41}$$

and

$$G(\mathbf{u}) = ||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2\left(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}}\right) \tag{42}$$

The proof will now go along the lines of the proof of Theorem 6.1 of [6] for Gaussian regression with necessary adaptations to GLM and different normalization conditions on the columns of X.

To prove (36) consider two cases.

Case 1. $\tilde{H}(\mathbf{u}) \le G(\mathbf{u})$. In this case

$$||u||_2 \le \frac{||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}})}{\sqrt{\sum_{j=1}^{d_0}\ln(2d/j)}}$$

and, therefore, combining (35) and (38)-(42) with probability at least $1-\Delta$ yields

$$\begin{aligned} KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) &\le \frac{1}{n}A\sqrt{\frac{\mathcal{U}}{a}}\frac{2c_0}{c_0-1}\;||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}}) \\ &\le \frac{1}{2n}\left(\frac{A^2\mathcal{U}}{\epsilon a}\left(\frac{2c_0}{c_0-1}\right)^2(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}})^2 + \epsilon||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2^2\right) \end{aligned} \tag{43}$$

for any $\epsilon > 0$.

Lemma 1 of [2] established the equivalence of the Kullback-Leibler divergence $KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope})$ and the squared quadratic norm $||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||^2$ under Assumption (A'):

$$\frac{\mathcal{L}}{2a}||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2^2 \le nKL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \le \frac{\mathcal{U}}{2a}||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2^2 \tag{44}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2018.2884963, IEEE Transactions on Information Theory

Hence, taking $\epsilon = \mathcal{L}/(2a)$ in (43) after a straightforward calculus yields

$$KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \le \frac{8}{n}\frac{c_0^2}{(c_0-1)^2}\,\frac{\mathcal{U}}{\mathcal{L}}\;A^2(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}})^2 \tag{45}$$

with probability at least $1-\Delta$.

Case 2. $\tilde{H}(\mathbf{u}) > G(\mathbf{u})$. Using the definition of λ_j's in (35) and (38)-(42), with probability at least $1-\Delta$ we have

$$\begin{aligned}
&KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \le \\
&\quad \frac{1}{n}\,40\sqrt{\frac{6\mathcal{U}}{a}}\left(||\mathbf{u}||_2\sqrt{\sum_{j=1}^{d_0}\ln(2d/j)}\right.\\
&\quad \left.+\sum_{j=d_0+1}^{d}|u|_{(j)}\sqrt{\ln(2d/j)}\right)\\
&\quad +\frac{1}{n}\left(\sqrt{\sum_{j=1}^{d_0}\lambda_j^2}\;||\mathbf{u}||_2-\sum_{j=d_0+1}^{d}\lambda_j|u|_{(j)}\right)\\
&\quad \le \frac{1}{n}\left(\frac{2c_0}{c_0+1}||\mathbf{u}||_2\sqrt{\sum_{j=1}^{d_0}\lambda_j^2}-\frac{2}{c_0+1}\sum_{j=d_0+1}^{d}\lambda_j|u|_{(j)}\right)
\end{aligned} \tag{46}$$

The $KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \ge 0$ and, therefore, the RHS of (46) is necessarily positive. Thus,

$$\begin{aligned}
&\sum_{j=1}^{d}|u|_{(j)}\sqrt{\ln(2d/j)} \le ||\mathbf{u}||_2\sqrt{\sum_{j=1}^{d_0}\ln(2d/j)}\\
&\quad +\sum_{j=d_0+1}^{d}|u|_{(j)}\sqrt{\ln(2d/j)} \le (1+c_0)||\mathbf{u}||\sqrt{\sum_{j=1}^{d_0}\ln(2d/j)}
\end{aligned}$$

and, therefore, by $WRE(d_0,c_0)$ condition, (46) implies

$$\begin{aligned}
KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) &\le \frac{1}{n}\,\frac{2c_0}{c_0+1}\;||\mathbf{u}||_2\sqrt{\sum_{j=1}^{d_0}\lambda_j^2}\\
&\le \frac{1}{n}\,\frac{2c_0}{c_0+1}\frac{||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2}{\kappa(c_0,d_0)}\sqrt{\sum_{j=1}^{d_0}\lambda_j^2}\\
&\le \frac{1}{n}\left(\frac{c_0^2}{\epsilon(c_0+1)^2}\,\frac{\sum_{j=1}^{d_0}\lambda_j^2}{\kappa^2(c_0,d_0)}+\epsilon||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2^2\right)
\end{aligned}$$

for any $\epsilon > 0$. Taking $\epsilon = \mathcal{L}/(4a)$ and exploiting the equivalence between $KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope})$ and $||\widehat{\boldsymbol{\theta}}_{Slope}-\boldsymbol{\theta}||_2^2$ in (44) imply that with probability at least $1-\Delta$,

$$\begin{aligned}
KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) &\le \frac{1}{n}\,\frac{1}{\mathcal{L}}\,\frac{8ac_0^2}{(c_0+1)^2}\frac{\sum_{j=1}^{d_0}\lambda_j^2}{\kappa^2(c_0,d_0)}\\
&\le \frac{8}{n}\frac{c_0^2}{(c_0-1)^2}\frac{\mathcal{U}}{\mathcal{L}}\;A^2\frac{d_0\ln(2de/d_0)}{\kappa^2(c_0,d_0)},
\end{aligned}$$

where we used the definition (35) of λ_j's and the upper bound $\sum_{j=1}^{d_0}\ln(2d/j) \le d_0\ln(2ed/d_0)$ (see, e.g., (2.7) of [6]).

To prove the second statement (37) of the theorem denote $C^* = 8A^2\frac{c_0^2}{(c_0-1)^2}$ and note that

$$\begin{aligned}
&C^*\frac{1}{n}\,\frac{\mathcal{U}}{\mathcal{L}}\max\left\{\left(\sqrt{\pi/2}+\sqrt{2\ln\Delta^{-1}}\right)^2,\frac{d_0}{\kappa^2(d_0,c_0)}\ln\left(\frac{2de}{d_0}\right)\right\}\\
&\le\; C^*\frac{1}{n}\,\frac{\mathcal{U}}{\mathcal{L}}\max\left\{2\pi,\;8\ln\Delta^{-1},\;\frac{d_0}{\kappa^2(d_0,c_0)}\ln\left(\frac{2de}{d_0}\right)\right\}\\
&\le\; C^*\frac{1}{n}\,\frac{\mathcal{U}}{\mathcal{L}}\\
&\times\max\left\{\max\left(\frac{2\pi}{\ln(2d)},\;\frac{1}{\kappa^2(d_0,c_0)}\right)d_0\ln\left(\frac{2de}{d_0}\right),8\ln\Delta^{-1}\right\}\\
&\le\; C^*\frac{1}{n}\,\frac{\mathcal{U}}{\mathcal{L}}\\
&\times\max\left\{\left(\frac{2\pi}{\ln(2d)}+\frac{1}{\kappa^2(d_0,c_0)}\right)d_0\ln\left(\frac{2de}{d_0}\right),\;8\ln\Delta^{-1}\right\}
\end{aligned} \tag{47}$$

Then, by integrating, (36) and (47) after a straightforward calculus yield

$$\begin{aligned}
\mathbb{E}KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) &= \int_0^\infty P\left(KL(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_{Slope}) \ge t\right)dt\\
&\le C^*\frac{1}{n}\,\frac{\mathcal{U}}{\mathcal{L}}\left(\left(\frac{2\pi}{\ln(2d)}+\frac{1}{\kappa^2(d_0,c_0)}\right)d_0\ln\left(\frac{2de}{d_0}\right)\right.\\
&\left.+8\left(\frac{2de}{d_0}\right)^{-\frac{d_0}{8}\max\{\frac{2\pi}{\ln(2d)},\kappa^{-2}(d_0,c_0)\}}\right)\\
&\le C^*\,\frac{\mathcal{U}}{\mathcal{L}}\left(\frac{2\pi+8}{\ln(2d)}+\frac{1}{\kappa^2(d_0,c_0)}\right)\frac{d_0}{n}\ln\left(\frac{2de}{d_0}\right)
\end{aligned}$$

□

Appendix D
Proof of Theorem 7

We first introduce several notations. Let $||g||_{L_2} = (\int_{\mathcal{X}} g^2(\mathbf{x})d\mathbf{x})^{1/2}$ be a standard L_2-norm of a function g and $||g||_{L_2(q)} = (\int_{\mathcal{X}} g^2(\mathbf{x})q(\mathbf{x})d\mathbf{x})^{1/2}$ be the L_2-norm of g weighted by the marginal distribution q of $\mathbf{X}$. In addition, the L_∞-norm $||g||_\infty = \sup_{\mathbf{x}\in\mathcal{X}}|g(\mathbf{x})|$.

Applying Theorem 3 of [5] for $\psi(g) = g^2/2$, Assumption (B1) implies that there exists $C > 0$ such that

$$\mathcal{E}(\hat{\eta}_{\widehat{M}},\eta^*) \le C\left(E||\hat{p}_{\widehat{M}}-p||^2_{L^2(q)}\right)^{\frac{\alpha+1}{\alpha+2}} \tag{48}$$

Furthermore, let $f_{\boldsymbol{\beta}}(\mathbf{x},y)$ be the joint distribution of $(\mathbf{X},Y)$ for a given $\boldsymbol{\beta}$, i.e. $f_{\boldsymbol{\beta}}(\mathbf{x},y) = p(\mathbf{x})^y(1-p(\mathbf{x}))^{1-y}q(\mathbf{x})$, where $p(\mathbf{x}) = \frac{\exp\{\boldsymbol{\beta}^t\mathbf{x}\}}{1+\exp\{\boldsymbol{\beta}^t\mathbf{x}\}}$. Consider the square Hellinger distance $H^2(Bin(1,p_1),Bin(1,p_2))$ between two Bernoulli distributions with success probabilities p_1 and p_2. It is easy to verify that $H^2(Bin(1,p_1),Bin(1,p_2)) \ge \frac{(p_1-p_2)^2}{4(1-\delta)} \ge \frac{(p_1-p_2)^2}{2}$.

Then for the square Hellinger distance $d_H^2(f_{\boldsymbol{\beta}_1}, f_{\boldsymbol{\beta}_2})$ between $f_{\boldsymbol{\beta}_1}$ and $f_{\boldsymbol{\beta}_2}$ we have

$$d_H^2(f_{\boldsymbol{\beta}_1}, f_{\boldsymbol{\beta}_2}) = \int H^2(Bin(1, p_1(\mathbf{x})), Bin(1, p_2(\mathbf{x}))q(\mathbf{x})d\mathbf{x} \geq \frac{1}{2}||p_1 - p_2||^2_{L^2(q)} \tag{49}$$

and from (48) it is, therefore, sufficient to bound the Hellinger risk $Ed_H^2(f_{\widehat{\boldsymbol{\beta}}_{\widehat{M}}}, f_{\boldsymbol{\beta}})$.

We will show that the penalty (23) falls within a general class of penalties considered in [27] and then apply their Theorem 1 to find an upper bound for $Ed_H^2(f_{\widehat{\boldsymbol{\beta}}_{\widehat{M}}}, f_{\boldsymbol{\beta}})$.

Using the standard inequality $\ln(1+t) \leq t$, under Assumption (A1) we have

$$\begin{aligned} &|\ln f_{\boldsymbol{\beta}_2}(\mathbf{x}, y) - \ln f_{\boldsymbol{\beta}_1}(\mathbf{x}, y)| \\ &= \left| y \ln \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} + (1-y) \ln \frac{1 - p_2(\mathbf{x})}{1 - p_1(\mathbf{x})} \right| \\ &\leq \max \left(\left| \ln \frac{p_2(\mathbf{x})}{p_1(\mathbf{x})} \right|, \left| \ln \frac{1 - p_2(\mathbf{x})}{1 - p_1(\mathbf{x})} \right| \right) \\ &\leq \frac{1}{\delta} |p_2(\mathbf{x}) - p_1(\mathbf{x})| \end{aligned}$$

Define $\rho(f_{\boldsymbol{\beta}_1}, f_{\boldsymbol{\beta}_2}) = || \ln f_{\boldsymbol{\beta}_2} - \ln f_{\boldsymbol{\beta}_1} ||_\infty$. Thus,

$$\rho(f_{\boldsymbol{\beta}_1}, f_{\boldsymbol{\beta}_2}) \leq \frac{1}{\delta} ||p_2 - p_1||_\infty \tag{50}$$

For a given model M consider the set of coefficients $\mathcal{B}_M$ defined in Section II. One can easily verify that under Assumption (A1), for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}_M$ and the corresponding $p_1(\mathbf{x}), p_2(\mathbf{x})$

$$\delta(1-\delta) \left| (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^t \mathbf{x} \right| \leq |p_2(\mathbf{x}) - p_1(\mathbf{x})| \leq \frac{1}{4} \left| (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^t \mathbf{x} \right| \tag{51}$$

for any $\mathbf{x} \in \mathcal{X}$.

In particular, (51) implies

$$\begin{aligned} ||p_2(\mathbf{x}) - p_1(\mathbf{x})||_{L_2(q)} &\geq \delta(1-\delta) \sqrt{(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)^t G (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)} \\ &\geq \delta(1-\delta) \sqrt{\lambda_{min}(G)} \; ||\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||_2, \end{aligned} \tag{52}$$

where recall that $G = E(\mathbf{X}\mathbf{X}^t)$ and $\lambda_{min}(G) > 0$ is its minimal eigenvalue.

For each $\boldsymbol{\beta}_0 \in \mathcal{B}_M$ consider the corresponding Hellinger ball $\mathcal{H}_{f_{\boldsymbol{\beta}_0}, r} = \{ f_{\boldsymbol{\beta}} : d_H(f_{\boldsymbol{\beta}}, f_{\boldsymbol{\beta}_0}) \leq r, \;\; \boldsymbol{\beta} \in \mathcal{B}_M \}$. From (49) and (52) it then follows that if $f_{\boldsymbol{\beta}} \in \mathcal{H}_{f_{\boldsymbol{\beta}_0}, r}$, the corresponding $\boldsymbol{\beta} \in \mathcal{B}_M$ lies in the Euclidean ball $\mathcal{B}_{\boldsymbol{\beta}_0, r'} = \{ \boldsymbol{\beta} \in \mathbb{R}^{|M|} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 \leq r' \}$ with $r' = \frac{\sqrt{2}r}{\delta(1-\delta)\sqrt{\lambda_{min}(G)}}$.

Furthermore, for any $||\mathbf{x}||_2 \leq 1$, (51) and Cauchy-Schwarz inequality imply that $|p_2(\mathbf{x}) - p_1(\mathbf{x})| \leq \frac{1}{4} ||\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||_2$ and, therefore, by (50)

$$\rho(f_{\boldsymbol{\beta}_1}, f_{\boldsymbol{\beta}_2}) \leq \frac{1}{4\delta} \; ||\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1||_2 \tag{53}$$

Let $N(\mathcal{B}_{\boldsymbol{\beta}_0, r'}, l_2, \epsilon)$ be the ϵ-covering number of $\mathcal{B}_{\boldsymbol{\beta}_0, r'}$ w.r.t. l_2-distance. It is well-known that $N(\mathcal{B}_{\boldsymbol{\beta}_0, r'}, l_2, \epsilon) \leq \left(1 + \frac{2r'}{\epsilon}\right)^{|M|} \leq \left(\frac{3r'}{\epsilon}\right)^{|M|}$ for any $\epsilon < r'$.

Thus, for the ϵ-covering number $N(\mathcal{H}_{f_{\boldsymbol{\beta}_0}, r}, \rho, \epsilon)$ of $\mathcal{H}_{f_{\boldsymbol{\beta}_0}, r}$ w.r.t. the distance $\rho(f_{\boldsymbol{\beta}_1}, f_{\boldsymbol{\beta}_2})$, from (53) we have

$$\begin{aligned} N(\mathcal{H}_{f_{\boldsymbol{\beta}_0}, r}, \rho, \epsilon) &\leq N(\mathcal{B}_{\boldsymbol{\beta}_0, r'}, l_2, 4\delta\epsilon) \\ &\leq \left(\frac{3\sqrt{2}}{4\delta^2(1-\delta)\sqrt{\lambda_{\min}(G)}} \; \frac{r}{\epsilon} \right)^{|M|} \end{aligned}$$

The considered family of sparse logistic regression models satisfies then Assumption 1 of [27] with $A_M = \frac{c}{\delta^2(1-\delta)\sqrt{\lambda_{\min}(G)}}$ for some $c > 0$ and $m_M = |M|$. Apply now their Theorem 1 for a penalized maximum likelihood model selection procedure (3) with a complexity penalty $Pen(|M|) = C_1 \; m_M \ln A_M + C_2 \cdot C_M \leq \tilde{C}_1 \ln\left(\frac{1}{\delta\lambda_{\min}(G)}\right) \;\; |M| + C_2 |M| \ln \frac{de}{|M|}$, where $C_M = |M| \ln \frac{de}{|M|}$, and the exact positive constants C_1 and C_2 are given in the paper. Thus,

$$Ed_H^2(f_{\widehat{\boldsymbol{\beta}}_{\widehat{M}}}, f_{\boldsymbol{\beta}}) \leq \tilde{C} \ln\left(\frac{1}{\delta\lambda_{\min}(G)}\right) \; \frac{Pen(d_0)}{n}$$

To complete the proof note that one can always find a constant C in the penalty (23) such that the resulting $Pen(|M|) = C \ln\left(\frac{1}{\delta\lambda_{\min}(G)}\right) |M| \ln \frac{de}{|M|} \geq \tilde{C}_1 \left(\frac{1}{\delta\lambda_{\min}(G)}\right) \; |M| + C_2 |M| \ln \frac{de}{|M|}$.

Acknowledgments

The work was supported by the Israel Science Foundation (ISF), grants ISF-820/13 and ISF-589/18. The authors would like to thank Noga Alon for his help in the proof of Lemma 1, Alexander Tsybakov for valuable remarks and Roi Granot for the real-data example.

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. (2016). Model selection and minimax estimation in generalized linear models. *IEEE Trans. Inf. Theory* **62**, 3721-3730.

[3] Audibert, J-Y. and Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *A*nn. Statist., **35**, 608–633.

[4] Barron, A., Birgé, L. and Massart P. (1999). Risk bounds for model selection via penalization. *Prob. Theory Relat. Fields*, **113**, 301-413.

[5] Bartlett, P.L., Jordan, M.I. and McAuliffe, J.D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, **101**, 138–156.

[6] Bellec, P.C., Lecué, G. and Tsybakov, A. (2018). Slope meets Lasso: improved oracle bounds and optimaility. *Ann. Statist.*, **46**, 3603–3642.

[7] Bickel, P. and Levina, E. (2004). Some theory for Fisher's discriminant function, 'naive Bayes', and some alternatives where there are more variables than observations. *Bernoulli*, **10**, 989-1010.

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

[9] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* **138**, 33–73.

[10] Bogdan, M., van den Berg, E., Sabatti, C., Su, W. and Candés, E. (2015). SLOPE – adaptive variable selection via convex programming. *Ann. Appl. Statist.*, **9**, 1103–1140.

[11] Boucheron, S., Bousquet, O., and Lugosh, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM: Prob. Statist.*, **9**, 323-375.

[12] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Methods, Theory and Applications*. Springer, Berlin.

[13] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

[14] Fan, J. and Fan, Y. (2008). High-dimensional classification using feature annealed independence rules. *Ann. Statist.*, **36**, 2605–2637.

[15] Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. CRC Press, Boca Raton.

[16] Koltchinskii, V. and Beznosova, O. (2005). Exponential convergence rates in classification. In Learning Theory. Lecture Notes in Comput. Sci. **3559**, 295-307. Springer, Berlin.

[17] Mammen, E. and Tsybakov, A. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 1808–1829.

[18] Massart, P. and Nédélec, E. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34**, 2326–2366.

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

[20] Painsky, A. and Wornell, G.W. (2018). On the universality of the logistic loss function. *2018 IEEE International Symposium on Information Theory*, 936–940.

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

[22] Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32**, 135–166.

[23] van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36**, 614–645.

[24] Vapnik, V.N. (2000). *The Nature of Statistical Learning*, 2nd ed. Springer, New York.

[25] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenon. *Electr. J. Statist.* **6**, 38–90.

[26] Yang, Y. (1999). Minimax nonparametric classification. Parts I and II. *IEEE Trans. Inf. Theory* **45**, 2271–2292.

[27] Yang, Y. and Barron, A.R. (1998) An asymptotic property of model selection criteria. *IEEE Trans. Inf. Theory* **44**, 95–116.

[28] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32**, 56–85.

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 a 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.