

On the stable recovery of deep structured linear networks under sparsity constraints

François Malgouyres

▶ **To cite this version:**

François Malgouyres. On the stable recovery of deep structured linear networks under sparsity constraints. Mathematical and Scientific Machine Learning, Jul 2020, Princeton, United States. hal-01526083v3

HAL Id: hal-01526083
https://hal.science/hal-01526083v3

Submitted on 15 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Proceedings of Machine Learning Research vol 107:1–21, 2020

On the stable recovery of deep structured linear networks under sparsity constraints

François Malgouyres MALGOUYRES@MATH.UNIV-TOULOUSE.FR

Institut de Mathématiques de Toulouse ; UMR5219
Université de Toulouse ; CNRS
UPS IMT F-31062 Toulouse Cedex 9, France
and
Institut de Recherche Technologique Saint-Exupéry

Abstract

We consider a deep structured linear network under sparsity constraints. We study sharp conditions guaranteeing the stability of the optimal parameters defining the network. More precisely, we provide sharp conditions on the network architecture and the sample under which the error on the parameters defining the network scales linearly with the reconstruction error (i.e. the risk). Therefore, under these conditions, the weights obtained with a successful algorithms are well defined and only depend on the architecture of the network and the sample. The features in the latent spaces are stably defined. The stability property is required in order to interpret the features defined in the latent spaces. It can also lead to a guarantee on the statistical risk. This is what motivates this study.

The analysis is based on the recently proposed Tensorial Lifting. The particularity of this paper is to consider a sparsity prior. This leads to a better stability constant. As an illustration, we detail the analysis and provide sharp stability guarantees for convolutional linear network under sparsity prior. In this analysis, we distinguish the role of the network architecture and the sample input. This highlights the requirements on the data in connection to parameter stability.

Keywords: Stable recovery, deep structured linear networks, convolutional linear networks, feature robustess.

1. Introduction

1.1. The stability property

Artificial neural networks have improved the state of the art and continue to improve it in a large number of applications in science and technology. Their empirical success far exceeds the understanding of their theoretical properties. In particular, despite the very significant efforts of a very active research community, some behaviors remain partially understood: Why do optimization algorithms find good solutions? Why do over-parameterized neural networks retain good generalization properties? What classes of functions can be approximated by neural networks? With which minimal network architecture?

The work presented in this paper is of a theoretical nature and focuses on a stability property for the parameters leading to a low objective function. The statements are for a regression problem. To explain this stability property in a simplified context, we consider a parameterized family of functions $f_{\mathbf{w}}$ (e.g. neural networks), the parameter being $\mathbf{w}$; the parameter space is equipped with

© 2020 F. Malgouyres.

a metric[1] d; we consider a sample $(x_i, y_i)_{i=1..n}$ of size $n \in \mathbb{N}$. The stability statement then takes the following form.

Informal theorem 1 Stability Guarantee

If a certain condition *on the family f and the sample is satisfied then we have the following* stability property*:*

There exists $C > 0$ such that for η sufficiently small: For any $\mathbf{w}$ and $\mathbf{w}'$ such that

$$\sum_{i=1}^{n} \|f_{\mathbf{w}}(x_i) - y_i\|^2 \leq \eta \qquad \textit{and} \qquad \sum_{i=1}^{n} \|f_{\mathbf{w}'}(x_i) - y_i\|^2 \leq \eta$$

we have

$$d(\mathbf{w}, \mathbf{w}') \leq C\eta.$$

Notice first that the above informal theorem provides a sufficient condition guaranteeing the stability property. Subsequently, depending on the nature of the network under consideration, necessary and sufficient conditions or necessary conditions will be stated. The interest of the stability property is that it guarantees:

- **Feature stability and interpretability:** When the optimal $\mathbf{w}$ is stable, the features in the latent spaces and the output of the network are stably defined in the sense that the parameters $\mathbf{w}$ and $\mathbf{w}'$ for which η is small define similar features and output. The features and the output only depend on the value of
$$\sum_{i=1}^{n} \|f_{\mathbf{w}}(x_i) - y_i\|^2$$
and do not depend on the algorithm used to find $\mathbf{w}$. In particular, they do not depend on its initialization, the numerical parameters, the order of the samples in the stochastic algorithm, the numerical tricks etc The parameter $\mathbf{w}$ and therefore the function $f_{\mathbf{w}}$ only depends on f (i.e.: the network architecture, for neural networks) and the sample $(x_i, y_i)_{i=1..n}$. For neural networks, this is a strong guaranty when interpreting the influence of the features on the output.

- **Stable recovery:** If we make the additional assumption that the data are generated from the family f for an ideal parameter $\mathbf{w}$ (up to an accuracy smaller than η), then the stability guaranty ensures that any parameter $\mathbf{w}'$ for which
$$\sum_{i=1}^{n} \|f_{\mathbf{w}'}(x_i) - y_i\|^2$$
is sufficiently small is close to the ideal $\mathbf{w}$.

  The above additional assumption can be provided by approximation theory statement. This is, for instance, the usual argument in compressed-sensing Elad (2010). When solving a linear inverse problem under sparsity constraints, the sparsity hypothesis is not so restrictive because many signals/images classes are compressible. We can expect the same phenomenon to

1. To be accurate, the metric is defined between equivalence classes reflecting invariance properties of the family f. For instance, in the case of neural networks, we would like to consider weight rescaling and/or neurons re-arrangement.

happen for neural networks for which such statements are often referred to as "expressivity" or "expressive power". For instance, we can expect to have such guaranties when the neural network approximates a smooth function Boölcskei et al. (2019); Gühring et al. (2019); Gribonval et al. (2019).

1.2. Existing results on the stable recovery

Establishing stable recovery guarantees for neural networks is a difficult subject which has not been addressed very often. The subject remains largely unexplored. To the best of our knowledge conditions guaranteeing the stability property for neural networks have been established in Arora et al. (2014); Brutzkus and Globerson (2017); Li and Yuan (2017); Sedghi and Anandkumar (2014); Zhong et al. (2017); Malgouyres and Landsberg (2016, 2019). A negative statement, exhibiting an unstable configuration when the weights go to infinity is given in Petersen et al. (2019).

Among them, Brutzkus and Globerson (2017); Li and Yuan (2017); Zhong et al. (2017) consider a family of networks with one hidden layer. The article Sedghi and Anandkumar (2014) focuses on the recovery of the parameters defining one layer in a arbitrarily deep networks. The articles Arora et al. (2014); Malgouyres and Landsberg (2016, 2019) consider networks without depth limitation.

In Brutzkus and Globerson (2017), the authors consider the minimization of the population risk. The input is assumed Gaussian and the output is generated by a network involving one linear layer followed by ReLU and a mean. The number of intermediate nodes is smaller than the input size. They provide conditions guaranteeing that, with high probability, a randomly initialized gradient descent algorithm converges to the true parameters. The authors of Li and Yuan (2017) consider a framework similar to Brutzkus and Globerson (2017). They show that the stochastic gradient descent converges to the true solution. In Zhong et al. (2017), the authors consider a non-linear layer followed by a linear layer. The size of the intermediate layer is smaller than the size of the input and the size of the output is 1. They prove that the gradient algorithm minimizing the empirical risk converges to the true parameters, for the particular initialization described in the article.

The authors of Sedghi and Anandkumar (2014) consider a feed-forward neural network and show that, if the input is Gaussian or its distribution is known, a method based on moments and sparse dictionary learning can retrieve the parameters defining the first linear transform. Nothing is said about the stability or the estimation of the other transformations.

The authors of Arora et al. (2014) consider deep feed-forward networks which are very sparse and randomly generated. They show that they can be learned with high probability one layer after another. However, very sparse and randomly generated networks are not used in practice and one might want to study more versatile structures.

The article Malgouyres and Landsberg (2016) studies deep structured linear networks and uses the same tensorial lifting we use here. This result has been extended in Malgouyres and Landsberg (2019), where necessary and sufficient conditions of stable recovery have been established for a general constraint on the parameters defining the network. In the present article, we specialize the analysis to the sparsity constraint. We also obtain necessary and sufficient conditions of stable recovery. However, we obtain a better stability constant (the constant C in Informal Theorem 1). The difference is of the same nature as when the smallest singular value is replaced by a lower RIP constant in compressed sensing Elad (2010). Moreover, in the analysis dedicated to convolutional linear networks, we separate the hypotheses on the data $(x_i)_{i=1..n}$ and the network architecture.

This highlights the importance of having a full row rank X, where X is the concatenation of the data $(x_i)_{i=1..n}$, and shows the role of the smallest singular value of X in this context. These are the two main contributions of the paper.

Finally, denoting H the number of factors/layers, the approach developed in this paper extends to $H \geq 3$ existing compressed sensing results for $H \leq 2$. In particular, when $H = 1$, the considered problems boils down to a compressed sensing problem Elad (2010). When $H = 2$ and when extended to other constraints on the parameters $\mathbf{w}$, the statements apply to already studied problems such as: low rank approximation Candes et al. (2013), Non-negative matrix factorization Lee and Seung (1999); Donoho and Stodden (2003); Laurberg et al. (2008); Arora et al. (2012), dictionary learning Jenatton et al. (2012), phase retrieval Candes et al. (2013), blind deconvolution Ahmed et al. (2014); Choudhary and Mitra (2014); Li et al. (2016). Most of these papers use the same lifting property we are using. They further propose to convexify the problem. A more general bilinear framework is considered in Choudhary and Mitra (2014).

1.3. The considered sparse networks

As in Malgouyres and Landsberg (2019), we consider *structured linear networks*. The layers can be *convolutional* or *feedforward*. The network has at least one hidden layer and can be *deep*. The network is not biased. We give in this section all the notations on networks.

Throughout the paper, we consider $H \geq 2$, $S \geq 2$, $m_0 \dots m_H \in \mathbb{N}$ and write $m_H = m$. We consider a network and assume its architecture fixed. It has $H-1$ hidden layers. The layer 0 corresponds to the inputs, the layer H to the output. The hidden layers correspond to the indexes $1, \cdots, H-1$. For $h \in \{0, \cdots, H\}$, m_h is the size of the layer h. We assume that the whole network is parameterized by an element of $\mathbb{R}^{S \times H}$, say $\mathbf{w} \in \mathbb{R}^{S \times H}$. The architecture of the network is defined by linear mappings

$$\begin{array}{rcl} M_h : \mathbb{R}^S & \longrightarrow & \mathbb{R}^{m_h \times m_{h-1}} \\ w & \longmapsto & M_h(w) \end{array} \tag{1}$$

for $h \in \{1, \cdots, H\}$. For all $h \in \{1, \cdots, H\}$, the linear part of the transformation that maps the content of the layer $h-1$ to the layer h is parameterized by $\mathbf{w}_h \in \mathbb{R}^S$ and is defined by $M_h(\mathbf{w}_h)$. Modeling the architecture of the network with the operators M_h, we can consider many kind of networks. Indeed, depending on the operators M_h, the network can include feedforward layers, convolutional layers and other structured layers tailored to particular structures in the data. The layers might not be fully connected.

The mapping from $\mathbb{R}^{m_0}$ to $\mathbb{R}^m$ defined by the network is called the *prediction* and it is defined for any $x \in \mathbb{R}^{m_0}$ by

$$f_{\mathbf{w}}(x) = M_H(\mathbf{w}_H) M_{H-1}(\mathbf{w}_{H-1}) \cdots M_2(\mathbf{w}_2) M_1(\mathbf{w}_1) x.$$

We use the same notation $f_{\mathbf{w}}$ when applying $f_{\mathbf{w}}$ to every column of $X \in \mathbb{R}^{m_0 \times n}$ and concatenating the results in a matrix in $\mathbb{R}^{m \times n}$. The abuse of notation is not ambiguous, once in context.

Again, the considered networks do not involve activation functions and biases. However, as indicated in Malgouyres and Landsberg (2019), the action of the ReLU activation function multiplies the content of any neuron by an element of $\{0, 1\}$. The choice of the element depends on $\mathbf{w}$ and x. However, considering x fixed, since $\{0, 1\}$ is finite, there is a finite set of possibilities for the action

of the ReLU activation function. Said differently, there is a finite number of possibilities for the choice of the neurons that are kept. Therefore, there exists a partition of $\mathbb{R}^{S\times H}$ such that, on every piece of the partition, the action of ReLU is constant. Therefore, on every piece of the partition, the network is a structured linear network as studied in the present paper. Notice moreover that the analysis in Choromanska et al. (2015a,b) take the expectation of the action of ReLU networks (under an un-realistic independence hypothesis) and obtain a *structured linear network*. Beside, structured linear network are significantly more general than the *deep linear networks* that are often considered (see, among other, Baldi and Hornik (1989); Kawaguchi (2016)).

Throughout the paper, we consider a family of possible supports $\mathcal{M} \subset \mathcal{P}(\{1,\cdots,S\}^H)$, where $\mathcal{P}(\{1,\cdots,S\}^H)$ denotes the set of all possible supports (the parts of $\{1,\cdots,S\}^H$). A classical example is $\mathcal{M} = \{\mathcal{S} | \forall h = 1..h, |\mathcal{S}_h| \leq S'\}$, for a given $S' \leq S$. We constrain the parameter $\mathbf{w}$ to satisfy a sparsity constraint of the form : there exists $\mathcal{S} = (\mathcal{S}_h)_{h=1..H} \in \mathcal{M}$ such that

$$\mathrm{supp}\,(\mathbf{w}) \subset \mathcal{S}$$

(i.e.: $\forall h$, $\mathrm{supp}\,(\mathbf{w}_h) \subset \mathcal{S}_h$). Specializing the analysis to sparsity constraints is one of the main differences between this paper and Malgouyres and Landsberg (2016, 2019). Sparse networks have been considered in many contexts Ranzato et al. (2007, 2008); Lee et al. (2008); Srinivas et al. (2017); Louizos et al. (2018); Zhang et al. (2016); sparse convolutional neural networks have also been considered Liu et al. (2015).

1.4. The solutions of the problem

We assume that data are collected in the columns of matrices $X \in \mathbb{R}^{m_0\times n}$ and $Y \in \mathbb{R}^{m_H\times n}$.

To establish the stability property, we consider throughout the paper $\overline{\mathcal{S}}$ and $\overline{\mathcal{S}}' \in \mathcal{M}$, $\overline{\mathbf{w}}$ and $\overline{\mathbf{w}}' \in \mathbb{R}^{S\times H}$ such that

$$\mathrm{supp}\,(\overline{\mathbf{w}}) \subset \overline{\mathcal{S}} \qquad \text{and} \qquad \mathrm{supp}\,\left(\overline{\mathbf{w}}'\right) \subset \overline{\mathcal{S}}'$$

and for which

$$\|f_{\overline{\mathbf{w}}}(X) - Y\| = \delta \qquad \text{and} \qquad \|f_{\overline{\mathbf{w}}'}(X) - Y\| = \eta \tag{2}$$

are small. Generic parameters are denoted without the over-line: $\mathcal{S}$, $\mathcal{S}'$, $\mathbf{w}$, $\mathbf{w}'$ etc

We want to establish a condition guaranteeing that, up to a multiplicative constant, the distance between such $\overline{\mathbf{w}}$ and $\overline{\mathbf{w}}'$ is upper-bound by $\delta + \eta$. As already said, using a true distance would be too restrictive and the true statements involve a distance between equivalence classes of parameters.

2. Notations and preliminaries on Tensorial Lifting

Set $[\![H]\!] = \{1,\dots,H\}$ and $\mathbb{R}^{S\times H}_* = \{\mathbf{w} \in \mathbb{R}^{S\times H} | \forall h = 1..H, \|\mathbf{w}_h\| \neq 0\}$, where we remind that $\mathbf{w}_h \in \mathbb{R}^S$ contains the parameters defining the transform between layers $h-1$ and h. Define an equivalence relation in $\mathbb{R}^{S\times H}_*$: for any $\mathbf{w}$, $\mathbf{v} \in \mathbb{R}^{S\times H}$, $\mathbf{w} \sim \mathbf{v}$ if and only if there exists $(\lambda_h)_{h=1..H} \in \mathbb{R}^H$ such that

$$\prod_{h=1}^{H} \lambda_h = 1 \qquad \text{and} \qquad \forall h = 1..H, \mathbf{w}_h = \lambda_h \mathbf{v}_h.$$

Denote the equivalence class of $\mathbf{w} \in \mathbb{R}_*^{S\times H}$ by $[\mathbf{w}]$. For any $p \in [1,\infty]$, we denote the usual ℓ^p norm by $\|.\|_p$ and define the mapping $d_p : \left((\mathbb{R}_*^{S\times H}/\sim) \times (\mathbb{R}_*^{S\times H}/\sim)\right) \to \mathbb{R}$ by

$$d_p([\mathbf{w}],[\mathbf{v}]) = \inf_{\substack{\mathbf{w}'\in[\mathbf{w}]\cap\mathbb{R}_{\text{diag}}^{S\times H}\\ \mathbf{v}'\in[\mathbf{v}]\cap\mathbb{R}_{\text{diag}}^{S\times H}}} \|\mathbf{w}'-\mathbf{v}'\|_p \qquad ,\forall \mathbf{w},\, \mathbf{v} \in \mathbb{R}_*^{S\times H}, \tag{3}$$

where

$$\mathbb{R}_{\text{diag}}^{S\times H} = \{\mathbf{w}\in\mathbb{R}_*^{S\times H} | \forall h = 1..H, \|\mathbf{w}_h\|_\infty = \|\mathbf{w}_1\|_\infty\}.$$

It is proved in Malgouyres and Landsberg (2019) that d_p is a metric on $\mathbb{R}_*^{S\times H}/\sim$.

The real valued tensors of order H whose axes are of size S are denoted by $T \in \mathbb{R}^{S\times\ldots\times S}$. The space of tensors is abbreviated $\mathbb{R}^{S^H}$. We say that a tensor $T \in \mathbb{R}^{S^H}$ is of *rank* 1 if and only if there exists a collection of vectors $\mathbf{w} \in \mathbb{R}^{S\times H}$ such that, for any $\mathbf{i} = (i_1,\ldots,i_H) \in [\![S]\!]^H$,

$$T_{\mathbf{i}} = \mathbf{w}_{1,i_1}\ldots\mathbf{w}_{H,i_H}.$$

The set of all the tensors of rank less than 1 is denoted by Σ_1. We denote $\Sigma_2 = \Sigma_1 + \Sigma_1$. Moreover, we parameterize $\Sigma_1 \subset \mathbb{R}^{S^H}$ using the Segre embedding

$$\begin{array}{rccc} P : \mathbb{R}^{S\times H} & \longrightarrow & \Sigma_1 \subset \mathbb{R}^{S^H} \\ \mathbf{w} & \longmapsto & (\mathbf{w}_{1,i_1}\mathbf{w}_{2,i_2}\ldots\mathbf{w}_{H,i_H})_{\mathbf{i}\in[\![S]\!]^H} \end{array} \tag{4}$$

As stated in the next two theorems, we can control the distortion of the distance induced by P and its 'inverse'.

Theorem 1 Stability of $[\mathbf{w}]$ from $P(\mathbf{w})$, see Malgouyres and Landsberg (2019)

Let $\mathbf{w}$ and $\mathbf{w}' \in \mathbb{R}_*^{S\times H}$ be such that $\|P(\mathbf{w}')-P(\mathbf{w})\|_\infty \leq \frac{1}{2}\max\left(\|P(\mathbf{w})\|_\infty, \|P(\mathbf{w}')\|_\infty\right)$. For all $p,q \in [1,\infty]$,

$$d_p([\mathbf{w}],[\mathbf{w}']) \leq 7(HS)^{\frac{1}{p}} \min\left(\|P(\mathbf{w})\|_\infty^{\frac{1}{H}-1}, \|P(\mathbf{w}')\|_\infty^{\frac{1}{H}-1}\right)\|P(\mathbf{w})-P(\mathbf{w}')\|_q. \tag{5}$$

Theorem 2 'Lipschitz' continuity of P, see Malgouyres and Landsberg (2019)

We have for any $q \in [1,\infty]$ and any $\mathbf{w}$ and $\mathbf{w}' \in \mathbb{R}_*^{S\times H}$,

$$\|P(\mathbf{w})-P(\mathbf{w}')\|_q \leq S^{\frac{H-1}{q}} H^{1-\frac{1}{q}} \max\left(\|P(\mathbf{w})\|_\infty^{1-\frac{1}{H}}, \|P(\mathbf{w}')\|_\infty^{1-\frac{1}{H}}\right) d_q([\mathbf{w}],[\mathbf{w}']). \tag{6}$$

The Tensorial Lifting (see Malgouyres and Landsberg (2019)) states that for any $M_1,\ldots,M_H$ and any X there exists a unique linear map

$$\mathcal{A} : \mathbb{R}^{S^H} \longrightarrow \mathbb{R}^{m\times n},$$

such that for all $\mathbf{w} \in \mathbb{R}^{S\times H}$

$$M_H(\mathbf{w}_H)\cdots M_1(\mathbf{w}_1)X = \mathcal{A}P(\mathbf{w}). \tag{7}$$

The intuition leading to this equality is that every entry in $M_H(\mathbf{w}_H)\cdots M_1(\mathbf{w}_1)X$ is a multivariate polynomial whose variables are in $\mathbf{w}$. Moreover, every monomial of the polynomials is of the form

$a_{\mathbf{i}}P(\mathbf{w})_{\mathbf{i}}$ for $\mathbf{i} \in [\![S]\!]^H$, where $a_{\mathbf{i}}$ is a coefficient which depends on $M_1, \ldots, M_H$ and X. The Tensorial Lifting expresses any deep structured linear network using the Segre Embedding and a linear operator $\mathcal{A}$. The Segre embedding is non-linear and might seem difficult to deal with at the first sight, but it is always the same whatever the network architecture, the sparsity pattern, the action of the ReLU activation function... These constituents of the problem only influence the lifting linear operator $\mathcal{A}$.

In the next section, we study what properties of $\mathcal{A}$ are required to obtain the stable recovery. In Section 4, we study these properties when $\mathcal{A}$ corresponds to a sparse convolutional linear network.

3. General conditions for the stable recovery under sparsity constraint

From now on, the analysis differs from the one presented in Malgouyres and Landsberg (2019). It is dedicated to models that enforce sparsity. In this particular situation, we can indeed have a different view of the geometry of the problem. In order to describe it, we first establish some notation.

We define a support by $\mathcal{S} = (\mathcal{S}_h)_{h=1..H}$, with $\mathcal{S}_h \subset [\![S]\!]$, and remind that we denote the set of all supports by $\mathcal{P}([\![S]\!]^H)$ (the parts of $[\![S]\!]^H$). For a given support $\mathcal{S} \in \mathcal{P}([\![S]\!]^H)$, we denote

$$\mathbb{R}^{S\times H}_{\mathcal{S}} = \{\mathbf{w} \in \mathbb{R}^{S\times H} \mid \mathbf{w}_{h,i} = 0, \text{ for all } h = 1..H \text{ and } i \notin \mathcal{S}_h\}$$

(i.e., for all h, $\operatorname{supp}(\mathbf{w}_h) \subset \mathcal{S}_h$) and

$$\mathbb{R}^{S^H}_{\mathcal{S}} = \{T \in \mathbb{R}^{S^H} \mid T_{\mathbf{i}} = 0, \text{ if } \exists h = 1..H, \text{ such that } \mathbf{i}_h \notin \mathcal{S}_h\}.$$

We also denote by $\mathbf{P}_{\mathcal{S}}$ the orthogonal projection from $\mathbb{R}^{S^H}$ onto $\mathbb{R}^{S^H}_{\mathcal{S}}$. It has the closed-form expression: for all $T \in \mathbb{R}^{S^H}$ and all $\mathbf{i} \in [\![S]\!]^H$

$$(\mathbf{P}_{\mathcal{S}}T)_{\mathbf{i}} = \begin{cases} T_{\mathbf{i}} & \text{, if } \mathbf{i} \in \mathcal{S}, \\ 0 & \text{, otherwise.} \end{cases} \tag{8}$$

We consider different operators and define for any $\mathcal{S} \in \mathcal{P}([\![S]\!]^H)$

$$\mathcal{A}_{\mathcal{S}} = \mathcal{A}\mathbf{P}_{\mathcal{S}}. \tag{9}$$

We will use later on that for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$ and for any $\mathbf{w} \in \mathbb{R}^{S\times H}_{\mathcal{S}}$, or any $\mathbf{w} \in \mathbb{R}^{S\times H}_{\mathcal{S}'}$, or any $\mathbf{w} \in \mathbb{R}^{S\times H}_{\mathcal{S}\cup\mathcal{S}'}$, we have

$$\begin{aligned} \mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}P(\mathbf{w}) &= \mathcal{A}P(\mathbf{w}), \\ &= M_H(\mathbf{w}_H)\cdots M_1(\mathbf{w}_1)X. \end{aligned} \tag{10}$$

The introduction of the different operators $\mathcal{A}_{\mathcal{S}}$ leads to an analysis different from the one conducted in Malgouyres and Landsberg (2019). Instead of considering the intersection of one linear space with a subset of Σ_2 (as in Malgouyres and Landsberg (2019)), we consider the intersection of many linear sets (the kernels of the operator $\mathcal{A}_{\mathcal{S}}$) with Σ_1.

The following property will turn out to be necessary and sufficient to guarantee the stable recovery property.

Definition 1 Sparse-Deep-Null Space Property

Let $\gamma \geq 1$ and $\rho > 0$, we say that $\mathcal{A}$ satisfies the sparse-deep-Null Space Property (sparse-deep-NSP) *with constants (γ, ρ) for $\mathcal{M}$ if and only if for all $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, any $T \in P(\mathbb{R}_{\mathcal{S}}^{S \times H}) + P(\mathbb{R}_{\mathcal{S}'}^{S \times H})$ satisfying $\|\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'} T\| \leq \rho$ and any $T' \in \operatorname{Ker}(\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'})$, we have*

$$\|T\| \leq \gamma \|T - \mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} T'\|. \tag{11}$$

Geometrically, the sparse-deep-NSP does not hold when $\mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} \operatorname{Ker}(\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'})$ intersects $P(\mathbb{R}_{\mathcal{S}}^{S \times H}) + P(\mathbb{R}_{\mathcal{S}'}^{S \times H})$ away from the origin or tangentially at 0. It holds when the two sets intersect "transversally" at 0. Despite an apparent abstract nature, we will be able to characterize precisely when the lifting operator corresponding to a convolutional linear network satisfies the sparse-deep-NSP (see Section 4). We will also be able to calculate the constants (γ, ρ).

Proposition 1 Sufficient condition for sparse-deep-NSP

If $\operatorname{Ker}(\mathcal{A}) \cap \mathbb{R}_{\mathcal{S} \cup \mathcal{S}'}^{S^H} = \{0\}$, for all $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, then $\mathcal{A}$ satisfies the sparse-deep-NSP with constants $(\gamma, \rho) = (1, +\infty)$ for $\mathcal{M}$.

Proof In order to prove the proposition, let us consider $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, $T' \in \operatorname{Ker}(\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'})$. We have $\mathcal{A} \mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} T' = 0$ and therefore $\mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} T' \in \operatorname{Ker}(\mathcal{A})$. Moreover, by definition, $\mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} T' \in \mathbb{R}_{\mathcal{S} \cup \mathcal{S}'}^{S^H}$. Therefore, applying the hypothesis of the proposition, we obtain $\mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} T' = 0$ and (11) holds for any T, when $\gamma = 1$. Therefore, $\mathcal{A}$ satisfies the sparse-deep-NSP with constants $(\gamma, \rho) = (1, +\infty)$ for $\mathcal{M}$.

■

If $[\![S]\!]^H \in \mathcal{M}$, the condition becomes $\operatorname{Ker}(\mathcal{A}) = \{0\}$, which is sufficient but obviously not necessary for the sparse-deep-NSP to hold. However, when $\mathcal{M}$ truly imposes sparsity, the condition $\operatorname{Ker}(\mathcal{A}) \cap \mathbb{R}_{\mathcal{S} \cup \mathcal{S}'}^{S^H} = \{0\}$ says that the elements of $\operatorname{Ker}(\mathcal{A})$ shall not be sparse in some (tensorial) way. This nicely generalizes the case $H = 1$.

Definition 2 Deep-lower-RIP constant

There exists a constant $\sigma_{\mathcal{M}} > 0$ such that for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$ and any T in the orthogonal complement of $\operatorname{Ker}(\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'})$

$$\sigma_{\mathcal{M}} \|\mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} T\| \leq \|\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'} T\|. \tag{12}$$

We call $\sigma_{\mathcal{M}}$ a Deep-lower-RIP constant of $\mathcal{A}$ with regard to $\mathcal{M}$.

Proof The existence of $\sigma_{\mathcal{M}}$ is a straightforward consequence of the fact that the restriction of $\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'}$ on the orthogonal complement of $\operatorname{Ker}(\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'})$ is injective. We therefore have for all T in the orthogonal complement of $\operatorname{Ker}(\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'})$

$$\|\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'} T\| \geq \sigma_{\mathcal{S} \cup \mathcal{S}'} \|T\| \geq \sigma_{\mathcal{S} \cup \mathcal{S}'} \|\mathbf{P}_{\mathcal{S} \cup \mathcal{S}'} T\|,$$

where $\sigma_{\mathcal{S} \cup \mathcal{S}'} > 0$ is the smallest non-zero singular value of $\mathcal{A}_{\mathcal{S} \cup \mathcal{S}'}$. The last inequality holds because $\mathbf{P}_{\mathcal{S} \cup \mathcal{S}'}$ is a contraction.

We obtain the existence of $\sigma_{\mathcal{M}}$ by taking the minimum of the constants $\sigma_{\mathcal{S} \cup \mathcal{S}'}$ over the finite family of $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$. ■

Theorem 3 Sufficient condition of stable recovery for structured linear networks

Consider a structured linear network defined by $M_1, \ldots, M_H$, sparsity constraints defined by a family of possible supports $\mathcal{M}$, data X and Y and the operator $\mathcal{A}$ satisfying (7).

Assume $\mathcal{A}$ satisfies the sparse-deep-NSP with the constants $\gamma \geq 1$, $\rho > 0$ for $\mathcal{M}$. For any $\overline{\mathcal{S}} \in \mathcal{M}$, $\overline{\mathbf{w}} \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}}$ and $\overline{\mathcal{S}}' \in \mathcal{M}$, $\overline{\mathbf{w}}' \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}'}$ as in (2) with $\eta + \delta \leq \rho$, we have

$$\|P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}})\| \leq \frac{\gamma}{\sigma_{\mathcal{M}}} (\delta + \eta),$$

where $\sigma_{\mathcal{M}}$ is the Deep-lower-RIP constant of $\mathcal{A}$ with regard to $\mathcal{M}$.

Moreover, if $\frac{\gamma}{\sigma_{\mathcal{M}}} (\delta + \eta) \leq \frac{1}{2} \max(\|P(\overline{\mathbf{w}}')\|_\infty, \|P(\overline{\mathbf{w}})\|_\infty)$, then

$$d_p([\overline{\mathbf{w}}'], [\overline{\mathbf{w}}]) \leq 7(HS)^{\frac{1}{p}} \min\left(\|P(\overline{\mathbf{w}})\|_\infty^{\frac{1}{H}-1}, \|P(\overline{\mathbf{w}}')\|_\infty^{\frac{1}{H}-1} \right) \frac{\gamma}{\sigma_{\mathcal{M}}} (\delta + \eta).$$

Proof Because $\mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}}$ is linear and then because $\overline{\mathbf{w}} \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}}$ and $\overline{\mathbf{w}}' \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}'}$, using (10), we have

$$\begin{aligned} \|\mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}}(P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}}))\| &= \|\mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} P(\overline{\mathbf{w}}') - \mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} P(\overline{\mathbf{w}})\| \\ &= \|\mathcal{A}P(\overline{\mathbf{w}}') - \mathcal{A}P(\overline{\mathbf{w}})\| \\ &\leq \|\mathcal{A}P(\overline{\mathbf{w}}') - X\| + \|\mathcal{A}P(\overline{\mathbf{w}}) - X\| \\ &\leq \delta + \eta \end{aligned} \tag{13}$$

If we further decompose (the decomposition is unique)

$$P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}}) = T + T', \tag{14}$$

where $T' \in \mathrm{Ker}\left(\mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}}\right)$ and T is orthogonal to $\mathrm{Ker}\left(\mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}}\right)$, we have

$$\|\mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}}(P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}}))\| = \|\mathcal{A}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} T\| \geq \sigma_{\mathcal{M}} \|\mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} T\|,$$

where $\sigma_{\mathcal{M}}$ is the Deep-lower-RIP constant of $\mathcal{A}$ with regard to $\mathcal{M}$. Combining with (13), we get

$$\|\mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} T\| \leq \frac{\delta + \eta}{\sigma_{\mathcal{M}}}.$$

Combining this inequality with $\mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} P(\overline{\mathbf{w}}') = P(\overline{\mathbf{w}}')$, $\mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} P(\overline{\mathbf{w}}) = P(\overline{\mathbf{w}})$ and (14), we obtain

$$\begin{aligned} \|P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}}) - \mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} T'\| &= \|\mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} \left(P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}}) - T'\right)\| \\ &= \|\mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} T\| \\ &\leq \frac{\delta + \eta}{\sigma_{\mathcal{M}}}. \end{aligned}$$

Combining the latter inequality with the hypotheses:$\mathcal{A}$ satisfies the sparse-deep-NSP with constants (γ, ρ) for $\mathcal{M}$ and $\delta + \eta \leq \rho$; we have

$$\begin{aligned} \|P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}})\| &\leq \gamma \|P(\overline{\mathbf{w}}') - P(\overline{\mathbf{w}}) - \mathbf{P}_{\overline{\mathcal{S}}' \cup \overline{\mathcal{S}}} T'\| \\ &\leq \gamma \frac{\delta + \eta}{\sigma_{\mathcal{M}}}. \end{aligned}$$

When $\delta + \eta$ satisfy the condition in the theorem, we can apply Theorem 1 and obtain the last inequality. ■

Theorem 3 differs from the analogous theorem in Malgouyres and Landsberg (2019). In particular, it is dedicated to sparsity constraints. The constant of the upper bound is different. We replace the smallest non-zero singular value of an operator by the min, over a finite number of linear space, of the smallest non-zero singular value of the restriction of the operator on the linear space (see Definition 2 and its proof). This is the usual role of the lower-RIP constant in compressed sensing Elad (2010), hence the name Deep-lower-RIP.

One might again ask whether the condition "$\mathcal{A}$ satisfies the sparse-deep-NSP " is sharp or not. As stated in the following theorem, the answer is affirmative.

Theorem 4 Necessary condition for stable recovery for structured linear networks

Consider a structured linear network defined by $M_1, \dots, M_H$, , sparsity constraints defined by a family of possible supports $\mathcal{M}$, data X and Y and the operator $\mathcal{A}$ satisfying (7).

Assume the stability property holds: There exists C and $\delta > 0$ such that for any $\overline{\mathcal{S}} \in \mathcal{M}$ and any $\overline{\mathbf{w}} \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}}$, any $Y = \mathcal{A}P(\overline{\mathbf{w}}) + e$, with $\|e\| \leq \delta$, and any $\overline{\mathcal{S}}' \in \mathcal{M}$ and $\overline{\mathbf{w}}' \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}'}$ such that

$$\|\mathcal{A}P(\overline{\mathbf{w}}') - Y\| \leq \|e\|$$

we have

$$d_2([\overline{\mathbf{w}}'], [\overline{\mathbf{w}}]) \leq C \ \min\left(\|P(\overline{\mathbf{w}})\|_\infty^{\frac{1}{H}-1}, \|P(\overline{\mathbf{w}}')\|_\infty^{\frac{1}{H}-1} \right) \|e\|.$$

Then, $\mathcal{A}$ satisfies the sparse-deep-NSP with constants

$$\gamma = CS^{\frac{H-1}{2}}\sqrt{H}\,\sigma_{max} \qquad \text{and} \qquad \rho = \delta,$$

for $\mathcal{M}$, where σ_{max} is the spectral radius of $\mathcal{A}$.

Proof Let $\overline{\mathcal{S}}$ and $\overline{\mathcal{S}}' \in \mathcal{M}$. Let $\overline{\mathbf{w}} \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}}$ and $\overline{\mathbf{w}}' \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}'}$ be such that $\|\mathcal{A}\left(P(\overline{\mathbf{w}}) - P(\overline{\mathbf{w}}')\right)\| \leq \delta$. We have, using (10),

$$\mathcal{A}\left(P(\overline{\mathbf{w}}) - P(\overline{\mathbf{w}}')\right) = \mathcal{A}_{\overline{\mathcal{S}} \cup \overline{\mathcal{S}}'}\left(P(\overline{\mathbf{w}}) - P(\overline{\mathbf{w}}')\right).$$

Throughout the proof, we also consider $T' \in \mathrm{Ker}\left(\mathcal{A}_{\overline{\mathcal{S}} \cup \overline{\mathcal{S}}'}\right)$. We assume that $\|P(\overline{\mathbf{w}})\|_\infty \leq \|P(\overline{\mathbf{w}}')\|_\infty$. When it is not the case, the proof is analogue. We denote

$$Y = \mathcal{A}P(\overline{\mathbf{w}}') \qquad \text{and} \qquad e = \mathcal{A}P(\overline{\mathbf{w}}') - \mathcal{A}P(\overline{\mathbf{w}}).$$

We have $Y = \mathcal{A}P(\overline{\mathbf{w}}) + e$ with $\|e\| \leq \delta$. Moreover, since $\overline{\mathbf{w}} \in \mathbb{R}^{S \times H}_{\overline{\mathcal{S}}}$, $\|e\| \leq \delta$ and since we obviously have $\|\mathcal{A}P(\overline{\mathbf{w}}') - Y\| \leq \|e\|$, the assumption that the stability property holds guaranties

$$d_2([\overline{\mathbf{w}}'], [\overline{\mathbf{w}}]) \leq C\|P(\overline{\mathbf{w}}')\|_\infty^{\frac{1}{H}-1}\|e\|.$$

Using (10) and the fact that $e = \mathcal{A}_{\overline{\mathcal{S}} \cup \overline{\mathcal{S}}'}(P(\overline{\mathbf{w}}) - P(\overline{\mathbf{w}}'))$, for any $T' \in \mathrm{Ker}\left(\mathcal{A}_{\overline{\mathcal{S}} \cup \overline{\mathcal{S}}'}\right)$

$$\begin{aligned} \|e\| &= \|\mathcal{A}_{\overline{\mathcal{S}} \cup \overline{\mathcal{S}}'}(P(\overline{\mathbf{w}}) - P(\overline{\mathbf{w}}') - T')\|, \\ &\leq \sigma_{max}\|\mathbf{P}_{\overline{\mathcal{S}} \cup \overline{\mathcal{S}}'}(P(\overline{\mathbf{w}}) - P(\overline{\mathbf{w}}') - T')\|, \\ &= \sigma_{max}\|P(\overline{\mathbf{w}}) - P(\overline{\mathbf{w}}') - \mathbf{P}_{\overline{\mathcal{S}} \cup \overline{\mathcal{S}}'}T'\|, \end{aligned}$$

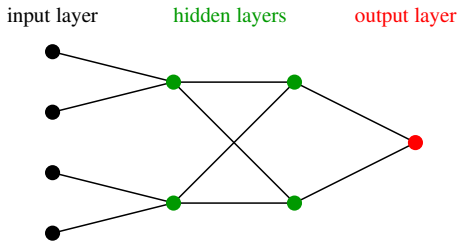


Figure 1: Example of a convolutional linear network. To every edge is attached a convolution kernel. The network does not involve non-linearities or sampling.

where σ_{max} is the spectral radius of $\mathcal{A}$. Therefore,

$$d_2([\overline{\mathbf{w}}'],[\overline{\mathbf{w}}]) \leq C\|P(\overline{\mathbf{w}}')\|_\infty^{\frac{1}{H}-1}\,\sigma_{max}\,\|P(\overline{\mathbf{w}})-P(\overline{\mathbf{w}}')-\mathbf{P}_{\overline{\mathcal{S}}\cup\overline{\mathcal{S}}'}T'\|,$$

Finally, using Theorem 2 and the fact that $\|P(\overline{\mathbf{w}})\|_\infty \leq \|P(\overline{\mathbf{w}}')\|_\infty$, we obtain

$$\begin{aligned}\|P(\overline{\mathbf{w}}')-P(\overline{\mathbf{w}})\| &\leq S^{\frac{H-1}{2}}H^{1-\frac{1}{2}}\|P(\overline{\mathbf{w}}')\|_\infty^{1-\frac{1}{H}}d_2([\overline{\mathbf{w}}'],[\overline{\mathbf{w}}])\\ &\leq CS^{\frac{H-1}{2}}\sqrt{H}\,\sigma_{max}\,\|P(\overline{\mathbf{w}})-P(\overline{\mathbf{w}}')-\mathbf{P}_{\overline{\mathcal{S}}\cup\overline{\mathcal{S}}'}T'\|\\ &= \gamma\|P(\overline{\mathbf{w}})-P(\overline{\mathbf{w}}')-\mathbf{P}_{\overline{\mathcal{S}}\cup\overline{\mathcal{S}}'}T'\|\end{aligned}$$

for $\gamma = CS^{\frac{H-1}{2}}\sqrt{H}\,\sigma_{max}$.

Summarizing, we conclude that under the hypothesis of the theorem: For any $\overline{\mathcal{S}}$ and $\overline{\mathcal{S}}' \in \mathcal{M}$ and any $T \in P(\mathbb{R}^{S\times H}_{\overline{\mathcal{S}}}) + P(\mathbb{R}^{S\times H}_{\overline{\mathcal{S}}'})$ (above $P(\overline{\mathbf{w}})-P(\overline{\mathbf{w}}')$ has the role of T) such that $\|\mathcal{A}T\| = \|\mathcal{A}_{\overline{\mathcal{S}}\cup\overline{\mathcal{S}}'}T\| \leq \delta$, we have for any $T' \in \mathrm{Ker}\left(\mathcal{A}_{\overline{\mathcal{S}}\cup\overline{\mathcal{S}}'}\right)$

$$\|T\| \leq \gamma\|T-\mathbf{P}_{\overline{\mathcal{S}}\cup\overline{\mathcal{S}}'}T'\|.$$

In words, $\mathcal{A}$ satisfies the sparse-deep-NSP for $\mathcal{M}$ with the constants of Theorem 4. ■

4. Application to convolutional linear network under sparsity prior

We consider a sparse convolutional linear network as depicted in Figure 1. Formally, the considered convolutional linear network is defined from a rooted directed acyclic graph $\mathcal{G}(\mathcal{E},\mathcal{N})$ composed of nodes $\mathcal{N}$ and edges $\mathcal{E}$. Each edge connects two nodes. The root of the graph is denoted by r (it contains the output signal) and the set containing all its leaves is denoted by $\mathcal{F}$ (the leaves contain the input signal). We denote by $\mathcal{P}$ the set of all paths connecting the leaves and the root. We assume, without loss of generality, that the length of any path between a leaf and the root is independent of the considered path and equal to $H \geq 0$. We also assume that, for any edge $e \in \mathcal{E}$, the length of the paths separating e and any leaf is constant. This length is called the depth of e. For any $h = 1..H$, we denote the set containing all the edges of depth h, by $\mathcal{E}(h)$.

Moreover, to any edge e is attached a convolution kernel of maximal support $\mathcal{S}_e \subset [\![N]\!]$. We assume (without loss of generality) that $\sum_{e\in\mathcal{E}(h)} |\mathcal{S}_e|$ is independent of h ($|\mathcal{S}_e|$ denotes the cardinality of $\mathcal{S}_e$). We take

$$S = \sum_{e\in\mathcal{E}(1)} |\mathcal{S}_e|.$$

For any edge e, we consider the mapping $\mathcal{T}_e : \mathbb{R}^S \longrightarrow \mathbb{R}^N$ that maps any $w \in \mathbb{R}^S$ into the convolution kernel $\mathcal{T}_e(w) \in \mathbb{R}^N$, attached to the edge e, whose support is $\mathcal{S}_e$. As in the previous section, we assume a sparsity constraint and will only consider a family $\mathcal{M}$ of possible supports $\mathcal{S} \subset [\![S]\!]^H$.

At each h, the convolutional linear network computes, for all $e \in \mathcal{E}(h)$, the convolution between the signal at the origin of e; then, it attaches to any ending node the sum of all the convolutions arriving at that node. Examples of such convolutional linear networks includes wavelets, wavelet packets Mallat (1998) or the fast transforms optimized in Chabiron et al. (2014, 2016). It is the usual convolutional neural network, without bias, in which the activation function is the identity and the supports are potentially scattered and not fixed. It is clear that the operation performed between any pair of consecutive layers depends linearly on parameters $w \in \mathbb{R}^S$. The convolutional linear network therefore depends on parameters $\mathbf{w} \in \mathbb{R}^{S\times H}$ and its prediction takes the form

$$f_{\mathbf{w}}(x) = M_H(\mathbf{w}_H)\cdots M_1(\mathbf{w}_1)x^{|\mathcal{F}|} \qquad \text{, for all } x \in \mathbb{R}^N$$

where the operators M_h satisfy the hypothesis of the present paper and $x^{|\mathcal{F}|} = M_0 x$ where M_0 concatenates vertically $|\mathcal{F}|$ identity matrix of size $N \times N$:

$$M_0 = \begin{pmatrix} Id \\ \vdots \\ Id \end{pmatrix} \in \mathbb{R}^{|\mathcal{F}|N\times N}. \tag{15}$$

Given a sample $(x_i, y_i)_{i=1..n} \in (\mathbb{R}^N \times \mathbb{R}^N)^n$ and reminding that X is the horizontal concatenation of the column vectors x_i, we also denote $X^{|\mathcal{F}|} = M_0 X \in \mathbb{R}^{N|\mathcal{F}|\times n}$.

Given $X^{|\mathcal{F}|}$ and a network architecture, this section applies the results of the preceding sections in order to identify sharp conditions guaranteeing that, for any supports $\overline{\mathcal{S}}$ and $\overline{\mathcal{S}}' \in \mathcal{M}$, any parameters $\overline{\mathbf{w}}$ and $\overline{\mathbf{w}}' \in \mathbb{R}^{S\times H}$ satisfying $\operatorname{supp}(\overline{\mathbf{w}}) \subset \overline{\mathcal{S}}$ and $\operatorname{supp}(\overline{\mathbf{w}})' \subset \overline{\mathcal{S}}'$, and such that

$$\|M_H(\overline{\mathbf{w}}_H)\cdots M_1(\overline{\mathbf{w}}_1)X^{|\mathcal{F}|} - Y\| = \delta \qquad \text{and} \qquad \|M_H(\overline{\mathbf{w}}'_H)\cdots M_1(\overline{\mathbf{w}}'_1)X^{|\mathcal{F}|} - Y\| = \eta$$

are small enough, we can guarantee that $\overline{\mathbf{w}}$ and $\overline{\mathbf{w}}'$ are close to each other.

In order to do so, we first establish a few simple properties and define relevant notations. Notice first that, we can apply the convolutional linear network to any input $u \in \mathbb{R}^{N|\mathcal{F}|}$, where u is the (vertical) concatenation of the signals $u^f \in \mathbb{R}^N$ for $f \in \mathcal{F}$. Therefore, $M_H(\mathbf{w}_H)\cdots M_1(\mathbf{w}_1)$ is the (horizontal) concatenation of $|\mathcal{F}|$ matrices $Z^f \in \mathbb{R}^{N\times N}$ such that

$$M_H(\mathbf{w}_H)\cdots M_1(\mathbf{w}_1)u = \sum_{f\in\mathcal{F}} Z^f u^f \qquad \text{, for all } u \in \mathbb{R}^{N|\mathcal{F}|}. \tag{16}$$

Let us consider the convolutional linear network defined by $\mathbf{w} \in \mathbb{R}^{S\times H}$ as well as $f \in \mathcal{F}$ and $n = 1..N$. The column of $M_H(\mathbf{w}_H)\cdots M_1(\mathbf{w}_1)$ corresponding to the leaf f and the entry n is the

translation by n of

$$\sum_{\mathbf{p}\in\mathcal{P}(f)} \mathcal{T}^{\mathbf{p}}(\mathbf{w}) \tag{17}$$

where $\mathcal{P}(f)$ contains all the paths of $\mathcal{P}$ starting from the leaf f and

$$\mathcal{T}^{\mathbf{p}}(\mathbf{w}) = \mathcal{T}_{e^H}(\mathbf{w}_H) * \ldots * \mathcal{T}_{e^1}(\mathbf{w}_1) \qquad \text{, where } \mathbf{p} = (e^1, \ldots, e^H)$$

and we remind that $\mathcal{T}_{e^h}(\mathbf{w}_h)$ is the convolution kernel on the edge e^h.

We define for any $h = 1..H$ the mapping $\mathbf{e}_h : [\![S]\!] \longrightarrow \mathcal{E}(h)$ which provides for any $i = 1..S$ the unique edge of $\mathcal{E}(h)$ such that the i^{th} entry of $w \in \mathbb{R}^S$ contributes to $\mathcal{T}_{\mathbf{e}_h(i)}(w)$. Also, for any $\mathbf{i} \in [\![S]\!]^H$, we denote $\mathbf{p}_\mathbf{i} = (\mathbf{e}_1(\mathbf{i}_1), \ldots, \mathbf{e}_H(\mathbf{i}_H))$ and, for any $\mathcal{S} \in \mathcal{M}$,

$$\mathbf{I}_\mathcal{S} = \left\{ \mathbf{i} \in [\![S]\!]^H | \mathbf{i} \in \mathcal{S} \text{ and } \mathbf{p}_\mathbf{i} \in \mathcal{P} \right\}.$$

The latter contains all the indices of $\mathcal{S}$ corresponding to a valid path in the network. For any set of parameters $\mathbf{w} \in \mathbb{R}^{S\times H}$ and any path $\mathbf{p} \in \mathcal{P}$, we also denote by $\mathbf{w}^\mathbf{p}$ the restriction of $\mathbf{w}$ to its indices contributing to the kernels on the path $\mathbf{p}$. We also define, for any $\mathbf{i} \in [\![S]\!]^H$, $\mathbf{w}^\mathbf{i} \in \mathbb{R}^{S\times H}$ by

$$\mathbf{w}^\mathbf{i}_{h,j} = \begin{cases} 1 & \text{, if } j = \mathbf{i}_h \\ 0 & \text{otherwise} \end{cases} \qquad \text{, for all } h = 1..H \text{ and } j = 1..S \tag{18}$$

so-that $P(\mathbf{w}^\mathbf{i})$ is a Dirac at position $\mathbf{i}$. The difference between $\mathbf{w}^\mathbf{p}$ and $\mathbf{w}^\mathbf{i}$ will not be ambiguous, once in context.

We can deduce from (17) that, when $\mathbf{i} \in \mathbf{I}_\mathcal{S}$, $M_H(\mathbf{w}^\mathbf{i}_H) \cdots M_1(\mathbf{w}^\mathbf{i}_1)$ simply convolves the entries at one leaf with a Dirac delta function. Therefore, all the entries of $M_H(\mathbf{w}^\mathbf{i}_H) \cdots M_1(\mathbf{w}^\mathbf{i}_1)$ are in $\{0, 1\}$ and we denote $\mathcal{D}_\mathbf{i} = \{(i, j) \in [\![N]\!] \times [\![N|\mathcal{F}|]\!] | \big(M_H(\mathbf{w}^\mathbf{i}_H) \cdots M_1(\mathbf{w}^\mathbf{i}_1)\big)_{i,j} = 1\}$.

We also denote $\mathbb{1} \in \mathbb{R}^S$ a vector of size S with all its entries equal to 1. For any edge $e \in \mathcal{E}$, $\mathbb{1}^e \in \mathbb{R}^S$ consists of zeroes except for the entries contributing to the convolution kernel on the edge e which are equal to 1. For any $\mathcal{S} \subset [\![S]\!]^H$, we define $\mathbb{1}^\mathcal{S} \in \mathbb{R}^{S\times H}$ which consists of zeroes except for the entries corresponding to the indexes in $\mathcal{S}$ which are equal to 1.

The equivalence relationship $\sim$, defined in Section 2, does not suffice to group parameters leading to the same network prediction. Indeed, with the considered convolutional networks, we can rescale the kernels on different path differently. Therefore, we say that two networks sharing the same architecture and defined by the parameters $\mathbf{w}$ and $\mathbf{w}' \in \mathbb{R}^{S\times H}$ are equivalent if and only if

$$\forall \mathbf{p} \in \mathcal{P}, \exists (\lambda_e)_{e\in\mathbf{p}} \in \mathbb{R}^\mathbf{p}, \text{ such that } \prod_{e\in\mathbf{p}} \lambda_e = 1 \text{ and } \forall e \in \mathbf{p}, \mathcal{T}_e(\mathbf{w}') = \lambda_e \mathcal{T}_e(\mathbf{w}).$$

The equivalence class of $\mathbf{w} \in \mathbb{R}^{S\times H}$ is denoted by $\{\mathbf{w}\}$. It is not difficult to see that the prediction of the networks defined by equivalent parameters are identical. For any $p \in [1, +\infty[$, we define

$$\Delta_p(\{\mathbf{w}\}, \{\mathbf{w}'\}) = \Big(\sum_{\mathbf{p}\in\mathcal{P}} d_p\big([\mathbf{w}^\mathbf{p}], [\mathbf{w}'^\mathbf{p}]\big)^p \Big)^{\frac{1}{p}}, \tag{19}$$

where we remind that d_p is defined in (3). Since d_p is a metric, Δ_p is a metric between network classes.

The equivalence classes we have defined do not take into the account the fact it is possible to modify $\mathbf{w}$ in a way that corresponds to permutation of the nodes of the network. Taking into account this invariant is difficult and remains an open question. It has not been addressed in Arora et al. (2014); Brutzkus and Globerson (2017); Li and Yuan (2017); Sedghi and Anandkumar (2014); Zhong et al. (2017); Malgouyres and Landsberg (2016, 2019).

Finally, we remind that because of (7), there exists a unique mapping

$$\mathcal{A} : \mathbb{R}^{S^H} \longrightarrow \mathbb{R}^{N \times n}$$

such that

$$\mathcal{A}P(\mathbf{w}) = M_H(\mathbf{w}_H) \cdots M_1(\mathbf{w}_1) X^{|\mathcal{F}|} \qquad \text{, for all } \mathbf{w} \in \mathbb{R}^{S \times H},$$

where P is the Segre embedding defined in (4).

Proposition 2 Necessary condition of identifiability of a sparse network

Only one of the two following alternatives can occur.

1. *Either, there exist $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$ such that some entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'}) \cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$ do not belong to $\{0,1\}$.*

   When this holds, $\{\overline{\mathbf{w}}\}$ is not always identifiable: there exists $\{\overline{\mathbf{w}}\} \neq \{\overline{\mathbf{w}}'\}$ such that

   $$M_H(\overline{\mathbf{w}}_H) \cdots M_1(\overline{\mathbf{w}}_1)M_0 = M_H(\overline{\mathbf{w}}'_H) \cdots M_1(\overline{\mathbf{w}}'_1)M_0.$$

2. *Or, for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'}) \cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$ belong to $\{0,1\}$. When this holds :*

   (a) *For any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, for any distinct $\mathbf{i} \in \mathcal{S}$ and $\mathbf{i}' \in \mathcal{S}'$, we have $\mathcal{D}_{\mathbf{i}} \cap \mathcal{D}_{\mathbf{i}'} = \emptyset$.*

   (b) *For any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, for any $\mathbf{w} \in \mathbb{R}_{\mathcal{S}}^{S\times H}$ and $\mathbf{w}' \in \mathbb{R}_{\mathcal{S}'}^{S\times H}$ and any distinct $\mathbf{p}$ and $\mathbf{p}' \in \mathcal{P}$, we have*

   $$\text{supp}\left(M_H(\mathbf{w}_H^{\mathbf{p}}) \cdots M_1(\mathbf{w}_1^{\mathbf{p}})M_0\right) \bigcap \text{supp}\left(M_H((\mathbf{w}')_H^{\mathbf{p}'}) \cdots M_1((\mathbf{w}')_1^{\mathbf{p}'})M_0\right) = \emptyset.$$

   (c) *If moreover $|\mathcal{P}| = 1$ and X is full row rank:*

   $$\text{Ker}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right) = \{T \in \mathbb{R}^{S^H} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}.$$

The proof is in Appendix A.

Proposition 2, Item 1, expresses a necessary condition of stability: for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'}) \cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$ belong to $\{0,1\}$. The condition is restrictive but not empty. We will see in the sequel that, when X is full row rank, the condition is sufficient to guarantee the stability. Notice that the condition can be computed at a low cost by applying the network to Dirac delta functions, when $|\mathcal{M}|$ is not too large.

Proposition 3 *If $|\mathcal{P}| = 1$ and X is full row rank. If, for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'}) \cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$ belong to $\{0,1\}$, then $\text{Ker}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right)$ is the orthogonal complement of $\mathbb{R}_{\mathcal{S}\cup\mathcal{S}'}^{S^H}$ and $\mathcal{A}$ satisfies the sparse-deep-NSP with constants $(\gamma,\rho) = (1,+\infty)$ for $\mathcal{M}$. Moreover, $\sigma_{\mathcal{M}} = \sqrt{N}\sigma_{min}(X)$, where $\sigma_{min}(X)$ is the smallest singular value of X, is a deep-lower-RIP constant of $\mathcal{A}$ with regard to $\mathcal{M}$.*

The proof of the proposition is in Appendix B.

Let us remind: we consider X and $Y \in \mathbb{R}^{N\times n}$, $\overline{\mathcal{S}}$ and $\overline{\mathcal{S}}' \in \mathcal{M}$ and parameters $\overline{\mathbf{w}}$ and $\overline{\mathbf{w}}' \in \mathbb{R}^{S\times H}$ satisfying

$$\operatorname{supp}(\overline{\mathbf{w}}) \subset \overline{\mathcal{S}} \qquad \text{and} \qquad \operatorname{supp}(\overline{\mathbf{w}}') \subset \overline{\mathcal{S}}' \tag{20}$$

and denote

$$\delta = \|M_H(\overline{\mathbf{w}}_H)\cdots M_1(\overline{\mathbf{w}}_1)X^{|\mathcal{F}|} - Y\| \qquad \text{and} \qquad \eta = \|M_H(\overline{\mathbf{w}}'_H)\cdots M_1(\overline{\mathbf{w}}'_1)X^{|\mathcal{F}|} - Y\|, \tag{21}$$

where we will assume in the theorem that δ and η are small.

For any path $\mathbf{p} \in \mathcal{P}$, we denote

$$\delta^{\mathbf{p}} = \|M_H(\overline{\mathbf{w}}^{\mathbf{p}}_H)\cdots M_1(\overline{\mathbf{w}}^{\mathbf{p}}_1) - M_H((\overline{\mathbf{w}}')^{\mathbf{p}}_H)\cdots M_1((\overline{\mathbf{w}}')^{\mathbf{p}}_1)\|$$

where we remind that $\overline{\mathbf{w}}^{\mathbf{p}}$ (resp $\overline{\mathbf{w}}'^{\mathbf{p}}$) denotes the restriction of $\overline{\mathbf{w}}$ (resp $\overline{\mathbf{w}}'$) to the path $\mathbf{p}$. Under the hypothesis of the following theorem, when $\delta+\eta$ is small, we will prove that $\delta^{\mathbf{p}}$ is small too for every $\mathbf{p} \in \mathcal{P}$ (see (25)).

Theorem 5 Sufficient condition of stability

Let X, Y, $\overline{\mathcal{S}}$, $\overline{\mathcal{S}}'$, $\overline{\mathbf{w}}$, $\overline{\mathbf{w}}'$, δ and η be as described above (see (20) and (21)). Assume X is full row rank.

If for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'})\cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$ belong to $\{0,1\}$ and if there exists $\varepsilon > 0$ such that for all $e \in \mathcal{E}$, $\|\mathcal{T}_e(\overline{\mathbf{w}})\|_\infty \geq \varepsilon$ and for all $\mathbf{p} \in \mathcal{P}$, $\frac{\delta^{\mathbf{p}}}{\sqrt{N}\sigma_{min}(X)} \leq \frac{1}{2}\max(\|P(\overline{\mathbf{w}}^{\mathbf{p}})\|_\infty, \|P((\overline{\mathbf{w}}')^{\mathbf{p}})\|_\infty)$, then $\overline{\mathbf{w}}$ and $\overline{\mathbf{w}}'$ are close to each other: for any $p \in [1,\infty[$

$$\Delta_p(\{\overline{\mathbf{w}}'\},\{\overline{\mathbf{w}}\}) \leq 7\frac{(HS)^{\frac{1}{p}}}{\sqrt{N}\sigma_{min}(X)^2\varepsilon^{H-1}}(\delta+\eta).$$

We remind that, according to Proposition 2, Item 1, the network is not identifiable when some entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'})\cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$ do not belong to $\{0,1\}$.

The proof of the theorem is in Appendix C.

5. Conclusion

We provide a necessary and sufficient condition of stability for the optimal weights of a sparse linear network. In the general setting, when no assumption is made on the architecture of the network, the stability constant C is improved when compared to un-specified weight models Malgouyres and Landsberg (2019). The gain is comparable to the gain obtained in compressed sensing when replacing the smallest singular value by the lower RIP constant Elad (2010). We then specialize the results to sparse convolutional linear networks. In this analyses, we detail the stability condition in terms of a condition on the architecture and a condition on the sample inputs. The condition on the architecture is restrictive but not empty. The condition on the sample inputs is rather weak and basically requires to have as many (diverse) samples as the dimension of the input space. The constant $\sigma_{min}(X)$ is a key component of the stability constant.

Acknowledgments

Francois Malgouyres is funded by the project DEEL (https://www.deel.ai/) and by the ANITI (https://aniti.univ-toulouse.fr/index.php/en/). The author would like to thank Joseph Landsberg for all his remarks.

References

Arif Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *ICML*, pages 584–592, 2014.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Helmut Boölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 605–614. JMLR. org, 2017.

Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

Olivier Chabiron, François Malgouyres, Jean-Yves Tourneret, and Nicolas Dobigeon. Toward fast transform learning. *International Journal of Computer Vision*, pages 1–22, 2014.

Olivier Chabiron, François Malgouyres, Herwig Wendt, and Jean-Yves Tourneret. Optimization of a fast transform structured as a convolutional tree. *preprint HAL*, (hal-01258514), 2016.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015a.

Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, pages 1756–1760, 2015b.

Sunav Choudhary and Urbashi Mitra. Identifiability scaling laws in bilinear inverse problems. *arXiv preprint arXiv:1402.2637*, 2014.

David L Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? 2003.

Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. working paper or preprint, June 2019. URL https://hal.inria.fr/hal-02117139.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep relu neural networks in $w^{s,p}$ norms. *arXiv preprint arXiv:1902.07896*, 2019. To appear in "Anal. and Appl.".

Rodolphe Jenatton, Rémi Gribonval, and Francis Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. arxiv, 2012.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

Hans Laurberg, Mads Græsbøll Christensen, Mark D Plumbley, Lars Kai Hansen, and Søren Holdt Jensen. Theorems on positive data: On the uniqueness of nmf. *Computational intelligence and neuroscience*, 2008, 2008.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Honglak Lee, Chaitanya Ekanadham, and Andrew Y Ng. Sparse deep belief net model for visual area v2. In *Advances in neural information processing systems*, pages 873–880, 2008.

Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *CoRR*, abs/1606.04933, 2016. URL http://arxiv.org/abs/1606.04933.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–814, 2015.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. In *Internatinal Conference on Learning Representation*, 2018.

François Malgouyres and Joseph Landsberg. On the identifiability and stable recovery of deep/multi-layer structured matrix factorization. In *IEEE, Info. Theory Workshop*, Sept. 2016.

François Malgouyres and Joseph Landsberg. Multilinear compressive sensing and an application to convolutional linear networks. *SIAM Journal on Mathematics of Data Science*, 1(3):446–475, 2019.

Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, Boston, 1998.

Philipp Petersen, Felix Voigtlaender, and Mones Raslan. Topological properties of the set of functions generated by neural networks of fixed size. *arXiv preprint 1806.08459*, 2019. To appear in "Fondation of computational Mathematics".

Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann L Cun. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144, 2007.

Marc'Aurelio Ranzato, Y-Lan Boureau, and Yann L Cun. Sparse feature learning for deep belief networks. In *Advances in neural information processing systems*, pages 1185–1192, 2008.

Hanie Sedghi and Anima Anandkumar. Provable methods for training neural networks with sparse connectivity. In *Deep Learning and representation learning workshop: NIPS*, 2014.

Suraj Srinivas, Akshayvarun Subramanya, and R Venkatesh Babu. Training sparse neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 138–145, 2017.

Shijin Zhang, Zidong Du, Lei Zhang, Huiying Lan, Shaoli Liu, Ling Li, Qi Guo, Tianshi Chen, and Yunji Chen. Cambricon-x: An accelerator for sparse neural networks. In *The 49th Annual IEEE/ACM International Symposium on Microarchitecture*, page 20. IEEE Press, 2016.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4140–4149, 2017.

Appendix A. Proof of Proposition 2

First notice that the entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'})\cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})$ are non-negative integers.

Let us first assume that: There exist $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$ and an entry of

$$M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'})\cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$$

that does not belong to $\{0,1\}$.

Using (15), (16) and (17), we know that there is $n = 1..N$ such that

$$\sum_{f\in\mathcal{F}}\sum_{\mathbf{p}\in\mathcal{P}(f)} \mathcal{T}^{\mathbf{p}}(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_n \geq 2.$$

As a consequence, there is $\mathbf{i}$ and $\mathbf{j} \in \mathcal{S}\cup\mathcal{S}'$ with $\mathbf{i}\neq\mathbf{j}$ and

$$\mathcal{T}^{\mathbf{p_i}}(\mathbf{w^i})_n = \mathcal{T}^{\mathbf{p_j}}(\mathbf{w^j})_n = 1.$$

Therefore, since both $\mathcal{T}^{\mathbf{p_i}}(\mathbf{w^i})$ and $\mathcal{T}^{\mathbf{p_j}}(\mathbf{w^j})$ are Diracs,

$$M_H(\mathbf{w}_H^{\mathbf{i}})\cdots M_1(\mathbf{w}_1^{\mathbf{i}})M_0 = M_H(\mathbf{w}_H^{\mathbf{j}})\cdots M_1(\mathbf{w}_1^{\mathbf{j}})M_0.$$

Since $\mathbf{i} \neq \mathbf{j}$, $\{\mathbf{w}^{\mathbf{i}}\} \neq \{\mathbf{w}^{\mathbf{j}}\}$ and the network is not identifiable. This proves Item 1.

Let us now assume that: For any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of

$$M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'})\cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$$

belong to $\{0,1\}$.

For any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$ and any distinct $\mathbf{i} \in \mathcal{S}$ and $\mathbf{i}' \in \mathcal{S}'$, since $\mathcal{T}^{\mathbf{p}}(\mathbf{w}^{\mathbf{i}})$ and $\mathcal{T}^{\mathbf{p}}(\mathbf{w}^{\mathbf{i}'})$ are Diracs, using (17), (16) and the hypothesis we establish Item 2a.

To prove Item 2b, we consider $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, $\mathbf{w} \in \mathbb{R}^{S\times H}_{\mathcal{S}}$ and $\mathbf{w}' \in \mathbb{R}^{S\times H}_{\mathcal{S}'}$, and distinct $\mathbf{p} \neq \mathbf{p}' \in \mathcal{P}$. We have

$$\operatorname{supp}\left(M_H(\mathbf{w}_H^{\mathbf{p}})\cdots M_1(\mathbf{w}_1^{\mathbf{p}})M_0\right) \subset \operatorname{supp}\left(M_H((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_H^{\mathbf{p}})\cdots M_1((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_1^{\mathbf{p}})M_0\right)$$

and

$$\operatorname{supp}\left(M_H((\mathbf{w}')_H^{\mathbf{p}'})\cdots M_1((\mathbf{w}')_1^{\mathbf{p}'})M_0\right) \subset \operatorname{supp}\left(M_H((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_H^{\mathbf{p}'})\cdots M_1((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_1^{\mathbf{p}'})M_0\right).$$

Using the hypothesis, we know (as in the proof of Item 2a) that

$$\operatorname{supp}\left(M_H((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_H^{\mathbf{p}})\cdots M_1((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_1^{\mathbf{p}})M_0\right) \bigcap \operatorname{supp}\left(M_H((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_H^{\mathbf{p}'})\cdots M_1((\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})_1^{\mathbf{p}'})M_0\right) = \emptyset$$

and conclude that Item 2b holds.

To prove the Item 2c, notice first that $(P(\mathbf{w}^{\mathbf{i}}))_{\mathbf{i}\notin\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}}$ forms a basis of $\{T \in \mathbb{R}^{S^H} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}$. We check using (17) and (9) that, for any $\mathbf{i} \notin \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}$,

$$\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}P(\mathbf{w}^{\mathbf{i}}) = \begin{cases} \mathcal{A}0 = 0 & \text{, if } \mathbf{i} \notin \mathcal{S}\cup\mathcal{S}' \\ M_H(\mathbf{w}_H^{\mathbf{i}})\cdots M_1(\mathbf{w}_1^{\mathbf{i}})X^{|\mathcal{F}|} = 0 & \text{, if } \mathbf{i} \in \mathcal{S}\cup\mathcal{S}' \text{ and } \mathbf{p}_{\mathbf{i}} \notin \mathcal{P}. \end{cases}$$

As a consequence,

$$\{T \in \mathbb{R}^{S^H} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\} \subset \operatorname{Ker}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right). \tag{22}$$

To prove the converse inclusion, we observe that

$$\begin{aligned} \operatorname{rk}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right) &= \dim\left(\operatorname{Span}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}P(\mathbf{w}^{\mathbf{i}})|\mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}\right)\right) \\ &= \dim\left(\operatorname{Span}\left(M_H(\mathbf{w}_H^{\mathbf{i}})\cdots M_1(\mathbf{w}_1^{\mathbf{i}})X^{|\mathcal{F}|}|\mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}\right)\right) \\ &= \dim\left(\operatorname{Span}\left(M_H(\mathbf{w}_H^{\mathbf{i}})\cdots M_1(\mathbf{w}_1^{\mathbf{i}})|\mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}\right)\right) \end{aligned}$$

where the last equality holds because, when $|\mathcal{P}| = 1$, $X^{|\mathcal{F}|} = X$ is full row rank. Moreover, under the hypothesis of the proposition, for any distinct $\mathbf{i}$ and $\mathbf{j} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}$, $\mathcal{D}_{\mathbf{i}} \cap \mathcal{D}_{\mathbf{j}} = \emptyset$, and therefore

$$\dim\left(\operatorname{Span}\left(M_H(\mathbf{w}_H^{\mathbf{i}})\cdots M_1(\mathbf{w}_1^{\mathbf{i}})|\mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}\right)\right) = |\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}|.$$

Therefore, $\operatorname{rk}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right) = |\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}|$; i.e.

$$S^H - \dim(\operatorname{Ker}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right)) = S^H - \dim(\{T \in \mathbb{R}^{S^H} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\})$$

and $\dim(\operatorname{Ker}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right)) = \dim(\{T \in \mathbb{R}^{S^H} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\})$. Combined with (22), we obtain

$$\operatorname{Ker}\left(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}\right) = \{T \in \mathbb{R}^{S^H} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}.$$

This proves Item 2c.

Appendix B. Proof of Proposition 3

The fact that, $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$ is the orthogonal complement of $\mathbb{R}^{S^H}_{\mathcal{S}\cup\mathcal{S}'}$ is a direct consequence of Proposition 2, Item 2c, and the fact that, when $|\mathcal{P}|=1$, $\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}=\mathcal{S}\cup\mathcal{S}'$. We then deduce that, for any $T'\in\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$, $\mathbf{P}_{\mathcal{S}\cup\mathcal{S}'}T'=0$. A straightforward consequence (see (11)) is that $\mathcal{A}$ satisfies the sparse-deep-NSP with constants $(\gamma,\rho)=(1,+\infty)$ for $\mathcal{M}$.

To calculate $\sigma_{\mathcal{M}}$, let us consider $\mathcal{S},\mathcal{S}'\in\mathcal{M}$ and T in the orthogonal complement of $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$. Using Proposition 2, Item 2c, we express T under the form $T=\sum_{\mathbf{i}\in\mathcal{S}\cup\mathcal{S}'}T_{\mathbf{i}}P(\mathbf{w}^{\mathbf{i}})$, where $\mathbf{w}^{\mathbf{i}}$ is defined by (18). Using (9) and (8), the linearity of $\mathcal{A}$ and the fact that, when $|\mathcal{P}|=1$, $X^{|\mathcal{F}|}=X$, we obtain

$$\begin{aligned}\|\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}T\|^2 &= \|\sum_{\mathbf{i}\in\mathbf{I}}T_{\mathbf{i}}\mathcal{A}P(\mathbf{w}^{\mathbf{i}})\|^2,\\ &= \|\sum_{\mathbf{i}\in\mathbf{I}}T_{\mathbf{i}}M_H(\mathbf{w}^{\mathbf{i}})\cdots M_1(\mathbf{w}^{\mathbf{i}})X\|^2,\\ &\geq \sigma^2_{min}(X)\|\sum_{\mathbf{i}\in\mathbf{I}}T_{\mathbf{i}}M_H(\mathbf{w}^{\mathbf{i}})\cdots M_1(\mathbf{w}^{\mathbf{i}})\|^2 \qquad (23)\end{aligned}$$

Let us remind that, applying Proposition 2, Item 2a, the supports of $M_H(\mathbf{w}^{\mathbf{i}})\cdots M_1(\mathbf{w}^{\mathbf{i}})$ (i.e. $\mathcal{D}_{\mathbf{i}}$) and $M_H(\mathbf{w}^{\mathbf{j}})\cdots M_1(\mathbf{w}^{\mathbf{j}})$ (i.e. $\mathcal{D}_{\mathbf{j}}$) are disjoint, when $\mathbf{i}\neq\mathbf{j}$. Let us also add that, since $\mathcal{A}P(\mathbf{w}^{\mathbf{i}})$ is the matrix of a convolution with a Dirac mass, we have $|\mathcal{D}_{\mathbf{i}}|=N$, for all $\mathbf{i}\in\mathbf{I}$. Combining these two properties with (23) and reminding that $\|.\|$ is the Frobinius norm, we obtain

$$\begin{aligned}\|\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}T\|^2 &\geq \sigma^2_{min}(X)\sum_{\mathbf{i}\in\mathbf{I}}T^2_{\mathbf{i}}\|M_H(\mathbf{w}^{\mathbf{i}})\cdots M_1(\mathbf{w}^{\mathbf{i}})\|^2\\ &= \sigma^2_{min}(X)N\sum_{\mathbf{i}\in\mathbf{I}}T^2_{\mathbf{i}}=\sigma^2_{min}(X)N\|T\|^2.\end{aligned}$$

Using that $P_{\mathcal{S}\cup\mathcal{S}'}T=T$, we deduce the value of $\sigma_{\mathcal{M}}$ in the proposition.

Appendix C. Proof of Theorem 5

Let us consider a path $\mathbf{p}\in\mathcal{P}$, using (17), since all the entries of $M_H(\mathbb{1}_H^{\mathcal{S}\cup\mathcal{S}'})\cdots M_1(\mathbb{1}_1^{\mathcal{S}\cup\mathcal{S}'})M_0$ belong to $\{0,1\}$, the restriction of the network to $\mathbf{p}$ satisfy the same property. Therefore, we can apply Proposition 3 and Theorem 3 to the restriction of the convolutional linear network to $\mathbf{p}$, with

$$X'=Id \qquad \text{and} \qquad Y'=M_H((\overline{\mathbf{w}}')^{\mathbf{p}}_H)\cdots M_1((\overline{\mathbf{w}}')^{\mathbf{p}}_1)$$

and obtain, when $\frac{\delta^{\mathbf{p}}}{\sqrt{N}\sigma_{min}(X)}\leq\frac{1}{2}\max(\|P(\overline{\mathbf{w}}^{\mathbf{p}})\|_\infty,\|P((\overline{\mathbf{w}}')^{\mathbf{p}})\|_\infty)$, for any $p\in[1,\infty[$

$$d_p([(\overline{\mathbf{w}}')^{\mathbf{p}}],[\overline{\mathbf{w}}^{\mathbf{p}}])\leq 7\frac{(HS)^{\frac{1}{p}}}{\sqrt{N}\sigma_{min}(X)}\varepsilon^{1-H}\delta^{\mathbf{p}}. \qquad (24)$$

We also have, using the definition of $X^{|\mathcal{F}|}$,

$$\begin{aligned}
\delta+\eta &= \|M_H(\overline{\mathbf{w}}_H)\cdots M_1(\overline{\mathbf{w}}_1)X^{|\mathcal{F}|}-Y\|+\|M_H(\overline{\mathbf{w}}'_H)\cdots M_1(\overline{\mathbf{w}}'_1)X^{|\mathcal{F}|}-Y\| \\
&\geq \|M_H(\overline{\mathbf{w}}_H)\cdots M_1(\overline{\mathbf{w}}_1)M_0X-M_H(\overline{\mathbf{w}}'_H)\cdots M_1(\overline{\mathbf{w}}'_1)M_0X\| \\
&\geq \sigma_{min}(X)\|M_H(\overline{\mathbf{w}}_H)\cdots M_1(\overline{\mathbf{w}}_1)M_0-M_H(\overline{\mathbf{w}}'_H)\cdots M_1(\overline{\mathbf{w}}'_1)M_0\| \\
&= \sigma_{min}(X)\sum_{\mathbf{p}\in\mathcal{P}}\|M_H(\overline{\mathbf{w}}_H^{\mathbf{p}})\cdots M_1(\overline{\mathbf{w}}_1^{\mathbf{p}})M_0-M_H((\overline{\mathbf{w}}')_H^{\mathbf{p}})\cdots M_1((\overline{\mathbf{w}}')_1^{\mathbf{p}})M_0\| \\
&= \sigma_{min}(X)\sum_{\mathbf{p}\in\mathcal{P}}\delta^{\mathbf{p}}
\end{aligned}$$

where the penultimate equality is due to Proposition 2, Item 2b. Combining this inequality, the definition of $\delta^{\mathbf{p}}$ and a standard norm inequality, we obtain

$$\left(\sum_{\mathbf{p}\in\mathcal{P}}(\delta^{\mathbf{p}})^p\right)^{\frac{1}{p}}\leq\sum_{\mathbf{p}\in\mathcal{P}}\delta^{\mathbf{p}}\leq\frac{\delta+\eta}{\sigma_{min}(X)}. \tag{25}$$

Finally, combining the definition of the metric Δ_p (19), (24) and the above inequality we obtain

$$\begin{aligned}
\Delta_p(\{\overline{\mathbf{w}}'\},\{\overline{\mathbf{w}}\}) &\leq 7\frac{(HS)^{\frac{1}{p}}}{\sqrt{N}\sigma_{min}(X)}\varepsilon^{1-H}\left(\sum_{\mathbf{p}\in\mathcal{P}}(\delta^{\mathbf{p}})^p\right)^{\frac{1}{p}}, \\
&\leq 7\frac{(HS)^{\frac{1}{p}}}{\sqrt{N}\sigma_{min}(X)^2}\varepsilon^{1-H}\,(\delta+\eta).
\end{aligned}$$