Journal of Machine Learning Research 20 (2019) 1-34 Submitted 3/17; Revised 1/19; Published 8/19

Binarsity: a penalization for one-hot encoded features in linear supervised learning

Mokhtar Z. Alaya MOKHTARZAHDI.ALAYA@GMAIL.COM
Laboratoire de Probabilités Statistique et Modélisation, CNRS UMR 8001
Sorbonne University
Paris, France

Simon Bussy SIMON.BUSSY@GMAIL.COM
Laboratoire de Probabilités Statistique et Modélisation, CNRS UMR 8001
Sorbonne University
Paris, France

Stéphane Gaïffas STEPHANE.GAIFFAS@LPSM.PARIS
Laboratoire de Probabilités Statistique et Modélisation, CNRS UMR 8001
Université Paris Diderot
Paris, France

Agathe Guilloux AGATHE.GUILLOUX@MATH.CNRS.FR
LaMME, UEVE and UMR 8071
Université Paris Saclay
Evry, France

Editor: John Shawe-Taylor

Abstract

This paper deals with the problem of large-scale linear supervised learning in settings where a large number of continuous features are available. We propose to combine the well-known trick of one-hot encoding of continuous features with a new penalization called *binarsity*. In each group of binary features coming from the one-hot encoding of a single raw continuous feature, this penalization uses total-variation regularization together with an extra linear constraint. This induces two interesting properties on the model weights of the one-hot encoded features: they are piecewise constant, and are eventually block sparse. Non-asymptotic oracle inequalities for generalized linear models are proposed. Moreover, under a sparse additive model assumption, we prove that our procedure matches the state-of-the-art in this setting. Numerical experiments illustrate the good performances of our approach on several datasets. It is also noteworthy that our method has a numerical complexity comparable to standard ℓ_1 penalization.

Keywords: Supervised learning, Features binarization, Sparse additive modeling, Total-variation, Oracle inequalities, Proximal methods

1. Introduction

In many applications, datasets used for linear supervised learning contain a large number of continuous features, with a large number of samples. An example is web-marketing, where features are obtained from bag-of-words scaled using tf-idf (Russell, 2013), recorded during the visit of users on websites. A well-known trick (Wu and Coggeshall, 2012; Liu et al., 2002) in this setting is to

replace each raw continuous feature by a set of binary features that one-hot encodes the interval containing it, among a list of intervals partitioning the raw feature range. This improves the linear decision function with respect to the raw continuous features space, and can therefore improve prediction. However, this trick is prone to over-fitting, since it increases significantly the number of features.

A new penalization. To overcome this problem, we introduce a new penalization called *binarsity*, that penalizes the model weights learned from such grouped one-hot encodings (one group for each raw continuous feature). Since the binary features within these groups are naturally ordered, the binarsity penalization combines a group total-variation penalization, with an extra linear constraint in each group to avoid collinearity between the one-hot encodings. This penalization forces the weights of the model to be as constant (with respect to the order induced by the original feature) as possible within a group, by selecting a minimal number of relevant cut-points. Moreover, if the model weights are all equal within a group, then the full block of weights is zero, because of the extra linear constraint. This allows to perform raw feature selection.

High-dimensional linear supervised learning. To address the high-dimensionality of features, sparse linear inference is now an ubiquitous technique for dimension reduction and variable selection, see for instance Bühlmann and van De Geer (2011) and Hastie et al. (2001) among many others. The principle is to induce sparsity (large number of zeros) in the model weights, assuming that only a few features are actually helpful for the label prediction. The most popular way to induce sparsity in model weights is to add a ℓ_1-penalization (Lasso) term to the goodness-of-fit (Tibshirani, 1996a). This typically leads to sparse parametrization of models, with a level of sparsity that depends on the strength of the penalization. Statistical properties of ℓ_1-penalization have been extensively investigated, see for instance Knight and Fu (2000); Zhao and Yu (2006); Bunea et al. (2007); Bickel et al. (2009) for linear and generalized linear models and Donoho and Huo (2001); Donoho and Elad (2002); Candès et al. (2008); Candès and Wakin (2008) for compressed sensing, among others.

However, the Lasso ignores ordering of features. In Tibshirani et al. (2005), a structured sparse penalization is proposed, known as fused Lasso, which provides superior performance in recovering the true model in such applications where features are ordered in some meaningful way. It introduces a mixed penalization using a linear combination of the ℓ_1-norm and the total-variation penalization, thus enforcing sparsity in both the weights and their successive differences. Fused Lasso has achieved great success in some applications such as comparative genomic hybridization (Rapaport et al., 2008), image denoising (Friedman et al., 2007), and prostate cancer analysis (Tibshirani et al., 2005).

Features discretization and cuts. For supervised learning, it is often useful to encode the input features in a new space to let the model focus on the relevant areas (Wu and Coggeshall, 2012). One of the basic encoding technique is *feature discretization* or *feature quantization* (Liu et al., 2002) that partitions the range of a continuous feature into intervals and relates these intervals with meaningful labels. Recent overviews of discretization techniques can be found in Liu et al. (2002) or Garcia et al. (2013).

Obtaining the optimal discretization is a NP-hard problem (Chlebus and Nguyen, 1998), and an approximation can be easily obtained using a greedy approach, as proposed in decision trees: CART (Breiman et al., 1984) and C4.5 (Quinlan, 1993), among others, that sequentially select pairs

of features and cuts that minimize some purity measure (intra-variance, Gini index, information gain are the main examples). These approaches build decision functions that are therefore very simple, by looking only at a single feature at a time, and a single cut at a time. Ensemble methods (boosting (Lugosi and Vayatis, 2004), random forests (Breiman, 2001)) improve this by combining such decisions trees, at the expense of models that are harder to interpret.

Main contribution. This paper considers the setting of linear supervised learning. The main contribution of this paper is the idea to use a total-variation penalization, with an extra linear constraint, on the weights of a generalized linear model trained on a binarization of the raw continuous features, leading to a procedure that selects multiple cut-points per feature, looking at all features simultaneously. Our approach therefore increases the capacity of the considered generalized linear model: several weights are used for the binarized features instead of a single one for the raw feature. This leads to a more flexible decision function compared to the linear one: when looking at the decision function as a function of a single raw feature, it is now piecewise constant instead of linear, as illustrated in Figure 2 below.

Organization of the paper. The proposed methodology is described in Section 2. Section 3 establishes an oracle inequality for generalized linear models and provides a convergence rate for our procedure in the particular case of a sparse additive model. Section 4 highlights the results of the method on various datasets and compares its performances to well known classification algorithms. Finally, we discuss the obtained results in Section 5.

Notations. Throughout the paper, for every $q > 0$, we denote by $\|v\|_q$ the usual ℓ_q-quasi norm of a vector $v \in \mathbb{R}^m$, namely $\|v\|_q = (\sum_{k=1}^m |v_k|^q)^{1/q}$, and $\|v\|_\infty = \max_{k=1,\dots,m} |v_k|$. We also denote $\|v\|_0 = |\{k : v_k \neq 0\}|$, where $|A|$ stands for the cardinality of a finite set A. For $u, v \in \mathbb{R}^m$, we denote by $u \odot v$ the Hadamard product $u \odot v = (u_1 v_1, \dots, u_m v_m)^\top$. For any $u \in \mathbb{R}^m$ and any $L \subset \{1, \dots, m\}$, we denote u_L as the vector in $\mathbb{R}^m$ satisfying $(u_L)_k = u_k$ for $k \in L$ and $(u_L)_k = 0$ for $k \in L^\complement = \{1, \dots, m\} \backslash L$. We write, for short, $\mathbf{1}$ (resp. $\mathbf{0}$) for the vector of $\mathbb{R}^m$ having all coordinates equal to one (resp. zero). Finally, we denote by $\mathrm{sign}(x)$ the set of subdifferentials of the function $x \mapsto |x|$, namely $\mathrm{sign}(x) = \{1\}$ if $x > 0$, $\mathrm{sign}(x) = \{-1\}$ if $x < 0$ and $\mathrm{sign}(0) = [-1, 1]$.

2. The proposed method

Consider a supervised training dataset $(x_i, y_i)_{i=1,\dots,n}$ containing features $x_i = [x_{i,1} \cdots x_{i,p}]^\top \in \mathbb{R}^p$ and labels $y_i \in \mathcal{Y} \subset \mathbb{R}$, that are independent and identically distributed samples of (X, Y) with unknown distribution $\mathbb{P}$. Let us denote $\boldsymbol{X} = [x_{i,j}]_{1 \leq i \leq n; 1 \leq j \leq p}$ the $n \times p$ features matrix vertically stacking the n samples of p raw features. Let $\boldsymbol{X}_{\bullet,j}$ be the j-th feature column of $\boldsymbol{X}$.

Binarization. The binarized matrix $\boldsymbol{X}^B$ is a matrix with an extended number $d > p$ of columns, where the j-th column $\boldsymbol{X}_{\bullet,j}$ is replaced by $d_j \geq 2$ columns $\boldsymbol{X}^B_{\bullet,j,1}, \dots, \boldsymbol{X}^B_{\bullet,j,d_j}$ containing only zeros and ones. Its i-th row is written

$$x_i^B = [x^B_{i,1,1} \cdots x^B_{i,1,d_1} x^B_{i,2,1} \cdots x^B_{i,2,d_2} \cdots x^B_{i,p,1} \cdots x^B_{i,p,d_p}]^\top \in \mathbb{R}^d,$$

where $d = \sum_{j=1}^p d_j$. In order to simplify the presentation of our results, we assume in the paper that all raw features $\boldsymbol{X}_{\bullet,j}$ are continuous, so that they are transformed using the following one-hot

encoding. For each raw feature j, we consider a partition of intervals $I_{j,1}, \dots, I_{j,d_j}$ of range$(\boldsymbol{X}_{\bullet,j})$, namely satisfying $\cup_{k=1}^{d_j} I_{j,k} = \text{range}(\boldsymbol{X}_{\bullet,j})$ and $I_{j,k} \cap I_{j,k'} = \varnothing$ for $k \neq k'$ and define

$$x^B_{i,j,k} = \begin{cases} 1 & \text{if } x_{i,j} \in I_{j,k}, \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \dots, n$, $j = 1, \dots, p$ and $k = 1, \dots, d_j$. An example is interquantiles intervals, namely $I_{j,1} = \big[q_j(0), q_j(\frac{1}{d_j})\big]$ and $I_{j,k} = \big(q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\big]$ for $k = 2, \dots, d_j$, where $q_j(\alpha)$ denotes a quantile of order $\alpha \in [0,1]$ for $\boldsymbol{X}_{\bullet,j}$. In practice, if there are ties in the estimated quantiles for a given feature, we simply choose the set of ordered unique values to construct the intervals. This principle of binarization is a well-known trick (Garcia et al., 2013), that allows to improve over the linear decision function with respect to the raw feature space: it uses a larger number of model weights, for each interval of values for the feature considered in the binarization. If training data contains also unordered qualitative features, one-hot encoding with ℓ_1-penalization can be used for instance.

Goodness-of-fit. Given a loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$, we consider the goodness-of-fit term

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, m_\theta(x_i)), \tag{1}$$

where $m_\theta(x_i) = \theta^\top x_i^B$ and $\theta \in \mathbb{R}^d$ where we recall that $d = \sum_{j=1}^p d_j$. We then have $\theta = [\theta_{1,\bullet}^\top \cdots \theta_{p,\bullet}^\top]^\top$, with $\theta_{j,\bullet}$ corresponding to the group of coefficients weighting the binarized raw j-th feature. We focus on generalized linear models (Green and Silverman, 1994), where the conditional distribution $Y|X = x$ is assumed to be from a one-parameter exponential family distribution with a density of the form

$$y|x \mapsto f^0(y|x) = \exp\Big(\frac{y m^0(x) - b(m^0(x))}{\phi} + c(y, \phi)\Big), \tag{2}$$

with respect to a reference measure which is either the Lebesgue measure (e.g. in the Gaussian case) or the counting measure (e.g. in the logistic or Poisson cases), leading to a loss function of the form

$$\ell\big(y_1, y_2\big) = -y_1 y_2 + b(y_2).$$

The density described in (2) encompasses several distributions, see Table 1. The functions $b(\cdot)$ and $c(\cdot)$ are known, while the natural parameter function $m^0(\cdot)$ is unknown. The dispersion parameter ϕ is assumed to be known in what follows. It is also assumed that $b(\cdot)$ is three times continuously differentiable. It is standard to notice that

$$\mathbb{E}[Y|X = x] = \int y f^0(y|x) dy = b'(m^0(x)),$$

where b' stands for the derivative of b. This formula explains how b' links the conditional expectation to the unknown m^0. The results given in Section 3 rely on the following Assumption.

Assumption 1 *Assume that b is three times continuously differentiable, that there is $C_b > 0$ such that $|b'''(z)| \leq C_b |b''(z)|$ for any $z \in \mathbb{R}$ and that there exist constants $C_n > 0$ and $0 < L_n \leq U_n$ such that $C_n = \max_{i=1,\dots,n} |m^0(x_i)| < \infty$ and $L_n \leq \max_{i=1,\dots,n} b''\big(m^0(x_i)\big) \leq U_n$.*

This assumption is satisfied for most standard generalized linear models. In Table 1, we list some standard examples that fit in this framework, see also van de Geer (2008) and Rigollet (2012).

Model	ϕ	$b(z)$	$b'(z)$	$b''(z)$	$b'''(z)$	C_b	L_n	U_n
Normal	σ^2	$\frac{z^2}{2}$	z	1	0	0	1	1
Logistic	1	$\log(1+e^z)$	$\frac{e^z}{1+e^z}$	$\frac{e^z}{(1+e^z)^2}$	$\frac{1-e^z}{1+e^z}b''(z)$	2	$\frac{e^{C_n}}{(1+e^{C_n})^2}$	$\frac{1}{4}$
Poisson	1	e^z	e^z	e^z	$b''(z)$	1	e^{-C_n}	e^{C_n}

Tab. 1: Examples of standard distributions that fit in the considered setting of generalized linear models, with the corresponding constants in Assumption 1.

Binarsity. Several problems occur when using the binarization trick described above:

(P1) The one-hot-encodings satisfy $\sum_{k=1}^{d_j} \boldsymbol{X}^B_{i,j,k} = 1$ for $j = 1, \dots, p$, meaning that the columns of each block sum to $\mathbf{1}$, making $\boldsymbol{X}^B$ not of full rank by construction.

(P2) Choosing the number of intervals d_j for binarization of each raw feature j is not an easy task, as too many might lead to overfitting: the number of model-weights increases with each d_j, leading to a over-parametrized model.

(P3) Some of the raw features $\boldsymbol{X}_{\bullet,j}$ might not be relevant for the prediction task, so we want to select raw features from their one-hot encodings, namely induce block-sparsity in θ.

A usual way to deal with (P1) is to impose a linear constraint (Agresti, 2015) in each block. In order to do so, let us introduce first $n_{j,k} = |\{i : x_{i,j} \in I_{j,k}\}|$ and the vector $n_j = [n_{j,1} \cdots n_{j,d_j}] \in \mathbb{N}^{d_j}$. In our penalization term, we impose the linear constraint

$$n_j^\top \theta_{j,\bullet} = \sum_{k=1}^{d_j} n_{j,k} \theta_{j,k} = 0 \tag{3}$$

for all $j = 1, \dots, p$. Note that if the $I_{j,k}$ are taken as interquantiles intervals, then for each j, we have that $n_{j,k}$ for $k = 1, \dots, d_j$ are equal and the constraint (3) becomes the standard constraint $\sum_{k=1}^{d_j} \theta_{j,k} = 0$.

The trick to tackle (P2) is to remark that within each block, binary features are ordered. We use a within block total-variation penalization

$$\sum_{j=1}^{p} \|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}$$

where

$$\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}|, \tag{4}$$

with weights $\hat{w}_{j,k} > 0$ to be defined later, to keep the number of different values taken by $\theta_{j,\bullet}$ to a minimal level.

Finally, dealing with (P3) is actually a by-product of dealing with (P1) and (P2). Indeed, if the raw feature j is not-relevant, then $\theta_{j,\bullet}$ should have all entries constant because of the penalization (4), and in this case all entries are zero, because of (3). We therefore introduce the following penalization, called *binarsity*

$$\mathrm{bina}(\theta) = \sum_{j=1}^{p} \Big(\sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}| + \delta_j(\theta_{j,\bullet}) \Big) \tag{5}$$

where the weights $\hat{w}_{j,k} > 0$ are defined in Section 3 below, and where

$$\delta_j(u) = \begin{cases} 0 & \text{if} \quad n_j^\top u = 0, \\ \infty & \text{otherwise.} \end{cases} \tag{6}$$

We consider the goodness-of-fit (1) penalized by (5), namely

$$\hat{\theta} \in \mathrm{argmin}_{\theta \in \mathbb{R}^d} \big\{ R_n(\theta) + \mathrm{bina}(\theta) \big\}. \tag{7}$$

An important fact is that this optimization problem is numerically cheap, as explained in the next paragraph. Figure 1 illustrates the effect of the binarsity penalization with a varying strength on an example.

In Figure 2, we illustrate on a toy example, when $p = 2$, the decision boundaries obtained for logistic regression (LR) on raw features, LR on binarized features and LR on binarized features with the binarsity penalization.

Proximal operator of binarsity. The proximal operator and proximal algorithms are important tools for non-smooth convex optimization, with important applications in the field of supervised learning with structured sparsity (Bach et al., 2012). The proximal operator of a proper lower semi-continuous (Bauschke and Combettes, 2011) convex function $g : \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\mathrm{prox}_g(v) \in \mathrm{argmin}_{u \in \mathbb{R}^d} \Big\{ \frac{1}{2} \|v - u\|_2^2 + g(u) \Big\}.$$

Proximal operators can be interpreted as generalized projections. Namely, if g is the indicator of a convex set $C \subset \mathbb{R}^d$ given by

$$g(u) = \delta_C(u) = \begin{cases} 0 & \text{if } u \in C, \\ \infty & \text{otherwise,} \end{cases}$$

then prox_g is the projection operator onto C. It turns out that the proximal operator of binarsity can be computed very efficiently, using an algorithm (Condat, 2013) that we modify in order to include weights $\hat{w}_{j,k}$. It applies in each group the proximal operator of the total-variation since binarsity penalization is block separable, followed by a simple projection onto $\mathrm{span}(n_j)^\perp$ the orthogonal of $\mathrm{span}(n_j)$, see Algorithm 1 below. We refer to Algorithm 2 in Section 6.2 for the weighted total-variation proximal operator.

Proposition 1 *Algorithm 1 computes the proximal operator of* $\mathrm{bina}(\theta)$ *given by* (5).

A proof of Proposition 1 is given in Section 6.1. Algorithm 1 leads to a very fast numerical routine, see Section 4. The next section provides a theoretical analysis of our algorithm with an oracle inequality for the prediction error, together with a convergence rate in the particular case of a sparse additive model.

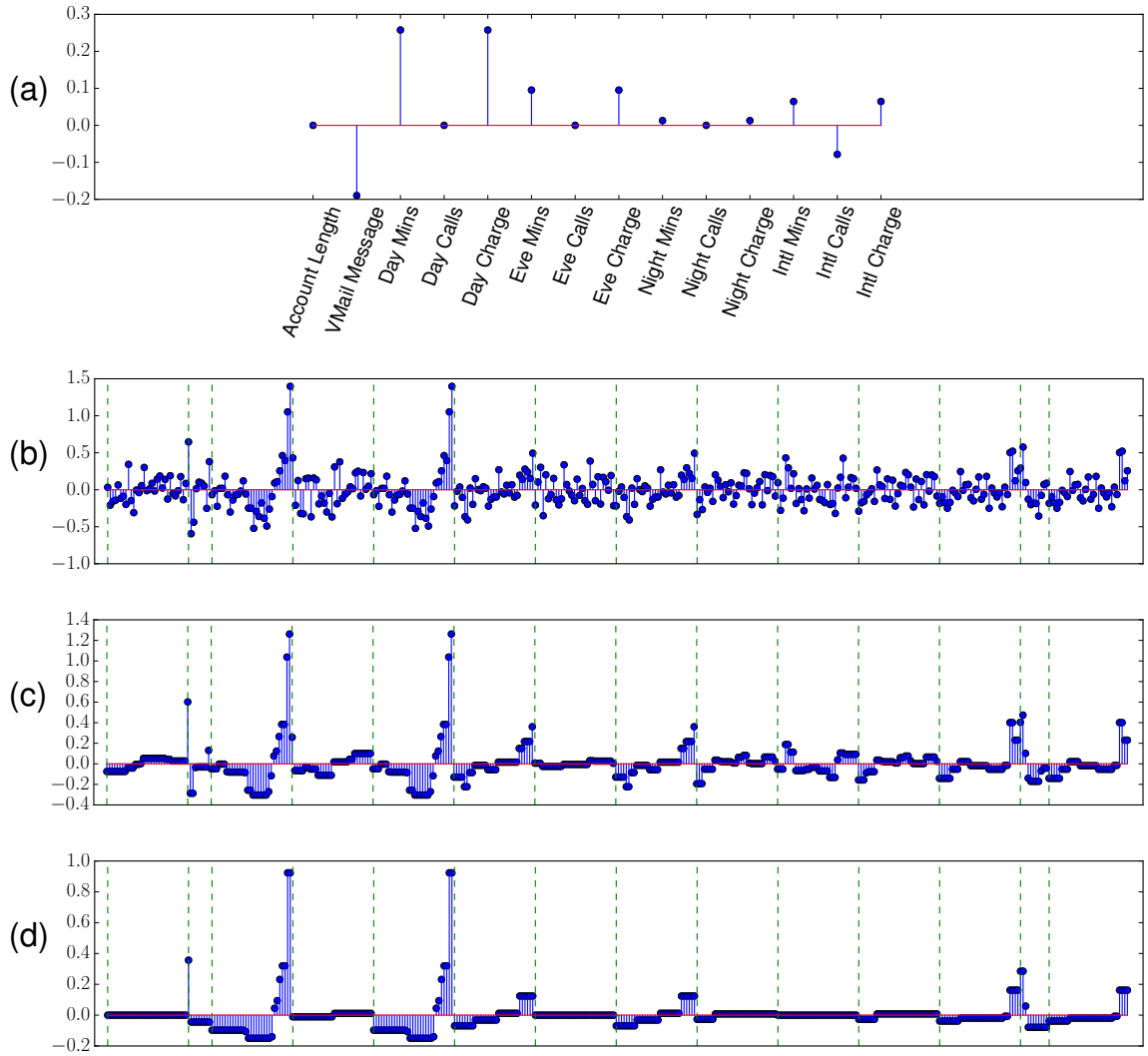

Fig. 1: Illustration of the binarsity penalization on the "Churn" dataset (see Section 4 for details) using logistic regression. Figure (a) shows the model weights learned by the Lasso method on the continuous raw features. Figure (b) shows the unpenalized weights on the binarized features, where the dotted green lines mark the limits between blocks corresponding to each raw features. Figures (c) and (d) show the weights with medium and strong binarsity penalization respectively. We observe in (c) that some significant cut-points start to be detected, while in (d) some raw features are completely removed from the model, the same features as those removed in (a).

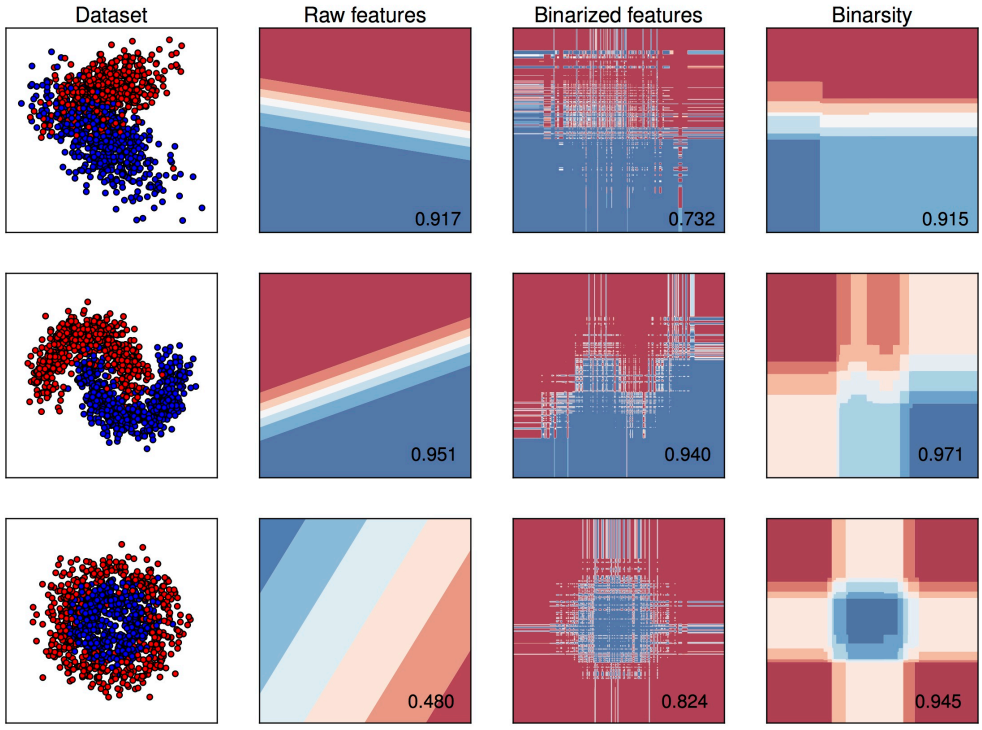

Fig. 2: Illustration of binarsity on 3 simulated toy datasets for binary classification with two classes (blue and red points). We set $n = 1000$, $p = 2$ and $d_1 = d_2 = 100$. In each row, we display the simulated dataset, followed by the decision boundaries for a logistic regression classifier trained on initial raw features, then on binarized features without regularization, and finally on binarized features with binarsity. The corresponding testing AUC score is given on the lower right corner of each figure. Our approach allows to keep an almost linear decision boundary in the first row, while a good decision boundaries are learned on the two other examples, which correspond to non-linearly separable datasets, without apparent overfitting.

Algorithm 1: Proximal operator of $\mathrm{bina}(\theta)$, see (5)

Input: vector $\theta \in \mathbb{R}^d$ and weights $\hat{w}_{j,k}$ and $n_{j,k}$ for $j = 1, \dots, p$ and $k = 1, \dots, d_j$
Output: vector $\eta = \mathrm{prox}_{\mathrm{bina}}(\theta)$
for $j = 1$ ***to*** p **do**
  $\beta_{j,\bullet} \leftarrow \mathrm{prox}_{\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})$ (TV-weighted prox in block j, see (4))
  $\eta_{j,\bullet} \leftarrow \beta_{j,\bullet} - \frac{n_j^\top \beta_{j,\bullet}}{\|n_j\|_2^2} n_j$ (projection onto $\mathrm{span}(n_j)^\perp$)
Return: η

3. Theoretical guarantees

We now investigate the statistical properties of (8) where the weights in the binarsity penalization have the form

$$\hat{w}_{j,k} = O\Big(\sqrt{\frac{\log d}{n}\,\hat{\pi}_{j,k}}\Big), \qquad \text{with} \qquad \hat{\pi}_{j,k} = \frac{\big|\{i = 1, \dots, n : x_{i,j} \in \cup_{k'=k}^{d_j} I_{j,k'}\}\big|}{n}$$

for all $k \in \{2, \dots, d_j\}$, see Theorem 2 for a precise definition of $\hat{w}_{j,k}$. Note that $\hat{\pi}_{j,k}$ corresponds to the proportion of ones in the sub-matrix obtained by deleting the first k columns in the j-th binarized block matrix $\boldsymbol{X}^B_{\bullet,j}$. In particular, we have $\hat{\pi}_{j,k} > 0$ for all j, k. We consider the risk measure defined by

$$R(m_\theta) = \frac{1}{n} \sum_{i=1}^{n} \big\{ - b'(m^0(x_i)) m_\theta(x_i) + b(m_\theta(x_i)) \big\},$$

which is standard with generalized linear models (van de Geer, 2008).

3.1. A general oracle inequality

We aim at evaluating how "close" to the minimal possible expected risk our estimated function $m_{\hat{\theta}}$ with $\hat{\theta}$ given by (8) is. To measure this closeness, we establish a non-asymptotic oracle inequality with a fast rate of convergence considering the excess risk of $m_{\hat{\theta}}$, namely $R(m_{\hat{\theta}}) - R(m^0)$. To derive this inequality, we consider for technical reasons the following problem instead of (7):

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in B_d(\rho)} \big\{ R_n(\theta) + \operatorname{bina}(\theta) \big\}, \tag{8}$$

where

$$B_d(\rho) = \Big\{ \theta \in \mathbb{R}^d : \sum_{j=1}^{p} \|\theta_{j,\bullet}\|_\infty \leq \rho \Big\}.$$

This constraint is standard in literature for the proof of oracle inequalities for sparse generalized linear models, see for instance van de Geer (2008), and is discussed in details below.

We also impose a restricted eigenvalue assumption on $\boldsymbol{X}^B$. For all $\theta \in \mathbb{R}^d$, let $J(\theta) = [J_1(\theta), \dots, J_p(\theta)]$ be the concatenation of the support sets relative to the total-variation penalization, that is

$$J_j(\theta) = \{k = 2, \dots, d_j \ : \ \theta_{j,k} \neq \theta_{j,k-1}\}.$$

Similarly, we denote $J^{\complement}(\theta) = \big[J_1^{\complement}(\theta), \dots, J_p^{\complement}(\theta)\big]$ the complementary of $J(\theta)$. The restricted eigenvalue condition is defined as follow.

Assumption 2 *Let $K = [K_1, \dots, K_p]$ be a concatenation of index sets such that*

$$\sum_{j=1}^{p} |K_j| \leq J^\star, \tag{9}$$

where $J^\star$ is a positive integer. Define

$$\kappa(K) \in \inf_{u \in \mathscr{C}_{\mathrm{TV},\hat{w}}(K) \setminus \{\mathbf{0}\}} \left\{ \frac{\|\boldsymbol{X}^B u\|_2}{\sqrt{n} \|u_K\|_2} \right\}$$

with

$$\mathscr{C}_{\mathrm{TV},\hat{w}}(K) = \left\{ u \in \mathbb{R}^d : \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j{}^{\complement}}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} \leq 2 \sum_{j=1}^{p} \|(u_{j,\bullet})_{K_j}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} \right\}. \tag{10}$$

We assume that the following condition holds

$$\kappa(K) > 0 \tag{11}$$

for any K satisfying (9).

The set $\mathscr{C}_{\mathrm{TV},\hat{w}}(K)$ is a cone composed by all vectors with a support "close" to K. Theorem 2 gives a risk bound for the estimator $m_{\hat{\theta}}$.

Theorem 2 *Let Assumptions 1 and 2 be satisfied. Fix $A > 0$ and choose*

$$\hat{w}_{j,k} = \sqrt{\frac{2U_n\phi(A + \log d)}{n}\,\hat{\pi}_{j,k}}. \tag{12}$$

Then, with probability at least $1 - 2e^{-A}$, any $\hat{\theta}$ given by (8) *satisfies*

$$R(m_{\hat{\theta}}) - R(m^0) \leq \inf_{\theta} \Big\{ 3(R(m_\theta) - R(m^0)) \\ + \frac{2560(C_b(C_n + \rho) + 2)}{L_n \kappa^2(J(\theta))}\, |J(\theta)|\, \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \Big\},$$

where the infimum is over the set of vectors $\theta \in B_d(\rho)$ such that $n_j^\top \theta_{j,\bullet} = 0$ for all $j = 1, \dots, p$ and such that $|J(\theta)| \leq J^*$.

The proof of Theorem 2 is given in Section 6.3 below. Note that the "variance" term or "complexity" term in the oracle inequality satisfies

$$|J(\theta)| \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \leq 2U_n\phi \frac{|J(\theta)|(A + \log d)}{n}. \tag{13}$$

The value $|J(\theta)|$ characterizes the sparsity of the vector θ, given by

$$|J(\theta)| = \sum_{j=1}^p |J_j(\theta)| = \sum_{j=1}^p |\{k = 1, \dots, d_j : \theta_{j,k} \neq \theta_{j,k-1}\}|.$$

It counts the number of non-equal consecutive values of θ. If θ is block-sparse, namely whenever $|\mathcal{J}(\theta)| \ll p$ where $\mathcal{J}(\theta) = \{j = 1, \dots, p : \theta_{j,\bullet} \neq 0_{d_j}\}$ (meaning that few raw features are useful for prediction), then $|J(\theta)| \leq |\mathcal{J}(\theta)| \max_{j \in \mathcal{J}(\theta)} |J_j(\theta)|$, which means that $|J(\theta)|$ is controlled by the block sparsity $|\mathcal{J}(\theta)|$.

The oracle inequality from Theorem 2 is stated uniformly for vectors $\theta \in B_d(\rho)$ satisfying $n_j^\top \theta_{j,\bullet} = 0$ for all $j = 1, \dots, p$ and $|J(\theta)| \leq J^*$. Writing this oracle inequality under the assumption $|J(\theta)| \leq J^*$ meets the standard way of stating sparse oracle inequalities, see e.g. Bühlmann and van De Geer (2011). Note that J^* is introduced in Assumption 2 and corresponds to a maximal sparsity for which the matrix $\boldsymbol{X}^B$ satisfies the restricted eigenvalue assumption. Also, the oracle inequality stated in Theorem 2 stands for vectors such that $n_j^\top \theta_{j,\bullet} = 0$, which is natural since the binarsity penalization imposes these extra linear constraints.

The assumption that $\theta \in B_d(\rho)$ is a technical one, that allows to establish a connection, via the notion of self-concordance, see Bach (2010), between the empirical squared ℓ_2-norm and the empirical Kullback divergence (see Lemma 9 in Section 6.3). It corresponds to a technical constraint which is commonly used in literature for the proof of oracle inequalities for sparse generalized linear models, see for instance van de Geer (2008), a recent contribution for the particular case of Poisson regression being Ivanoff et al. (2016). Also, note that

$$\max_{i=1,\dots,n} |\langle x_i^B, \theta \rangle| \leq \sum_{j=1}^p \|\theta_{j,\bullet}\|_\infty \leq |\mathcal{J}(\theta)| \times \|\theta\|_\infty, \tag{14}$$

where $\|\theta\|_\infty = \max_{j=1,\dots,p} \|\theta_{j,\bullet}\|_\infty$. The first inequality in (14) comes from the fact that the entries of $\boldsymbol{X}^B$ are in $\{0,1\}$, and it entails that $\max_{i=1,\dots,n} |\langle x_i^B, \theta\rangle| \leq \rho$ whenever $\theta \in B_d(\rho)$. The second inequality in (14) entails that ρ can be upper bounded by $|\mathcal{J}(\theta)| \times \|\theta\|_\infty$, and therefore the constraint $\theta \in B_d(\rho)$ becomes only a box constraint on θ, which depends on the dimensionality of the features through $|\mathcal{J}(\theta)|$ only. The fact that the procedure depends on ρ, and that the oracle inequality stated in Theorem 2 depends linearly on ρ is commonly found in literature about sparse generalized linear models, see van de Geer (2008); Bach (2010); Ivanoff et al. (2016). However, the constraint $B_d(\rho)$ is a technicality which is not used in the numerical experiments provided in Section 4 below.

In the next Section, we exhibit a consequence of Theorem 2, whenever one considers the Gaussian case (least-squares loss) and where m^0 has a sparse additive structure defined below. This structure allows to control the bias term from Theorem 2 and to exhibit a convergence rate.

3.2. Sparse linear additive regression

Theorem 2 allows to study a particular case, namely an additive model, see e.g. Hastie and Tibshirani (1990); Horowitz et al. (2006) and in particular a sparse additive linear model, which is of particular interest in high-dimensional statistics, see Meier et al. (2009); Ravikumar et al. (2009); Bühlmann and van De Geer (2011). We prove in Theorem 3 below that our procedure matches the convergence rates previously known from literature. In this setting, we work under the following assumptions.

Assumption 3 *We assume to simplify that $x_i \in [0,1]^d$ for all $i = 1,\dots,n$. We consider the Gaussian setting with the least-squares loss, namely $\ell(y,y') = \frac{1}{2}(y-y')^2$, $b(y) = \frac{1}{2}y^2$ and $\phi = \sigma^2$ (noise variance) in Equation* (2)*, with $L_n = U_n = 1$, $C_b = 0$ in Assumption 1. Moreover, we assume that m^0 has the following sparse additive structure*

$$m^0(x) = \sum_{j \in \mathcal{J}_*} m_j^0(x_j)$$

for $x = [x_1 \cdots x_p] \in \mathbb{R}^p$, where $m_j^0 : \mathbb{R} \to \mathbb{R}$ are L-Lipschitz functions, namely satisfying $|m_j^0(z) - m_j^0(z')| \leq L|z-z'|$ for any $z, z' \in \mathbb{R}$, and where $\mathcal{J}_* \subset \{1,\dots,p\}$ is a set of active features (sparsity means that $|\mathcal{J}_*| \ll p$). Also, we assume the following identifiability condition

$$\sum_{i=1}^n m_j^0(x_{i,j}) = 0$$

for all $j = 1,\dots,p$.

Assumption 3 contains identifiability and smoothness requirements that are standard when studying additive models, see e.g. Meier et al. (2009). We restrict the functions m_j^0 to be Lipschitz and not smoother, since our procedure produces a piecewise constant decision function with respect to each j, that can approximate optimally only Lipschitz functions. For more regular functions, our procedure would lead to suboptimal rates, see also the discussion below the statement of Theorem 3.

Theorem 3 *Consider procedure* (7) *with $d_j = D$, where D is the integer part of $n^{1/3}$, and $I_{j,1} = [0, \frac{1}{D}]$, $I_{j,k} = (\frac{k-1}{D}, \frac{k}{D}]$ for all $k = 2,\dots,D$ and $j = 1,\dots,p$, and keep the weights $\hat{w}_{j,k}$ the same*

as in Theorem 2. Introduce also $\theta^*_{j,k} = \sum_{i=1}^n m_j^0(x_{i,j})\mathbf{1}_{I_k}(x_{i,j}) / \sum_{i=1}^n \mathbf{1}_{I_k}(x_{i,j})$ *for* $j \in \mathcal{J}_*$ *and* $\theta^*_{j,\bullet} = \mathbf{0}_D$ *for* $j \notin \mathcal{J}_*$*. Then, under Assumption 2 with* $J^* = J(\theta^*)$ *and Assumption 3, we have*

$$\|m_{\hat{\theta}} - m^0\|_n^2 \leq \Big(3L^2|\mathcal{J}_*| + \frac{5120 M_n \sigma^2 (A + \log(pn^{1/3}M_n)}{\kappa^2(J(\theta^*))}\Big)\frac{|\mathcal{J}_*|}{n^{2/3}},$$

where $M_n = \max_{j=1,\dots,p} \max_{i=1,\dots,n} |m_j^0(x_{i,j})|$.

The proof of Theorem 3 is given in Section 6.8 below. It is an easy consequence of Theorem 2 under the sparse additive model assumption. It uses Assumption 2 with $J^* = J(\theta^*)$, since $\theta^*_{j,\bullet}$ is the minimizer of the bias for each $j \in \mathcal{J}_*$, see the proof of Theorem 3 for details.

The rate of convergence is, up to constants and logarithmic terms, of order $|\mathcal{J}_*|^2 n^{-2/3}$. Recalling that we work under a Lipschitz assumption, namely Hölder smoothness of order 1, the scaling of this rate w.r.t. to n is $n^{-2r/(2r+1)}$ with $r = 1$, which matches the one-dimensional minimax rate. This rate matches the one obtained in Bühlmann and van De Geer (2011), see Chapter 8 p. 272, where the rate $|\mathcal{J}_*|^2 n^{-2r/(2r+1)} = |\mathcal{J}_*|^2 n^{-4/5}$ is derived under a C^2 smoothness assumption, namely $r = 2$. Hence, Theorem 3 shows that, in the particular case of a sparse additive model, our procedure matches in terms of convergence rate the state of the art. Further improvements could consider more general smoothness (beyond Lipschitz) and adaptation with respect to the regularity, at the cost of a more complicated procedure which is beyond the scope of this paper.

4. Numerical experiments

In this section, we first illustrate the fact that the binarsity penalization is roughly only two times slower than basic ℓ_1-penalization, see the timings in Figure 3. We then compare binarsity to a large number of baselines, see Table 2, using 9 classical binary classification datasets obtained from the UCI Machine Learning Repository (Lichman, 2013), see Table 3.

For each method, we randomly split all datasets into a training and a test set (30% for testing), and all hyper-parameters are tuned on the training set using V-fold cross-validation with $V = 10$. For support vector machine with radial basis kernel (SVM), random forests (RF) and gradient boosting (GB), we use the reference implementations from the `scikit-learn` library (Pedregosa et al., 2011), and we use the `LogisticGAM` procedure from the `pygam` library[1] for the GAM baseline. The binarsity penalization is proposed in the `tick` library (Bacry et al., 2018), we provide sample code for its use in Figure 4. Logistic regression with no penalization or ridge penalization gave similar or lower scores for all considered datasets, and are therefore not reported in our experiments.

The binarsity penalization does not require a careful tuning of d_j (number of bins for the one-hot encoding of raw feature j). Indeed, past a large enough value, increasing d_j even further barely changes the results since the cut-points selected by the penalization do not change anymore. This is illustrated in Figure 5, where we observe that past 50 bins, increasing d_j even further does not affect the performance, and only leads to an increase of the training time. In all our experiments, we therefore fix $d_j = 50$ for $j = 1, \dots, p$.

The results of all our experiments are reported in Figures 6 and 7. In Figure 6 we compare the performance of binarsity with the baselines on all 9 datasets, using ROC curves and the Area Under the Curve (AUC), while we report computing (training) timings in Figure 7. We observe that binarsity consistently outperforms Lasso, as well as Group L1: this highlights the importance of the TV

1. `https://github.com/dswah/pyGAM`

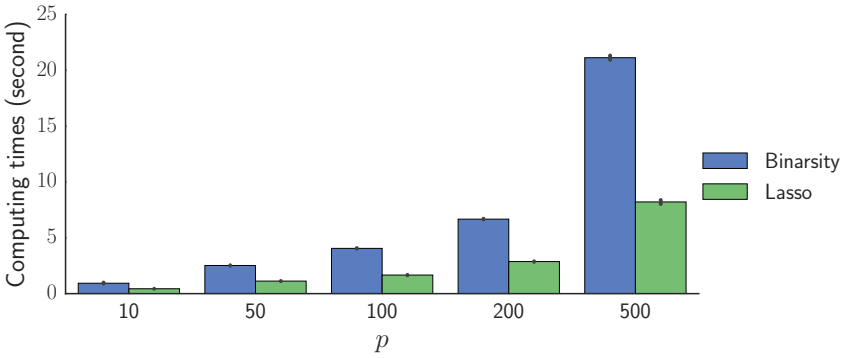

Fig. 3: Average computing time in second (with the black lines representing $\pm$ the standard deviation) obtained on 100 simulated datasets for training a logistic model with binarsity VS Lasso penalization, both trained on $\boldsymbol{X}^B$ with $d_j = 10$ for all $j \in 1, \ldots, p$. Features are Gaussian with a Toeplitz covariance matrix with correlation 0.5 and $n = 10000$. Note that the computing time ratio between the two methods stays roughly constant and equal to 2.

Name	Description	Reference
Lasso	Logistic regression (LR) with ℓ_1 penalization	Tibshirani (1996b)
Group L1	LR with group ℓ_1 penalization	Meier et al. (2008)
Group TV	LR with group total-variation penalization	
SVM	Support vector machine with radial basis kernel	Schölkopf and Smola (2002)
GAM	Generalized additive model	Hastie and Tibshirani (1990)
RF	Random forest classifier	Breiman (2001)
GB	Gradient boosting	Friedman (2002)

Tab. 2: Baselines considered in our experiments. Note that Group L1 and Group TV are considered on binarized features.

Dataset	#Samples	#Features	Reference
Ionosphere	351	34	Sigillito et al. (1989)
Churn	3333	21	Lichman (2013)
Default of credit card	30000	24	Yeh and Lien (2009)
Adult	32561	14	Kohavi (1996)
Bank marketing	45211	17	Moro et al. (2014)
Covertype	550088	10	Blackard and Dean (1999)
SUSY	5000000	18	Baldi et al. (2014)
HEPMASS	10500000	28	Baldi et al. (2016)
HIGGS	11000000	24	Baldi et al. (2014)

Tab. 3: Basic informations about the 9 considered datasets.

```python
# input: features X, labels y
from tick.inference import LogisticRegression
from tick.preprocessing import FeaturesBinarizer
from sklearn.model_selection import train_test_split

# binarize data
binarizer = FeaturesBinarizer(n_cuts=50)
X = binarizer.fit_transform(X)

# shuffle and split training and test sets
X, X_test, y, y_test = train_test_split(X, y, stratify=y)

# fit the model
learner = LogisticRegression(penalty='binarsity', C=1,
                             blocks_start=binarizer.blocks_start,
                             blocks_length=binarizer.blocks_length)
learner.fit(X, y)

# predict on test set
y_pred = learner.predict_proba(X_test)[:, 1]
```

Fig. 4: Sample python code for the use of binarsity with logistic regression in the `tick` library, with the use of the `FeaturesBinarizer` transformer for features binarization.

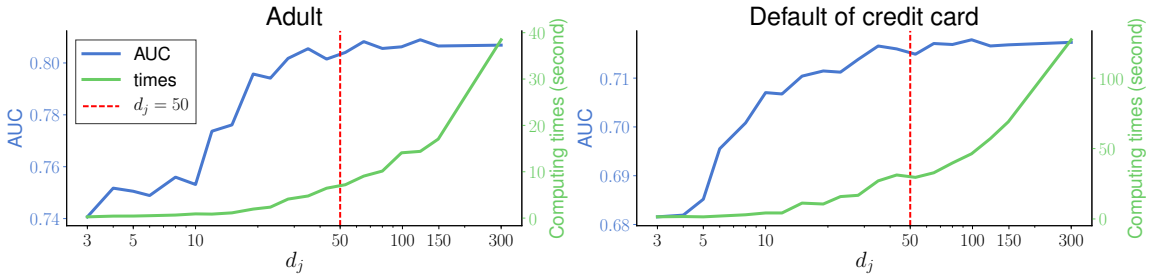

Fig. 5: Impact of the number of bins used in each block (d_j) on the classification performance (measured by AUC) and on the training time using the "Adult" and "Default of credit card" datasets. All d_j are equal for $j = 1, \ldots, p$, and we consider in all cases the best hyper-parameters selected after cross validation. We observe that past $d_j = 50$ bins, performance is roughly constant, while training time strongly increases.

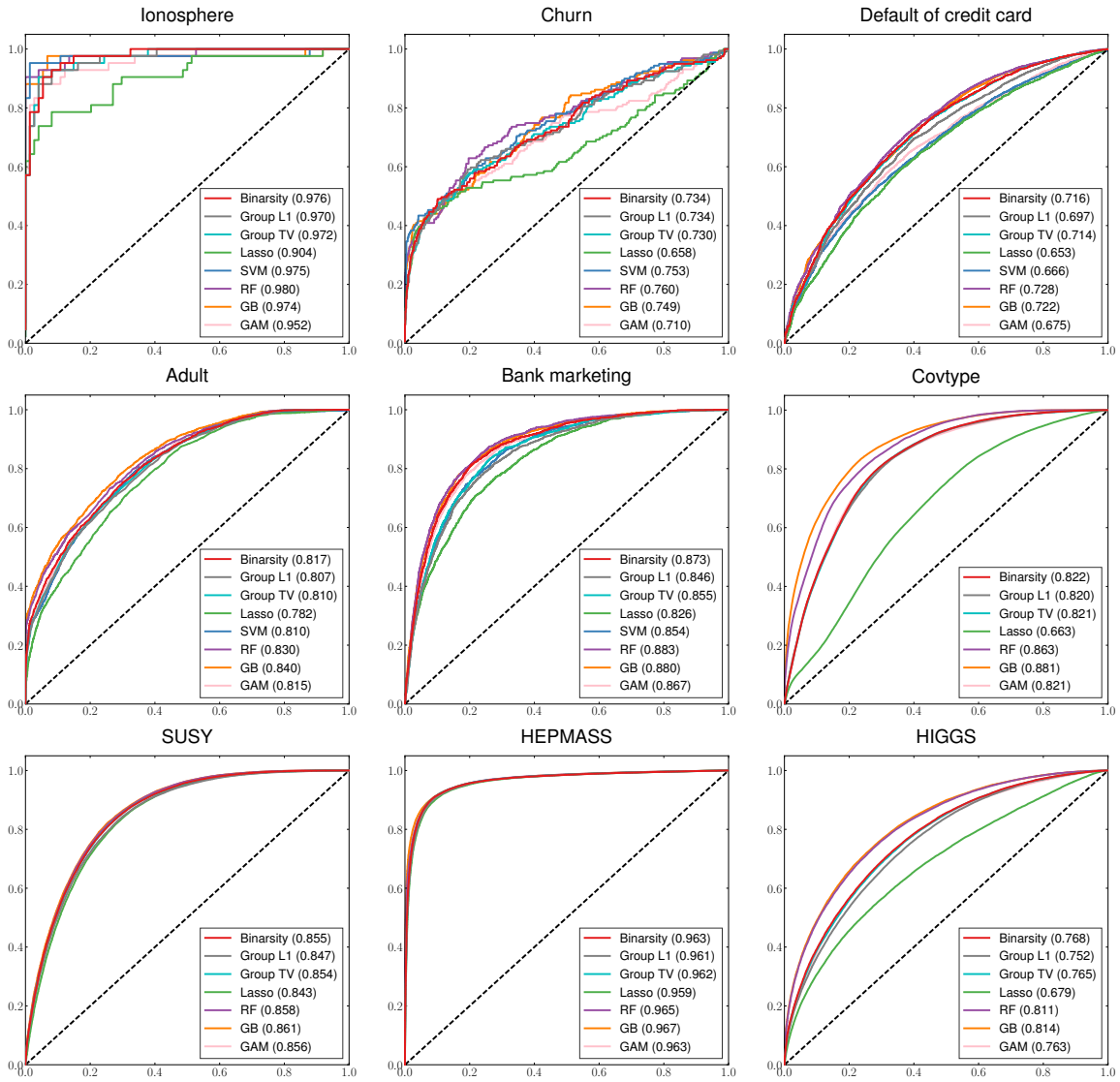

Fig. 6: Performance comparison using ROC curves and AUC scores (given between parenthesis) computed on test sets. The 4 last datasets contain too many examples for SVM (RBF kernel). Binarsity consistently does a better job than Lasso, Group L1, Group TV and GAM. Its performance is comparable to SVM, RF and GB but with computational timings that are orders of magnitude faster, see Figure 7.

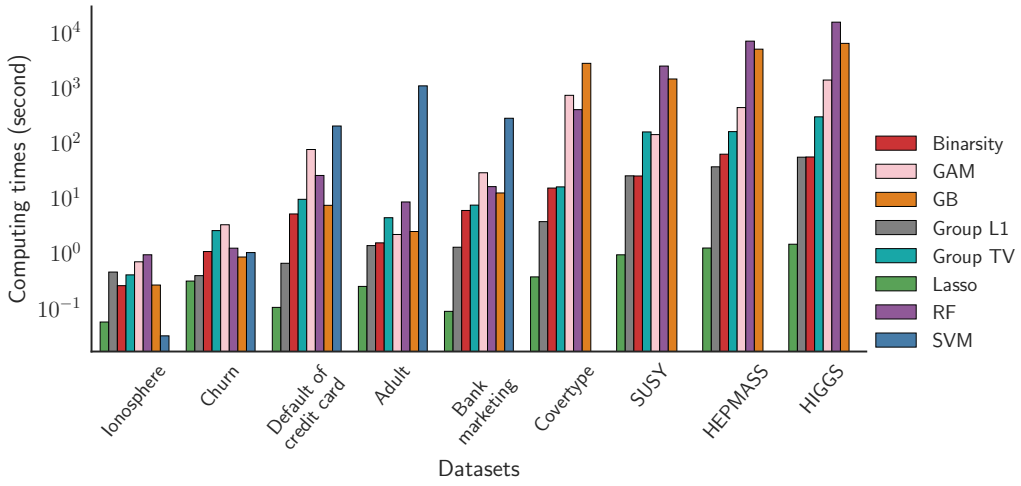

Fig. 7: Computing time comparisons (in seconds) between the methods on the considered datasets. Note that the time values are log-scaled. These timings concern the learning task for each model with the best hyper parameters selected, after the cross validation procedure. The 4 last datasets contain too many examples for the SVM with RBF kernel to be trained in a reasonable time. Roughly, binarsity is between 2 and 5 times slower than ℓ_1 penalization on the considered datasets, but is more than 100 times faster than random forests or gradient boosting algorithms on large datasets, such as HIGGS.

norm within each group. The AUC of Group TV is always slightly below the one of binarsity, and more importantly it involves a much larger training time: convergence is slower for Group TV, since it does not use the linear constraint of binarsity, leading to a ill-conditioned problem (sum of binary features equals 1 in each block). Finally, binarsity outperforms also GAM and its performance is comparable in all considered examples to RF and GB, with computational timings that are orders of magnitude faster, see Figure 7. All these experiments illustrate that binarsity achieves an extremely competitive compromise between computational time and performance, compared to all considered baselines.

5. Conclusion

In this paper, we introduced the binarsity penalization for one-hot encodings of continuous features. We illustrated the good statistical properties of binarsity for generalized linear models by proving non-asymptotic oracle inequalities. We conducted extensive comparisons of binarsity with state-of-the-art algorithms for binary classification on several standard datasets. Experimental results illustrate that binarsity significantly outperforms Lasso, Group L1 and Group TV penalizations and also generalized additive models, while being competitive with random forests and boosting. Moreover, it can be trained orders of magnitude faster than boosting and other ensemble methods. Even more importantly, it provides interpretability. Indeed, in addition to the raw feature selection ability of binarsity, the method pinpoints significant cut-points for all continuous feature. This leads to a much more precise and deeper understanding of the model than the one provided by Lasso on raw features. These results illustrate the fact that binarsity achieves an extremely competitive compromise between computational time and performance, compared to all considered baselines.

6. Proofs

In this Section we gather the proofs of all the theoretical results proposed in the paper. Throughout this Section, we denote by $\partial(\phi)$ the subdifferential mapping of a convex function ϕ.

6.1. Proof of Proposition 1

Recall that the indicator function δ_j is given by (6). For any fixed $j = 1, \dots, p$, we prove that $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}+\delta_j}$ is the composition of $\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}}$ and prox_{δ_j}, namely

$$\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}+\delta_j}(\theta_{j,\bullet}) = \mathrm{prox}_{\delta_j}\left(\mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}_{j,\bullet}}}(\theta_{j,\bullet})\right)$$

for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$. Using Theorem 1 in Yu (2013), it is sufficient to show that for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$, we have

$$\partial\big(\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}\big) \subseteq \partial\big(\|\,\mathrm{prox}_{\delta_j}(\theta_{j,\bullet})\|_{\mathrm{TV},\hat{w}_{j,\bullet}}\big). \tag{15}$$

We have $\mathrm{prox}_{\delta_j}(\theta_{j,\bullet}) = \Pi_{\mathrm{span}\{n_j\}^\perp}(\theta_{j,\bullet})$, where $\Pi_{\mathrm{span}\{n_j\}^\perp}(\cdot)$ stands for the projection onto the orthogonal of $\mathrm{span}\{n_j\}$. This projection simply writes

$$\Pi_{\mathrm{span}\{n_j\}^\perp}(\theta_{j,\bullet}) = \theta_{j,\bullet} - \frac{n_j^\top \theta_{j,\bullet}}{\|n_j\|_2^2} n_j$$

Now, let us define the $d_j \times d_j$ matrix D_j by

$$D_j = \begin{bmatrix} 1 & 0 & & 0 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{d_j} \times \mathbb{R}^{d_j}. \tag{16}$$

We then remark that for all $\theta_{j,\bullet} \in \mathbb{R}^{d_j}$,

$$\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \sum_{k=2}^{d_j} \hat{w}_{j,k}|\theta_{j,k} - \theta_{j,k-1}| = \|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1. \tag{17}$$

Using subdifferential calculus (see details in the proof of Proposition 5 below), one has

$$\partial\big(\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}\big) = \partial\big(\|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1\big) = {D_j}^\top \hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j\theta_{j,\bullet}).$$

Then, the linear constraint $n_j^\top \theta_{j,\bullet} = 0$ entails

$${D_j}^\top \hat{w}_{j,\bullet} \odot \mathrm{sign}(D_j\theta_{j,\bullet}) = {D_j}^\top \hat{w}_{j,\bullet} \odot \mathrm{sign}\left(D_j\big(\theta_{j,\bullet} - \frac{n_j^\top \theta_{j,\bullet}}{\|n_j\|_2^2} n_j\big)\right),$$

which leads to (15) and concludes the proof of the Proposition. □

6.2. Proximal operator of the weighted TV penalization

We recall in Algorithm 2 an algorithm provided in Alaya et al. (2015) for the computation of the proximal operator of the weighted total-variation penalization

$$\beta = \mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}(\theta) \in \mathrm{argmin}_{\theta\in\mathbb{R}^m} \left\{ \frac{1}{2}\|\beta-\theta\|_2^2 + \|\theta\|_{\mathrm{TV},\hat{w}} \right\}. \tag{18}$$

A quick explanation of this algorithm is as follows. The algorithm runs forwardly through the input vector $(\theta_1, \ldots, \theta_m)$. Using Karush-Kuhn-Tucker (KKT) optimality conditions (Boyd and Vandenberghe, 2004), we have that at a location k, the weight β_k stays constant whenever $|u_k| < \hat{w}_{k+1}$, where u_k is a solution to a dual problem associated to the primal problem (18). If not possible, it goes back to the last location where a jump can be introduced in β, validates the current segment until this location, starts a new segment, and continues.

6.3. Proof of Theorem 2

The proof relies on several technical properties that are described below. From now on, we consider $\boldsymbol{y} = [y_1 \cdots y_n]^\top$, $\boldsymbol{X} = [x_1 \cdots x_n]^\top$, $m^0(\boldsymbol{X}) = [m^0(x_1) \cdots m^0(x_n)]^\top$, and recalling that $m_\theta(x_i) = \theta^\top x_i^B$) we introduce $m_\theta(\boldsymbol{X}) = [m_\theta(x_1) \cdots m_\theta(x_n)]^\top$ and $b'(m_\theta(\boldsymbol{X})) = [b'(m_\theta(x_1)) \cdots b'(m_\theta(x_n))]^\top$.

Let us now define the Kullback-Leibler divergence between the true probability density funtion f^0 defined in (2) and a candidate f_θ within the generalized linear model $f_\theta(y|x) = \exp\big(ym_\theta(x) - b(m_\theta(x))$ as follows

$$\begin{aligned}\mathrm{KL}_n(f^0, f_\theta) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}_{\mathbf{y}|X}}\Big[\log\frac{f^0(y_i|x_i)}{f_\theta(y_i|x_i)}\Big] \\ &:= \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),\end{aligned}$$

where $\mathbb{P}_{\mathbf{y}|X}$ is the joint distribution of $\boldsymbol{y}$ given $\boldsymbol{X}$. We then have the following Lemma.

Lemma 4 *The excess risk satisfies*

$$R(m_\theta) - R(m^0) = \phi \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$

where we recall that ϕ is the dispertion parameter of the generalized linear model, see (2)*.*

Proof. If follows from the following simple computation

$$\begin{aligned}&\mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\ &\quad= \phi^{-1}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}_{\mathbf{y}|X}}\Big[\big(-y_i m_\theta(x_i) + b(m_\theta(x_i))\big) - \big(-y_i m^0(x_i) + b(m^0(x_i))\big)\Big] \\ &\quad= \phi^{-1}\big(R(m_\theta) - R(m^0)\big)\end{aligned}$$

which proves the Lemma. □

Algorithm 2: Proximal operator of weighted TV penalization

Input: vector $\theta = \big(\theta_1, \dots, \theta_m\big)^\top \in \mathbb{R}^m$ and weights $\hat{w} = (\hat{w}_1, \dots, \hat{w}_m) \in \mathbb{R}_+^m$.
Output: vector $\beta = \mathrm{prox}_{\|\cdot\|_{\mathrm{TV},\hat{w}}}(\theta)$

Set $k = k_0 = k_- = k_+ \leftarrow 1$
$\beta_{\min} \leftarrow \theta_1 - \hat{w}_2$; $\beta_{\max} \leftarrow \theta_1 + \hat{w}_2$
$u_{\min} \leftarrow \hat{w}_2$; $u_{\max} \leftarrow -\hat{w}_2$

if $k = m$ **then**
  $\beta_m \leftarrow \beta_{\min} + u_{\min}$

if $\theta_{k+1} + u_{\min} < \beta_{\min} - \hat{w}_{k+2}$ **then** `/* negative jump */`
  $\beta_{k_0} = \dots = \beta_{k_-} \leftarrow \beta_{\min}$
  $k = k_0 = k_- = k_+ \leftarrow k_- + 1$
  $\beta_{\min} \leftarrow \theta_k - \hat{w}_{k+1} + \hat{w}_k$; $\beta_{\max} \leftarrow \theta_k + \hat{w}_{k+1} + \hat{w}_k$
  $u_{\min} \leftarrow \hat{w}_{k+1}$; $u_{\max} \leftarrow -\hat{w}_{k+1}$

else if $\theta_{k+1} + u_{\max} > \beta_{\max} + \hat{w}_{k+2}$ **then** `/* positive jump */`
  $\beta_{k_0} = \dots = \beta_{k_+} \leftarrow \beta_{\max}$
  $k = k_0 = k_- = k_+ \leftarrow k_+ + 1$
  $\beta_{\min} \leftarrow \theta_k - \hat{w}_{k+1} - \hat{w}_k$; $\beta_{\max} \leftarrow \theta_k + \hat{w}_{k+1} - \hat{w}_k$
  $u_{\min} \leftarrow \hat{w}_{k+1}$; $u_{\max} \leftarrow -\hat{w}_{k+1}$

else `/* no jump */`
  set $k \leftarrow k + 1$
    $u_{\min} \leftarrow \theta_k + \hat{w}_{k+1} - \beta_{\min}$
    $u_{\max} \leftarrow \theta_k - \hat{w}_{k+1} - \beta_{\max}$ **if** $u_{\min} \geq \hat{w}_{k+1}$ **then**
    $\beta_{\min} \leftarrow \beta_{\min} + \frac{u_{\min} - \hat{w}_{k+1}}{k - k_0 + 1}$
    $u_{\min} \leftarrow \hat{w}_{k+1}$
    $k_- \leftarrow k$
  if $u_{\max} \leq -\hat{w}_{k+1}$ **then**
    $\beta_{\max} \leftarrow \beta_{\max} + \frac{u_{\max} + \hat{w}_{k+1}}{k - k_0 + 1}$
    $u_{\max} \leftarrow -\hat{w}_{k+1}$
    $k_+ \leftarrow k$

if $k < m$ **then**
  go to **3.**

if $u_{\min} < 0$ **then**
  $\beta_{k_0} = \dots = \beta_{k_-} \leftarrow \beta_{\min}$
  $k = k_0 = k_- \leftarrow k_- + 1$
  $\beta_{\min} \leftarrow \theta_k - \hat{w}_{k+1} + \hat{w}_k$
  $u_{\min} \leftarrow \hat{w}_{k+1}$; $u_{\max} \leftarrow \theta_k + \hat{w}_k - v_{\max}$
  go to **2.**

else if $u_{\max} > 0$ **then**
  $\beta_{k_0} = \dots = \beta_{k_+} \leftarrow \beta_{\max}$
  $k = k_0 = k_+ \leftarrow k_+ + 1$
  $\beta_{\max} \leftarrow \theta_k + \hat{w}_{k+1} - \hat{w}_k$
  $u_{\max} \leftarrow -\hat{w}_{k+1}$; $u_{\min} \leftarrow \theta_k - \hat{w}_k - u_{\min}$
  go to **2.**

else
  $\beta_{k_0} = \dots = \beta_m \leftarrow \beta_{\min} + \frac{u_{\min}}{k - k_0 + 1}$

6.4. Optimality conditions

As explained in the following Proposition, a solution to problem (8) can be characterized using the Karush-Kuhn-Tucker (KKT) optimality conditions (Boyd and Vandenberghe, 2004).

Proposition 5 *A vector $\hat{\theta} = [\hat{\theta}_{1,\bullet}^\top \cdots \hat{\theta}_{p,\bullet}^\top]^\top \in \mathbb{R}^d$ is an optimum of the objective function* (8) *if and only if there are subgradients $\hat{h} = [\hat{h}_{j,\bullet}]_{j=1,\dots,p} \in \partial\|\hat{\theta}\|_{\mathrm{TV},\hat{w}}$ and $\hat{g} = [\hat{g}_{j,\bullet}]_{j=1,\dots,p} \in \partial[\delta_j(\hat{\theta}_{j,\bullet})]_{j=1,\dots,p}$ such that*

$$\nabla R_n(\hat{\theta}_{j,\bullet}) + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} = \mathbf{0},$$

where

$$\begin{cases} \hat{h}_{j,\bullet} = D_j^\top\big(\hat{w}_{j,\bullet} \odot \operatorname{sign}(D_j\hat{\theta}_{j,\bullet})\big) & \text{if } j \in J(\hat{\theta}), \\ \hat{h}_{j,\bullet} \in D_j^\top\big(\hat{w}_{j,\bullet} \odot [-1,+1]^{d_j}\big) & \text{if } j \in J^{\complement}(\hat{\theta}), \end{cases} \tag{19}$$

and where we recall that $J(\hat{\theta})$ is the support set of $\hat{\theta}$. The subgradient $\hat{g}_{j,\bullet}$ belongs to

$$\partial\big(\delta_j(\hat{\theta}_{j,\bullet})\big) = \big\{\mu_{j,\bullet} \in \mathbb{R}^{d_j} : \mu_{j,\bullet}^\top\theta_{j,\bullet} \le \mu_{j,\bullet}^\top\hat{\theta}_{j,\bullet} \text{ for all } \theta_{j,\bullet} \text{ such that } n_j^\top\theta_{j,\bullet} = 0\big\}.$$

For the generalized linear model, we have

$$\frac{1}{n}\big(\boldsymbol{X}^B_{\bullet,j}\big)^\top\big(b'(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{y}\big) + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} + \hat{f}_{j,\bullet} = \mathbf{0}, \tag{20}$$

where $\hat{f} = [\hat{f}_{j,\bullet}]_{j=1,\dots,p}$ belongs to the normal cone of the ball $B_d(\rho)$.

Proof. The function $\theta \mapsto R_n(\theta)$ is differentiable, so the subdifferential of $R_n(\cdot) + \operatorname{bina}(\cdot)$ at a point $\theta = (\theta_{j,\bullet})_{j=1,\dots,p} \in \mathbb{R}^d$ is given by

$$\partial(R_n(\theta) + \operatorname{bina}(\theta)) = \nabla R_n(\theta) + \partial(\operatorname{bina}(\theta)),$$

where $\nabla R_n(\theta) = \Big[\frac{\partial R_n(\theta)}{\partial\theta_{1,\bullet}} \cdots \frac{\partial R_n(\theta)}{\partial\theta_{p,\bullet}}\Big]^\top$ and

$$\partial\operatorname{bina}(\theta) = \Big[\partial\|\theta_{1,\bullet}\|_{\mathrm{TV},\hat{w}_{1,\bullet}} + \partial\delta_j(\theta_{1,\bullet}) \ \cdots \ \partial\|\theta_{p,\bullet}\|_{\mathrm{TV},\hat{w}_{p,\bullet}} + \partial\delta_j(\theta_{p,\bullet})\Big]^\top.$$

We have $\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \|\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}\|_1$ for all $j = 1,\dots,p$. Then, by applying some properties of the subdifferential calculus, we get

$$\partial\|\theta_{j,\bullet}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} = \begin{cases} D_j^\top \operatorname{sign}(\hat{w}_{j,\bullet} \odot D_j\theta_{j,\bullet}) & \text{if } D_j\theta \neq \mathbf{0}, \\ D_j^\top\big(\hat{w}_{j,\bullet} \odot v_j\big) & \text{otherwise,} \end{cases} \tag{21}$$

where $v_j \in [-1,+1]^{d_j}$ for all $j = 1,\dots,p$. For generalized linear models, we rewrite

$$\hat{\theta} \in \operatorname{argmin}_{\theta\in\mathbb{R}^d}\big\{R_n(\theta) + \operatorname{bina}(\theta) + \delta_{B_d(\rho)}(\theta)\big\}, \tag{22}$$

where $\delta_{B_d(\rho)}$ is the indicator function of $B_d(\rho)$. Now, $\hat{\theta} = [\hat{\theta}_{1,\bullet}^\top \cdots \hat{\theta}_{p,\bullet}^\top]^\top$ is an optimum of (22) if and only if $\mathbf{0} \in \nabla R_n(m_{\hat{\theta}}) + \partial\|\hat{\theta}\|_{\mathrm{TV},\hat{w}} + \partial\delta_{B_d(\rho)}(\hat{\theta})$. Recall that the subdifferential of $\delta_{B_d(\rho)}(\cdot)$ is the normal cone of $B_d(\rho)$, namely

$$\partial\delta_{B_d(\rho)}(\hat{\theta}) = \big\{\eta \in \mathbb{R}^d : \eta^\top\theta \le \eta^\top\hat{\theta} \text{ for all } \theta \in B_d(\rho)\big\}. \tag{23}$$

One has

$$\frac{\partial R_n(\theta)}{\partial\theta_{j,\bullet}} = \frac{1}{n}(\boldsymbol{X}^B_{\bullet,j})^\top(b'(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{y}), \tag{24}$$

so that together with (24) and (23) we obtain (20), which concludes the proof of Proposition 5. □

6.5. Compatibility conditions

Let us define the block diagonal matrix $\mathbf{D} = \mathrm{diag}(D_1, \ldots, D_p)$ with D_j defined in (16). We denote its inverse T_j which is defined by the $d_j \times d_j$ lower triangular matrix with entries $(T_j)_{r,s} = 0$ if $r < s$ and $(T_j)_{r,s} = 1$ otherwise. We set $\mathbf{T} = \mathrm{diag}(T_1, \ldots, T_p)$, so that one has $\mathbf{D}^{-1} = \mathbf{T}$.

In order to prove Theorem 2, we need the following results which give a compatibility property (van de Geer, 2008; van de Geer and Lederer, 2013; Dalalyan et al., 2017) for the matrix $\mathbf{T}$, see Lemma 6 below and for the matrix $\boldsymbol{X}^B\mathbf{T}$, see Lemma 7 below. For any concatenation of subsets $K = [K_1, \ldots, K_p]$, we set

$$K_j = \{\tau_j^1, \ldots, \tau_j^{b_j}\} \subset \{1, \ldots, d_j\} \tag{25}$$

for all $j = 1, \ldots, p$ with the convention that $\tau_j^0 = 0$ and $\tau_j^{b_j+1} = d_j + 1$.

Lemma 6 *Let $\gamma \in \mathbb{R}_+^d$ be given and $K = [K_1, \ldots, K_p]$ with K_j given by* (25) *for all $j = 1, \ldots, p$. Then, for every $u \in \mathbb{R}^d \backslash \{\mathbf{0}\}$, we have*

$$\frac{\|\mathbf{T}u\|_2}{|\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1|} \geq \kappa_{\mathbf{T},\gamma}(K),$$

where

$$\kappa_{\mathbf{T},\gamma}(K) = \left\{ 32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2|K_j| \|\gamma_{j,\bullet}\|_\infty^2 \Delta_{\min,K_j}^{-1} \right\}^{-1/2},$$

and $\Delta_{\min,K_j} = \min_{r=1,\ldots b^j} |\tau_j^{r_j} - \tau_j^{r_j-1}|$.

Proof. Using Proposition 3 in Dalalyan et al. (2017), we have

$$\begin{aligned}
&\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1 \\
&\qquad = \sum_{j=1}^{p} \|u_{K_j} \odot \gamma_{K_j}\|_1 - \|u_{K_j^\complement} \odot \gamma_{K_j^\complement}\|_1 \\
&\qquad \leq \sum_{j=1}^{p} 4\|T_j u_{j,\bullet}\|_2 \left\{ 2 \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2(b_j+1)\|\gamma_{j,\bullet}\|_\infty^2 \Delta_{\min,K_j}^{-1} \right\}^{1/2}.
\end{aligned}$$

Using Hölder's inequality for the right hand side of the last inequality gives

$$\begin{aligned}
&\|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1 \\
&\qquad \leq \|\mathbf{T}u\|_2 \left\{ 32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2|K_j| \|\gamma_{j,\bullet}\|_\infty^2 \Delta_{\min,K_j}^{-1} \right\}^{1/2},
\end{aligned}$$

which completes the proof of the Lemma. □

Combining Assumption 2 and Lemma 6 allows to establish a compatibility condition satisfied by $\boldsymbol{X}^B\mathbf{T}$.

Lemma 7 *Let $\gamma \in \mathbb{R}_+^d$ be given and $K = [K_1, \dots, K_p]$ with K_j given by* (25) *for $j = 1, \dots, p$. Then, if Assumption 2 holds, one has*

$$\inf_{u \in \mathscr{C}_{1,\hat{w}}(K) \setminus \{\mathbf{0}\}} \left\{ \frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n} \mid \|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1|} \right\} \geq \kappa_{\mathbf{T},\gamma}(K) \kappa(K), \tag{26}$$

where

$$\mathscr{C}_{1,\hat{w}}(K) = \Big\{ u \in \mathbb{R}^d : \sum_{j=1}^p \|(u_{j,\bullet})_{K_j^\complement}\|_{1,\hat{w}_{j,\bullet}} \leq 2 \sum_{j=1}^p \|(u_{j,\bullet})_{K_j}\|_{1,\hat{w}_{j,\bullet}} \Big\}. \tag{27}$$

Proof. Lemma 6 gives

$$\frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n} | \|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1|} \geq \kappa_{\mathbf{T},\gamma}(K) \frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n} \|\mathbf{T} u\|_2}.$$

Now, we note that if $u \in \mathscr{C}_{1,\hat{w}}(K)$, then $\mathbf{T} u \in \mathscr{C}_{\mathrm{TV},\hat{w}}(K)$. Hence, Assumption 2 entails

$$\frac{\|\boldsymbol{X}^B \mathbf{T} u\|_2}{\sqrt{n} | \|u_K \odot \gamma_K\|_1 - \|u_{K^\complement} \odot \gamma_{K^\complement}\|_1|} \geq \kappa_{\mathbf{T},\gamma}(K) \kappa(K),$$

which concludes the proof of the Lemma. □

6.6. Connection between the empirical Kullback-Leibler divergence and the empirical squared norm

The next Lemma is from Bach (2010) (see Lemma 1 herein).

Lemma 8 *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a three times differentiable convex function such that for all $t \in \mathbb{R}$, $|\varphi'''(t)| \leq M |\varphi''(t)|$ for some $M \geq 0$. Then, for all $t \geq 0$, one has*

$$\frac{\varphi''(0)}{M^2} \psi(-Mt) \leq \varphi(t) - \varphi(0) - \varphi'(0) t \leq \frac{\varphi''(0)}{M^2} \psi(Mt),$$

with $\psi(u) = e^u - u - 1$.

This Lemma entails the following in our setting.

Lemma 9 *Under Assumption 1, one has*

$$\frac{L_n \psi(-2(C_n + \rho))}{4\phi(C_n + \rho)^2} \frac{1}{n} \|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \leq \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$
$$\frac{U_n \psi(2(C_n + \rho))}{4\phi(C_n + \rho)^2} \frac{1}{n} \|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \geq \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$

for all $\theta \in B_d(\rho)$.

Proof. Let us consider the function $G_n : \mathbb{R} \to \mathbb{R}$ defined by $G_n(t) = R_n(m^0 + t m_\eta)$, with m_η to be defined later, which writes

$$G_n(t) = \frac{1}{n} \sum_{i=1}^n b(m^0(x_i) + t m_\eta(x_i)) - \frac{1}{n} \sum_{i=1}^n y_i (m^0(x_i) + t m_\eta(x_i)).$$

We have

$$G_n'(t) = \frac{1}{n}\sum_{i=1}^n m_\eta(x_i) b'(m^0(x_i) + t m_\eta(x_i)) - \frac{1}{n}\sum_{i=1}^n y_i m_\eta(x_i),$$
$$G_n''(t) = \frac{1}{n}\sum_{i=1}^n m_\eta^2(x_i) b''(m^0(x_i) + t m_\eta(x_i)),$$
$$\text{and} \quad G_n'''(t) = \frac{1}{n}\sum_{i=1}^n m_\eta^3(x_i) b'''(m^0(x_i) + t m_\eta(x_i)).$$

Using Assumption 1, we have $|G_n'''(t)| \leq C_b \|m_\eta\|_\infty |G_n''(t)|$ where $\|m_\eta\|_\infty := \max_{i=1,\dots,n} |m_\eta(x_i)|$. Lemma 8 with $M = C_b\|m_\eta\|_\infty$ gives

$$G_n''(0)\frac{\psi(-C_b\|m_\eta\|_\infty t)}{C_b^2\|m_\eta\|_\infty^2} \leq G_n(t) - G_n(0) - tG_n'(0) \leq G_n''(0)\frac{\psi(C_b\|m_\eta\|_\infty t)}{C_b^2\|m_\eta\|_\infty^2}$$

for all $t \geq 0$ and $t = 1$ leads to

$$G_n''(0)\frac{\psi(-C_b\|m_\eta\|_\infty)}{C_b^2\|m_\eta\|_\infty^2} \leq R_n(m^0 + m_\eta) - R_n(m^0) - G_n'(0) \leq G_n''(0)\frac{\psi(C_b\|m_\eta\|_\infty)}{C_b^2\|m_\eta\|_\infty^2}.$$

An easy computation gives

$$-G_n'(0) = \frac{1}{n}\sum_{i=1}^n m_\eta(x_i)\big(y_i - b'(m^0(x_i))\big) \;\text{ and }\; G_n''(0) = \frac{1}{n}\sum_{i=1}^n m_\eta^2(x_i) b''(m_\eta(x_i)),$$

and since obviously $\mathbb{E}_{\mathbb{P}_{\mathbf{y}|X}}[G_n'(0)] = 0$, we obtain

$$G_n''(0)\frac{\psi(-C_b\|m_\eta\|_\infty)}{C_b^2\|m_\eta\|_\infty^2} \leq R(m^0 + m_\eta) - R(m^0) \leq G_n''(0)\frac{\psi(C_b\|m_\eta\|_\infty)}{C_b^2\|m_\eta\|_\infty^2}.$$

Now, choosing $m_\eta = m_\theta - m^0$ and combining Assumption 1 with Equation (14) gives

$$C_b\|m_\eta\|_\infty \leq C_b \max_{i=1,\dots,n}(|\langle x_i^B, \theta\rangle| + |m^0(x_i)|) \leq C_b(\rho + C_n).$$

Hence, since $x \mapsto \psi(x)/x^2$ is an increasing function on $\mathbb{R}^+$, we end up with

$$G_n''(0)\frac{\psi(-C_b(C_n+\rho))}{C_b^2(C_n+\rho)^2} \leq R(m_\theta) - R(m_0) = \phi \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$
$$G_n''(0)\frac{\psi(C_b(C_n+\rho))}{C_b^2(C_n+\rho)^2} \geq R(m_\theta) - R(m_0) = \phi \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$

and since $G_n''(0) = n^{-1}\sum_{i=1}^n (m_\theta(x_i) - m^0(x_i))^2 b''(m^0(x_i))$, we obtain

$$\frac{L_n\psi(-C_b(C_n+\rho))}{C_n^2\phi(C_n+\rho)^2}\frac{1}{n}\|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \leq \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$
$$\frac{U_n\psi(C_b(C_n+\rho))}{C_b^2\phi(C_n+\rho)^2}\frac{1}{n}\|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2^2 \geq \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$

which concludes the proof of the Lemma. □

6.7. Proof of Theorem 2

Let us recall that

$$R_n(m_\theta) = \frac{1}{n}\sum_{i=1}^n b(m_\theta(x_i)) - \frac{1}{n}\sum_{i=1}^n y_i m_\theta(x_i)$$

for all $\theta \in \mathbb{R}^d$ and that

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in B_d(\rho)} \big\{ R_n(\theta) + \operatorname{bina}(\theta) \big\}. \tag{28}$$

Proposition 5 above entails that there is $\hat{h} = [\hat{h}_{j,\bullet}]_{j=1,\dots,p} \in \partial \|\hat{\theta}\|_{\mathrm{TV},\hat{w}}$, $\hat{g} = [\hat{g}_{j,\bullet}]_{j=1,\cdots,p} \in [\partial \delta_j(\hat{\theta}_{j,\bullet})]_{j=1,\dots,p}$ and $\hat{f} = [\hat{f}_{j,\bullet}]_{j=1,\dots,p} \in \partial \delta_{B_d(\rho)}(\hat{\theta})$ such that

$$\Big\langle \frac{1}{n}(\boldsymbol{X}^B)^\top (b'(m_{\hat{\theta}}(\boldsymbol{X})) - \boldsymbol{y}) + \hat{h} + \hat{g} + \hat{f}, \hat{\theta} - \theta \Big\rangle = 0$$

for all $\theta \in \mathbb{R}^d$. This can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle \\
&\qquad - \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle + \langle \hat{h} + \hat{g} + \hat{f}, \hat{\theta} - \theta\rangle = 0.
\end{aligned}$$

For any $\theta \in B_d(\rho)$ such that $n_j^\top \theta_{j,\bullet} = 0$ for all j and $h \in \partial \|\theta\|_{\mathrm{TV},\hat{w}}$, the monotony of the subdifferential mapping implies $\langle \hat{h}, \theta - \hat{\theta}\rangle \le \langle h, \theta - \hat{\theta}\rangle$, $\langle \hat{g}, \theta - \hat{\theta}\rangle \le 0$, and $\langle \hat{f}, \theta - \hat{\theta}\rangle \le 0$, so that

$$\begin{aligned}
&\frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle \\
&\qquad \le \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta} - \theta\rangle.
\end{aligned} \tag{29}$$

Now, consider the function $H_n : \mathbb{R} \to \mathbb{R}$ defined by

$$H_n(t) = \frac{1}{n}\sum_{i=1}^n b(m_{\hat{\theta}+t\eta}(x_i)) - \frac{1}{n}\sum_{i=1}^n b'(m^0(x_i)) m_{\hat{\theta}+t\eta}(x_i),$$

where η will be defined later. We use again the same arguments as in the proof of Lemma 9. We differentiate H_n three times with respect t, so that

$$\begin{aligned}
H_n'(t) &= \frac{1}{n}\sum_{i=1}^n m_\eta(x_i) b'(m_{\hat{\theta}+t\eta}(x_i)) - \frac{1}{n}\sum_{i=1}^n b'(m^0(x_i)) m_\eta(x_i), \\
H_n''(t) &= \frac{1}{n}\sum_{i=1}^n m_\eta^2(x_i) b''(m_{\hat{\theta}+t\eta}(x_i)), \\
\text{and} \quad H_n'''(t) &= \frac{1}{n}\sum_{i=1}^n m_\eta^3(x_i) b'''(m_{\hat{\theta}+t\eta}(x_i)),
\end{aligned}$$

and in the way as in the proof of Lemma 9, we have $|H_n'''(t)| \le C_b(C_n + \rho)|H_n''(t)|$, and Lemma 8 entails

$$H_n''(0)\frac{\psi(-C_b t(C_n+\rho))}{C_b^2(C_n+\rho)^2} \le H_n(t) - H_n(0) - tH_n'(0) \le H_n''(0)\frac{\psi(C_b t(C_n+\rho))}{C_b^2(C_n+\rho)^2},$$

for all $t \geq 0$. Taking $t = 1$ and $\eta = \theta - \hat{\theta}$ implies

$$H_n(1) = \frac{1}{n}\sum_{i=1}^{n} b(m_\theta(x_i)) - \frac{1}{n}\sum_{i=1}^{n} b'(m^0(x_i))m_\theta(x_i) = R(m_\theta),$$
$$\text{and} \quad H_n(0) = \frac{1}{n}\sum_{i=1}^{n} b(m_{\hat{\theta}}(x_i)) - \frac{1}{n}\sum_{i=1}^{n} b'(m^0(x_i))m_{\hat{\theta}}(x_i) = R(m_{\hat{\theta}}).$$

Moreover, we have

$$\begin{aligned} H_n'(0) &= \frac{1}{n}\sum_{i=1}^{n} \langle x_i^B, \theta - \hat{\theta}\rangle b'(m_{\hat{\theta}}(x_i)) - \frac{1}{n}\sum_{i=1}^{n} b'(m^0(x_i))\langle x_i^B, \hat{\theta} - \theta\rangle \\ &= \frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), \boldsymbol{X}^B(\theta - \hat{\theta})\rangle, \\ \text{and} \quad H_n''(0) &= \frac{1}{n}\sum_{i=1}^{n} \langle x_i^B, \hat{\theta} - \theta\rangle^2 b''(m_{\hat{\theta}}(x_i)). \end{aligned}$$

Then, we deduce that

$$\begin{aligned} H_n''(0)\frac{\psi(-C_b(C_n+\rho))}{C_b^2(C_n+\rho)^2} &\leq R(m_\theta) - R(m_{\hat{\theta}}) - \frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), \boldsymbol{X}^B(\theta - \hat{\theta})\rangle \\ &= \phi\mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) - \phi\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \\ &\quad + \frac{1}{n}\langle b'(m_{\hat{\theta}}(\boldsymbol{X})) - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle. \end{aligned}$$

Then, with Equation (29), one has

$$\begin{aligned} &\phi\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) + H_n''(0)\frac{\psi(-C_b(C_n+\rho))}{C_b^2(C_n+\rho)^2} \\ &\quad \leq \phi\mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta} - \theta\rangle. \end{aligned} \tag{30}$$

As $H_n''(0) \geq 0$, it implies that

$$\begin{aligned} \phi\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) &\leq \phi\mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\ &\quad + \frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta} - \theta\rangle. \end{aligned} \tag{31}$$

If $\frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), \boldsymbol{X}^B(\hat{\theta} - \theta)\rangle - \langle h, \hat{\theta} - \theta\rangle < 0$, it follows that

$$\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})),$$

then Theorem 2 holds. From now on, let us assume that

$$\frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle - \langle h, \hat{\theta} - \theta\rangle \geq 0. \tag{32}$$

We first derive a bound on $\frac{1}{n}\langle \boldsymbol{y} - b'(m^0(\boldsymbol{X})), m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\rangle$. Recall that $\mathbf{D}^{-1} = \mathbf{T}$ (see beginning of Section 6.5). We focus on finding out a bound for $\frac{1}{n}\langle (\boldsymbol{X}^B\mathbf{T})^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X}))), \mathbf{D}(\hat{\theta} - \theta)\rangle$. On the one hand, one has

$$\begin{aligned}
&\frac{1}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X}))), \hat{\theta} - \theta\rangle \\
&\qquad = \frac{1}{n}\langle (\boldsymbol{X}^B\mathbf{T})^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X}))), \mathbf{D}(\hat{\theta} - \theta)\rangle \\
&\qquad \leq \frac{1}{n}\sum_{j=1}^{p}\sum_{k=1}^{d_j} |((\boldsymbol{X}^B_{\bullet,j}T_j)_{\bullet,k})^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X})))|\, |(D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}))_k|
\end{aligned}$$

where $(\boldsymbol{X}^B_{\bullet,j}T_j)_{\bullet,k} = [(\boldsymbol{X}^B_{\bullet,j}T_j)_{1,k}\cdots(\boldsymbol{X}^B_{\bullet,j}T_j)_{n,k}]^\top \in \mathbb{R}^n$ is the k-th column of the matrix $\boldsymbol{X}^B_{\bullet,j}T_j$. Let us consider the event

$$\mathcal{E}_n = \bigcap_{j=1}^{p}\bigcap_{k=2}^{d_j}\mathcal{E}_{n,j,k}, \text{ where } \mathcal{E}_{n,j,k} = \Big\{\frac{1}{n}|(\boldsymbol{X}^B_{\bullet,j}T_j)^\top_{\bullet,k}(\boldsymbol{y} - b'(m^0(\boldsymbol{X})))| \leq \hat{w}_{j,k}\Big\},$$

so that, on $\mathcal{E}_n$, we have

$$\begin{aligned}
\frac{1}{n}\langle (\boldsymbol{X}^B)^\top(\boldsymbol{y} - b'(m^0(\boldsymbol{X})), \hat{\theta} - \theta\rangle &\leq \sum_{j=1}^{p}\sum_{k=1}^{d_j}\hat{w}_{j,k}|(D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet}))_k| \\
&\leq \sum_{j=1}^{p}\|\hat{w}_{j,\bullet}\odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})\|_1. \qquad (33)
\end{aligned}$$

On the other hand, from the definition of the subgradient $[h_{j,\bullet}]_{j=1,\dots,p} \in \partial\|\theta\|_{\mathrm{TV},\hat{w}}$ (see Equation (19)), one can choose h such that

$$h_{j,k} = (D_j^\top(\hat{w}_{j,\bullet}\odot \mathrm{sign}(D_j\theta_{j,\bullet})))_k$$

for all $k \in J_j(\theta)$ and

$$h_{j,k} = (D_j^\top(\hat{w}_{j,\bullet}\odot \mathrm{sign}(D_j\hat{\theta}_{j,\bullet}))_k = (D_j^\top(\hat{w}_{j,\bullet}\odot \mathrm{sign}(D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})))_k$$

for all $k \in J_j^\complement(\theta)$. Using a triangle inequality and the fact that $\mathrm{sign}(x)^\top x = \|x\|_1$, we obtain

$$\begin{aligned}
-\langle h, \hat{\theta} - \theta\rangle &\leq \sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j(\theta)}\odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1 \\
&\quad - \sum_{j=1}^{p}\|(\hat{w}_{j,\bullet})_{J_j^\complement(\theta)}\odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^\complement(\theta)}\|_1 \\
&\leq \sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} - \sum_{j=1}^{p}\|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^\complement(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}. \qquad (34)
\end{aligned}$$

Combining inequalities (33) and (34), we get

$$\sum_{j=1}^{p} \|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^{\complement}(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}} \leq 2 \sum_{j=1}^{p} \|(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_{\mathrm{TV},\hat{w}_{j,\bullet}}$$

on $\mathcal{E}_n$. Hence

$$\sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j^{\complement}(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j^{\complement}(\theta)}\|_1 \leq 2 \sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1.$$

This means that

$$\hat{\theta} - \theta \in \mathscr{C}_{\mathrm{TV},\hat{w}}(J(\theta)) \text{ and } \mathbf{D}(\hat{\theta} - \theta) \in \mathscr{C}_{1,\hat{w}}(J(\theta)), \tag{35}$$

see (10) and (27). Now, going back to (31) and taking into account (35), the compatibility of $\boldsymbol{X}^B\mathbf{T}$ given in Equation (26) provides the following on the event $\mathcal{E}_n$:

$$\begin{aligned}\phi \mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq \phi \mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) \\ + 2 \sum_{j=1}^{p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)} \odot D_j(\hat{\theta}_{j,\bullet} - \theta_{j,\bullet})_{J_j(\theta)}\|_1.\end{aligned}$$

Then

$$\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq \mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\theta}(\boldsymbol{X})) + \frac{\|m_{\hat{\theta}}(\boldsymbol{X}) - m_{\theta}(\boldsymbol{X})\|_2}{\sqrt{n}\, \phi\, \kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))}, \tag{36}$$

where $\hat{\gamma} = (\hat{\gamma}_{1,\bullet}^\top, \ldots, \hat{\gamma}_{p,\bullet}^\top)^\top$ is such that

$$\hat{\gamma}_{j,k} = \begin{cases} 2\hat{w}_{j,k} & \text{if } k \in J_j(\theta), \\ 0 & \text{if } k \in J_j^{\complement}(\theta), \end{cases}$$

for all $j = 1, \ldots, p$ and

$$\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta)) = \left\{ 32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\hat{\gamma}_{j,k+1} - \hat{\gamma}_{j,k}|^2 + 2|J_j(\theta)| \|\hat{\gamma}_{j,\bullet}\|_\infty^2 \Delta_{\min,J_j(\theta)}^{-1} \right\}^{-1/2}.$$

Now, we find an upper bound for

$$\frac{1}{\kappa_{\mathbf{T},\hat{\gamma}}^2(J(\theta))} = 32 \sum_{j=1}^{p} \sum_{k=1}^{d_j} |\hat{\gamma}_{j,k+1} - \hat{\gamma}_{j,k}|^2 + 2|J_j(\theta)| \|\hat{\gamma}_{j,\bullet}\|_\infty^2 \Delta_{\min,J_j(\theta)}^{-1}.$$

Note that $\|\hat{\gamma}_{j,\bullet}\|_\infty \leq 2\|\hat{w}_{j,\bullet}\|_\infty$. Let us write $J_j(\theta) = \left\{k_j^1, \ldots, k_j^{|J_j(\theta)|}\right\}$ and set $B_r = [\![k_j^{r-1}, k_j^r[\![= \{k_j^{r-1}, k_j^{r-1} + 1, \ldots, k_j^r - 1\}$ for $r = 1, \ldots, |J_j(\theta)| + 1$ with the convention that $k_j^0 = 0$ and

$k_j^{|J_j(\theta)|+1} = d_j + 1$. Then

$$\begin{aligned}
\sum_{k=1}^{d_j} |\hat{\gamma}_{j,k+1} - \hat{\gamma}_{j,k}|^2 &= \sum_{r=1}^{|J_j(\theta)|+1} \sum_{k \in B_r} |\hat{\gamma}_{j,k+1} - \hat{\gamma}_{j,k}|^2 \\
&= \sum_{r=1}^{|J_j(\theta)|+1} |\hat{\gamma}_{j,k_j^{r-1}+1} - \hat{\gamma}_{j,k_j^{r-1}}|^2 + |\hat{\gamma}_{j,k_j^r} - \hat{\gamma}_{j,k_j^r-1}|^2 \\
&= \sum_{r=1}^{|J_j(\theta)|+1} \hat{\gamma}_{j,k_j^{r-1}}^2 + \hat{\gamma}_{j,k_j^r}^2 \\
&= \sum_{r=1}^{|J_j(\theta)|} 2\, \hat{\gamma}_{j,k_j^r}^2 \\
&\leq 8\, |J_j(\theta)|\, \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2.
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{1}{\kappa_{\mathbf{T},\hat{\gamma}}^2(J(\theta))} &\leq 512 \sum_{j=1}^{p} \Big(|J_j(\theta)|\, \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 + |J_j(\theta)|\, \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \Delta_{\min,J_j(\theta)}^{-1} \Big) \\
&\leq 512 \sum_{j=1}^{p} \Big(1 + \frac{1}{\Delta_{\min,J_j(\theta)}} \Big) |J_j(\theta)| \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2 \\
&\leq 512 |J(\theta)| \max_{j=1,\dots,p} \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|_\infty^2.
\end{aligned} \tag{37}$$

Now, we use the connection between the empirical norm and Kullback-Leibler divergence. Indeed, using Lemma 9, we get

$$\begin{aligned}
&\frac{\|m_{\hat{\theta}}(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2}{\sqrt{n}\phi\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))} \\
&\quad\leq \frac{1}{\sqrt{\phi}\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))} \Big(\frac{1}{\sqrt{n}} \|m_{\hat{\theta}}(\boldsymbol{X}) - m^0(\boldsymbol{X})\|_2 + \frac{1}{\sqrt{n}} \|m^0(\boldsymbol{X}) - m_\theta(\boldsymbol{X})\|_2 \Big) \\
&\quad\leq \frac{2}{\sqrt{\phi}\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))\sqrt{C_n(\rho, L_n)}} \Big(\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X}))^{1/2} \\
&\qquad\qquad + \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X}))^{1/2} \Big),
\end{aligned}$$

where we defined $C_n(\rho, L_n) = \frac{L_n \psi(-C_b(C_n+\rho))}{C_b^2 \phi(C_n+\rho)^2}$, so that combined with Equation (36), we obtain

$$\begin{aligned}
\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) &\leq \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) \\
&\quad + \frac{2}{\sqrt{\phi}\kappa_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa(J(\theta))\sqrt{C_n(\rho, L_n)}} \big(\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X}))^{1/2} \\
&\qquad\qquad + \mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X}))^{1/2} \big).
\end{aligned}$$

This inequality entails the following upper bound

$$\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq 3\mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{5}{\phi\kappa^2_{\mathbf{T},\hat{\gamma}}(J(\theta))\kappa^2(J(\theta))C_n(\rho, L_n)},$$

since whenever we have $x \leq c + b\sqrt{x}$ for some $x, b, c > 0$, then $x \leq 2c + b^2$. Introducing $g(x) = x^2/\psi(-x) = x^2/(e^{-x} + 1 - x)$, we note that

$$\frac{1}{C_n(\rho, L_n)} = \frac{\phi}{L_n} g(C_b(C_n + \rho)) \leq \frac{\phi}{L_n}(C_b(C_n + \rho) + 2),$$

since $g(x) \leq x + 2$ for any $x > 0$. Finally, by using also (37), we end up with

$$\mathrm{KL}_n(m^0(\boldsymbol{X}), m_{\hat{\theta}}(\boldsymbol{X})) \leq 3\mathrm{KL}_n(m^0(\boldsymbol{X}), m_\theta(\boldsymbol{X})) + \frac{2560(C_b(C_n + \rho) + 2)}{L_n\kappa^2(J(\theta))} \, |J(\theta)| \, \|(\hat{w}_{j,\bullet})_{J_j(\theta)}\|^2_\infty,$$

which is the statement provided in Theorem 2. The only thing remaining is to control the probability of the event $\mathcal{E}_n^\complement$. This is given by the following:

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_n^\complement] &\leq \sum_{j=1}^{p}\sum_{k=2}^{d_j} \mathbb{P}\Big[\frac{1}{n}|(\boldsymbol{X}^B_{\bullet,j}T_j)^\top_{\bullet,k}(\boldsymbol{y} - b'(m^0(\boldsymbol{X})))| \geq \hat{w}_{j,k}\Big] \\
&\leq \sum_{j=1}^{p}\sum_{k=2}^{d_j} \mathbb{P}\Big[\sum_{i=1}^{n} |(\boldsymbol{X}^B_{\bullet,j}T_j)_{i,k}(y_i - b'(m^0(x_i)))| \geq n\hat{w}_{j,k}\Big].
\end{aligned}$$

Let $\xi_{i,j,k} = (\boldsymbol{X}^B_{\bullet,j}T_j)_{i,k}$ and $Z_i = y_i - b'(m^0(x_i))$. Note that conditionally on x_i, the random variables (Z_i) are independent. It can be easily shown (see Theorem 5.10 in Lehmann and Casella (1998)) that the moment generating function of Z (copy of Z_i) is given by

$$\mathbb{E}[\exp(tZ)] = \exp(\phi^{-1}\{b(m^0(x) + t) - tb'(m^0(x) - b(m^0(x)))\}). \tag{38}$$

Applying Lemma 6.1 in Rigollet (2012), using (38) and Assumption 1, we can derive the following Chernoff-type bounds

$$\mathbb{P}\Big[\sum_{i=1}^{n} |\xi_{i,j,k}Z_i| \geq n\hat{w}_{j,k}\Big] \leq 2\exp\Big(-\frac{n^2\hat{w}^2_{j,k}}{2U_n\phi\|\xi_{\bullet,j,k}\|_2^2}\Big), \tag{39}$$

where $\xi_{\bullet,j,k} = [\xi_{1,j,k} \cdots \xi_{n,j,k}]^\top \in \mathbb{R}^n$. We have

$$\boldsymbol{X}^B_{\bullet,j}T_j = \begin{bmatrix}
1 & \sum_{k=2}^{d_j} x^B_{1,j,k} & \sum_{k=3}^{d_j} x^B_{1,j,k} & \cdots & \sum_{k=d_{j-1}}^{d_j} x^B_{1,j,k} & x^B_{1,j,d_j} \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & \sum_{k=2}^{d_j} x^B_{n,j,k} & \sum_{k=3}^{d_j} x^B_{n,j,k} & \cdots & \sum_{k=d_{j-1}}^{d_j} x^B_{n,j,k} & x^B_{n,j,d_j}
\end{bmatrix},$$

therefore

$$\|\xi_{\bullet,j,k}\|_2^2 = \sum_{i=1}^{n}(\boldsymbol{X}^B_{\bullet,j}T_j)^2_{\bullet,k} = \#\Big(\Big\{i : x_{i,j} \in \bigcup_{r=k}^{d_j} I_{j,r}\Big\}\Big) = n\hat{\pi}_{j,k}. \tag{40}$$

So, using the weights $\hat{w}_{j,k}$ given by (12) together with (39) and (40), we obtain that the probability of $\mathcal{E}_n^\complement$ is smaller than $2e^{-A}$. This concludes the proof of the first part of Theorem 2. □

6.8. Proof of Theorem 3

First, let us note that in the least squares setting, we have $R(m_\theta) - R(m^0) = \|m_\theta - m^0\|_n^2$ for any $\theta \in \mathbb{R}^d$ where $\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g(x_i)^2$, and that $b(y) = \frac{1}{2}y^2$, $\phi = \sigma^2$ (noise variance) in Equation (2), and $L_n = U_n = 1$, $C_b = 0$. Theorem 2 provides

$$\|m_{\hat{\theta}} - m^0\|_n^2 \leq 3\|m_\theta - m_{\theta^0}\|_n^2 + \frac{5120\sigma^2}{\kappa^2(J(\theta))}\frac{|J(\theta)|(A + \log d)}{n}$$

for any $\theta \in \mathbb{R}^d$ such that $n_j^\top \theta_{j\bullet} = 0$ and $J(\theta) \leq J_*$. Since $d_j = D$ for all $j = 1, \ldots, p$, we have $d = Dp$ and

$$|J(\theta)| = \sum_{j=1}^p |\{k = 2, \ldots, D : \theta_{j,k} \neq \theta_{j,k-1}\}| \leq (D-1)|\mathcal{J}(\theta)|\|\theta\|_\infty \leq Dp\|\theta\|_\infty \tag{41}$$

for any $\theta \in \mathbb{R}^d$, where we recall that $\mathcal{J}(\theta) = \{j = 1, \ldots, p : \theta_{j,\bullet} \neq \mathbf{0}_D\}$. Also, recall that $I_{j,1} = I_1 = [0, \frac{1}{D}]$ and $I_{j,k} = I_k = (\frac{k-1}{D}, \frac{k}{D}]$ for $k = 2, \ldots, D$ and $j = 1, \ldots, p$. Also, we consider $\theta = \theta^*$, where $\theta^*_{j,\bullet}$ is defined, for any $j \in \mathcal{J}_*$, as the minimizer of

$$\sum_{i=1}^n \Big(\sum_{k=1}^D (\theta_{j,k} - m_j^0(x_{i,j}))\mathbf{1}_{I_k}(x_{i,j}) \Big)^2$$

over the set of vectors $\theta_{j,\bullet} \in \mathbb{R}^D$ satisfying $n_j^\top \theta_{j,\bullet} = 0$, and we put $\theta^*_{j,\bullet} = \mathbf{0}_D$ for $j \notin \mathcal{J}_*$. It is easy to see that the solution is given by

$$\theta^*_{j,k} = \frac{\sum_{i=1}^n m_j^0(x_{i,j})\mathbf{1}_{I_k}(x_{i,j})}{n_{j,k}},$$

where we recall that $n_{j,k} = \sum_{i=1}^n \mathbf{1}_{I_k}(x_{i,j})$. Note in particular that the identifiability assumption $\sum_{i=1}^n m_j^0(x_{i,j}) = 0$ entails that $n_j^\top \theta^*_{j,\bullet} = 0$. In order to control the bias term, an easy computation gives that, whenever $x_{i,j} \in I_k$

$$|\theta^*_{j,k} - m_j^0(x_{i,j})| \leq \frac{\sum_{i'=1}^n |m_j^0(x_{i',j}) - m_j^0(x_{i,j})|\mathbf{1}_{I_k}(x_{i',j})}{n_{j,k}} \leq L|I_k| = \frac{L}{D},$$

where we used the fact that m_j^0 is L-Lipschitz, so that

$$\begin{aligned}
\|m_{\theta^*} - m^0\|_n^2 &= \frac{1}{n}\sum_{i=1}^n (m_{\theta^*}(x_{i,j}) - m^0(x_{i,j}))^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big(\sum_{j\in\mathcal{J}_*} \sum_{k=1}^D (\theta^*_{j,k} - m^0(x_{i,j}))\mathbf{1}_{I_k} \Big)^2 \\
&\leq \frac{|\mathcal{J}_*|}{n}\sum_{i=1}^n \Big(\sum_{j\in\mathcal{J}_*} \sum_{k=1}^D (\theta^*_{j,k} - m^0(x_{i,j}))\mathbf{1}_{I_k} \Big)^2 \\
&\leq \frac{|\mathcal{J}_*|}{n}\sum_{i=1}^n \sum_{j\in\mathcal{J}_*} \sum_{k=1}^D (\theta^*_{j,k} - m^0(x_{i,j}))^2 \mathbf{1}_{I_k}(x_{i,j}) \\
&\leq |\mathcal{J}_*| \sum_{j\in\mathcal{J}_*} \sum_{k=1}^D L^2 |I_k|^2 \mathbf{1}_{I_k}(x_{i,j}) \leq \frac{L^2|\mathcal{J}_*|^2}{D^2}.
\end{aligned}$$

Note that $|\theta^*_{j,k}| \leq \|m_j^0\|_{n,\infty}$ where $\|m_j^0\|_{n,\infty} = \max_{i=1,\dots,n} |m_j^0(x_{i,j})|$. This entails that $\|\theta^*\|_\infty \leq \max_{j=1,\dots,p} \|m_j^0\|_{n,\infty} = M_n$. So, using also (41), we end up with

$$\|m_{\hat{\theta}} - m^0\|_n^2 \leq \frac{3L^2|\mathcal{J}_*|^2}{D^2} + \frac{5120\sigma^2}{\kappa^2(J(\theta^*))}\frac{D\mathcal{J}_* M_n(A + \log(DpM_n))}{n},$$

which concludes the proof Theorem 3 using $D = n^{1/3}$. □

References

A. Agresti. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015.

M. Z. Alaya, S. Gaïffas, and A. Guilloux. Learning the intensity of time events with change-points. *Information Theory, IEEE Transactions on*, 61(9):5148–5171, 2015.

F. Bach. Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414, 2010.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

E. Bacry, M Bompaire, P. Deegan, S. Gaïffas, and S. V. Poulsen. tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *Journal of Machine Learning Research*, 18(214):1–5, 2018. URL `http://jmlr.org/papers/v18/17-381.html`.

P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5, 2014.

P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76(5):1–7, Apr 2016.

H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. ISSN 0090-5364.

J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. ISBN 0-521-83378-7.

L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

P. Bühlmann and S. van De Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, 1:169–194, 2007.

E. J. Candès and M. B. Wakin. An Introduction To Compressive Sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

B. Chlebus and S. H. Nguyen. On finding optimal discretizations for two attributes. In Lech Polkowski and Andrzej Skowron, editors, *Rough Sets and Current Trends in Computing*, volume 1424 of *Lecture Notes in Computer Science*, pages 537–544. Springer Berlin Heidelberg, 1998.

L. Condat. A Direct Algorithm for 1D Total Variation Denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, 2013.

A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

D. L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. In *PROC. NATL ACAD. SCI. USA 100 2197202*, 2002.

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845–2862, 2001.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4): 367–378, 2002.

S. Garcia, J. Luengo, J. A. Saez, V. Lopez, and F. Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2013.

P. J. Green and B. W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall, London, 1994.

T. Hastie and R. Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

J. Horowitz, J. Klemelä, and E. Mammen. Optimal estimation in additive regression models. *Bernoulli*, 12(2):271–298, 2006.

S. Ivanoff, F. Picard, and V. Rivoirard. Adaptive lasso and group-lasso for functional poisson regression. *The Journal of Machine Learning Research*, 17(1):1903–1948, 2016.

K. Knight and W. Fu. Asymptotics for Lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378, 2000.

R. Kohavi. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer texts in statistics. Springer, New York, 1998.

M. Lichman. UCI Machine Learning Repository, 2013.

H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: an enabling technique. *Data Min. Knowl. Discov.*, 6(4):393–423, 2002.

G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Annals of Statistics*, pages 30–55, 2004.

L. Meier, S. van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

L. Meier, S. Van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. R. Quinlan. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, 1 edition, 1993.

F. Rapaport, E. Barillot, and J. P. Vert. Classification of arraycgh data using fused SVM. *Bioinformatics*, 24(13):i375–i382, 2008.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

P. Rigollet. Kullback Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 2012.

M. A. Russell. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, 2013.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266, 1989.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1): 267–288, 1996a.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996b.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

S. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2): 614–645, 2008.

S. van de Geer and J. Lederer. *The Lasso, correlated design, and improved oracle inequalities*, volume Volume 9 of *Collections*, pages 303–316. Institute of Mathematical Statistics, 2013.

J. Wu and S. Coggeshall. *Foundations of Predictive Analytics (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 1st edition, 2012.

I. C. Yeh and C. H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

Y. L. Yu. On decomposing the proximal map. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 91–99. 2013.

P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.