Multilinear compressive sensing and an application to convolutional linear networks

François Malgouyres, Joseph Landsberg

Multilinear compressive sensing and an application to convolutional linear networks

François Malgouyres[1] and Joseph Landsberg[2]

[1]Institut de Mathématiques de Toulouse ; UMR5219, Université de Toulouse ; CNRS, UPS IMT, F-31062 Toulouse Cedex 9, France
[2]Department of Mathematics , Mailstop 3368, Texas A& M University, College Station, TX 77843-3368

July 3, 2018

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 the layers). Each factor is obtained by applying a fixed linear operator to a vector of parameters satisfying a constraint. The number of factors is not limited. In machine learning, the error between the product of the estimated factors and X (i.e. the reconstruction error) relates to the statistical risk.

In this paper, we provide necessary and sufficient conditions on the network topology under which stable recovery holds. This means that 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 task leads to stably defined features and therefore interpretable layers/network.

In order to do so, we first evaluate how the Segre embedding and its inverse distort distances. Then, we show that any deep linear network can be cast as a generic multilinear problem (that uses the Segre embedding). We call this method *tensorial lifting*. Using the tensorial lifting, we provide necessary and sufficient conditions for the identifiability of the factors (up to a scale rearrangement). We finally provide the necessary and sufficient condition called deep-Null Space Property (because of the analogy with the usual Null Space Property in the compressed sensing framework) which guarantees that the stable recovery of the factors holds.

We illustrate the theory with a practical example where the deep linear network is a convolutional linear network. As expected, the conditions are rather strong but not empty. A simple test on the network topology can be implemented to test if the condition holds.

1 Introduction

1.1 Content of the paper

We consider the following matrix factorization problem: let $K \in \mathbb{N}$, $m_1 \dots m_{K+1} \in \mathbb{N}$, write $m_1 = m$, $m_{K+1} = n$. We are given a matrix $X \in \mathbb{R}^{m \times n}$ which is (approximately) the product of factors $X_k \in \mathbb{R}^{m_k \times m_{k+1}}$:

$$X = X_1 \cdots X_K.$$

This paper investigates models/constraints imposed on the factors X_k for which we can (up to obvious scale rearrangement) identify or stably recover the factors X_k from X.

This question is of paramount importance in many fields including statistics and machine learning, vision, signal and image processing, information theory and numerical linear algebra. As will be detailed in Section 1.2, important problems sharing this structure include: under-determined linear inverse problems, non-negative matrix factorization, dictionary learning, source separation, blind demixing, phase retrieval, low rank approximation, self-calibration, etc. In practical applications X contains data or represents a linear operator. It is often only specified indirectly and/or approximately. Notice that X might be a simple vector.

The motivation for studying identifiability and stable recovery conditions strongly depends on the application and the semantic attached to the factors. We illustrate it with two examples. In most applications coming from signal and image processing, one factor corresponds to an unknown input and the matrix product models an imperfectly known measuring device. In those examples, recovering the factors allows one to calibrate the device and/or recover the input. In most machine learning examples (for instance non-negative matrix factorization), the columns and rows of X correspond to objects of a different nature; X contains a function of these two variables. For instance, for a movie recommendation system, each column correspond to a movie, each row to a person and the corresponding element in X is the palatability of the person for the movie. When factoring X, we find on the columns and rows of the factors features that describe each movie and the palatability of the person for these features. Stable recovery in that context guarantees that the features are stably defined and reliable. This a strong property that allows the interpretation of the features and makes it possible to explain the decisions that are based on the factorization.

We now describe the structures imposed on the factors that we investigate in this paper.

The factors are required to be structured matrices defined by a number $S \in \mathbb{N}$ of unknown parameters. More precisely, for $k = 1 \dots K$, let

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

be a linear map. Doing so, the product $M_1(h_1) \cdots M_K(h_K)$ is a *linear network*.

For instance, the linear maps M_k might for instance map the values in h to prescribed entries of the matrix $M_k(h)$. Such a factorization is usually called a *feed-forward linear network*. Another insightful example is when M_k uses h to construct a convolution matrix (either Toeplitz or circulant). More complex examples include those where the matrix $M_k(h)$ is obtained by combining several smaller Toeplitz or circulant matrices. In the latter case, the product $M_1(h_1) \cdots M_K(h_K)$ applies a *convolutional linear network*. This example is presented in Section 7. Another insightful example is obtained when M_K (or equivalently M_1) is defined by the product $M'_K D$ (or $D M'_1$), where M'_K (or M'_1) has the form (1) and the columns of D and X contain input/output pairs. In this case, the product $M_1(h_1) \cdots M'_K(h_K)$

maps D onto X. Finally, we will see in Section 1.3.2 other linear networks that naturally emerge when considering neural networks using the rectified linear unit activation function (ReLU)[1].

In addition to the structure induced by the operators M_k, we also consider structure imposed on the vectors h. We assume that we know a collection of models $\mathcal{M} = (\mathcal{M}^L)_{L\in\mathbb{N}}$ with the property that for every L, $\mathcal{M}^L \subset \mathbb{R}^{S\times K}$ is a given subset. We will assume that the parameters $\mathbf{h} \in \mathbb{R}^{S\times K}$ defining the factors are such that there exists $L \in \mathbb{N}$ such that $\mathbf{h} \in \mathcal{M}^L$. For instance, the constraint $\mathbf{h} \in \mathcal{M}^L$ might be used to impose sparsity, grouped sparsity or co-sparsity. One might also use the constraint $\mathbf{h} \in \mathcal{M}^L$ to impose non-negativity, orthogonality, equality (in phase retrieval), compactness, etc.

We now precisely state the problem considered in this paper. Or goal is to obtain a statement of the form below for linear networks and convolutional linear networks.

Target theorem (informal version). Stable recovery

We assume a collection of models $\mathcal{M}$ and the mappings M_k of (1) *known. Given a metric*[2] *d between parameter pairs, we establish a necessary and sufficient condition on $\mathcal{M}$ and the mappings M_k guaranteeing that there is $C > 0$ such that for any $X \in \mathbb{R}^{m\times n}$, any $\overline{L}$, L^* and any*

$$(\overline{h}_k)_{k=1..K} \in \mathcal{M}^{\overline{L}} \qquad \text{and} \qquad (h^*_k)_{k=1..K} \in \mathcal{M}^{L^*}$$

such that

$$\delta = \|X - M_1(\overline{h}_1)\cdots M_K(\overline{h}_K)\|,$$

and

$$\eta = \|X - M_1(h^*_1)\cdots M_K(h^*_K)\|.$$

are sufficiently small, we have

$$d(\overline{h}, h^*) \leq C(\delta + \eta). \tag{2}$$

The constituents of this statement are : the metric d; the condition on $\mathcal{M}$ and the mappings M_k; the size condition on δ and η; the constant C. The formal statements for deep linear networks are in Theorem 5, Theorem 6. The statements for convolutional linear networks are in Proposition 8 and Theorem 7.

Let us interpret this statement in the context of signal processing and machine learning. In signal processing, we usually know that $\overline{h}$ exists. The difference $X - M_1(\overline{h}_1)\cdots M_K(\overline{h}_K)$ is an error: typically the sum of a modeling error and noise. The inequality (2) guarantees that, when the condition is satisfied, even an approximative minimizer of

$$\text{argmin}_{L\in\mathbb{N},(h_k)_{k=1..K}\in\mathcal{M}^L} \ \|M_1(h_1)\cdots M_K(h_K) - X\|^2. \tag{3}$$

leads to a solution h^* close $\overline{h}$. When $\delta = 0$ (i.e. the data exactly fit the model and is not noisy) and $\eta = 0$ (i.e. (3) is perfectly solved) this is an *identifiability guarantee*. This is a necessary condition of stable recovery.

In the machine learning context, the interpretation is slightly different. Considering a regression problem, the values δ and η can be interpreted as the risk for the parameters $\overline{h}$ and h^*. The inequality (2) therefore guarantees that the set made of the parameters leading to a small risk has a small diameter. The features defined using such parameters are therefore stably defined. Again, this seems to be the minimal condition allowing the interpretation of these features.

[1] ReLU is the most common activation function.

[2] The metric will take into account layer-wise rescaling.

The minimization problem (3) is non-convex because the product $M_1(h_1)\cdots M_K(h_K)$ is not (jointly) linear. The constraint $\cup_{L\in\mathbb{N}}\mathcal{M}^L$ might also be non-convex. As a consequence, solving or even finding an efficient heuristic solving (3) might be difficult or impossible for some instances of the problem. We do not address the numerical issues related to the minimization of (3). There is significant empirical evidence suggesting that (3) can be minimized efficiently in a surprisingly large number of situations. However, despite an increasing activity related to that question [44, 30, 17, 18, 56], the theory explaining this phenomenon is still far from satisfactory, in particular when $K \geq 3$. Notice that, feed-forward linear networks (which are in general not identifiable) have been closely investigated and the success of the minimization has been explained in that context [7, 8, 35, 63]. In this regard, although the identifiability is desired to interpret the solution, it implies that the minimizer of (3) is unique (up to scale invariance). Intuitively, this is expected to reduce the size of the convergence basin and complicate the numerical resolution of (3). In that sense, a sharp condition of identifiability separates identifiable problems and problems which better lend themselves to global optimization. Outside of this crude intuition, we do not investigate whether (3) can actually be minimized or not.

The main contributions of this paper are:

- In Section 4, we describe the *tensorial lifting*. It expresses any matrix factorization problem of the above structure in a generic multilinear format. The latter composes a linear lifting operator and the Segre embedding.
- In the absence of noise (see Section 5):
  - We establish a simple geometric condition on the intersection of two sets which are necessary and sufficient to guarantee the identifiability of the parameters $\overline{h}_k$ defining the factors (Proposition 7).
  - We provide simpler conditions which involve the rank of the Lifting operator (defined in Section 4) such that:
    - If the rank of the Lifting operator is large (e.g. larger than $2K(S-1)+2$, when $\mathcal{M} = \mathbb{R}^{S\times K}$) and the Lifting operator is random, for almost every Lifting operator, the solution of
    $$M_1(h_1)\cdots M_K(h_K) = X$$
    is identifiable (Theorem 3).
    - If the rank of the Lifting operator is small (e.g. smaller than $2S-1$, when $\mathcal{M} = \mathbb{R}^{S\times K}$), the solution of
    $$M_1(h_1)\cdots M_K(h_K) = X$$
    is not identifiable (Theorem 4);
  - We also provide a simple algorithm to compute the rank of the Lifting operator (Proposition 4).
- Stable recovery statements for the general problem are in Section 6:
  - We define the deep-Null Space Property (Definition 3): a generalization of the usual Null Space Property [20] that also applies to the deep matrix factorization problem.
  - We establish that when the deep-Null Space Property holds we can recover the factors with an accuracy bounded above by the sum of $\delta+\eta$ (see the informal statement above or Theorem 5).

  - We establish the converse statement: if we are able to recover the factors with an accuracy upper bounded by δ then the deep-Null Space Property holds (Theorem 6).
- We specialize the above results to convolutional linear networks and establish a simple condition, that can be computed in many contexts, such that
  - If the condition is satisfied the convolutional linear networks can be stably recovered (see Theorem 7);
  - If the condition is not satisfied, the convolutional linear network is not identifiable (see Proposition 8).

In order to establish these results, we investigate and recall several results on tensors, tensor rank and the Segre embedding (see Section 3). In particular, we investigate how the Segre embedding distort distances.

1.2 Bibliographical landmarks

Matrix factorization problems are ubiquitous in statistics, information theory and data representation. It is not possible to give an bibliography on problems fitting the general framework of the paper. We give however an extensive description of the bibliography on the subject. It shows in particular that the studied framework includes many interesting problems.

To simplify notations, from now on, the parameters defining the factors are gathered in a single matrix and are denoted $\mathbf{h} \in \mathbb{R}^{S\times K}$ (i.e., using bold fonts). The k^{th} vector containing the parameters for the layer k is denoted $\mathbf{h}_k \in \mathbb{R}^S$.

In this section, we distinguish the cases $K = 1$, $K = 2$ and $K \geq 3$.

1.2.1 $K = 1$: Linear inverse problems

The simplest version consists of a model with one layer (i.e., $K = 1$) and $\mathcal{M} = \mathbb{R}^{S\times K}$. Problem (3) is then a linear inverse problem. The data X can be vectorized to form a column vector and the operator M_1 simply multiplies the column vector $\mathbf{h}_1$ by a fixed (rectangular) matrix. Typically, when the linear inverse problem is over-determined, the latter matrix has more rows than columns, the uniqueness of a solution to (3) depends on the column rank of the matrix and the stable recovery constant depends on the smallest singular value of M_1.

When the matrix is not full colum rank, the identifiability and stable recovery for this problem has been intensively studied for many constraints $\mathcal{M}$. In particular, for sparsity constraints this is the compressed/compressive sensing problem (see the seminal articles [10, 23]). Some compressed sensing statements (especially the ones guaranteeing that any minimizer of the ℓ^0 problem stably recovers the unknown) are special cases ($K = 1$) of the statements provided in this paper. We will not perform a complete review on compressed sensing but would like to highlight the Null Space Property described in [20]. The fundamental limits of compressed sensing (for a solution of the ℓ^0 problem) have been analyzed in detail in [9].

Although the main novelty of the paper is to investigate stable recovery properties for any $K \geq 1$, we will always specialize the statements made for $K \geq 1$ in the case $K = 1$. The goal is to illustrate the new statements and to provide a way of comparison with well known results.

1.2.2 $K = 2$: Bilinear inverse problems and bilinear parameterizations

When $k \geq 2$, the problem becomes non-linear because of the product in (3). This significantly complicates the analysis. Let us describe below the main instances studied in the literature when $K = 2$.

Non-negative Matrix factorization (NMF) and low rank prior: In Non-negative matrix factorization [39], the mapping M_1 and M_2 maps the entries in $\mathbf{h}_1$ and $\mathbf{h}_2$ at prescribed locations in the factors (say, one column after another). The constraints $\mathcal{M}$ imposes that all the entries in $\mathbf{h}_1$ and $\mathbf{h}_2$ are non-negative. The (NMF) has been widely used for many applications.

Conditions guaranteeing that the factors provided by the (NMF) identify (not stably recover) the correct factors (up to rescaling and permutation) were first established in the pioneering work [22]. To the best of our knowledge, this is the first paper addressing recovery guarantees for a problem of depth $K = 2$. It emphasizes a separability condition that guarantees identifiability. The proof is purely geometric and relies on the analysis of inclusions of simplicial cones. This result is significantly extended in [38]. In this paper, the continuity of the NMF estimator is established. Concerning computational aspects, the (NMF) is NP-complete [62]. However, under the separability hypothesis of [22], the solution of the (NMF) problem can be computed in polynomial time [4].

Notice that, if we slightly generalize[3] the problem and introduce a linear degradation operator

$$H : \mathbb{R}^{m\times n} \longrightarrow \mathbb{R}^{m\times n}.$$

Using the same mapping M_1 and M_2 as for the (NMF), with $\mathcal{M} = \mathbb{R}^{S\times K}$, but with a small number of lines (resp columns) in $M_2(\mathbf{h}_2)$ (resp. $M_1(\mathbf{h}_1)$). Any solution of the problem

$$(\mathbf{h}_1^*, \mathbf{h}_2^*) \in \operatorname{argmin}_{\mathbf{h}\in\mathbb{R}^{S\times K}} \ \|H(M_1(\mathbf{h}_1)M_2(\mathbf{h}_2)) - X\|^2.$$

leads to a low rank approximation $M_1(\mathbf{h}_1^*)M_2(\mathbf{h}_2^*)$ of an inverse of H, at X. Again, a large corpus of literature exists on the low rank prior [52, 12, 25, 14].

Phase Retrieval: Phase retrieval fits the framework described in the present paper when we take

$$M_1(\mathbf{h}_1) = \operatorname{diag}(\mathcal{F}\mathbf{h}_1) \qquad M_2(\mathbf{h}_2) = (\mathcal{F}\mathbf{h}_2)^*$$

and

$$\mathcal{M} = \{(h,h) \in \mathbb{R}^{S\times K} \mid h \in \mathbb{R}^S\}$$

where S is the size of the signal, $\mathcal{F}$ computes N linear measurements of any element in $\mathbb{R}^S$ (typically Fourier measurements), $\operatorname{diag}(.)$ creates an $N \times N$ diagonal matrix whose diagonal contains the input and * is the (entry-wise) complex conjugate.

The tensorial lifting at the core of the present paper generalizes the lifting used in the inspiring work on PhaseLift [41, 13, 11]. As is often the case when $K = 2$, PhaseLift is a semidefinite program that can be efficiently solved when the unknown is of moderate size. Also, these papers provide conditions on the measurements guaranteeing that the phases are stably recovered by PhaseLift.

The benefit of the generalization introduced with the tensorial lifting is that it applies to any multilinear inverse problem.

[3] The interested readers can check that this generalization only leads to a small change of the Lifting operator introduced in Section 4. It is therefore done at no cost.

Self-calibration and de-mixing Measuring operators often depend linearly on parameters that are not perfectly known. The estimation of these parameters is crucial to restore the data measured by the device. This is the self-calibration problem. This naturally fits the setting of this article: - we let $\mathbf{h}_1$ be the parameters defining the sensing matrix and $M_1(\mathbf{h}_1)$ be the sensing matrix. The parameter $\mathbf{h}_2$ then defines the signal (or signals) contained in the column(s) of $M_2(\mathbf{h}_2)$.

Many instances of this problem have been studied and much progress has been made to obtain algorithms that can be applied to problems of larger and larger size. This leads to a very interesting line of research.

To the best of our knowledge, the first stable recovery statements concern the blind-deconvolution problem. In [2], the authors use a lifting to transform the blind-deconvolution problem into a semidefinite program with an unknown whose size is the product of the sizes[4] of $\mathbf{h}_1$ and $\mathbf{h}_2$. Such problems can be solved for unknowns of moderate size. The authors of [2] provide explicit conditions guaranteeing the stable recovery with high probability. This idea has been generalized and applied to other similar problems in [19, 6]. The authors of [43] consider a significantly more general calibration model. In this model, $M_1(\mathbf{h}_1)$ is diagonal and its diagonal contains the entries of $\mathbf{h}_1$. $M_2(\mathbf{h}_2)$ simply multiplies $\mathbf{h}_2$ by a fixed known matrix (the theorems consider a random matrix). The constraint on the parameters imposes $\mathbf{h}_2$ to be sparse. For this problem, they prove that with high probability the numerical method called SparseLift is stable with a controlled accuracy. SparseLift returns the left and right singular vectors of the solutions of an ℓ^1 optimization problem whose unknown is the same as in [2]. However, solving an ℓ^1 minimization problem is much simpler than a semi-definite problem. This is a very significant practical improvement.

As emphasized in [40] in order to motivate its non-convex approach, the only drawback of the numerical methods described in [2, 43] is their complexity. The extra complexity is due to the fact that they optimize a variable in the product space $\mathbb{R}^{S\times S}$ and then deduce an approximate solution of the "un-lifted" problem. This is what motivates the authors of [40] to propose a non-convex approach. The constructed algorithm provably stably recovers the sensing parameters and the signals with a geometric onvergence rate.

Sparse coding and dictionary learning: Sparse coding and dictionary learning is another kind of bilinear problem (see [53] for an overview on the subject). In that framework, the columns of X contain the data. Most often, people consider two layers: $K = 2$. The layer $M_1(\mathbf{h}_1)$ is an optimized dictionary of atoms defined by the parameters $\mathbf{h}_1$ and each column of $M_2(\mathbf{h}_2)$ contains the code (or coordinates) of the corresponding column in X. Most often, $\mathbf{h}_2$ is assumed sparse.

The identifiability and stable recovery of the factors has been studied in many dictionary learning contexts and provides guarantees on the approximate recovery of both an incoherent dictionary and sparse coefficients when the number of samples is sufficiently large (i.e., n is large, in our setting). In [29], the authors developed local optimality conditions in the noiseless case, as well as sample complexity bounds for local recovery when $M_1(\mathbf{h}_1)$ is square and $M_2(\mathbf{h}_2)$ are iid Bernoulli-Gaussian. This was extended to overcomplete dictionaries in [26] (see also [57] for tight frames) and to the noisy case in [34]. The authors of [59] provide exact recovery results for dictionary learning, when the coefficient matrix has Bernoulli-Gaussian entries and the dictionary matrix has full column rank. This was extended to overcomplete dictionaries in [1] and in [5] but only for approximate recovery. Finally, [28] provides such guarantees under general conditions which cover many practical

[4] With our notations this is simply $S \times S$ but this can be much more favorable.

settings.

Contributions in these frameworks The present article considers the identifiability and stability of the recovery for any $K \geq 1$ in a general and unifying framework. As was already mentioned, we do not investigate the possibility to build proved algorithms. As will appear in the sequel of the paper, the analogue of the lifting at the core of the algorithms described in the above papers (in particular the papers on phase retrieval and self-calibration) is a *tensorial lifting* (see Section 4) and involves tensors that cannot be manipulated in practice. Also, even when we are able to manipulate the tensors, the computation of the best rank 1 approximation of such tensors is an open non-convex problem. Therefore, there is no numerically efficient and reliable way to extract the "un-lifted" parameters from an optimized tensor. Because of that, we have not yet pursued the construction of a numerical scheme based on the tensorial lifting when $K \geq 3$. As was already mentioned, at this writing, the success of algorithms for $K \geq 3$ is mostly supported by empirical evidence. Proving their efficiency is a wide open problem (see [44, 30, 35, 17, 18, 56, 8, 63]). The purpose of the paper is rather to provide guarantees on the stability of the solution when such an empirical success occurs.

The specialization of the presented results to problems with $K = 2$ leads to necessary and sufficient conditions for the stable recovery. This is slightly different from the usual approach. Usually, authors provide sufficient conditions and argue their sharpness by comparing the number of samples required by their method and the information theoretic limit (typically, the number of independent variables of the problem).

It would of course be interesting to see how far it is possible to unify the different problems with $K = 2$ using the framework of this paper. We have however not pursued this route and instead focused on the situation $K \geq 3$.

1.2.3 $K \geq 3$.

To the best of our knowledge, little is known concerning the identifiability and the stability of matrix factorization when $K \geq 3$. The uniqueness of the factorization corresponding to the Fast Fourier Transform was proved in [46]. Other results consider the identifiability of the factors which are sparse and random [51] and might even consider the presence of non-linearities between the layers to include the deep classification architectures [3]. The authors of the present paper have announced preliminary versions of the results described here in [49]. They are significantly extended here.

The difficulties, when $K \geq 3$, come from the fact that some of the tools used for problems with $K = 2$ cannot be used. In particular, we cannot use the usual lifting, the singular value decomposition, the sin-θ theorem in [21]. Often, these tools need to be replaced by analogous objects involving tensors. This complicates the analysis and prohibits the use of numerical schemes that manipulate lifted variables.

1.3 Motivations

1.3.1 Motivating examples with $K \geq 3$

The use of deep matrix factorization is classical. In particular handcrafted deep matrix factorization of a few particular matrices are used in many fields of mathematics and engineering. Most fast transforms, such as the Cooley-Tukey Fast Fourier Transform, the Discrete

Cosine Transform and the Wavelet transform, are matrix products involving a large number of factors.

The construction of deep matrix factorization only started recently (see [15, 16] and references therein). In [15, 16, 48], the authors consider compositions of sparse convolutions organized according to a convolutional tree. In the simplified case studied in [15], X is a vector, the vectors $\mathbf{h}_k$ define the convolution kernels and each operator M_k maps $\mathbf{h}_k$ to a circulant (or block-circulant) matrix. The convolutional networks studied in Section 7 include this example. The first layer corresponds to the coordinates/code of X in the frame obtained by computing the compositions of convolutions along the unique branch of the tree. In [48], the authors consider a factorization involving several sparse layers. In that work, the authors simultaneously estimate the support and the coefficients of each sparse factor. They use this factorization to define an analogue of the Fast Fourier Transform for signals living on graphs [47] and latter reworked on this application using compositions of Given's rotations [27]. In [36], the authors consider a (deep) multi-resolution matrix factorization, inspired by the wavelet decomposition, where the factors are orthogonal and sparse. In [54, 55], the authors consider factors based on householder reflectors and Givens rotation. In [45], the authors study a multi-layer Non-negative matrix factorization. Finally, deep factorizations based on Kroneker products have been considered in [61].

1.3.2 Connections with deep neural networks

In order to clearly express the links between deep neural networks and the deep linear networks considered in (3), we introduce samples $(x_l, y_l)_{1\leq l\leq L}$ according to a distribution law. The regression task aims at constructing a function that predicts y from x for a new realization (x, y) of the same random variable. As usual, we take X whose columns contain the samples y_l. We also use the samples x_l to define the columns of a matrix D and define

$$M_K(\mathbf{h}_K) = M'_K(\mathbf{h}_K)D$$

where M'_k is a matrix describing the linear part of the first layer of the network. Practically, for any layer $k = 1..K$, a feed-forward layers is defined by a mapping M_k that writes the entry of $\mathbf{h}_k$ corresponding to an edge of the network in the corresponding entry in $M_k(\mathbf{h}_k)$. For convolutional layerss, M_k and M'_K concatenate convolution matrices [5] defined by a portion of the entries in $\mathbf{h}_k$. Each convolution matrix is at the location corresponding to a prescribed edge.

However, in addition to this linear structure, deep neural networks usually include a non-linear mapping at each layer. Many non-linearities have been tested and implemented (addition of a constant term, non-linear poolings, activation functions...). These non-linearities and in particular non-linear activation functions are at the core of the efficiency and versatility/expressiveness of deep neural networks (see [24] which nicely illustrates this fact).

In the deep learning community, networks that do not include non-linearities are called *deep linear networks*. They are sometimes studied in place of neural networks [17, 18, 35]. To support this fact, the authors use a moderately convincing argument (see [18]) based on the independence of the action of the activation function to the input.

The main argument for studying deep linear networks comes from a remark in [56]. For the rectified linear unit activation function (ReLU)[6], the action of the ReLU activation

[5] Depending on the situation: Toeplitz, block-Toeplitz, circulant or block-circulant matrices. The matrices often involve a subsampling.

[6] ReLU is the most common activation function.

function at the layer k treats every entry independently of the other entries and multiplies by either 1 or 0. More precisely, the action of the Relu activation function on the layer k applies the mapping $A_k : \mathbb{R}^{m_k \times n} \longmapsto \mathbb{R}^{m_k \times n}$ (where $m_k \times n$ is the size of the layer k) is such that :

$$(A_k M)_{i,j} = a_k(\mathbf{h})_{i,j} M_{i,j}, \qquad \text{for } (i,j) \in [m_k] \times [n]$$

where $a_k(\mathbf{h}) \in \{0,1\}^{m_k \times n}$ is defined by

$$a_k(\mathbf{h})_{i,j} = \begin{cases} 1 & \text{if } \Big(M_{k+1}(\mathbf{h}_{k+1}) A_{k+1} M_{k+2}(\mathbf{h}_{k+2}) \cdots A_{K-1} M_K(\mathbf{h}_K)\Big)_{i,j} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The function

$$\begin{array}{rcl} a_k : \mathbb{R}^{S \times K} & \longrightarrow & \{0,1\}^{m_k \times n} \\ \mathbf{h} & \longmapsto & a_k(\mathbf{h}) \end{array}$$

is piecewise constant because $\{0,1\}^{m_k \times n}$ is finite. As a consequence, $\mathbb{R}^{S \times K}$ is partitioned into subsets and on every subset all the functions a_k are constant. As a consequence, on every subset the action of the non-linear network coincides with the action of a linear network that groups at every layer A_k and M_{k+1}. Further, the landscape of the objective function of the neural network that uses ReLU coincides, on every part of the partition, with the landscape of a linear network.

This is a strong argument in favor of the study deep linear networks. Indeed, it seems impossible to establish properties of the objective function for neural network if we do not already understand them for deep linear networks.

2 Notation and summary of the hypotheses

We continue to use the notation introduced in the introduction. For an integer $k \in \mathbb{N}$, set $[k] = \{1, \cdots, k\}$.

We consider $K \geq 1$ and $S \geq 2$ and real valued tensors of order K whose axes are of size S, denoted $T \in \mathbb{R}^{S \times \cdots \times S}$. The space of tensors is abbreviated $\mathbb{R}^{S^K}$. The entries of T are denoted $T_{i_1,\cdots,i_K}$, where $(i_1, \cdots, i_K) \in [S]^K$. For $\mathbf{i} \in [S]^K$, the entries of $\mathbf{i}$ are $\mathbf{i} = (i_1, \cdots, i_K)$ (for $\mathbf{j} \in [S]^K$ we let $\mathbf{j} = (j_1, \cdots, j_K)$, *etc*). We either write $T_{\mathbf{i}}$ or $T_{i_1,\cdots,i_K}$.

We recall that parameters are denoted $\mathbf{h} \in \mathbb{R}^{S \times K}$ (i.e., using bold fonts). They gather K vectors of size S and the k^{th} vector is denoted $\mathbf{h}_k \in \mathbb{R}^S$. The i^{th} entry of the k^{th} vector is denoted $\mathbf{h}_{k,i} \in \mathbb{R}$. A vector not related to an element in $\mathbb{R}^{S \times K}$ is denoted $h \in \mathbb{R}^S$ (i.e., using a light font). Throughout the paper we assume

$$\mathcal{M} = (\mathcal{M}^L)_{L \in \mathbb{N}}, \text{ with } \mathcal{M}^L \subset \mathbb{R}^{S \times K}.$$

We also assume that, for all $L \in \mathbb{N}$, $\mathcal{M}^L \neq \emptyset$. They can however be equal or constant after a given L'.

All the vector spaces $\mathbb{R}^{S^K}$, $\mathbb{R}^{S \times K}$, $\mathbb{R}^S$ *etc.* are equipped with the usual Euclidean norm. This norm is denoted $\|.\|$ and the scalar product $\langle .,. \rangle$. In the particular case of matrices, $\|.\|$ corresponds to the Frobenius norm. We also use the usual p norm, for $p \in [1, \infty]$, and denote it by $\|.\|_p$. In particular, for $\mathbf{h} \in \mathbb{R}^{S \times K}$ and $T \in \mathbb{R}^{S^K}$, we have for $p < +\infty$

$$\|\mathbf{h}\|_p = \left(\sum_{k=1}^{K} \sum_{i=1}^{S} |\mathbf{h}_{k,i}|^p \right)^{1/p} \qquad , \qquad \|T\|_p = \left(\sum_{\mathbf{i} \in [S]^K} |T_{\mathbf{i}}|^p \right)^{1/p}$$

and

$$\|\mathbf{h}\|_{+\infty} = \max_{\substack{k\in[K]\\ i\in[S]}} |\mathbf{h}_{k,i}| \qquad , \qquad \|T\|_{+\infty} = \max_{\mathbf{i}\in[S]^K} |T_{\mathbf{i}}|$$

Set

$$\mathbb{R}_*^{S\times K} = \{\mathbf{h} \in \mathbb{R}^{S\times K} \mid \forall k \in [K], \|\mathbf{h}_k\| \neq 0\}. \tag{4}$$

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 exist $(\lambda_k)_{k\in[K]} \in \mathbb{R}^K$ such that

$$\prod_{k=1}^{K} \lambda_k = 1 \qquad \text{and} \qquad \mathbf{h}_k = \lambda_k \mathbf{g}_k, \forall k \in [K]. \tag{5}$$

Denote the equivalence class of $\mathbf{h} \in \mathbb{R}_*^{S\times K}$ by $\langle \mathbf{h} \rangle$.

We say that the zero tensor is of rank 0. We say that a non-zero tensor $T \in \mathbb{R}^{S^K}$ is of *rank* 1 (or decomposable) if and only if there exists $\mathbf{h} \in \mathbb{R}_*^{S\times K}$ such that T is the outer product of the vectors $\mathbf{h}_k$, for $k \in [K]$. That is, for any $\mathbf{i} \in [S]^K$,

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

Let $\Sigma_1 \subset \mathbb{R}^{S^K}$ denote the set of tensors of rank 0 or 1.

The *rank* of any tensor $T \in \mathbb{R}^{S^K}$ is defined to be

$$\mathrm{rk}(T) = \min\{r \in \mathbb{N} \mid \text{ there exists } T_1, \cdots, T_r \in \Sigma_1 \text{ such that } T = T_1 + \cdots + T_r\}.$$

For $r \in \mathbb{N}$, let

$$\Sigma_r = \{T \in \mathbb{R}^{S^K} \mid \mathrm{rk}(T) \leq r\}.$$

The $*$ superscript refers to optimal solutions. A set with a $*$ subscript means that 0 is ruled out of the set. In particular, $\Sigma_{1,*}$ denotes the non-zero tensors of rank 1. Attention should be paid to $\mathbb{R}_*^{S\times K}$ (see (4)).

3 Facts on the Segre embedding and tensors of rank 1 and 2

Parametrize $\Sigma_1 \subset \mathbb{R}^{S^K}$ by the map

$$\begin{array}{rccc} P : \mathbb{R}^{S\times K} & \longrightarrow & \Sigma_1 \subset \mathbb{R}^{S^K} \\ \mathbf{h} & \longmapsto & (\mathbf{h}_{1,i_1}\mathbf{h}_{2,i_2}\cdots\mathbf{h}_{K,i_K})_{\mathbf{i}\in[S]^K} \end{array} \tag{6}$$

The map P is called the Segre embedding and is often denoted $\widehat{Seg}$ in the algebraic geometry literature.

Standard Facts:

1. **Identifiability of $\langle \mathbf{h} \rangle$ from $P(\mathbf{h})$:** For $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{S\times K}$, $P(\mathbf{h}) = P(\mathbf{g})$ if and only if $\langle \mathbf{h} \rangle = \langle \mathbf{g} \rangle$.
2. **Geometrical description of $\Sigma_{1,*}$:** $\Sigma_{1,*}$ is a smooth (i.e., C^∞) manifold of dimension $K(S-1)+1$ (see, e.g., [37], chapter 4, pp. 103).

3. **Geometrical description of Σ_2:** We recall that the singular locus $(\Sigma_2)_{sing}$ of (Σ_2) has dimension strictly less than that of Σ_2 and that $\Sigma_2 \backslash (\Sigma_2)_{sing}$ is a smooth manifold. The dimension of $\Sigma_2 \backslash (\Sigma_2)_{sing}$ is $2K(S-1)+2$ when $K > 2$, and is $4(S-1)$ when $K = 2$ (see, e.g., [37], chapter 5).

We can improve Standard Fact 1 and obtain a stability result guaranteeing, that if we know a rank 1 tensor sufficiently close to $P(\mathbf{h})$, we approximately know $\langle \mathbf{h} \rangle$. In order to state this, we need to define a metric on $\mathbb{R}_*^{S\times K}/\sim$ (where $\sim$ is defined by (5)). This has to be considered with care since, whatever $\mathbf{h} \in \mathbb{R}_*^{S\times K}$, the subset $\{h \mid h \in \langle \mathbf{h} \rangle\}$ is not compact. In particular, considering

$$\mathbf{h}'_k = \begin{cases} \lambda\, \mathbf{h}_k & \text{if } k = 1 \\ \lambda^{-\frac{1}{K-1}}\, \mathbf{h}_k & \text{otherwise} \end{cases}$$

when λ goes to infinity, we easily construct examples that make the standard metric on equivalence classes useless[7].

This leads us to consider

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

The interest in this set comes from the fact that, whatever $\mathbf{h} \in \mathbb{R}_*^{S\times K}$, the set $\langle \mathbf{h} \rangle \cap \mathbb{R}_{\text{diag}}^{S\times K}$ is finite. Indeed, if $\mathbf{g} \in \langle \mathbf{h} \rangle \cap \mathbb{R}_{\text{diag}}^{S\times K}$ the $(\lambda_k)_{k\in[K]} \in \mathbb{R}^K$ such that, for all $k \in [K]$, $\mathbf{h}_k = \lambda_k \mathbf{g}_k$ must all satisfy $|\lambda_k| = 1$, i.e. $\lambda_k = \pm 1$.

Definition 1. *For any $p \in [1,\infty]$, we define the mapping $d_p : (\mathbb{R}_*^{S\times K}/\sim \times \mathbb{R}_*^{S\times K}/\sim) \to \mathbb{R}$ by*

$$d_p(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle) = \inf_{\substack{\mathbf{h}' \in \langle \mathbf{h} \rangle \cap \mathbb{R}_{diag}^{S\times K} \\ \mathbf{g}' \in \langle \mathbf{g} \rangle \cap \mathbb{R}_{diag}^{S\times K}}} \|\mathbf{h}' - \mathbf{g}'\|_p \qquad , \forall \mathbf{h}, \mathbf{g} \in \mathbb{R}_*^{S\times K}.$$

Proposition 1. *For any $p \in [1,\infty]$, d_p is a metric on $\mathbb{R}_*^{S\times K}/\sim$.*

The proof is in Appendix 9.1.

Using this metric, we can state that not only $\langle \mathbf{h} \rangle$ is uniquely determined by $P(\mathbf{h})$, but this operation is stable.

Theorem 1. Stability of $\langle \mathbf{h} \rangle$ from $P(\mathbf{h})$

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

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

[7] For instance, if $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{S\times K}$ are such that $\mathbf{h}_1 = \mathbf{g}_1$, we have

$$\inf_{\mathbf{h}' \in \langle \mathbf{h} \rangle, \mathbf{g}' \in \langle \mathbf{g} \rangle} \|\mathbf{h}' - \mathbf{g}'\|_p = 0$$

even though we might have $\mathbf{h}_2 \neq \mathbf{g}_2$ (and therefore $\langle \mathbf{h} \rangle \neq \langle \mathbf{g} \rangle$). This does not define a metric.

Also, when $\mathbf{h}$ and $\mathbf{g}$ are such that $\mathbf{h}_k \neq \mathbf{g}_k$, whatever $k \in [K]$, we have

$$\sup_{\mathbf{h}' \in \langle \mathbf{h} \rangle} \inf_{\mathbf{g}' \in \langle \mathbf{g} \rangle} \|\mathbf{h}' - \mathbf{g}'\|_p = +\infty.$$

Therefore, the Hausdorff distance between $\langle \mathbf{h} \rangle$ and $\langle \mathbf{g} \rangle$ is infinite for almost every pair $(\mathbf{h}, \mathbf{g})$. This metric is therefore not very useful in the present context.

The theorem is proved in Appendix 9.2.

In the final result, the bound established in Theorem 1 plays a role similar to the $sin-\theta$ Theorem of [21] in [43, 13, 2].

The following proposition shows that the upper bound in (7) cannot be improved by a significant factor, in particular when q is large.

Proposition 2. *There exist* $\mathbf{h}$ *and* $\mathbf{g} \in \mathbb{R}_*^{S\times K}$ *such that* $\|P(\mathbf{g})\|_\infty \leq \|P(\mathbf{h})\|_\infty$, $\|P(\mathbf{g})-P(\mathbf{h})\|_\infty \leq \frac{1}{2}\,\|P(\mathbf{h})\|_\infty$ *and*

$$7(KS)^{\frac{1}{p}}\|P(\mathbf{h})\|_\infty^{\frac{1}{K}-1}\|P(\mathbf{h})-P(\mathbf{g})\|_q \leq C_q\, d_p(\langle\mathbf{h}\rangle,\langle\mathbf{g}\rangle),$$

where

$$C_q = \begin{cases} 28(KS)^{\frac{1}{q}} & \text{if } q<+\infty, \\ 28 & \text{if } q=+\infty. \end{cases}$$

The proposition is proved in Appendix 9.3.

As stated in the following theorem, we have a more valuable upper bound in the general case.

Theorem 2. "Lipschitz continuity" of P

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

$$\|P(\mathbf{h})-P(\mathbf{g})\|_q \leq S^{\frac{K-1}{q}}K^{1-\frac{1}{q}}\max\left(\|P(\mathbf{h})\|_\infty^{1-\frac{1}{K}},\|P(\mathbf{g})\|_\infty^{1-\frac{1}{K}}\right)d_q(\langle\mathbf{h}\rangle,\langle\mathbf{g}\rangle). \tag{8}$$

The theorem is proved in Appendix 9.4.

Notice that, considering $\mathbf{h}$ and $\mathbf{g}\in\mathbb{R}^{S\times K}$ such that $\mathbf{h}_{k,i}=1$ and $\mathbf{g}_{k,i}=\varepsilon$, for all $k\in[K]$ and $i\in[S]$ and for a $0<\varepsilon\ll 1$, we easily calculate

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

As a consequence, the upper bound in Theorem 2 is tight up to at most a factor K.

4 The tensorial lifting

The following proposition is clear (it can be shown by induction on K):

Proposition 3. *The entries of the matrix*

$$M_1(\mathbf{h}_1)M_2(\mathbf{h}_2)\cdots M_K(\mathbf{h}_K)$$

are multivariate polynomials whose variables are the entries of $\mathbf{h}\in\mathbb{R}^{S\times K}$. *Moreover, every entry is the sum of monomials of degree K. Each monomial is a constant times* $\mathbf{h}_{1,i_1}\cdots\mathbf{h}_{K,i_K}$, *for some* $\mathbf{i}\in[S]^K$.

Notice that any monomial $\mathbf{h}_{1,i_1}\cdots\mathbf{h}_{K,i_K}$ is the entry $P(\mathbf{h})_\mathbf{i}$ in the tensor $P(\mathbf{h})$. Therefore every polynomial in the previous proposition takes the form $\sum_{\mathbf{i}\in[S]^K}c_\mathbf{i}P(\mathbf{h})_\mathbf{i}$ for some constants $(c_\mathbf{i})_{\mathbf{i}\in[S]^K}$ independent of $\mathbf{h}$. In words, every entry of the matrix $M_1(\mathbf{h}_1)M_2(\mathbf{h}_2)\cdots M_K(\mathbf{h}_K)$ is obtained by applying a linear form to $P(\mathbf{h})$. Moreover, the polynomial coefficients defining the linear form are uniquely determined by the linear maps $M_1,\cdots,M_K$. This leads to the following statement.

Corollary 1. Tensorial Lifting

Let M_k, $k \in [K]$ be as in (1)*. The map*

$$(\mathbf{h}_1, \dots, \mathbf{h}_K) \longmapsto M_1(\mathbf{h}_1) M_2(\mathbf{h}_2) \cdots M_K(\mathbf{h}_K),$$

uniquely determines a linear map

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

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

$$M_1(\mathbf{h}_1) M_2(\mathbf{h}_2) \cdots M_K(\mathbf{h}_K) = \mathcal{A}P(\mathbf{h}). \tag{9}$$

We call (9) and its use the tensorial lifting. When $K = 1$, we simply have $\mathcal{A} = M_1$. When $K = 2$ it corresponds to the usual lifting already exploited to establish stability results for phase recovery, blind-deconvolution, self-calibration, sparse coding, etc. Notice that, when $K \geq 2$, it may be difficult to provide a closed form expression for the operator $\mathcal{A}$. We can however determine simple properties of $\mathcal{A}$. In most reasonable cases, $\mathcal{A}$ is sparse. If the operators M_k simply embed the values of h in a matrix, the matrix representing $\mathcal{A}$ only contains zeros and ones. Also, since the operators M_k are known, we can compute $\mathcal{A}P(\mathbf{h})$, whatever $\mathbf{h} \in \mathbb{R}^{S \times K}$, using (9). Said differently, we can compute $\mathcal{A}$ for any rank 1 tensor. Therefore, since $\mathcal{A}$ is linear, we can compute $\mathcal{A}T$ for any low rank tensor T. If the dimensions of the problem permit, one can manipulate $\mathcal{A}$ in a basis of $\mathbb{R}^{S^K}$.

Since $\mathrm{rk}(\mathcal{A})$ is an important quantity, let us emphasize that we always have $\mathrm{rk}(\mathcal{A}) \leq mn$. It is also possible to compute $\mathrm{rk}(\mathcal{A})$, when mn is not too large, using the following proposition.

Proposition 4. *If we consider R independent random $\mathbf{h}^r$, with $r = 1..R$, according to the normal distribution in $\mathbb{R}^{S \times K}$, we have (with probability* 1*)*

$$\dim(\mathrm{Span}\,((\mathcal{A}P(\mathbf{h}^r))_{r=1..R})) = \begin{cases} R & \text{if } R \leq \mathrm{rk}(\mathcal{A}) \\ \mathrm{rk}(\mathcal{A}) & \text{otherwise.} \end{cases} \tag{10}$$

The proof is provided in Appendix 9.5

Using Corollary 1, when (3) has a minimizer, we rewrite in the form

$$\mathbf{h}^* \in \mathrm{argmin}_{L \in \mathbb{N}, \mathbf{h} \in \mathcal{M}^L} \ \|\mathcal{A}P(\mathbf{h}) - X\|^2. \tag{11}$$

We now decompose this problem into two sub-problems: A least-squares problem

$$T^* \in \mathrm{argmin}_{T \in \mathbb{R}^{S^K}} \|\mathcal{A}T - X\|^2 \tag{12}$$

and a non-convex problem

$$\mathbf{h}'^* \in \mathrm{argmin}_{L \in \mathbb{N}, \mathbf{h} \in \mathcal{M}^L} \|\mathcal{A}(P(\mathbf{h}) - T^*)\|^2. \tag{13}$$

Proposition 5. *For any X, $\mathcal{A}$, when* (3) *has a minimizer:*

1. *Let $\mathbf{h}^*$ be a solution of* (11)*. Then, for any solution T^* of* (12)*, $\mathbf{h}^*$ also minimizes* (13)*.*
2. *Let T^* be a solution of* (12) *and $\mathbf{h}'^*$ a solution of* (13)*. Then, $\mathbf{h}'^*$ also minimizes* (11)*.*

The Proposition is proved in Appendix 9.6.

From now on, because of the equivalence between solutions of (13) and (11), we stop using the notation $\mathbf{h}'^*$ and write $\mathbf{h}^* \in \mathrm{argmin}_{L \in \mathbb{N}, \mathbf{h} \in \mathcal{M}^L} \|\mathcal{A}(P(\mathbf{h}) - T^*)\|^2$.

5 Identifiability (error free case)

Throughout this section, we assume that X is such that there exists $\overline{L}$ and $\overline{\mathbf{h}} \in \mathcal{M}^{\overline{L}}$ such that

$$X = M_1(\overline{\mathbf{h}}_1)\cdots M_K(\overline{\mathbf{h}}_K). \tag{14}$$

Under this assumption, $X = \mathcal{A}P(\overline{\mathbf{h}})$, so

$$P(\overline{\mathbf{h}}) \in \operatorname{argmin}_{T\in\mathbb{R}^{S^K}} \|\mathcal{A}T - X\|^2.$$

Moreover, we trivially have $P(\overline{\mathbf{h}}) \in \Sigma_1$ and therefore $\overline{\mathbf{h}}$ minimizes (13), (3) and (11). As a consequence, (3) has a minimizer.

We ask whether there exist guarantees that the resolution of (3) allows one to recover $\overline{\mathbf{h}}$ (up to the usual uncertainties).

In this regard, for any $\mathbf{h} \in \langle\overline{\mathbf{h}}\rangle$, we have $P(\mathbf{h}) = P(\overline{\mathbf{h}})$ and therefore $\mathcal{A}P(\mathbf{h}) = \mathcal{A}P(\overline{\mathbf{h}}) = X$. Thus unless we make further assumptions on $\overline{\mathbf{h}}$, we cannot expect to distinguish any particular element of $\langle\overline{\mathbf{h}}\rangle$ using only X. In other words, recovering $\langle\overline{\mathbf{h}}\rangle$ is the best we can hope for.

Definition 2. Identifiability

We say that $\langle\overline{\mathbf{h}}\rangle$ is identifiable *if the elements of* $\langle\overline{\mathbf{h}}\rangle$ *are the only solutions of* (3).

We say that $\mathcal{M}$ is identifiable *if for every* $L \in \mathbb{N}$ *and every* $\overline{\mathbf{h}} \in \mathcal{M}^L$, $\langle\overline{\mathbf{h}}\rangle$ *is identifiable.*

Proposition 6. Characterization of the global minimizers

For any $L^* \in \mathbb{N}$ *and any* $\mathbf{h}^* \in \mathcal{M}^{L^*}$, $(L^*, \mathbf{h}^*) \in \operatorname{argmin}_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}} \|\mathcal{A}P(\mathbf{h}) - X\|^2$ *if and only if*

$$P(\mathbf{h}^*) \in P(\overline{\mathbf{h}}) + \operatorname{Ker}(\mathcal{A}).$$

The Proposition is proved in Appendix 9.7.

In order to state the following proposition, we define for any L and $L' \in \mathbb{N}$

$$P(\mathcal{M}^L) - P(\mathcal{M}^{L'}) := \left\{ P(\mathbf{h}) - P(\mathbf{g}) \mid \ \mathbf{h} \in \mathcal{M}^L \text{ and } \mathbf{g} \in \mathcal{M}^{L'} \right\} \subset \mathbb{R}^{S^K}.$$

Proposition 7. Necessary and sufficient conditions of identifiability

1. *For any* $\overline{L}$ *and* $\overline{\mathbf{h}} \in \mathcal{M}^{\overline{L}}$: $\langle\overline{\mathbf{h}}\rangle$ *is identifiable if and only if for any* $L \in \mathbb{N}$

$$\left(P(\overline{\mathbf{h}}) + \operatorname{Ker}(\mathcal{A})\right) \cap P(\mathcal{M}^L) \ \subset \ \{P(\overline{\mathbf{h}})\}.$$

2. $\mathcal{M}$ *is identifiable if and only if for any* L *and* $L' \in \mathbb{N}$

$$\operatorname{Ker}(\mathcal{A}) \cap \left(P(\mathcal{M}^L) - P(\mathcal{M}^{L'})\right) \subset \{0\}. \tag{15}$$

The Proposition is proved in Appendix 9.8.

In the context of the usual compressed sensing (i.e., when $K = 1$, $\mathcal{M}$ contains L-sparse signals, $\mathcal{A}$ is a rectangular matrix with full row rank and X is a vector), the proposition is already stated in Lemma 3.1 of [20].

In reasonably small cases and when $P(\mathcal{M})$ is algebraic, one can use tools from numerical algebraic geometry such as those described in [32, 33] to check whether the condition (15) holds or not. The drawback of Proposition 7 is that, given a factorization model described by $\mathcal{A}$, the condition (15) might be difficult to verify.

We therefore establish simpler conditions related to the identifiability of $\mathcal{M}$. First we establish a condition such that for almost every $\mathcal{A}$ satisfying it, $\mathcal{M}$ is identifiable. The main benefit of this condition is that its constituents can be computed in many practical situations.

Before that, we recall a few facts of algebraic geometry, for $X, Y \subset \mathbb{R}^N$, the *join* of X and Y (see, e.g., [31, Ex. 8.1]) is

$$J(X,Y) := \overline{\{sx+ty \mid x \in X,\ y \in Y,\ s,t \in \mathbb{R}\}}^{Zar}.$$

If for all $L \in \mathbb{N}$, $\mathcal{M}^L$ is Zariski closed and invariant under rescaling (e.g., if they are all linear spaces), then $P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ is a Zariski open subset of $J(P(\mathcal{M}^L), P(\mathcal{M}^{L'}))$. In general, it is contained in this join.

Recall the following fact (*): for complex algebraic varieties $X, Y \subset \mathbb{C}^N$, any component Z of $X \cap Y$ has $\dim(Z) \geq \dim(X) + \dim(Y) - N$, and equality holds generically (we make "generically" precise in our context below). Moreover, if X, Y are invariant under rescaling, since $0 \in X \cap Y$, we have $X \cap Y \neq \emptyset$. (See, e.g., [58, §I.6.2].)

This intersection result indicates that if there exists L, L' such that

$$\mathrm{rk}(\mathcal{A}) < \dim\left(P(\mathcal{M}^L) - P(\mathcal{M}^{L'})\right)$$

we expect to have non-identifiability; and if the rank is larger, for all pair L, L', we expect identifiability.

It is straightforward to make the identifiability assertion precise:

Theorem 3. Almost surely sufficient condition for Identifiability
For almost every $\mathcal{A}$ such that

$$\mathrm{rk}(\mathcal{A}) \geq \dim\left(J(P(\mathcal{M}^L), P(\mathcal{M}^{L'}))\right), \qquad \textit{for all } L, L',$$

$\mathcal{M}$ is identifiable.

The theorem is proved in Appendix 9.9.

Since $\dim\left(J(P(\mathcal{M}^L), P(\mathcal{M}^{L'}))\right) \leq \dim\left(P(\mathcal{M}^L)\right) + \dim\left(P(\mathcal{M}^{L'})\right)$, if D_{max} is the maximum dimension of $P(\mathcal{M}^L)$ over all L, one has the same conclusion if $\mathrm{rk}(\mathcal{A}) \geq 2D_{max}$.

When $K = 1$, we illustrate this result by interpreting it in the context of compressive sensing, where $\mathbf{h}$ is a vector, X is a vector, $\mathcal{A}$ is a rectangular sampling matrix of full row rank and $\mathrm{Ker}(\mathcal{A})$ is large. The statement analogous to Theorem 3 in the compressive sensing framework takes the form: "For almost every sampling matrix, any L sparse signal $\mathbf{h}$ can be recovered from $\mathcal{A}\mathbf{h}$ as soon as $2L \leq \mathrm{rk}(\mathcal{A})$." Moreover, the constituent of the ℓ^0 minimization model used to recover the signal are also the constituents of (11). Again, the main novelty is to extend this result to the identifiability of the factors of a deep matrix products.

In order to establish a necessary condition for identifiability, first note that if we extend $P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ to be scale invariant, this will not affect whether or not it intersects $\ker(\mathcal{A})$ outside of the origin. We immediately conclude that in the complex setting where $\mathcal{M}^L, \mathcal{M}^{L'}$ are both Zariski closed, that $\mathcal{M}$ is non-identifiable whenever $\mathrm{rk}(\mathcal{A}) < \dim\left(P(\mathcal{M}^L) - P(\mathcal{M}^{L'})\right)$. This indicates that we should always expect non-identifiability whenever $\mathrm{rk}(\mathcal{A}) < \dim\left(P(\mathcal{M}^L) - P(\mathcal{M}^{L'})\right)$ but is not adequate to prove it because real algebraic varieties need not satisfy (*). However it is true for real linear spaces, so we immediately conclude the following weak result:

Theorem 4. Necessary condition for Identifiability

Let $C(P(\mathcal{M}^L) - P(\mathcal{M}^{L'}))$ be the set of all points on all lines through the origin intersecting $P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$, and let q be the maximal dimension of a linear space on $C(P(\mathcal{M}^L) - P(\mathcal{M}^{L'}))$. Then if $q > \mathrm{rk}(\mathcal{A})$, $\mathcal{M}$ is not identifiable. In particular when the $\mathcal{M}^L$'s contain linear space and if we let S' the be the largest dimension of these vector space, if $2S' > \mathrm{rk}(\mathcal{A})$, then $\mathcal{M}$ is not identifiable.

6 Stable recovery in the noisy case

In this section, we consider errors of different natures. More precisely, we assume that we know a collection $\mathcal{M} = (\mathcal{M}^L)_{L\in\mathbb{N}}$ of models $\mathcal{M}^L \subset \mathbb{R}^{S\times K}$, for $L \in \mathbb{N}$. We also assume that there exists $\overline{L}$ and $L^* \in \mathbb{N}$, $\overline{\mathbf{h}} \in \mathcal{M}^{\overline{L}}$ and $\mathbf{h}^* \in \mathcal{M}^{L^*}$, such that

$$\|M_1(\overline{\mathbf{h}}_1)\cdots M_K(\overline{\mathbf{h}}_K) - X\| \leq \delta, \tag{16}$$

and

$$\|M_1(\mathbf{h}_1^*)\cdots M_K(\mathbf{h}_K^*) - X\| \leq \eta, \tag{17}$$

for δ and η typically small.

Again, in signal processing applications, this correspond to existing unknown parameters $\overline{\mathbf{h}}$ that we estimate from a noisy observation X, using an inaccurate solution $\mathbf{h}^*$ of (3) (as in [9] where the case $K = 1$ is studied). In machine learning application, $\overline{\mathbf{h}}$ and $\mathbf{h}^*$ shall be interpreted as different learned parameters; δ and η are the corresponding risks.

Also, notice that the above hypothesis does not even require (3) to have a solution. Also, algorithms which do not come with a guarantee sometimes manage to reach small δ and η values. In those cases, the analysis we conduct in this section permits to guarantee the stable recovery of $\overline{\mathbf{h}}$ despite the lack of a guarantee of the algorithm. Finally, the hypotheses (16)and (17) permit to obtain guarantees for algorithms that, instead of minimizing (3), minimize an objective function which approximates the one in (3). This is particularly relevant for machine learning applications when (3) can be an empirical risk that need to be regularized or is not truely minimized (for instance, when using *dropout* [60]).

A necessary and sufficient condition for the identifiability of $\mathcal{M}$ is stated in Proposition 7. The condition is on the way $\mathrm{Ker}\,(\mathcal{A})$ and $P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ intersect. In order to guarantee the stable recovery of the elements of $\mathcal{M}$, we need a stronger condition on the geometry of this intersection to hold for every L and $L' \in \mathbb{N}$. This condition is provided in the next definition.

Definition 3. Deep-Null Space Property

Let $\gamma > 0$ and $\rho > 0$, we say that $\mathrm{Ker}\,(\mathcal{A})$ *satisfies the* deep-Null Space Property (deep-NSP) *with respect to the collection of models $\mathcal{M}$ with constants (γ,ρ) if for any L and $L' \in \mathbb{N}$, any $T \in P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ satisfying $\|\mathcal{A}T\| \leq \rho$ and any $T' \in \mathrm{Ker}\,(\mathcal{A})$, we have*

$$\|T\| \leq \gamma\|T - T'\|. \tag{18}$$

The deep-NSP implies that, for $T \in P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ close to $\mathrm{Ker}\,(\mathcal{A})$ in the sense that $\|\mathcal{A}T\| \leq \rho$ we must have, by decomposing $T = T' + T''$, with $T' \in \mathrm{Ker}\,(\mathcal{A})$ and T'' in its orthogonal complement

$$\|T\| \leq \gamma\|T - T'\| = \gamma\|T''\| \leq \frac{\gamma}{\sigma_{min}}\|\mathcal{A}T''\| \leq \frac{\gamma}{\sigma_{min}}\rho,$$

where σ_{min} is the smallest non-zero singular value of $\mathcal{A}$. In words, $\|T\|$ must be small. We can conclude that under the deep-NSP, $P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ and $\{T \in \mathbb{R}^{SK} \mid \|\mathcal{A}T\| \leq \rho\}$ only intersect in the vicinity of 0.

Additionally, (18) implies that in the vicinity of 0, $\mathrm{Ker}\,(\mathcal{A})$ and $P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ are not tangential. Their intersection is transverse.

Let us mention that if $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constants (γ, ρ), we have for all $T' \in \mathrm{Ker}\,(\mathcal{A})$ and all $T \in P(\mathcal{M}^L) - P(\mathcal{M}^{L'})$ satisfying $\|\mathcal{A}T\| \leq \rho$

$$\|T'\| \leq \|T\| + \|T' - T\| \leq (\gamma + 1)\|T' - T\|.$$

Therefore,

$$\forall T' \in \mathrm{Ker}\,(\mathcal{A}), \qquad \|T'\| \leq (\gamma + 1) d_{loc}(T', P(\mathcal{M}^L) - P(\mathcal{M}^{L'})) \tag{19}$$

where we have set for any $C \subset \mathbb{R}^{SK}$

$$d_{loc}(T', C) = \inf_{T \in C, \|\mathcal{A}T\| \leq \rho} \|T' - T\|.$$

The converse is also true, if $\mathrm{Ker}\,(\mathcal{A})$ satisfies (19), it satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with appropriate constants. In the context of the usual compressed sensing (i.e., when $K = 1$, $\mathcal{M}^L$ contains L-sparse signals, $\mathcal{A}$ is a rectangular matrix with full row rank and X is a vector), the localization appearing in d_{loc} can be discarded since the inequality must hold when T' is small and since in this case this localization has no effect. Therefore, in the compressed sensing context, (19) (and therefore deep-NSP) is the usual Null Space Property with respect to L-sparse vectors, as defined in [20]. However, deep-NSP is generalized to take into account deep factorization problems. This motivates the name.

In the general case, the deep-NSP can be understood as a local version of the generalized-NSP for $\mathcal{A}$ relative to $P(\cup_{L \in \mathbb{N}} \mathcal{M}^L) - P(\cup_{L \in \mathbb{N}} \mathcal{M}^L)$, as defined in [9]. Our interest for the locality (as imposed by the constraint $\|\mathcal{A}T\| \leq \rho$) is motivated by the fact that we want to use the deep-NSP when the signal to noise ratio is controlled (i.e., the hypotheses of Theorem 1 are satisfied). Our stable recovery property therefore includes such hypotheses. Such locality hypotheses are needed to obtain Theorem 6.

Also, we have not adapted the robust-NSP defined in [9]. The benefit not to use this definition is to obtain a simpler definition for deep-NSP. In particular (18) does not involve the geometry of $\mathcal{A}$ in the orthogonal complement of $\mathrm{Ker}\,(\mathcal{A})$. Looking in detail at the benefit of this adaptation is of course, of a great interest.

Finally, notice that we trivially have the following two facts:

- If $\mathrm{Ker}\,(\mathcal{A}) = \{0\}$, then $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with respect to the model $\mathbb{R}^{S \times K}$ with constant $(1, +\infty)$.
- For any $\gamma' \geq \gamma$: If $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constants (γ, ρ), then $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constant γ'.
- For any $\mathcal{M}' \subset \mathcal{M}$: If $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$ with constant (γ, ρ), then $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}'$ with constant (γ, ρ). In particular, if $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with respect to the model $\mathbb{R}^{S \times K}$ with constant (γ, ρ), it satisfies the deep-NSP with respect to any collection of models, with constant (γ, ρ).

Theorem 5. Sufficient condition for stable recovery

Assume $\mathrm{Ker}(\mathcal{A})$ *satisfies the deep-NSP with respect to the collection of models* $\mathcal{M}$ *and with the constant* (γ,ρ)*. For any* $\mathbf{h}^*$ *as in* (17) *with* η *and* δ *(see* (17) *and* (16)*) such that* $\delta+\eta\leq\rho$*, we have*

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

where σ_{min} *is the smallest non-zero singular value of* $\mathcal{A}$*. Moreover, if* $\overline{\mathbf{h}}\in\mathbb{R}_*^{S\times K}$ *and* $\frac{\gamma}{\sigma_{min}}\,(\delta+\eta)\leq\frac{1}{2}\,\max\left(\|P(\overline{\mathbf{h}})\|_\infty,\|P(\mathbf{h}^*)\|_\infty\right)$ *then*

$$d_p(\langle\mathbf{h}^*\rangle,\langle\overline{\mathbf{h}}\rangle)\leq\frac{7(KS)^{\frac{1}{p}}\gamma}{\sigma_{min}}\min\left(\|P(\overline{\mathbf{h}})\|_\infty^{\frac{1}{K}-1},\|P(\mathbf{h}^*)\|_\infty^{\frac{1}{K}-1}\right)(\delta+\eta).\tag{20}$$

The first part of the proof is very similar to usual proofs in the Compressed Sensing and stable recovery literature. The second part simply uses Theorem 1. The detailed proof of the theorem is provided in Appendix 9.10.

When $K=2$, the first upper bound obtained in Theorem 5 is similar to the bound in Theorem 2 of [2] and Theorem 3.3 in [42]. Notice that the existing bounds in [2, 42] are degraded because the estimator is numerically realistic. The second bound is to be compared with the bound in Corollary 1 of [2] and Corollary 3.4 of [42].

This theorem provides a sufficient condition to get stable recovery. The only significant hypothesis made on the factorization problem is that $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect to the collection of models $\mathcal{M}$. One might ask whether this hypothesis is sharp or not. As expressed in the next theorem, the answer to this question is positive.

Theorem 6. Necessary condition for stable recovery

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

$$\|\mathcal{A}P(\mathbf{h}^*)-X\|^2\leq\|e\|$$

we have

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

Then, $\mathrm{Ker}(\mathcal{A})$ *satisfies the deep-NSP with respect to the collection of models* $\mathcal{M}$ *with constants*

$$(\gamma,\rho)=(CS^{\frac{K-1}{2}}\sqrt{K}\,\sigma_{max},\delta)$$

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

The first part of the proof is inspired by and close to the proof of the analogous converse statement in [20]. The second part simply uses Theorem 2. The detailed proof of the theorem is provided in Appendix 9.11.

The sharpness of the known results when $K=2$ is usually argued by comparing the number of samples necessary for the recovery and the information theoretic limit of the problem. As far as the authors know, the above theorem is therefore new even when $K=2$.

7 Application to convolutional linear network

We consider a 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

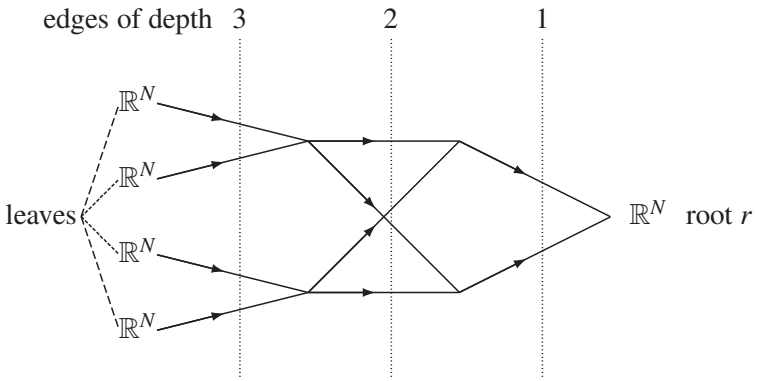

Figure 1: Example of the considered convolutional linear network. To every edge is attached a convolution kernel. The network does not involve non-linearities or sampling.

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 $\mathcal{P}$ 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. It is called the depth of e. We also say that e belongs to the layer k. For any $k \in [K]$, we denote the set containing all the edges of depth k, by $\mathcal{E}(k)$.

Moreover, to any edge e is attached a convolution kernel of support $\mathcal{S}_e \subset [N]$. 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.

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 [50] or the fast transforms optimized in [15, 16]. 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

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

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

This section aims at identifying conditions such that any unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}^{S\times K}$ can be identified or stably recovered from $X = M_1(\overline{\mathbf{h}}_1)\cdots M_K(\overline{\mathbf{h}}_K)$ (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 sig-

nals $x^f \in \mathbb{R}^N$ 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 $n \in [N]$. The column of X corresponding to the entry n in the leaf f is the translation by n of

$$\sum_{p\in\mathcal{P}(f)} \mathcal{T}^p(\mathbf{h}) \tag{21}$$

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

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

is the composition of convolutions along the path p.

Moreover, we define for any $k \in [K]$ the mapping $\mathbf{e}_k : [S] \longrightarrow \mathcal{E}(k)$ which provides for any $i \in [S]$ 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 $\mathbf{i} \in [S]^K$, we denote $\mathbf{p_i} = (\mathbf{e}_1(\mathbf{i}_1), \cdots, \mathbf{e}_K(\mathbf{i}_K))$ and

$$\mathbf{I} = \left\{\mathbf{i} \in [S]^K | \mathbf{p_i} \in \mathcal{P}\right\}.$$

The latter contains all the indices corresponding to a valid path in the network. For any set of parameters $\mathbf{h} \in \mathbb{R}^{S\times K}$ and any path $\mathbf{p} \in \mathcal{P}$, 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 denote $\mathbb{1} \in \mathbb{R}^S$ a vector of size S with all its entries equal to 1. For any edge 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 $\mathbf{p} = (e^1, \cdots, e^K) \in \mathcal{P}$, the support of $M_1(\mathbb{1}^{e_1}) \cdots M_K(\mathbb{1}^{e_K})$ is denoted by $\text{Supp}(\mathbf{p})$.

Finally, we recall that because of Corollary 1, there exists a unique mapping

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

such that

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

where P is the Segre embedding (defined in (6)).

Proposition 8. Necessary condition of identifiability of convolutional linear network

- *Either all the entries of* $M_1(\mathbb{1}) \cdots M_K(\mathbb{1})$ *belong to* $\{0,1\}$ *and then*
  1. *for any distinct* $\mathbf{p}$ *and* $\mathbf{p}' \in \mathcal{P}$*, we have* $\text{Supp}(\mathbf{p}) \cap \text{Supp}(\mathbf{p}') = \emptyset$.
  2. $\text{Ker}(\mathcal{A}) = \{T \in \mathbb{R}^{S^K} | \forall \mathbf{i} \in \mathbf{I}, T_{\mathbf{i}} = 0\}$.
- *or some of the entries of* $M_1(\mathbb{1}) \cdots M_K(\mathbb{1})$ *do not belong to* $\{0,1\}$ *and then* $\mathbb{R}^{S\times K}$ *is not identifiable.*

The proof of the proposition is in Appendix 9.12.

The interest of the condition in Proposition 8 is that it can easily be computed when $N \times N|\mathcal{F}|$ is not too large. Notice that, beside the known examples in blind-deconvolution (i.e. when $K = 2$ and $|\mathcal{P}| = 1$) [2, 6], there are known (truly deep) convolutional linear networks that satisfy the condition of the first statement of Proposition 8. For instance, the

convolutional linear network corresponding to the un-decimated Haar (wavelet)[8] transform is a tree and for any of its leaves $f \in \mathcal{F}$, $|\mathcal{P}(f)| = 1$. Moreover, the support of the kernel living on the edge e, of depth k, on this path is $\{0, 2^k\}$. It is therefore not difficult to check that the first condition of Proposition 8 holds.

We also have the following proposition.

Proposition 9. *If $|\mathcal{P}| = 1$ and all the entries of $M_1(\mathbb{1}) \cdots M_K(\mathbb{1})$ belong to $\{0,1\}$, then $\mathrm{Ker}(\mathcal{A}) = \{0\}$ and $\mathrm{Ker}(\mathcal{A})$ satisfies the deep-NSP with respect to any model collection $\mathcal{M}$ with constant $(\gamma, \rho) = (1, +\infty)$. Moreover, we have $\sigma_{min} = \sqrt{N}$.*

The proof of the proposition is in Appendix 9.13.

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 $\mathcal{G}(\mathcal{E}, \mathcal{N})$ and $(\mathcal{S}_e)_{e \in \mathcal{E}}$ but defined by unknown parameters $\overline{\mathbf{h}} \in \mathbb{R}^{S \times K}$. We consider the noisy situation where

$$\|M_1(\overline{\mathbf{h}}_1) \cdots M_K(\overline{\mathbf{h}}_K) - X\| \leq \delta,$$

and an estimate $\mathbf{h}^* \in \mathbb{R}^{S \times K}$ such that

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

We say that two networks sharing the same structure and defined by $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}^{S \times K}$ are equivalent if and only if

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

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

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

where we recall 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 Definition 1. Since d_p is a metric, we easily prove that δ_p is a metric between network classes.

Theorem 7. Sufficient condition of stable recovery of convolutional linear network

If all the entries of $M_1(\mathbb{1}) \cdots M_K(\mathbb{1})$ belong to $\{0,1\}$, if there exists $\varepsilon > 0$ such that for all $e \in \mathcal{E}$, $\|\mathcal{T}_e(\overline{\mathbf{h}})\|_\infty \geq \varepsilon$, and if $\delta + \eta \leq \frac{\sqrt{N}\varepsilon^K}{2}$ then

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

where

$$S' = \max_{e \in \mathcal{E}} |\mathcal{S}_e|$$

The proof of the Theorem is in Appendix 9.14.

[8] Un-decimated means computed with the "Algorithme à trous", [50], Section 5.5.2 and 6.3.2. The Haar wavelet is described in [50], Section 7.2.2, p. 247 and Example 7.7, p. 235

8 Conclusion and perspectives

In this paper, we have established necessary and sufficient conditions for the identifiability and stable recovery of deep linear networks. They rely on the lifting of the problem in a tensor space. The technique is called *tensorial lifting*. The main results are proved using compressed sensing technics and properties of the Segre embedding (the embedding that maps the parameters in the tensor space). The general results are then particularized to establish necessary and sufficient conditions for the stable recovery of a convolutional linear network of any depth $K \geq 1$.

To the best of our knowledge, this is the first time stable recovery statements are obtained for a "deep factorization" problem (i.e. a problem with $K \geq 3$). It paves the way for the study of deep factorization problems (a list is in Section 1.3) and promises to be an essential step towards the theoretical understanding of deep neural networks.

Acknowledgment

Joseph Landsberg is supported by NSF DMS-1405348 and AF1814254 .

9 Appendices

9.1 Proof of Proposition 1

Notice that, the sets $\langle \mathbf{h} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}$ and $\langle \mathbf{g} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}$ are finite and therefore the infimum in the definition of d is reached. We also have whatever $\mathbf{h}, \mathbf{g} \in \mathbb{R}^{S\times K}_{*}$

$$d_p(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle) = \inf_{\mathbf{h}' \in \langle \mathbf{h} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}} \Big(\inf_{\mathbf{g}' \in \langle \mathbf{g} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}} \|\mathbf{h}' - \mathbf{g}'\|_p \Big). \tag{22}$$

Moreover, whatever $\mathbf{h} \in \mathbb{R}^{S\times K}_{*}$ and $\mathbf{h}'$ and $\mathbf{h}'' \in \langle \mathbf{h} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}$ there exist $(s_k)_{k\in[K]} \in \{-1,1\}^K$ such that $\prod_{k\in[K]} s_k = 1$ and

$$\mathbf{h}'_k = s_k \mathbf{h}''_k \qquad , \forall k \in [K].$$

Using the above two properties, we can check that

$$\inf_{\mathbf{g}' \in \langle \mathbf{g} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}} \|\mathbf{h}' - \mathbf{g}'\|_p = \inf_{\mathbf{g}' \in \langle \mathbf{g} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}} \|\mathbf{h}'' - \mathbf{g}'\|_p$$

As a consequence, the outer infimum in (22) is irrelevant and we have

$$d_p(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle) = \inf_{\mathbf{g}' \in \langle \mathbf{g} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}} \|\mathbf{h}' - \mathbf{g}'\|_p \qquad , \forall \mathbf{h}, \mathbf{g} \in \mathbb{R}^{S\times K}_{*} \text{ and } \mathbf{h}' \in \langle \mathbf{h} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}.$$

Using this last property, we easily check that d_p is a metric on $\mathbb{R}^{S\times K}_{*}/\sim$. □

9.2 Proof of Theorem 1

Notice first that when $K = 1$ the inequality is a straightforward consequence of the usual inequalities between l^p norms. We therefore assume from now on that $K \geq 2$.

All along the proof, we consider $\mathbf{h}$ and $\mathbf{g} \in \mathbb{R}_*^{S\times K}$ and assume that $\|P(\mathbf{h})\|_\infty \geq \|P(\mathbf{g})\|_\infty$. We also assume that $\|P(\mathbf{g}) - P(\mathbf{h})\|_\infty \leq \frac{1}{2}\|P(\mathbf{h})\|_\infty$. We first prove the inequality when $p = q = +\infty$.

In order to do so, we consider

$$\mathbf{i} \in \text{argmax}_{\mathbf{j}\in[S]^K} \ |P(\mathbf{h})_\mathbf{j}|$$

and assume, without lost of generality (otherwise, we can multiply one vector of $\mathbf{h}$ and $\mathbf{g}$ by -1 to get this property and multiply back once the inequality have been established), that $P(\mathbf{h})_\mathbf{i} \geq 0$. We therefore have $P(\mathbf{h})_\mathbf{i} = \|P(\mathbf{h})\|_\infty$. Notice also that we have, under the above hypotheses,

$$\|P(\mathbf{g})\|_\infty \geq P(\mathbf{g})_\mathbf{i} \geq P(\mathbf{h})_\mathbf{i} - \|P(\mathbf{g}) - P(\mathbf{h})\|_\infty \geq \frac{1}{2}\|P(\mathbf{h})\|_\infty > 0. \tag{23}$$

Moreover, we consider the operator $E_\mathbf{i}$ that extracts the K signals of size S that are obtained when freezing, at the index $\mathbf{i}$ in a tensor T, all coordinates but one. Formally, we denote

$$\begin{array}{rcl} E_\mathbf{i} : \mathbb{R}^{S^K} & \longrightarrow & \mathbb{R}^{S\times K} \\ T & \longmapsto & E_\mathbf{i}(T) \end{array}$$

where for all $k \in [K]$ and all $j \in [S]$

$$E_\mathbf{i}(T)_{k,j} = T_{\mathbf{i}_1,\ldots,\mathbf{i}_{k-1},j,\mathbf{i}_{k+1},\ldots,\mathbf{i}_K}.$$

We consider

$$\mathbf{h}' = (P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}} \ E_\mathbf{i}(P(\mathbf{h})) \qquad \text{and} \qquad \mathbf{g}' = (P(\mathbf{g})_\mathbf{i})^{-1+\frac{1}{K}} \ E_\mathbf{i}(P(\mathbf{g})).$$

We have for all $\mathbf{j} \in [S]^K$

$$\begin{array}{rcl} P(\mathbf{h}')_\mathbf{j} & = & (P(\mathbf{h})_\mathbf{i})^{-K+1} \ P\left(E_\mathbf{i}(P(\mathbf{h}))\right)_\mathbf{j}, \\ & = & (P(\mathbf{h})_\mathbf{i})^{-K+1} \ \displaystyle\prod_{k=1}^{K} P(\mathbf{h})_{\mathbf{i}_1,\ldots,\mathbf{i}_{k-1},\mathbf{j}_k,\mathbf{i}_{k+1},\ldots,\mathbf{i}_K} \\ & = & (P(\mathbf{h})_\mathbf{i})^{-K+1} \ \displaystyle\prod_{k=1}^{K} \mathbf{h}_{1,\mathbf{i}_1} \ldots \mathbf{h}_{k-1,\mathbf{i}_{k-1}} \mathbf{h}_{k,\mathbf{j}_k} \mathbf{h}_{k+1,\mathbf{i}_{k+1}} \ldots \mathbf{h}_{K,\mathbf{i}_K} \\ & = & \displaystyle\prod_{k=1}^{K} \mathbf{h}_{k,\mathbf{j}_k} = P(\mathbf{h})_\mathbf{j}. \end{array}$$

We therefore have $P(\mathbf{h}') = P(\mathbf{h})$. This can be written $\mathbf{h}' \in \langle \mathbf{h} \rangle$. Similarly, we have $\mathbf{g}' \in \langle \mathbf{g} \rangle$.

Also, because of the definition of $\mathbf{i}$ and $\mathbf{h}'$, we are guaranteed that, whatever $k \in [K]$,

$$\begin{array}{rcl} \|\mathbf{h}'_k\|_\infty & = & (P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}} \|E_\mathbf{i}(P(\mathbf{h}))\|_\infty \\ & = & \|P(\mathbf{h})\|_\infty^{-1+\frac{1}{K}} \|P(\mathbf{h})\|_\infty = \|P(\mathbf{h})\|_\infty^{\frac{1}{K}} \end{array}$$

The latter being independent of k, we have $\mathbf{h}' \in \mathbb{R}^{S\times K}_{\text{diag}}$. Unfortunately, unless for instance $\mathbf{i} \in \text{argmax}_{\mathbf{j}\in[S]^K} |P(\mathbf{g})_\mathbf{j}|$, it might occur that $\mathbf{g}' \notin \mathbb{R}^{S\times K}_{\text{diag}}$. However, if we consider

$$\mathbf{g}'' \in \text{argmin}_{\mathbf{f}\in\langle\mathbf{g}\rangle\cap\mathbb{R}^{S\times K}_{\text{diag}}} \|\mathbf{f}-\mathbf{g}'\|_\infty,$$

we have since $\mathbf{h}' \in \langle\mathbf{h}\rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}$ and $\mathbf{g}'' \in \langle\mathbf{g}\rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}$

$$\begin{aligned} d_\infty(\langle\mathbf{h}\rangle,\langle\mathbf{g}\rangle) &\leq \|\mathbf{h}'-\mathbf{g}''\|_\infty \\ &\leq \|\mathbf{h}'-\mathbf{g}'\|_\infty + \|\mathbf{g}'-\mathbf{g}''\|_\infty. \end{aligned} \tag{24}$$

In the sequel we will successively calculate upper bounds of $\|\mathbf{h}'-\mathbf{g}'\|_\infty$ and $\|\mathbf{g}'-\mathbf{g}''\|_\infty$ in order to find an upper bound of $d_\infty(\langle\mathbf{h}\rangle,\langle\mathbf{g}\rangle)$.

Upper bound of $\|\mathbf{h}'-\mathbf{g}'\|_\infty$:
We have

$$\begin{aligned} \|\mathbf{h}'-\mathbf{g}'\|_\infty &= \|(P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}}\, E_\mathbf{i}(P(\mathbf{h})) - (P(\mathbf{g})_\mathbf{i})^{-1+\frac{1}{K}}\, E_\mathbf{i}(P(\mathbf{g}))\|_\infty \\ &\leq \|(P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}}\,(E_\mathbf{i}(P(\mathbf{h})) - E_\mathbf{i}(P(\mathbf{g})))\,\|_\infty \\ &\qquad +\|\left((P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}} - (P(\mathbf{g})_\mathbf{i})^{-1+\frac{1}{K}}\right) E_\mathbf{i}(P(\mathbf{g}))\|_\infty \\ &\leq \|P(\mathbf{h})\|_\infty^{-1+\frac{1}{K}}\|E_\mathbf{i}(P(\mathbf{h})) - E_\mathbf{i}(P(\mathbf{g}))\|_\infty + \|P(\mathbf{g})\|_\infty|(P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}} - (P(\mathbf{g})_\mathbf{i})^{-1+\frac{1}{K}}| \\ &\leq \|P(\mathbf{h})\|_\infty^{-1+\frac{1}{K}}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty + \|P(\mathbf{h})\|_\infty|(P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}} - (P(\mathbf{g})_\mathbf{i})^{-1+\frac{1}{K}}| \end{aligned}$$

But we also have using the mean value theorem and (23)

$$\begin{aligned} |(P(\mathbf{h})_\mathbf{i})^{-1+\frac{1}{K}} - (P(\mathbf{g})_\mathbf{i})^{-1+\frac{1}{K}}| &\leq \left(1-\frac{1}{K}\right) P(\mathbf{g})_\mathbf{i}^{-2+\frac{1}{K}}|P(\mathbf{h})_\mathbf{i} - P(\mathbf{g})_\mathbf{i}| \\ &\leq \left(1-\frac{1}{K}\right)\left(\frac{1}{2}\|P(\mathbf{h})\|_\infty\right)^{-2+\frac{1}{K}}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty \\ &\leq 4\,\|P(\mathbf{h})\|_\infty^{-2+\frac{1}{K}}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty \end{aligned}$$

We therefore finally obtain that

$$\|\mathbf{h}'-\mathbf{g}'\|_\infty \leq 5\|P(\mathbf{h})\|_\infty^{-1+\frac{1}{K}}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty. \tag{25}$$

Upper bound of $\|\mathbf{g}'-\mathbf{g}''\|_\infty$:
First, since $\mathbf{g}'' \in \langle\mathbf{g}\rangle = \langle\mathbf{g}'\rangle$, we know that there exists $(\lambda_k)_{k\in[K]} \in \mathbb{R}^K$ such that

$$\prod_{k=1}^K \lambda_k = 1 \tag{26}$$

and

$$\mathbf{g}''_k = \lambda_k \mathbf{g}'_k \qquad \text{, for all } k \in [K].$$

Furthermore, we have for all $k \in [K]$

$$\|\mathbf{g}'_k - \mathbf{g}''_k\|_\infty = |1-\lambda_k|\;\|\mathbf{g}'_k\|_\infty. \tag{27}$$

Also, if there is k' such that $\lambda_{k'} < 0$, since (26) holds, there necessarily exist another k'' such that $\lambda_{k''} < 0$. If we replace $\mathbf{g}''_{k'}$ by $-\mathbf{g}''_{k'}$ and replace $\mathbf{g}''_{k''}$ by $-\mathbf{g}''_{k''}$ we remain in $\langle \mathbf{g} \rangle \cap \mathbb{R}^{S\times K}_{\text{diag}}$ and can only make $\|\mathbf{g}' - \mathbf{g}''\|_\infty$ decrease. Repeating this process until all the λ_k's are non-negative, we can assume without loss of generality that

$$\lambda_k \geq 0 \qquad \text{, whatever } k \in [K].$$

This being said, we establish two other simple facts that motivate the structure of the proof. First, in order to find an upper bound for (27), we easily establish (using (23)) that

$$\begin{aligned} \|\mathbf{g}'_k\|_\infty &= (P(\mathbf{g})_{\mathbf{i}})^{-1+\frac{1}{K}} \, \|E_{\mathbf{i}}(P(\mathbf{g}))\|_\infty \\ &\leq (\frac{1}{2}\|P(\mathbf{h})\|_\infty)^{-1+\frac{1}{K}} \, \|P(\mathbf{h})\|_\infty \\ &\leq 2\|P(\mathbf{h})\|_\infty^{\frac{1}{K}}. \end{aligned} \tag{28}$$

Second, the value λ_k appearing in (27), can be bounded by using bounds on $\|\mathbf{g}'_k\|_\infty$ and the identity

$$\|\mathbf{g}''_k\|_\infty = \|P(\mathbf{g})\|_\infty^{\frac{1}{K}} = \lambda_k \, \|\mathbf{g}'_k\|_\infty. \tag{29}$$

Qualitatively, the latter identity indeed guarantees that, as $\|P(\mathbf{g}) - P(\mathbf{h})\|_\infty$ goes to 0, λ_k goes to 1. Let us now establish this quantitatively.

Recalling that

$$\mathbf{g}' = (P(\mathbf{g})_{\mathbf{i}})^{-1+\frac{1}{K}} \, E_{\mathbf{i}}(P(\mathbf{g})),$$

and using (23) again, we obtain

$$\|\mathbf{g}'_k\|_\infty \leq \left(\|P(\mathbf{h})\|_\infty - \frac{1}{2}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty \right)^{-1+\frac{1}{K}} \, \|P(\mathbf{g})\|_\infty.$$

We also have (again, using (23))

$$\begin{aligned} \|\mathbf{g}'_k\|_\infty &\geq (P(\mathbf{g})_{\mathbf{i}})^{-1+\frac{1}{K}} |P(\mathbf{g})_{\mathbf{i}}| \\ &= (P(\mathbf{g})_{\mathbf{i}})^{\frac{1}{K}} \\ &\geq \left(\|P(\mathbf{h})\|_\infty - \frac{1}{2}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty \right)^{\frac{1}{K}}. \end{aligned}$$

Plugging the upper bound of $\|\mathbf{g}'_k\|_\infty$ in (29), using successively (23), the mean value

theorem and the hypothesis on the size of $P(\mathbf{h}) - P(\mathbf{g})$ gives:

$$\begin{aligned}
\lambda_k - 1 &= \frac{\|P(\mathbf{g})\|_\infty^{\frac{1}{K}}}{\|\mathbf{g}'_k\|_\infty} - 1 \\
&\geq \|P(\mathbf{g})\|_\infty^{-1+\frac{1}{K}} \left(\|P(\mathbf{h})\|_\infty - \frac{1}{2}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty \right)^{1-\frac{1}{K}} - 1 \\
&\geq \left(1 - \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \right)^{1-\frac{1}{K}} - 1 \\
&\geq -(1-\frac{1}{K}) \left(1 - \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \right)^{-\frac{1}{K}} \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \\
&\geq -\left(1 - \frac{1}{4} \right)^{-\frac{1}{K}} \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \\
&\geq -\frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{\|P(\mathbf{h})\|_\infty}.
\end{aligned}$$

Similarly, plugging the lower bound of $\|\mathbf{g}'_k\|_\infty$ in (29), we obtain using successively (23), the mean value theorem and the hypothesis on the size of $P(\mathbf{h}) - P(\mathbf{g})$:

$$\begin{aligned}
\lambda_k - 1 &\leq \|P(\mathbf{g})\|_\infty^{\frac{1}{K}} \left(\|P(\mathbf{h})\|_\infty - \frac{1}{2}\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty \right)^{-\frac{1}{K}} - 1 \\
&\leq \left(1 - \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \right)^{-\frac{1}{K}} - 1 \\
&\leq \frac{1}{K} \left(1 - \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \right)^{-1-\frac{1}{K}} \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \\
&\leq \frac{1}{K} \left(1 - \frac{1}{4} \right)^{-1-\frac{1}{K}} \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{2\|P(\mathbf{h})\|_\infty} \\
&\leq \frac{4^2}{2K3^2} \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{\|P(\mathbf{h})\|_\infty} \\
&\leq \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{\|P(\mathbf{h})\|_\infty}.
\end{aligned}$$

Finally, we get

$$|\lambda_k - 1| \leq \frac{\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty}{\|P(\mathbf{h})\|_\infty}. \tag{30}$$

By combining (27), (28) and (30), we obtain

$$\|\mathbf{g}'_k - \mathbf{g}''_k\|_\infty \leq 2 \, \|P(\mathbf{h})\|_\infty^{-1+\frac{1}{K}} \, \|P(\mathbf{h}) - P(\mathbf{g})\|_\infty.$$

Combining the latter inequality with (24) and (25) provides

$$d_\infty(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle) \leq 7\|P(\mathbf{h})\|_\infty^{-1+\frac{1}{K}} \|P(\mathbf{h}) - P(\mathbf{g})\|_\infty,$$

and concludes the proof when $p = q = +\infty$.

In order to establish the property when $1 \leq p \leq +\infty$ and $1 \leq q \leq +\infty$, we simply use the fact that

$$d_p(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle) \leq (KS)^{\frac{1}{p}} \, d_\infty(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle)$$

and

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_\infty \leq \|P(\mathbf{h}) - P(\mathbf{g})\|_q.$$

□

9.3 Proof of Proposition 2

In the example, we consider $\mathbf{h}$ and $\mathbf{g}$ such that for all $k \in [K]$ and all $i \in [S]$

$$\mathbf{h}_{k,i} = \begin{cases} 1 & \text{if } i = 0, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \mathbf{g}_{k,i} = \begin{cases} \left(\frac{1}{2}\right)^{\frac{1}{K}} & \text{if } i = 0, \\ \varepsilon_q & \text{otherwise,} \end{cases}$$

where $\varepsilon_{+\infty} = \left(\frac{1}{2}\right)^{\frac{1}{K}}$ and $\varepsilon_q = \min\left(\left(\frac{1-\left(\frac{1}{2}\right)^{\frac{q}{K}}}{S-1}\right)^{\frac{1}{q}}, \left(\frac{1}{2}\right)^{\frac{1}{K}}\right)$, if $q < +\infty$. We immediately obtain

$$\|P(\mathbf{h})\|_\infty = 1, \qquad \|P(\mathbf{g})\|_\infty = \frac{1}{2} \qquad \text{and} \qquad \|P(\mathbf{h}) - P(\mathbf{g})\|_\infty = \frac{1}{2}.$$

We also have,

$$d_p(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle)^p = \|\mathbf{h} - \mathbf{g}\|_p^p \geq K(S-1)\, \varepsilon_q^p \geq \frac{KS}{2}\, \varepsilon_q^p.$$

Decomposing the sum necessary to the calculation of the l^q norm of a tensor according to number of index different from 0 (which corresponds to l in the sum below), we obtain

$$\begin{aligned} \|P(\mathbf{h}) - P(\mathbf{g})\|_q^q &= \sum_{l=0}^{K} \binom{l}{K} (S-1)^l \varepsilon_q^{lq} \left(\frac{1}{2}\right)^{\frac{(K-l)q}{K}}, \\ &= \left(\left(\frac{1}{2}\right)^{\frac{q}{K}} + (S-1)\varepsilon_q^q\right)^K \leq 1. \end{aligned}$$

We then easily obtain that

$$\begin{aligned} 7\|P(\mathbf{h})\|_\infty^{-1+\frac{1}{K}} (KS)^{\frac{1}{p}} \|P(\mathbf{h}) - P(\mathbf{g})\|_q &\leq 7(KS)^{\frac{1}{p}}, \\ &\leq 7\, \frac{d_p(\langle \mathbf{h} \rangle, \langle \mathbf{g} \rangle)}{\varepsilon_q}\, 2^{\frac{1}{p}}. \end{aligned} \tag{31}$$

(32)

We first calculate a lower bound of ε_q when $\varepsilon_q = \left(\frac{1-\left(\frac{1}{2}\right)^{\frac{q}{K}}}{S-1}\right)^{\frac{1}{q}}$ (which, in particular, rules out $q = +\infty$). Using the mean value theorem, we obtain

$$1 - \left(\frac{1}{2}\right)^{\frac{q}{K}} \geq \min_{t \in [\frac{1}{2}, 1]} \left(\frac{q}{K} t^{\frac{q}{K}-1}\right)\left(1 - \frac{1}{2}\right).$$

Distinguishing, whether $q \leq K$ or not, we find after a short calculation that, since $q \geq 1$,

$$1 - \left(\frac{1}{2}\right)^{\frac{q}{K}} \geq \min\left(\frac{1}{2K}, \frac{1}{K}\left(\frac{1}{2}\right)^{\frac{q}{K}}\right) = \frac{1}{K}\min\left(\left(\frac{1}{2}\right)^{\frac{1}{q}}, \left(\frac{1}{2}\right)^{\frac{1}{K}}\right)^q \geq \frac{1}{K2^q}.$$

We then deduce

$$\varepsilon_q \geq \frac{1}{2\,(KS)^{\frac{1}{q}}}.$$

Of course, when $\varepsilon_q = \left(\frac{1}{2}\right)^{\frac{1}{K}}$ (which includes $q = +\infty$), we immediately obtain

$$\varepsilon_q \geq \frac{1}{2}.$$

Using this lower bound in (31) leads to the bounds stated in the proposition. □

9.4 Proof of Theorem 2

Before starting the proof, we define for any $k \in \{0, \ldots, K\}$

$$P_k(\mathbf{h}, \mathbf{g})_{\mathbf{i}} = \mathbf{g}_{1,\mathbf{i}_1} \ldots \mathbf{g}_{k,\mathbf{i}_k} \mathbf{h}_{k+1,\mathbf{i}_{k+1}} \ldots \mathbf{h}_{K,\mathbf{i}_K} \qquad \text{, for all } \mathbf{h}, \mathbf{g} \in \mathbb{R}^{S \times K} \text{ and all } \mathbf{i} \in [S]^K.$$

We consider $\mathbf{g}$ and $\mathbf{h} \in \mathbb{R}^{S \times K}$. Let us first assume that $\|\mathbf{g}\|_\infty \leq \|\mathbf{h}\|_\infty = 1$. We have for any $\mathbf{i} \in [S]^K$, using this hypothesis and standard inequalities between l^p norms, when $q < +\infty$

$$\begin{aligned}
|P(\mathbf{g})_{\mathbf{i}} - P(\mathbf{h})_{\mathbf{i}}|^q &= \left|\sum_{k=0}^{K-1} (P_{k+1}(\mathbf{h}, \mathbf{g})_{\mathbf{i}} - P_k(\mathbf{h}, \mathbf{g})_{\mathbf{i}})\right|^q \\
&\leq K^{q-1} \sum_{k=0}^{K-1} |P_{k+1}(\mathbf{h}, \mathbf{g})_{\mathbf{i}} - P_k(\mathbf{h}, \mathbf{g})_{\mathbf{i}}|^q \\
&\leq K^{q-1} \sum_{k=0}^{K-1} \left|\mathbf{g}_{k+1,\mathbf{i}_{k+1}} - \mathbf{h}_{k+1,\mathbf{i}_{k+1}}\right|^q
\end{aligned}$$

The same calculation when $q = +\infty$ leads to

$$|P(\mathbf{g})_{\mathbf{i}} - P(\mathbf{h})_{\mathbf{i}}| \leq K \max_{k=1..K} \left|\mathbf{g}_{k,\mathbf{i}_k} - \mathbf{h}_{k,\mathbf{i}_k}\right|.$$

Therefore, we have when $q < +\infty$

$$\begin{aligned}
\|P(\mathbf{h}) - P(\mathbf{g})\|_q^q &= \sum_{\mathbf{i} \in [S]^K} |P(\mathbf{h})_{\mathbf{i}} - P(\mathbf{g})_{\mathbf{i}}|^q \\
&\leq K^{q-1} \sum_{k=1}^{K} \sum_{\mathbf{i} \in [S]^K} \left|\mathbf{g}_{k,\mathbf{i}_k} - \mathbf{h}_{k,\mathbf{i}_k}\right|^q \\
&= K^{q-1} \sum_{k=1}^{K} S^{K-1} \|\mathbf{g}_k - \mathbf{h}_k\|_q^q \\
&= K^{q-1} S^{K-1} \|\mathbf{g} - \mathbf{h}\|_q^q
\end{aligned}$$

and therefore

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_q \leq K^{1-\frac{1}{q}} S^{\frac{K-1}{q}} \|\mathbf{g} - \mathbf{h}\|_q.$$

Again, a similar calculus for $q = +\infty$ leads to

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_{+\infty} \leq K \|\mathbf{g} - \mathbf{h}\|_{+\infty}.$$

Remember that the two last inequalities hold for $\mathbf{g}$ and $\mathbf{h} \in \mathbb{R}^{S\times K}$ such that $\|\mathbf{g}\|_\infty \leq \|\mathbf{h}\|_\infty = 1$.

Let us now consider any $\mathbf{g}'$ and $\mathbf{h}' \in \mathbb{R}^{S\times K}$ and any $\mathbf{g} \in \mathbb{R}^{S\times K}_{\text{diag}} \cap \langle \mathbf{g}' \rangle$ and $\mathbf{h} \in \mathbb{R}^{S\times K}_{\text{diag}} \cap \langle \mathbf{h}' \rangle$. We denote $\delta = \max(\|\mathbf{g}\|_{+\infty}, \|\mathbf{h}\|_{+\infty})$. Notice first that $\|\mathbf{g}\|_{+\infty} = \|P(\mathbf{g}')\|_{+\infty}^{\frac{1}{K}}$ and $\|\mathbf{h}\|_{+\infty} = \|P(\mathbf{h}')\|_{+\infty}^{\frac{1}{K}}$. Therefore

$$\delta = \max(\|P(\mathbf{g}')\|_{+\infty}, \|P(\mathbf{h}')\|_{+\infty})^{\frac{1}{K}}. \tag{33}$$

We can apply the above inequality to $\frac{\mathbf{h}}{\delta}$ and $\frac{\mathbf{g}}{\delta}$ (we might need to switch $\mathbf{h}$ and $\mathbf{g}$ but it does not change the final inequality) and obtain when $q < +\infty$

$$\|P(\frac{\mathbf{h}}{\delta}) - P(\frac{\mathbf{g}}{\delta})\|_q \leq K^{1-\frac{1}{q}} S^{\frac{K-1}{q}} \|\frac{\mathbf{g}}{\delta} - \frac{\mathbf{h}}{\delta}\|_q.$$

This leads to

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_q \leq K^{1-\frac{1}{q}} S^{\frac{K-1}{q}} \delta^{K-1} \|\mathbf{g} - \mathbf{h}\|_q.$$

Similarly, when $q = +\infty$, we obtain

$$\|P(\mathbf{h}) - P(\mathbf{g})\|_{+\infty} \leq K \delta^{K-1} \|\mathbf{g} - \mathbf{h}\|_{+\infty}.$$

The fact that these two last inequalities hold for any $\mathbf{g} \in \mathbb{R}^{S\times K}_{\text{diag}} \cap \langle \mathbf{g}' \rangle$ and any $\mathbf{h} \in \mathbb{R}^{S\times K}_{\text{diag}} \cap \langle \mathbf{h}' \rangle$, together with (33), leads to the statement provided in Theorem 2. □

9.5 Proof of Proposition 4

The span of the Segre variety $P(\mathbb{R}^{S\times K})$ is the full ambient space $\mathbb{R}^{S^K}$, so there exists sets of $R \leq S^K$ points on it that are linearly independent. The set of R-tuples of points on $P(\mathbb{R}^{S\times K})$ that fail to be linearly independent is a proper subvariety of the variety of sets of R-tuples of points on $P(\mathbb{R}^{S\times K})$ because being a linearly independent set of points is an open condition and there exists sets of points that are linearly independent. Therefore $R \leq S^K$ independent and randomly chosen points according to a continuous distribution on $P(\mathbb{R}^{S\times K})$ will be linearly independent.

The intersection $P(\mathbb{R}^{S\times K}) \cap \text{Ker}\,(\mathcal{A})$ is a proper subvariety of $P(\mathbb{R}^{S\times K})$, so with probability one, $R \leq S^K$ independent randomly chosen points according to a continuous distribution will not intersect it and be linearly independent. This is indeed the intersection of two non-empty open conditions. Therefore, all spans of subsets of the points will intersect $\text{Ker}\,(\mathcal{A})$ transversely (in particular, the span of fewer than $\text{rk}\,(\mathcal{A})$ points will not intersect it). Thus there image under $\mathcal{A}$ will have dimension as large as possible. The same argument works if $R > S^K$. □

9.6 Proof of Proposition 5

The proof relies on the fact that for any $T^* \in \operatorname{argmin}_{T\in\mathbb{R}^{sK}} \|\mathcal{A}T - X\|^2$, we have

$$\mathcal{A}^t(\mathcal{A}T^* - X) = 0,$$

where $\mathcal{A}^t : \mathbb{R}^{n\times m} \to \mathbb{R}^{sK}$ is the adjoint linear map. This implies that for any

$$T^* \in \operatorname{argmin}_{T\in\mathbb{R}^{sK}} \|\mathcal{A}T - X\|^2,$$

any $L \in \mathbb{N}$ and any $\mathbf{h} \in \mathcal{M}^L$

$$\begin{aligned}
\|\mathcal{A}P(\mathbf{h}) - X\|^2 &= \|\mathcal{A}(P(\mathbf{h}) - T^*) + (\mathcal{A}T^* - X)\|^2, \\
&= \|\mathcal{A}(P(\mathbf{h}) - T^*)\|^2 + \|\mathcal{A}T^* - X\|^2 + 2\langle \mathcal{A}(P(\mathbf{h}) - T^*), \mathcal{A}T^* - X\rangle, \\
&= \|\mathcal{A}(P(\mathbf{h}) - T^*)\|^2 + \|\mathcal{A}T^* - X\|^2.
\end{aligned}$$

In words, $\|\mathcal{A}P(\mathbf{h}) - X\|^2$ and $\|\mathcal{A}(P(\mathbf{h}) - T^*)\|^2$ only differ by an additive constant. Moreover, since the value of the objective function $\|\mathcal{A}T^* - X\|^2$ is independent of the particular minimizer T^* we are considering, this additive constant is independent of T^*. As a consequence, a minimizer of $\|\mathcal{A}P(\mathbf{h}) - X\|^2$ also minimizes $\|\mathcal{A}(P(\mathbf{h}) - T^*)\|^2$ and vice versa. □

9.7 Proof of Proposition 6

Write $\overline{T} = P(\overline{\mathbf{h}})$ and let L^* and $\mathbf{h}^*$ be a minimizer of (3). Proposition 5 and the fact that $\overline{T}$ minimizes (12) implies that $(L^*, \mathbf{h}^*) \in \operatorname{argmin}_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}^L} \|\mathcal{A}(P(\mathbf{h}) - \overline{T})\|^2$. As a consequence,

$$\|\mathcal{A}(P(\mathbf{h}^*) - \overline{T})\|^2 = 0$$

and

$$P(\mathbf{h}^*) \in \overline{T} + \operatorname{Ker}(\mathcal{A}),$$

proving the first implication.

Conversely, let $L^* \in \mathbb{N}$ and $\mathbf{h}^* \in \mathcal{M}^{L^*}$ be such that $P(\mathbf{h}^*) \in \overline{T} + \operatorname{Ker}(\mathcal{A})$, then

$$\|\mathcal{A}(P(\mathbf{h}^*) - \overline{T})\|^2 = 0 = \min_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}^L} \|\mathcal{A}(P(\mathbf{h}) - \overline{T})\|^2.$$

As a consequence, $(L^*, \mathbf{h}^*) \in \operatorname{argmin}_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}^L} \|\mathcal{A}(P(\mathbf{h}) - \overline{T})\|^2$ and, using Proposition 5, $\mathbf{h}^*$ is a minimizer of (3). □

9.8 Proof of Proposition 7

- Proof of the first statement of Proposition 7:
  We first assume that $\langle\overline{\mathbf{h}}\rangle$ is identifiable. We consider L^* and h^* such that there is L^* such that $P(\mathbf{h}^*) \in (P(\overline{\mathbf{h}}) + \operatorname{Ker}(\mathcal{A})) \cap P(\mathcal{M}^{L^*})$. We know from Proposition 6 that $\mathbf{h}^* \in \operatorname{argmin}_{L\in\mathbb{N},\mathbf{h}\in\mathcal{M}^L} \|\mathcal{A}P(\mathbf{h}) - X\|^2$. Using that $\langle\overline{\mathbf{h}}\rangle$ is identifiable, $\langle\mathbf{h}^*\rangle = \langle\overline{\mathbf{h}}\rangle$ and, from Standard Fact 1 (at the beginning of Section 3), we get $P(\mathbf{h}^*) = P(\overline{\mathbf{h}})$. Finally, we can conclude, that if $\langle\overline{\mathbf{h}}\rangle$ is identifiable we have $(P(\overline{\mathbf{h}}) + \operatorname{Ker}(\mathcal{A})) \cap P(\mathcal{M}) \subset \{P(\overline{\mathbf{h}})\}$.

Let us assume now that for all $L \in \mathbb{N}$, $(P(\overline{\mathbf{h}}) + \mathrm{Ker}\,(\mathcal{A})) \cap P(\mathcal{M}^L) \subset \{P(\overline{\mathbf{h}})\}$ and consider

$$(L^*, \mathbf{h}^*) \in \mathrm{argmin}_{L \in \mathbb{N}, \mathbf{h} \in \mathcal{M}} \|\mathcal{A}P(\mathbf{h}) - X\|^2.$$

Using Proposition 6, we know that $P(\mathbf{h}^*) \in (P(\overline{\mathbf{h}}) + \mathrm{Ker}\,(\mathcal{A})) \cap P(\mathcal{M}^{L^*})$. Using the hypothesis, we have $P(\mathbf{h}^*) = P(\overline{\mathbf{h}})$ and using Standard Fact 1, we finally conclude that $\langle \mathbf{h}^* \rangle = \langle \overline{\mathbf{h}} \rangle$. This completes the proof of the first statement.

- Proof of the second statement of Proposition 7:
  Assume that there is L and $L' \in \mathbb{N}$ such that $\mathrm{Ker}\,(\mathcal{A}) \cap \big(P(\mathcal{M}^L) - P(\mathcal{M}^{L'})\big) \not\subset \{0\}$ then there exist $\mathbf{h} \in \mathcal{M}^L$ and $\overline{\mathbf{h}} \in \mathcal{M}^{L'}$ such that $P(\mathbf{h}) \neq P(\overline{\mathbf{h}})$ and $P(\mathbf{h}) - P(\overline{\mathbf{h}}) \in \mathrm{Ker}\,(\mathcal{A})$. Using the first statement of the proposition, we obtain that $\overline{\mathbf{h}}$ is not identifiable. As a conclusion, $\mathcal{M}$ is not identifiable.
  Conversely, assume that there exists L' and some non-identifiable $\overline{\mathbf{h}} \in \mathcal{M}^{L'}$. Using the first statement of the proposition, we know that there exists $L \in \mathbb{N}$ and $\mathbf{h} \in \mathcal{M}^L$ such that $P(\mathbf{h}) \neq P(\overline{\mathbf{h}})$ and $P(\mathbf{h}) - P(\overline{\mathbf{h}}) \in \mathrm{Ker}\,(\mathcal{A})$. Therefore $\mathrm{Ker}\,(\mathcal{A}) \cap \big(P(\mathcal{M}) - P(\mathcal{M})\big) \not\subset \{0\}$.

□

9.9 Proof of Theorem 3

We first make the "equality holds generically" statement precise in our context. Fix any variety X and assume Y is a linear space, say of dimension y. Let $G(y, \mathbb{C}^N)$ denote the Grassmannian of y-planes through the origin in $\mathbb{C}^N$. The Grassmannian is both a smooth manifold and an algebraic variety. We can interpret "equality holds generically" in this context as saying for a Zariski open subset of $G(y, \mathbb{C}^N)$, equality will hold. In our situation, if we fix $\mathrm{rk}(\mathcal{A})$ and allow $\ker(\mathcal{A})$ to vary as a point in the Grassmannian, with probability one, it will intersect $J(P(\mathcal{M}^L), P(\mathcal{M}^{L'}))$ only in the origin, and this assertion is also true over $\mathbb{R}$ because complex numbers are only needed to assure existence of intersections, not non-existence. □

9.10 Proof of Theorem 5

We have

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

Geometrically, this means that $P(\mathbf{h}^*)$ belongs to a cylinder centered at $P(\overline{\mathbf{h}})$ whose direction is $\mathrm{Ker}\,(\mathcal{A})$ and whose section is defined using the operator $\mathcal{A}$. If we further decompose (the decomposition is unique)

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

where $T' \in \mathrm{Ker}\,(\mathcal{A})$ and T is orthogonal to $\mathrm{Ker}\,(\mathcal{A})$, we have

$$\|\mathcal{A}(P(\mathbf{h}^*) - P(\overline{\mathbf{h}}))\| = \|\mathcal{A}T\| \geq \sigma_{min}\|T\|, \tag{34}$$

where σ_{min} is the smallest non-zero singular value of $\mathcal{A}$. We finally obtain

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

The term on the left-hand side corresponds to the distance between a point in $P(\mathcal{M}^{L^*}) - P(\mathcal{M}^{\overline{L}})$ (namely $P(\mathbf{h}^*) - P(\overline{\mathbf{h}})$) and a point in $\mathrm{Ker}\,(\mathcal{A})$ (namely T').

Since $\mathrm{Ker}\,(\mathcal{A})$ satisfies the deep-NSP with constants (γ,ρ), when $\delta+\eta \leq \rho$, we obtain the first inequality of the theorem

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

When $\overline{\mathbf{h}} \in \mathbb{R}_*^{S\times K}$, for $\frac{\gamma}{\sigma_{min}}\,(\delta+\eta) \leq \frac{1}{2}\,\max\left(\|P(\overline{\mathbf{h}})\|_\infty, \|P(\mathbf{h}^*)\|_\infty\right)$, we can apply Theorem 1 and obtain (20). □

9.11 Proof of Theorem 6

Let $\overline{L}$ and $\overline{L}' \in \mathbb{N}$ and $\overline{\mathbf{h}} \in \mathcal{M}^{\overline{L}}$ and $\overline{\mathbf{h}}' \in \mathcal{M}^{\overline{L}'}$ be such that $\|\mathcal{A}\left(P(\overline{\mathbf{h}}) - P(\overline{\mathbf{h}}')\right)\| \leq \delta$. We also consider throughout the proof $T' \in \mathrm{Ker}\,(\mathcal{A})$. We assume that $\|P(\overline{\mathbf{h}})\|_\infty \leq \|P(\overline{\mathbf{h}}')\|_\infty$. If it is not the case, we simply switch $\overline{\mathbf{h}}$ 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 $X = \mathcal{A}P(\overline{\mathbf{h}}') + e$ and $\|e\| \leq \delta$. Therefore, the hypothesis of the theorem (applied with $\mathbf{h}^* = \overline{\mathbf{h}}$ and $L^* = \overline{L}$) guarantees that

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

Using the fact that $e = \mathcal{A}P(\overline{\mathbf{h}}) - \mathcal{A}P(\overline{\mathbf{h}}')$ and $T' \in \mathrm{Ker}\,(\mathcal{A})$ we obtain

$$\|e\| = \|\mathcal{A}(P(\overline{\mathbf{h}}) - P(\overline{\mathbf{h}}') - T')\| \leq \sigma_{max}\,\|P(\overline{\mathbf{h}}) - P(\overline{\mathbf{h}}') - T'\|.$$

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

$$d_2(\langle\overline{\mathbf{h}}\rangle, \langle\overline{\mathbf{h}}'\rangle) \leq C\|P(\overline{\mathbf{h}}')\|_\infty^{\frac{1}{K}-1}\,\sigma_{max}\,\|P(\overline{\mathbf{h}}) - P(\overline{\mathbf{h}}') - T'\|.$$

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

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

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

Summarizing, we conclude that under the hypothesis of the theorem: For any $T \in P(\mathcal{M}^{\overline{L}}) - P(\mathcal{M}^{\overline{L}'})$ such that $\|\mathcal{A}T\| \leq \delta$ we have for any $T' \in \mathrm{Ker}\,(\mathcal{A})$

$$\|T\| \leq \gamma\|T - T'\|.$$

□

9.12 Proof of Proposition 8

Throughout the proof, we define, for any $\mathbf{i} \in [S]^K$, $\mathbf{h}^{\mathbf{i}} \in \mathbb{R}^{S \times K}$ by

$$\mathbf{h}^{\mathbf{i}}_{k,j} = \begin{cases} 1 & \text{if } j = \mathbf{i}_k \\ 0 & \text{otherwise} \end{cases} \quad \text{, for all } k \in [K] \text{ and } j \in [S]. \tag{35}$$

This notation shall not be confused with $\mathbf{h}^{\mathbf{p}}$, with $\mathbf{p} \in \mathcal{P}$.

- Let us first prove the first statement:
  We can easily check that $(P(\mathbf{h}^{\mathbf{i}}))_{\mathbf{i} \notin \mathbf{I}}$ forms a basis of $\{T \in \mathbb{R}^{S^K} \mid \forall \mathbf{i} \in \mathbf{I}, T_{\mathbf{i}} = 0\}$. We can also easily check using (21) that, for any $\mathbf{i} \notin \mathbf{I}$,

  $$\mathcal{A}P(\mathbf{h}^{\mathbf{i}}) = M_1(\mathbf{h}^{\mathbf{i}}_1) \dots M_K(\mathbf{h}^{\mathbf{i}}_K) = 0.$$

  Therefore, $\{T \in \mathbb{R}^{S^K} \mid \forall \mathbf{i} \in \mathbf{I}, T_{\mathbf{i}} = 0\} \subset \mathrm{Ker}(\mathcal{A})$.
  Conversely, for any $\mathbf{i} \in \mathbf{I}$, we can deduce from (21) and the hypotheses of the proposition that all the entries of $\mathcal{A}P(\mathbf{h}^{\mathbf{i}})$ are in $\{0,1\}$. We denote $D_{\mathbf{i}} = \{(i,j) \in [N] \times [N|\mathcal{F}|] \mid \mathcal{A}P(\mathbf{h}^{\mathbf{i}})_{i,j} = 1\}$. Using (again) the hypothesis of the proposition and (21), we can prove that, for any distinct $\mathbf{i}$ and $\mathbf{j} \in \mathbf{I}$, we have $D_{\mathbf{i}} \cap D_{\mathbf{j}} = \emptyset$. This easily leads to the item 1 of the first statement. We also deduce that

  $$\mathrm{rk}(\mathcal{A}) \geq |\mathbf{I}| = S^K - \dim(\{T \in \mathbb{R}^{S^K} \mid \forall \mathbf{i} \in \mathbf{I}, T_{\mathbf{i}} = 0\}).$$

  Finally, we deduce that $\dim(\mathrm{Ker}(\mathcal{A})) \leq \dim(\{T \in \mathbb{R}^{S^K} \mid \forall \mathbf{i} \in \mathbf{I}, T_{\mathbf{i}} = 0\})$ and therefore

  $$\mathrm{Ker}(\mathcal{A}) = \{T \in \mathbb{R}^{S^K} \mid \forall \mathbf{i} \in \mathbf{I}, T_{\mathbf{i}} = 0\}.$$

- Let us now prove the second statement:
  Using the hypothesis of the second statement and (21), we know that there is $f \in \mathcal{F}$ and $n \in [N]$ such that

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

  As a consequence, there is $\mathbf{i}$ and $\mathbf{j} \in [S]^K$ 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.

□

9.13 Proof of Proposition 9

The fact that, under the hypotheses of the proposition, $\mathrm{Ker}(\mathcal{A}) = \{0\}$ is a direct consequence of Proposition 8. The deep-NSP property and the value of γ also immediately follow from the definition of the deep-NSP.

To calculate σ_{min}, let us consider $T \in \mathbb{R}^{S^K}$ and express it under the form $T = \sum_{\mathbf{i}\in\mathbf{I}} T_{\mathbf{i}} P(\mathbf{h}^{\mathbf{i}})$, where $\mathbf{h}^{\mathbf{i}}$ is defined (35). Let us also remind that, applying Proposition 8, 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, its support is of size 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 σ_{min}. □

9.14 Proof of Theorem 9

Let us consider a path $\mathbf{p} \in \mathcal{P}$, using (21), since all the entries of $M_1(\mathbb{1})\dots M_K(\mathbb{1})$ belong to $\{0,1\}$, all the entries of $M_1(\mathbb{1}^{\mathbf{p}})\dots M_K(\mathbb{1}^{\mathbf{p}})$ belong to $\{0,1\}$. Therefore, we can apply Proposition 9 and Theorem 5 to the restriction of the convolutional linear network to $\mathbf{p}$ and obtain

$$d_p(\langle(\mathbf{h}^*)^{\mathbf{p}}\rangle, \langle\overline{\mathbf{h}}^{\mathbf{p}}\rangle) \leq \frac{7(KS')^{\frac{1}{p}}}{\sqrt{N}} \min\left(\|P(\overline{\mathbf{h}}^{\mathbf{p}})\|_\infty^{\frac{1}{K}-1}, \|P((\mathbf{h}^*)^{\mathbf{p}})\|_\infty^{\frac{1}{K}-1}\right)(\delta^{\mathbf{p}} + \eta^{\mathbf{p}}),$$

where $\delta^{\mathbf{p}}$ and $\eta^{\mathbf{p}}$ are the restrictions of the errors on $\mathrm{Supp}(\mathbf{p})$.

We therefore have

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

and finally

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

□

References

[1] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *stat*, 1050:8, 2013.

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

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

[4] 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.

[5] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *COLT*, pages 779–806, 2014.

[6] 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.

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

[8] Pierre F Baldi and Kurt Hornik. Learning in linear neural networks: A survey. *IEEE Transactions on neural networks*, 6(4):837–858, 1995.

[9] Anthony Bourrier, Mike Davies, Tomer Peleg, Patrick Pérez, and Rémi Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *IEEE Transactions on Information Theory*, 60(12):7928–7946, 2014.

[10] Emmanuel Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[11] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

[12] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[13] 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.

[14] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

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

[16] 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.

[17] 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, 2015.

[18] 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, 2015.

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

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

[21] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[22] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, pages 1141–1148, 2004.

[23] David L. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

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

[25] M Fazel, E Candes, B Recht, and P Parrilo. Compressed sensing and robust recovery of low rank matrices. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1043–1047. IEEE, 2008.

[26] Quan Geng, Huan Wangy, and John Wright. On the local correctness of ℓ_1-minimization for dictionary learning. In *International Symposium on Information Theory (ISIT)*, 2014.

[27] Wallace Givens. Computation of plain unitary rotations transforming a general matrix to triangular form. *Journal of the Society for Industrial and Applied Mathematics*, 6(1):26–50, 1958.

[28] Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. Sample complexity of dictionary learning and other matrix factorizations. *Information Theory, IEEE Transactions on*, 61(6):3469–3486, June 2015.

[29] Rémi Gribonval and Karin Schnass. Dictionary identification - sparse matrix-factorisation via l1-minimisation. *IEEE transaction on information theory*, 56(7):3523–3539, 2010.

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

[31] Joe Harris. *Algebraic geometry*, volume 133 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995. A first course, Corrected reprint of the 1992 original.

[32] Jonathan D. Hauenstein and Andrew J. Sommese. Witness sets of projections. *Applied Mathematics and Computation*, 217(7):3349 – 3354, 2010.

[33] Jonathan D. Hauenstein and Andrew J. Sommese. Membership tests for images of algebraic sets by linear projections. *Applied Mathematics and Computation*, 219(12):6809 – 6818, 2013.

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

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

[36] Risi Kondor, Nedelina Teneva, and Vikas Garg. Multiresolution matrix factorization. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1620–1628, 2014.

[37] Joseph M. Landsberg. *Tensors: Geometry and Applications*, volume 128 of *Graduate studies in mathematics*. American Mathematical Soc., 2012.

[38] 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.

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

[40] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *CoRR*, abs/1606.04933, 2016.

[41] Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[42] Shuyang Ling and Thomas Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *arXiv preprint arXiv:1512.07730*, 2015.

[43] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[44] 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.

[45] Siwei Lyu and Xin Wang. On algorithms for sparse multi-factor nmf. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 602–610, USA, 2013. Curran Associates Inc.

[46] Luc Le Magoarou. *Matrices efficientes pour le traitement du signal et l'apprentissage automatique*. PhD thesis, Université Bretagne Loire, 2016.

[47] Luc Le Magoarou and Rémi Gribonval. Are there approximate fast fourier transforms on graphs? In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4811–4815, 2016.

[48] Luc Le Magoarou and Rémi Gribonval. Flexible multi-layer sparse approximations of matrices and applications. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):688–700, 2016.

[49] 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.

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

[51] Behnam Neyshabur and Rina Panigrahy. Sparse matrix factorization. *arXiv preprint arXiv:1311.3315*, 2013.

[52] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[53] Ron Rubinstein, Alfred Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proc. IEEE - Special issue on applications of sparse representation and compressive sensing*, 98(6):1045–1057, 2010.

[54] Cristian Rusu, Nuria González-Prelcic, and Robert W Heath. Fast orthonormal sparsifying transforms based on householder reflectors. *IEEE Transactions on Signal Processing*, 64(24):6589–6599, 2016.

[55] Cristian Rusu and John Thompson. Learning fast sparsifying transforms. *arXiv preprint arXiv:1611.08230*, 2016.

[56] 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.

[57] Karin Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd. *Applied and Computational Harmonic Analysis*, 37(3):464–491, 2014.

[58] Igor R. Shafarevich. *Basic algebraic geometry. 1*. Springer, Heidelberg, third edition, 2013. Varieties in projective space.

[59] Daniel Spielmana, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *COLT*, pages 37–1, 2012.

[60] 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.

[61] Theodoros Tsiligkaridis, Alfred O Hero III, and Shuheng Zhou. On convergence of kronecker graphical lasso algorithms. *IEEE transactions on signal processing*, 61(7):1743–1755, 2013.

[62] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

[63] L. Venturi, A. S. Bandeira, and J. Bruna. Neural Networks with Finite Intrinsic Dimension have no Spurious Valleys. *ArXiv e-prints*, February 2018.