Sparse hierarchical interaction learning with epigraphical projection

Mingyuan Jiu, Nelly Pustelnik, Member, IEEE, Stefan Janaqi, Mériam Chebre, Philippe Ricoux

Abstract—This work focus on regression optimization problem with hierarchical interactions between variables, which is beyond the additive models in the traditional linear regression. We investigate more specifically two different fashions encountered in the literature to deal with this problem: “hierNet” and structural-sparsity regularization, and study their connections. We propose a primal-dual proximal algorithm based on epigraphical projection to optimize a general formulation of this learning problem. The experimental setting first highlights the improvement of the proposed procedure compared to state-of-the-art methods based on fast iterative shrinkage-thresholding algorithm (i.e. FISTA) or alternating direction method of multipliers (i.e. ADMM) and second we provide fair comparisons between the different hierarchical penalizations. The experiments are conducted both on the synthetic and real data, and they clearly show that the proposed primal-dual proximal algorithm based on epigraphical projection is efficient and effective to solve and investigate the question of the hierarchical interaction learning problem.

Index Terms—Regression learning, hierarchical interaction, primal-dual algorithm, epigraphical projection

I. INTRODUCTION

LEARN interactions between features is of main interest in data processing such as for genomics [1], [2]. Graphical Lasso [3] is an efficient strategy to obtain a graph of interactions from the inverse covariance matrix but its limitation is to not being adapted for a specific task such as regression, which is the core of this contribution. To learn the interactions in the context of regression, additive sparse regression model has to be extended. Several work have been dedicated to this subject in a recent literature [1], [4]–[7]. In this contribution we improve on these recent works by providing a common framework and an efficient minimization strategy based on epigraphical projection and primal-dual proximal algorithms.

One challenge for feature interaction learning is that the interaction number quadratically increases as the feature size, for example, 1000 features will have one million possible interactions, this therefore results into an over-fitting problem due to insufficient observation samples in most situations. To overcome this weakness, the regression weights and the quadratic interactions can be assumed to be sparse such that most features would not contribute to the decision. Sparsity-based regularization is known to be mainly suitable when the feature size is larger than the training samples. Various sparsity regularizations have been proposed and extensively studied in the additive model, for instance, involving $\ell_0$-pseudo-norm [8], $\ell_1$-norm [9], $\ell_\infty$-norm [10], or structural sparsity norm [11]. See [12] for an exhaustive list of sparse-based regularization in learning and [13] to refer to structural sparsity.

This work focuses on classical regression problem, which is still a subject of active research, for instance, [14] in order to reduce the computational cost of support vector regression (SVR) by considering the effective patterns in the training data according to the distribution of nearest neighbours; but also [15] providing a comparisons between $\ell_1$-norm SVR and sparse coding algorithm.

In this work, instead of linear regression, we include a quadratic behaviour in the regression learning to improve the performance. The idea to integrate quadratic interactions in the learning problem including discrimination task is not new, for example, discriminant quadratic learning [16], [17] learns the covariance matrix in the quadratic term to improve the discrimination ability. However, the idea to learn covariance matrix combined with sparsity regularization is more recent with the preliminary work by Zhao et al. [4] and the following works [1], [5]–[7]. Formally, if we denote the regression training set $\mathcal{R} = \{(y_\ell, x_\ell) \in \mathbb{R} \times \mathbb{R}^N | \ell \in \{1, \ldots, L\}\}$, the criterion involved in sparse quadratic interactions learning takes the form

$$\min_{v \in \mathbb{R}^N, \Theta \in \mathbb{R}^{N \times N}} \frac{1}{2} \sum_{\ell = 1}^{L} (y_\ell - x_\ell^\top v - x_\ell^\top \Theta x_\ell)^2 + \Omega(v, \Theta) \tag{1}$$

where $v = (v_{ij}^{(\ell)})_{1 \leq i \leq N, 1 \leq j \leq N}$ models the regression weight vector, $\Theta = (\Theta_{ij}^{(\ell)})_{1 \leq i \leq N, 1 \leq j \leq N}$ denotes the matrix of interactions, and $\Omega(v, \Theta)$ is the regularization term including sparse regularization and other constraints such as hierarchy constraints.

Our work focusses on feature selection, which aims to remove the redundant features and only keep the informative ones. The relevant features are usually correlated and belong to the same group. Many different works have been done in this respect, for instance, multilayer perceptron (MLP) neural network with “controlled redundancy” [18] is applied to feature selection. Xu et al. [19] proposed a semisupervised feature selection by max-relevance and min-redundancy criterion, as a result, the selected features are strong relevance to the labels under supervised manner and avoid redundancy under unsupervised manner. Feature selection by ranking with SVM with a sparse regularization is proposed in [20], and a reweighted
Both regularizations can be written as $C_{\text{FISTA}}$ [22]. This procedure appears to be very efficient when $C = S$, $z(\cdot) = \| \cdot \|_r$ with $r = \{1, +\infty\}$ and $q = +\infty$, i.e.

$$\Omega(v, \Theta) = \frac{\lambda}{2}\|\Theta\|_1 + \lambda \sum_{i=1}^{N} \|\Theta(\ell^i)\|_1 + \iota C(\Theta).$$

(4)

The algorithmic procedure does not resort to inner iterations leading to faster algorithmic procedure.

Following arguments provided in [1]. Section II derives an equivalent writing of (4) using epigraphical formulation and recall the relationship with [21]. Section III presents the primal-dual proximal algorithm iterations and the convergence guarantees. Its specification for solving the minimization problem (1) with regularization function (4) is specified and new epigraphical projections are provided. Section IV evaluates the performance of the proposed strategy compared to FISTA and ADMM formulations proposed in [1], [21] both on synthetic data and real data. Based on the new algorithmic procedure, fair comparisons between a penalization involving $r = 1$ or $r = +\infty$ can be provided. Finally Section V gives our conclusion.

II. EPIGRAPHICAL FORMULATION

The minimization of (1) with the regularization term defined as in Eq. (2) is difficulty, since it involves non-smooth convex functions and symmetry constraint. In [21], Jenatton et al. solve the problem for $r = \infty$ in the weak hierarchy formulation by means of proximal algorithm. However, their algorithm is hardly adaptive to strong hierarchy. In [1], the authors reformulate (4) when $r = 1$ under an epigraphical formulation, which represents the problem as a set of hard constraints. This reformulation helps in the interpretation and it highlights that a reduction in the shrinkage of certain main effects and an increase in the shrinkage of certain interactions. Its second advantage is to help in the design of the ADMM algorithmic scheme. The merit of this formulation is to split the problem as a set of convex constraints, and it allows us to deal with them by computing their projection onto convex sets. The following proposition gives the epigraphical formulation of our minimization problem.

**Proposition 1.** The minimization problem

\[
\min_{v, \Theta} \frac{1}{2} \sum_{\ell=1}^{L} (y_{\ell} - x_{\ell}^T v - x_{\ell}^T \Theta x_{\ell})^2 + \Omega(v, \Theta)
\]

(5)

2Matlab codes will be made available at the time of the publication.
where $\Omega(v, \Theta)$ is defined in (4) can be equivalently formulated as

$$\begin{align*}
\text{minimize}_{(v^+,-v^-,\Theta)} & \frac{1}{2} \sum_{i=1}^{L} (y_{t} - x_i^T (v^+ - v^-) - x_i^T \Theta x_i)^2 \\
& + \frac{\lambda}{2} \|\Theta\|_1 + \lambda \Theta^T (v^+ - v^-) + \sum_{i=1}^{N} \epsilon_{e_r} (v^{(i)}, v^{-(i)}, \Theta^{(i)})
\end{align*}$$

(6)

where $O$ denotes the positive orthant and $E_r = \{ (w^+, w^-, u) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N \mid \|u\|, \leq \omega^+ + \omega^-\}$. The proof follows arguments derived in [1] for the specific case $r = 1$. Let $v = v^+ - v^-$, (6) can be equivalently written as

$$\begin{align*}
\text{minimize}_{v, v^+, v^-} & \frac{1}{2} \sum_{i=1}^{L} (y_{t} - x_i^T v - x_i^T \Theta x_i)^2 \\
& + \frac{\lambda}{2} \|\Theta\|_1 + \lambda \Theta^T (2v^+ - v) \\
\text{subject to} & \|\Theta^{(i)}\|_r \leq 2v^{(i)} + v^{-(i)} \quad (\forall i \in \{1, \ldots, N\}) \\
& (v^+, v^-, \Theta) \in O \times O \times C.
\end{align*}$$

(7)

The constraints involving $(v, v^+, v^-)$ can be reformulated as

$$\begin{align*}
(\forall i \in \{1, \ldots, N\}) & \left\{ \begin{array}{l}
v^{(i)} + v^{-(i)} \geq \frac{1}{2} \|\Theta^{(i)}\|_r + v^{(i)} \\
v^{(i)} \geq v^{(i)} \\
v^{(i)} \geq 0.
\end{array} \right.
\end{align*}$$

yielding to $v^{(i)} \geq \max\{0, v^{(i)}\} + \frac{1}{2} \|\Theta^{(i)}\|_r + v^{(i)}$. Using $|v^{(i)}| = 2 \max\{0, v^{(i)}\} - v^{(i)}$, we get the expected result:

$$\begin{align*}
\text{minimize}_{v, \Theta} & \frac{1}{2} \sum_{i=1}^{L} (y_{t} - x_i^T v - x_i^T \Theta x_i)^2 \\
& + \frac{\lambda}{2} \|\Theta\|_1 + \lambda \sum_{i=1}^{N} \max\{|v^{(i)}|, \|\Theta^{(i)}\|_r\} + \epsilon_{C}(\Theta).
\end{align*}$$

(8)

The algorithmic strategy designed in next section focus on the epigraphical formulation (9).

III. PRIMAL-DUAL PROXIMAL ALGORITHM

Proximal algorithms are derived from two main frameworks that are forward-backward scheme and Douglas-Rachford iterations (both being deduced from Krasnosel’skii-Mann scheme) [24]. FISTA can be presented as an accelerated version of forward-backward iterations [22], while ADMM can be viewed as a Douglas-Rachford in the dual [26]. In the literature dedicated to sparse regression the most encountered algorithm viewed as a Douglas-Rachford in the dual [26]. In the literature that are forward-backward scheme and Douglas-Rachford iterations [22]. The algorithmic scheme designed in next section focus on the epigraphical formulation (6).

Algorithm 1 Primal-dual splitting algorithm

Initialization: Set $\tau > 0$, $\sigma > 0$, $(w^{[0]}, u^{[0]}) \in H \times G$

For $i = 0, 1, \ldots$

$$\begin{align*}
w^{[i+1]} & = \text{prox}_{H} (w^{[i]} - \tau \nabla f(w^{[i]}) - \tau H^* u^{[i]}) \\
u^{[i+1]} & = \text{prox}_{\sigma h_\tau} (u^{[i]} + \sigma H (2w^{[i+1]} - w^{[i]}))
\end{align*}$$

(10)

to often lead to an approximate solution even if global convergence can be obtained in specific cases [23]. Another class of algorithmic procedures allowing to minimize a criterion with more that two functions, possibly including a differentiable function with a Lipschitz gradient, is the class of primal-dual proximal approaches [27], [30]. Several primal-dual proximal schemes have been derived but one of the most popular is based on forward-backward iterations [29], [30] in order to estimate

$$\begin{align*}
\tilde{w} \in \text{Argmin}_{w \in H} f(w) + g(w) + h(H w),
\end{align*}$$

(9)

where $H : H \rightarrow G$ denotes a bounded linear operator, where $H$ and $G$ being two Hilbert spaces, $f : \mathcal{H} \rightarrow \mathcal{G}$, $g : \mathcal{H} \rightarrow \mathcal{R} + \mathcal{R}$ and $h : \mathcal{G} \rightarrow \mathcal{R}$ denote convex, l.s.c and proper functions. We additionally assume that $f$ is differentiable with a Lipschitz constant denoted $\beta > 0$. Iterations are summarized in Algorithm 1 involving the proximity operator defined as

$$\begin{align*}
(\forall \tilde{w} \in \mathcal{H}) \text{ prox}_g (\tilde{w}) = \text{argmin}_{w \in \mathcal{H}} \|w - \tilde{w}\|_2^2 + g(w)
\end{align*}$$

and $h^*$ denotes the Fenchel-Rockafellar conjugate of a function $h$. The proximity operator of the conjugate can be computed according to Moreau identity that is $\text{prox}_{\lambda h^*} (w) = w - \lambda \text{prox}_{\frac{1}{\lambda} h} (w / \lambda)$ for $\sigma > 0$.

Algorithm 1 involves two parameters $\tau$ and $\sigma$. In finite dimensions, the convergence of the sequence $(w^{[i+1]})_{i \in \mathbb{N}}$ to $\tilde{w}$ is insured when

$$\begin{align*}
1 - \sigma \|H\|_2^2 \geq \frac{\beta}{2}.
\end{align*}$$

Additional technical assumptions can be found in [29][Theorem 3.1].

This algorithmic scheme can be extended for dealing with more than three functions by setting $h(H x) = \sum_{k=1}^{K} h_k(H k x)$ involving the computation of $\text{prox}_{h^*}$. The interest of this scheme compared to ADMM is twofold: it first makes the possibility to deal with the gradient of the differentiable function and second it avoids to invert $\sum_{k} H_k^* H_k$.

A. Specificity for minimizing (6)

We first provide the iterations of Algorithm 1 specified to the minimization of (6). We set $w = (v^+, v^-, \Theta) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N$, for weak hierarchy, we can split it as follows:

$$\begin{align*}
f(w) = \frac{1}{2} \sum_{i=1}^{L} (y_{t} - x_i^T (v^+ - v^-) - x_i^T \Theta x_i)^2 \\
& + \lambda \Theta^T (v^+ + v^-),
\end{align*}$$

(10)

$$\begin{align*}
g(w) = \epsilon(\nabla v^+) + \epsilon(\nabla v^-) + \frac{\lambda}{2} \|\Theta\|_1,
\end{align*}$$

(10)

$$\begin{align*}
h(w) = \sum_{i=1}^{N} \epsilon_{e_r} (v^{(i)}, v^{-(i)}, \Theta^{(i)}).
\end{align*}$$
and for strong hierarchy:

\[
\begin{aligned}
  f(w) &= \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T v - x_i^T \Theta(x_i))^2 \\
  g(w) &= \epsilon_o(v^+) + \epsilon_o(v^-) + \epsilon_S(\Theta), \\
  h_1(w) &= \frac{1}{2} \| \Theta \|_1, \\
  h_2(w) &= \sum_{i=1}^{N} \epsilon_{E_i} (v^{(i)} + v^{(-i)}, \Theta^{(i)})
\end{aligned}
\]

(11)

The respective iterations are summarized in Algorithm 2 and 3. The projection onto the positive orthant and on \(S\) have well known closed form expressions [24] that are:

\[
\begin{aligned}
P_O(\cdot) &= \max \{0, \cdot\}, \\
P_S(\Theta) &= \Theta + \Theta^T.
\end{aligned}
\]

The difficulty comes from the computation of \(P_{E_i}\) whose expressions are provided in next section.

B. New epigraphical projection

We first focus on the derivation of \(P_{E_\infty}\).

**Proposition 2.** Let \(u = (u^{(i)})_{1 \leq i \leq N}\). The projection of \((\omega^+, \omega^-, \nu) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N\) on the epigraphic set \(E_\infty\) reads

\[
(\eta^+, \eta^-, \nu) = P_{E_\infty} (\omega^+ + \omega^-, \nu)
\]

with

\[
\begin{aligned}
\eta^- &= \frac{\omega^- - (N-\bar{n}+1)(\omega^+ - \omega^-) + \sum_{i=1}^{N} \nu^{(i)}}{1 + 2(N-\bar{n}+1)} \\
\eta^+ &= \eta^- + \omega^+ - \omega^- \\
p &= P_{[-\eta^- - \eta^+ - \eta^-]}(u)
\end{aligned}
\]

(12)

where \((\nu^{(1)}, \ldots, \nu^{(N)})\) is an ordered version of \(\{u^{(i)}\}_{1 \leq i \leq N}\) in an ascending order and with \(\nu^{(0)} = -\infty\), and \(\bar{n} \in \{1, \ldots, N\}\) such that \(u^{(\bar{n}-1)} < \eta^+ + \eta^- \leq \nu^{(\bar{n})}\).

**Proof.** According to the definition of the epigraphical projection, we solve the following minimization problem:

\[
P_{E_\infty}(\omega^+, \omega^-, \nu) = \arg \min_{\eta^+, \eta^- \geq 0} \left\{ (\eta^+ - \omega^+)^2 + (\eta^- - \omega^-)^2 + \min_{|p^{(i)}| < \eta^+ + \eta^-} \|p - u\|^2 \right. \\
\left. + \sum_{i=1}^{N} \max\{|u^{(i)}| - \eta^+ - \eta^-, 0|^2 \} \right\}
\]

(13)

The inner minimization yields a simple projection for \(u\) to \([-\eta^+ - \eta^-, \eta^+ + \eta^-]\), i.e. \(p = P_{[-\eta^- - \eta^+ - \eta^-]}(u)\), thus Eq. (13) becomes:

\[
P_{E_\infty}(\omega^+, \omega^-) = \arg \min_{\eta^+, \eta^- \geq 0} \left\{ (\eta^+ - \omega^+)^2 + (\eta^- - \omega^-)^2 \\
+ \sum_{i=1}^{N} \max\{|u^{(i)}| - \eta^+ - \eta^- , 0|^2 \} \right\}
\]

(14)

and thus

\[
P_{E_\infty}(\omega^+, \omega^-) = \text{prox}_{\phi_{\omega^+, \omega^-}}(\omega^+, \omega^-)
\]

(15)

where \(\phi(\omega^+, \omega^-) = \sum_{i=1}^{N} \max\{|u^{(i)}| - \eta^+ - \eta^- , 0|^2 \} \). We now focus on the computation of \(\text{prox}_{\phi}(\omega^+, \omega^-)\).

First, we sort \(|u^{(i)}|\) \(1 \leq i \leq N\) in ascending order to be \((\nu^{(1)}, \ldots, \nu^{(N)})\), such that \(|\nu^{(1)}| \leq \ldots \leq |\nu^{(N)}|\), and also \(\nu^{(0)} = -\infty\), and second \(\bar{n} \in [1, N]\) can be found such that \(\nu^{(\bar{n}-1)} \leq \eta^+ + \eta^- \leq \nu^{(\bar{n})}\), then \(\phi(\omega^+, \omega^-) = \sum_{i=\bar{n}}^{N} (\eta^+ - \eta^-)^2\), so the proximity operator of \(\phi\) has:

\[
\begin{aligned}
\nu^{(\bar{n}-1)} &\leq \eta^+ + \eta^- \leq \nu^{(\bar{n})}, \\
\omega^+ - \eta^+ &\leq (N-\bar{n}+1)(\eta^+ + \eta^-) - \sum_{i=\bar{n}}^{N} \nu^{(i)}, \\
\omega^- - \eta^- &\leq \omega^+ - \eta^+,
\end{aligned}
\]

(16)

and that leads to

\[
\begin{aligned}
\eta^+ &= \frac{\omega^- - (N-\bar{n}+1)(\omega^+ - \omega^-) + \sum_{i=\bar{n}}^{N} \nu^{(i)}}{1 + 2(N-\bar{n}+1)}, \\
\eta^- &= \nu^- + \omega^+ - \omega^-.
\end{aligned}
\]

(17)

The above result is obtained from arguments closed than the ones derived in [31] [Proposition 5] for an epigraphical constraint of the form \(\{u, \omega \in \mathbb{R} \times \mathbb{R}^N | \max \{\tau_i |u^{(i)}| , \ldots, \tau_N |u^{(N)}|\} \leq \omega \} \) where \((\tau_i)_{1 \leq i \leq N}\) denotes positive weights.

Next proposition is a preliminary result to derive \(P_{E_i}\).

**Proposition 3.** Let \(E_i^+ = \{ (\omega^+, \omega^-, \nu) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N | \|u\|_1 = \omega^+ + \omega^- + \nu \geq 0 \}\). The projection of \((\omega^+, \omega^-, \nu) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N\) on the epigraphic set \(E_i^+\) reads

\[
(\eta^+, \eta^-, \nu) = P_{E_i^+} (\omega^+, \omega^-, \nu)
\]

with

\[
\begin{aligned}
\eta^- &= \frac{\nu^{(i)} - \omega^-}{i + (N-i+1)n} \\
\eta^+ &= \nu^+ - \omega^+ \\
p &= u - \max \{0, \sum_{i=1}^{N} \nu^{(i)}(\eta^+ + \eta^-) \}
\end{aligned}
\]

(18)

with

\[
\bar{n} = \max\{n \in \{1, \ldots, N\} | \mu^{(n)} - \sum_{i=1}^{n} \nu^{(i)}(\eta^+ + \eta^-) > 0\}
\]

(19)

and where \(\mu = (\mu^{(i)})_{1 \leq i \leq N}\) is an ordered version of \(u = (u^{(i)})_{1 \leq i \leq N}\) in an descending order.

**Proof.** The Lagrangian associated to the function involved in \(P_{E_i^+}\) is:

\[
L(\eta^+, \eta^-, \nu, \alpha, \xi) = \frac{1}{2} \|p - u\|^2 + \frac{1}{2} (\eta^+ - \omega^+)^2 \\
+ \frac{1}{2} (\eta^- - \omega^-)^2 + \alpha \left( \sum_{i=1}^{N} p^{(i)} - \eta^+ - \eta^- \right) - \xi^T p.
\]

(20)

The KKT conditions are:

\[
\begin{aligned}
& \quad \langle \nu \in \{1, \ldots, N\} \rangle \quad \nu^{(i)} - u^{(i)} + \alpha - \xi_i = 0 \\
& \quad \eta^+ - \omega^+ + \alpha = 0 \\
& \quad \eta^- - \omega^- - \alpha = 0 \\
& \quad \nu \in \{1, \ldots, N\} \quad \nu^{(i)}p^{(i)} = 0 \\
& \quad \sum_{i=1}^{N} p^{(i)} - \eta^+ - \eta^- = 0.
\end{aligned}
\]

(21)
From the fourth condition in Eq. (21), we know that if $p^{(i)} > 0$ then $\xi^{(i)} = 0$ and $p^{(i)} = u^{(i)} - \alpha$. Thus, if we denote $\pi = (\pi^{(i)})_{1 \leq i \leq N}$ is an ordered version of $(p^{(i)})_{1 \leq i \leq N}$ in a decreasing order, we can write

$$N \sum_{i=1}^{N} p^{(i)} = \sum_{i=1}^{\bar{n}} \pi^{(i)} = \sum_{i=1}^{\bar{n}} (\mu^{(i)} - \alpha) = \eta^+ + \eta^-.$$ \hfill (22)

From similar arguments than in [32][Lemma 2] , $\bar{n}$ is computed as [19] and thus

$$\alpha = \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} \mu^{(i)} - (\eta^+ + \eta^-).$$ \hfill (23)

From the second and third equality of the KKT conditions, we have $\eta^+ = \eta^- + \omega^+ - \omega^-$. Finally, combining (23) with $\eta^- = \omega^- + \alpha$, yields to

$$\eta^- = \frac{1}{\bar{n}(1 + 2/\bar{n})} \left( \sum_{i=1}^{\bar{n}} \mu^{(i)} - \omega^+ + (\bar{n} + 1)\omega^- \right).$$ \hfill (24)

Next we get the solution of $P_{E_1}$ from the ones of projection to $E_1^+$ according to [32][Lemma 3] and have the following proposition:

**Proposition 4.** The projection of $(u^+, u^-, w) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N$ on the epigraphic set $E_1$ reads

$$(\eta^+, \eta^-, \hat{p}) = P_{E_1}(u^+, u^-, w)$$

with

$$\hat{p} = \text{sign}(w) \hat{p}$$

where $\text{abs}(-)$ and $\text{sign}(-)$ denote the componentwise absolute value and the componentwise sign.

This result can be directly derived from [32][Lemma 3] and the proof is therefore omitted here.

Integrating the projection derived in Proposition 2 into Algorithm 2 or Proposition 3 and Proposition 4 into Algorithm 3 with

$$\text{prox}_{\sigma E_1}(u^+, u^-, w) = (u^+, u^-, w) - \sigma P_{E_1}(\frac{u^+}{\sigma}, \frac{u^-}{\sigma}, \frac{w}{\sigma})$$

insure the convergence to a mimimizer of (6) and, from Proposition 1 to a mimimizer of (5), respectively for weak hierarchy and strong hierarchy.

**IV. EXPERIMENTS**

In this section, we provide comparison results of both $\ell_1$ and $\ell_\infty$ proposed approaches with the two closest state-of-the-art procedures, that are “hierNet” [11] and Jenatton’s framework [21]. Comparisons are of two types, comparisons in terms of convergence behavior and in term of performing estimation.

**A. Simulated data**

1) *Dataset:* The dataset is created according to [5]. It is initially composed of $N$ main features and of $N(N-1)/2$ interactions (due to symmetry). We denote $(\mathbf{v}, \Theta) \in \mathbb{R}^N \times \mathbb{R}^{N \times N}$
the ground truth generated according to strong hierarchy. $\mathbf{v}$ denotes a sparse vector where the non-zero values are associated to randomly selected indexes. The value for the non-zero $\mathbf{v}^{(i)}$ is randomly selected from the set $\{-5, -4, \ldots, -1, 1, \ldots, 5\}$. Due to strong hierarchy constraint, $\mathbf{e}_i$ is only non-zero for those whose main effects are not zero. The values are randomly chosen from the set $\{-10, -8, \ldots, -2, 2, \ldots, 8, 10\}$.

The algorithmic performance is evaluated on two simulated datasets. The first one is composed with $N = 30$ features, the first 10 main features are non-zeros and the interaction ratio $\rho = 100$ (Dataset30-345). The second dataset is of size $N = 100$, the first 30 features are non-zeros and the interaction ratio is $\rho = 30\%$ (Dataset100-030).

The datasets are composed with a training, a validation and a testing set with 100 samples each. $x_t$ is randomly generated according to normal distribution $\mathcal{N}(0, I_N)$. The observation value for each sample is set by $y_t = x_t^\top \mathbf{v} + x_t^\top \mathbf{e}_t + \epsilon_t$, where $\epsilon_t$ is independent Gaussian noise to make the signal-to-noise ratio approximately equals to 5dB.

2) Comparison with $\ell^r$ - $r = \infty$ and weak hierarchy: In order to provide fair comparison with the work in [21], we first investigate the performance of proposed primal-dual proximal algorithm (Algorithm 2) when $r = \infty$ and in the configuration of weak hierarchy (no symmetry constraint is considered, i.e. $C = \mathbb{R}^{N \times N}$); “Weak-PD-$\ell^\infty$”. The algorithmic procedure designed in [21] is based on “FISTA” and it will be named “Weak-FISTA-$\ell^\infty$”.

The comparisons are displayed in Fig. 2 for Dataset30-345 and in Fig. 3 for Dataset100-030. The convergence behaviour are presented both w.r.t iteration number (first row) and time (second row). It is observed that:

i) “Weak-PD-$\ell^\infty$” and “Weak-FISTA-$\ell^\infty$” converge to the same value as shown in the first column of both Fig. 2 and 3.

ii) From the objective function convergence point-of-view, “Weak-FISTA-$\ell^\infty$” converges faster than “Weak-PD-$\ell^\infty$” w.r.t iteration numbers and also time.

iii) From the convergence of the iterates point of view (second column of Fig. 2 and 3), “Weak-PD-$\ell^\infty$” converges much faster than “Weak-FISTA-$\ell^\infty$” to the optimal solution, especially from the comparison in terms of time (bottom figures). It shows that the proposed algorithm enables to be closest to the optimal solution with less time;

iv) The conclusions are very close for both datasets (Dataset30-345 and Dataset100-030) leading to the conclusion that the proposed method is robust to dimensionality.

v) Contrary to “Weak-FISTA-$\ell^\infty$”, the proposed solution has a “Strong-PD-$\ell^\infty$” counterpart without inner iterations.

3) Comparison with $\ell^r$ - $r = 1$ and strong hierarchy: Similar experiments are conducted for $r = 1$ and strong hierarchy (i.e., $C = S$). The proposed algorithm (Algorithm 3 with $r = 1$) is called “Strong-PD-$\ell^1$” and it is compared with “hierNet” whose iterations are derived from an ADMM scheme, named “Strong-ADMM-$\ell^1$”, in our experiments.

The comparisons between these two schemes are displayed in Fig. 4 (for Dataset30-345) and Fig. 5 (for Dataset100-030). Both are compared in three respects: the convergence of the objective function $f$, the convergence of the iterates (i.e. $\|w_t^k - w_t^{[\infty]}\|$), and the distance to the set $S$. The evolutions are displayed w.r.t iteration number and time. It is observed that:

i) “Strong-PD-$\ell^1$” and “Strong-ADMM-$\ell^1$” converge to the same solution but the convergence of the objective to the optimum solution is very sensitive to the trade-off hyper-parameter for “Strong-ADMM-$\ell^1$”.

ii) “Strong-PD-$\ell^1$” is always faster either in iteration or
Fig. 4. Comparison between the proposed “Strong-PD-ℓ₁” and “Strong-ADMM-ℓ₁” on Dataset30-345 for λ = 8. (top-left) Objective function (5) w.r.t iterations, (bottom-left) Objective function (5) w.r.t time, (top-middle) Distance ∥w[k] − w[∞]∥ w.r.t iterations, (bottom-middle) Distance ∥w[k] − w[∞]∥ w.r.t time, (top-right) Distance to the strong hierarchy constraint S w.r.t iterations, (bottom-right) Distance to the strong hierarchy constraint S w.r.t. time.

Fig. 5. Comparison between the proposed “Strong-PD-ℓ₁” and “Strong-ADMM-ℓ₁” on Dataset100-030 for λ = 10. (top-left) Objective function (5) w.r.t iterations, (bottom-left) Objective function (5) w.r.t time, (top-middle) Distance ∥w[k] − w[∞]∥ w.r.t iterations, (bottom-middle) Distance ∥w[k] − w[∞]∥ w.r.t time, (top-right) Distance to the strong hierarchy constraint S w.r.t iterations, (bottom-right) Distance to the strong hierarchy constraint S w.r.t. time.

in time and either in terms of functional or in terms of iterations. The explanation mainly comes from the inner iterations required with “Strong-ADMM-ℓ₁” when dealing with C = S;

iii) Third column of Fig. 4 and 5 highlights that the constraint violations (i.e. distance to S) with the proposed method is always smaller than with ADMM.

4) Discussion regarding the choice of r: From the above algorithmic comparisons we have observed that the proposed method either for ℓ₁ or ℓ∞ deliver an accurate solution (as other algorithmic strategies) but faster and with the possibility to include the strong-hierarchy constraint without inner iterations. In the following experiments, we will thus focus on comparisons between weak/strong ℓ₁ or ℓ∞ using the proposed algorithm (“Weak-PD-ℓ₁”, “Weak-PD-ℓ∞”, “Strong-PD-ℓ₁”, “Strong-PD-ℓ∞”). We perform 20 simulations and the average
performance with the associated variances are presented in Fig. 6 for different values of the regularization parameter \( \lambda \). Moreover, the performance obtained by cross-validation are presented Table 1.

i) The good behaviour of the strong hierarchy constraint clearly appear for both datasets. Indeed, we recall that the data have been created with strong hierarchy structure and we can clearly observe that for both datasets the performance with the strong hierarchy constraint are always better either for \( r = 1 \) or \( r = +\infty \).

ii) In our set of experiments, the regularization with \( r = 1 \) always leads to better performance than with \( r = +\infty \).

iii) MSE values associated with the Dataset100-030 are larger, probably due to overfitting. This assumption can be validated by the results obtained on the training dataset.

B. Application to HIV Data

In this section, we apply the proposed algorithms to HIV dataset on the susceptibility of the HIV-1 virus to six nucleoside reverse transcriptase inhibitors (NRTIs). This dataset is collected by [33] and it is used to model HIV-1 susceptibility to the drugs because HIV-1 virus can become resistant through genome mutations for different subjects. In the dataset, there are 639 subjects and 240 genomic locations. For each observation, the mutation status at each genomic location is recorded and also give six drugs (log) susceptibility responses. In this work, we focus on the prediction model for 3TC drug.

1) Reduced dataset: Followed by [5], we first work on a reduced dataset over the bins of ten adjacent loci, rather than all 240 genomic locations, resulting from the fact that nearby genomes have similar effects to the drug susceptibility, and also leading to less sparse data. So for the first set of experiments we have \( N = 24 \) features. The value for each bin is set to 1 if one of genomes in that bin occurs mutation.

The dataset is randomly split into two half sets for training and test respectively as adopted by [1], [5]. We report the average performance of MSE on the test set over 5 splits under different \( \lambda \) for “Weak-PD-\( \ell_\infty \)”, “Strong-PD-\( \ell_\infty \)”, “Weak-PD-\( \ell_1 \)” and “Strong-PD-\( \ell_1 \)” in Fig. 7 (left).

It can be observed that “Strong-PD-\( \ell_\infty \)” and “Strong-PD-\( \ell_1 \)” are slightly better than their weak counterparts, and the best \( \lambda \) for “Strong-PD-\( \ell_\infty \)” and “Strong-PD-\( \ell_1 \)” are 12 and 10 respectively. The interactions for one realization are visualized in the Fig. 8 (left). In order to better visualize the effect of the penalizations, we also plot the \( \ell_\infty \)-norm and \( \ell_1 \)-norm over each row of \( \Theta \) for strong hierarchy, as shown in the middle and right of Fig. 8. Both have the strong effect for \( 1^{th} \) feature, which is consistent with the results observed in [5]. We also observe an interesting fact from the visualization of interactions, it can be seen that “Strong-PD-\( \ell_1 \)” is able to learn a sparser interactions matrix than “Strong-PD-\( \ell_\infty \)”, however, the maximum strength (i.e \( \ell_\infty \)-norm) and the sum of strength (\( \ell_1 \)-norm) in Fig. 8 have similar distributions for both cases. The possible reason is that the extra interactions learned by “Strong-PD-\( \ell_\infty \)” are subtle, it can be also confirmed both performance are very similar in the Fig. 7 (left).

2) Full dataset: We further work on the full data without binning operation leading to \( N = 240 \) features and each feature gives the indicator of mutation. The data splitting is the same as the one previously described. The average performance of MSE on the test set under four situations are shown in Fig. 7 (right). It is clear to demonstrate that strong hierarchy obtains better performance than weak one. The learned interactions and their feature effects from “Strong-PD-\( \ell_\infty \)” and “Strong-PD-\( \ell_1 \)” with best \( \lambda = 1 \) are shown in Fig. 9. It is found that both algorithms can detect that 184th genomic location has the most strong effect. From the maximum strength and sum of strength for each gene in Fig. 9 we can see the different properties for both algorithms. For “Strong-PD-\( \ell_\infty \)”, maximum strength for each gene has a more sparse distribution than sum of strength, which is consistent with its objective that the maximum of absolute values of \( \Theta(i,:i) \) is not larger that \( v(i) \), whereas a more sparse distribution over sum of strength is obtained for “Strong-PD-\( \ell_1 \)”, whose objective is to ensure that the sum of absolute values of \( \Theta(i,:i) \) is not larger that \( v(i) \). From Fig. 7 (right), it also can be observed that “Strong-PD-\( \ell_1 \)” is slightly better than “Strong-PD-\( \ell_\infty \)” in that it imposes a stronger penalization for \( \Theta \) than “Strong-PD-\( \ell_\infty \)”.

V. CONCLUSION

In this paper, we propose a primal-dual proximal algorithm based on epigraphical projection to optimize the sparse regression learning problem when hierarchical constraints are made between the main effects and the interactions. Four algorithms are derived (“Weak-PD-\( \ell_\infty \)” , “Weak-PD-\( \ell_1 \)” , “Strong-PD-\( \ell_\infty \)” , “Strong-PD-\( \ell_1 \)” ) allowing to deal and compare properly.
different configurations of the regularization term (4). Compared to other state-of-the-art algorithms, for instance, FISTA or ADMM, the proposed algorithm is computationally more efficient (find a solution belonging to $S$, convergence in terms of iterates is faster). The experiments on the synthetic and real data validate its effectiveness and efficiency.

ACKNOWLEDGMENT

The authors would like to thank Marie-France Benassy, Mickael Bonnard, Laurent Grosset, Georges Oppenheim, Maxime Durot, Leire Oro-Urea from TOTAL for helpful discussion and Prof. Lin Qi from Zhengzhou University for providing technical support to this work.

REFERENCES

[1] J. Bien, J. Taylor, and R. Tibshirani, “A lasso for hierarchical interactions,” *Annals of Statistics*, vol. 41, no. 3, pp. 1111–1141, 2013.
[2] A. Pirayre, C. Couprie, F. Bidard, L. Duval, and J.-C. Pesquet, “BRANE Cut: Biologically-related apriori network enhancement with graph cuts for gene regulatory network inference,” *BMC Bioinformatics*, vol. 16, no. 1, 2015.
[3] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
[4] P. Zhao, G. Rocha, and B. Yu, “The composite absolute penalties family for groupes and hierarchical variable selection,” *The Annals of Statistics*, vol. 37, no. 6A, pp. 3468–3497, 2009.
[5] A. Haris, D. Witten, and N. Simon, “Convex modeling of interactions with strong heredity,” *Journal of Computational and Graphical Statistics*, pp. 1–31, 2014.
[6] M. Lim and T. Hastie, “Learning interactions through hierarchical group-lasso regularization,” *Journal of Computational and Graphical Statistics*, vol. 24, pp. 627–654, 2015.
[7] Y. She and H. Jiang, “Group regularized estimation under structural hierarchy,” *Journal of the American Statistical Association*, 2016.
[8] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, “Use of the zero-norm with linear models and kernel methods,” *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.
[9] R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society, Series B, vol. 58, pp. 267–288, 1994.
[10] H. Zou and M. Yuan, “The F∞-norm support vector machine,” Statistica Sinica, vol. 18, pp. 379–398, 2008.
[11] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, “Structured sparsity through convex optimization,” Statistical Science, vol. 27, no. 4, pp. 450–468, 2012.
[12] ——, “Optimization with sparsity-inducing penalties,” Foundations and Trends in Machine Learning, vol. 4, no. 1, pp. 1–106, 2012.
[13] J. Gui, Z. Sun, S. Ji, S. Member, D. Tao, and T. Tan, “Feature selection based on structured sparsity: a comprehensive study,” IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 7, pp. 1828–1833, 2015.
[14] J. Xu, B. Tang, H. He, and H. Man, “Semisupervised feature selection based on relevance and redundancy criteria,” IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 9, pp. 1974–1984, 2017.
[15] Q. Zhang, X. Hu, and B. Zhang, “Comparison of ℓ1-norm SVR and sparse coding algorithms for linear regression,” IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 8, pp. 1118–1130, 2014.
[16] R. Chakraborty and N. R. Pal, “Feature selection using a neural framework with controlled redundancy,” IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 1, pp. 35–50, 2015.
[17] J. A. Anderson, “Quadratic logistic discrimination,” Biometrika, vol. 62, pp. 149–154, 1975.
[18] H. R. Ronald, D. B. James, S. R. John, and V. H. Robert, “Generalized linear and quadratic discriminant functions using robust estimates,” Journal of the American Statistical Association, vol. 73, pp. 564–568, 1978.
[19] R. Chambolle and N. R. Pal, “Feature selection using a neural network framework with controlled redundancy,” IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 1, pp. 35–50, 2015.
[20] H. He, R. Tibshirani, and M. Yuan, “Epigraphical projection and proximal tools for solving constrained image recovery problems,” Journal of Mathematical Imaging and Vision, vol. 40, no. 1, pp. 120–145, 2011.
[21] A. Chambolle and P. Combettes, “Convex analysis and monotone operator theory in hilbert spaces,” Springer, New York, 2011.
[22] A. Chambolle and C. Dossal, “On the convergence of the iterates of the fast iterative shrinkage-thresholding algorithm,” Journal of Optimization Theory and Applications, vol. 166, no. 3, pp. 968–982, Sep. 2015.
[23] S. Setzer, Split Bregman algorithm, Douglas-Rachford splitting and frame shrinkage, 2009, vol. 5567, ch. Scale Space and Variational Methods in Computer Vision. SSVM 2009. Lecture Notes in Computer Science, pp. 464–476.
[24] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” Journal of Mathematical Imaging and Vision, vol. 40, no. 1, pp. 120–145, 2011.
[25] N. Komodakis and J. C. Pesquet, “Playing with duality: an overview of recent primal-dual approaches for solving large-scale optimization problems,” IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 31–54, Nov 2015.
[26] L. Condat, “A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms,” Journal of Optimization Theory and Applications, vol. 158, no. 2, pp. 460–479, 2013.
[27] B. C. Vu, “A splitting algorithm for dual monotone inclusions involving cocoercive operators,” Advances in Computational Mathematics, vol. 38, no. 3, pp. 667–681, Apr. 2013.
[28] M. Chierchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu, “Epigraphical projection and proximal tools for solving constrained convex optimization problems,” Signal, Image and Video Processing, vol. 9, no. 8, pp. 1737–1749, 2015.
[29] J. C. Duchi, S.-S. Shai, Y. Singer, and T. Chandra, “Efficient projections onto the l1-ball for learning in high dimensions,” in ICML, ser. ACM International Conference Proceeding Series, vol. 307, 2008, pp. 272–279.