Stable recovery of deep linear networks under sparsity constraints

François Malgouyres, Joseph Landsberg

Proceedings of Machine Learning Research vol 75:1–15, 2018

Stable recovery of deep 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 Address 1

Joseph Landsberg JML@MATH.TAMU.EDU
Department of Mathematics
Mailstop 3368
Texas A& M University

Abstract

We study a deep linear network expressed under the form of a matrix factorization problem. It takes as input a matrix X obtained by multiplying K matrices (called factors and corresponding to the action of a layer). Each factor is obtained by applying a fixed linear operator to a vector of parameters satisfying a sparsity constraint. In machine learning, the error between the product of the estimated factors and X (i.e. the reconstruction error) relates to the statistical risk. The stable recovery of the parameters defining the factors is required in order to interpret the factors and the intermediate layers of the network.

In this paper, we provide sharp conditions on the network topology under which the error on the parameters defining the factors (i.e. the stability of the recovered parameters) scales linearly with the reconstruction error (i.e. the risk). Therefore, under these conditions on the network topology, any successful learning tasks leads to robust and therefore interpretable layers.

The analysis is based on the recently proposed Tensorial Lifting. The particularity of this paper is to consider a sparse prior. As an illustration, we detail the analysis and provide sharp guarantees for the stable recovery of convolutional linear network under sparsity prior. As expected, the condition are rather strong.

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

1. Introduction

Let $K \in \mathbb{N}^*$, $m_1 \dots m_{K+1} \in \mathbb{N}$, write $m_1 = m$, $m_{K+1} = n$. We impose the factors to be structured matrices defined by a number S of unknown parameters. More precisely, for $k = 1 \dots K$, let

$$\begin{array}{rcl} M_k : \mathbb{R}^S & \longrightarrow & \mathbb{R}^{m_k \times m_{k+1}}, \\ h & \longmapsto & M_k(h) \end{array}$$

be a linear map. We assume that we know the matrix $X \in \mathbb{R}^{m \times n}$ which is provided by

$$X = M_1(\overline{\mathbf{h}}_1) \cdots M_K(\overline{\mathbf{h}}_K) + e, \tag{1}$$

for an unknown error term e, such that $\|e\| \leq \delta$, and parameters $\overline{\mathbf{h}} = (\overline{\mathbf{h}}_k)_{k=1..K} \in \mathbb{R}^{S \times K}$. Moreover, considering a family of possible supports $\mathcal{M}$ (e.g., all the supports of size S', for a

given $S' \leq S$), we assume that the $\overline{\mathbf{h}}$ satisfy a sparsity constraint of the form : there exists $\overline{\mathcal{S}} = (\overline{\mathcal{S}}_k)_{k=1..K} \in \mathcal{M}$ such that $\operatorname{supp}\left(\overline{\mathbf{h}}\right) \subset \overline{\mathcal{S}}$ (i.e.: $\forall k$, $\operatorname{supp}\left(\overline{\mathbf{h}}_k\right) \subset \overline{\mathcal{S}}_k$).

This work investigates necessary and sufficient conditions imposed on the constituents of (1) for which we can (up to obvious scale rearrangement) stably recover the parameters $\overline{\mathbf{h}}$ from X. Beside these conditions, we assume that we have a way to find $\mathcal{S}^* \in \mathcal{M}$ and $\mathbf{h}^* \in \mathbb{R}^{S\times K}_{\mathcal{S}^*}$ such that

$$\eta = \|M_1(\mathbf{h}_1^*)\dots M_K(\mathbf{h}_K^*) - X\| \qquad \text{, is small.} \tag{2}$$

As we will discuss later, at the writing, the success of algorithms for constructing $\mathbf{h}^*$ is mostly supported by empirical evidence and lack theoretical justifications. These aspects of the problem are out of the scope of the present paper. However, in machine learning problems, the reconstruction error η represents the risk. There is therefore no point in analyzing the properties of $\mathbf{h}^*$, if η is large.

The established upper-bound on the recovery error of the parameters linearly depends on $\delta + \eta$. Therefore, when the learning algorithm is successful (i.e. the sum of the risk η and noise δ is sufficiently small), if the deep linear network satisfies the conditions established in this paper the estimation of the parameters is stable. The latter property is required if one wants to interpret the features provided by the machine learning algorithm. That is the main interest of the proposed analysis. Notice that we also establish that the conditions are sharp.

Also, the study considers deep linear networks instead of deep neural networks. As can be deduced from Eldan and Shamir (2016), this significantly diminishes the expressiveness of the network. The main argument for studying deep linear networks (as is done in the present paper) comes from a remark in Safran and Shamir (2016). For the rectified linear unit activation function (ReLU)[1], between each layer every entry is multiplied by an element of the discrete set $\{0, 1\}$. As a consequence, the parameter space $\mathbb{R}^{S\times K}$ can be partitioned into subsets such that, on every subset, the action of the non-linear network is the same deep linear network (i.e. the activation function has a constant action when $\mathbf{h}$ varies in the subset). Therefore, the objective function optimized in deep learning is made of pieces and on every piece it is the objective function of a deep linear network. As a consequence, properties of the objective function for deep neural networks generalize properties of the objective function for deep linear networks. Restricting the analysis to linear networks is legitimate as a step towards the study of deep neural networks.

Notice, that the authors of Choromanska et al. (2015a,b); Kawaguchi (2016) use a different argument but also end-up studying deep linear networks. The simplifying assumption assumes the independence of the activation to the input. Taking the expectation then leads to linear networks that the authors analyse. As explained by the same authors in Choromanska et al. (2015b), this is however a moderatly convincing argument. We prefer to say clearly that we consider deep linear networks.

Finally, $\mathcal{S}^*$ and $\mathbf{h}^*$ are typically found by an algorithm (most often a heuristic) that tries to lower

$$\|M_1(\mathbf{h}_1)\dots M_K(\mathbf{h}_K) - X\|^2 \tag{3}$$

while avoiding overfit. A classical strategy is the *dropout* of Srivastava et al. (2014). This is perfectly compatible with the assumption (2). However, even if we ignore the overfit issue and restrict the analysis to the minimization of (3), we see that it is non-convex. Again, we do not address this minimization issue but there is significant empirical evidence suggesting that (3) can be minimized efficiently in a surprisingly large number of situations. Despite an increasing theoretical activity

1. ReLU is the most common activation function.

related to that question the theory explaining this phenomenon is still far from satisfactory when $K \geq 3$ (see Livni et al. (2014); Haeffele and Vidal (2015); Kawaguchi (2016); Choromanska et al. (2015a,b); Safran and Shamir (2016)).

The approach developed in this paper extends to $K \geq 3$ existing results for $K \leq 2$. In particular, when $K = 1$, the considered problems boils down to a compressed sensing problem Elad (2010). When $K = 2$ and when extended to other constraints on the parameters $\overline{\mathbf{h}}$, 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). The only existing statements when $K \geq 3$ are very recent Malgouyres and Landsberg (2017). They are also applied to deep linear networks but do not include sparsity constraint.

The present work describes an alternative analysis, specialized to sparsity constraints, of the results exposed in Malgouyres and Landsberg (2017). Doing so, we obtain better bounds (defined with an analogue of the lower-RIP) and weaker constraints on the model. Its application to sparse convolutional linear networks leads to simple necessary and sufficient conditions of stable recovery, for a large class of solvers. The stability inequality (see Theorem 5) only involves explicit and simple ingredients of the problem. The condition on the network topology is rather strong but takes an simple format. Implementing a test checking if the condition is met is easy and the test only requires to apply the networks as many times as the network has leaves, for every couple of supports.

2. Notations and preliminaries on Tensorial Lifting

Set $\mathbb{N}_K = \{1, \dots, K\}$ and $\mathbb{R}_*^{S\times K} = \{\mathbf{h} \in \mathbb{R}^{S\times K}, \forall k = 1..K, \|\mathbf{h}_k\| \neq 0\}$. Define an equivalence relation in $\mathbb{R}_*^{S\times K}$: for any $\mathbf{h}$, $\mathbf{g} \in \mathbb{R}^{S\times K}$, $\mathbf{h} \sim \mathbf{g}$ if and only if there exists $(\lambda_k)_{k=1..K} \in \mathbb{R}^K$ such that

$$\prod_{k=1}^{K} \lambda_k = 1 \qquad \text{and} \qquad \forall k = 1..K, \mathbf{h}_k = \lambda_k \mathbf{g}_k.$$

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

$$d_p([\mathbf{h}], [\mathbf{g}]) = \inf_{\substack{\mathbf{h}'\in[\mathbf{h}]\cap\mathbb{R}_{\text{diag}}^{S\times K} \\ \mathbf{g}'\in[\mathbf{g}]\cap\mathbb{R}_{\text{diag}}^{S\times K}}} \|\mathbf{h}' - \mathbf{g}'\|_p \qquad , \forall \mathbf{h}, \mathbf{g} \in \mathbb{R}_*^{S\times K}, \tag{4}$$

where

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

It is proved in Malgouyres and Landsberg (2017) that d_p is a metric on $\mathbb{R}_*^{S\times K}/\sim$.

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

$$T_{\mathbf{i}} = \mathbf{h}_{1,i_1} \dots \mathbf{h}_{K,i_K}.$$

The set of all the tensors of rank 1 is denoted by Σ_1. Moreover, we parametrize $\Sigma_1 \subset \mathbb{R}^{S^K}$ using the Segre embedding

$$\begin{array}{rcl} P : \mathbb{R}^{S\times K} & \longrightarrow & \Sigma_1 \subset \mathbb{R}^{S^K} \\ \mathbf{h} & \longmapsto & (\mathbf{h}_{1,i_1}\mathbf{h}_{2,i_2}\dots\mathbf{h}_{K,i_K})_{\mathbf{i}\in\mathbb{N}_S^K} \end{array} \tag{5}$$

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

Theorem 1 Stability of $[\mathbf{h}]$ from $P(\mathbf{h})$, see Malgouyres and Landsberg (2017)
Let $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{S\times K}$ be such that $\|P(\mathbf{g}) - P(\mathbf{h})\|_\infty \leq \frac{1}{2}\max\left(\|P(\mathbf{h})\|_\infty, \|P(\mathbf{g})\|_\infty\right)$. For all $p, q \in [1,\infty]$,

$$d_p([\mathbf{h}],[\mathbf{g}]) \leq 7(KS)^{\frac{1}{p}} \min\left(\|P(\mathbf{h})\|_\infty^{\frac{1}{K}-1}, \|P(\mathbf{g})\|_\infty^{\frac{1}{K}-1}\right)\|P(\mathbf{h}) - P(\mathbf{g})\|_q. \tag{6}$$

Theorem 2 Lipschitz continuity of P, see Malgouyres and Landsberg (2017)
We have for any $q \in [1,\infty]$ and any $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{S\times K}$,

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_q \leq S^{\frac{K-1}{q}} K^{1-\frac{1}{q}} \max\left(\|P(\mathbf{h})\|_\infty^{1-\frac{1}{K}}, \|P(\mathbf{g})\|_\infty^{1-\frac{1}{K}}\right) d_q([\mathbf{h}],[\mathbf{g}]). \tag{7}$$

The Tensorial Lifting (see Malgouyres and Landsberg (2017)) states that there exists a unique linear map

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

such that for all $\mathbf{h} \in \mathbb{R}^{S\times K}$

$$M_1(\mathbf{h}_1)M_2(\mathbf{h}_2)\dots M_K(\mathbf{h}_K) = \mathcal{A}P(\mathbf{h}). \tag{8}$$

The intuition leading to this equality is that every entry in $M_1(\mathbf{h}_1)M_2(\mathbf{h}_2)\dots M_K(\mathbf{h}_K)$ is a multivariate polynomial whose variables are in $\mathbf{h}$. Moreover, every monomial of the polynomials is of the form $a_{\mathbf{i}}P(\mathbf{h})_{\mathbf{i}}$ for $\mathbf{i} \in \mathbb{N}_S^K$, where $a_{\mathbf{i}}$ is a coefficient depending on $M_1, \dots, M_K$. The great property of the Tensorial Lifting is to express any deep 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 topology, 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 (2017). 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 $\mathcal{S} = (\mathcal{S}_k)_{k=1..K}$, with $\mathcal{S}_k \subset \mathbb{N}_S$, and denote the set of all supports by $\mathcal{P}(\mathbb{N}_S^K)$ (the parts of $\mathbb{N}_S^K$). For a given support $\mathcal{S} \in \mathcal{P}(\mathbb{N}_S^K)$, we denote

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

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

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

We also denote by $\mathbf{P}_{\mathcal{S}}$ the orthogonal projection from $\mathbb{R}^{S^K}$ onto $\mathbb{R}_{\mathcal{S}}^{S^K}$. We trivially have for all $T \in \mathbb{R}^{S^K}$ and all $\mathbf{i} \in \mathbb{N}_S^K$

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

As explained in the introduction, we assume that there exists a known family of admissible supports $\mathcal{M} \subset \mathcal{P}(\mathbb{N}_S^K)$, an unknown support $\overline{\mathcal{S}} \in \mathcal{M}$ and unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}_{\overline{\mathcal{S}}}^{S\times K}$ that we would like to estimate from the noisy matrix product

$$X = M_1(\overline{\mathbf{h}}_1)\dots M_K(\overline{\mathbf{h}}_K) + e. \tag{9}$$

We assume that there exists $\delta \geq 0$ such that the error satisfies

$$\|e\| \leq \delta. \tag{10}$$

Also, we consider an inexact minimization and assume that we have a way to find $\mathcal{S}^* \in \mathcal{M}$ and $\mathbf{h}^* \in \mathbb{R}_{\mathcal{S}^*}^{S\times K}$

$$\eta = \|M_1(\mathbf{h}_1^*)\dots M_K(\mathbf{h}_K^*) - X\| \qquad \text{is small.}$$

We remind that, in machine learning problems, η represents the risk.

In the geometrical view described in the sequel, we consider different linear operators $\mathcal{A}_{\mathcal{S}}$, with $\mathcal{S} \in \mathcal{P}(\mathbb{N}_S^K)$, such that for all $\mathbf{h} \in \mathbb{R}_{\mathcal{S}}^{S\times K}$

$$\mathcal{A}_{\mathcal{S}}P(\mathbf{h}) = M_1(\mathbf{h}_1)\dots M_K(\mathbf{h}_K).$$

In order to achieve that, considering (8), we simply define for any $\mathcal{S} \in \mathcal{P}(\mathbb{N}_S^K)$

$$\mathcal{A}_{\mathcal{S}} = \mathcal{A}\mathbf{P}_{\mathcal{S}}. \tag{11}$$

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

Definition 1 Deep-$\mathcal{M}$-Null Space Property

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

$$\|T\| \leq \gamma\|T - \mathbf{P}_{\mathcal{S}\cup\mathcal{S}'}T'\|. \tag{12}$$

Geometrically, the deep-$\mathcal{M}$-NSP does not hold when $\mathbf{P}_{\mathcal{S}\cup\mathcal{S}'}\operatorname{Ker}(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$ intersects $P(\mathbb{R}_{\mathcal{S}}^{S\times K}) + P(\mathbb{R}_{\mathcal{S}'}^{S\times K})$ 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 deep-$\mathcal{M}$-NSP (see Section 4). We will also be able to calculate the constants (γ, ρ).

Proposition 1 Sufficient condition for deep-$\mathcal{M}$-NSP

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

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, $\mathbf{P}_{\mathcal{S}\cup\mathcal{S}'}T' \in \mathbb{R}_{\mathcal{S}\cup\mathcal{S}'}^{SK}$. Therefore, applying the hypothesis of the proposition, we obtain $\mathbf{P}_{\mathcal{S}\cup\mathcal{S}'}T' = 0$ and (12) holds for any T, when $\gamma = 1$. Therefore, $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-NSP with constants $(\gamma, \rho) = (1, +\infty)$. ■

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

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* $\operatorname{Ker}(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$

$$\sigma_{\mathcal{M}}\|\mathbf{P}_{\mathcal{S}\cup\mathcal{S}'}T\| \leq \|\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}T\|. \tag{13}$$

We call $\sigma_{\mathcal{M}}$ *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 $\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}$. 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 for stable recovery

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

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

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

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

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

Proof We have

$$\begin{aligned}\|\mathcal{A}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}(P(\mathbf{h}^*)-P(\overline{\mathbf{h}}))\| &= \|\mathcal{A}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}P(\mathbf{h}^*)-\mathcal{A}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}P(\overline{\mathbf{h}})\| \\ &= \|\mathcal{A}P(\mathbf{h}^*)-\mathcal{A}P(\overline{\mathbf{h}})\| \\ &\leq \|\mathcal{A}P(\mathbf{h}^*)-X\|+\|\mathcal{A}P(\overline{\mathbf{h}})-X\| \\ &\leq \delta+\eta\end{aligned}$$

If we further decompose (the decomposition is unique)

$$P(\mathbf{h}^*)-P(\overline{\mathbf{h}})=T+T',$$

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

$$\|\mathcal{A}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}(P(\mathbf{h}^*)-P(\overline{\mathbf{h}}))\| = \|\mathcal{A}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}T\| \geq \sigma_{\mathcal{M}}\|\mathbf{P}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}T\|,$$

where $\sigma_{\mathcal{M}}$ is the Deep-lower-RIP constant of $\mathcal{A}$ with regard to $\mathcal{M}$. We finally obtain, since $\mathbf{P}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}P(\mathbf{h}^*)=P(\mathbf{h}^*)$ and $\mathbf{P}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}P(\overline{\mathbf{h}})=P(\overline{\mathbf{h}})$,

$$\|P(\mathbf{h}^*)-P(\overline{\mathbf{h}})-\mathbf{P}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}T'\| = \|\mathbf{P}_{\mathcal{S}^*\cup\overline{\mathcal{S}}}T\| \leq \frac{\delta+\eta}{\sigma_{\mathcal{M}}}.$$

Since $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-NSP with constants (γ,ρ) and $\delta+\eta\leq\rho$, we have

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

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 (2017). In particular, it is dedicated to sparsity constraint with much weaker hypotheses on $\mathcal{A}$. The constant of the upper bound is also different.

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

Theorem 4 Necessary condition for stable recovery

Assume the stable recovery property holds: There exists $\mathcal{M}$, C and $\delta>0$ such that for any $\mathcal{S}\in\mathcal{M}$ and any $\overline{\mathbf{h}}\in\mathbb{R}^{S\times K}_{\mathcal{S}}$, any $X=\mathcal{A}P(\overline{\mathbf{h}})+e$, with $\|e\|\leq\delta$, and any $\mathcal{S}^*\in\mathcal{M}$ and $\mathbf{h}^*\in\mathbb{R}^{S\times K}_{\mathcal{S}^*}$ such that

$$\|\mathcal{A}P(\mathbf{h}^*)-X\|\leq\|e\|$$

we have

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

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

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

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

The proof is very similar to the proof of the Theorem 6, in Malgouyres and Landsberg (2017) and the proof of the analogous converse statement in Cohen et al. (2009). It is provided in Appendix A.

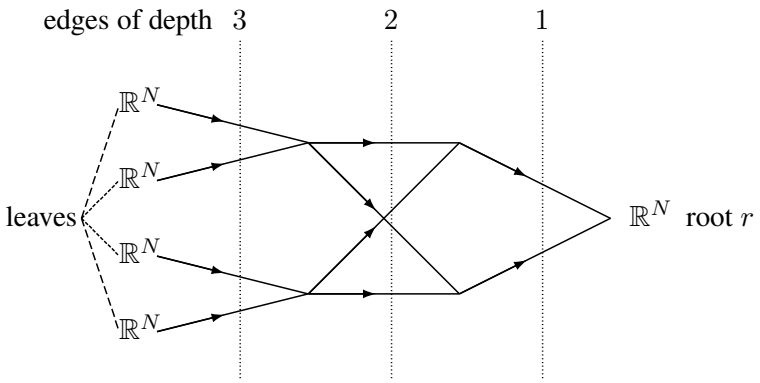

Figure 1: Example of the considered convolutional linear network. To every edge/neuron is attached a convolution kernel. The network does not involve non-linearities or sampling.

4. Application to convolutional linear network under sparsity prior

We consider a sparse convolutional linear network as depicted in Figure 1. The network typically aims at performing a linear analysis or synthesis of a signal living in $\mathbb{R}^N$. 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 and the set containing all its leaves is denoted by $\mathcal{F}$. We denote by $\mathcal{P}_a$ the set of all paths connecting the leaves and the root. We assume, without loss of generality, that the length of any path between any leaf and the root is independent of the considered leaf and equal to some constant $K \geq 0$. We also assume that, for any edge $e \in \mathcal{E}$, the number of edges separating e and the root is the same for all paths between e and r. This length is called the depth of e. For any $k = 1..K$, we denote the set containing all the edges of depth k, by $\mathcal{E}(k)$. For $e \in \mathcal{E}(k)$, we also say that e belongs to the layer k.

Moreover, to any edge e is attached a convolution kernel of maximal support $\mathcal{S}_e \subset \mathbb{N}_N$. We assume (without loss of generality) that $\sum_{e\in\mathcal{E}(k)} |\mathcal{S}_e|$ is independent of k ($|\mathcal{S}_e|$ 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 $h \in \mathbb{R}^S$ into the convolution kernel h_e, attached to the edge e, whose support is $\mathcal{S}_e$. It simply writes at the right location (i.e. those in $\mathcal{S}_e$) the entries of h defining the kernel on the edge e. As in the previous section, we assume a sparsity constraint and will only consider a family $\mathcal{M}$ of possible supports $\mathcal{S} \subset \mathbb{N}_S^K$.

At each layer k, the convolutional linear network computes, for all $e \in \mathcal{E}(k)$, 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 similar to the usual convolutional neural network except that the linear network does not involve any non-linearity and the supports are not fixed. It is clear that the operation performed at any layer depends linearly on the parameters $h \in \mathbb{R}^S$ and that its results serves as inputs for the next layer.

The convolutional linear network therefore depends on parameters $\mathbf{h} \in \mathbb{R}^{S\times K}$ and takes the form

$$X = M_1(\mathbf{h}_1)\dots M_K(\mathbf{h}_K),$$

where the operators M_k satisfy the hypothesis of the present paper.

This section applies the results of the preceding section in order to identify conditions such that any unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}$ satisfying $\operatorname{supp}\left(\overline{\mathbf{h}}\right) \subset \overline{\mathcal{S}}$, for a given $\overline{\mathcal{S}} \in \mathcal{M}$, can be stably recovered from $X = M_1(\overline{\mathbf{h}}_1)\dots M_K(\overline{\mathbf{h}}_K)$ (possibly corrupted by an error).

In order to do so, let us define a few notations. Notice first that, we apply the convolutional linear network to an input $x \in \mathbb{R}^{N|\mathcal{F}|}$, where x is the concatenation of the signals $x^f \in \mathbb{R}^N$ for $f \in \mathcal{F}$. Therefore, X is the (horizontal) concatenation of $|\mathcal{F}|$ matrices $X^f \in \mathbb{R}^{N\times N}$ such that

$$Xx = \sum_{f\in\mathcal{F}} X^f x^f \qquad \text{, for all } x \in \mathbb{R}^{N|\mathcal{F}|}.$$

Let us consider the convolutional linear network defined by $\mathbf{h} \in \mathbb{R}^{S\times K}$ as well as $f \in \mathcal{F}$ and $n = 1..N$. The column of X corresponding to the entry n in the leaf f is the translation by n of

$$\sum_{p\in\mathcal{P}_a(f)} \mathcal{T}^p(\mathbf{h}) \tag{14}$$

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

$$\mathcal{T}^p(\mathbf{h}) = \mathcal{T}_{e^1}(\mathbf{h}_1) * \dots * \mathcal{T}_{e^K}(\mathbf{h}_K) \qquad \text{, where } p = (e^1,\dots,e^K).$$

Moreover, we define for any $k = 1..K$ the mapping $\mathbf{e}_k : \mathbb{N}_S \longrightarrow \mathcal{E}(k)$ which provides for any $i = 1..S$ the unique edge of $\mathcal{E}(k)$ such that the i^{th} entry of $h \in \mathbb{R}^S$ contributes to $\mathcal{T}_{\mathbf{e}_k(i)}(h)$. Also, for any $\mathbf{i} \in \mathbb{N}_S^K$, we denote $\mathbf{p_i} = (\mathbf{e}_1(\mathbf{i}_1),\dots,\mathbf{e}_K(\mathbf{i}_K))$ and, for any $\mathcal{S} \in \mathcal{M}$,

$$\mathbf{I}_\mathcal{S} = \left\{\mathbf{i} \in \mathbb{N}_S^K | \mathbf{i} \in \mathcal{S} \text{ and } \mathbf{p_i} \in \mathcal{P}_a\right\}.$$

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

$$\mathbf{h}^{\mathbf{i}}_{k,j} = \begin{cases} 1 & \text{, if } j = \mathbf{i}_k \\ 0 & \text{otherwise} \end{cases} \qquad \text{, for all } k = 1..K \text{ and } j = 1..S. \tag{15}$$

We can deduce from (14) that, when $\mathbf{i} \in \mathbf{I}_\mathcal{S}$, $\mathcal{A}P(\mathbf{h^i})$ simply convolves the entries at one leaf with a dirac delta function. Thefore, all the entries of $\mathcal{A}P(\mathbf{h^i})$ are in $\{0,1\}$ and we denote $\mathcal{D}_\mathbf{i} = \{(i,j) \in \mathbb{N}_N \times \mathbb{N}_{N|\mathcal{F}|} | \mathcal{A}P(\mathbf{h^i})_{i,j} = 1\}$.

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 corresponding to the edge e which are equal to 1. For any $\mathcal{S} \subset \mathbb{N}_S$, we define $\mathbb{1}^\mathcal{S} \in \mathbb{R}^S$ which consists of zeroes except for the entries corresponding to the indexes in $\mathcal{S}$. For any $\mathbf{p} = (e^1,\dots,e^K) \in \mathcal{P}_a$, the support of $M_1(\mathbb{1}^{e_1})\dots M_K(\mathbb{1}^{e_K})$ is denoted by $\mathcal{D}^\mathbf{p}$.

Finally, we remind that because of (8), there exists a unique mapping

$$\mathcal{A} : \mathbb{R}^{S^K} \longrightarrow \mathbb{R}^{N\times N|\mathcal{F}|}$$

such that

$$\mathcal{A}P(\mathbf{h}) = M_1(\mathbf{h}_1)\dots M_K(\mathbf{h}_K) \qquad \text{, for all } \mathbf{h} \in \mathbb{R}^{S\times K},$$

where P is the Segre embedding defined in (5).

Proposition 2 Necessary condition of identifiability of a sparse network

Either $\mathbb{R}^{S\times K}$ is not identifiable or, for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\dots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$. When the latter holds :

1. *For any distinct $\mathbf{p}$ and $\mathbf{p}' \in \mathcal{P}_a$, we have $\mathcal{D}^{\mathbf{p}} \cap \mathcal{D}^{\mathbf{p}'} = \emptyset$.*
2. $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}) = \{T \in \mathbb{R}^{S^K} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}$.

Proof

Let us assume that: There exist $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$ and an entry of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\dots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ that does not belong to $\{0,1\}$.

Using (14), we know that there is $f \in \mathcal{F}$ and $n = 1..N$ such that

$$\sum_{p\in\mathcal{P}_a(f)} \mathcal{T}^p(\mathbb{1})_n \geq 2.$$

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

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

Therefore,

$$\mathcal{A}P(\mathbf{h}^{\mathbf{i}}) = \mathcal{A}P(\mathbf{h}^{\mathbf{j}})$$

and the network is not identifiable. This proves the first statement.

Let us assume that: For any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\dots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$.

We immediately observe that (14) leads to the item 1 of the Proposition.

To prove the second item, we can easily check that $(P(\mathbf{h}^{\mathbf{i}}))_{\mathbf{i}\notin\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}}$ forms a basis of $\{T \in \mathbb{R}^{S^K} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}$. We can also easily check using (14) and (11) that, for any $\mathbf{i} \notin \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}$,

$$\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}P(\mathbf{h}^{\mathbf{i}}) = \begin{cases} 0 & \text{, if } \mathbf{i} \notin \mathcal{S}\cup\mathcal{S}' \\ M_1(\mathbf{h}^{\mathbf{i}}_1)\dots M_K(\mathbf{h}^{\mathbf{i}}_K) = 0 & \text{, if } \mathbf{i} \in \mathcal{S}\cup\mathcal{S}' \text{ and } \mathbf{p_i} \notin \mathcal{P}_a \end{cases}$$

As a consequence, $\{T \in \mathbb{R}^{S^K} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\} \subset \mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$.

To prove the converse inclusion, we observe that for any distinct $\mathbf{i}$ and $\mathbf{j} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}$, we have $\mathcal{D}_{\mathbf{i}} \cap \mathcal{D}_{\mathbf{j}} = \emptyset$. This implies that

$$\mathrm{rk}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'}) \geq |\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}| = S^K - \dim(\{T \in \mathbb{R}^{S^K} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\}).$$

Finally, we deduce that $\dim(\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})) \leq \dim(\{T \in \mathbb{R}^{S^K} | \forall \mathbf{i} \in \mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}, T_{\mathbf{i}} = 0\})$ and therefore

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

■

Proposition 2 extends Proposition 8 of Malgouyres and Landsberg (2017) by considering several possible supports. Said differently, Proposition 8 of Malgouyres and Landsberg (2017) corresponds to Proposition 2 when $\mathcal{M} = \{\mathbb{N}_S^K\}$.

The interest of the condition in Proposition 2 is that it can easily be computed by applying the network to dirac delta functions, when $|\mathcal{M}|$ is not too large. Notice that, beside the known examples in blind-deconvolution (i.e. when $K = 2$ and $|\mathcal{P}_a| = 1$) Ahmed et al. (2014); Bahmani and Romberg (2015), there are known convolutional linear networks (with $K \geq 2$) that satisfy the condition of the first statement of Proposition 2. For instance, the convolutional linear network corresponding to the un-decimated Haar wavelet[2] transform is a tree and for any of its leaves $f \in \mathcal{F}$, $|\mathcal{P}_a(f)| = 1$. Moreover, the support of the kernel living on the edge e, of depth k, on this path is $\{0, 2^k\}$. It is not difficult to check that the first condition of Proposition 2 holds. Otherwise, it is clear that the necessary condition will be rarely satisfied.

Proposition 3 *If $|\mathcal{P}_a| = 1$ and if, for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'}) \dots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$, then $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$ is the orthogonal complement of $\mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'}$ and $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-NSP with constants $(\gamma,\rho) = (1,+\infty)$. Moreover, the deep-lower-RIP of $\mathcal{A}$ with regard to $\mathcal{M}$ is $\sigma_{\mathcal{M}} = \sqrt{N}$.*

Proof The fact that, $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$ is the orthogonal complement of $\mathbb{R}^{S^K}_{\mathcal{S}\cup\mathcal{S}'}$ is a direct consequence of Proposition 2 and the fact that, when $|\mathcal{P}_a| = 1$, $\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'} = \mathcal{S}\cup\mathcal{S}'$. We then trivialy 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 is that $\mathcal{A}$ satisfies the deep-$\mathcal{M}$-NSP with constants $(\gamma,\rho) = (1,+\infty)$.

To calculate $\sigma_{\mathcal{M}}$, let us consider $\mathcal{S},\mathcal{S}' \in \mathcal{M}$ and T in the orthogonal complement of $\mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$. We express T under the form $T = \sum_{\mathbf{i}\in\mathbf{I}_{\mathcal{S}\cup\mathcal{S}'}} T_{\mathbf{i}} P(\mathbf{h}^{\mathbf{i}})$, where $\mathbf{h}^{\mathbf{i}}$ is defined (15). Let us also remind that, applying Proposition 2, the supports of $\mathcal{A}P(\mathbf{h}^{\mathbf{i}})$ and $\mathcal{A}P(\mathbf{h}^{\mathbf{j}})$ are disjoint, when $\mathbf{i} \neq \mathbf{j}$. Let us finally add that, since $\mathcal{A}P(\mathbf{h}^{\mathbf{j}})$ is the matrix of a convolution with a Dirac mass, we have $|\mathcal{D}_{\mathbf{j}}| = N$. We finally have

$$\begin{aligned} \|\mathcal{A}T\|^2 &= \|\sum_{\mathbf{i}\in\mathbf{I}} T_{\mathbf{i}}\mathcal{A}P(\mathbf{h}^{\mathbf{i}})\|^2, \\ &= N\sum_{\mathbf{i}\in\mathbf{I}} T_{\mathbf{i}}^2 = N\|T\|^2, \end{aligned}$$

from which we deduce the value of $\sigma_{\mathcal{M}}$. ■

In the sequel, we establish stability results for a convolutional linear network estimator. In order to do so, we consider a convolutional linear network of known structure $\mathcal{G}(\mathcal{E},\mathcal{N})$, $(\mathcal{S}_e)_{e\in\mathcal{E}}$ and $\mathcal{M}$. The convolutional linear network is defined by unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}$ satisfying a constraint $\mathrm{supp}\,(\overline{\mathbf{h}}) \subset \overline{\mathcal{S}}$ for an unknown support $\overline{\mathcal{S}} \in \mathcal{M}$. We consider the noisy situation where

$$X = M_1(\overline{\mathbf{h}}_1)\dots M_K(\overline{\mathbf{h}}_K) + e,$$

2. Un-decimated means computed with the "Algorithme à trous", Mallat (1998), Section 5.5.2 and 6.3.2. The Haar wavelet is described in Mallat (1998), Section 7.2.2, p. 247 and Example 7.7, p. 235

with $\|e\| \leq \delta$ and an estimate $\mathbf{h}^* \in \mathbb{R}^{S\times K}$ such that

$$\|M_1(\mathbf{h}_1^*)\dots M_K(\mathbf{h}_K^*) - X\| \leq \eta.$$

The equivalence relationship $\sim$ does not suffise to group parameters leading to the same network action. Indeed, with networks, we can rescale the kernels on different path differently. Therefore, we say that two networks sharing the same topology and defined by the parameters $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}^{S\times K}$ are equivalent if and only if

$$\forall \mathbf{p} \in \mathcal{P}_a, \exists (\lambda_e)_{e\in\mathbf{p}} \in \mathbb{R}^{\mathbf{p}}, \text{ such that } \prod_{e\in\mathbf{p}} \lambda_e = 1 \text{ and } \forall e \in \mathbf{p}, \mathcal{T}_e(\mathbf{g}) = \lambda_e \mathcal{T}_e(\mathbf{h}).$$

We trivially observe that applying the networks defined by equivalent parameters lead to the same result. The equivalence class of $\mathbf{h} \in \mathbb{R}^{S\times K}$ is denoted by $\{\mathbf{h}\}$. For any $p \in [1, +\infty]$, we define

$$\delta_p(\{\mathbf{h}\},\{\mathbf{g}\}) = \left(\sum_{\mathbf{p}\in\mathcal{P}_a} d_p([\mathbf{h}^{\mathbf{p}}],[\mathbf{g}^{\mathbf{p}}])^p\right)^{\frac{1}{p}},$$

where we remind that $\mathbf{h}^{\mathbf{p}}$ (resp $\mathbf{g}^{\mathbf{p}}$) denotes the restriction of $\mathbf{h}$ (resp $\mathbf{g}$) to the path $\mathbf{p}$ and d_p is defined in (4). Since d_p is a metric, we easily prove that δ_p is a metric between network classes.

Theorem 5 *If for any $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$, all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\dots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ belong to $\{0,1\}$ and if there exists $\varepsilon > 0$ such that for all $e \in \mathcal{E}$, $\|\mathcal{T}_e(\overline{\mathbf{h}})\|_\infty \geq \varepsilon$ then*

$$\delta_p(\{\mathbf{h}^*\},\{\overline{\mathbf{h}}\}) \leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}\varepsilon^{K-1}}(\delta+\eta).$$

Proof Let us consider a path $\mathbf{p} \in \mathcal{P}_a$, using (14), since all the entries of $M_1(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})\dots M_K(\mathbb{1}^{\mathcal{S}\cup\mathcal{S}'})$ 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 $\mathbf{p}$ and obtain for any $p \in [1,\infty]$

$$d_p([(\mathbf{h}^*)^{\mathbf{p}}],[\overline{\mathbf{h}}^{\mathbf{p}}]) \leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}}\varepsilon^{1-K}(\delta^{\mathbf{p}}+\eta^{\mathbf{p}}),$$

where $\delta^{\mathbf{p}}$ and $\eta^{\mathbf{p}}$ are the restrictions of the errors on $\mathcal{D}^{\mathbf{p}}$. Finally, using item 1 of Proposition 2

$$\begin{aligned}\delta_p(\{\mathbf{h}^*\},\{\overline{\mathbf{h}}\}) &\leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}}\varepsilon^{1-K}\left(\sum_{\mathbf{p}\in\mathcal{P}_a}(\delta^{\mathbf{p}}+\eta^{\mathbf{p}})^p\right)^{\frac{1}{p}},\\ &\leq 7\frac{(KS)^{\frac{1}{p}}}{\sqrt{N}\varepsilon^{K-1}}(\delta+\eta).\end{aligned}$$

■

Acknowledgments

Joseph Landsberg is supported by NSF DMS-1405348.

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.

Sohail Bahmani and Justin Romberg. Lifting for blind deconvolution in random mask imaging: Identifiability and convex relaxation. *SIAM Journal on Imaging Sciences*, 8(4):2203–2238, 2015.

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.

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best -term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

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.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on Learning Theory*, pages 907–940, 2016.

Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

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.

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.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

François Malgouyres and Joseph Landsberg. Stable recovery of the factors from a deep matrix product and application to convolutional network. *arXiv preprint arXiv:1703.08044*, 2017.

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

Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782, 2016.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014.

Appendix A. Proof of Theorem 4

Proof Let $\mathcal{S}$ and $\mathcal{S}' \in \mathcal{M}$. Let $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}_{\mathcal{S}}$ and $\overline{\mathbf{h}}' \in \mathbb{R}^{S\times K}_{\mathcal{S}'}$ be such that $\|\mathcal{A}\left(P(\overline{\mathbf{h}}) - P(\overline{\mathbf{h}}')\right)\| \leq \delta$. Throughout the proof, we also consider $T' \in \mathrm{Ker}\,(\mathcal{A}_{\mathcal{S}\cup\mathcal{S}'})$. We assume that $\|P(\overline{\mathbf{h}})\|_\infty \leq \|P(\overline{\mathbf{h}}')\|_\infty$. If it is not the case, we simply switch the role of $\overline{\mathbf{h}}$ and $\overline{\mathbf{h}}'$ in the definition of X and e, below. We denote

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

We have $X = \mathcal{A}P(\overline{\mathbf{h}}') + e$ with $\|e\| \leq \delta$. Moreover, when $\mathcal{S}$ and $\overline{\mathbf{h}}$ play the role of $\mathcal{S}^*$ and $\mathbf{h}^*$ in the hypothesis, since $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}_{\mathcal{S}}$ and $\|e\| \leq \delta$, we have

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

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

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

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

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

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

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

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

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

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

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