Classification by sparse generalized additive models

Felix Abramovich

Department of Statistics and Operations Research,
Tel Aviv University,
Israel
e-mail: felix@tauex.tau.ac.il

Abstract: We consider (nonparametric) sparse (generalized) additive models (SpAM) for classification. The design of a SpAM classifier is based on minimizing the logistic loss with a sparse group Lasso/Slope-type penalties on the coefficients of univariate additive components' expansions in orthonormal series (e.g., Fourier or wavelets). The resulting classifier is inherently adaptive to the unknown sparsity and smoothness. We show that under certain sparse group restricted eigenvalue condition it is nearly-minimax (up to log-factors) simultaneously across the entire range of analytic, Sobolev and Besov classes. The performance of the proposed classifier is illustrated on a simulated and a real-data examples.

Contents

1. Introduction

Consider a binary classification setup with a vector of features $\mathbf{X} \in \mathbb{R}^d$ and the outcome class label $Y|(\mathbf{X} = \mathbf{x}) \sim B(1, p(\mathbf{x}))$, where $p(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$. The quality of a classifier $\eta : \mathbb{R}^d \to \{0, 1\}$ is quantified by its misclassification error $R(\eta) = P(Y \neq \eta(\mathbf{X}))$ and the optimal Bayes classifier is $\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}$. However, the conditional probability $p(\mathbf{x})$ is typically unknown. One then estimates it from the data $D = (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ by some $\widehat{p}(\mathbf{x})$ and designs a plug-in classifier $\widehat{\eta} = I\{\widehat{p}(\mathbf{x}) \geq 1/2\}$ with the conditional risk $R(\widehat{\eta}) = P(Y \neq \widehat{\eta}(\mathbf{X})|D)$. The goodness of $\widehat{\eta}$ is measured by the misclassification excess risk $\mathcal{E}(\widehat{\eta}, \eta^*) = \mathbb{E}R(\widehat{\eta}) - R(\eta^*)$, where the expectation is w.r.t. the distribution of D.

Probably the most commonly used model for $p(\mathbf{x})$ is logistic regression, where it is assumed that the log-odds (logit)

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = \boldsymbol{\beta}^T \mathbf{x} \tag{1.1}$$

and $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector of (unknown) regression coefficients. The corresponding Bayes (linear) classifier for the model (1.1) is $\eta^*(\mathbf{x}) = I\{\boldsymbol{\beta}^T \mathbf{x} \geq 0\}$. Assuming $d < n$ and estimating $\boldsymbol{\beta}$ by the maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\beta}}$ yields the plug-in classifier $\widehat{\eta}_L(\mathbf{x}) = I\{\widehat{\boldsymbol{\beta}}^T \mathbf{x} \geq 0\}$. Logistic regression classification has been well-studied and is one of the main tools used by practitioners. In particular, it is well-known known that $\mathcal{E}(\widehat{\eta}_L, \eta^*) \sim \sqrt{d/n}$ which is rate-optimal (in the minimax sense) over all linear classifiers (see, e.g. [1]).

Yet, a linear model for $g(\mathbf{x})$ in (1.1) might be too restrictive and oversimplified in a variety of applications. Nonparametric models allow one more flexibility by assuming much more general assumptions on $g(\mathbf{x})$. One can use then various nonparametric techniques (e.g., kernel estimators, local polynomial regression, orthogonal series or nearest neighbours) to estimate the unknown $p(\mathbf{x})$ (or, equivalently, $g(\mathbf{x})$) and define a nonparametric plug-in classifier $\widehat{\eta}_N(\mathbf{x}) = I\{\widehat{p}(\mathbf{x}) \geq 1/2\}$. [26] showed that with the properly chosen tuning parameters, $\widehat{\eta}_N(\mathbf{x})$ achieves the minimax rates across various smoothness classes. Thus, assuming that the unknown $p(\mathbf{x})$ has a smoothness s, $\mathcal{E}(\widehat{\eta}_N, \eta^*) = O(n^{-\frac{s}{2s+d}})$.

A general nonparametric classifier $\widehat{\eta}_N$ suffers, however, from a "curse of dimensionality" problem, where the convergence rates rapidly slow down as d increases and the required sample size for consistent classification grows exponentially with d. To handle it one needs some additional assumptions on the model.

Note that classification is particularly challenging near the decision boundary $\{\mathbf{x} : p(\mathbf{x}) = 1/2\}$ or, equivalenlty, $\{\mathbf{x} : g(\mathbf{x}) = 0\}$, where it is especially hard to predict the class label accurately, while at points far from the boundary, it is easier. Thus, assume that for all $0 < h < h^*$, $P(|p(\mathbf{X}) - 1/2| \leq h) \leq Ch^\alpha$ for some $C > 0, \alpha \geq 0$ and $0 < h^* < 1/2$ known as the margin or low-noise condition [23]. The case $\alpha = 0$ essentially corresponds to no assumption, while the second extreme case $\alpha = \infty$ implies the existence of a hard margin of size h^* separating between two classes. [3] showed that under the above margin assumption, the

rates of misclassification excess risk of $\widehat{\eta}_N$ can be indeed improved. Assuming again that $p(\mathbf{x})$ has a smoothness s, under some mild conditions on the marginal distribution $\mathbb{P}_X$ of $\mathbf{X}$, the rate for the misclassification excess risk becomes of the order $n^{-\frac{s(1+\alpha)}{(2+\alpha)s+d}}$ approaching the parametric rate n^{-1} as $\alpha \to \infty$. It can even be reduced further under stronger assumptions on $\mathbb{P}_X$ ([3]).

Although the margin assumption improves the convergence rates, it provides only a partial solution to the curse of dimensionality problem since computational burden for fitting a general nonparametric classifier remains prohibitive even for relatively small and moderate d. In addition, it is difficult to interpret such a model, to perform feature selection, etc.

To overcome these challenges one needs additional structural assumptions on the model. A common approach is to consider (generalized) *additive* models (GAM)([11]). In particular, for logistic regression, assume that

$$g(\mathbf{x}) = \mu + \sum_{j=1}^{d} g_j(x_j), \tag{1.2}$$

where $g_j(\cdot)$'s are univariate "smooth" functions. To make the model (1.2) identifiable impose $\mathbb{E}g_j(X_j) = 0$ for all $j = 1, \ldots, d$. The GAM (1.2) is a natural nonparametric generalization of the linear logistic regression model (1.1). Additive models have become a standard tool in high-dimensional nonparametric regression and classification, and can be efficiently fitted by the backfitting algorithm [11]. Additivity assumption drastically improves the convergence rates of the resulting classifier $\widehat{\eta}_A$. Thus, adapting the results of [26] and [19] one can show that $\mathcal{E}(\widehat{\eta}_A, \eta^*) = O(\sqrt{d}\, n^{-\frac{s}{2s+1}})$ if all univariate components g_j are of smoothness s, or, more generally, $\mathcal{E}(\widehat{\eta}_A, \eta^*) = O\left(\sqrt{\sum_{j=1}^{d} n^{-\frac{2s_j}{2s_j+1}}}\right)$ if g_j are of different smoothness.

Nevertheless, in the era of "Big Data", the number of features d might be very large and even larger that the sample size n (large d, small n setups). The GAM (1.2) requires $d \ll n$ and cannot cope with a curse of dimensionality in this case. Reducing dimensionality of a feature space by selecting a sparse subset of significant active features becomes essential. Thus, consider a *sparse* generalized additive model (SpAM) assuming that

$$g(\mathbf{x}) = \mu + \sum_{j \in \mathcal{J}} g_j(x_j), \tag{1.3}$$

where $\mathcal{J} \subseteq \{1, \ldots, d\}$ is an (unknown) subset of active features of cardinality $|\mathcal{J}| = d_0$ that have a "significant" impact on the outcome. We are mainly interested in sparse setups, where $d_0 \ll \min(d, n)$.

SpAMs have been intensively studied in the context of nonparametric regression w.r.t. the quadratic risk, where a common approach is based on penalized least squares estimation with different combinations of sparsity-induced and regularized smoothness convex penalties. See, e.g., [16, 17, 21, 13, 19, 2] for various

SpAM regression estimators and their properties. Their comparison is discussed in [2] and [10]. [13] and [10] extended this approach to more general convex losses relevant, in particular, for the considered logistic regression model. However, their estimators are not adaptive to the (usually unknown) smoothness of g_j's.

Classification by sparse linear logistic regression was studied in [1]. They showed that the corresponding minimax misclassification excess risk is of the order $\sqrt{d_0 \ln(de/d_0)/n}$ and designed penalized MLE classifiers that achieve it.

In this paper we investigate (nonparametric) SpAM classifiers. We propose a SpAM estimator of g in (1.3) defined in the Fourier/wavelet domain w.r.t. the penalized logistic loss with *sparse group Lasso* or more general *sparse group Slope* penalties on the Fourier/wavelet coefficients. The estimator and the resulting plug-in classifier are inherently *adaptive* to the unknown sparsity and smoothness classes. We establish the minimax rates for the misclassification excess risk for SpAMs across analytic, Sobolev and Besov classes of functions and show that there is a phase transition between sparse and dense SpAMs. In particular, if g_j, $j \in \mathcal{J}$ has a smoothness s_j, the minimax misclassificitaion excess risk is of the order $\max\left(\sqrt{\frac{d_0 \ln(de/d_0)}{n}}, \sqrt{\sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}}\right)$. We prove that with a proper choice of tuning parameters and under certain restricted sparse group eigenvalue condition, the sparse group Lasso/Slope SpAM classifiers $\widehat{\eta}_{sgL}$ and $\widehat{\eta}_{sgS}$ are simultaneously nearly-minimax (up to log-factors) across the entire range of those classes.

The rest of the paper is organized as follows. Section 2 presents the SpAM model for classification and defines sparse group Lasso and Slope classifiers. In Section 3 we establish the minimaxity of the proposed SpAM classifiers across Sobolev, analytic and Besov classes. Their performance is illustrated on a simulated and a real-data examples in Section 4. All the proofs are given in Appendix.

2. SpAM classifier

Consider a sparse additive logistic regression model:

$$Y|(\mathbf{X} = \mathbf{x}) \sim B(1, p(\mathbf{x})), \tag{2.1}$$

where $\mathbf{X} \in \mathbb{R}^d$ is a vector of linearly independent features with a marginal probability distribution $\mathbb{P}_X$ on a bounded support $\mathcal{X}$. Without loss of generality assume that $\mathcal{X} \subseteq [0,1]^d$. The logit function $g(\mathbf{x}) = \ln \frac{p(\mathbf{x})}{1-p(\mathbf{x})} \in L_2(\mathcal{X})$ has a sparse additive form (1.3), where g_j lies in some class of functions $\mathcal{F}_j$ on the unit interval (e.g., Sobolev $H^{s_j}[0,1]$ or more general Besov $B^{s_j}_{p_j,q_j}[0,1]$), $\mathbb{E}g_j(X_j) = 0$ and the subset of active features $\mathcal{J}$ is a priori unknown.

Expand the unknown univariate functions g_j, $j = 1, \ldots, d$ in some orthonormal basis $\{\psi_\ell\}_{l=0}^\infty$ in $\mathcal{F}_j$ as

$$g_j(x_j) = \sum_{\ell=0}^\infty \beta_{j\ell} \psi_\ell(x_j),$$

where $\beta_{j\ell} = \int g_j(x_j)\psi_\ell(x_j)dx_j$. For simplicity of exposition we consider the same basis for all g_j's. Due to the identifiability conditions $\mathbb{E}g_j(X_j) = 0$, so we consider $\beta_{j0} = 0$ for all j. By Parseval's identity $||g_j||^2_{L_2[0,1]} = \sum_{l=1}^{\infty} |\beta_{j\ell}|^2$ and $||g||^2_{L_2[0,1]^d} = \mu^2 + \sum_{j=1}^{d} \sum_{l=1}^{\infty} |\beta_{j\ell}|^2$.

Estimate the unknown coefficients $\beta_{j\ell}$ from the data sample $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$. Obviously, for all $j = 1, \ldots, d$, we set $\widehat{\beta}_{j\ell} = 0$ for all $\ell > n$. To estimate $\beta_{j\ell}$ for $1 \leq \ell \leq n$ consider additive truncated estimators of the form

$$\widetilde{g}_n(\mathbf{x}) = \widetilde{\mu} + \sum_{j=1}^{d}\sum_{\ell=1}^{n} \widetilde{\beta}_{j\ell}\psi_\ell(x_j)$$

and the logistic loss

$$\begin{aligned}\ell(Y, \widetilde{g}_n(\mathbf{X})) &= \ln\left(1 + \exp(\widetilde{g}_n(\mathbf{X}))\right) - \widetilde{g}_n(\mathbf{X})Y \\ &= \ln\left(1 + \exp\left(\widetilde{\mu} + tr(\Psi(\mathbf{X})^T\widetilde{B})\right)\right) - \left(\widetilde{\mu} + tr(\Psi(\mathbf{X})^T\widetilde{B})\right)Y,\end{aligned} \quad (2.2)$$

where $\widetilde{B} \in \mathbb{R}^{d \times n}$ is the regression coefficients matrix and $\Psi(\mathbf{X}) \in \mathbb{R}^{d \times n}$ is the matrix with entries $\Psi_{j\ell}(\mathbf{X}) = \psi_\ell(X_j)$.

Let $B \in \mathbb{R}^{d \times n}$ be the matrix with the true (unknown) regression coefficients $\beta_{j\ell}$, $j = 1, \ldots, d$; $\ell = 1, \ldots n$. Sparse representation of g and smoothness properties of g_j's in (1.3) are directly expressed in terms of B. Evidently, B has zero rows for all $j \notin \mathcal{J}$. In addition, if the chosen basis $\{\psi_\ell\}_{l=0}^{\infty}$ allows sparse representation of g_j, $j \in \mathcal{J}$, such that they can be well-approximated by a small number of basis functions, then the decreasingly ordered $|\beta|_{j(\ell)}$ tend rapidly to zero as ℓ increases for $j \in \mathcal{J}$. These arguments induce a double row-wise sparsity structure of B: it has only d_0 non-zero rows (global row-wise sparsity) and even those are "approximately sparse" in the sense that they have only a few "large" entries (local row-wise approximate sparsity). To capture such type of sparsity we consider a *logistic sparse group Lasso* estimator of B. Sparse group Lasso was proposed in [22] for linear regression. Its logistic version was used in multiclass classification by multinomial linear logistic regression in [25] and [14]. Within the considered SpAM binary classification setup we define the logistic sparse group Lasso estimator $\widehat{B}_{sgL}$ of B as follows:

$$\begin{aligned}(\widehat{B}_{sgL}, \widehat{\mu}_{sgL}) &= \arg\min_{\widetilde{B},\widetilde{\mu}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\ell(Y_i, \widetilde{g}_n(\mathbf{X}_i)) + \lambda\sum_{j=1}^{d}|\widetilde{B}_{j\cdot}|_2 + \kappa\sum_{j=1}^{d}|\widetilde{B}_{j\cdot}|_1\right\} \\ &= \arg\min_{\widetilde{B},\widetilde{\mu}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\ln\left(1 + \exp\left(\widetilde{\mu} + tr(\Psi(\mathbf{X}_i)^T\widetilde{B})\right)\right)\right.\right. \\ &\quad\left.\left. - \left(\widetilde{\mu} + tr(\Psi(\mathbf{X}_i)^T\widetilde{B})\right)Y_i\right) + \lambda\sum_{j=1}^{d}|\widetilde{B}_{j\cdot}|_2 + \kappa\sum_{j=1}^{d}|\widetilde{B}_{j\cdot}|_1\right\},\end{aligned}$$

$$(2.3)$$

where $|\widetilde{B}_{j\cdot}|_2$ and $|\widetilde{B}_{j\cdot}|_1$ are respectively the l_2 and l_1-norms of a j-th row of $\widetilde{B}$, and $\lambda > 0$ and $\kappa > 0$ are tuning parameters.

The corresponding estimator

$$\widehat{g}_{sgL}(\mathbf{x}) = \widehat{\mu}_{sgL} + \sum_{j=1}^{d}\sum_{l=1}^{n}\widehat{\beta}_{sgL}\psi_{\ell}(x_j) \tag{2.4}$$

and the resulting logistic sparse group Lasso classifier

$$\widehat{\eta}_{sgL}(\mathbf{x}) = I\{\widehat{g}_{sgL}(\mathbf{x}) \geq 0\}. \tag{2.5}$$

A more flexible generalization of (2.3) is a *logistic sparse group Slope* estimator that utilizes *sorted* norms in the penalty:

$$\widehat{B}_{sgS} = \arg\min_{\widetilde{B},\widetilde{\mu}}\left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\ln\left(1 + \exp\left(\widetilde{\mu} + tr(\Psi(\mathbf{X}_i)^T\widetilde{B})\right)\right)\right.\right.$$
$$\left. - \left(\widetilde{\mu} + tr(\Psi(\mathbf{X}_i)^T\widetilde{B})\right)Y_i\right) \tag{2.6}$$
$$\left. + \sum_{j=1}^{d}\lambda_j|B|_{(j)2} + \sum_{j=1}^{d}\sum_{\ell=1}^{n}\kappa_{\ell}|\widetilde{B}|_{j(\ell)}\right\},$$

where with some ambiguity of notations $|\widetilde{B}|_{(1)2} \geq \ldots \geq |\widetilde{B}|_{(d)2}$ are the descendingly ordered l_2-norms of the rows of $\widetilde{B}$, $|\widetilde{B}|_{j(1)} \geq \ldots \geq |\widetilde{B}|_{j(n)}$ are the descendingly ordered absolute values of entries of its j-row, and $\lambda_1 \geq \ldots \geq \lambda_d > 0$ and $\kappa_1 \geq \ldots \geq \kappa_n > 0$ are sequences of tuning parameters. Evidently, $\widehat{B}_{sgL}$ is a particular case of $\widehat{B}_{sgS}$ with equal λ_j's and κ_{ℓ}'s.

Similarly to sparse group Lasso classification, we consider

$$\widehat{g}_{sgS}(\mathbf{x}) = \widehat{\mu}_{sgS} + \sum_{j=1}^{d}\sum_{l=1}^{n}\widehat{\beta}_{sgS}\psi_{\ell}(x_j)$$

and the corresponding sparse group Slope classifier

$$\widehat{\eta}_{sgS}(\mathbf{x}) = I\{\widehat{g}_{sgS}(\mathbf{x}) \geq 0\}. \tag{2.7}$$

3. Minimaxity across various function classes

We now show that with the proper choice of tuning parameters the proposed logistic sparse group Lasso and Slope classifiers are adaptively nearly-minimax (up to log-factors) across various classes of functions $\mathcal{F}_j$ for g_j's in (1.3).

As usual with convex penalization, one needs some (mild) assumptions on the design. Consider a truncated version g_n of g:

$$g_n(\mathbf{x}) = \mu + \sum_{j=1}^{d}\sum_{\ell=1}^{n}\beta_{j\ell}\psi_{\ell}(x_j).$$

Let random vectors $\boldsymbol{\Psi}(X_j) = (\psi_1(X_j), \ldots, \psi_n(X_j))^T$, $j = 1, \ldots, d$ be the rows of the matrix $\Psi(\mathbf{X})$ and define the cross-covariance matrices

$$V_{jk} = E_X(\boldsymbol{\Psi}(X_j)\boldsymbol{\Psi}(X_k)^T).$$

By simple algebra,

$$\int g_n(\mathbf{x})^2 d\mathbb{P}_X = \mu^2 + \sum_{j=1}^{d}\sum_{k=1}^{d} B_{j\cdot} V_{jk} B_{k\cdot}^T.$$

For a given $1 \leq d_0 \leq d$, denote a vector of integers $\mathbf{m} = (m_1, \ldots, m_{d_0}) \in \mathbb{R}^{d_0}$ with $m_j \in \{1, \ldots, n\}$.

Assumption 3.1. ($WRE(d_0, \mathbf{m}, c_0)$-condition). Given non-increasing sequences $\{\lambda_j\}_{j=1}^{d}$ and $\{\kappa_\ell\}_{\ell=1}^{n}$, consider the cone of matrices

$$\mathcal{S}(d_0, \mathbf{m}, c_0) = \left\{ A \in \mathbb{R}^{d \times n} : ||A||_{\boldsymbol{\lambda},\boldsymbol{\kappa}} \leq c_0 \left(\sqrt{\sum_{j=1}^{d_0} \lambda_j^2} + \sqrt{\sum_{j=1}^{d_0}\sum_{\ell=1}^{m_j} \kappa_\ell^2} \right) ||A||_F \right\},$$

where $||A||_{\boldsymbol{\lambda},\boldsymbol{\kappa}} = \sum_{j=1}^{d} \lambda_j |A|_{(j)2} + \sum_{j=1}^{d}\sum_{\ell=1}^{n} \kappa_\ell |A|_{j(\ell)}$ is the sparse group Slope norm of A and $||A||_F$ is its Frobenius norm, and assume that

$$\nu_{gS}(d_0) = \inf_{A \in \mathcal{S}(d_0, \mathbf{m}, c_0) \setminus \{0_{n \times d}\}} \frac{\sum_{j=1}^{d}\sum_{k=1}^{d} A_{j\cdot} V_{jk} A_{k\cdot}^T}{||A||_F^2} > 0.$$

The weighted restricted eigenvalue (WRE) condition 3.1 is an extension of the WRE condition introduced for Slope in [4] to sparse group Slope and random design. Such type of assumption is common for convex penalties (see [4] for discussion). In particular, for isotropic $\mathbf{X}$, due to orthogonality of the basis, $V_{jj} = I_n$ and $V_{jk} = 0_{n \times n}$ for $j \neq k$. Thus, $\sum_{j=1}^{d}\sum_{k=1}^{d} A_{j\cdot} V_{jk} A_{k\cdot}^T = ||A||_F^2$ and Assumption 3.1 is trivially satisfied.

Evidently, for sparse group Lasso with constant λ and κ

$$\mathcal{S}(d_0, \mathbf{m}, c_0) = \Big\{ A \in \mathbb{R}^{d \times n} : \lambda \sum_{j=1}^{d} |A_{j\cdot}|_2 + \kappa \sum_{j=1}^{n} |A_{j\cdot}|_1 \leq$$
$$c_0 \left(\lambda\sqrt{d_0} + \kappa\sqrt{\sum_{j=1}^{d_0} m_j} \right) ||A||_F \Big\}.$$

3.1. Sobolev classes

Consider the orthonormal cosine series $\psi_0(x_j) = 1$, $\psi_\ell(x_j) = \sqrt{2}\cos(\pi\ell x_j)$, $\ell = 1, 2, \ldots$ and let $\beta_{j\ell} = \sqrt{2}\int_0^1 g_j(t)\cos(\pi\ell t)dt$ be the cosine Fourier coefficients of g_j.

Suppose that g_j belongs to a periodic Sobolev class
$\widetilde{H}^{s_j}[0,1] = \left\{ g_j : \sum_{\ell=0}^{\infty} \ell^{2s_j} \beta_{j\ell}^2 \leq R_j, \ s_j > \frac{1}{2} \right\}$. In particular, for an integer s_j it is equivalent to

$$\left\{ g_j : \int_0^1 \left(g_j^{(s_j)}(t) \right)^2 dt \leq R_j, \ g_j^{(m)}(0) = g_j^{(m)}(1) = 0, \ m = 1, \ldots, s_j - 1 \right\}$$

(e.g., [24], Section 1.7.1).

Consider a set of sparse additive functions

$$\mathcal{G}_{\widetilde{H}}(d_0, \mathbf{s}) = \left\{ g(\mathbf{x}) \in L_2(\mathcal{X}) : \ g(\mathbf{x}) = \mu + \sum_{j \in \mathcal{J}} g_j(x_j), \ |\mathcal{J}| \leq d_0, \right.$$
$$\left. g_j \in \widetilde{H}^{s_j}[0,1], \ j \in \mathcal{J}, \ \mathbb{E}g_j(X_j) = 0 \right\}$$

and the corresponding set of logistic SpAM classifiers

$$\mathcal{C}_{\widetilde{H}}(d_0, \mathbf{s}) = \left\{ \eta(\mathbf{x}) = I\{g(\mathbf{x}) \geq 0\} : g(\mathbf{x}) \in \mathcal{G}_{\widetilde{H}}(d_0, \mathbf{s}) \right\}.$$

The following theorem establishes the upper bounds for the misclassification excess risk for logistic sparse group Lasso and Slope classifiers over periodic Sobolev classes:

Theorem 3.2. *Consider the SpAM (1.1)-(1.3), where $g \in \mathcal{G}_{\widetilde{H}}(d_0, \mathbf{s})$, and assume $WRE(d_0, \mathbf{m}, c_0)$-condition 3.1 with $m_j = n^{\frac{1}{2s_j+1}}$, $j = 1, \ldots, d_0$ and c_0 that can be derived from the proof.*

1. *Apply the sparse group Lasso classifier (2.5) with $\lambda = C_1 \sqrt{\frac{\ln d}{n}}$ and $\kappa = C_2 \sqrt{\frac{\ln n}{n}}$ for some $C_1, C_2 > 0$ that can be derived from the proof. Then,*

$$\sup_{\eta^* \in \mathcal{C}_{\widetilde{H}}(d_0, \mathbf{s})} \mathcal{E}(\widehat{\eta}_{sgL}, \eta^*) \leq C \sqrt{\frac{1}{\nu_{sg}(d_0)} \left(\frac{d_0 \ln d}{n} + \ln n \sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}} \right)}$$

for some $C > 0$.

2. *Apply the sparse group Slope classifier (2.7) with $\lambda_j = C_1 \sqrt{\frac{\ln(de/j)}{n}}$, $j = 1, \ldots, d$ and $\kappa_\ell = C_2 \sqrt{\frac{\ln(ne/\ell)}{n}}$, $\ell = 1, \ldots, n$. Then,*

$$\sup_{\eta^* \in \mathcal{C}_{\widetilde{H}}(d_0, \mathbf{s})} \mathcal{E}(\widehat{\eta}_{sgS}, \eta^*) \leq C \sqrt{\frac{1}{\nu_{sg}(d_0)} \left(\frac{d_0 \ln \left(\frac{de}{d_0} \right)}{n} + \ln n \sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}} \right)}.$$

The proposed SpAM classifiers are inherently adaptive to a sparsity parameter d_0, a set of active features $\mathcal{J}$ and smoothness parameters s_j's. Theorem 3.2 implies that a more flexible Slope-type penalty allows one to reduce the log-factor in the first term of the upper bounds. We now show that the upper bounds in Theorem 3.2 are nearly-minimax (up to log-factors) over $\mathcal{C}_{\widetilde{H}}(d_0, \mathbf{s})$:

Theorem 3.3. *Consider the SpAM (1.1)-(1.3), where $d_0 \ln(\frac{de}{d_0}) \leq n$ and $g \in \mathcal{G}_{\widetilde{H}}(d_0, \mathbf{s})$. Then,*

$$\inf_{\widetilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{\widetilde{H}}(d_0, \mathbf{s}), \mathbb{P}_X} \mathcal{E}(\widetilde{\eta}, \eta^*) \geq \tilde{C} \sqrt{\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n} + \sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j + 1}}}$$

for some $\tilde{C} > 0$, where the infimum is taken over all classifiers $\widetilde{\eta}$ based on the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$.

The obtained misclassification excess risk bounds contain two terms and indicate on a phase transition between sparse and dense SpAMs. The term $\frac{d_0 \ln(de/d_0)}{n} \sim \frac{1}{n} \ln \binom{d}{d_0}$ corresponds to the error of selecting d_0 nonzero univariate components g_j's out of d and is common for model selection. The term $\sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}$ is due to simultaneous nonparametric estimating d_0 functions $g_j \in \widetilde{H}^{s_j}([0,1])$. Assume for simplicity that $s_j = s$ for all $j \in \mathcal{J}$, so that the second term becomes $d_0 n^{-\frac{2s}{2s+1}}$. One then immediately realizes that the model selection error is dominating for sparse cases, where $\frac{d_0}{d} \lesssim e^{-n^{\frac{1}{2s+1}}}$, while for less sparse cases nonparametric estimation is the main error source. A similar phase transition phenomenon for sparse linear classification was shown in [1].

It can be shown that the exact minimax risk (without the extra $\ln n$-factor) can be achieved by the penalized SpAM estimator of [10] applied to the classification setup. However, it is not adaptive to smoothness, as both penalties and optimal values of the smoothing parameters depend on s_j's. For a particular case of design on a regular lattice, [2] proposed an adaptive SpAM estimator based on a certain complexity penalty on the number of nonzero rows of B and the numbers of their nonzero entries, with the exact minimax risk within the Gaussian additive models framework. In fact, sparse group Lasso/Slope can be viewed as a convex surrogate of their complexity penalty. We conjecture that under certain conditions the results of [2] can be extended to a general design and the logistic loss, but the use of a complexity penalty will make it computationally infeasible in this case.

3.2. Analytic functions

Assume now that g_j's are very smooth, where g_j belongs to a class of analytic functions $\mathcal{A}^{\alpha_j}[0,1] = \left\{ g_j : \sum_{\ell=0}^{\infty} e^{2\alpha_j \ell} \beta_{j\ell}^2 \leq R_j, \ \alpha_j > 0 \right\}$. Such functions admit analytic continuation into a band of width α_j of the complex plane (e.g., [9]) and, in particular, are infinitely differentiable.

Define a set of functions

$$\mathcal{G}_{\mathcal{A}}(d_0, \boldsymbol{\alpha}) = \left\{ g(\mathbf{x}) \in L_2(\mathcal{X}) : \ g(\mathbf{x}) = \mu + \sum_{j \in \mathcal{J}} g_j(x_j), \ |\mathcal{J}| \leq d_0, \right.$$
$$\left. g_j \in \mathcal{A}^{\alpha_j}[0,1], \ j \in \mathcal{J}, \ \mathbb{E}g_j(X_j) = 0 \right\}$$

and the corresponding set of SpAM classifiers

$$\mathcal{C}_{\mathcal{A}}(d_0, \boldsymbol{\alpha}) = \{\eta(\mathbf{x}) = I\{g(\mathbf{x}) \geq 0\} : g(\mathbf{x}) \in \mathcal{G}_{\mathcal{A}}(d_0, \boldsymbol{\alpha})\}.$$

We now show that the logistic sparse group Lasso classifier $\widehat{\eta}_{sgL}$ in (2.5) applied to the Fourier cosine coefficients $\beta_{j\ell}$'s with the same choice for tuning parameters as in Theorem 3.2 for Sobolev classes is also nearly-minimax over $\mathcal{C}_{\mathcal{A}}(d_0, \boldsymbol{\alpha})$, while the logistic sparse group Slope classifier $\widehat{\eta}_{sgS}$ in (2.7) achieves the exact minimax rate for these classes.

Theorem 3.4. *Consider the SpAM (1.1)-(1.3), where $g \in \mathcal{G}_{\mathcal{A}}(d_0, \boldsymbol{\alpha})$, and assume $WRE(d_0, \mathbf{m}, c_0)$-condition 3.1 with $m_j = \frac{1}{2\alpha_j} \ln\left(\frac{n}{\ln n}\right)$, $j = 1, \ldots, d_0$ and c_0 that can be derived from the proof.*

1. *Apply the sparse group Lasso classifier (2.5) with $\lambda = C_1 \sqrt{\frac{\ln d}{n}}$ and $\kappa = C_2 \sqrt{\frac{\ln n}{n}}$ for some $C_1, C_2 > 0$ that can be derived from the proof. Then,*

$$\sup_{\eta^* \in \mathcal{C}_{\mathcal{A}}(d_0, \boldsymbol{\alpha})} \mathcal{E}(\widehat{\eta}_{sgL}, \eta^*) \leq C \sqrt{\frac{1}{\nu_{sg}(d_0)} \left(\frac{d_0 \ln d}{n} + \sum_{j \in \mathcal{J}} \frac{1}{\alpha_j} \frac{\ln n}{n} \right)}$$

   for some $C > 0$.

2. *Apply the sparse group Slope classifier (2.7) with $\lambda_j = C_1 \sqrt{\frac{\ln(de/j)}{n}}$, $j = 1, \ldots, d$ and $\kappa_\ell = C_2 \sqrt{\frac{\ln(ne/\ell)}{n}}$, $\ell = 1, \ldots, n$. Then,*

$$\sup_{\eta^* \in \mathcal{C}_{\mathcal{A}}(d_0, \boldsymbol{\alpha})} \mathcal{E}(\widehat{\eta}_{sgS}, \eta^*) \leq C \sqrt{\frac{1}{\nu_{sg}(d_0)} \left(\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n} + \sum_{j \in \mathcal{J}} \frac{1}{\alpha_j} \frac{\ln n}{n} \right)}.$$

Theorem 3.5. *Consider the SpAM (1.1)-(1.3), where $d_0 \ln(\frac{de}{d_0}) \leq n$ and $g \in \mathcal{A}(d_0, \boldsymbol{\alpha})$. Then,*

$$\inf_{\widetilde{\eta}} \sup_{\eta^* \in \mathcal{C}_{\widetilde{H}}(d_0, \boldsymbol{\alpha}), \mathbb{P}_X} \mathcal{E}(\widetilde{\eta}, \eta^*) \geq \widetilde{C} \sqrt{\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n} + \sum_{j \in \mathcal{J}} \frac{1}{\alpha_j} \frac{\ln n}{n}}$$

for some $\widetilde{C} > 0$.

Similar to Sobolev classes there is the phase transition phenomenon for the misclassification excess risk. Assuming for simplicity equal $\alpha_j = \alpha$, the model selection error term $\frac{d_0 \ln(de/d_0)}{n}$ dominates for sparse cases where $\frac{d_0}{d} \lesssim n^{-\frac{1}{\alpha}}$.

3.3. Besov classes

Consider now more general Besov classes $B_{p,q}^s$ of functions. The precise formal definition of Besov spaces can be found, e.g. in [18]. On the intuitive level, (not

necessarily integer) s measures the number of function's derivatives, where their existence is required in the L_p-sense, while q provides a further finer gradation. The Besov spaces include, in particular, the traditional Sobolev H^s and Hölder C^s classes of smooth functions ($B_{2,2}^s$ and $B_{\infty,\infty}^s$ respectively) but also various classes of spatially inhomogeneous functions like functions of bounded variation, sandwiched between $B_{1,1}^1$ and $B_{1,\infty}^1$ ([18, 8]).

Besov classes can be equivalently characterized by coefficients of function's expansion in orthonormal wavelet series. Given a compactly supported scaling function ϕ of regularity r and the corresponding mother wavelet ψ, one generates an orthonormal wavelet basis on the unit interval with $\psi_{hk}(t) = 2^{h/2}\psi(2^h t - k)$, $h \geq 0$, $k = 0, \ldots, 2^h - 1$ (see [6]). Assume that $g_j \in B_{p_j,q_j}^{s_j}[0,1]$, $1 \leq p_j, q_j \leq \infty$, $(\frac{1}{p_j} - \frac{1}{2})_+ < s_j < r$ and let $\beta_{j,hk} = \int_0^1 g_j(t)\psi_{hk}(t)dt$ be its wavelet coefficients. Then,

$$\sum_{h=0}^{\infty} \left(2^{h(s_j+1/2-1/p_j)} \left(\sum_{k=0}^{2^h-1} |\beta_{j,hj}|^{p_j} \right)^{1/p_j} \right)^{q_j} \leq R_j \tag{3.1}$$

with l_{p_j} and/or l_{q_j} norms in (3.1) being replaced by the corresponding l_∞-norms for $p_j = \infty$ and/or $q_j = \infty$ (e.g., [18, 8]).

Moreover, let $\ell = 2^h + k$, $h \geq 0$, $0 \leq k \leq 2^h - 1$ and re-write the triple-indexed wavelet coefficients $\beta_{j,hk}$ in terms of double-indexed $\beta_{j\ell}$. Then,

$$|\beta|_{j(\ell)} \leq C\ell^{-(s_j+1/2)} \tag{3.2}$$

([7], Lemma 2).

Define the set of sparse additive functions

$$\mathcal{G}_{\mathcal{B}}(d_0, \mathbf{s}, \mathbf{p}, \mathbf{q}) = \left\{ g(\mathbf{x}) \in L_2(\mathcal{X}) : g(\mathbf{x}) = \mu + \sum_{j \in \mathcal{J}} g_j(x_j) \geq 0\}, \ |\mathcal{J}| \leq d_0, \right.$$
$$\left. g_j \in B_{p_j,q_j}^{s_j}[0,1], \ j \in \mathcal{J}, \ \mathbb{E}g_j(X_j) = 0 \right\},$$

where $1 \leq p_j, q_j \leq \infty$, $(\frac{1}{p_j} - \frac{1}{2})_+ < s_j < r$ for all $j \in \mathcal{J}$, and the corresponding set of SpAM classifiers

$$\mathcal{C}_{\mathcal{B}}(d_0, \mathbf{s}, \mathbf{p}, \mathbf{q}) = \{\eta(\mathbf{x}) = I\{g(\mathbf{x}) \geq 0\} : g(\mathbf{x}) \in \mathcal{G}_{\mathcal{B}}(d_0, \mathbf{s}, \mathbf{p}, \mathbf{q})\}.$$

Theorem 3.6. *Consider the SpAM (1.1)-(1.3), where $g \in \mathcal{G}_{\mathcal{B}}(d_0, \mathbf{s}, \mathbf{p}, \mathbf{q})$, and assume $WRE(d_0, \mathbf{m}, c_0)$-condition 3.1 with $m_j = n^{\frac{1}{2s_j+1}}$, $j = 1, \ldots, d_0$ and c_0 that can be derived from the proof.*

1. *Apply the sparse group Lasso classifier (2.5) with $\lambda = C_1\sqrt{\frac{\ln d}{n}}$ and $\kappa = C_2\sqrt{\frac{\ln n}{n}}$ for some $C_1, C_2 > 0$ that can be derived from the proof. Then,*

$$\sup_{\eta^* \in \mathcal{C}_{\mathcal{B}}(d_0, \mathbf{s}, \mathbf{p}, \mathbf{q})} \mathcal{E}(\widehat{\eta}_{sgL}, \eta^*) \leq C \sqrt{\frac{1}{\nu_{sg}(d_0)} \left(\frac{d_0 \ln d}{n} + \ln n \sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}} \right)}$$

for some $C > 0$.

2. *Apply the sparse group Slope classifier (2.7) with $\lambda_j = C_1\sqrt{\frac{\ln(de/j)}{n}}$, $j = 1, \ldots, d$ and $\kappa_\ell = C_2\sqrt{\frac{\ln(ne/\ell)}{n}}$, $\ell = 1, \ldots, n$. Then,*

$$\sup_{\eta^* \in \mathcal{C}_\mathcal{B}(d_0, \mathbf{s}, \mathbf{p}, \mathbf{q})} \mathcal{E}(\widehat{\eta}_{sgS}, \eta^*) \leq C\sqrt{\frac{1}{\nu_{sg}(d_0)}\left(\frac{d_0\ln\left(\frac{de}{d_0}\right)}{n} + \ln n \sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}\right)}.$$

Theorem 3.7. *Consider the SpAM (1.1)-(1.3), where $d_0\ln(\frac{de}{d_0}) \leq n$ and $g \in \mathcal{G}_\mathcal{B}(d_0, \mathbf{s})$. Then,*

$$\inf_{\widetilde{\eta}} \sup_{\eta^* \in \mathcal{C}_\mathcal{B}(d_0, \mathbf{s}, \mathbf{p}, \mathbf{q}), \mathbb{P}_X} \mathcal{E}(\widetilde{\eta}, \eta^*) \geq \tilde{C}\sqrt{\frac{d_0\ln\left(\frac{de}{d_0}\right)}{n} + \sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}}$$

for some $\tilde{C} > 0$.

Thus, sparse group Lasso and Slope wavelet-based SpAM classifiers with the tuning parameters from Theorem 3.6 are simultaneously nearly-optimal (up to log-factors) across the entire range of Besov classes including smooth and non-smooth functions.

4. Examples

In this section we present experimental results for the developed SpAM classifiers applied to both simulated and real data. Solving numerically (2.3) and (2.6) implies convex programming and various methods can be used. We could not find an available software for fitting logistic sparse group Slope and therefore tested Lasso-type classifiers. In particular, we adapted the R package **sparsegl** from CRAN based on the group-wise majorization-minimization algorithm (see [15] for details)[1]. Vanilla Lasso and group Lasso were fitted as particular cases corresponding respectively to $\lambda = 0$ and $\kappa = 0$ in (2.3).

4.1. Simulations

We generated n random samples of feature vectors $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^T \in \mathbb{R}^d$ according to

$$X_{ij} = \frac{W_{ij} + U_i}{2}, \quad j = 1, \ldots, d; \ \ i = 1, \ldots, n,$$

where W_{ij} and U_i are i.i.d. $U[0, 1]$. Such a design results in correlations 0.5 between all features. We considered a SpAM with a logit function

$$g(\mathbf{X}) = g_1(X_1) + g_2(X_2) + g_3(X_3),$$

---

[1]Note that the tuning parameters λ and κ of sprase group Lasso are defined slightly different in **sparsegl.**

where

$$g_1(x_1) = |x_1 - 0.6|, \quad g_2(x_2) = (2x_2 - 1)^2, \quad g_3(x_3) = \frac{\sin(2\pi x_3)}{2 - \sin(2\pi x_3)}$$

while other additive components are identically zero. The resulting vectors $g_1(\mathbf{X}_1), g_2(\mathbf{X}_2), g_3(\mathbf{X}_3) \in \mathbb{R}^n$ were centred and scaled to have unit Euclidean norm. For a generated $\mathbf{X}_i$, the class label Y_i was sampled from a Bernoulli distribution $B(1, p(\mathbf{X}_i))$, $i = 1, \ldots, n$ with $p(\mathbf{X}_i) = \frac{e^{g(\mathbf{X}_i)}}{1 + e^{g(\mathbf{X}_i)}}$. We tried all combinations of $d = 10, 30, 100$ and $n = 30, 100, 300, 1000$ including, in particular, cases with $d \gtrsim n$.

We used cosine series basis, and applied Lasso, group Lasso and sparse group Lasso (2.3) to find $\widehat{g}_{sgL}$ in (2.4). The tuning parameters were chosen by 10-fold cross-validation. The performance of the resulting SpAM classifiers (2.5) was measured by misclassification excess risk to compare the misclassification errors of the considered classifiers w.r.t. those of the Bayes (oracle) classifier $\eta^*(\mathbf{X}) = I\{g(\mathbf{X}) \geq 0\}$ on the independently generated test sets of size 100. The procedure was replicated 10 times for each combination of d and n.

Table 1 presents the misclassification excess risks on the test sets, the numbers of selected features (nonzero rows of the regression coefficient matrix B) and the overall numbers of selected coefficients (nonzero entries in B) averaged over 10 replications.

Table 1 demonstrates that a more flexible sparse group Lasso consistently outperforms vanilla Lasso and group Lasso. Misclassification excess risks for all methods converge to zero as n increases for all d. Group Lasso converges more slowly and results in a larger number of selected coefficients, often leading to zero training errors. This behavior can be explained by the method keeping all coefficients of nonzero rows.

It should be mentioned that while Lasso-type methods perform well for prediction, they may not be as successful in feature selection (support recovery) for strongly correlated features which are common for high-dimensional data (see, e.g., [5], Section 7.2). Thus, although in the vast majority of cases, all the procedures correctly selected X_1, X_2 and X_3, they often also included spurious features and coefficients especially for large d. Sparse group Lasso typically selected less features.

4.2. Real-data example

To illustrate the performance of various Lasso-type classifiers on real data we considered the well-known benchmark email spam example from Example 1 of [12]. The data consists of 4601 real and spam email messages, where for each email the outcome (**mail** or **spam**) and $d = 57$ numeric attributes including relative frequencies of the most commonly occurring words and other email characters are available. The goal is to design a classifier for automatic spam detection based on this data.

TABLE 1
Average misclassification excess risks for various Lasso-type SpAM classifiers. The average numbers of selected features and of nonzero coefficients are given in parentheses.

n	group Lasso	Lasso	sparse group Lasso
		d=10	
30	0.223	0.266	0.209
	(3.7; 111)	(2.4; 4.3)	(3.6; 75.4)
100	0.213	0.200	0.187
	(7.6; 760)	(6.2; 16)	(6; 213.4)
300	0.221	0.086	0.085
	(8.9; 2670)	(9.7; 45.1)	(9.1; 108.9)
1000	0.154	0.079	0.049
	(9.8; 9800)	(9.8; 49.3)	(9.3; 94.6)
		d=30	
30	0.258	0.319	0.251
	(6.7; 201)	(1.9; 2.3)	(4.6; 96.3)
100	0.232	0.178	0.177
	(10.4; 1040)	(14.9; 26)	(10.6;3 3.9)
300	0.192	0.067	0.070
	(17.2; 5160)	(16.3; 33.1)	(14.5; 40.6)
1000	0.154	0.032	0.028
	(9.8; 9800)	(24.2; 77.6)	(15.6; 106.2)
		d=100	
30	0.244	0.278	0.280
	(6; 180)	(9; 10.4)	(4.9; 73.1)
100	0.257	0.207	0.192
	(19.7; 1370)	(6.5; 7.5)	(5.4; 7.7)
300	0.203	0.101	0.092
	(25.8; 7740)	(37.5; 53.2)	(38; 90.5)
1000	0.174	0.052	0.043
	(43.1; 43100)	(62.2; 142.0)	(37.8; 127.8)

Following [20] we used only $n = 300$ randomly selected samples for the training set to study the performance for relatively small n (w.r.t. d) case. We used cosine series basis, and applied Lasso, group Lasso and sparse group Lasso (2.3) to find $\widehat{g}_{sgL}$ in (2.4). The features were scaled first to have unit Euclidean norms and the test set was used as the hold-out set for choosing the tuning parameters. The performance of the resulting SpAM classifiers (2.5) was measured by misclassification errors on the remaining 4301 observations from the test set. For comparison, we also present the results reported in [20] for their SpAM classifier, which is based on a certain additive smoothing procedure.

Table 2 provides misclassification errors for the test test, the subsets of selected features and the overall numbers of selected coefficients for the considered classifiers. Lasso-type classifiers resulted in smaller misclassification errors than SpAM of [20]. One realizes that sparse group Lasso classifier yielded the smallest misclassification error and both group-based classifiers outperformed vanilla Lasso. The latter disregarded the global row-wise sparsity of the matrix B and identified 32 active features (nonzero rows), although it resulted in the smallest overall number of nonzero coefficients. Grouping coefficients from the same

TABLE 2
Misclassification errors and feature selection for various SpAM classifiers.

Method	Error	# Selected features	# Nonzero coeff.
Lasso	0.0879	32	1703
		$\{1-3, 5-13, 16-19, 21, 24, 25, 27,$	
		$33, 37, 42, 44-46, 50, 52, 53, 55-57\}$	
group Lasso	0.0844	22	6600
		$\{3, 5, 7, 12, 13, 16, 17, 19, 21, 24, 25,$	
		$27, 37, 44-46, 50, 52-55, 57\}$	
sparse group Lasso	0.0842	22	6389
		$\{3, 5, 7, 12, 13, 16, 17, 19, 21, 24, 25,$	
		$27, 37, 44-46, 50, 52-55, 57\}$	
SpAM ([20])	0.1083	24	
		$\{4, 6-10, 14-22, 26, 27, 38, 53-58\}$	

features allowed one to reduce the number of active features to 22, where both sparse group Lasso and group Lasso selected the same sets of features. While all entries of nonzero rows for group Lasso are nonzeroes (hence, $22 \times 300 = 6600$), sparse group Lasso, in addition, thresholded part of them as well.

Summarizing, the results of this section indicate that Lasso-type classifiers demonstrate good performance across various types of data sets, and confirm the established theoretical findings. A more flexible sparse group Lasso overall outperforms vanilla Lasso which overlooks a grouped structure of coefficients, and group Lasso, which fails to account for within-row sparsity and retains the entire nonzero rows.

Appendix

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

A.1. Proofs of the upper bounds

We start from a proof for a general upper bound and then apply it to correspondingly Sobolev, analytic and Besov classes of functions.

Consider a general logistic SpAM model (2.1)-(1.3), where $g_j \in \mathcal{F}_j[0,1]$, and let $\{\psi_\ell\}_{l=0}^{\infty}$ be an orthonormal basis in $\mathcal{F}_j$. Then,

$$g(\mathbf{x}) = \mu + \sum_{j=1}^{d} \sum_{\ell=1}^{\infty} \beta_{j\ell} \psi_\ell(x_j).$$

Recall that g_n is a truncated version of g:

$$g_n(\mathbf{x}) = \mu + \sum_{j=1}^{d} \sum_{\ell=1}^{n} \beta_{j\ell} \psi_\ell(x_j)$$

and consider the corresponding truncated Bayes classifier $\eta_n^* = I\{g_n(\mathbf{x}) \geq 0\}$.

Let $\widehat{g}$ and $\widehat{\eta} = I\{\widehat{g} \geq 0\}$ be respectively the logistic sparse group Lasso or Slope estimator of g and the corresponding classifier. The misclassification excess risk $\mathcal{E}(\widehat{\eta}, \eta^*)$ can be decomposed as

$$\mathcal{E}(\widehat{\eta}, \eta^*) = \mathbb{E}R(\widehat{\eta}) - R(\eta_n^*) + R(\eta_n^*) - R(\eta^*) = \mathcal{E}(\widehat{\eta}, \eta_n^*) + R(\eta_n^*) - R(\eta^*). \quad \text{(A.1)}$$

The logistic loss $\ell(y, g_n(\mathbf{x}))$ in (2.2) is Lipschitz. In addition, one can show that for the bounded support $\mathcal{X}$, $\mathbb{E}\left(\ell(Y, g_{1n}(\mathbf{X}) - \ell(Y, g_{2n}(\mathbf{X}))\right) \geq C||g_{1n}(\mathbf{x}) - g_{2n}(\mathbf{x})||^2_{L_2(\mathbb{P}_X)}$ for some $C > 0$ and any pair (g_{1n}, g_{2n}). Furthermore, assume that it is possible to approximate d_0 "approximately sparse" nonzero rows $B_{j\cdot}$, $j \in \mathcal{J}$ of the regression coefficients matrix B by truly sparse vectors $B'_{j\cdot}$ with the properly chosen numbers of nonzero entries m_j's such that $\kappa|B_{j\cdot} - B'_{j\cdot}|_1 = \kappa \sum_{\ell=1}^{n} |\beta_{j\ell} - \beta'_{j\ell}| \leq \ln n \; m_j n^{-1}$ for Lasso and $\sum_{\ell=1}^{n} \kappa_j|B - B'|_{j(l)} \leq \ln(ne/m_j)m_j n^{-1}$ for Slope for κ's given in the theorems. Similar to B, complete the matrix B' with zero rows for $j \notin \mathcal{J}$. One can then exploit the results of [14] (Section 3.2 and Remark 8) for sparse group Lasso and Slope to get under the $WRE(d_0, \mathbf{m}, c_0)$-condition 3.1[2]

$$\mathcal{E}(\widehat{\eta}_{sgL}, \eta_n^*) \leq C\sqrt{\frac{1}{\nu_{sg}(d_0)} \; \frac{d_0 \ln d + \ln n \sum_{j \in \mathcal{J}} m_j}{n}} \quad \text{(A.2)}$$

and

$$\mathcal{E}(\widehat{\eta}_{sgS}, \eta_n^*) \leq C\sqrt{\frac{1}{\nu_{sg}(d_0)} \; \frac{d_0 \ln\left(\frac{de}{d}\right) + \sum_{j \in \mathcal{J}} m_j \ln\left(\frac{ne}{m_j}\right)}{n}}. \quad \text{(A.3)}$$

To bound the second term in the RHS of (A.1) note that

$$R(\eta_n^*) - R(\eta^*) \leq 2||g_n - g||_{L_2(\mathbb{P}_X)} \leq C\sqrt{\sum_{j \in \mathcal{J}} \sum_{\ell \geq n} |\beta_{j\ell}|^2} \quad \text{(A.4)}$$

([27]). We will show that this term is negligible w.r.t. $\mathcal{E}(\widehat{\eta}, \eta_n^*)$.

We now adapt the presented general scheme for Sobolev, analytic and Besov classes by finding the corresponding matrices B'.

A.1.1. Proof of Theorem 3.2

For $g_j \in \widetilde{H}^{s_j}[0,1]$, $s_j > \frac{1}{2}$ construct the corresponding sparse approximations $B'_{j\cdot}$ for nonzero rows $B_{j\cdot}$ of B as follows. Take $m_j = n^{\frac{1}{2s_j+1}}$ and set $\beta'_{j\ell} = \beta_{j\ell}$ for the first m_j entries of $B_{j\cdot}$ and zeroes for others. Note that for $g_j \in \widetilde{H}^{s_j}[0,1], s_j >$

---

[2]Although the original assumption required in [14] is somewhat stronger, it can be relaxed to the WRE condition 3.1.

1/2 the cosine series coefficients satisfy $|\beta|_{j\ell} \leq C\ell^{-(s_j+\frac{1}{2})}$. For Lasso with $\kappa = C_2\sqrt{\frac{\ln n}{n}}$ we then have

$$\kappa|B_{j\cdot} - B'_{j\cdot}|_1 = \kappa \sum_{\ell=m_j+1}^{n} |\beta|_{j\ell} \leq C\kappa \; m_j^{-s_j+\frac{1}{2}} \leq C\sqrt{\ln n} \; n^{-\frac{2s_j}{2s_j+1}} \leq C\frac{\ln n}{n}m_j.$$

Similarly, for Slope with $\kappa_j = C_2\sqrt{\frac{\ln(ne/j)}{n}}$,

$$\sum_{\ell=1}^{n} \kappa_\ell |B - B'|_{j(\ell)} = \sum_{\ell=m_j+1}^{n} \kappa_l |\beta|_{j\ell} \leq C\kappa_{m_j} \; m_j^{-s_j+\frac{1}{2}} \leq C\sqrt{\ln\left(\frac{ne}{m_j}\right)}n^{-\frac{2s_j}{2s_j+1}}$$

$$\leq C\frac{\ln\left(\frac{ne}{m_j}\right)}{n}m_j.$$

Hence, applying the results of [14] (Section 3.2 and Remark 8) with λ's and κ's in Theorem 3.2 and constants C_1, C_2 given in their proofs, (A.2) implies

$$\mathcal{E}(\widehat{\eta}_{sgL}, \eta_n^*) \leq C\sqrt{\frac{1}{\nu_{sg}(d_0)} \; \frac{d_0 \ln d + \ln n \sum_{j\in\mathcal{J}} m_j}{n}}$$

$$\leq C\sqrt{\frac{1}{\nu_{sg}(d_0)} \; \left(\frac{d_0 \ln d}{n} + \ln n \sum_{j\in\mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}\right)}.$$

Similarly, by (A.3)

$$\mathcal{E}(\widehat{\eta}_{sgS}, \eta_n^*) \leq C\sqrt{\frac{1}{\nu_{sg}(d_0)} \; \frac{d_0 \ln\left(\frac{de}{d_0}\right) + \sum_{j\in\mathcal{J}} m_j \ln\left(\frac{ne}{m_j}\right)}{n}}$$

$$\leq C\sqrt{\frac{1}{\nu_{sg}(d_0)} \; \left(\frac{d_0 \ln\left(\frac{de}{d_0}\right)}{n} + \ln n \sum_{j\in\mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}\right)}.$$

Finally, for the tail sums in (A.4) we have

$$\sqrt{\sum_{j\in\mathcal{J}} \sum_{\ell>n} \beta_{j\ell}^2} \leq C\sqrt{\sum_{j\in\mathcal{J}} n^{-2s_j}} = o(\mathcal{E}(\widehat{\eta}, \eta_n^*)).$$

A.1.2. Proof of Theorem 3.4

For an analytic $g_j \in \mathcal{A}^{\alpha_j}[0,1]$, the cosine coefficients $|\beta|_{j\ell} \leq Ce^{-\alpha_j\ell}$. Define the nonzero rows of B' by truncating nonzero rows of the regression matrix B at

$m_j = \frac{1}{2\alpha_j} \ln \frac{n}{\ln n}$, $j \in \mathcal{J}$. We then have

$$\kappa |B_{j\cdot} - B'_{j\cdot}|_1 = \kappa \sum_{\ell=m_j+1}^{n} |\beta|_{j\ell} \le C\kappa e^{-\alpha_j m_j} \le C\frac{\ln n}{n} m_j$$

for Lasso and

$$\sum_{\ell=1}^{n} \kappa_\ell |B - B'|_{j(\ell)} = \sum_{\ell=m_j+1}^{n} \kappa_\ell |\beta|_{j\ell} \le C\kappa_{m_j} e^{-\alpha_j m_j} \le C\frac{\ln\left(\frac{ne}{m_j}\right)}{n} m_j$$

for Slope.

Going along the lines of the proof for Sobolev classes, (A.2) and (A.3) yield

$$\mathcal{E}(\widehat{\eta}_{sgL}, \eta_n^*) \le C\sqrt{\frac{1}{\nu_{sg}(d_0)}\frac{d_0\ln d + \ln n\sum_{j\in\mathcal{J}}m_j}{n}}$$

$$\le C\sqrt{\frac{1}{\nu_{sg}(d_0)}\left(\frac{d_0\ln d}{n} + \sum_{j\in\mathcal{J}}\frac{1}{\alpha_j}\frac{\ln n}{n}\right)}$$

and

$$\mathcal{E}(\widehat{\eta}_{sgS}, \eta_n^*) \le C\sqrt{\frac{1}{\nu_{sg}(d_0)}\frac{d_0\ln\left(\frac{de}{d_0}\right) + \sum_{j\in\mathcal{J}}m_j\ln\left(\frac{ne}{m_j}\right)}{n}}$$

$$\le C\sqrt{\frac{1}{\nu_{sg}(d_0)}\left(\frac{d_0\ln\left(\frac{de}{d_0}\right)}{n} + \sum_{j\in\mathcal{J}}\frac{1}{\alpha_j}\frac{\ln n}{n}\right)}.$$

Evidently, $\sqrt{\sum_{j\in\mathcal{J}}\sum_{\ell>n}\beta_{j\ell}^2} = o\left(\mathcal{E}(\widehat{\eta}_{sgS}, \eta_n^*)\right)$.

A.1.3. Proof of Theorem 3.6

Recall that for $g_j \in B_{p_j,q_j}^{s_j}[0,1]$ we consider orthonormal wavelet series $\{\psi_\ell\}_{l=0}^{\infty}$ of regularity $r > s_j$, where $\ell = 2^h + k$, $h \ge 0$, $0 \le k \le 2^h - 1$, and $|\beta|_{j(\ell)} \le C\ell^{-(s_j+\frac{1}{2})}$ (see (3.2)). For a given nonzero row of B design its sparse approximation $B'_{j\cdot}$ as follows: keep $m_j = n^{\frac{1}{2s_j+1}}$ coefficients of $B_{j\cdot}$ with the largest absolute values and zero others.

Hence,

$$\kappa|B_{j\cdot} - B'_{j\cdot}|_1 = \kappa \sum_{\ell=m_j+1}^{n} |\beta|_{j(\ell)} \le C\kappa\, m_j^{-s_j+\frac{1}{2}} \le C\sqrt{\ln n}\, n^{-\frac{2s_j}{2s_j+1}} \le C\frac{\ln n}{n} m_j$$

for Lasso and

$$\sum_{\ell=1}^{n} \kappa_\ell |B - B'|_{j(\ell)} = \sum_{\ell=m_j+1}^{n} \kappa_l |\beta|_{j(\ell)} \leq C\kappa_{m_j} \; m_j^{-s_j+\frac{1}{2}} \leq C\sqrt{\ln\left(\frac{ne}{m_j}\right)} \; n^{-\frac{2s_j}{2s_j+1}}$$

$$\leq C\frac{\ln\left(\frac{ne}{m_j}\right)}{n} m_j$$

for Slope. To complete the proof one repeats the arguments from the proof for Sobolev classes.

A.2. Proofs of the lower bounds (Theorems 3.3, 3.5 and 3.7)

To prove the lower bounds for the minimax misclassification excess risk we utilize the general approach developed in [26].

Assume that the unknown logit function $g(\mathbf{x})$ belongs to some class of functions $\mathcal{G}$ with a metric entropy $\mathcal{M}(\mathcal{G}, \epsilon)$ – the logarithm of the ϵ-packing number of $\mathcal{G}$ w.r.t. $L_2(\mathbb{P}_X)$-norm. Consider the corresponding class $\mathcal{P}$ for $p(\mathbf{x})$: $\mathcal{P} = \{p : p = \frac{e^g}{1+e^g}, \; g \in \mathcal{G}\}$. Let ϵ^* satisfy

$$n\epsilon^{*2} \gtrsim \mathcal{M}(\mathcal{P}, \epsilon^*). \tag{A.5}$$

Then if the class $\mathcal{P}$ is rich enough in terms of $\mathcal{M}(\mathcal{P}, \epsilon)$ to satisfy the assumptions given in [26], the minimax misclassification excess risk

$$\inf_{\widetilde{\eta}} \sup_{p \in \mathcal{P}, \mathbb{P}_X} \mathcal{E}(\widetilde{\eta}, \eta^*) \asymp \epsilon^*, \tag{A.6}$$

where recall that the Bayes classifier $\eta^*(\mathbf{x}) = I\{p(\mathbf{x}) \geq 1/2\}$.

By Taylor expansion, one immediately verifies that for any probability functions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ and their logit functions $g_1(\mathbf{x}), g_2(\mathbf{x})$,

$$||p_2 - p_1||_{L_2(\mathbb{P}_X)} \leq \frac{1}{4}||g_2 - g_1||_{L_2(\mathbb{P}_X)}.$$

Hence, $\mathcal{M}(\mathcal{P}, \epsilon) \leq \mathcal{M}(\mathcal{G}, 4\epsilon)$.

For the considered general SpAM setup with $d_0 < d/4$ by Lemma 4 of Raskutti *et al.* (2012),

$$\mathcal{M}(\mathcal{G}, 4\epsilon) \leq C\left(d_0 \ln \frac{de}{d_0} + \sum_{j \in \mathcal{J}} \mathcal{M}\left(\mathcal{G}_j, \frac{4\epsilon}{\sqrt{d_0}}\right)\right), \tag{A.7}$$

where for the considered Sobolev, analytic and Besov classes $\mathcal{G}_j$'s we have $\mathcal{M}(\widetilde{H}^{s_j}[0,1], \epsilon) \simeq \mathcal{M}(B_{p_j, q_j}^{s_j}[0,1], \epsilon) \simeq \left(\frac{1}{\epsilon}\right)^{\frac{1}{s_j}}$ and $\mathcal{M}(\mathcal{A}^{\alpha_j}[0,1], \epsilon) \simeq \frac{1}{\alpha_j} \ln\left(\frac{1}{\epsilon}\right)$ (e.g., [26, 19]). Moreover, [26] showed that the required conditions on metric entropies to apply (A.6) are satisfied for these classes.

Solving (A.5) and (A.7) after some algebra implies that the suitable solutions are $\epsilon^{*2} = C\left(\frac{d_0 \ln \frac{de}{d_0}}{n} + \sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}\right)$ and $\epsilon^{*2} = C\left(\frac{d_0 \ln \frac{de}{d_0}}{n} + \sum_{j \in \mathcal{J}} \frac{1}{\alpha_j}\frac{\ln n}{n}\right)$ respectively.

For the dense case $d/4 \leq d_0 \leq d$ (hence, $d_0 \sim d$), the model selection error is negligible w.r.t. nonparametric estimation term. Evidently, no classifier can perform better than an oracle who knows the subset of truly active features $\mathcal{J}$. Therefore, for Sobolev and Besov classes in this case $d_0 \ln \frac{de}{d_0} \sim d_0 \lesssim \sum_{j \in \mathcal{J}} n^{\frac{1}{2s_j+1}}$ and $\epsilon^{*2} = C\sum_{j \in \mathcal{J}} n^{-\frac{2s_j}{2s_j+1}}$. Similarly, for analytic functions, $d_0 \ln \frac{de}{d_0} \lesssim \ln n \sum_{j \in \mathcal{J}} \frac{1}{\alpha_j}$ and $\epsilon^{*2} = C\frac{\ln n}{n}\sum_{j \in \mathcal{J}} \frac{1}{\alpha_j}$.

Applying (A.6) completes the proof.

Acknowledgments

Funding

References

[1] Abramovich, F. and Grinshtein, V. (2019). High-dimensional classification by sparse logistic regression *IEEE Trans. Inf. Theory* **65**, 3068–3079. MR3951383

[2] Abramovich, F. and Lahav, T. (2015). Sparse additive estimation on a regular lattice. *J.R. Statist. Soc.* B **77**, 443–459. MR3310534

[3] Audibert, J-Y. and Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *A*nn. Statist., **35**, 608–633. MR2336861

[4] Bellec, P.C., Lecué, G. and Tsybakov, A. (2018). Slope meets Lasso: improved oracle bounds and optimaility. *A*nn. Statist. **46**, 3603–3642. MR3852663

[5] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Methods, Theory and Applications.* Springer, Berlin Heidelberg. MR2807761

[6] Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harm. Anal.*, **1**, 54–81. MR1256527

[7] Donoho, D.L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Comp. Harm. Anal.* **1**, 100–115. MR1256530

[8] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Stat. Assoc.* **90**, 1200–1224. MR1379464

[9] Golubev, Y.K., Levit, B. and Tsybakov, A. (1996). Asymptotically efficient estimation of analytic functions in Gaussian white noise. *Bernoulli* **2**, 167–181 MR1410136

[10] Haris, A., Simon, N. and Shojaie, A. (2022). Generalized sparse additive models. *J. Mach. Learn. Research* **23**, 1–56. MR4576655

[11] Hastie, T. and Tibshirani, R. (1999). *Generalized Additive Models*. New York: Chapman and Hall. MR1082147

[12] Hastie, T., Tisbhirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning* (Second edition), Springer. MR2722294

[13] Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38**, 3660–3695. MR2766864

[14] Levy, T. and Abramovich, F. (2023). Generalization error bounds for multiclass sparse linear classifiers. *J. Mach. Learn. Research* **24**, 1–35. MR4596098

[15] Liang, X., Cohen, A., Heinsfeld, A.S., Pestilly, F. and McDonald, D.J. (2022). **sparsegl**: an R package for estimating sparse group Lasso. *arXiv:2208.02942*.

[16] Lin, Y. and Zhang, H.H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272–2297. MR2291500

[17] Meier, L., van de Geer, S. and Bühlmann, P. (2009). High-dimensional additive modelling. *Ann. Statist.* **37**, 3779–3821. MR2572443

[18] Meyer, Y. (1992). *Wavelets and Operators*. Cambridge University Press. MR1228209

[19] Raskutti, G., Wainwright, M.J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Research* **13**, 389–427. MR2913704

[20] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2007). SpAM: sparse additive models. In *Advances in Neural Information Processing Systems* **20**. MR2750255

[21] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *J.R. Statist. Soc.* B **71**, 1009-1030. MR2750255

[22] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22**, 231–245. MR3173712

[23] Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32**, 135–166. MR2051002

[24] Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[25] Vincent, M. and Hansen, N.R. (2014). Sparse group lasso and high dimensional multinomial classification. *Comput. Statist. Data Anal.* **71**, 771–786. MR3132005

[26] Yang, Y. (1999). Minimax nonparametric classification – Part I: Rates of convergence. *IEEE Trans. Inf. Theory* **45**, 2271–2292. MR1725115

[27] Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.* **32**, 56–85. MR2051001