

Multi-level projection with exponential parallel speedup; Application to sparse auto-encoders neural networks

Guillaume Perez and Michel Barlaud
I3S Laboratory, University of Cote d’Azur , France

May 6, 2024

Keywords— Bi-level Projection, Structured sparsity, low computational complexity, exponential parallel speedup

Abstract

The $\ell_{1,\infty}$ norm is an efficient structured projection but the complexity of the best algorithm is unfortunately $\mathcal{O}(nm \log(nm))$ for a matrix in $\mathbb{R}^{n \times m}$. In this paper, we propose a new bi-level projection method for which we show that the time complexity for the $\ell_{1,\infty}$ norm is only $\mathcal{O}(nm)$ for a matrix in $\mathbb{R}^{n \times m}$, and $\mathcal{O}(n+m)$ with full parallel power. We generalize our method to tensors and we propose a new multi-level projection, having an induced decomposition that yields a linear parallel speedup up to an exponential speedup factor, resulting in a time complexity lower-bounded by the sum of the dimensions. Experiments show that our bi-level $\ell_{1,\infty}$ projection is 2.5 times faster than the actual fastest algorithm provided by *Chu et. al.* while providing same accuracy and better sparsity in neural networks applications.

1 Introduction

1.1 Motivations and related methods

Sparsity requirement appears in many machine learning applications, such as the identification of biomarkers in biology [1, 2] or the recovery of sparse signals in compressed sensing [3, 4]. It is well known that the impressive performance of neural networks is achieved at the cost of a high-processing complexity and large memory requirement. Recently, advances in sparse recovery and deep learning have shown that training neural networks with sparse weights not only improves the processing time, but most importantly improves the robustness and test accuracy of the learned models [5, 6, 7, 8, 9][10, 11]. Regularizing techniques have been proposed to sparsify neural networks, such as the popular *LASSO* method [12, 13]. The *LASSO* considers the ℓ_1 norm as regularization. Note that projecting onto the ℓ_1 norm ball is of linear-time complexity [14, 15]. Unfortunately, these methods generally produce sparse weight

matrices, but this sparsity is not structured and thus is not computationally processing efficient.

Let's consider again the sparsity requirement. In neural networks processing, a lot of hardware can compute multiply-add operations in a single instruction. We therefore require low MACCs (multiply-accumulate operations) as computational cost (one multiplication and one addition are counted as a single instruction). Thus a structured sparsity is required (i.e. a sparsity able to set a whole set of columns to zero). To address this issue, extension of the ℓ_1 to $\ell_{1,q}$ norms with $q \in 1, 2, \infty$ promotes group sparsity, i.e., the solution is not only sparse but group variables can be removed simultaneously. Group-LASSO originally proposed in [16], was used in order to sparsify neural networks without loss of performance [17, 18, 19]. Unfortunately, the classical Group-LASSO algorithm is based on Block coordinate descent [20, 21] which requires high computational cost. Note that the $\ell_{1,1}$ constraint was also used in order to sparsify convolutional neural networks [22].

The $\ell_{1,\infty}$ norm is of particular interest because it is able to set a whole set of columns to zero, instead of spreading zeros as done by the ℓ_1 norm. This makes it particularly interesting for reducing computational cost. Many projection algorithms were proposed [23, 24, 25]. However complexity of these algorithms remains an issue. The worst-case time complexity of this algorithm is $\mathcal{O}(nm * \log(nm))$ for a matrix in $\mathbb{R}^{n \times m}$. This complexity is an issue, and to the best of our knowledge, no current publication reports the use of the $\ell_{1,\infty}$ projection for sparsifying large neural networks.

1.2 Contributions

In this paper we focus on the computational issue for the $\ell_{1,\infty}$ norm. First, in order to cope with this issue, we propose a new bi-level projection method. The main motivation for this work is the direct independent splitting made by the bi-level optimization which take into account the structured sparsity requirement. Second, we generalize the bi-level projection to multi-level, and shows its theoretical guaranties, such as an exponential parallel speedup. The paper is organized as follows. First we present our general bi-level projection framework using our splitting approach 3. Then, we provide in section 4 the application to the bi-level $\ell_{1,\infty}$ projection. In section 5, we apply our bi-level framework to well known constraints such as $\ell_{1,1}$ constraint and to any combination of p, q norm such as the group-Lasso and the exclusive LASSO. We extend our bi-level projection method to tensors using multi-level projections in section 6. In Section 7, we finally compare different projection methods experimentally. Our experimental section is split in two parts. First, we provide an empirical analysis of the projection algorithms onto the bi-level projection $\ell_{1,\infty}$ ball. This section shows the benefit of the proposed method, especially in the context of sparsity. Second, we apply our framework to the diagnosis (or classification) using a supervised autoencoder on a synthetic dataset and a biological dataset.

2 Definitions and Notations

In this paper we use the following notations: lowercase Greek symbol for scalars, scalar i, j, c, m, n are indices of vectors and matrices, lowercase for vectors, capital for matrices, and calligraphic for tensors. For every $p, q \in \mathbb{R}$, the $\ell_{p,q}$ norm of a real matrix

$X = [x_1 \ \cdots \ x_m] \in \mathbb{R}^{n \times m}$ with columns x_j and elements $X_{i,j}$ is given by:

$$\|X\|_{p,q} := \left(\sum_{j=1}^m \|x_j\|_q^p \right)^{\frac{1}{p}}, \quad (1)$$

where the ℓ_q norm of the column vector $x_j \in \mathbb{R}^n$ is:

$$\|x_j\|_q := \left(\sum_{i=1}^n |X_{i,j}|^q \right)^{\frac{1}{q}}. \quad (2)$$

Given a real value η , we denote by $\mathcal{B}_\eta^p$ the ball of radius η for the norm p :

$$\mathcal{B}_\eta^p = \{X \in \mathbb{R}^n \mid \|X\|_p \leq \eta\}, \quad (3)$$

Using the definition of the ball, the euclidean projection is defined by:

$$P_\eta^{p,q}(Y) = \arg \min_{X \in \mathcal{B}_\eta^{p,q}} \|X - Y\|_2 \quad (4)$$

In the general case, p and q are set and dedicated algorithms are used to process the projection.

3 A new bi-level projection formulation

Processing the projection of matrices or other mathematical objects, such as tensors, is becoming harder and harder as dedicated algorithms satisfying equation (4) must be defined for each possible combination of p and q . of equation (4) and to reformulate the projection as a bi-level problem. In this reformulated problem, the $\ell_{p,q}$ projection will be split into an aggregation using the q norm followed by a simpler and well defined projection onto the p norm of the aggregated vector. Let $v_q = (\|Y_1\|_q, \dots, \|Y_n\|_q)$ be the vector composed of the q norms of the columns of a matrix.

Given a real positive value η . The $\ell_{p,q}$ bi-level projection optimization problem is defined by:

$$\begin{aligned} BP_\eta^{p,q}(Y) = \{X \mid \forall i, x_i = \arg \min_{x_i \in \mathcal{B}_{\hat{u}_i}^q} \|X_i - Y_i\|_2 \\ \text{such that } \hat{u} \in \arg \min_{u \in \mathcal{B}_\eta^p} \|u - v_q\|_2\} \end{aligned} \quad (5)$$

This problem is composed of two simpler problems. The first one, the most inner one is:

$$\hat{u} \in \arg \min_{u \in \mathcal{B}_\eta^p} \|u - v_q\|_2 \quad (6)$$

Once the columns of the matrix have been aggregated to a vector v_q using the q norm, the problem becomes a usual ℓ_p ball projection problem. Such problems are solved by definition by:

$$u \leftarrow P_\eta^p((\|Y_1\|_q, \dots, \|Y_n\|_q)) \quad (7)$$

It is interesting to note that for some values of p , projection algorithms have been defined already. For example, linear algorithms exist for the ℓ_1 and its weighted version ℓ_{w1} , the ℓ_2 and ℓ_∞ . In addition all of them are strongly convex, which make the solution unique. Note that it is not necessarily the case for any p .

Then, the second part of the bi-level optimization problem, once vector $\hat{u}$ is known is given by:

$$x_i = \min_{x_i \in \mathcal{B}_{\hat{u}_i}^q} \|x_i - y_i\|_2 \quad (8)$$

For each column i of the original matrix. Here, for each x_i , an independent optimization problem is defined. Each of them is optimally solved by definition of the projection on the ℓ_q ball:

$$x_i \leftarrow P_{u_i}^q(y_i) \forall i \in 1, \dots, n \quad (9)$$

The following algorithm is a possible implementation.

Algorithm 1 Bi-level $\ell_{p,q}$ projection ($BP_{\eta}^{p,q}(Y)$).

Input: Y, η
 $u \leftarrow P_{\eta}^p((\|y_1\|_q, \|y_2\|_q, \dots, \|y_n\|_q))$
for $i \in [1, \dots, n]$ **do**
 $x_i \leftarrow P_{u_i}^q(y_i)$
end for
Output: X

Here again, the projection onto the ℓ_q should be known and well defined. It is important to remark that even if the projection is not the closest from a Euclidean point of view, the resulting matrix satisfies the constraint (i.e. $X \in \mathcal{B}_{\eta}^{p,q}$). In the following sections provide our bi-level projections and compare with the existing projections in the literature. It is important to note that we focused on norms providing structured sparsity (removing columns), namely bi-level $\ell_{1,\infty}$, bi-level $\ell_{1,1}$ and bi-level $\ell_{1,2}$. Yet, this is not exhaustive as more norms using different values for p and q can be defined.

4 Bi-level $\ell_{1,\infty}$ projection

4.1 A new bi-level projection

The $\ell_{1,\infty}$ ball projection has gained a lot of interest in the recent years. The main reason being its efficiency to enforce sparsity in the weights of neural networks, while keeping high accuracy.[23, 26, 27, 25] and the classical approach is given as follows. Let $Y \in \mathbb{R}^{n \times m}$ be a real matrix of dimensions $m \geq 1$, $n \geq 1$, with elements $Y_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, m$. The $\ell_{1,\infty}$ norm of Y is

$$\|Y\|_{1,\infty} := \sum_{j=1}^m \max_{i=1,\dots,n} |Y_{i,j}|. \quad (10)$$

Given a radius $\eta \geq 0$, the goal is to project Y onto the $\ell_{1,\infty}$ norm ball of radius C , denoted by

$$\mathcal{B}_{1,\infty}^{\eta} := \left\{ X \in \mathbb{R}^{n \times m} : \|X\|_{1,\infty} \leq \eta \right\}. \quad (11)$$

The projection $P_{\mathcal{B}_{1,\infty}^{\eta}}$ onto $\mathcal{B}_{1,\infty}^{\eta}$ is given by:

$$P_{\mathcal{B}_{1,\infty}^{\eta}} = \arg \min_{X \in \mathcal{B}_{1,\infty}^{\eta}} \frac{1}{2} \|X - Y\|_F^2 \quad (12)$$

Synthetic	(bi-level/usual) $\ell_{1,2}$	$\ell_{1,1}$	bi-level $\ell_{1,1}$	$\ell_{1,\infty}$	bi-level $\ell_{1,\infty}$
Complexity	$O(mn)$	$O(mn)$	$O(mn)$	$O(nm \log(nm))$	$O(nm)$
LP Complexity	$O(m+n)$	$O(mn)$	$O(m+n)$	$O(nm \log(nm))$	$O(n+m)$

Table 1: **Theoretical complexity** Complexity assuming the ℓ_1 and ℓ_2 projection costs are linear [30] and complexity of $\ell_{1,\infty}$ is $O(nm \log(nm))$ [23]. LP (i.e. longest path) complexity is the complexity of the longest path in the computational graph of the bi-level projection. This is the complexity limit of a generic parallel implementation.

where $\|\cdot\|_F = \|\cdot\|_{2,2}$ is the Frobenius norm.

In this paper we propose the following alternative new bi-level method. Recall that $v_\infty = (\|Y_1\|_\infty, \dots, \|Y_n\|_\infty)$ is the vector composed of the infinity norms of the columns of matrix Y . Let the infinity norm projection $P_{u_i}^\infty(y) = (\min(y_i, u_i), \forall y_i \in y)$. The bi-level $\ell_{1,\infty}$ projection optimization problem is defined by:

$$BP_\eta^{1,\infty}(Y) = \{X | \forall i, X_i = \arg \min_{X_i \in \mathcal{B}_{\hat{u}_i}^\infty} \|X_i - Y_i\|_2$$

$$\text{such that } \hat{u} \in \arg \min_{u \in \mathcal{B}_\eta^1} \|u - v_\infty\|_2\} \quad (13)$$

Algorithm 2 is a possible implementation. It is important to remark that usual bi-level optimization requires many iterations [28, 29] while our model reaches the optimum in one iteration.

Algorithm 2 Bi-level $\ell_{1,\infty}$ projection ($BP_\eta^{1,\infty}(Y)$).

Input: Y, η
 $u \leftarrow P_\eta^1((\|y_1\|_\infty, \|y_2\|_\infty, \dots, \|y_n\|_\infty))$
for $i \in [1, \dots, n]$ **do**
 $x_i \leftarrow P_{u_i}^\infty(y_i)$
end for
Output: X

4.2 Computational complexity

The best computational complexity of the projection of a matrix in $\mathbb{R}^{nm}$ onto the $\ell_{1,\infty}$ ball is usually $O(nm \log(nm))$ [23, 25]. The computational complexity of the bi-level projection here is $O(nm)$. Indeed, consider Algorithm 2. Step 1) complexity is $O(nm)$, step 2) complexity is $O(n)$ [14, 30], step 3) complexity is $O(n)$, and step 4 complexity is $O(nm)$. Moreover, the bi-level projection explicit independent processing. While the time complexity without any parallel processing will remain the same, the time complexity with a full parallel power is only $O(n+m)$ as steps 1) and 4) can be run in a parallel.

5 Extension to other norms

5.1 Extension to the $\ell_{1,1}$ bi-level projection

The ℓ_1 ball is famous for being robust and yielding sparsity. Nevertheless, its extension to matrix, the $\ell_{1,1}$ does not yield structured sparsity. We propose to define the $\ell_{1,1}$ bi-level optimization problem:

$$\begin{aligned} BP_{\eta}^{1,1}(Y) = \{X | \forall i, X_i = \arg \min_{X_i \in \mathcal{B}_{\hat{u}_i}^1} \|X_i - Y_i\|_2 \\ \text{such that } \hat{u} \in \arg \min_{u \in \mathcal{B}_{\eta}^1} \|u - v_1\|_2\} \end{aligned} \quad (14)$$

This bi-level $\ell_{1,1}$ projection yields structured sparsity. A possible implementation of the bi-level $\ell_{1,1}$ is given in Algorithm 3. This bi-level $\ell_{1,1}$ projection has the same advantage as the bi-level $\ell_{1,\infty}$ which is the induced parallel decomposition leading to a time complexity with a full parallel power of $O(n + m)$.

Algorithm 3 Bi-level $\ell_{1,1}$ projection. ($BP_{\eta}^{1,1}(Y)$)

Input: Y, η
 $u \leftarrow P_{\eta}^1((\|y_1\|_1, \dots, \|y_n\|_1))$
for $i \in [1, \dots, n]$ **do**
 $x_i \leftarrow P_{u_i}^1(y_i)$
end for
Output: X

5.2 Extension to bi-level $\ell_{1,2}$ and $\ell_{2,1}$ projections

Group LASSO and Exclusive LASSO are well-known algorithm with impressive properties for grouping variables and solving inverse problems [16, 31, 32, 33]. For example, the $\ell_{1,2}$ constraint has been proposed in the context of Group Lasso regularization [34], where the main idea was to enforce parameters of different classes to share common features. We propose to consider their bi-level counter-parts (i.e. $\ell_{1,2}$ and $\ell_{2,1}$). Let the $\ell_{1,2}$ bi-level optimization problem be:

$$\begin{aligned} BP_{\eta}^{1,2}(Y) = \{X | \forall i, X_i = \arg \min_{X_i \in \mathcal{B}_{\hat{u}_i}^2} \|X_i - Y_i\|_2 \\ \text{such that } \hat{u} \in \arg \min_{u \in \mathcal{B}_{\eta}^1} \|u - v_2\|_2\} \end{aligned} \quad (15)$$

First, the bi-level projection algorithms for $\ell_{1,2}$ is given by algorithm 4.

It is interesting to note that the bi-level $\ell_{1,2}$ projection and usual Group LASSO formulation are different. The usual Group LASSO formulation is provided in Appendix Section A. For the bi-level $\ell_{2,1}$ projection the algorithm is given in appendix. This second algorithm, and the many others that can be defined using our framework, is considered out of the topic of this paper as we focus on structured sparsity for the columns for our matrix norm experiments. Nevertheless, they can be defined and have the same advantages as any assignment of p and q .

Algorithm 4 Bi-level $\ell_{1,2}$ projection. ($BP_\eta^{1,2}(Y)$)

Input: Y, η
 $u \leftarrow P_\eta^1((\|Y_2\|_2, \dots, \|Y_n\|_2))$
for $i \in [1, \dots, n]$ **do**
 $x_i \leftarrow P_{u_i}^2(y_i) \forall i \in 1, \dots, n$
end for
Output: X

6 Tensor Generalization

6.1 Tri-level Projection Algorithms

Matrices are not the only mathematical objects that are used in neural networks. Images, for examples, are often represented as order 3 tensors in $\mathbb{R}^{c,n,m}$ with c representing the channels of the image. Moreover, the current development of deep-learning framework for image compression are leaning toward tensor efficient regularization methods [35, 36], and not only vectors or matrices methods. For example, the new image compression standard JPEG AI [37] uses tensor representation of images in the latent space of an autoencoder. That is the reason why we propose to generalize the bi-level projection, first to tri-level, then to multi-level. We consider tensors as multi-dimensional arrays with subscripts indexing. First, consider order 3 tensors of the form $\mathcal{Y} \in \mathbb{R}^{c,n,m}$. For example, $\mathcal{Y} \in \mathbb{R}^{3,m,n}$ is used to represent an image of width $m \times n$ and having 3 channels. We define by $\forall i \in \langle c, n, m \rangle$ all the tuples of the Cartesian product of the indices. For example $\forall i \in \langle 3, 256, 256 \rangle \equiv \forall i \in ((1, 1, 1), (1, 1, 2), \dots, (3, 256, 256))$. Given a tensor $\mathcal{Y} \in \mathbb{R}^{c,n,m}$, we denote by $V_q = (\|\mathcal{Y}_{1,1}\|_q, \dots, \|\mathcal{Y}_{n,m}\|_q)$ the matrix of aggregation of the channels by the norm q .

Definition 6.1. Given a tensor $\mathcal{Y} \in \mathbb{R}^{c,n,m}$ and a radius η , the tri-level $\ell_{1,\infty,\infty}$ projection optimization problem is equal to:

$$\begin{aligned} TP_\eta^{1,\infty,\infty}(\mathcal{Y}) = \{ \mathcal{X} | \forall i \in \langle n, m \rangle, \mathcal{X}_i = \\ \arg \min_{\mathcal{X}_i \in \mathcal{B}_{\hat{U}_i}^\infty} \|\mathcal{X}_i - \mathcal{Y}_i\|_2. \end{aligned} \quad (16)$$

such that $\hat{U} \in BP_\eta^{1,\infty}(V_\infty)$

This tri-level projection first uses the ℓ_∞ norm as a multi-channel aggregator. Then, for each column of the resulting matrix, it aggregates again using the ℓ_∞ norm. Finally, the resulting vector is projected onto the ℓ_1 ball. A possible implementation is given in Algorithm 5, where $BP_\eta^{1,\infty}(V_\infty)$ is given in Algorithm 2. First, line 2 shows the aggregation of the tensor into a vector. The infinity norm aggregation is applied to the channels, then the previously defined bi-level $\ell_{1,\infty}$ is applied and stored in U^2 . Finally, for each couple of coordinate (i, j) , the resulting channels is equal to the original channels at the coordinate projected onto the ℓ_∞ ball of radius $U_{i,j}$. An iterative implementation of the algorithm is given in Algorithm 9 in appendix.

6.2 Multi-level Projection Algorithms

As shown in the previous section, the tri-level projection yields a recursive pattern that is generalized in this section.

Algorithm 5 Tri-level projection $\ell_{1,\infty,\infty} \cdot (MP_\eta^{1,\infty,\infty}(\mathcal{Y}))$

- 1: **Input:** $\mathcal{Y} \in \mathbb{R}^{c,n,m}, \eta$
 - 2: $U \leftarrow BP_\eta^{1,\infty}(V_\infty)$ {Algorithm 2}
 - 3: **for** $i \in \langle n, m \rangle$ **do**
 - 4: $x_i \leftarrow P_{U_i}^\infty(y_i)$
 - 5: **end for**
 - 6: **Output:** $\mathcal{X}$
-

We consider tensors as multi-dimensional arrays with subscripts indexing. Let $T = (d_1, \dots, d_r)$ a list of r dimensions, usually called shape. Let $|s|$ and $s_{i:j}$ respectively denote the cardinality and sub-list operator. Let $\mathcal{Y} \in \mathbb{R}^T$ denotes a tensor of order $|T|$. For example if $T = (c, n, m)$, then $\mathcal{Y} \in \mathbb{R}^T$ is an order 3 tensor.

Let $\nu \in \mathcal{N}^{|T|}$ a list of norms design of an order $|T|$ tensor. Let each norm $\nu^i \in \nu$ be of the form $\nu^i = (\nu_1^i, \dots, \nu_j^i)$. Let $\forall i \in \langle T_{k:l} \rangle$ denote all the tuples of the Cartesian product of the indices from the sub-list of dimensions $T_{k:l}$. This is an extension to the the order 3 definition of the indexes enumeration.

Given $\mathcal{Y} \in \mathbb{R}^T$ let $\mathcal{V}_{\nu^i} \in \mathbb{R}^{T_{1:|T|-|\nu^i|}}$ be the aggregation of $\mathcal{Y}$ using the ν^i norm.

Definition 6.2. Given a tensor $\mathcal{Y} \in \mathbb{R}^T$, a list of norms ν , and a radius η , the multi-level projection optimization problem is equal to:

$$\begin{aligned}
 MP_\eta^\nu(\mathcal{Y}) = \{ \mathcal{X} | \forall i \in \langle T_{|\nu_1|:|T|} \rangle, \mathcal{X}_i = \\
 \arg \min_{\mathcal{X}_i \in \mathcal{B}_{\mathcal{U}_t}^{\nu_1}} \|\mathcal{X}_i - \mathcal{Y}_i\|_2. \\
 \text{such that } \hat{\mathcal{U}} \in MP_\eta^{\nu_{2:|\nu|}}(\mathcal{V}_{\nu_1}) \}
 \end{aligned} \tag{17}$$

and $MP_\eta^\nu(\mathcal{Y}) = P_\eta^{\nu_1}(\mathcal{Y})$ if $|\nu| = 1$.

For example $\nu = ((1, \infty))$ is be the usual $\ell_{1,\infty}$ norm while $\nu = ((\infty), (1))$ is be the bi-level $\ell_{1,\infty}$ norm. Finally, if $\nu = ((\infty), (\infty), (1))$ it is the tri-level defined in the previous section.

Proposition 6.3. *The multi-level projection is a generalization of the usual projection.*

Proof is in appendix. In the general case, the computational worst-case time complexity of a projection is lower-bounded by the product of the dimensions of $\mathcal{Y}$ given by $O(\prod_{d \in T_r} d)$. For example, the best known computational complexity of the projection of a matrix in $\mathbb{R}^{n \times m}$ onto the $\ell_{1,\infty}$ ball is usually $O(nm \log(nm))$ [23, 25]. Yet sometimes, for some norms, a parallel version can be defined, or at least with independent sub-parts. Multi-level projection aims at reducing this complexity.

Algorithm 6 is a possible implementation of the multi-level projection. At the first line, the recursive call extracts the multi-level projection of the ν_1 -aggregated tensor onto the list of norms $\nu_{2:|\nu|}$. This list consists of all the norms except the first one. For example, for the tri-level $\nu = ((\infty), (\infty), (1))$, the first line is the bi-level projection of the tensor aggregated using the ∞ norm onto the list of norm $((\infty), (1))$. Then each tuple of indices $t \in \langle T_{|\nu_1|:|T|} \rangle$ if considered. It is the set of indices of all the dimensions except the $|\nu_1|$ first ones. Consider again the tri-level example, the t was among the set of indices $\langle n, m \rangle$. For each of these tuples t , the local projection onto ν_1 of the sub-part $\mathcal{Y}_t$ of the tensor $\mathcal{Y}$ is processed and stored in the sub-part $\mathcal{X}_t$ the tensor $\mathcal{X} \in \mathbb{R}^T$. In the end, $\mathcal{X}$ contained the multi-level projection of $\mathcal{Y}$.

Proposition 6.4. *Using infinite parallel processing power, the lower-bound worst-case time complexity of the multi-level projection is reduced from $O(\prod_{d \in T_r} d)$ to $O(\sum_{d \in T_r} d)$, resulting in an exponential speedup.*

Proof is given in appendix. This implies that with infinite parallel processing power, the time complexity of the aggregations of all the recursive calls is $O(\sum_i f_i(|\nu_i|))$.

Algorithm 6 Multi-level projection ($MP_\eta^\nu(\mathcal{Y})$).

```

1: Input:  $\mathcal{Y}, \nu, \eta$ 
2:  $\mathcal{U} \leftarrow MP_\eta^{\nu_{2:|\nu|}}(\mathcal{V}_{\nu_1})$ 
3: for  $i \in \llbracket T_{|\nu_1|:|T|} \rrbracket$  do
4:    $x_i \leftarrow P_{\mathcal{U}_i}^{\nu_1}(y_i)$ 
5: end for
6: Output:  $\mathcal{X}$ 

```

7 Experimental results

7.1 Benchmark times using Pytorch C++ extension using a Macbook Labtop with with a i9 processor; Comparison with the best actual projection method

This section presents experimental results of the projection operation alone. The experiments were run on laptop with a I9 processor having 32 GB of memory. The state of the art on such is pretty large, starting with [23] who proposed the first algorithm, the Newton-based root-finding method and column elimination method [26, 25], and the recent paper of *Chu et. al.* [27] which outperforms all of the other state-of-the-art methods. We compare our bi-level method against the best actual algorithm proposed by *Chu et. al.* which uses a semi-smooth Newton algorithm for the projection. The Pytorch C++ extension implementation used is the one generously provided by the authors. All other methods usually take order of magnitude more times, hence are not present in most of our figures.

The Pytorch C++ implementation of our bi-level $\ell_{1,\infty}$ method is based on fast ℓ_1 projection algorithms of [14, 15] which are of linear complexity. The code is available online¹.

Figure 1 shows the running time as a function of the radius. The size of the matrices is 1000x10000, values between 0 and 1 uniformly sampled and the radius are in $[0.25, 4]$. As we can see, the running time of our bi-level method is at least 2.5 times faster than the actual fastest method *Chu et al.*. Note that this behaviour is the same for all the $\ell_{1,\infty}$ projection algorithms we tested. The running time of the bi-level one is almost not impacted by the sparsity, which make it more stable in general.

Figure 2 shows the running time as a function of the matrix size. Here the radius has been fixed to $\eta = 1.$. As we can see, the running time of our bilevel method is at least 2.5 faster than the actual fastest method, and this factor remains the same even when increasing the matrix size both in number of columns and rows. In

¹<https://github.com/memo-p/projection>

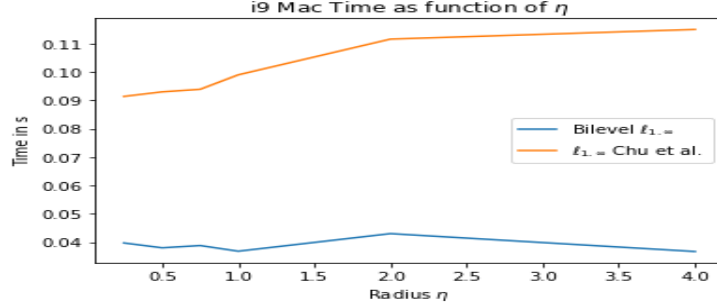


Figure 1: Comparison of our bi-level method with *Chu et al.*. Processing time as a function of the radius (matrice fixed 1000x10000)

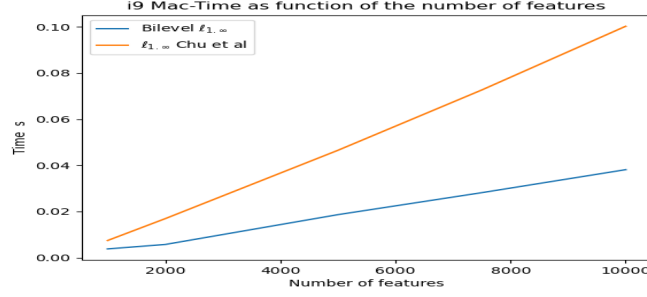


Figure 2: Processing time using C++: our bi-level projection method versus *Chu et. al.* method. $m=1000$ and $\eta = 1$ fixed .

conclusion to this comparison, using the bi-level $\ell_{1,\infty}$ is faster in our experiments and smoother in the selection of a radius. Note that pytorch c++ extension is 20 times faster than the standard pytorch implementation.

Figure 3 shows the running times of our Tri-level algorithm as a function of the tensor size with dimension m for two projections $\ell_{1,1,1}$ and $\ell_{1,\infty,\infty}$. We can see that the running times of both projections are similar and grow linearly with the increase of the dimension m . Note that we have the same behavior for the other dimensions of the tensor.

7.2 Benchmark using C++ parallel implementation

One of the arguments of using the bi-level version of projection instead of the usual one is not only that it can be faster and easier to implement, but also that a native parallel decomposition of the work can be extracted from the computation tree. We implemented the parallel version of the bi-level $\ell_{1,\infty}$ using a basic Thread-pool implementation using native future of C++. These experiments were run on a 12-Core Processor an *AMD Ryzen 9 5900X 12-Core Processor 3.70 GHz* desktop machine having 32 GB of memory. , thus we used 12 as a maximum number of workers. In this run, the compiler

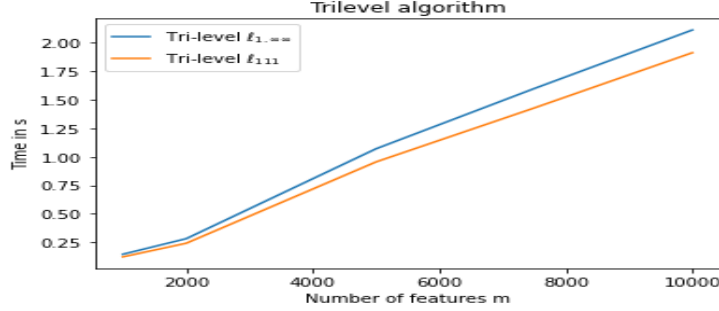


Figure 3: Processing time using C++: our Tri-level projection method $d=32$ and $n=1000$ fixed.

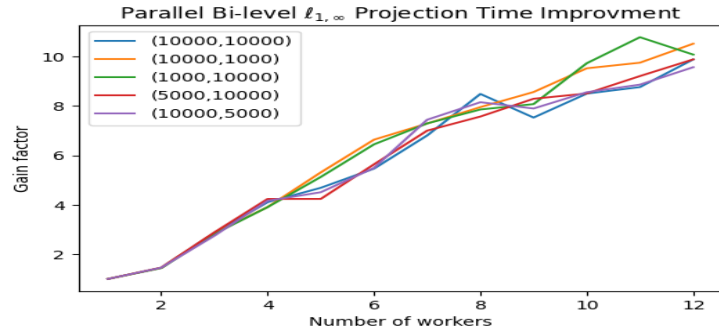


Figure 4: Gain factor of the native parallel workload decomposition for various matrix sizes.

optimization have been deactivated. The reason is that once activated, the best speed factor is 2, larger instances are required to see better speedup, and the time is spent moving memory around and saturating the few memory channels. Yet, future work involving CPU and GPU expert might solve these issues. As shown in Figure 4, the workload is easy to balance between workers by definition of the processing tree. That is the reason why the gain factor grows linearly with the number of workers. This implies that in our computer, our naive parallel implementation using CPU is between 20 to 30 times faster than the best current $\ell_{1,\infty}$ projection algorithm.

7.3 Experimental results on classification using a supervised autoencoder neural network

7.3.1 Supervised Autoencoder (SAE) framework

Autoencoders were introduced within the field of neural networks decades ago, their most efficient application being dimensionality reduction [38, 35]. Autoencoders were used in application ranging from unsupervised deep-clustering [39, 40, 41] to supervised learning adding a classification loss in order to improve classification performance

[42, 43]. In this paper, we use the cross entropy as the added classification loss. Let X be the dataset in $\mathbb{R}^d$, and Y the labels in $\{0, \dots, k\}$, with k the number of classes. Let $Z \in \mathbb{R}^k$ be the encoded latent vectors, $\hat{X} \in \mathbb{R}^d$ the reconstructed data and W the weights of the neural network. Note that the dimension of the latent space k corresponds to the number of classes.

The goal is to learn the network weights W minimizing the total loss. In order to sparsify the neural network, we propose to use the different bi-level projection methods as a constraint to enforce sparsity in our model. The global criterion to minimize

$$\underset{W}{\text{minimize}} \quad \phi(X, Y) \quad \text{subject to} \quad \|W\|_{1,\infty} \leq \eta \quad (18)$$

where $\phi(X, Y) = \alpha\psi(X, \hat{X}) + \mathcal{H}(Y, Z)$. We use the Cross Entropy Loss $\mathcal{H}$ as the added classification loss and the robust Smooth ℓ_1 (Huber) Loss [44] as the reconstruction loss ψ . Parameter α is a linear combination factor used to define the final loss. Note that the constrained approach [45] avoids computing the Lagrangian parameter with the "lasso path" in the case of a Lagrangian approach [46, 47], which is computationally costly [48]. We compute the mask by using the various bilevel projection methods and we use the double descent algorithm [49, 50] for minimizing the criterion 18. We implemented our SAE method using the PyTorch framework for the model, optimizer, schedulers and loss functions. We chose the ADAM optimizer [51], as the standard optimizer in PyTorch. We used a symmetric linear fully connected network with the encoder comprised of an input layer of d neurons, one hidden layer followed by a ReLU or SiLU activation function and a latent layer of dimension $k = 2$.

7.3.2 Experimental accuracy results

We generate artificial biological data to benchmark our bi level projection using the *make_classification* utility from *scikit-learn*. We generate $n = 1,000$ samples with a number $m = 2000$ features because this is the typical range for biological data. We chose a low number of informative features (64) and a separability = 0.8 realistically with biological databases.

We provide the classical accuracy metric and the sparsity score in %: number of columns or features set to zero

The biological **LUNG** dataset was provided by Mathe et al. [52]. The goal of this experiment is to propose a diagnosis of the Lung cancer from urine samples. This dataset includes metabolomic data concerning urine samples from 10005 samples: 469 Non-Small Cell Lung Cancer (NSCLC) patients prior to treatment and 536 control patients. Each sample is described by $m = 2944$ metabolomic features. We apply to this metabolic dataset the classical log-transform for reducing heteroscedasticity. Table 2 shows accuracy classification. The baseline is an implementation that does not process any projection. Compared to the baseline the SAE using the $\ell_{1,\infty}$ projection improves the accuracy by 7%.

Table 3 shows that accuracy results of the bi-level $\ell_{1,\infty}$ and classical $\ell_{1,\infty}$ are similar. Again, compared to the baseline the SAE using the $\ell_{1,\infty}$ projection improves the accuracy by 3.5%.

Figure 5 shows the impact of the radius (η) on synthetic and lung dataset using the bilevel projection. It can be seen that tuning the projection radius, and thus the sparsity, is necessary to improve the accuracy for synthetic data. From Table 2 on synthetic dataset and Table 3 on Lung dataset it can be seen that the best accuracy is obtained for $\eta = 0.75$ for $\ell_{1,\infty}$ and for $\eta = 1$. for the bilevel $\ell_{1,\infty}$ projection, maximum

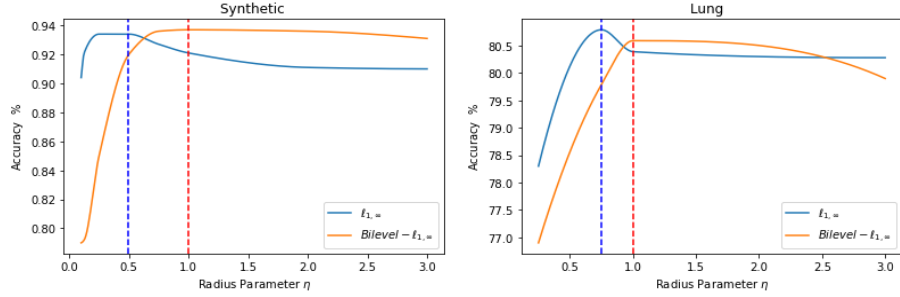


Figure 5: Accuracy as a function of the radius parameter η ; Left Synthetic data;, Right Lung dataset

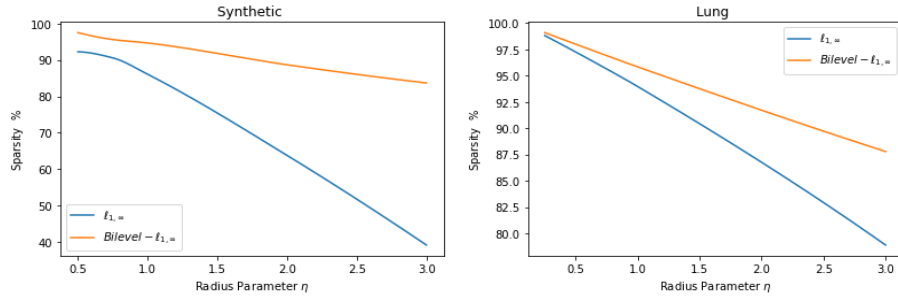


Figure 6: Sparsity as a function of the radius parameter η : Left Synthetic data, Right Lung dataset

accuracy of both method are similar. However sparsity and computation time are better for bilevel $\ell_{1,\infty}$ than for regular $\ell_{1,\infty}$.

Synthetic	Baseline	$\ell_{1,\infty}$	bi-level $\ell_{1,\infty}$
Best Radius		0.75	1.0
Accuracy %	86.6 ± 1.2	94.4 ± 1.45	$94. \pm 1.45$
Sparsity %	-	90.57 ± 0.021	94.66 ± 0.02

Table 2: **Synthetic** dataset. SiLU activation, Accuracy versus sparsity : comparison of $\ell_{1,\infty}$ and bi-level $\ell_{1,\infty}$.

The table 4 on synthetic dataset and the Table 5 on Lung dataset show that the projection bi-level $\ell_{1,1}$ and $\ell_{1,2}$ have almost the same accuracy performance as the projection bi-level $\ell_{1,\infty}$ but the sparsity is much lower.

So bilevel $\ell_{1,\infty}$ outperforms all other projections in terms of computation cost and sparsity.

Lung	Baseline	$\ell_{1,\infty}$ (Chu)	bi-level $\ell_{1,\infty}$
Best Radius	-	0.75	1.0
Accuracy %	77.12 $\pm$ 3	80.79 $\pm$ 0.4	80.59 $\pm$ 0.43
Sparsity %	-	95.63 $\pm$ 0.75	95.83 $\pm$ 0.6

Table 3: **Lung** dataset. SiLU activation, Accuracy versus sparsity : comparison of $\ell_{1,\infty}$ $\eta = 0.75$ and bi-level $\ell_{1,\infty}$ $\eta = 1$.

Synthetic	Baseline	$\ell_{1,2}$	bi-level $\ell_{1,1}$
Best Radius		75	75
Accuracy %	86.6 $\pm$ 1.2	94.4 $\pm$ 1.45	95.1 $\pm$ 1.45
Sparsity %	-	76 $\pm$ 0.021	75.6 $\pm$ 0.02

Table 4: **Synthetic** dataset. SiLU activation, Accuracy versus sparsity : comparison of $\ell_{1,2}$ and bi-level $\ell_{1,1}$.

Lung	Baseline	$\ell_{1,2}$	bi-level $\ell_{1,1}$
Best Radius		75	200
Accuracy %	79.4 $\pm$ 3	80.59 $\pm$ 0.4	80.79 $\pm$ 0.43
Sparsity %	-	76.2 $\pm$ 0.75	76.90 $\pm$ 0.6

Table 5: **Lung** dataset. SiLU activation, Accuracy versus sparsity : comparison of $\ell_{1,2}$ and bi-level $\ell_{1,1}$.

8 Conclusion

Although many projection algorithms were proposed for the projection of the $\ell_{1,\infty}$ norm, complexity of these algorithms remain an issue. The worst-case time complexity of these algorithms is $\mathcal{O}(n \times m \times \log(n \times m))$ for a matrix in $\mathbb{R}^{n \times m}$. In order to cope with this complexity issue, we have proposed a new bi-level (and multilevel) projection method. The main motivation of our work is the direct independent splitting made by the bi-level optimization which takes into account the structured sparsity requirement. We showed that the theoretical computational cost of our new bi-level method is only $\mathcal{O}(n \times m)$ for a matrix in $\mathbb{R}^{n \times m}$. Experiments on synthetic data show that our bi-level method is 2.5 times faster than the actual fastest algorithm provided by *Chu et. al.*. Moreover our bi-level $\ell_{1,\infty}$ projection outperforms other bi-level projections $\ell_{1,1}$ and $\ell_{1,2}$. Our parallel implementation with Nw workers improves the computational time by a factor close to Nw. Note that our extension to multilevel projection can be applied for sparsifying large convolutional neural networks.

References

- [1] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, “Robust biomarker identification for cancer diagnosis with ensemble feature selection methods,” *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2009.

- [2] Z. He and W. Yu, “Stable feature selection for biomarker discovery,” *Computational biology and chemistry*, vol. 34, no. 4, pp. 215–225, 2010.
- [3] D. L. Donoho *et al.*, “Compressed sensing,” *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [4] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, “Sparse reconstruction by separable approximation,” *IEEE Transactions on signal processing*, vol. 57, no. 7, pp. 2479–2493, 2009.
- [5] J. M. Alvarez and M. Salzmann, “Learning the number of neurons in deep networks,” in *Advances in Neural Information Processing Systems*, 2016, pp. 2270–2278.
- [6] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” in *Advances in neural information processing systems*, 2015, pp. 1135–1143.
- [7] E. Tartaglione, S. Lepsøy, A. Fiandrotti, and G. Francini, “Learning sparse neural networks via sensitivity-driven regularization,” in *Advances in Neural Information Processing Systems*, 2018, pp. 3878–3888.
- [8] A. N. Gomez, I. Zhang, K. Swersky, Y. Gal, and G. E. Hinton, “Learning sparse networks using targeted dropout,” *arXiv :1905.13678*, 2019.
- [9] U. Oswal, C. Cox, M. Lambon-Ralph, T. Rogers, and R. Nowak, “Representational similarity learning with application to brain networks,” in *International Conference on Machine Learning*, 2016, pp. 1041–1049.
- [10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [11] J. Cavazza, P. Morerio, B. Haeffele, C. Lane, V. Murino, and R. Vidal, “Dropout as a low-rank regularizer for matrix factorization,” in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018, pp. 435–444.
- [12] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [13] T. Hastie, R. Tibshirani, and M. Wainwright, “Statistical learning with sparsity: The lasso and generalizations,” *CRC Press*, 2015.
- [14] L. Condat, “Fast projection onto the simplex and the ℓ_1 ball,” *Mathematical Programming Series A*, vol. 158, no. 1, pp. 575–585, 2016.
- [15] G. Perez, M. Barlaud, L. Fillatre, and J.-C. Régis, “A filtered bucket-clustering method for projection onto the simplex and the ℓ_1 -ball,” *Mathematical Programming*, May 2019.
- [16] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.
- [17] Z. Huang and N. Wang, “Data-driven sparse structure selection for deep neural networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 304–320.
- [18] J. Yoon and S. J. Hwang, “Combined group and exclusive sparsity for deep neural networks,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3958–3966.

- [19] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, “Group sparse regularization for deep neural networks,” *Neurocomputing*, vol. 241, pp. 81–89, 2017.
- [20] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A sparse-group lasso,” *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.
- [21] I. Yasutoshi, F. Yasuhiro, and K. Hisashi, “Fast sparse group lasso,” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.
- [22] M. Barlaud and F. Guyard, “Learning sparse deep neural networks using efficient structured projections on convex constraints for green ai,” *International Conference on Pattern Recognition, Milan*, pp. 1566–1573, 2020.
- [23] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, “An efficient projection for $\ell_{1,\infty}$ regularization,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 857–864.
- [24] A. Rakotomamonjy, R. Flamary, G. Gasso, and S. Canu, “lp-lq penalty for sparse linear and sparse multiple kernel multitask learning,” *IEEE Transactions on Neural Networks*, vol. 22, p. 307–1320, 2011.
- [25] B. Bejar, I. Dokmanić, and R. Vidal, “The fastest $\ell_{1,\infty}$ prox in the West,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3858–3869, 2021.
- [26] G. Chau, B. Wohlberg, and P. Rodriguez, “Efficient projection onto the $\ell_{1,\infty}$ mixed-norm ball using a newton root search method,” *SIAM Journal on Imaging Sciences*, vol. 12, no. 1, pp. 604–623, 2019.
- [27] D. Chu, C. Zhang, S. Sun, and Q. Tao, “Semismooth newton algorithm for efficient projections onto $\ell_{1,\infty}$ -norm ball,” in *International Conference on Machine Learning*, 2020, pp. 1974–1983.
- [28] A. Sinha, P. Malo, and K. Deb, “A review on bilevel optimization: From classical to evolutionary approaches and applications,” *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, pp. 276–295, 2018.
- [29] K. Bennett, J. Hu, X. Ji, G. Kunapuli, and J.-S. Pang, “Model selection via bilevel optimization,” *IEEE International Conference on Neural Networks - Conference Proceedings*, 2006.
- [30] G. Perez, L. Condat, and M. Barlaud, “Near-linear time projection onto the $\ell_{1,\infty}$ ball application to sparse autoencoders,” *arXiv: 2307.09836*, 2023.
- [31] D. Kong, R. Fujimaki, J. Liu, F. Nie, and C. Ding, “Exclusive feature learning on arbitrary structures via $\ell_{1,2}$ -norm,” in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 1655–1663.
- [32] D. Gregoratti, X. Mestre, and C. Buelga, “Exclusive group lasso for structured variable selection,” *arXiv:2108.10284*, 2021.
- [33] Y. Zhou, R. Jin, and S. C. Hoi, “Exclusive group lasso for multi-task feature selection,” *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.
- [34] M. Barlaud, A. Chambolle, and J.-B. Caillaud, “Classification and feature selection using a primal-dual method and projection on structured constraints,” *International Conference on Pattern Recognition, Milan*, pp. 6538–6545, 2020.

- [35] L. Theis, W. Shi, A. Cunningham, and F. Huszár, “Lossy image compression with compressive autoencoders,” *ICLR Conference Toulon*, 2017.
- [36] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, “High-fidelity generative image compression,” *NEURIPS*, 2020.
- [37] J. Ascenso, E. Alshina, and T. Ebrahimi, “The jpeg ai standard: Providing efficient human and machine visual data consumption,” *IEEE MultiMedia*, vol. 30, no. 1, pp. 100–111, 2023.
- [38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016, vol. 1.
- [39] D. Kingma and M. Welling, “Auto-encoding variational bayes,” *International Conference on Learning Representation*, 2014.
- [40] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” *Advances in neural information processing systems*, vol. 27, 2014.
- [41] J. Snoek, R. Adams, and H. Larochelle, “On nonparametric guidance for learning autoencoder representations,” in *Artificial Intelligence and Statistics*. PMLR, 2012, pp. 1073–1080.
- [42] L. Le, A. Patterson, and M. White, “Supervised autoencoders: Improving generalization performance with unsupervised regularizers,” *Advances in Neural Information Processing Systems*, 2018.
- [43] M. Barlaud and F. Guyard, “Learning a sparse generative non-parametric supervised autoencoder,” *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada*, June 2021.
- [44] P. J. Huber, *Robust statistics*. Wiley, New York, 1981.
- [45] M. Barlaud, W. Belhajali, P. Combettes, and L. Fillatre, “Classification and regression using an outer approximation projection-gradient method,” vol. 65, no. 17, 2017, pp. 4635–4643.
- [46] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, “The entire regularization path for the support vector machine,” *Journal of Machine Learning Research*, vol. 5, pp. 1391–1415, 2004.
- [47] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization path for generalized linear models via coordinate descent,” *Journal of Statistical Software*, vol. 33, pp. 1–122, 2010.
- [48] J. Mairal and B. Yu, “Complexity analysis of the lasso regularization path,” in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012, pp. 353–360.
- [49] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” *arXiv preprint arXiv:1803.03635*, 2018.
- [50] H. Zhou, J. Lan, R. Liu, and J. Yosinski, “Deconstructing lottery tickets: Zeros, signs, and the supermask,” in *Advances in Neural Information Processing Systems 32*, 2019, pp. 3597–3607.
- [51] D. Kingma and J. Ba, “a method for stochastic optimization.” *International Conference on Learning Representations*, pp. 1–13, 2015.
- [52] E. Mathé *et al.*, “Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer,” *Cancer research*, vol. 74, no. 12, p. 3259–3270, June 2014.

A Group LASSO and Exclusive LASSO

Let first recall the definition of the usual $\ell_{1,2}$ Group LASSO projection be:

$$P_\eta^{1,2}(Y) = \arg \min \|X - Y\|_2 \text{ s.t. } \|X\|_{1,2} \leq \eta. \quad (19)$$

Note that the formulation of the Group-LASSO is different from our bi-level approach, and their optimal points are in the general case different. The classical Group-Lasso algorithm is based on Block coordinate descent (BCD) [20] which requires high computational cost. Recent paper, provided algorithms for fast Group LASSO, yet the algorithm is quite complicated [21].

The exclusive LASSO was first introduced in [31, 32, 33]. The main idea is that if one feature in a class is selected (large weight), the method tends to assign small weights to the other features in the same class. Using our bi-level framework, we obtain the following bi-level projection algorithms.

Algorithm 7 Bi-level $\ell_{2,1}$ projection.

Input: Y, η
 $t \leftarrow P_\eta^2((\|Y_2\|_1, \dots, \|Y_n\|_1))$
for $i \in [1, \dots, n]$ **do**
 $X_i \leftarrow P_{t_i}^1(Y_i) \forall i \in 1, \dots, n$
end for
Output: X

B Double descent algorithm

Originally proposed as follows: after training a network, set all weights smaller than a given threshold to zero, rewind the rest of the weights to their initial configuration, and then retrain the network from this starting configuration while keeping the zero weights frozen (untrained). A possible implementation of the double descent algorithm is given in Algorithm 8.

Algorithm 8 Double-descent algorithm.

1: **Input:** W^*, γ, η
2: **for** $n = 1, \dots, N(\text{epochs})$ **do**
3: $V \leftarrow f(W, \gamma, \nabla \psi(W))$
4: **end for**
5: $W := BP_\eta^{p,q}(V)$
6: $(M_0)_{ij} \leftarrow \mathbb{1}_{x \neq 0}(w_{ij})$
7: **for** $n = 1, \dots, N(\text{epoch})$ **do**
8: $W \leftarrow f(W, \gamma, \nabla \phi(W, M_0))$
9: **end for**
10: **Output:** W

In line 3 the first descent is made. Then in line 5 the projection is made and line 6 extract the mask of non-zero entries in the projected matrix. Finally line 8 is the second descent. Note that low values of η imply high sparsity of the network.

C Iterative form of the multi-level projection

In the paper we provided a recursive form of the tri-level algorithm and an iterative form for the multi-level as they are easier to understand. We provided here their iterative version.

Algorithm 9 Tri-level iterative projection $\ell_{1,\infty,\infty}$.

```

1: Input:  $\mathcal{Y} \in \mathbb{R}^{c,n,m}, \eta$ 
2:  $u^3 \leftarrow P_\eta^1(((v_{\infty,\infty})_t, \forall t \in m))$ 
3: for  $t \in \langle m \rangle$  do
4:    $U_t^2 \leftarrow P_{u_t^3}^\infty((V_\infty)_t)$ 
5: end for
6: for  $t \in \langle n, m \rangle$  do
7:    $\mathcal{U}_t^1 \leftarrow P_{U_t^2}^\infty(\mathcal{Y}_t)$ 
8: end for
9: Output:  $\mathcal{U}^1$ 

```

Algorithm 10 Multi-level iterative projection.

```

1: Input:  $\mathcal{Y}, \nu, \eta$ 
2:  $\mathcal{U}^{|\nu|} \leftarrow P_\eta^{\nu_{|\nu|}}((\mathcal{V}_{\nu_1, \dots, \nu_{|\nu|-1}})_t, \forall t \in T_{|T| - |\nu_{|\nu|}| : |T|})$ 
3: for  $\nu_i \in (\nu_{|\nu|-1}, \dots, \nu_1)$  do
4:   for  $t \in \langle T_{|T| - \sum_{j=i+1}^{|\nu|} |\nu_j| : |T|} \rangle$  do
5:      $\mathcal{U}_t^i \leftarrow P_{\mathcal{U}_t^{i-1}}^{\nu_i}((\mathcal{V}_{\nu_1, \dots, \nu_{i-1}})_t)$ 
6:   end for
7: end for
8: Output:  $\mathcal{U}^1$ 

```

D Proofs of propositions

Proof of proposition 6.3. For any norm $\mathcal{N}$ and radius η , let $\nu = (\mathcal{N})$. Then, since $|\nu| = 1$, we have $MP_\eta^\nu(Y) = P_\eta^\mathcal{N}(Y)$.

Proof of proposition 6.4 The complexity of the recursive algorithm 6 is split in two parts: **1) Aggregations**) At line 2 tensor $\mathcal{Y}$ is aggregated using the first norm presents in ν . This will be done by each call of the $MP_\eta^\nu(\mathcal{Y})$ until ν is a singleton. For a given norm ν_i , the current aggregated tensor is split into independent sub-parts and the norm ν_i of each of these sub-parts is processed. For the two-level case, the aggregation is made of many independent processes that can run in parallel. The time complexity of aggregation for norm ν_i is the sum of the time complexity of processing the norm ν_i for each sub-part t $O(\sum_t f_i(|\nu_i|))$. With infinite parallel processing power, the time complexity can be reduced to $O(f_i(|\nu_i|))$, as each sub-part is independent. **2) Projections**) Once the aggregated tensor is processed, it is projected. The result of each of these projections is stored into t . The most inner one, in the deepest call of the recursive processing, is the usual projection of the aggregated tensor onto the

norm $\nu_{|\nu|}$ of radius η . The time complexity of this line is $O(f_p(|\nu_{|\nu|}|))$ which is the time to project onto the norm $\nu_{|\nu|}$. Then, the algorithm uses the result of the previous projection (t^2) to project its current aggregation of the tensor, given as input argument in the recursive calls. This is done in line 4 where the result of the previous projection (t_t^2 , a value) is used to compute the current projection (t_t^1 , a tensor). Again, the loop on the indices t , at line 3, is made of independent sub-parts that can be run in parallel. This implies that with infinite parallel processing power, the time complexity of the projections is $O(\sum_i f_{p,i}(|\nu_i|))$. Compared to the general case, the parallel computation may provide an exponential speedup. An iterative implementation of the algorithm is given in Algorithm 10 to explicitly show the different computations.