Binarity: a penalization for one-hot encoded features

Mokhtar Z. Alaya
Theoretical and applied statistics laboratory
Pierre and Marie Curie University
Paris, France

Simon Bussy
Theoretical and applied statistics laboratory
Pierre and Marie Curie University
Paris, France

Stéphane Gaïffas
Centre de Mathématiques Appliquées, École Polytechnique
CNRS UMR 7641
91128 Palaiseau, France

Agathe Guilloux
LaMME, UEVE and UMR 8071
Université Paris Saclay
Evry, France

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 binarity. 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 to avoid collinearity within groups. Non-asymptotic oracle inequalities for generalized linear models are proposed, and 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 $\ell_1$ penalization.

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

1. Introduction
In many applications, datasets used for 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 leads to a non-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 dimension of the problem.
A new penalty. 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.

Sparsity. To address the high-dimensionality of features, sparse 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 \( \ell_1 \)-penalization (Lasso) term to the goodness-of-fit (Tibshirani, 1996). This typically leads to sparse parametrization of models, with a level of sparsity that depends on the strength of the penalization. Statistical properties of \( \ell_1 \)-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 features ordering. 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 \( \ell_1 \)-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.
Organization of the paper. 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 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. The proposed methodology is described in Section 2. Section 3 establishes an oracle inequality for generalized linear models. 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 $\ell_q$-quasi norm of a vector $v \in \mathbb{R}^m$, namely $\|v\|_q = (\sum_{k=1}^m |v_k|^q)^{1/q}$, and $\|v\|_\infty = \max_{k=1,...,m} |v_k|$. 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 $u \odot v = (u_1v_1, \ldots, u_mv_m)^\top$. For any $u \in \mathbb{R}^m$ and any $L \subseteq \{1, \ldots, m\}$, we denote by $u_L$ the vector in $\mathbb{R}^m$ satisfying $(u_L)_k = u_k$ for $k \in L$ and $(u_L)_k = 0$ for $k \in \mathbb{R}^m \setminus \{1, \ldots, m\} \setminus L$. We write $1_m$ (resp. $0_m$) for the vector of $\mathbb{R}^m$ having all coordinates equal to one (resp. zero). Finally, we denote by $\text{sign}(x)$ the sub-differential of the function $x \mapsto |x|$, namely $\text{sign}(x) = \{1\}$ if $x > 0$, $\text{sign}(x) = \{-1\}$ if $x < 0$ and $\text{sign}(0) = [-1, 1]$.

2. The proposed method

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

Binarization. The binarized matrix $X_B$ is a matrix with an extended number $d > p$ of columns, where the $j$-th column $X_{i,j}$ is replaced by $d_j \geq 2$ columns $X_{i,j,1}, \ldots, X_{i,j,d_j}$ containing only zeros and ones. Its $i$-th row is written

$$x_i^B = (x_{i,1,1}^B, \ldots, x_{i,1,d_1}^B, x_{i,2,1}^B, \ldots, x_{i,2,d_2}^B, \ldots, x_{i,p,1}^B, \ldots, x_{i,p,d_p}^B)^\top \in \mathbb{R}^d.$$

In order to simplify presentation of our results, we assume in the paper that all raw features $X_{i,j}$ are continuous, so that they are transformed using the following one-hot encoding. We consider a partition of intervals $I_{j,1}, \ldots, I_{j,d_j}$ of the range of the coordinates of $X_{i,j}$ and define

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

for $i = 1, \ldots, n$ and $k = 1, \ldots, d_j$. A natural choice of intervals is given by quantiles, namely $I_{j,1} = [q_j(0), q_j(1/d_j)]$ and $I_{j,k} = (q_j((k-1)/d_j), q_j(k/d_j))$ for $k = 2, \ldots, d_j$, where $q_j(\alpha)$ denotes a quantile of order $\alpha \in [0, 1]$ for $X_{i,j}$. 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 construct a non-linear decision with respect to the raw features space. If training data contains also unordered qualitative features, one-hot encoding with $\ell_1$-penalization can be used for instance.

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

$$R_m(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, m_\theta(x_i)), \quad (1)$$
where \( m_\theta(x_i) = \theta^T x_i^B \) and \( \theta \in \mathbb{R}^d \) with \( d = \sum_{j=1}^p d_j \). We then have \( \theta = (\theta_{1, \bullet}, \ldots, \theta_{p, \bullet})^T \), 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 \( Y|X = x \) has a density

\[
y|x \mapsto f^0(y|x) = \exp \left( \frac{y m^0(x) - b(m^0(x)) + c(y)}{\phi} \right),
\]

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

\[
\ell(y, y^\prime) = -yy^\prime + b(y^\prime).
\]

The density described in (2) encompasses several distributions, see Table 1. Only \( m^0(\cdot) \) is unknown in the density (2), the parameter \( \phi \) is a scale parameter, and it is assumed that \( b(\cdot) \) is three times continuously differentiable. It is standard to notice that

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

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, and that there exist constants \( C_n > 0 \), and \( 0 < L_n \leq U_n \) such that \( C_n = \max_{i=1,\ldots,n} |m^0(x_i)| < \infty \) and \( L_n \leq \max_{i=1,\ldots,n} b''(m^0(x_i)) \leq U_n \).

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); Rigollet (2012).

<table>
<thead>
<tr>
<th></th>
<th>( \phi )</th>
<th>( b(z) )</th>
<th>( b'(z) )</th>
<th>( b''(z) )</th>
<th>( L_n )</th>
<th>( U_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( \sigma^2 )</td>
<td>( \frac{z^2}{2} )</td>
<td>( z )</td>
<td>1</td>
<td>1</td>
<td>( C_n )</td>
</tr>
<tr>
<td>Logistic</td>
<td>1</td>
<td>( \log(1 + e^z) )</td>
<td>( e^z )</td>
<td>1</td>
<td>( C_n )</td>
<td>( e^{-C_n} )</td>
</tr>
<tr>
<td>Poisson</td>
<td>1</td>
<td>( e^z )</td>
<td>( e^z )</td>
<td>( e^z )</td>
<td>( e^{-C_n} )</td>
<td>( e^{C_n} )</td>
</tr>
</tbody>
</table>

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

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

(P1) The one-hot-encodings satisfy \( \sum_{k=1}^{d_j} X_{ij,k}^B = 1 \) for \( j = 1, \ldots, p \), meaning that the columns of each block sum to \( 1_n \), making \( X^B \) 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.
(P3) Some of the raw features $X_{\bullet,j}$ 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 $\theta$.

A usual way to deal with (P1) is to impose a linear constraint (Agresti, 2015) in each block. In our penalization term, we impose

$$
\sum_{k=1}^{d_j} \theta_{j,k} = 0
$$

(3)

for all $j = 1, \ldots, p$. Now, the trick to tackle (P2) is to remark that within each block, binary features are ordered. We use a within block total-variation penalization

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

where

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

(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

$$
bina(\theta) = \sum_{j=1}^{p} \left( \sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}| + \delta_1(\theta_{j,\bullet}) \right)
$$

(5)

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

$$
\delta_1(u) = \begin{cases} 0 & \text{if } 1^\top u = 0, \\ \infty & \text{otherwise}. \end{cases}
$$

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

$$
\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ R_n(\theta) + bina(\theta) \right\}.
$$

(6)

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.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Account Length</th>
<th>VMail Message</th>
<th>Day Mins</th>
<th>Day Calls</th>
<th>Day Charge</th>
<th>Eve Mins</th>
<th>Eve Calls</th>
<th>Eve Charge</th>
<th>Night Mins</th>
<th>Night Calls</th>
<th>Night Charge</th>
<th>Intl Mins</th>
<th>Intl Calls</th>
<th>Intl Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−0.2</td>
<td>−0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>−1.0</td>
<td>−0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>−0.4</td>
<td>−0.2</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td>1.2</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>−0.2</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1: Illustration of the binarsity penalty 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 feature. 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).

**Proximal operator of binarity.** The proximal operator and proximal algorithms are important tools for non-smooth convex optimization, with important applications in the field of supervised
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 binarity. 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 non-linear decision boundaries are learned on the two other examples, without apparent overfitting.

learning with structured sparsity (Bach et al., 2012). The proximal operator of a proper lower semi-continuous (Bauschke and Combettes, 2011) convex function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined by

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

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

$$g(u) = \delta_C(u) = \begin{cases} 
0 & \text{if } u \in C, \\
\infty & \text{otherwise},
\end{cases}$$
Algorithm 1: Proximal operator of \( \text{bina}(\theta) \), see (5)

**Input:** vector \( \theta \in \mathbb{R}^d \) and weights \( \hat{w}_{j,k} \) for \( j = 1, \ldots, p \) and \( k = 1, \ldots, d_j \)

**Output:** vector \( \eta = \text{prox}_{\text{bina}}(\theta) \)

**for** \( j = 1 \) **to** \( p \) **do**
- \( \beta_{j,\bullet} \leftarrow \text{prox}_{\|\theta_{j,\bullet}\|_{TV}}(\theta_{j,\bullet}) \) (TV-weighted prox in block \( j \), see (4))
- \( \eta_{j,\bullet} \leftarrow \beta_{j,\bullet} - \frac{1}{d_j} \sum_{k=1}^{d_j} \beta_{j,k} \) (within-block centering)

**Return:** \( \eta \)

then \( \text{prox}_g \) is the projection operator onto \( C \). It turns out that the proximal operator of binarity 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 binarity penalty is block separable, followed by a centering within each block to satisfy the sum-to-zero constraint, see Algorithm 1 below. We refer to Algorithm 2 in Section B for the weighted total-variation proximal operator.

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

A proof of Proposition 1 is given in Section A. 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.

3. **Theoretical guarantees**

We now investigate the statistical properties of (6) where the weights in the binarity penalty have the form

\[
\hat{w}_{j,k} = O\left(\sqrt{\frac{\log d}{n}} \hat{\pi}_{j,k}\right),
\]

with

\[
\hat{\pi}_{j,k} = \left| \left\{ i = 1, \ldots, n : x_{i,j} \in \left( q_j \left( \frac{k}{d_j} \right), q_j(1) \right) \right\} \right| \frac{1}{n}
\]

for all \( k \in \{2, \ldots, d_j\} \), 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 \( X_{\bullet,j}^B \). We consider the risk measure defined by

\[
R(m_\theta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ -b'(m_0(x_i))m_\theta(x_i) + b(m_\theta(x_i)) \right\},
\]

which is standard with generalized linear models (Van de Geer, 2008). We aim at evaluating how “close” to the minimal possible expected risk our estimated function \( m_\hat{\theta} \) with \( \hat{\theta} \) given by (6) 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 need to impose a restricted eigenvalue assumption on \( X^B \).
For all $\theta \in \mathbb{R}^d$, let $J(\theta) = [J_1(\theta), \ldots, J_p(\theta)]$ be the concatenation of the support sets relative to the total-variation penalization, that is

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

Similarly, we denote $J^c_i(\theta) = [J^c_1(\theta), \ldots, J^c_p(\theta)]$ the complementary of $J(\theta)$. The restricted eigenvalue condition is defined as follow.

**Assumption 2** Let $K = [K_1, \ldots, K_p]$ be a concatenation of index sets. We consider

$$\kappa(K) \in \inf_{u \in C_{TV,\hat{w}}(K) \setminus \{0_d\}} \left\{ \frac{\| X^T u \|_2}{\sqrt{n} \| u_K \|_2} \right\}$$

with

$$C_{TV,\hat{w}}(K) = \left\{ u \in \mathbb{R}^d : \sum_{j=1}^p \| (u_{j,\ast})_{K_j} \|_{TV,\hat{w}_{j,\ast}} \leq 2 \sum_{j=1}^p \| (u_{j,\ast})_{K_j} \|_{TV,\hat{w}_{j,\ast}} \right\}.$$ 

We suppose that the following condition holds

$$\kappa(K) > 0. \quad (8)$$

The set $C_{TV,\hat{w}}(K)$ is a cone composed by all vectors with similar support $K$. Let us now work locally on

$$B_d(\rho) = \{ \theta \in \mathbb{R}^d : \| \theta \|_2 \leq \rho \},$$

the $\ell_2$-ball of radius $\rho > 0$ in $\mathbb{R}^d$. This restriction has already been considered in the case of high-dimensional generalized linear models (Van de Geer, 2008). It allows us to establish a connection, via the notion of self-concordance (Bach, 2010), between the empirical squared $\ell_2$-norm and the empirical Kullback divergence (see Lemma 9 in Section C). 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

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

Let $C_n(\rho, p) = 2(C_n + \rho \sqrt{p})$, $\psi(u) = e^u - u - 1$, and consider the following constants

$$C_n(\rho, p, L_n) = \frac{L_n \psi(-C_n(\rho, p))}{C_n^2(\rho, p)} , \quad \epsilon > \frac{2}{C_n(\rho, p, L_n)} \quad \text{and} \quad \zeta = \frac{4}{\epsilon C_n(\rho, p, L_n) - 2}.$$ 

Then, with probability at least $1 - 2e^{-A}$, any solution $\hat{\theta}$ of problem (6) restricted on $B_d(\rho)$ fulfills the following risk bound

$$R(m_{\hat{\theta}}) - R(m^0) \leq (1 + \zeta) \inf_{\theta \in B_d(\rho)} \left\{ R(m_{\theta}) - R(m^0) + \frac{\xi |J(\theta)|}{\kappa(\theta)} + \max_{j=1,\ldots,p} \| (\hat{w}_{j,\ast})_{J(\theta)} \|_{\infty}^2 \right\}, \quad (10)$$

where

$$\xi = \frac{512 \epsilon^2 C_n(\rho, p, L_n)}{\epsilon C_n(\rho, p, L_n) - 2}.$$
A proof of Theorem 2 is given in Section C. The second term in the right-hand side of (10) can be viewed as a variance term, and its dominant term satisfies

$$\frac{|J(\theta)|}{\kappa^2(J(\theta))} \max_{j=1,...,p} \| (\hat{w}_j, \cdot) J_j(\theta) \|_2^2 \leq \frac{A' U_n \phi}{\kappa^2(J(\theta))} \frac{|J(\theta)| \log d}{n},$$

(11)

for some positive constant $A'$. The complexity term in (11) depends on both the sparsity and the restricted eigenvalues of the binarized matrix. The value $|J(\theta)|$ characterizes the sparsity of the vector $\theta$, that is the smaller $|J(\theta)|$, the sparser $\theta$. The rate of convergence of the estimator $m_\hat{\theta}$ has the expected shape $\log d/n$. Moreover, for the case of least squares regression, the oracle inequality in Theorem 2 is sharp, in the sense that $\zeta = 0$ (see Remark 10 in Section C).

4. Numerical experiments

In this section, we first illustrate the fact that the binarsity penalization is roughly only two times slower than basic $\ell_1$-penalization, see the timings in Figure 3. We then apply our method on 9 classical binary classification datasets obtained from the UCI Machine Learning Repository (Lichman, 2013).

Fig. 3: Average computing time in second (with the black lines representing ± the standard deviation) obtained on 100 simulated datasets for training a logistic model with binarsity VS Lasso penalization, both trained on $X^B$ 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 tends to stay roughly constant and equal to 2.

Let us consider the datasets presented in Table 2. We use our method with a logistic regression model, and compare it with a Lasso logistic regression model (ridge logistic regression, as well as logistic regression without regularization, gives similar or lower scores for all considered datasets), support vector machine (SVM) (Schölkopf and Smola, 2002) with radial basis function kernel, and finally random forest (Breiman, 2001) and gradient boosting (Friedman, 2002) algorithms, denoted
Fig. 4: Performance comparison between the five methods on the considered datasets in terms of ROC curve (AUC score is also given in parenthesis) calculated from predictions made on test sets. The 4 last datasets contain too many examples for the SVM with RBF kernel to be trained in a reasonable time. Binarsity consistently does a better job than the Lasso, and outperforms it in some cases. Its performance is comparable in all the considered examples to random forests and gradient boosting, with a computational time which is orders of magnitude faster, see Figure 5.
Fig. 5: 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 $\ell_1$ 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, thus achieving a nice compromise between computational time and performance.
Tab. 2: Basic informations about the 9 datasets considered: the number of examples \( n \), the initial dimension \( p \) and the corresponding reference.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( n )</th>
<th>( p )</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>Sigillito et al. (1989)</td>
</tr>
<tr>
<td>Churn</td>
<td>3333</td>
<td>21</td>
<td>Lichman (2013)</td>
</tr>
<tr>
<td>Default of credit card</td>
<td>30000</td>
<td>24</td>
<td>Yeh and Lien (2009)</td>
</tr>
<tr>
<td>Adult</td>
<td>32561</td>
<td>14</td>
<td>Kohavi (1996)</td>
</tr>
<tr>
<td>Bank marketing</td>
<td>45211</td>
<td>17</td>
<td>Moro et al. (2014)</td>
</tr>
<tr>
<td>Covertype</td>
<td>550088</td>
<td>10</td>
<td>Blackard and Dean (1999)</td>
</tr>
<tr>
<td>SUSY</td>
<td>5000000</td>
<td>18</td>
<td>Baldi et al. (2014)</td>
</tr>
<tr>
<td>HEPMASS</td>
<td>10500000</td>
<td>28</td>
<td>Baldi et al. (2016)</td>
</tr>
<tr>
<td>HIGGS</td>
<td>11000000</td>
<td>24</td>
<td>Baldi et al. (2014)</td>
</tr>
</tbody>
</table>

RF and GB respectively. 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 a \( K \)-fold cross-validation procedure with \( K = 10 \).

Note that we do not need to tune the \( d_j \) parameters (number of bins for the one-hot encoding of raw feature \( j \)). Indeed, considering the \( j \)-th block and assuming that \( d_j \) is “high enough”, then increasing \( d_j \) barely changes the results since the number of cut-points does not change and their positions barely move: only the strength of the penalty changes. We then fix \( d_j = 50 \) for all \( j \in 1, \ldots, p \) but in practice, blocks with different sizes appear since we only keep distinct cut-points (estimated quantiles often lead to ties with probabilities increments of 1/50), see Figure 1 for instance.

All the methodology discussed in this paper is implemented in Python/C++ and will be available shortly on a GitHub repository in the form of annotated programs, together with notebook tutorials. For random forests and gradient boosting, we use the reference implementations from the scikit-learn library (Pedregosa et al., 2011).

5. Conclusion

In this paper, the binarsity penalty has been introduced and applied to one-hot encodings of continuous features. We evidenced the good statistical properties of the binarsity estimation procedure by proving non-asymptotic oracle inequalities for general linear models, and conducted extensive comparison to state-of-the-art prediction algorithms on several standard datasets. Experimental results illustrate that our method significantly outperforms Lasso, while being competitive with random forests and boosting. Moreover, it runs orders of magnitude faster than boosting and 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 interpretation than, for instance, the one provided by the Lasso on raw features, by giving insights on the relevant thresholds for each feature.
Appendix A. Proof of Proposition 1: proximal operator of binarity

This section corresponds to the proof of Proposition 1 that is for any fixed $j = 1, \ldots, p$, we aim to prove that $\text{prox}_{\| \cdot \|_{TV, \hat{w}_j, \cdot} + \delta_1}$ is the composite proximal operators of $\text{prox}_{\| \cdot \|_{TV, \hat{w}}} \text{ and } \text{prox}_{\delta_1}$, namely

$$\text{prox}_{\| \cdot \|_{TV, \hat{w}_j, \cdot} + \delta_1}(\theta_{j, \cdot}) = \text{prox}_{\| \cdot \|_{TV, \hat{w}_j, \cdot}}\left(\text{prox}_{\| \cdot \|_{TV, \hat{w}}}(\theta_{j, \cdot})\right)$$

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

$$\partial \left( \| \theta_{j, \cdot} \|_{TV, \hat{w}_j, \cdot} \right) \subseteq \partial \left( \| \text{prox}_{\delta_1}(\theta_{j, \cdot}) \|_{TV, \hat{w}_j, \cdot} \right). \quad (12)$$

By the definition of the proximal operator, we have

$$\text{prox}_{\delta_1}(\theta_{j, \cdot}) = \Pi_{1_{d_j}}(\theta_{j, \cdot}),$$

where $\Pi_{1_{d_j}}(\cdot)$ stands for the projection onto the vector $1_{d_j}$. Besides, we know that

$$\Pi_{1_{d_j}}(\theta_{j, \cdot}) = \langle \theta_{j, \cdot}, 1_{d_j} \rangle / \| 1_{d_j} \|_2^2 1_{d_j} = (1_{d_j} \sum_{k=1}^{d_j} \theta_{j,k}) 1_{d_j} =: \bar{\theta}_{j, \cdot} 1_{d_j}. \quad (13)$$

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

$$\| \theta_{j, \cdot} \|_{TV, \hat{w}_j, \cdot} = \sum_{k=2}^{d_j} \hat{w}_{j,k} |\theta_{j,k} - \theta_{j,k-1}| = \| \hat{w}_{j, \cdot} \odot D_j \theta_{j, \cdot} \|_1. \quad (14)$$

By using subdifferential calculus (see details in the proof of Proposition 4 below), one has

$$\partial \left( \| \theta_{j, \cdot} \|_{TV, \hat{w}_j, \cdot} \right) = \partial \left( \| \hat{w}_{j, \cdot} \odot D_j \theta_{j, \cdot} \|_1 \right) = D_j^\top \hat{w}_{j, \cdot} \odot \text{sign}(D_j \theta_{j, \cdot}).$$

Then, one has

$$D_j^\top \hat{w}_{j, \cdot} \odot \text{sign}(D_j \theta_{j, \cdot}) = D_j^\top \hat{w}_{j, \cdot} \odot \text{sign}(D_j (\theta_{j, \cdot} - \bar{\theta}_{j, \cdot} 1_{d_j})),$$

that entails (12). Hence, setting $\beta_{j, \cdot} = \text{prox}_{\| \cdot \|_{TV, \hat{w}_j, \cdot}}(\theta_{j, \cdot})$ and $\bar{\beta}_{j, \cdot} = \frac{1}{d_j} \sum_{k=1}^{d_j} \beta_{j,k}$ we get

$$\text{prox}_{\| \cdot \|_{TV, \hat{w}_j, \cdot} + \delta_1}(\theta_{j, \cdot}) = \beta_{j, \cdot} - \bar{\beta}_{j, \cdot} 1_{d_j}$$

which gives Algorithm 1.
Appendix B. Algorithm of computing proximal operator of weighted TV penalization

We recall here the algorithm given in Alaya et al. (2015) for computing the proximal operator of weighted total-variation penalization. The latter is defined as follows

\[ \beta = \text{prox}_{\parallel \cdot \parallel_{TV,w}}(\theta) = \arg\min_{\theta \in \mathbb{R}^m} \left\{ \frac{1}{2} \|\beta - \theta\|_2^2 + \|\theta\|_{TV,\hat{w}} \right\}. \]  

The proposed algorithm consists in running forwardly through the samples \((\theta_1, \ldots, \theta_m)\). Using the Karush-Kuhn-Tucker (KKT) optimality conditions for a convex optimization (Boyd and Vandenberghe, 2004), at location \(k\), \(\beta_k\) stays constant where \(|u_k| < \hat{w}_{k+1}\). Here \(u_k\) is a solution to a dual problem associated to the primal problem (15). If this is not possible, it goes back to the last location where a jump can be introduced in \(\beta\), validates the current segment until this location, starts a new segment, and continues. This algorithm is described precisely in Algorithm 2.

Appendix C. Proof of Theorem 2: fast oracle inequality under binarsity

In this section, we prove Theorem 2. The proof relies on some technical properties given below.

Additional notation. Hereafter, we use the following vector notations: \(y = (y_1, \ldots, y_n)^\top\), \(m^0(X) = (m^0(x_1), \ldots, m^0(x_n))^\top\), \(m_\theta(X) = (m_\theta(x_1), \ldots, m_\theta(x_n))^\top\) (recall that \(m_\theta(x_i) = \theta^\top x_i^B\)), and \(b'(m_\theta(X)) = (b'(m_\theta(x_1)), \ldots, b'(m_\theta(x_n)))^\top\).

C.1 Empirical Kullback-Leibler divergence.

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

\[ KL_n(f^0, f_\theta) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_{y|x}} \left[ \log \frac{f^0(y_i|x_i)}{f_\theta(y_i|x_i)} \right] := KL_n(m^0(X), m_\theta(X)), \]

where \(P_{y|x}\) is the joint distribution of \(y = (y_1, \ldots, y_n)^\top\) given \(X = (x_1, \ldots, x_n)^\top\). We then have the following property.

Lemma 3 The excess risk verifies \(R(m_\theta) - R(m^0) = \phi KL_n(m^0(X), m_\theta(X))\).

Proof. Straightforwardly, one has

\[ KL_n(m^0(X), m_\theta(X)) = \phi^{-1} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_{y|x}} \left[ ( - y_i m_\theta(x_i) + b(m_\theta(x_i))) - ( - y_i m^0(x_i) + b(m^0(x_i))) \right] \]

\[ = \phi^{-1} (R(m_\theta) - R(m^0)). \]
Algorithm 2: Proximal operator of weighted TV penalization

**Input:** vector $\theta = (\theta_1, \ldots, \theta_m)^T \in \mathbb{R}^m$ and weights $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m) \in \mathbb{R}_+^m$.

**Output:** vector $\beta = \text{prox}_{||\cdot||_{TV}, \hat{w}}(\theta)$

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

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

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

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

5. 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
   
   $\beta_{\max} \leftarrow \beta_{\max} + \frac{u_{\max} + \hat{w}_{k+1}}{k - k_0 + 1}$
   
   $u_{\max} \leftarrow -\hat{w}_{k+1}$
   
   $k_+ \leftarrow k$

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

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

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

9. else
   
   $\beta_{k_0} = \ldots = \beta_m \leftarrow \beta_{\min} + \frac{u_{\min}}{k - k_0 + 1}$
C.2 Optimality conditions.

To characterize the solution of the problem (6), the following result can be straightforwardly obtained using the Karush-Kuhn-Tucker (KKT) optimality conditions for a convex optimization (Boyd and Vandenberghe 2004).

**Proposition 4** A vector \( \hat{\theta} = (\hat{\theta}_1^\top, \ldots, \hat{\theta}_p^\top)^\top \in \mathbb{R}^d \) is an optimum of the objective function in (6) if and only if there exists a sequence of subgradients \( \hat{h} = (\hat{h}_{j, \bullet})_{j=1, \ldots, p} \in \partial(\|\hat{\theta}\|_{TV, \hat{w}}) \) and \( \hat{g} = (\hat{g}_{j, \bullet})_{j=1, \ldots, p} \in \partial(\delta_1(\hat{\theta}_{j, \bullet}))_{j=1, \ldots, p} \) such that

\[
\nabla R_n(\hat{\theta}_{j, \bullet}) + \hat{h}_{j, \bullet} + \hat{g}_{j, \bullet} = 0_{d_j},
\]

where

\[
\begin{align*}
\hat{h}_{j, \bullet} & = D_j^\top (\hat{w}_{j, \bullet} \odot \text{sign}(D_j \hat{\theta}_{j, \bullet})) & \text{if } j \in J(\hat{\theta}), \\
\hat{h}_{j, \bullet} & = D_j^\top (\hat{w}_{j, \bullet} \odot [-1, +1]^{d_j}) & \text{if } j \in J^c(\hat{\theta}),
\end{align*}
\]

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

\[
\partial(\delta_1(\hat{\theta}_{j, \bullet})) = \{ \mu_{j, \bullet} \in \mathbb{R}^{d_j} : \langle \mu_{j, \bullet}, \hat{\theta}_{j, \bullet} \rangle \leq \langle \mu_{j, \bullet}, \hat{\theta}_{j, \bullet} \rangle \text{ for all } \theta_{j, \bullet} \text{ such that } 1_j^\top \theta_{j, \bullet} = 0 \}.
\]

For the generalized linear model, we have

\[
\frac{1}{n} X_{B}^\top \left( b'(m_{\hat{\theta}}(X)) - y \right) + \hat{h}_{j, \bullet} + \hat{g}_{j, \bullet} + \hat{f}_{j, \bullet} = 0_{d_j},
\]

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

**Proof.** We denote by \( \partial(\phi) \) the subdifferential mapping of a convex functional \( \phi \). The function \( \theta \mapsto R_n(\theta) \) is differentiable, so the subdifferential of \( R_n(\cdot) + \text{bina}(\cdot) \) at a point \( \theta = (\theta_{j, \bullet})_{j=1, \ldots, p} \in \mathbb{R}^d \) is given by

\[
\partial(\nabla R_n(\theta) + \partial(\text{bina}(\theta))) = \nabla R_n(\theta) + \partial(\text{bina}(\theta)),
\]

where

\[

abla R_n(\theta) = \left( \frac{\partial R_n(\theta)}{\partial \theta_{1, \bullet}}, \ldots, \frac{\partial R_n(\theta)}{\partial \theta_{p, \bullet}} \right)^\top
\]

and

\[
\partial(\text{bina}(\theta)) = \left( \partial(\|\theta_{1, \bullet}\|_{TV, \hat{w}_{1, \bullet}}) + \partial(\delta_1(\theta_{1, \bullet})), \ldots, \partial(\|\theta_{p, \bullet}\|_{TV, \hat{w}_{p, \bullet}}) + \partial(\delta_1(\theta_{p, \bullet})) \right)^\top.
\]

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

\[
\partial(\|\theta_{j, \bullet}\|_{TV, \hat{w}_{j, \bullet}}) = \begin{cases} 
D_j^\top \text{sign}(\hat{w}_{j, \bullet} \odot D_j \theta_{j, \bullet}) & \text{if } D_j \theta \neq 0_{d_j}, \\
D_j^\top (\hat{w}_{j, \bullet} \odot v_j) & \text{otherwise},
\end{cases}
\]

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

\[
\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ R_n(\theta) + \text{bina}(\theta) + \delta_{B_d(\rho)}(\theta) \right\},
\]
where \( \delta_{B_d(\rho)} \) is the indicator function for \( B_d(\rho) \). Now, \( \hat{\theta} = (\hat{\theta}_1^T, \ldots, \hat{\theta}_{p^*}^T) \) is an optimum of Problem (19) if and only if \( 0_d \in \nabla R_n(m_{\hat{\theta}}) + \partial \left( \| \hat{\theta} \|_{TV, \hat{w}} + \partial (\delta_{B_d(\rho)}(\hat{\theta})) \right) \). Recall that the subdifferential of \( \delta_{B_d(\rho)}(\cdot) \) is the normal cone of \( B_d(\rho) \), that is
\[
\partial (\delta_{B_d(\rho)}(\hat{\theta})) = \{ \eta \in \mathbb{R}^d : \langle \eta, \theta \rangle \leq \langle \eta, \hat{\theta} \rangle \text{ for all } \theta \in B_d(\rho) \}. \tag{20}
\]

Straightforwardly, one obtains
\[
\frac{\partial (R_n(\theta))}{\partial (\theta_{j,*})} = \frac{1}{n}(X_{j}^{B})^T(b'(m_{\hat{\theta}}(X)) - y), \tag{21}
\]
and equalities (21) and (20) give equation (17), which ends the proof of Proposition 4.

\[ \square \]

**C.3 Compatibility condition for \( T \).**

We need, in addition to Assumption 2, the following result which gives a compatibility condition satisfied by the matrix \( T \). To this end, for any concatenation of subsets \( K = [K_1, \ldots, K_p] \), we set
\[
K_j = \{ \tau_j^1, \ldots, \tau_j^{b_j} \} \subset \{ 1, \ldots, d_j \} \tag{22}
\]
for all \( j = 1, \ldots, p \) and with the convention that \( \tau_j^0 = 0 \) and \( \tau_j^{b_j+1} = d_j + 1 \).

**Lemma 5** Let \( \gamma \in \mathbb{R}^d \) be a given vector of weights and \( K = [K_1, \ldots, K_p] \) with \( K_j \) given by (22) for all \( j = 1, \ldots, p \). Then for every \( u \in \mathbb{R}^d \setminus 0_d \), we have
\[
\frac{\| Tu \|_2}{\| u_K \odot \gamma_K \|_1 - \| u_{K^c} \odot \gamma_{K^c} \|_1} \geq \kappa_{T, \gamma}(K),
\]
where
\[
\kappa_{T, \gamma}(K) = \left\{ \frac{32}{\Delta_{\min, K_j}} \sum_{j=1}^p \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2K_j \| \gamma_{j,*} \|_\infty \Delta_{\min, K_j}^{-1} \right\}^{-1/2},
\]
and \( \Delta_{\min, K_j} = \min_{r=1, \ldots, b_j} |\tau_j^r - \tau_j^{r-1}|. \)

**Proof.** Using Proposition 3 in Dalalyan et al. (2014), we have
\[
\| u_K \odot \gamma_K \|_1 - \| u_{K^c} \odot \gamma_{K^c} \|_1 = \sum_{j=1}^p \| u_{K_j} \odot \gamma_{K_j} \|_1 - \| u_{K^c_j} \odot \gamma_{K^c_j} \|_1 \\
\leq \sum_{j=1}^p 4\| T_j u_{j,*} \|_2 \left\{ 2 \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2(b_j + 1) \| \gamma_{j,*} \|_\infty \Delta_{\min, K_j}^{-1} \right\}^{1/2}.
\]

Applying Hölder’s inequality for the right hand side of the last inequality gives
\[
\| u_K \odot \gamma_K \|_1 - \| u_{K^c} \odot \gamma_{K^c} \|_1 \\
\leq \| Tu \|_2 \left\{ 32 \sum_{j=1}^p \sum_{k=1}^{d_j} |\gamma_{j,k+1} - \gamma_{j,k}|^2 + 2K_j \| \gamma_{j,*} \|_\infty \Delta_{\min, K_j}^{-1} \right\}^{1/2}.
\]

This completes the proof. \[ \square \]
C.4 Compatibility condition for $X^BT$.

Now, using Assumption 2 and Lemma 5, we establish a compatibility condition satisfied by the product of matrices $X^BT$.

**Lemma 6** Let Assumption 2 holds. Let $\gamma \in \mathbb{R}_+^d$ be a given vector of weights, and $K = [K_1, \ldots, K_p]$ such that $K_j$ is given by (22) for all $j = 1, \ldots, p$. Then, one has

$$\inf_{u \in \mathcal{C}_{1,\hat{w}}(K) \setminus \{0_d\}} \left\{ \frac{\|X^B Tu\|_2}{\sqrt{n} \|u_K \otimes \gamma_K\|_1 - \|u_{K^c} \otimes \gamma_{K^c}\|_1} \right\} \geq \kappa_{T,\gamma}(K)\kappa(K),$$

(23)

where

$$\mathcal{C}_{1,\hat{w}}(K) = \left\{ u \in \mathbb{R}^d : \sum_{j=1}^p \|(u_{j,\bullet})_{K_j} \|_1,\hat{w}_{j,\bullet} \leq 2 \sum_{j=1}^p \|(u_{j,\bullet})_{K_j} \|_1,\hat{w}_{j,\bullet} \right\},$$

(24)

with $\| \cdot \|_{1,a}$ denoting the weighted $\ell_1$-norm.

**Proof.** By Lemma 5, we have that

$$\frac{\|X^B Tu\|_2}{\sqrt{n} \|u_K \otimes \gamma_K\|_1 - \|u_{K^c} \otimes \gamma_{K^c}\|_1} \geq \kappa_{T,\gamma}(K)\|X^B Tu\|_2 \sqrt{n}\|Tu\|_2.$$

Now, we note that if $u \in \mathcal{C}_{1,\hat{w}}(K)$, then $Tu \in \mathcal{C}_{TV,\hat{w}}(K)$. Hence, by Assumption 2, we get

$$\frac{\|X^B Tu\|_2}{\sqrt{n} \|u_K \otimes \gamma_K\|_1 - \|u_{K^c} \otimes \gamma_{K^c}\|_1} \geq \kappa_{T,\gamma}(K)\kappa(K).$$

\[
\]

C.5 Connection between empirical Kullback-Leibler divergence and the empirical squared norm.

We remark that the binarized matrix $X^B$ satisfies $\max_{i=1,\ldots,n} \|x_i^B\|_2 = \sqrt{p}$. A direct consequence of this remark is given in the next lemma.

**Lemma 7** One has

$$\max_{i=1,\ldots,n} \sup_{\theta \in B_d(\rho)} |\langle x_i^B, \theta \rangle| \leq \rho\sqrt{p}.$$

(25)

To compare the empirical Kullback-Leibler divergence and the empirical squared norm, we use Lemma 1 in Bach (2010), that we recall here.

**Lemma 8** Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex three times differentiable 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 $\psi(u) = e^u - u - 1$.

Now, we give a version of the previous Lemma in our setting.
Lemma 9 Under Assumption 1 and with \( C_n(p, p) \) as defined in Theorem 2 one has
\[
\frac{L_n}{\phi C_n^2(p, p)} \frac{1}{n} \| m_0(X) - m_\theta(X) \|_2^2 \leq KL_n(m_0(X), m_\theta(X)),
\]
\[
\frac{U_n}{\phi C_n^2(p, p)} \frac{1}{n} \| m_0(X) - m_\theta(X) \|_2^2 \geq KL_n(m_0(X), m_\theta(X)),
\]
for all \( \theta \in B_d(p) \).

**Proof.** Let us consider the function \( G_n : \mathbb{R} \to \mathbb{R} \) defined by \( G_n(t) = R_n(m^0 + tm_\eta) \), then
\[
G_n(t) = \frac{1}{n} \sum_{i=1}^{n} b(m^0(x_i) + tm_\eta(x_i)) - \frac{1}{n} \sum_{i=1}^{n} y_i(m^0(x_i) + tm_\eta(x_i)).
\]

By differentiating \( G_n \) three times with respect to \( t \), we obtain
\[
G''_n(t) = \frac{1}{n} \sum_{i=1}^{n} m_\eta(x_i)b'(m^0(x_i) + tm_\eta(x_i)) - \frac{1}{n} \sum_{i=1}^{n} y_im_\eta(x_i),
\]
\[
G'''_n(t) = \frac{1}{n} \sum_{i=1}^{n} m_\eta^2(x_i)b''(m^0(x_i) + tm_\eta(x_i)),
\]
and \( G''''_n(t) = \frac{1}{n} \sum_{i=1}^{n} m_\eta^3(x_i)b'''(m^0(x_i) + tm_\eta(x_i)) \).

In all the considered models, we have \( |b'''(z)| \leq 2|b''(z)| \), see the following table

<table>
<thead>
<tr>
<th>Model</th>
<th>( \phi )</th>
<th>( b(z) )</th>
<th>( b'(z) )</th>
<th>( b''(z) )</th>
<th>( b'''(z) )</th>
<th>( L_n )</th>
<th>( U_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>( \sigma^2 )</td>
<td>( \frac{z^2}{2} )</td>
<td>( z )</td>
<td>( 1 )</td>
<td>( 0 )</td>
<td>( 1 )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>Logistic</td>
<td>1</td>
<td>( \log(1 + e^z) )</td>
<td>( \frac{e^z}{1+e^z} )</td>
<td>( \frac{e^z}{(1+e^z)^2} )</td>
<td>( \frac{1-e^z}{1+e^z} b''(z) )</td>
<td>( \frac{1}{1+e^Cn} )</td>
<td>( \frac{1}{eCn} )</td>
</tr>
<tr>
<td>Poisson</td>
<td>1</td>
<td>( e^z )</td>
<td>( e^z )</td>
<td>( e^z )</td>
<td>( b''(z) )</td>
<td>( e^{-C_n} )</td>
<td>( e^{C_n} )</td>
</tr>
</tbody>
</table>

Then, we get \( |G''''_n(t)| \leq 2\|m_\eta\|_\infty|G_n(t)| \) where \( \|m_\eta\|_\infty := \max_{i=1,...,n} |m_\eta(x_i)| \). Applying Lemma 8 with \( M = 2\|m_\eta\|_\infty \), we obtain
\[
G''_n(0) \frac{\psi(-2\|m_\eta\|_\infty t)}{4\|m_\eta\|_\infty^2} \leq G_n(t) - G_n(0) - tG'_n(0) \leq G''_n(0) \frac{\psi(2\|m_\eta\|_\infty t)}{4\|m_\eta\|_\infty^2}.
\]

for all \( t \geq 0 \). Taking \( t = 1 \) leads to
\[
G''_n(0) \frac{\psi(-2\|m_\eta\|_\infty)}{4\|m_\eta\|_\infty^2} \leq R_n(m^0 + m_\eta) - R_n(m^0) - G'_n(0),
\]
\[
G''_n(0) \frac{\psi(2\|m_\eta\|_\infty)}{4\|m_\eta\|_\infty^2} \geq R_n(m^0 + m_\eta) - R_n(m^0) - G'_n(0).
\]

A short calculation gives that
\[
-G'_n(0) = \frac{1}{n} \sum_{i=1}^{n} m_\eta(x_i)(y_i - b'(m^0(x_i))), \quad \text{and} \quad G''_n(0) = \frac{1}{n} \sum_{i=1}^{n} m_\eta^2(x_i)b''(m_\theta(x_i)).
\]
According to Proposition 4, Equation (26) involves that there is $\hat{\theta}$ and $\hat{\eta} = \frac{1}{n} \sum_{i=1}^{n} y_{i} m_{\theta}(x_{i})$ such that

$$ 2\|m_{\eta}\|_{\infty} \leq \sum_{i=1}^{n} |\langle x_{i}, \theta \rangle| + |m^{0}(x_{i})| \leq 2(\rho \sqrt{p} + C_{n}) = C_{n}(\rho, p). $$

Hence, we obtain

$$ G_{n}(\theta, m_{\theta}) \leq R(m_{\theta}) - R(m_{\theta}) = \phi KL_{n}(m^{0}(X), m_{\theta}(X)), $$

$$ C_{n}(\theta, m_{\theta}) \geq \phi KL_{n}(m^{0}(X), m_{\theta}(X)), $$

with $G_{n}(\theta) = C_{n}(\theta, m_{\theta})$. It entails that

$$ L_{n}\psi(-C_{n}(\theta, m_{\theta})) \leq \frac{1}{\phi C_{n}^{2}(\theta, m_{\theta})} \sum_{i=1}^{n} (m_{\theta}(x_{i}) - m^{0}(x_{i}))^{2} \leq KL_{n}(m^{0}(X), m_{\theta}(X)) $$

and

$$ \hat{\theta} \in \arg\min_{\hat{\theta} \in B_{d}(\rho)} \{ R_{n}(\theta) + bina(\theta) \}. $$

**C.6 Proof of Theorem 2**

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

$$ R_{n}(m_{\theta}) = \frac{1}{n} \sum_{i=1}^{n} b(m_{\theta}(x_{i})) - \frac{1}{n} \sum_{i=1}^{n} y_{i} m_{\theta}(x_{i}) $$

and

$$ \hat{\theta} \in \arg\min_{\hat{\theta} \in B_{d}(\rho)} \{ R_{n}(\theta) + bina(\theta) \}. $$

According to Proposition 4, Equation (26) involves that there is $\hat{h} = (\hat{h}_{j})_{j=1,\ldots,p} \in \partial(\|\hat{\theta}\|_{TV}, \hat{\omega})$, $\hat{g} = (\hat{g}_{j})_{j=1,\ldots,p} \in \partial(\delta_{1}(\hat{\theta}_{j}, \cdot))$ and $\hat{f} = (\hat{f}_{j})_{j=1,\ldots,p} \in \partial(\delta_{B_{d}(\rho)}(\hat{\theta}))$ such that

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

for all $\theta \in \mathbb{R}^{d}$, which can be written

$$ \frac{1}{n} \langle b^{\prime}(m_{\hat{\theta}}(X)) - b^{\prime}(m^{0}(X)), m_{\hat{\theta}}(X) - m_{\theta}(X) \rangle $$

$$ - \frac{1}{n} \langle y - b^{\prime}(m^{0}(X)), m_{\hat{\theta}}(X) - m_{\theta}(X) \rangle + \langle \hat{h} + \hat{g} + \hat{f}, \hat{\theta} - \theta \rangle = 0.$$
For any $\theta \in B_d(\rho)$ such that $1^T \theta = 0$, and $h \in \partial(\|\theta\|_{TV, \omega})$, the monotony of the subdifferential mapping implies $\langle h, \theta - \hat{\theta} \rangle \leq \langle h, \theta - \hat{\theta} \rangle$, $\langle \hat{g}, \theta - \hat{\theta} \rangle \leq 0$, and $\langle \hat{f}, \theta - \hat{\theta} \rangle \leq 0$. Therefore

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

(27)

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

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

By differentiating $H_n$ three times with respect to $t$, we obtain

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

$$H''_n(t) = \frac{1}{n} \sum_{i=1}^{n} m_{\eta}^2(x_i)b''(m_{\hat{\theta} + t\eta}(x_i)),$$

and

$$H'''_n(t) = \frac{1}{n} \sum_{i=1}^{n} m_{\eta}^3(x_i)b'''(m_{\hat{\theta} + t\eta}(x_i)).$$

Using Lemma 7, we have $|H'''_n(t)| \leq 2\rho\sqrt{p}|H'_n(t)|$. Applying now Lemma 8 with $M(\rho, p) = 2\rho\sqrt{p}$, we obtain

$$H'''_n(0) \frac{\psi(-tM(\rho, p))}{M^2(\rho, p)} \leq H'_n(t) - H'_n(0) - tH''_n(0) \leq H'''_n(0) \frac{\psi(tM(\rho, p))}{M^2(\rho, p)},$$

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

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

and

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

Moreover, we have

$$H'_n(0) = \frac{1}{n} \sum_{i=1}^{n} \langle x_i^B, \theta - \hat{\theta} \rangle b'(m_{\hat{\theta}}(x_i)) - \frac{1}{n} \sum_{i=1}^{n} b'(m^0(x_i))\langle x_i^B, \hat{\theta} - \theta \rangle$$

$$= \frac{1}{n} \langle b'(m_{\hat{\theta}}(X)) - b'(m^0(X)), X^B(\theta - \hat{\theta}) \rangle,$$

and

$$H''_n(0) = \frac{1}{n} \sum_{i=1}^{n} \langle x_i^B, \hat{\theta} - \theta \rangle^2 b''(m_{\hat{\theta}}(x_i)).$$

Then, we deduce that

$$H'''_n(0) \frac{\psi(-M(\rho, p))}{M(\rho, p)^2} \leq R(m_{\theta}) - R(m_{\hat{\theta}}) - \frac{1}{n} \langle b'(m_{\hat{\theta}}(X)) - b'(m^0(X)), X^B(\theta - \hat{\theta}) \rangle$$

$$= \phi KL_n(m^0(X), m_{\theta}(X)) - \phi KL_n(m^0(X), m_{\hat{\theta}}(X)) + \frac{1}{n} \langle b'(m_{\hat{\theta}}(X)) - b'(m^0(X)), m_{\hat{\theta}}(X) - m_{\theta}(X) \rangle.$$
Then, with Equation (27), one has
\[
\phi KL_n(m^0(X), m_{\hat{\theta}}(X)) + H''_n(0) \frac{\psi(-M(\rho, p))}{M^2(\rho, p)} \leq \phi KL_n(m^0(X), m_{\theta}(X)) + \frac{1}{n} \langle y - b'(m^0(X)), m_{\hat{\theta}}(X) - m_{\theta}(X) \rangle - \langle h, \hat{\theta} - \theta \rangle. \tag{28}
\]
As $H''_n(0) \geq 0$, it implies that
\[
\phi KL_n(m^0(X), m_{\hat{\theta}}(X)) \leq \phi KL_n(m^0(X), m_{\theta}(X)) + \frac{1}{n} \langle y - b'(m^0(X)), m_{\hat{\theta}}(X) - m_{\theta}(X) \rangle - \langle h, \hat{\theta} - \theta \rangle. \tag{29}
\]
If $\frac{1}{n} \langle y - b'(m^0(X)), X^B(\hat{\theta} - \theta) \rangle - \langle h, \hat{\theta} - \theta \rangle < 0$, it follows that
\[
KL_n(m^0(X), m_{\hat{\theta}}(X)) \leq KL_n(m^0(X), m_{\theta}(X)),
\]
then Theorem 2 holds. From now on, let us assume that
\[
\frac{1}{n} \langle y - b'(m^0(X)), m_{\hat{\theta}}(X) - m_{\theta}(X) \rangle - \langle h, \hat{\theta} - \theta \rangle \geq 0. \tag{30}
\]
We first derive a bound on $\frac{1}{n} \langle y - b'(m^0(X)), m_{\hat{\theta}}(X) - m_{\theta}(X) \rangle$. Let us define the block diagonal matrix $D = \text{diag}(D_1, \ldots, D_p)$, with $D_j$, defined in (13), being invertible. 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 $T = \text{diag}(T_1, \ldots, T_p)$. Using $TD = I_d$, we focus on finding out a bound of $\frac{1}{n} \langle (X^B T)\top (y - b'(m^0(X))), D(\hat{\theta} - \theta) \rangle$. In one hand, using that $D^{-1} = T$, one has
\[
\frac{1}{n} \langle (X^B)\top (y - b'(m^0(X)), \hat{\theta} - \theta) = \frac{1}{n} \langle (X^B T)\top (y - b'(m^0(X)), D(\hat{\theta} - \theta) \rangle
\leq \frac{1}{n} \sum_{j=1}^{p} \sum_{k=1}^{d_j} \left\| ((X_{\bullet j}^B T_j)_{\bullet k}, y - b'(m^0(X)) \right\| \left( (D_j(\hat{\theta}_{\bullet} - \theta_{\bullet}))_{k} \right),
\]
where $(X_{\bullet j}^B T_j)_{\bullet k} = ((X_{\bullet j}^B T_j)_{1,k}, \ldots, (X_{\bullet j}^B T_j)_{n,k})\top \in \mathbb{R}^n$ is the $k$-th column of the matrix $(X_{\bullet j}^B T_j)$. Let us consider the event
\[
\mathcal{E}_n = \bigcap_{j=1}^{p} \bigcap_{k=2}^{d_j} \mathcal{E}_{n,j,k}, \text{ where } \mathcal{E}_{n,j,k} = \left\{ \frac{1}{n} \left\| ((X_{\bullet j}^B T_j)_{\bullet k}, y - b'(m^0(X))) \right\| \leq \hat{w}_{j,k} \right\}.
\]
Then, on $\mathcal{E}_n$, we have
\[
\frac{1}{n} \langle (X^B)\top (y - b'(m^0(X)), \hat{\theta} - \theta) \leq \sum_{j=1}^{p} \sum_{k=1}^{d_j} \hat{w}_{j,k} \left\| (D_j(\hat{\theta}_{\bullet} - \theta_{\bullet}))_{k} \right\|
\leq \sum_{j=1}^{p} \| \hat{w}_{j,\bullet} \odot D_j(\hat{\theta}_{\bullet} - \theta_{\bullet}) \|_1. \tag{31}
\]
In another hand, from the definition of the subgradient \((h_j, \theta)_{j=1,\ldots,p} \in \partial(\|\theta\|_{TV, \phi})\) (see Equation (16)), one can choose \(h\) such that
\[
h_{j,k} = \left( D^\top_j (\hat{\omega}_{j,\theta} \odot \text{sign}(D_j \theta_j)) \right)_k,
\]
for all \(k = 1, \ldots, J_j(\theta)\) and
\[
h_{j,k} = \left( D^\top_j (\hat{\omega}_{j,\theta} \odot \text{sign}(D_j \hat{\theta}_j)) \right)_k = \left( D^\top_j (\hat{\omega}_{j,\theta} \odot \text{sign}(D_j (\hat{\theta}_j - \theta_j))) \right)_k
\]
for all \(k = 1, \ldots, J^p_j(\theta)\). Using a triangle inequality and the fact that \((\text{sign}(x), x) = \|x\|_1\), we obtain
\[
-(h, \hat{\theta} - \theta) \leq \sum_{j=1}^p \|((\hat{\omega}_{j,\theta})_{J_j(\theta)} \odot D_j (\hat{\theta}_j - \theta_j))_{J_j(\theta)}\|_1
\]
\[-\sum_{j=1}^p \|((\hat{\omega}_{j,\theta})_{J^p_j(\theta)} \odot D_j (\hat{\theta}_j - \theta_j))_{J^p_j(\theta)}\|_1
\]
\[\leq \sum_{j=1}^p \|((\hat{\theta}_j - \theta_j)_{J_j(\theta)})_{TV, \hat{\omega}_j} - \sum_{j=1}^p \|((\hat{\theta}_j - \theta_j)_{J_j(\theta)})_{TV, \hat{\omega}_j}\|_1.
\]
Combining inequalities (31) and (32), we get
\[
\sum_{j=1}^p \|((\hat{\theta}_j - \theta_j)_{J_j(\theta)})_{TV, \hat{\omega}_j}\|_1 \leq 2 \sum_{j=1}^p \|((\hat{\theta}_j - \theta_j)_{J_j(\theta)})_{TV, \hat{\omega}_j}\|_1.
\]
on \(\mathcal{E}_n\). Hence
\[
\sum_{j=1}^p \|((\hat{\omega}_{j,\theta})_{J^p_j(\theta)} \odot D_j (\hat{\theta}_j - \theta_j))_{J^p_j(\theta)}\|_1 \leq 2 \sum_{j=1}^p \|((\hat{\omega}_{j,\theta})_{J_j(\theta)} \odot D_j (\hat{\theta}_j - \theta_j))_{J_j(\theta)}\|_1.
\]
This means that
\[
\hat{\theta} - \theta \in \mathcal{C}_{TV, \phi} (J(\theta)) \text{ and } D (\hat{\theta} - \theta) \in \mathcal{C}_{1, \phi} (J(\theta)),
\]
see (7) and (24). Now, going back to (29) and taking into account (33), the compatibility of \(X^{B, T}\) (see (23)), on \(\mathcal{E}_n\) the following holds
\[
\phi KL_n(m^0(X), m_{\hat{\theta}}(X)) \leq \phi KL_n(m^0(X), m_{\theta}(X)) + 2 \sum_{j=1}^p \|((\hat{\omega}_{j,\theta})_{J_j(\theta)} \odot D_j (\hat{\theta}_j - \theta_j))_{J_j(\theta)}\|_1.
\]
Then
\[
KL_n(m^0(X), m_{\hat{\theta}}(X)) \leq KL_n(m^0(X), m_{\theta}(X)) + \frac{\|m_{\hat{\theta}}(X) - m_{\theta}(X)\|_2}{\sqrt{n} \phi \kappa_{T, \hat{\gamma}} (J(\theta)) \kappa (J(\theta))},
\]
where \(\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)\) such that
\[
\hat{\gamma}_{j,k} = \begin{cases} 
2\hat{\omega}_{j,k} & \text{if } k \in J_j(\theta), \\
0 & \text{if } k \in J^p_j(\theta),
\end{cases}
\]
for all $j = 1, \ldots, p$ and

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

Next, we find an upper bound for $1/\kappa_{T, \tilde{\gamma}}^2(J(\theta))$. We have

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

Note that $||\gamma_{j,\cdot}||_\infty \leq 2||\hat{\gamma}_{j,\cdot}||_\infty$. We write the set $J_j(\theta) = \{ k_j^1, \ldots, k_j^{\delta J_j(\theta)} \}$ and we set $B_r = [k_j^{-1}, k_j^r] = \{ k_j^{-1}, k_j^{-1} + 1, \ldots, k_j^r - 1 \}$ for $r = 1, \ldots, |J_j(\theta)| + 1$ with the convention that $k_j^0 = 0$ and $k_j^{\delta J_j(\theta)+1} = d_j + 1$. Then

$$
\sum_{k=1}^{d_j} |\gamma_{j,k+1} - \hat{\gamma}_{j,k}|^2 = \sum_{r=1}^{|J_j(\theta)|+1} \sum_{k \in B_r} |\gamma_{j,k} - \hat{\gamma}_{j,k}|^2 = \sum_{r=1}^{|J_j(\theta)|+1} |\hat{\gamma}_{j,k_r+1} - \hat{\gamma}_{j,k_r}|^2 + |\hat{\gamma}_{j,k_r} - \gamma_{j,k_r-1}|^2
\leq 8 |J_j(\theta)| \| (\hat{\gamma}_{j,\cdot}, J_j(\theta)) \|_\infty^2.
$$

Therefore

$$
\frac{1}{\kappa_{T, \tilde{\gamma}}^2(J(\theta))} \leq 32 \sum_{j=1}^{p} \left\{ 8 |J_j(\theta)| \| (\hat{\gamma}_{j,\cdot}, J_j(\theta)) \|_\infty^2 + 8 |J_j(\theta)| \| (\hat{\gamma}_{j,\cdot}, J_j(\theta)) \|_\infty^2 \Delta_{\min, J_j(\theta)}^{-1} \right\} + 8 |J_j(\theta)| \| (\hat{\gamma}_{j,\cdot}, J_j(\theta)) \|_\infty^2 \\
\leq (32 \times 8) \sum_{j=1}^{p} \left\{ 1 + \frac{1}{\Delta_{\min, J_j(\theta)}} \right\} |J_j(\theta)| \| (\hat{\gamma}_{j,\cdot}, J_j(\theta)) \|_\infty^2 \\
\leq 512 |J(\theta)| \max_{j=1,\ldots,p} \| (\hat{\gamma}_{j,\cdot}, J_j(\theta)) \|_\infty^2.
$$
Remark 10 For the case of least squares regression where \( y_i | x_i \) has Gaussian distribution with mean \( m^0(x_i) \) and variance \( \phi = \sigma^2 \). Using inequalities (28) and (34), we get

\[
\phi KL_n(m^0(X), m_\theta(X)) + \frac{\psi(-M(\rho, p))}{M^2(\rho, p)} \frac{1}{n} \|m_\theta(X) - m_\theta(X)\|_2^2 \\
\leq \phi KL_n(m^0(X), m_\theta(X)) + \frac{\|m_\theta(X) - m_\theta(X)\|_2}{\sqrt{n\phi KL_n(J(\theta))\kappa(J(\theta))}} \\
\leq \phi KL_n(m^0(X), m_\theta(X)) + 2 \frac{\sqrt{\psi(-M(\rho, p))}}{M(\rho, p)} \frac{1}{\sqrt{n}} \|m_\theta(X) - m_\theta(X)\|_2 \frac{M(\rho, p)}{\sqrt{\psi(-M(\rho, p))\kappa KL_n(J(\theta))\kappa(J(\theta))}}
\]

Using the fact that \( 2uv \leq u^2 + v^2 \) it yields

\[
\phi KL_n(m^0(X), m_\theta(X)) \leq \phi KL_n(m^0(X), m_\theta(X)) + \frac{M^2(\rho, p)}{\psi(-M(\rho, p))\kappa KL_n(J(\theta))\kappa(J(\theta))}
\]

Hence, we derive the following sharp oracle inequality

\[
R(m_\theta) - R(m^0) \leq \inf_{\theta \in B_d(\rho)} \left\{ R(m_\theta) - R(m^0) + \frac{\xi|J(\theta)|}{\kappa^2(J(\theta))} \max_{j=1,\ldots,p} \|\hat{w}_{j,\bullet}J_{j,\theta}\|_\infty^2 \right\},
\]

where

\[
\xi = \frac{512M^2(\rho, p)}{\psi(-M(\rho, p))}.
\]

Now for generalized linear models, we use the connection between the empirical norm and the Kullback-Leibler divergence. First, we have

\[
\frac{\|m_\theta(X) - m_\theta(X)\|_2}{\sqrt{n\phi KL_n(J(\theta))\kappa(J(\theta))}} \\
\leq \frac{1}{\phi KL_n(J(\theta))\kappa(J(\theta))} \left( \frac{1}{\sqrt{n}} \|m_\theta(X) - m^0(X)\|_2 + \frac{1}{\sqrt{n}} \|m^0(X) - m_\theta(X)\|_2 \right).
\]

Therefore, by Lemma 9, we get

\[
\frac{\|m_\theta(X) - m_\theta(X)\|_2}{\sqrt{n\phi KL_n(J(\theta))\kappa(J(\theta))}} \\
\leq \frac{2}{\sqrt{\phi KL_n(J(\theta))\kappa(J(\theta))}} \left( \sqrt{C_n(\rho, p, L_n)^{-1} KL_n(m^0(X), m_\theta(X))} \\
+ \sqrt{C_n(\rho, p, L_n)^{-1} KL_n(m^0(X), m_\theta(X))} \right).
\]

We now use the elementary inequality \( 2uv \leq eu^2 + v^2/\epsilon \) with \( \epsilon > 0 \). Therefore (34) becomes

\[
KL_n(m^0(X), m_\theta(X)) \leq KL_n(m^0(X), m_\theta(X)) + \frac{\epsilon}{\phi KL_n(J(\theta))\kappa^2(J(\theta))} \\
+ 2 \left( \epsilon C_n(\rho, p, L_n)^{-1} KL_n(m^0(X), m_\theta(X)) \right) \\
+ 2 \left( \epsilon C_n(\rho, p, L_n)^{-1} KL_n(m^0(X), m_\theta(X)) \right).
\]
By choosing $2(\varepsilon C_n(\rho, p, L_n))^{-1} < 1$, we get

$$KL_n(m^0(X), m_\hat{\theta}(X)) \leq \frac{1 + 2(\varepsilon C_n(\rho, p, L_n))^{-1} KL_n(m^0(X), m_\theta(X))}{1 - 2(\varepsilon C_n(\rho, p, L_n))^{-1}}$$

$$+ \frac{\epsilon^2 (1 - 2(\varepsilon C_n(\rho, p, L_n))^{-1}) \phi \kappa_n^2(J(\theta)) \kappa_n^2(J(\theta))}{(\varepsilon C_n(\rho, p, L_n) - 2) \phi \kappa_n^2(J(\theta)) \kappa_n^2(J(\theta))}$$

Setting

$$\frac{\varepsilon C_n(\rho, p, L_n) + 2}{\varepsilon C_n(\rho, p, L_n) - 2} = 1 + \frac{4}{\varepsilon C_n(\rho, p, L_n) - 2} = 1 + \zeta,$$

we get the desired result in (10).

Finally, we have to compute the probability of the complementary of the event $\mathcal{E}_n$. This is given by the following:

$$\mathbb{P}[\mathcal{E}_n^c] \leq \sum_{j=1}^{p} \sum_{k=1}^{d_j} \mathbb{P}\left[\left|\left(\mathbf{X}^B_{i,j,k} T_j\right)_{\bullet,k} - b'(m^0(X))\right| \geq \hat{\omega}_{j,k}\right]$$

$$\leq \sum_{j=1}^{p} \sum_{k=2}^{d_j} \mathbb{P}\left[\sum_{i=1}^{n} \left|\left(\mathbf{X}^B_{i,j,k} T_j\right)_{i,k}(y_i - b'(m^0(x_i)))\right| \geq n\hat{\omega}_{j,k}\right].$$

Let $\xi_{i,j,k} = (\mathbf{X}^B_{i,j,k} T_j)_{i,k}$, 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

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

(35)

Applying Lemma 6.1 in Rigollet (2012), using (35) and Assumption 1, we can derive the following Chernoff-type bounds

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

(36)

where $\xi_{i,j,k} = (\xi_{1,j,k}, \ldots, \xi_{n,j,k})^\top \in \mathbb{R}^n$. We have

$$\mathbf{X}^B_{i,j,k} T_j = \begin{bmatrix}
1 & \sum_{k=2}^{d_j} x^B_{1,j,k} & \sum_{k=3}^{d_j} x^B_{1,j,k} & \cdots & \sum_{k=d_j-1}^{d_j} x^B_{1,j,k} & x^B_{1,j,d_j} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
1 & \sum_{k=2}^{d_j} x^B_{n,j,k} & \sum_{k=3}^{d_j} x^B_{n,j,k} & \cdots & \sum_{k=d_j-1}^{d_j} x^B_{n,j,k} & x^B_{n,j,d_j}
\end{bmatrix}.$$ 

Therefore,

$$\|\xi_{j,k}\|^2_2 = \sum_{i=1}^{n} (\mathbf{X}^B_{i,j,k} T_j)_{i,k}^2 = \#\left\{i \in [n] : x_{i,j} \in \bigcup_{r=k}^{d_j} I_{j,r}\right\} = n\tilde{\pi}_{j,k}.$$ 

(37)
Using weights $\hat{w}_{j,k}$ (see (9) in Theorem 2), and (36) together with (37), we find that the probability of the complementary event $\mathcal{E}_n^c$ is smaller than $2e^{-A}$. This concludes the proof of Theorem 2.

References


D. L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via $\ell_1$ minimization. In *PROC. NATL ACAD. SCI. USA 100 2197202*, 2002.


