# Orthogonal regularizers in deep learning: how to handle rectangular matrices?

Estelle Massart
ICTEAM, Université catholique de Louvain,
Avenue Georges Lemaitre, 4/L4.05.01
1348 Louvain-la-Neuve, Belgium
Email: estelle.massart@uclouvain.be

***Abstract*—Orthogonal regularizers typically promote column orthonormality of some matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$, by measuring the discrepancy between $\mathbf{W}^\top\mathbf{W}$ and the identity according to some matrix norm. This paper explores the behavior of these regularizers when $\mathbf{W}$ is horizontal ($n < p$), so that column orthonormality cannot be achieved. Our motivation comes from orthogonal regularization of feed-forward neural networks: it is there desired to regularize all (vertical and horizontal) weight matrices of the model.**

**One possible solution to address this issue is to transpose horizontal matrices before regularization. We prove that transposition is useless for the Frobenius norm (squared), as the corresponding regularizer promotes simultaneously orthonormality of the rows and of the columns of $\mathbf{W}$. On the other hand, we highlight important qualitative differences with newer regularizers, including the MC and SRIP orthogonal regularizers. We conclude the paper with some numerical results supporting our theoretical findings.**

## I. Introduction

Recent years have seen several works integrating orthogonal regularizers/constraints in deep neural networks. For example, orthogonal regularization of weight matrices or convolutional filters has been proposed to reduce correlation between neurons of the same layer in deep neural network training, thereby diminishing model overfitting [1]. Orthogonality was also shown to improve signal propagation in deep neural networks (DNNs): the authors of [2] show that initializing the weight matrices at random following a uniform distribution on the orthogonal group (apart from a common factor scaling the norm of all columns of the matrix) leads for some activation functions to dynamical isometry, a regime where the spectrum of the input-output Jacobian of the network at initialization concentrates around the unity. This type of weight initialization was further proven in [3] to result in a faster training for deep linear neural networks, compared to standard Gaussian initialization. Regarding model generalization, the authors of [4] derived a bound on the generalization error, using the machinery of algorithmic robustness for feed-forward ReLU neural networks, and showed that their bound is minimized if the weight matrices are orthogonal. Finally, let us mention that orthogonality constraints have also been used in recurrent neural network training [5], Wasserstein generative adversarial network training [6], unsupervised extraction of disantangled features [7], and interpretable learning [8].

These applications motivated the development of efficient algorithms to train deep learning models under orthogonality constraints, see, e.g., [9] and references therein. This paper addresses the alternative question of orthogonal regularization. Indeed, in some cases, imposing orthogonality constraints is thought to reduce model expressivity and harm model accuracy, so that orthogonal regularization is preferred [4].

### A. Mathematical formulation

Let us consider a feed-forward neural network $f : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$, defined as

$$f(\mathbf{x}) = \phi(\mathbf{W}_L\sigma(\mathbf{W}_{L-1}\dots\sigma(\mathbf{W}_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1})+\mathbf{b}_L), \tag{1}$$

where $\mathbf{W}_1,\dots,\mathbf{W}_L$ are weight matrices, with $\mathbf{W}_l \in \mathbb{R}^{d_l\times d_{l-1}}$, $\mathbf{b}_1,\dots,\mathbf{b}_L$, with $\mathbf{b}_l \in \mathbb{R}^{d_l}$, bias vectors, and $\sigma$ and $\phi$ are nonlinear functions (e.g., ReLU and softmax, respectively). Given a set of training data $\mathcal{T} = \{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1,\dots,N}$, the parameters of the model are typically learned by solving the regularized optimization problem

$$\min_{\substack{\mathbf{W}_l,\mathbf{b}_l,\\ l=1,\dots,L}} \sum_{i=1}^{N} \mathcal{L}(\mathbf{y}_i, f(\mathbf{x}_i)) + \lambda\sum_{j=1}^{L}\mathcal{R}_j(\mathbf{W}_j),$$

where $\mathcal{L} : \mathbb{R}^{d_L}\times\mathbb{R}^{d_L} \to \mathbb{R}_{\geq 0}$ is a loss function, $\mathcal{R}_j : \mathbb{R}^{d_j\times d_{j-1}} \to \mathbb{R}_{\geq 0}$ a regularizer, and $\lambda > 0$ a weight balancing the loss and the regularizer in the final objective. For example, one may choose $\mathcal{R}_j(\mathbf{W}_j) = \|\mathbf{W}_j^\top\mathbf{W}_j - \mathbf{I}_{d_{j-1}}\|_{\mathrm{F}}$ for all $j$, where $\mathbf{I}_d$ is the $d\times d$ identity matrix. Often, when we are not considering one layer in particular, we drop the layer indices and simply write $\mathcal{R}(\mathbf{W})$ for the regularizer applied to some matrix $\mathbf{W}$, with $\mathbf{W} \in \mathbb{R}^{n\times p}$ (so $n$ and $p$ are standard variables referring to the number of rows and columns of the matrix, respectively). In this work, matrices are displayed in capital bold letters and vectors in lowercase bold letters.

### B. Orthogonal regularizers proposed in the literature

Table I presents the most common orthogonal regularizers used in the literature (note that we added a factor $1/2$ in the definition of the DSO regularizer for comparison purposes). These regularizers typically impose the product $\mathbf{W}^\top\mathbf{W}$ to be close to the identity, and differ in the norm chosen. We use the following notation: for $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\|\mathbf{A}\|_2$ is the spectral norm, $\|\mathbf{A}\|_1 = \sum_{i,j}|a_{ij}|$ the (entrywise) $l^1$ norm, $\|\mathbf{A}\|_{\mathrm{F}} =$

$\left(\sum_{i,j} a_{ij}^2\right)^{1/2}$ the Frobenius norm and $\|\mathbf{A}\|_\infty := \max_{i,j} |a_{ij}|$ the entrywise $\infty$-norm (we stress that it differs in general from both the induced and Schatten $\infty$-norms).

As indicated in Table I, most orthogonal regularizers promote column orthogonality, by measuring some norm of the matrix $\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p$. If the neural network contains horizontal weight matrices ($n < p$), column orthogonality cannot be achieved. Alternatives have been proposed: promoting orthogonality of the rows instead of the columns (see, for example, the DSO regularizer in Table I), or splitting the set of columns into smaller subsets, and promoting orthogonality of each subset of columns separately [10].

TABLE I: Main orthogonal regularizers proposed so far, for $\mathbf{W} \in \mathbb{R}^{n\times p}$.

| NAME | DESCRIPTION |
|---|---|
| $\mathrm{Orth}_{l^1}$ [8] | $\mathcal{R}_{\mathrm{Orth}_{l^1}}(\mathbf{W}) = \|\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p\|_1^2$ |
| SO [11] | $\mathcal{R}_{\mathrm{SO}}(\mathbf{W}) = \|\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p\|_{\mathrm{F}}^2$ |
| DSO [11] | $\mathcal{R}_{\mathrm{DSO}}(\mathbf{W}) = \frac{1}{2}(\|\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p\|_{\mathrm{F}}^2 + \|\mathbf{W}\mathbf{W}^\top - \mathbf{I}_n\|_{\mathrm{F}}^2)$ |
| MC [11] | $\mathcal{R}_{\mathrm{MC}}(\mathbf{W}) = \|\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p\|_\infty$ |
| SRIP [11] | $\mathcal{R}_{\mathrm{SRIP}}(\mathbf{W}) = \|\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p\|_2$ |

The $\mathrm{Orth}_{l^1}$ and SO (*Soft Orthogonality*) regularizers are simply obtained by using the entrywise $l^1$ and $l^2$ norms to measure column non-orthonormality. The DSO (*Double Soft Orthogonality*) was proposed to account for the case of horizontal matrices: it penalizes explicitly both non-orthonormality of the rows and columns of the matrix. The MC regularizer was originally motivated by the notion of *mutual coherence* of $\mathbf{W} \in \mathbb{R}^{n\times p}$, defined as

$$\mu_{\mathbf{W}} = \max_{i\neq j} \frac{|\mathbf{w}_i^\top \mathbf{w}_j|}{\|\mathbf{w}_i\| \cdot \|\mathbf{w}_j\|},$$

where $\mathbf{w}_i \in \mathbb{R}^n$ is the $i$th column of $\mathbf{W}$ and $\|\cdot\|$ is the $l^2$ norm. The SRIP (*Spectral Restricted Isometry Property*) regularizer was also proposed in [11], building on the well-known restricted isometry property widely used in compressed sensing; it is there also suggested to approximate the spectral norm with a few iterations of the power method. The authors of [11] highlight that the DSO, MC and SRIP regularizers are well-defined both in the case $n \leq p$ and $n > p$.

This paper mostly addresses orthogonal regularizers for feed-forward neural networks, though our findings apply equally to convolutional neural networks where orthogonal regularizers are applied to matrix-reshaped versions of the convolutional filters, as suggested in [11]. Note that there exist other ways to promote orthogonality of convolutional filter tensors, that we do not consider in this work, see, e.g., [12], [13] and references therein.

### C. Contributions

We compare mathematically the different orthogonal regularizers in Table I, by characterizing their optimal solutions. Obviously, the $\mathrm{Orth}_{l^1}$, SO, MC and SRIP regularizers are optimal if and only if $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_p$, so that the columns of $\mathbf{W}$ are orthonormal. In the case $n < p$, it is not possible to impose orthonormality of the columns of $\mathbf{W}$, and we show that in this case the regularizers in Table I exhibit substantial differences. More precisely, we prove that:

- The SO and DSO regularizers are mathematically equivalent: when using the Frobenius norm (squared), penalizing non-orthonormality of the rows or of the columns yields the same gradient flow. The SO regularizer can thus be used both for $n \geq p$ and $n < p$, without having to transpose the matrix in the last case.
- In the case $n = p - 1$, the $\mathrm{Orth}_{l^1}$ regularizer is minimized if and only if $\mathbf{W}$ has orthogonal rows and one column equal to zero; the set of minimizers of the $\mathrm{Orth}_{l^1}$ regularizer is therefore a proper subset of the set of minimizers of the SO regularizer.
- In the case $n < p$, the SRIP regularizer is minimized if and only if all the singular values of the matrix $\mathbf{W}$ lie in the interval $[0, \sqrt{2}]$: the set of minimizers of the SO regularizer is thus a proper subset of the set of minimizers of the SRIP regularizer.
- In the case $n = p - 1$, the MC regularizer is minimized if and only if $\mathbf{W} = \mathbf{Q}\mathbf{M}$, where $\mathbf{Q}$ is any $n \times n$ orthogonal matrix and $\mathbf{M}$ is a matrix whose columns have identical norms and are separated by identical angles. Graphically, the columns of $\mathbf{M}$ are the $n + 1$ vertices of an $n$-dimensional regular simplex (generalization of the equilateral triangle, or regular tetrahedron, to $n$ dimensions). The set of minimizers of the MC regularizer is thus in general disjoint from the set of minimizers of the SO regularizer.

Our findings indicate that, analogously to LASSO regression, resorting to the $\mathrm{Orth}_{l^1}$ regularizer for promoting row orthogonality of the weight matrices tends to sparsify the model when the layer size is increasing. Let us assume that $d_l = d_{l-1} + 1$ for some layer $l$. Row orthogonality of $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ is not achievable and, as an immediate corollary of the discussion above, one row of the optimal $\mathbf{W}_l$ for the $\mathrm{Orth}_{l^1}$ regularizer is zero. There follows that, at optimality of the $\mathrm{Orth}_{l^1}$ regularizer, one activation of layer $l$ is zero. We also illustrate this sparsification induced by the $\mathrm{Orth}_{l^1}$ regularizer in the last section of the paper for various architectures with increasing layer dimension (beyond $d_l = d_{l-1}+1$), for feed-forward, but also convolutional neural networks.

## II. Comparison of orthogonal regularizers

This section aims to characterize the optimal solutions to the regularizers presented in Table I, in the case $n < p$.

### A. Regularizers based on the Frobenius norm

We first prove the following result, that relates the SO and DSO regularizers.

**Proposition II.1.** *Let $\mathbf{W} \in \mathbb{R}^{n\times p}$. Then,*

$$\|\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p\|_{\mathrm{F}}^2 = \|\mathbf{W}\mathbf{W}^\top - \mathbf{I}_n\|_{\mathrm{F}}^2 + p - n.$$

*Proof.* There holds

$$\begin{aligned}\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\|_{\mathrm{F}}^2 &= \mathrm{Tr}\left(\left(\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\right)^\top\left(\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\right)\right)\\ &= \mathrm{Tr}\left(\mathbf{W}^\top\mathbf{W}\mathbf{W}^\top\mathbf{W}\right)-2\mathrm{Tr}\left(\mathbf{W}^\top\mathbf{W}\right)+\mathrm{Tr}(\mathbf{I}_p)\\ &= \mathrm{Tr}\left(\mathbf{W}\mathbf{W}^\top\mathbf{W}\mathbf{W}^\top\right)-2\mathrm{Tr}\left(\mathbf{W}\mathbf{W}^\top\right)+\mathrm{Tr}(\mathbf{I}_p)\\ &= \|\mathbf{W}\mathbf{W}^\top-\mathbf{I}_n\|_{\mathrm{F}}^2-\mathrm{Tr}(\mathbf{I}_n)+\mathrm{Tr}(\mathbf{I}_p).\end{aligned}$$

□

Proposition II.1 indicates that the SO and DSO regularizers are equivalent: using either of them provides the same objective up to a constant, hence the same gradient flow. It follows that, when using the SO regularizer, transposing the matrix when it is horizontal leads to exactly the same gradient flow than without transposition, this step is thus not needed.

Let us now consider a weight matrix $\mathbf{W}\in\mathbb{R}^{n\times p}$, and assume that $n<p$. An immediate consequence of Proposition II.1 is that the SO (and DSO) optimizer is minimized if and only if $\mathbf{W}\mathbf{W}^\top=\mathbf{I}_n$, i.e., the rows of the weight matrix are orthonormal. In terms of singular value decomposition, the set of minimizers of the SO and DSO regularizers, in the case $n<p$, is given by the following result.

**Corollary II.2.** *Let $\mathbf{W}\in\mathbb{R}^{n\times p}$, with $n<p$. Then, the SO (and DSO) regularizers are minimized if and only if*

$$\mathbf{W}=\mathbf{U}\begin{pmatrix}\mathbf{I}_n & \mathbf{0}_{p-n}\end{pmatrix}\mathbf{V}^\top, \tag{2}$$

*where $\mathbf{U}\in\mathbb{R}^{n\times n}$ and $\mathbf{V}\in\mathbb{R}^{p\times p}$ are orthogonal, i.e., all the singular values of $\mathbf{W}$ are equal to one.*

### *B. Regularizer based on the entrywise $l^1$ norm*

We restrict here our analysis to the case $n=p-1$, for which we characterize the set of global minimizers as a proper subset of the set of minimizers of the SO and DSO regularizers: additionally to having orthonormal rows, minimizers of the $\mathrm{Orth}_{l^1}$ regularizer have one zero column.

**Proposition II.3.** *Let $\mathbf{W}\in\mathbb{R}^{n\times p}$, with $n=p-1$. Then, the $\mathrm{Orth}_{l^1}$ regularizer is minimized if and only if*

$$\mathbf{W}=\mathbf{U}\begin{pmatrix}\mathbf{I}_n & \mathbf{0}_1\end{pmatrix}\mathbf{V}^\top, \tag{3}$$

*where $\mathbf{V}\in\mathbb{R}^{p\times p}$ is orthogonal and such that $\mathbf{V}=[\tilde{\mathbf{V}}\ \mathbf{e}_k]$ with $\mathbf{e}_k$ the $k$th canonical vector for some $k\in\{1,\dots,p\}$. This implies that the $k$th row of $\tilde{\mathbf{V}}$ (hence the $k$th column of $\mathbf{W}$) is zero.*

*Proof.* Let us first assume that $\mathbf{W}$ is a minimizer of the SO regularizer, so that $\mathbf{W}=\mathbf{U}\begin{pmatrix}\mathbf{I}_{p-1} & \mathbf{0}_1\end{pmatrix}\mathbf{V}^\top=\mathbf{U}\tilde{\mathbf{V}}^\top$, where $\tilde{\mathbf{V}}\in\mathbb{R}^{p\times p-1}$ is the submatrix containing the $p-1$ first columns of $\mathbf{V}$. Then,

$$\begin{aligned}\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\|_1 &= \|\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top-\mathbf{I}_p\|_1\\ &\geq \|\mathrm{diag}(\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top-\mathbf{I}_p)\|_1\\ &= \sum_i|1-(\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top)_{i,i}|\\ &= \sum_i(1-(\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top)_{i,i})\\ &= p-\mathrm{tr}(\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top),\end{aligned}$$

where $\mathrm{diag}(\mathbf{A})$ is the diagonal part of $\mathbf{A}$, and where the third equality results from the fact that the rows of $\tilde{\mathbf{V}}$ have norm upper bounded by one as $\tilde{\mathbf{V}}$ is a submatrix of the orthogonal matrix $\mathbf{V}$. Note also that $\mathrm{tr}(\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top)=\mathrm{tr}(\tilde{\mathbf{V}}^\top\tilde{\mathbf{V}})=p-1$ by construction of $\tilde{\mathbf{V}}$, so that $\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\|_1\geq 1$ for all $\mathbf{W}$ of the form (2). There holds $\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\|_1=1$ if and only if $\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top$ is diagonal, or equivalently, $\mathbf{W}$ is of the form (3).

Note finally that, for all $\mathbf{W}$, there holds $\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\|_1\geq\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\|_{\mathrm{F}}\geq 1$, where the first inequality is a well-known property of the entrywise $l^1$ and $l^2$ norms, and the second is a consequence of the Eckart-Young theorem for the Frobenius norm (since $\mathbf{W}^\top\mathbf{W}$ has rank at most $p-1$). There follows that, for any $\mathbf{W}$ that is not of the form (2), i.e., not a minimizer of the SO regularizer, $\|\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p\|_1>1$.

□

As a consequence, in the case $n=p-1$, the $\mathrm{Orth}_{l^1}$ regularizer promotes one hidden unit of the previous layer to be simply forgotten in the next layer. Note that any matrix of the form (3) satisfies $\mathbf{W}\mathbf{W}^\top=\mathbf{I}_n$, i.e., similarly as for the SO regularizer, the $\mathrm{Orth}_{l^1}$ regularizer promotes orthogonality of the rows of the weight matrix. Note finally that relaxing the assumption $n=p-1$ makes the analysis substantially more complex; we leave this question for further research.

### *C. Regularizer based on the spectral norm*

We now show that, unlike the regularizers considered so far, the SRIP regularizer does not promote row orthogonality if $n<p$.

**Proposition II.4.** *Let $\mathbf{W}\in\mathbb{R}^{n\times p}$, with $n<p$. Then, the SRIP regularizer is minimized if and only if*

$$\mathbf{W}=\mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{4}$$

*with $\mathbf{U}\in\mathbb{R}^{n\times n}$ and $\mathbf{V}\in\mathbb{R}^{p\times p}$ orthogonal, and where $\mathbf{D}\in\mathbb{R}^{n\times p}$ is a diagonal matrix whose diagonal elements are in the interval $[0,\sqrt{2}]$.*

*Proof.* By unitarily invariance of the spectral norm, there holds $\sigma(\mathbf{W}^\top\mathbf{W}-\mathbf{I}_p)=\sigma(\mathbf{V}\mathbf{D}^\top\mathbf{D}\mathbf{V}^\top-\mathbf{I}_p)=\sigma(\mathbf{D}^\top\mathbf{D}-\mathbf{I}_p)=\max_i|\mathbf{D}_{ii}^2-1|$. Note that, since $\mathbf{W}^\top\mathbf{W}$ has rank at most $n<p$, $\mathbf{D}$ has at least $p-n$ diagonal elements equal to zero, so that $\sigma(\mathbf{D}^\top\mathbf{D}-\mathbf{I}_p)\geq 1$. Note finally that $\sigma(\mathbf{D}^\top\mathbf{D}-\mathbf{I}_p)=1$ for all $\mathbf{D}$ with diagonal elements in the interval $[0,\sqrt{2}]$, which concludes the proof. □

This result indicates that, in the case $n<p$, the SRIP regularizer simply promotes the spectrum of the weight matrix to be in some "reasonably small" interval. We have thus proven that using this regularizer on horizontal matrices does not promote neither orthonormality of the rows nor of the columns.

### *D. Regularizer based on the entrywise $l^\infty$ norm*

Again, instead of providing a full characterization of the minimizers, we restrict our analysis to the case $n=p-1$ to ease the analysis. Our main conclusion is that the MC regularizer acts very differently than the regularizers considered so far; in the case $n=p-1$ it is minimized when all columns

of $\mathbf{W}$ have the same length and are pairwise separated by the same angle.

Let us start by exhibiting an optimal solution for the case $n = 2$, $p = 3$. A minimizer of the MC regularizer is given by:

$$\mathbf{W}^* = \sqrt{\frac{2}{3}} \begin{pmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{pmatrix},$$

i.e., the columns of $\mathbf{W}$ are three equal-norm vectors separated by angles of $2\pi/3$. Indeed, note that

$$\mathbf{W}^{*\top}\mathbf{W}^* - \mathbf{I}_3 = \begin{pmatrix} -1/3 & -1/3 & -1/3 \\ -1/3 & -1/3 & -1/3 \\ -1/3 & -1/3 & -1/3 \end{pmatrix},$$

so that $\|\mathbf{W}^{*\top}\mathbf{W} - \mathbf{I}_3\|_\infty = 1/3$. This is tight since

$$\|\mathbf{W}^{*\top}\mathbf{W} - \mathbf{I}_3\|_\infty^2 \geq \|\mathbf{W}^{*\top}\mathbf{W} - \mathbf{I}_3\|_{\mathrm{F}}^2/3^2 \geq 1/9,$$

where the lower bound on the Frobenius norm is a consequence of the Eckart-Young theorem [14], using the fact that $\mathbf{W}^{*\top}\mathbf{W}^*$ has rank at most $n = 2$. For any $p$ and $n = p - 1$, we get the following result.

**Proposition II.5.** *Let $\mathbf{W} \in \mathbb{R}^{n\times p}$, with $n = p - 1$. The* MC *regularizer is minimized at:*

$$\mathbf{W}^* = \sqrt{\frac{n}{p}}\mathbf{M},$$

*where $\mathbf{M} \in \mathbb{R}^{n\times p}$ is constructed as follows.*

$$\begin{aligned} &\mathbf{M}_{11} = 1 \\ &\mathbf{M}_{ij} = -\mathbf{M}_{ii}/(p-i) && \forall j > i \\ &\mathbf{M}_{ii} = \sqrt{1 - \textstyle\sum_{k<i} \mathbf{M}_{ki}^2} \\ &\mathbf{M}_{ij} = 0 && \forall j < i. \end{aligned} \tag{5}$$

*Proof.* Note that, by construction, $(\mathbf{W}^{*\top}\mathbf{W}^* - \mathbf{I}_p)_{i,j} = -1/p$ for all $i, j \in \{1, \ldots, p\}$. This is optimal since

$$\|\mathbf{W}^{*\top}\mathbf{W}^* - \mathbf{I}_p\|_\infty^2 = 1/p^2,$$

which needs to be larger than or equal to

$$\|\mathbf{W}^{*\top}\mathbf{W}^* - \mathbf{I}_p\|_{\mathrm{F}}^2/p^2 \geq (p-n)/p^2 = 1/p^2,$$

where the last inequality results from the Eckart-Young theorem [14]. □

Note further that all minimizers of the MC regularizer can be obtained as a simple rotation of the solution given in Proposition II.5. Indeed, we show the following result.

**Proposition II.6.** *Let $\mathbf{W} \in \mathbb{R}^{n\times p}$, with $n = p - 1$. The* MC *regularizer is minimized at $\mathbf{W}$ if and only if*

$$\mathbf{W} = \sqrt{\frac{n}{p}}\mathbf{Q}\mathbf{M},$$

*where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is orthogonal and $\mathbf{M}$ is defined in* (5)*.*

*Proof.* It is readily checked that $\mathbf{W}$ is optimal if and only if $\mathbf{W}^\top\mathbf{W} - \mathbf{I}_p = -1/p\mathbf{J}_p$, where $\mathbf{J}_p$ is the $p \times p$ matrix whose entries are equal to one, so that $\mathbf{W}^\top\mathbf{W} = -1/p\mathbf{J}_p + \mathbf{I}_p$. The conclusion simply follows from results characterizing all possible solutions to the low-rank factorization problem $\mathbf{W}^\top\mathbf{W} = \mathbf{A}$, for some $\mathbf{A} \in \mathbb{R}^{p\times p}$ of rank $n$, see, e.g., [15, Prop. 2.1]. □

*Geometric intuition:* Note that the optimal solution of the MC regularizer, described in Proposition II.5 has a geometrical interpretation. In the case $n = 2$, $p = 3$, the columns of the matrix $\mathbf{M}$ defined in (5) are three vectors in $\mathbb{R}^2$ whose extremities define an equilaterial triangle centered at the origin. Similarly, for $n = 3$, $p = 4$, the columns of $\mathbf{M}$ form a regular tetrahedron with barycenter at the origin. In higher dimensions, the columns of $\mathbf{M}$ define a $n$-dimensional regular simplex.

## III. Interpretation on the $\mathrm{Orth}_{l^1}$ regularizer

Consider the feed-forward neural network $f$ described in (1), and assume that there exists a layer $l$ such that $d_l = d_{l-1} - 1$ (layer of decreasing dimension). We showed in Section II-B that the $\mathrm{Orth}_{l^1}$ regularizer is optimal at $\mathbf{W}_l^* \in \mathbb{R}^{d_l \times d_{l-1}}$ if and only if $\mathbf{W}_l^*$ has orthonormal rows and one column equal to zero. Up to column permutation, $\mathbf{W}_l^*$ can thus be written as:

$$\mathbf{W}_l^* = \begin{pmatrix} \mathbf{Q}_{d_l} & \mathbf{0}_1 \end{pmatrix},$$

where $\mathbf{Q}_{d_l}$ is a $d_l \times d_l$ orthogonal matrix. This implies that the activations at layer $l$ are linear combinations of the $d_l$ first activations of layer $l-1$, the last activation being simply dropped by the model.

Note that conversely, if we consider the case $d_l = d_{l-1} + 1$ and use the entrywise $l_1$ norm to promote *orthonormality of the rows instead of the columns*, i.e., if we consider the regularizer

$$\tilde{\mathcal{R}}_{\mathrm{Orth}_{l^1}}(\mathbf{W}) = \|\mathbf{W}\mathbf{W}^\top - \mathbf{I}_{d_l}\|_1^2 \tag{6}$$

with $\mathbf{W} \in \mathbb{R}^{d_l \times d_{l-1}}$, optimal solutions are of the form (up to a permutation of the rows of $\mathbf{W}_l^*$):

$$\mathbf{W}_l^* = \begin{pmatrix} \mathbf{Q}_{d_{l-1}} \\ \mathbf{0}_1 \end{pmatrix},$$

so that the last activation of layer $l$ is set to zero. Thus, regularizer (6) induces a sparsification of the model. The next section illustrates numerically a similar sparsification of the model for layers of expanding dimension (i.e., $d_l > d_{l-1}$, but not necessarily assuming $d_l = d_{l-1} + 1$).

## IV. Numerical experiments

In this section, we validate numerically our claim regarding the sparsification induced by the $\mathrm{Orth}_{l^1}$ regularizer. We first consider the MNIST classification problem. As the focus is on regularizer comparison instead of classification performance, we use simple 3-hidden-layers feed-forward neural networks. We consider two different architectures, differing in the number of hidden units per layer: 1000, 1200, and 100 for the first model, and 750, 750, and 100 for the second. Since MNIST pictures are $28 \times 28$, the input dimension is $28^2 = 784$, so the first model corresponds to an architecture with two expanding layers (in which we hope to see the sparsification behaviour described above), while the second

has a monotonically decreasing layer dimension. Layers of both models are endowed with a sigmoid nonlinearity, while the last hidden layer is sent to a softmax with 10 output units. A similar architecture has been used in [16] to illustrate the difference in the distribution of the hidden units with or without batch normalization. In our case, we use this problem setting and architecture to illustrate the impact on training of regularizer (6) compared to the row-orthogonality counterpart of the SO regularizer:

$$\tilde{\mathcal{R}}_{\mathrm{SO}}(\mathbf{W}) = \|\mathbf{W}\mathbf{W}^\top - \mathbf{I}\|_{\mathrm{F}}^2. \tag{7}$$

We trained each architecture with the cross-entropy loss, regularized with either (6) or (7), using stochastic gradient descent with learning rate 0.1, momentum 0.9, batchsize 60, and default initialization for all variables. We did not use any other form or regularization (no weight decay nor dropout). The regularizer weight in (1) was set to $\lambda = 10^{-6}$.

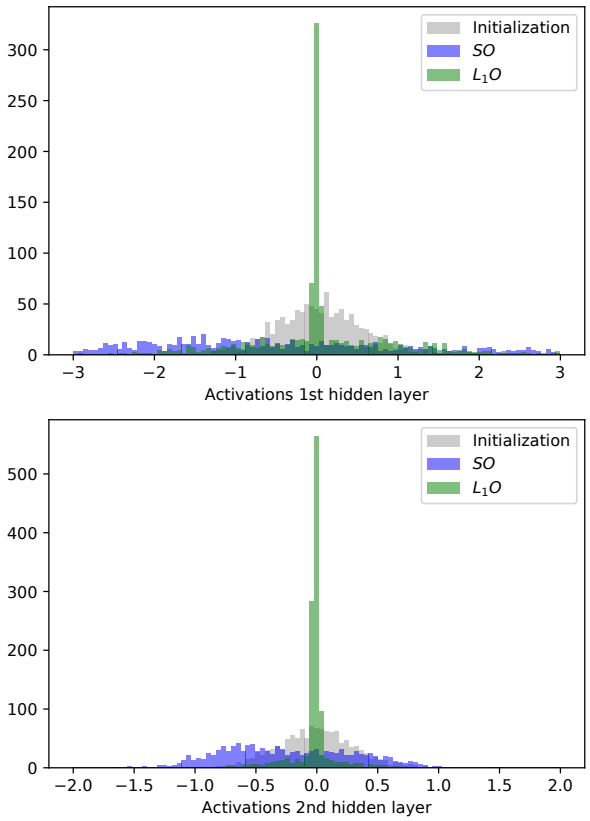


Fig. 1: Histogram of the magnitude of the activations (first and second hidden layers) for the 1000-1200-100 feed-forward neural network trained on MNIST classification. As expected, for the regularizer defined in (6), a substantial amount of activations are very close to zero.

We plot in Figure 1 and Figure 2 histograms representing the distribution of the activations of the two first hidden layers over an arbitrarily selected mini-batch, respectively for the architectures 1000-1200-100 and 750-750-100, after 20 training epochs. As a reference, we also display the histogram of the activations at initialization. Note that a substantial number of activations are very close to zero for the architecture 1000-1200-100, which is not the case for the architecture 750-750-100, consistently with the discussion in Section III.

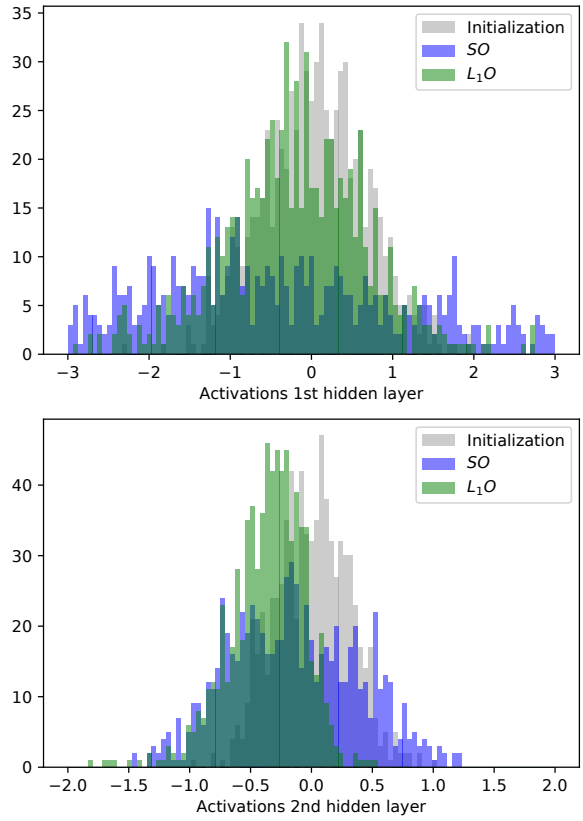


Fig. 2: Histogram of the magnitude of the activations (first and second hidden layers) for the 750-750-100 feed-forward neural network trained on MNIST classification. There is no peak around zero in this case.

### A. *Experiments on CIFAR10 using a WideResNet*

As a second experiment, we considered CIFAR-10 classification using a residual neural network (ResNet), namely, a convolutional neural network endowed with skip connections, allowing to train very deep architectures. We use a Wide ResNet 28-10 architecture, which has 28 layers whose dimensions are progressively increased by a widening factor 10. Due to their layers of increasing dimension, Wide ResNets are particularly suitable to illustrate our claims.

We follow the approach used in [11] to regularize convolutional filters: each convolutional filter $\mathbf{W} \in \mathbb{R}^{C_{out} \times C_{in} \times k \times k}$, with $C_{out}$ and $C_{in}$ output/input channels, and $k$, the filter spatial dimension, is reshaped into a matrix with $C_{out}$ rows and $C_{in}k^2$ columns, to which the regularizer is applied.

We rely on the model architecture implementation of [11]. The model is trained for the cross-entropy loss regularized with either (6) or (7), using SGD with learning rate 0.1, no momentum, batchsize 128, $\lambda = 10^{-6}$, and dropout with

dropping probability 0.3. We did not use any other form of regularization (no weight decay).

According to Section III, we expect, in the case $C_{out} \geq C_{in}k^2$, some rows of the matrix $W \in \mathbb{R}^{C_{out} \times C_{in}k^2}$ representing the convolutional filters to be close to zero. For the sake of comparison, we select two convolutional layers in our model, represented by matrices $W_A \in \mathbb{R}^{160 \times 1440}$, and $W_B \in \mathbb{R}^{160 \times 16}$. As the second has more rows than columns, we expect some of its rows to have a small norm due to the impact of the regularizer (6).

Figure 3 illustrates the distribution of the norms of the rows of $W_A$ and $W_B$ at initialization and after 20 epochs using regularizers (6) and (7). As expected, the norm of several rows of the matrix $W_B$ has substantially decreased over the 20 first epochs.

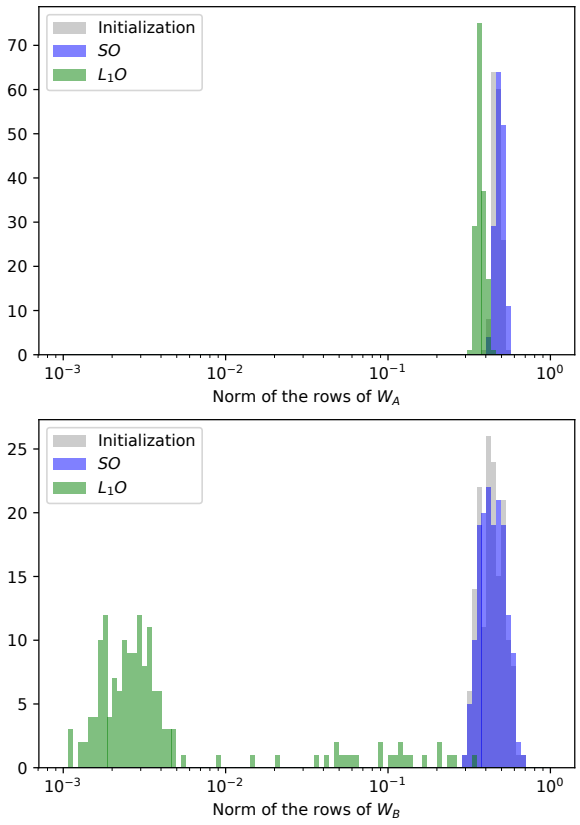


Fig. 3: Histogram of the norm of the rows of the matrices $W_A \in \mathbb{R}^{160 \times 1440}$ and $W_B \in \mathbb{R}^{160 \times 16}$, namely, reshaped convolutional filters of a Wide ResNet 28-10 trained on CIFAR-10 classification for 20 epochs. The regularizer (6) leads to many rows decreasing to zero, unlike regularizer (7).

## V. Conclusion

We characterize the set of optimal solutions of several orthogonal regularizers for horizontal weight matrices, exhibiting important differences between existing regularizers. We hope that this work will pave the way towards a better understanding of the impact of orthogonal regularization in deep learning, regarding model training, generalization, and interactions with other explicit/implicit regularization mechanisms.


## Acknowledgment

Most of this work was done when the author was with the University of Oxford and the National Physical Laboratory (Teddington, UK); the author was then funded by the National Physical Laboratory. The author is now funded by the Fonds de la Recherche Scientifique-FNRS, Belgium.


## References


[1] P. Rodriguez, J. Gonzalez, G. Cucurull, J. M. Gonfaus, and X. Roca, "Regularizing CNNs with locally constrained decorrelations," in *5th International Conference on Learning Representations (ICLR)*, 2017.

[2] J. Pennington, S. S. Schoenholz, and S. Ganguli, "Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice," in *31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[3] W. Hu, L. Xiao, and J. Pennington, "Provable benefit of orthogonal initialization in optimizing deep linear networks," in *8th International Conference on Learning Representations (ICLR)*, virtual, 2020.

[4] K. Jia, S. Li, Y. Wen, T. Liu, and D. Tao, "Orthogonal deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1352 – 1368, 2020.

[5] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," in *33rd International Conference on Machine Learning (ICML)*, New York, NY, USA, 2016.

[6] J. Müller, R. Klein, and M. Weinmann, "Orthogonal wasserstein gans," arxiv: 1911.13060, 2019.

[7] F. Tonin, P. Patrinos, and J. A. Suykens, "Unsupervised learning of disentangled representations in deep restricted kernel machines with orthogonality constraints," *Neural Networks*, vol. 142, pp. 661–679, 2021.

[8] N. Schaaf, M. Huber, and J. Maucher, "Enhancing decision tree based interpretation of deep neural networks through l1-orthogonal regularization," in *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019.

[9] E. Massart and V. Abrol, "Coordinate descent on the orthogonal group for recurrent neural network training," in *36th AAAI conference on artificial intelligence*, virtual, 2022.

[10] L. Huang, X. Liu, B. Lang, A. W. Yu, Y. Wang, and B. Li, "Orthogonal Weight Normalization: Solution to Optimization over Multiple Dependent Stiefel Manifolds in Deep Neural Networks," in *32nd AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018.

[11] N. Bansal, X. Chen, and Z. Wang, "Can we gain more from orthogonality regularizations in training deep CNNs?" in *32nd Conference on Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, 2018.

[12] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, virtual, 2020.

[13] A. Trockman and J. Z. Kolter, "Orthogonalizing Convolutional Layers with the Cayley Transform," in *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[14] G. H. Golub and C. F. V. Loan, *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 2013.

[15] E. Massart and P.-A. Absil, "Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 41, no. 1, pp. 171–198, 2020.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32th International Conference on Machine Learning (ICML)*, Lille, France, 2015.