Implicit Regularization with Polynomial Growth in Deep Tensor Factorization

Kais Hariz, Hachem Kadri, Stéphane Ayache, Maher Moakher, Thierry Artières

Implicit Regularization with Polynomial Growth in Deep Tensor Factorization

Kais Hariz [1 2] **Hachem Kadri** [1] **Stéphane Ayache** [1] **Maher Moakher** [2] **Thierry Artières** [1 3]

Abstract

We study the implicit regularization effects of deep learning in tensor factorization. While implicit regularization in deep matrix and 'shallow' tensor factorization via linear and certain type of non-linear neural networks promotes low-rank solutions with at most quadratic growth, we show that its effect in deep tensor factorization grows polynomially with the depth of the network. This provides a remarkably faithful description of the observed experimental behaviour. Using numerical experiments, we demonstrate the benefits of this implicit regularization in yielding a more accurate estimation and better convergence properties.

1. Introduction

A major challenge in deep learning is to understand the underlying mechanisms behind the ability of deep neural networks to generalize. This is of fundamental importance to reconcile the observation that deep neural networks generalize well even for situations where the number of learnable parameters is much larger than the number of training data. Starting with the report by (Neyshabur et al., 2014), a body of work has emerged exploring the role of implicit regularization in deep learning (Gunasekar et al., 2017; Arora et al., 2019; Kumar & Poole, 2020; Razin & Cohen, 2020; Li et al., 2021; Razin et al., 2021; Milanesi et al., 2021; Zou et al., 2021). Our work contributes to this effort by providing insight into the behaviour of implicit regularization in deep tensor factorization where we focus on deep versions of canonical rank and Canonical-Polyadic (CP) factorizations (Kolda & Bader, 2009).

Attempts of studying implicit regularization in deep learning have identified matrix completion as a suitable testbed (Arora et al., 2019). Gunasekar et al. (2017) observed that for matrix factorization when there are no constraints on the rank, the solution of the optimization problem via gradient descent turns out to be a low-rank matrix. Furthermore, they conjectured that, with small enough learning rate and initialization, gradient descent on full-dimensional matrix factorization converges to the solution with minimal nuclear norm. Arora et al. (2019) and Razin & Cohen (2020) extended the analysis to deep matrix factorization and showed in this case that implicit regularization of gradient descent cannot be formulated as a norm-minimization problem. By studying the dynamics of gradient descent, they found theoretically and experimentally that it instead promotes sparsity of the singular values of the learned matrix, indicating that implicit regularization in deep learning has to be studied from a dynamical point of view. Moreover, Razin et al. (2021) studied implicit regularization in 'shallow' tensor decomposition and showed an equivalence between a tensor completion task and a prediction problem with a nonlinear neural network, stressing the interest of studying the tensor completion task.

Our main contributions focus on implicit regularization in deep Canonical-Polyadic tensor factorization and can be summarized as follows:

- we prove that the effect of the implicit regularization in deep tensor CP factorization via gradient descent grows polynomially with the depth of the factorization (Theorem 3.2),
- we theoretically show under some conditions that this effect in the overparameterized regime leads to produce solutions with low tensor rank (Theorem 3.3),
- we perform numerical experiments that support our theoretical results, illustrating that the implicit regularization could yield more accurate estimations and better convergence properties (Section 5).

2. Problem Setup and Background

We study implicit regularization in deep tensor factorization, and so we consider the problem of tensor completion with overparameterized factorization. By 'overparameterized'

[1]Aix Marseille University, CNRS, LIS, Marseille, France [2]LAMSIN, National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia [3]Ecole Centrale de Marseille, Marseille, France. Correspondence to: Kais Hariz <kais.hariz@univ-amu.fr>, Hachem Kadri <hachem.kadri@univ-amu.fr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

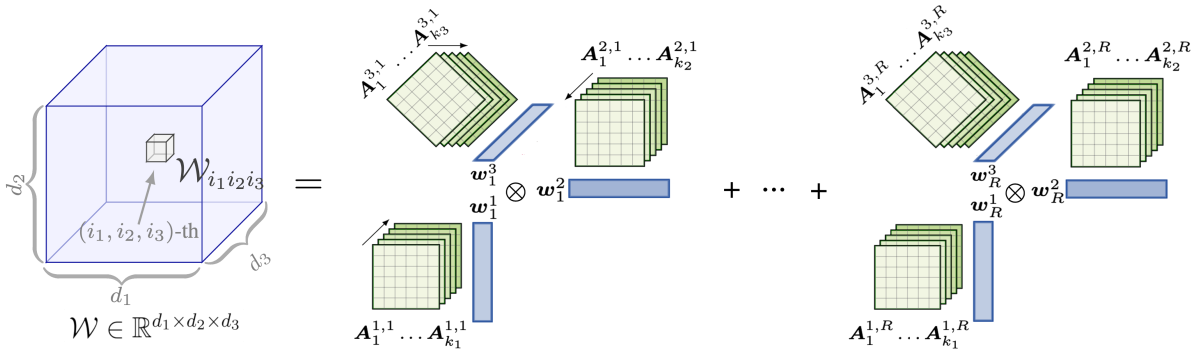


Figure 1. Overparameterized deep CP factorization.

we mean that no assumptions are made on the rank of the tensor. This is crucial in order to analyze the effect of the implicit regularization on the learned tensor.

Notation and terminology. Throughout the paper, boldface lowercase letters such as $\boldsymbol{w}$ denote vectors, boldface capital letters such as $\boldsymbol{A}$, $\boldsymbol{W}$ denote matrices, and calligraphic letters such as $\mathcal{W}$ denote tensors. The $(i_1, i_2, \dots, i_N)$-th entry of an N-order tensor $\mathcal{W} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ will be denoted by $\mathcal{W}_{i_1, i_2, \dots, i_N}$ where $i_n = 1, 2, \dots, d_n$, for all $n = [1, \dots, N]$. Given two tensors $\mathcal{V}, \mathcal{W} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$ their scalar product writes

$$\langle \mathcal{V}, \mathcal{W} \rangle = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} \mathcal{V}_{i_1, i_2 \cdots i_N} \mathcal{W}_{i_1, i_2 \cdots i_N},$$

and the Frobenius norm of $\mathcal{W}$ is defined as $\|\mathcal{W}\| := \sqrt{\langle \mathcal{W}, \mathcal{W} \rangle}$. We also write $\|\boldsymbol{A}\|$ and $\|\boldsymbol{w}\|$ for the Frobenius norm of a matrix $\boldsymbol{A}$ and the Euclidean norm of a vector $\boldsymbol{w}$, respectively.

Given N vectors $\boldsymbol{w}_1 \in \mathbb{R}^{d_1}, \boldsymbol{w}_2 \in \mathbb{R}^{d_2}, \dots, \boldsymbol{w}_N \in \mathbb{R}^{d_N}$, their outer product is the tensor whose $(i_1, \dots, i_N)$-th entry writes $(\boldsymbol{w}_1 \otimes \boldsymbol{w}_2 \otimes \cdots \otimes \boldsymbol{w}_N)_{i_1, i_2 \dots, i_N} = (\boldsymbol{w}_1)_{i_1} (\boldsymbol{w}_2)_{i_2} \dots (\boldsymbol{w}_N)_{i_N}$ for all $i_n \in [1, \dots, d_n]$, $n \in [1, \dots, N]$. An N-th order tensor $\mathcal{W}$ is called a rank-1 tensor if it can be written as the outer product of N vectors, i.e. $\mathcal{W} = \boldsymbol{w}_1 \otimes \boldsymbol{w}_2 \otimes \cdots \bigotimes \boldsymbol{w}_N$. This leads to the notion of canonical rank and Canonical-Polyadic (CP) factorization (Kolda & Bader, 2009).

The canonical rank of an arbitrary N-th order tensor $\mathcal{W}$ is the minimal number of rank-1 tensors that sum up to $\mathcal{W}$. Decompositions into rank-1 terms are called CP factorization. A rank R tensor $\mathcal{W}$ can be written as:

$$\mathcal{W} = \sum_{r=1}^{R} \boldsymbol{w}_r^1 \otimes \ldots \otimes \boldsymbol{w}_r^N. \tag{1}$$

We will use the term 'block' to refer to a rank-1 tensor of the CP decomposition so that $\mathcal{W}$ in Eq. (1) has R blocks and $\boldsymbol{w}_r^1 \otimes \ldots \otimes \boldsymbol{w}_r^N$ is its r-th block. In this work we focus only on the CP factorization of incomplete tensor. We did not consider any other tensor decomposition, such as Tucker and TensorTrain (Kolda & Bader, 2009; Grasedyck et al., 2013), which remain then out of the scope of this paper.

Overparameterized deep CP factorization. Let $\mathcal{A} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ the ground truth tensor we want to recover and let us denote by $\Omega \subset \{1, 2, \dots d_1\} \times \cdots \times \{1, 2, \dots, d_N\}$ the set which indexes the observed entries. To tackle the problem of tensor completion (Gandy et al., 2011; Song et al., 2019), we minimize the reconstruction loss defined by

$$\mathcal{L}(\mathcal{W}) = \frac{1}{|\Omega|} \sum_{(i_1, \dots, i_N) \in \Omega} \ell\left(\mathcal{W}_{i_1, \dots, i_N} - \mathcal{A}_{i_1, \dots, i_N}\right),$$

where ℓ is differentiable and locally smooth. The square loss, which we used in our experiments, is obtained when $\ell(z) = \frac{1}{2} z^2, \forall z \in \mathbb{R}$. In order to be able to investigate the mechanisms of implicit regularization in deep tensor factorization, we consider the following overparameterized deep CP decomposition (see Figure 1):

$$\mathcal{W} = \sum_{r=1}^{R} \left(\boldsymbol{A}_1^{1,r} \dots \boldsymbol{A}_{k_1}^{1,r} \boldsymbol{w}_r^1\right) \otimes \dots \otimes \left(\boldsymbol{A}_1^{N,r} \dots \boldsymbol{A}_{k_N}^{N,r} \boldsymbol{w}_r^N\right), \tag{2}$$

where $\boldsymbol{w}_r^n \in \mathbb{R}^{d_n}$ and $\boldsymbol{A}_i^{n,r} \in \mathbb{R}^{d_n \times d_n}$, $\forall n = 1, \dots, N$ and $i = 1, \dots, k_n$. We take a large R value to avoid any restriction of the CP rank. The matrices $\boldsymbol{A}_1^{n,r}, \dots, \boldsymbol{A}_{k_n}^{n,r}$ can be seen as a deep matrix factorization for the n-th mode and the r-th block of the CP decomposition and k_n is the depth of the factorization on the mode n. All these matrices are of dimension $d_n \times d_n$ such that no constraint on the rank is imposed (Arora et al., 2019). When they are fixed to the identity matrix, we recover the standard CP decomposition as defined in Eq. (1). In the sequel, we use the following

compact form to rewrite Eq. (2):

$$\mathcal{W} = \sum_{r=1}^{R} \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r} \boldsymbol{w}_r^n. \tag{3}$$

Our main aim is to highlight the role of implicit regularization in deep CP tensor factorization and characterize its dependence on the depth of the factorization. In the following, we consider learning a tensor $\mathcal{W}$ which has the form (3) by minimizing the loss function $\mathcal{L}(\mathcal{W}) = \Phi\Big(\{\boldsymbol{w}_r^n\}_{r=1\ n=1}^{R\ N}, \{\boldsymbol{A}_i^{n,r}\}_{r=1\ n=1\ i=1}^{R\ N\ k_n}\Big)$ using gradient descent. With infinitesimally small learning rate and non zero initialization, we have

$$\frac{d}{dt}\boldsymbol{w}_r^n(t) = -\frac{\partial}{\partial \boldsymbol{w}_r^n}\Phi\Big(\{\boldsymbol{w}_{r'}^{n'}(t)\}_{r',n'}, \{\boldsymbol{A}_i^{n',r'}(t)\}_{r',n',i}\Big),$$

and

$$\frac{d}{dt}\boldsymbol{A}_i^{n,r}(t) = -\frac{\partial}{\partial \boldsymbol{A}_i^{n,r}}\Phi\Big(\{\boldsymbol{w}_{r'}^{n'}(t)\}_{r',n'}, \{\boldsymbol{A}_i^{n',r'}(t)\}_{r',n',i}\Big).$$

Note that Razin et al. (2021) have made a connection between tensor completion via CP tensor factorization and a certain-type of non-linear one hidden layer neural network, motivating their work as a important step towards the study of implicit regularization in standard neural networks. From this perspective, our overparameterized CP factorization may be viewed as an extension of this statement to the deep setting.

Related work. The works that are most related to ours are Arora et al. (2019); Razin et al. (2021). Arora et al. (2019) considered deep matrix factorization, which consists in parameterizing the learned matrix $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$ as $\boldsymbol{W} = \boldsymbol{W}_k \boldsymbol{W}_{k-1} \dots \boldsymbol{W}_2 \boldsymbol{W}_1$ for some $k \in \mathbb{N}$ and with $\{\boldsymbol{W}_i\}_{i=1}^k$ to be such that no constraint on the rank is present. Notice that deep matrix factorization is a generalization of the shallow matrix factorization setup investigated by (Gunasekar et al., 2017), which corresponds to the case where $k = 2$. They observed that depth enhances recovery performances. This led them to study the dynamics in optimization and they found out that gradient descent promotes sparsity of singular values of $\boldsymbol{W}$, as summarized in the theorem below.

Theorem 2.1 (Arora et al., 2019)**.** *For depth $k \geq 2$, for any $r = 1, \dots, \min(d_1, d_2)$,*

$$\frac{d}{dt}\sigma_r(t) = k\alpha_r(t) \cdot (\sigma_r(t))^{2-\frac{2}{k}}, \tag{4}$$

where $\alpha_r(t) = \big\langle -\nabla\mathcal{L}(\boldsymbol{W}(t)), \boldsymbol{u}_r(t)\boldsymbol{v}_r^\top(t)\big\rangle$, $\sigma_r(t)$ is the r-th singular value of $\boldsymbol{W}$ at time $t \geq 0$, and $\boldsymbol{u}_r(t)$ and $\boldsymbol{v}_r(t)$ are its r-th singular vectors.

Razin et al. (2021) extended this analysis to CP tensor factorization (see Eq. 1) and showed the following result.

Theorem 2.2 (Razin et al., 2021)**.** *Under certain assumptions, for any $r = 1, \dots, R$,*

$$\frac{d}{dt}\left\|\bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t)\right\| = N\gamma_r(t) \cdot \left\|\bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t)\right\|^{2-\frac{2}{N}}, \tag{5}$$

where $\gamma_r(t) = \left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n=1}^{N} \frac{\boldsymbol{w}_r^n(t)}{\|\boldsymbol{w}_r^n(t)\|} \right\rangle$, N is the order of the tensor $\mathcal{W}$ to be learned, and $\boldsymbol{w}_r^n$ is the vector of the n-th mode and the r-th block of the CP decomposition.

This shows that training a CP tensor factorization via gradient descent with small learning rate and near-zero initialization tends to produce tensors with low canonical rank. Note that the result in Eq. (4) depends on the depth of the factorization k, while the one in Eq. (5) depends on the order of the tensor N. In both cases the impact of implicit regularization *grows* at most *quadratically*, with either k or N.

As far as we are aware, the only work on implicit regularization in deep tensor factorization appears in Milanesi et al. (2021), in which the authors considered Tucker and Tensor Train (TT) decompositions and observed that, even in the case where the rank is not constrained, only a small number of higher-order singular values (De Lathauwer et al., 2000) and TT singular values (Oseledets, 2011) are retained by a gradient-based neural network.[1] However, no theoretical justification is given there.

3. Main Results

We now present our main results, to be proved in Section 4. Following the idea of Razin et al. (2021), we provide a dynamical characterization of the trajectories of the norm of each block of the deep CP factorization. To proceed we need the following definition.

Definition 3.1. The unbalancedness magnitude at time $t \geq 0$ of the weight vectors and matrices of the CP factorization in Eq. (3) is defined as :

$$\varepsilon(t) = \max_{\substack{r\in\{1,\dots,R\},(n,m)\in\{1,\dots,N\}^2 \\ i\in\{1,\dots,k_m\}}} \Big| \|\boldsymbol{w}_r^n(t)\|^2 - \|\boldsymbol{A}_i^{m,r}(t)\|^2 \Big|.$$

Note that this notion of *unbalancedness magnitude* of the deep CP factorization is inspired from Razin et al. (2021), where the unbalancedness magnitude (Du et al., 2018) of the weight vectors of the CP decomposition was introduced.

[1] After this paper was submitted, a paper by Razin et al. (2022) was released on arXiv, which studied implicit regularization in hierarchical tensor factorization.

Note that $\varepsilon(0)$ is per definition purely determined by the initialization. We will show later that $\varepsilon(t)$ is constant during the gradient descent optimization, which is crucial to show our first main result.

Theorem 3.2. *Assume that $\varepsilon(0) = 0$. Then, for any $r \in \{1, \dots, R\}$ and time $t \geq 0$ at which $\left\| \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\| \neq 0$:*

(i) The weight vectors of the CP factorization in Eq. (3) evolve according to

$$\frac{d}{dt} \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\| = N \delta_r(t) \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|^{2 - \frac{2}{N} + \frac{k_1 + \dots + k_N}{N}},$$

where $\delta_r(t) := \left\langle -\nabla \mathcal{L}(\mathcal{W}(t)), \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \widehat{\boldsymbol{A}}_i^{n,r}(t) \widehat{\boldsymbol{w}}_r^n(t) \right\rangle$, $\widehat{\boldsymbol{w}}_r^n(t) := \frac{\boldsymbol{w}_r^n}{\|\boldsymbol{w}_r^n(t)\|}$ and $\widehat{\boldsymbol{A}}_i^{n,r}(t) := \frac{\boldsymbol{A}_i^{n,r}(t)}{\|\boldsymbol{A}_i^{n,r}(t)\|}$.

(ii) If in addition $\{\boldsymbol{A}_i^{n,r}(0)\}_{r=1}^{R}{}_{n=1}^{N}{}_{i=1}^{k_n}$ are rank-one matrices satisfying

$$\boldsymbol{A}_i^{n,r}(0)^\top \boldsymbol{A}_i^{n,r}(0) = \boldsymbol{A}_{i+1}^{n,r}(0) \boldsymbol{A}_{i+1}^{n,r}(0)^\top,$$

for all $i \in \{1, \dots, k_n - 1\}$ with $k_n \geq 2$ and $n \in \{1, \dots, N\}$, then

$$\delta_r(t) = \left\langle -\nabla \mathcal{L}(\mathcal{W}(t)), \bigotimes_{n=1}^{N} \frac{\prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t)}{\left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\|} \right\rangle \zeta_r(t),$$

where $\zeta_r(t) := \left| \left\langle \bigotimes_{n=1}^{N} \tilde{\boldsymbol{v}}_r^n(t), \bigotimes_{n=1}^{N} \widehat{\boldsymbol{w}}_r^n(t) \right\rangle \right|$ and $\tilde{\boldsymbol{v}}_r^n(t)$ is the first right singular vector of $\boldsymbol{A}_{k_n}^{n,r}(t)$.

By Theorem 3.2, if all the depths $k_1, \dots, k_N$ are equal to the same value, say k, we obtain:

$$\frac{d}{dt} \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\| = N \delta_r(t) \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|^{2 - \frac{2}{N} + k}. \quad (6)$$

Note that Part (ii) of Theorem 3.2 allows us to express $\delta_r(t)$ defined in Part (i) in terms of quantities with norms that do not depend with depths. This is achieved by characterizing the evolution of the singular values of the product of the matrix parameters $\boldsymbol{A}_i^{n,r}(t)$ (see Lemma A.2).

This shows that the evolution rates of the norm of the blocks of the tensor $\sum_{r=1}^{R} \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n$ are proportional to their norm raised to the power of $2 - \frac{2}{N} + k$. This is in line with the observation that CP block norms move faster when large and slower when small, as reported by Razin et al. (2021). More interestingly, this effect is more pronounced with larger depths. When the depth k increases, one block r, likely the one whose norm is maximum, will see its norm increases significantly faster (the bigger the value of k the faster) than the norm of all other blocks, up to a stability stage when $\delta_r(t)$ converges towards 0, meaning this block is somehow optimized.

This sequential block optimization mechanism promotes low-rankness in a greedy fashion where the blocks are selected and optimized one after the other, which will be confirmed experimentally. Interestingly, similar observations were made in Arora et al. (2019); Li et al. (2021); Ge et al. (2021). An interesting property of deep CP factorization that our theoretical analysis reveals, lies in that the effect of the depth of the factorization on the implicit regularization is *polynomial*, while it is *quadratic* in deep matrix and 'shallow' CP decomposition, as shown in Theorems 2.1 and 2.2.

One consequence of the greedy sequential optimization of the blocks is that when a sufficient number of blocks of $\sum_{r=1}^{R} \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n$ are effectively used (having a norm significantly different from zero) and have been optimized the other block remain ignored with a very small norm. Also, the number of effective blocks quickly decreases with increasing depth k.

Yet, Theorem 3.2 does not explicitly specify the effect of the implicit regularization on the learned tensor $\mathcal{W}$ by the deep CP factorization. The following theorem shows that the dynamical characterization provided above would favor selecting only a few number of the blocks of $\mathcal{W}$, promoting low canonical rank solutions.

Theorem 3.3. *Assume that $\varepsilon(0) = 0$. Then, for any time $t \geq 0$ of the optimization of the CP factorization in Eq. (3), the following inequality holds for all $r \in \{1, \dots, R\}$:*

$$\left\| \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\| \leq \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|^{1 + \frac{k_1 + \dots + k_N}{N}}.$$

If $k_1 = \dots = k_N$, let us say k, then:

$$\left\| \bigotimes_{n=1}^{N} \prod_{i=1}^{k} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\| \leq \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|^{1+k}. \quad (7)$$

Let us denote by $N_1(t)$ the number of blocks with non zero norm of the factorized tensor $\mathcal{W} = \sum_{r=1}^{R} \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r} \boldsymbol{w}_r^n$ at iteration t, and let $N_2(t)$ be the number of blocks with non zero norm of the factorized tensor $\sum_{r=1}^{R} \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n$ at iteration t. Theorem 3.3 shows that the term $\left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|$ controls the norm of the r-th

block of the deep CP factorization. If it is close to zero, then $\left\| \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\|$ is close to zero as well. Then, whatever the iteration t, $N_1(t) \leq N_2(t)$. Considering that $\mathrm{rank}(\mathcal{W}) \leq N_1(t)$, we can conclude that the rank of the learned tensor $\mathcal{W}$ is bounded by $N_2(t)$. Putting all together Theorem 3.2 shows that the depth k of the factorization ensures the convergence of the tensor $\sum_{r=1}^{R} \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n$ towards a tensor with a small $N_2(t)$ value, hence leads the deep CP factorization to converge to a tensor $\mathcal{W}$ that has a low canonical rank.

4. Proof Overview

To prove Theorem 3.2, we need the following result.

Lemma 4.1. *For all $r \in \{1, \dots, R\}$, $n, m \in \{1, \dots, N\}$, $i \in \{1, \dots, k_n\}$ and $j \in \{1, \dots, k_m\}$, the followings hold $\forall t \geq 0$,*

(i) $\|\boldsymbol{A}_i^{n,r}(t)\|^2 - \|\boldsymbol{A}_j^{m,r}(t)\|^2 = \|\boldsymbol{A}_i^{n,r}(0)\|^2 - \|\boldsymbol{A}_j^{m,r}(0)\|^2$,

(ii) $\|\boldsymbol{w}_r^n(t)\|^2 - \|\boldsymbol{w}_r^m(t)\|^2 = \|\boldsymbol{w}_r^n(0)\|^2 - \|\boldsymbol{w}_r^m(0)\|^2$,

(iii) $\|\boldsymbol{w}_r^n(t)\|^2 - \|\boldsymbol{A}_i^{n,r}(t)\|^2 = \|\boldsymbol{w}_r^n(0)\|^2 - \|\boldsymbol{A}_i^{n,r}(0)\|^2$.

The proof of Lemma 4.1 is in Appendix A. The essential idea of the proof comes from Razin et al. (2021). The key steps are as follows. We first show that $\frac{d}{dt}\|\boldsymbol{A}_i^{n,r}(t)\|^2$ is independent of n and i. So $\forall (n,m) \in [\![1, N]\!]^2$ and $\forall (i,j) \in [\![1, k_n]\!] \times [\![1, k_m]\!]$, the derivatives of $\|\boldsymbol{A}_i^{n,r}(t)\|^2$ and $\|\boldsymbol{A}_j^{m,r}(t)\|^2$ are equal, which means that $\|\boldsymbol{A}_i^{n,r}(t)\|^2 - \|\boldsymbol{A}_j^{m,r}(t)\|^2$ does not vary with t. The assertion (ii) is shown by the same arguments as above. The proof of (iii) is based on the observation that, $\forall n, m \in [\![1, N]\!]$ and $\forall i \in [\![1, k_n]\!]$, $\frac{d}{dt}\|\boldsymbol{w}_r^n(t)\|^2 = \frac{d}{dt}\|\boldsymbol{A}_i^{n,r}(t)\|^2$.

We now present a sketch of the proof of Theorem 3.2. Full details of the proof is provided in Appendix A. The ideas of the proof are similar to the ideas in Razin et al. (2021), with the difference that the extension to deep CP factorization will result in some technical complications due to the presence of the weight matrices $\boldsymbol{A}_i^{n,r}(t)$.

Sketch of the proof of Theorem 3.2. The main step of the proof of part (i) of the theorem is to show that

$$\frac{d}{dt}\left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\| = N \delta_r(t) \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|^{2 - \frac{2}{N} N} \prod_{n=1}^{N} \prod_{i=1}^{k_n} \|\boldsymbol{A}_i^{n,r}(t)\|. \tag{8}$$

This result is obtained by deriving upper and lower bounds of the first term in (8), which converges to the same value when the unbalancedness magnitude assumption is satisfied. Then, using Lemma 4.1 and the assumption $\epsilon(0) = 0$, we prove that $\|\boldsymbol{A}_i^{n,r}(t)\| = \left\| \bigotimes_{m=1}^{N} \boldsymbol{w}_r^m(t) \right\|^{\frac{1}{N}}$. Plugging the result into Eq. (8) completes the proof.

To prove part (ii) of the theorem, we state Lemma A.2 which characterizes the evolution of the singular values of the product matrix $\prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t)$. The proof of this lemma is inspired by Theorem 3 of Arora et al. (2019). This allows us to express $\left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\|$ in terms of $\left(\prod_{i=1}^{k_n} \|\boldsymbol{A}_i^{n,r}(t)\| \right) \|\boldsymbol{w}_r^n(t)\|$.

Proof of Theorem 3.3. Let us first recall that $\|\boldsymbol{A}_i^{n,r}(t)\| = \left\| \bigotimes_{m=1}^{N} \boldsymbol{w}_r^m(t) \right\|^{\frac{1}{N}}$, when $\epsilon(0) = 0$. We have,

$$\begin{aligned}
\left\| \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\| &= \prod_{n=1}^{N} \left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\| \\
&\leq \prod_{n=1}^{N} \left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \right\| \|\boldsymbol{w}_r^n(t)\| \\
&\leq \prod_{n=1}^{N} \left(\prod_{i=1}^{k_n} \|\boldsymbol{A}_i^{n,r}(t)\| \right) \|\boldsymbol{w}_r^n(t)\| \\
&= \prod_{n=1}^{N} \left(\prod_{i=1}^{k_n} \left\| \bigotimes_{m=1}^{N} \boldsymbol{w}_r^m(t) \right\|^{\frac{1}{N}} \right) \prod_{n=1}^{N} \|\boldsymbol{w}_r^n(t)\| \\
&= \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|^{\frac{k_1 + \dots + k_N}{N}} \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\| \\
&= \left\| \bigotimes_{n=1}^{N} \boldsymbol{w}_r^n(t) \right\|^{1 + \frac{k_1 + \dots + k_N}{N}}.
\end{aligned}$$

5. Experimental Analysis

We present here an experimental analysis that helps understanding our main theoretical results. We first detail the experimental settings and investigate the main trends that we observed during learning. We then report a more extensive analysis of the low rank inducing feature of deep over-parameterized tensor optimization. Finally, we explore how and when depth may help improving loss minimization.

5.1. Experimental Settings

We focus on tensor completion task: given a partially observed tensor $\mathcal{A}$, we learn a model to match inputs indices (tuples of size N, the order of $\mathcal{A}$) to the observed values. We generated synthetic data in order to analyze and

control the phenomena of implicit regularization in tensor factorization.

Synthetic data. We generated a $10\times10\times10\times10$ tensor with CP-rank equal to 5 and entries sampled from a centered and reduced normal distribution. From this complete tensor, we randomly split the set of (indices, values) to build training and testing sets. In the following, we comment experiments for various ratio of observed values (from 25% to 10%).

Model initialization. As we just said, entries of $\boldsymbol{w}_r^n$ are sampled from a centered normal distribution with a small variance σ_w. In our deep CP formulation, we also need to initialize entries of $\boldsymbol{A}_i^{n,r}$ from a centered normal distribution with small variance σ_A. Alike in previous works studying implicit regularization, we remarked the sensitivity of implicit regularization to model's initialization. With a too large σ_w the model does not converge to a low-rank solution, while with a too small σ_w a solution might exists, however, after a prohibitive number of gradient descent iterations. With our deep formulation, the product of multiple - small norm - matrices may lead to numerical instabilities and/or to the well known vanishing gradient. However, we observed implicit regularization with highest values in matrix initialization. In Jing et al. (2020), authors empirically found that standard Kaiming (He) initialization (He et al., 2015) of multiple matrices stacked after the encoder is able to yield implicit regularization in the latent space. In order to be close to theoretical conditions, we use in our simulations zero mean and near zero standard deviation for initializing the parameters to be learned. We have also considered matrices that are initialized with small values except on the diagonal to speed up learning.

5.2. Investigating Learning Behaviour

First, we investigate typical phenomenon that we observed during the learning. In all the experiments that we report here the percentage of observed and unobserved inputs in the tensor are 20% and 80%. Importantly, the learning is very sensitive to initial settings, meaning whatever the depth the learning may converge towards low-rank or high-rank solutions (see Figure 4).

Second, Figures 2 and 3 show typical learning behaviour with respect to depth, for two different initialization settings, where one sees in both cases that shallow architectures tends to converge to high-rank solutions with many blocks exhibiting a non negligible norm, while increasing depth makes most blocks' norm converge to 0 yielding a much more relevant low-rank solution.

Third, one may note (see e.g. Figures 2 (c) and 3 (d)) that blocks emerge sequentially, in a greedy fashion along the training process, one at a time. This phenomenon has already been observed in e.g. Razin & Cohen (2020) and

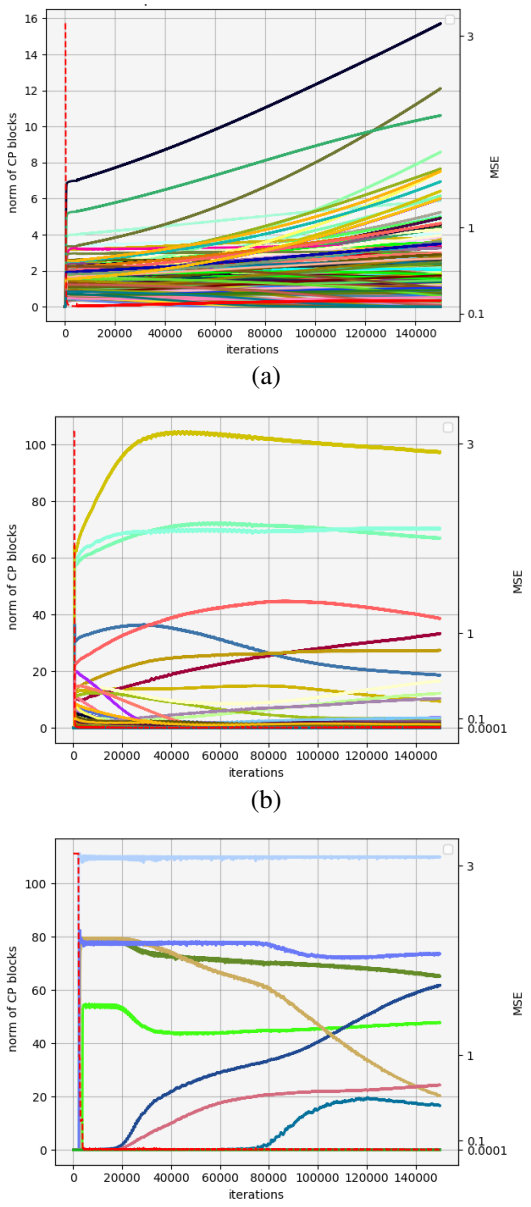
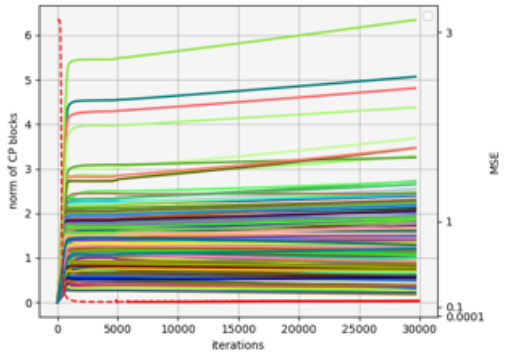
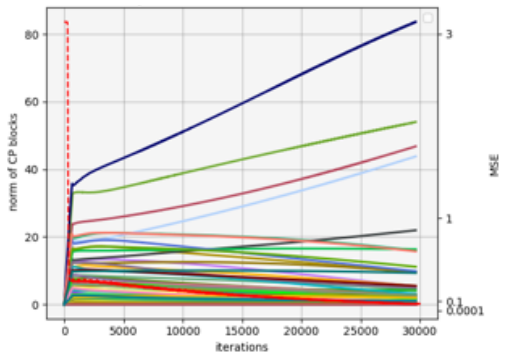


Figure 2. This series of figures compares the learning behaviour for shallow to deep tensors through the evolution of the norms of the blocks along training epochs for depth 0 (a), 1 (b), and 2 (c) for a particular initialization ($\sigma_w = 0.01$, $\sigma_A = 0.01$). Each figure shows 500 curves corresponding to the norm of blocks (y-scale on the left of the plot), plus an additional curve (dotted red line) which stands for the loss (with y-scale on the right of the plot in log-scale).

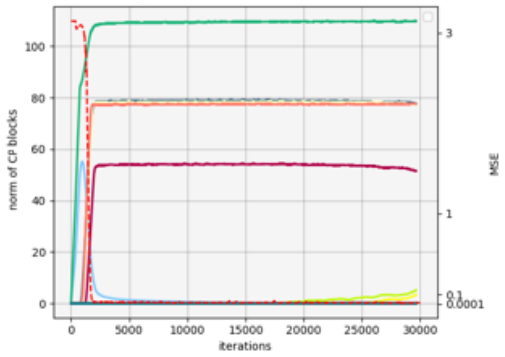
is a consequence of the dynamics rule in Theorem 3.2, as discussed in Section 3. We also observed blocks whose norm rises suddenly then decrease to converge to their final norm which may be also a consequence of the polynomial dynamics as well, this is illustrated in Figure 2.



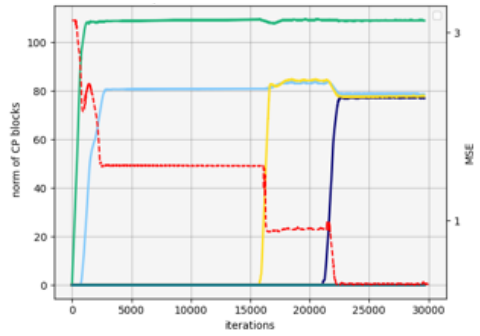
(a)



(b)



(c)



(d)

Figure 3. This series of figures compares the learning behaviour for shallow to deep tensors through the evolution of the norms of the blocks along training epochs for depth 0 (a), 1 (b), 3 (c), and 5 (d) for a particular initialization ($\sigma_w = 0.005$ and $\sigma_A = 0.1$).

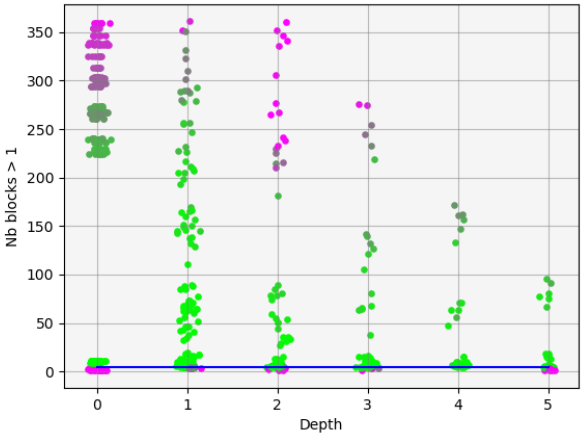


Figure 4. Analysis of the impact of the depth on the rank of the learned tensor. The figure shows for shallow (on the left, depth=0) to deep models (up to depth=5, on the right) the effective rank of the learned tensors for a number of runs that differ from initialization setting. Every single point stands for a learning experiment. Points are plotted with a small random displacement in x and y coordinates to better see point clouds. The color represents the test loss of the model, from green to purple respectively for small to large loss (ranging from 5.10^{-5} to 3.35). Few runs diverged and some led to 0 rank due to lacks of iterations (11 over 1500), those runs are not reported here. The blue line shows the real tensor rank (5) to approximate. Runs below this line lead to high losses. Conversely, most of runs which converged to higher ranks are able to minimize the loss objective.

5.3. How Depth Yields Low-Rank Solutions

We summarize in Figure 4 a number of experiments that illustrate the effectiveness of deeper architectures to consistently converge to low-rank solutions whose rank is close to the true tensor rank, whatever the initial conditions. We launched more than 1500 learning experiments using various initialization parameters and seeds, for depth ranging from 0 to 5. We report for each experiment the effective CP-rank of the model, i.e. the number of blocks that emerged at convergence (blocks that have a norm greater than 1). Again in all experiments reported here the percentage of observed and unobserved inputs in the tensor are 20% and 80%.

One can observe much more variability on the learned tensor rank when the depth is limited. For shallower architectures, the impact of initialization is huge and the solutions are mostly of high rank. For deeper architectures, the learned tensor rank is much more stable, close to the true tensor rank, hence showing lower dependency to the initialization setting.

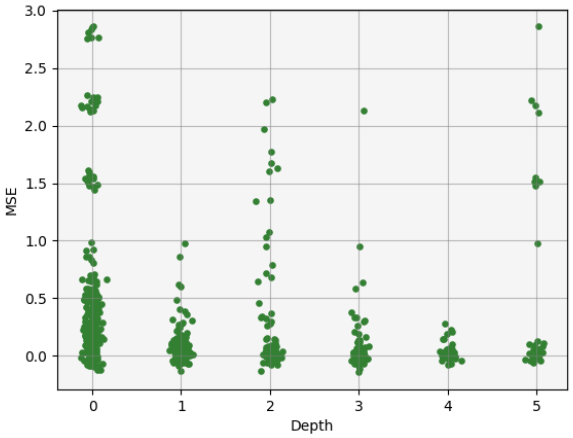


Figure 5. Analysis of the impact of the depth (x-axis) on the generalization loss (y-axis). The figure shows from shallow (on the left, depth=0) to deep models (up to depth=5, on the right), the generalization losses achieved in a number of runs that differ from initialization setting. Every single point stands for a learning experiment. Points are plotted with a small random displacement in x and y coordinates to better see point clouds.

5.4. Impact of the Depth on Generalization Loss

Finally, we explore how and when the depth may help achieving generalization. Figure 5 is a similar plot as Figure 4 but where the y-axis stands for the generalization loss. We again run many experiments for depth from 0 to 5 and for various initialization settings. In this figure all experiments reported have been obtained using percentages of observed and unobserved inputs equal to 20% and 80%.

As may be observed, whatever the depth, a small generalization loss may be achieved. However, increasing the depth makes the optimization much more robust and stable with respect to initialization. Depth consistently helps reaching best achievable generalization loss whatever the initialization. Hence, increasing the depth allows reaching low-rank approximation as well as low generalization loss.

To go deeper in the analysis, Table 1 reports for the depth ranging from 0 to 5, and for a percentage of unobserved values ranging from 75% to 90%, the smallest loss obtained on validation data whatever the initialization, and the rank of the corresponding learned tensor (in brackets). As we have already discussed, when running many experiments with various initialization settings one may most often get a low-rank solution whatever the depth (see bottom points for every depth in Figure 4). This explains why many of these best performing solutions are of low rank, whatever the depth (e.g. first line with depth=0) and the percentage of unobserved inputs. Yet, these results show that depth often allows reaching low-rank solutions as well as low generalization loss, and thus achieving a good trade-off between tensor rank and generalization error.

	75	85	90
0	7.925e-05 (5)	2.598e-05 (14)	5.836e-06 (11)
1	1.479e-05 (67)	2.685e-04 (12)	1.42e-05 (7)
2	3.607e-06 (10)	5.215e-05 (17)	2.37e-04 (13)
3	2.073e-06 (13)	8.048e-04 (10)	2.933e-04 (6)
4	8.053e-04 (5)	2.238e-04 (6)	2.350e-04 (9)
5	8.116e-03 (5)	8.169e-01 (3)	5.265e-01 (4)

Table 1. Comparison of best performing tensor factorization for a number of cases corresponding to few architecture depths and to various percentages of missing tensor inputs at training time, ranging from 75% to 90%. The best generalization and the effective rank of the best tensor factorization (in brackets) are reported.

5.5. Real-World Data

We also run experiments on two real data sets. We used Meteo-UK[2] and CCDS[3] data sets (Lozano et al., 2009), which contain monthly measurements of temporal variables in various stations across UK and North America, resulting in tensors of dimension $(50, 16, 5)$ and $(50, 15, 25)$, respectively. Figure 6 shows completion performance with 30% of observed data for multiple runs varying with initialisation std in $[0.0005, \ldots, 0.001]$. The obtained experimental results on real-world data corroborate the simulations and confirm our theoretical findings. We also used in this experiment random initialization with rank-one matrices and observed that the same experimental behavior is reproduced (see Appendix B).

6. Conclusion

We provided a theoretical analysis of implicit regularization in deep tensor factorization building on previous advances studying tensor and deep matrix factorization. Our results suggest a form of greedy low tensor rank search, but where the impact of implicit regularization is polynomially dependent of the depth. Experiments confirmed our theoretical results and provided insights on the main role of initialization, especially for shallow architectures. While shallow architectures may converge to low-rank and high-rank solutions, deeper factorization consistently converge to low-rank solutions, close to the true tensor rank. Finally, deeper architectures seem to help reaching more consistently best performing solutions with respect to the generalization error.

[2] https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data.

[3] https://viterbi-web.usc.edu/~liu32/data.html.

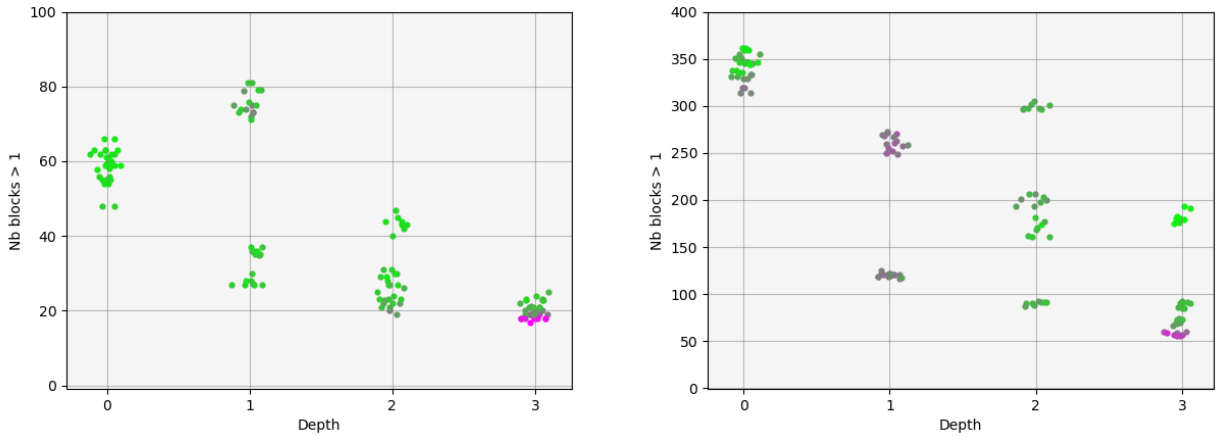


Figure 6. Tensor completion using Meteo-UK (left) and CCDS (right) data sets: analysis of the impact of the depth on the rank of the learned tensor. The figure shows for shallow (on the left, depth=0) to deep models (up to depth=5, on the right) the effective rank of the learned tensors for a number of runs that differ from initialization setting. Every single point stands for a learning experiment. Points are plotted with a small random displacement in x and y coordinates to better see point clouds. The color represents the test loss of the model, from green to purple respectively for small to large loss (ranging from 0.3 to 1.1 for Meteo-UK and from 0.59 to 1.12 for CCDS).

Acknowledgements

We thank the reviewers and the meta-reviewer for their helpful comments and for suggesting a way to improve Theorem 3.2. Research reported in this paper was partially supported by PHC Utique no. 44318NJ granted by the Ministry of Higher Education and Scientific Research of Tunisia and the Ministry of Foreign Affairs in France.

References

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning (ICML)*, pp. 244–253, 2018.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

De Lathauwer, L., De Moor, B., and Vandewalle, J. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

Du, S., Hu, W., and Lee, J. D. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Gandy, S., Recht, B., and Yamada, I. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse problems*, 27(2):025010, 2011.

Ge, R., Ren, Y., Wang, X., and Zhou, M. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Grasedyck, L., Kressner, D., and Tobler, C. A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78, 2013.

Gunasekar, S., Woodworth, B., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Jing, L., Zbontar, J., et al. Implicit rank-minimizing autoencoder. *Advances in Neural Information Processing Systems*, 33, 2020.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Kumar, A. and Poole, B. On implicit regularization in β-VAEs. In *International Conference on Machine Learning (ICML)*, 2020.

Li, Z., Luo, Y., and Lyu, K. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representation (ICLR)*, 2021.

Lozano, A. C., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J., and Abe, N. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 587–596, 2009.

Milanesi, P., Kadri, H., Ayache, S., and Artières, T. Implicit regularization in deep tensor factorization. In *International Joint Conference on Neural Networks (IJCNN)*, 2021.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Oseledets, I. V. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Razin, N. and Cohen, N. Implicit regularization in deep learning may not be explainable by norms. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Razin, N., Maman, A., and Cohen, N. Implicit regularization in tensor factorization. In *International Conference on Machine Learning (ICML)*, 2021.

Razin, N., Maman, A., and Cohen, N. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. *arXiv preprint arXiv:2201.11729*, 2022.

Song, Q., Ge, H., Caverlee, J., and Hu, X. Tensor completion algorithms in big data analytics. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(1):1–48, 2019.

Zou, D., Wu, J., Braverman, V., Gu, Q., Foster, D. P., and Kakade, S. The benefits of implicit regularization from SGD in least squares problems. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

A. Proofs

We provide here the proofs of Lemma 4.1 and Theorem 3.2. First let us recall that we consider learning a tensor $\mathcal{W}$ which has the following form

$$\mathcal{W} = \sum_{r=1}^{R} \bigotimes_{n=1}^{N} \Big(\prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r} \boldsymbol{w}_r^n \Big), \quad \boldsymbol{A}_i^{n,r} \in \mathbb{R}^{d_n \times d_n}, \boldsymbol{w}_r^n \in \mathbb{R}^{d_n},$$

by minimizing the loss function $\mathcal{L}(\mathcal{W}) = \Phi\Big(\{\boldsymbol{w}_r^n\}_{r=1\ n=1}^{R\ N}, \{\boldsymbol{A}_i^{n,r}\}_{r=1\ n=1\ i=1}^{R\ N\ k_n} \Big)$ using gradient descent. Then with infinitesimally small learning rate and non-zero initialization, we have

$$\frac{d}{dt}\boldsymbol{w}_r^n(t) = -\frac{\partial}{\partial \boldsymbol{w}_r^n} \Phi\Big(\{\boldsymbol{w}_{r'}^{n'}(t)\}_{r'=1\ n'=1}^{R\ N}, \{\boldsymbol{A}_i^{n',r'}(t)\}_{r'=1\ n'=1\ i=1}^{R\ N\ k'_n} \Big),$$

and

$$\frac{d}{dt}\boldsymbol{A}_i^{r,n}(t) = -\frac{\partial}{\partial \boldsymbol{A}_i^{r,n}} \Phi\Big(\{\boldsymbol{w}_{r'}^{n'}(t)\}_{r'=1\ n'=1}^{R\ N}, \{\boldsymbol{A}_i^{n',r'}(t)\}_{r'=1\ n'=1\ i=1}^{R\ N\ k'_n} \Big).$$

To show Lemma 4.1, we will use the following result shown in Razin et al. (2021).

Lemma A.1. *$\forall \mathcal{A} \in \mathbb{R}^{d_1 \times \ldots \times d_N}$ and $\{\boldsymbol{w}^n \in \mathbb{R}^{d_n}\}_{n=1}^N$ where $d_1, \ldots d_N \in \mathbb{N}$, it holds that*

$$\Big\langle \mathcal{A}, \bigotimes_{n'=1}^{N} \boldsymbol{w}^{n'} \Big\rangle = \Big\langle [\mathcal{A}]_{(n)} \cdot \bigodot_{n' \neq n} \boldsymbol{w}^{n'}, \boldsymbol{w}^n \Big\rangle, \ n = 1, \ldots, N$$

where $[\mathcal{A}]_{(n)}$ is matricization of the tensor $\mathcal{A}$ in the mode n, and $\odot$ is the kronecker product.

A.1. Proof of Lemma 4.1

A.1.1. Proof of (i)

We compute $\frac{d}{dt}\|\boldsymbol{A}_i^{n,r}(t)\|^2$. We assume that $\{\boldsymbol{w}_r^n\}_{r=1\ n=1}^{R\ N}$ and $\{\boldsymbol{A}_j^{n',s}\}_{(j,n',s)\neq(i,n,r)}$ are fixed, and consider $\Phi_{i,n,r}(\boldsymbol{A}_i^{n,r}) = \Phi\Big(\{\boldsymbol{w}_r^{n'}\}_{r=1\ n'=1}^{R\ N}, \Big\{\boldsymbol{A}_j^{n',s}\Big\}_{s=1\ n'=1\ j=1}^{R\ N\ k_{n'}} \Big)$.

For $\Delta \in \mathbb{R}^{d_n \times d_n}$, using Taylor approximation we have

$$\begin{aligned}
\Phi_{i,n,r}(\boldsymbol{A}_i^{n,r} + \Delta) &= \mathcal{L}\Big(\mathcal{W} + \bigotimes_{n'=1}^{n-1} \prod_{j=1}^{k_{n'}} \boldsymbol{A}_j^{n',r} \boldsymbol{w}_r^{n'} \otimes \prod_{j=1}^{i-1} \boldsymbol{A}_j^{n,r} \Delta \prod_{j=i+1}^{k_n} \boldsymbol{A}_j^{n,r} \boldsymbol{w}_r^n \otimes \bigotimes_{n'=n+1}^{N} \prod_{j=1}^{k_{n'}} \boldsymbol{A}_j^{n',r} \boldsymbol{w}_r^{n'} \Big) \\
&= \mathcal{L}(\mathcal{W}) + \Big\langle \nabla \mathcal{L}(\mathcal{W}), \bigotimes_{n'=1}^{n-1} \prod_{j=1}^{k_{n'}} \boldsymbol{A}_j^{n',r} \boldsymbol{w}_r^{n'} \otimes \prod_{j=1}^{i-1} \boldsymbol{A}_j^{n,r} \Delta \prod_{j=i+1}^{k_n} \boldsymbol{A}_j^{n,r} \boldsymbol{w}_r^n \otimes \bigotimes_{n'=n+1}^{N} \prod_{j=1}^{k_{n'}} \boldsymbol{A}_j^{n',r} \boldsymbol{w}_r^{n'} \Big\rangle \\
&+ o(\|\Delta\|) \\
&= \mathcal{L}(\mathcal{W}) + \Big\langle [\nabla \mathcal{L}(\mathcal{W})]_{(n)} \cdot \bigodot_{n' \neq n} \prod_{j=1}^{k_{n'}} \boldsymbol{A}_j^{n',r} \boldsymbol{w}_r^{n'}, \prod_{j=1}^{i-1} \boldsymbol{A}_j^{n,r} \Delta \prod_{j=i+1}^{k_n} \boldsymbol{A}_j^{n,r} \boldsymbol{w}_r^n \Big\rangle + o(\|\Delta\|) \\
&= \mathcal{L}(\mathcal{W}) + \Big\langle \Big(\prod_{j=1}^{i-1} \boldsymbol{A}_j^{n,r} \Big)^\top [\nabla \mathcal{L}(\mathcal{W})]_{(n)} \Big(\bigodot_{n' \neq n} \prod_{j=1}^{k_{n'}} \boldsymbol{A}_j^{n',r} \boldsymbol{w}_r^{n'} \Big) \boldsymbol{w}_r^{n\top} \Big(\prod_{j=i+1}^{k_n} \boldsymbol{A}_j^{n,r} \Big)^\top, \Delta \Big\rangle \\
&+ o(\|\Delta\|).
\end{aligned}$$

This implies that

$$\frac{\partial}{\partial \boldsymbol{A}_i^{n,r}} \Phi\Big(\{\boldsymbol{w}_{r'}^{n'}\}_{r'=1\ n'=1}^{R\ N}, \Big\{\boldsymbol{A}_i^{n',r'}\Big\}_{r'=1\ n'=1\ i'=1}^{R\ N\ k'_n} \Big) = \Big(\prod_{j=1}^{i-1} \boldsymbol{A}_j^{n,r} \Big)^\top [\nabla \mathcal{L}(\mathcal{W})]_{(n)} \Big(\bigodot_{n' \neq n} \prod_{j=1}^{k_{n'}} \boldsymbol{A}_j^{n',r} \boldsymbol{w}_r^{n'} \Big) \boldsymbol{w}_r^{n\top} \Big(\prod_{j=i+1}^{k_n} \boldsymbol{A}_j^{n,r} \Big)^\top.$$

Since $\frac{d}{dt}\boldsymbol{A}_i^{n,r}(t) = -\frac{\partial}{\partial \boldsymbol{A}_i^{n,r}}\Phi\Big(\{\boldsymbol{w}_{r'}^{n'}(t)\}_{r'=1\ n'=1}^{R\ N}, \Big\{\boldsymbol{A}_{i'}^{n',r'}(t)\Big\}_{r'=1\ n'=1\ i'=1}^{R\ N\ k_{n'}}\Big)$, we have

$$\frac{d}{dt}\boldsymbol{A}_i^{n,r}(t) = -\Big(\prod_{j=1}^{i-1}\boldsymbol{A}_j^{n,r}(t)\Big)^\top [\nabla\mathcal{L}(\mathcal{W}(t))]_{(n)}\Big(\bigodot_{n'\neq n}\prod_{j=1}^{k_{n'}}\boldsymbol{A}_j^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\Big)\boldsymbol{w}_r^n(t)^\top\Big(\prod_{j=i+1}^{k_n}\boldsymbol{A}_j^{n,r}(t)\Big)^\top. \tag{9}$$

Then

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{A}_i^{n,r}(t)\|^2 &= 2\left\langle \boldsymbol{A}_i^{n,r}(t), \frac{d}{dt}\boldsymbol{A}_i^{n,r}(t)\right\rangle \\
&= -2\left\langle \boldsymbol{A}_i^{n,r}(t), \frac{\partial}{\partial \boldsymbol{A}_i^{n,r}(t)}\Phi\Big(\{\boldsymbol{w}_{r'}^{n'}(t)\}_{r'=1\ n'=1}^{R\ N}, \Big\{\boldsymbol{A}_{i'}^{n',r'}(t)\Big\}_{r'=1\ n'=1\ i'=1}^{R\ N\ k_{n'}}\Big)\right\rangle \\
&= -2\left\langle \boldsymbol{A}_i^{n,r}(t), \Big(\prod_{j=1}^{i-1}\boldsymbol{A}_j^{n,r}(t)\Big)^\top [\nabla\mathcal{L}(\mathcal{W}(t))]_{(n)}\Big(\bigodot_{n'\neq n}\prod_{j=1}^{k_{n'}}\boldsymbol{A}_j^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\Big)\boldsymbol{w}_r^n(t)^\top\Big(\prod_{j=i+1}^{k_n}\boldsymbol{A}_j^{n,r}(t)\Big)^\top\right\rangle \\
&= -2\left\langle \prod_{j=1}^{i-1}\boldsymbol{A}_j^{n,r}(t)\cdot\boldsymbol{A}_i^{n,r}(t)\prod_{j=i+1}^{k_n}\boldsymbol{A}_j^{n,r}(t)\boldsymbol{w}_r^n(t), [\nabla\mathcal{L}(\mathcal{W}(t))]_{(n)}\bigodot_{n'\neq n}\prod_{j=1}^{k_{n'}}\boldsymbol{A}_j^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle \\
&= -2\left\langle \prod_{j=1}^{k_n}\boldsymbol{A}_j^{n,r}(t)\boldsymbol{w}^{n,r}(t), [\nabla\mathcal{L}(\mathcal{W}(t))]_{(n)}\bigodot_{n'\neq n}\prod_{j=1}^{k_{n'}}\boldsymbol{A}_j^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle. \\
&= -2\left\langle \nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{j=1}^{k_{n'}}\boldsymbol{A}_j^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle.
\end{aligned}$$

$\frac{d}{dt}\|\boldsymbol{A}_i^{n,r}(t)\|^2$ is independent of n and i, then $\forall(n,m)\in[\![1,N]\!]^2\ \forall(i,j)\in[\![1,k_n]\!]\times[\![1,k_m]\!]\ \|\boldsymbol{A}_i^{n,r}(t)\|^2$ and $\|\boldsymbol{A}_j^{m,r}(t)\|^2$ have the same derivative, which implies that $\|\boldsymbol{A}_i^{n,r}(t)\|^2 - \|\boldsymbol{A}_j^{m,r}(t)\|^2$ does not vary with time t. Thus

$$\|\boldsymbol{A}_i^{n,r}(t)\|^2 - \|\boldsymbol{A}_j^{m,r}(t)\|^2 = \|\boldsymbol{A}_i^{n,r}(0)\|^2 - \|\boldsymbol{A}_j^{m,r}(0)\|^2.$$

A.1.2. PROOF OF (ii)

We now compute $\frac{d}{dt}\|\boldsymbol{w}_r^n(t)\|^2$. We assume that $\{\boldsymbol{w}_s^{n'}\}_{(n',s)\neq(n,r)}$ and $\{\boldsymbol{A}_j^{n',s}\}_{s=1\ n'=1\ j=1}^{R\ N\ k_{n'}}$ are fixed, and consider $\Phi_r^n(\boldsymbol{w}_r^n) = \Phi\left(\{\boldsymbol{w}_{r'}^{n'}\}_{r'=1\ n'=1}^{R\ N}, \{\boldsymbol{A}_{i'}^{n',r'}\}_{r'=1\ n'=1\ i'=1}^{R\ N\ k_{n'}}\right)$.

For $\Delta\in\mathbb{R}^{d_n}$, using Taylor approximation we have

$$\begin{aligned}
\Phi_r^n(\boldsymbol{w}_r^n+\Delta) &= \mathcal{L}\left(\mathcal{W} + \bigotimes_{n'=1}^{n-1}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}\boldsymbol{w}_r^{n'}\otimes\prod_{i=1}^{k_n}\boldsymbol{A}_i^{n,r}\Delta\otimes\bigotimes_{n'=n+1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}\boldsymbol{w}_r^{n'}\right) \\
&= \mathcal{L}(\mathcal{W}) + \left\langle \nabla\mathcal{L}(\mathcal{W}), \bigotimes_{n'=1}^{n-1}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}\boldsymbol{w}_r^{n'}\otimes\prod_{i=1}^{k_n}\boldsymbol{A}_i^{n,r}\Delta\otimes\bigotimes_{n'=n+1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}\boldsymbol{w}_r^{n'}\right\rangle + o(\|\Delta\|) \\
&= \mathcal{L}(\mathcal{W}) + \left\langle [\nabla\mathcal{L}(\mathcal{W})]_{(n)}\cdot\Big(\bigodot_{n'\neq n}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}\boldsymbol{w}_r^{n'}\Big), \prod_{i=1}^{k_n}\boldsymbol{A}_i^{n,r}\Delta\right\rangle + o(\|\Delta\|) \\
&= \mathcal{L}(\mathcal{W}) + \left\langle \Big(\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n,r}\Big)^\top[\nabla\mathcal{L}(\mathcal{W})]_{(n)}\cdot\Big(\bigodot_{n'\neq n}\prod_{i=1}^{k_n}\boldsymbol{A}_i^{n',r}\boldsymbol{w}_r^{n'}\Big), \Delta\right\rangle + o(\|\Delta\|).
\end{aligned}$$

This implies that $\frac{\partial}{\partial\boldsymbol{w}_r^n}\Phi\left(\{\boldsymbol{w}_{r'}^{n'}\}_{r'=1\ n'=1}^{R\ N}, \{A_{i'}^{n',r'}\}_{r'=1\ n'=1\ i'=1}^{R\ N\ k_{n'}}\right) = \Big(\prod_{i=1}^{k_n}\boldsymbol{A}_i^{n,r}\Big)^\top[\nabla\mathcal{L}(\mathcal{W})]_{(n)}\cdot\Big(\bigodot_{n'\neq n}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}\boldsymbol{w}_r^{n'}\Big)$.

Then

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{w}_r^n(t)\|^2 &= 2\left\langle \boldsymbol{w}_r^n(t), \frac{d}{dt}\boldsymbol{w}_r^n(t)\right\rangle = -2\left\langle \boldsymbol{w}_r^n(t), \frac{\partial}{\partial \boldsymbol{w}_r^n}\Phi\left(\{\boldsymbol{w}_{r'}^{n'}(t)\}_{r'=1}^{R}{}_{n'=1}^{N}, \{A_{i'}^{n',r'}(t)\}_{r'=1}^{R}{}_{n'=1}^{N}{}_{i'=1}^{k_{n'}}\right)\right\rangle \\
&= -2\left\langle \boldsymbol{w}_r^n(t), \Big(\prod_{i=1}^{k_n}\boldsymbol{A}_i^{n,r}(t)\Big)^\top [\nabla\mathcal{L}(\mathcal{W}(t))]_{(n)}\cdot\Big(\bigodot_{n'\neq n}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\Big)\right\rangle \\
&= -2\left\langle \prod_{i=1}^{k_n}\boldsymbol{A}_i^{n,r}(t).\boldsymbol{w}_r^n(t), [\nabla\mathcal{L}(\mathcal{W}(t))]_{(n)}\cdot\Big(\bigodot_{n'\neq n}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\Big)\right\rangle \\
&= 2\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle
\end{aligned}$$

$\frac{d}{dt}\|\boldsymbol{w}_r^n(t)\|^2$ is independent of n then $\forall n, m \in [\![1, N]\!]$, $\|\boldsymbol{w}_r^n(t)\|^2$ and $\|\boldsymbol{w}_r^m(t)\|^2$ have the same derivative, which implies that $\|\boldsymbol{w}_r^n(t)\|^2 - \|\boldsymbol{w}_r^m(t)\|^2$ dos not vary with time t. Thus

$$\|\boldsymbol{w}_r^n(t)\|^2 - \|\boldsymbol{w}_r^m(t)\|^2 = \|\boldsymbol{w}_r^n(0)\|^2 - \|\boldsymbol{w}_r^m(0)\|^2.$$

A.1.3. PROOF OF (iii)

From the proofs of (i) and (ii), we have, $\forall n, m \in [\![1, N]\!]$ and $i \in [\![1, k_n]\!]$,

$$\frac{d}{dt}\|\boldsymbol{w}_r^n(t)\|^2 = \frac{d}{dt}\|\boldsymbol{A}_i^{n,r}(t)\|^2 = 2\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle,$$

which implies that

$$\|\boldsymbol{w}_r^n(t)\|^2 - \|\boldsymbol{A}_i^{n,r}(t)\|^2 = \|\boldsymbol{w}_r^n(0)\|^2 - \|\boldsymbol{A}_i^{n,r}(0)\|^2.$$

A.2. Proof of Theorem 3.2

A.2.1. PROOF OF (i)

First note that

$$\begin{aligned}
\frac{d}{dt}\|\boldsymbol{w}_r^n(t)\| &= \frac{1}{2\|\boldsymbol{w}_r^n(t)\|}\frac{d}{dt}\Big(\|\boldsymbol{w}_r^n(t)\|^2\Big) \\
&= \frac{1}{\|\boldsymbol{w}_r^n(t)\|}\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle.
\end{aligned}$$

We now compute $\frac{d}{dt}\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\|$.

$$\begin{aligned}
&\frac{d}{dt}\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\| = \frac{d}{dt}\Big(\prod_{n=1}^{N}\|\boldsymbol{w}_r^n(t)\|\Big) \\
&= \sum_{n=1}^{N}\frac{d}{dt}\|\boldsymbol{w}_r^n(t)\|\prod_{n'\neq n}\|\boldsymbol{w}_r^{n'}(t)\| \\
&= \sum_{n=1}^{N}\frac{1}{\|\boldsymbol{w}_r^n(t)\|}\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle\prod_{n'\neq n}\|\boldsymbol{w}_r^{n'}(t)\| \\
&= \sum_{n=1}^{N}\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\widehat{\boldsymbol{w}}_r^{n'}(t)\right\rangle\prod_{n'\neq n}\|\boldsymbol{w}_r^{n'}(t)\|^2,
\end{aligned}$$

where $\widehat{\boldsymbol{w}}_r^{n'}(t) = \frac{\boldsymbol{w}_r^{n'}(t)}{\|\boldsymbol{w}_r^{n'}(t)\|}$.

Let $\gamma_r(t) := \left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}} \boldsymbol{A}_i^{n',r}(t)\widehat{\boldsymbol{w}}_r^{n'}(t) \right\rangle$ and assume that $\gamma_r(t) \geq 0$, we have

$$\|\boldsymbol{w}_r^n(t)\|^2 \leq \min_{n'\in[\![1,N]\!]} \|\boldsymbol{w}_r^{n'}(t)\|^2 + \|\boldsymbol{w}_r^n(t)\|^2 - \min_{n'\in[\![1,N]\!]} \|\boldsymbol{w}_r^{n'}(t)\|^2$$

So,

$$\|\boldsymbol{w}_r^n(t)\|^2 \leq \min_{n'\in[\![1,N]\!]} \|\boldsymbol{w}_r^{n'}(t)\|^2 + \varepsilon_1(t) = \Big(\Big(\min_{n'\in[\![1,N]\!]} \|\boldsymbol{w}_r^{n'}(t)\|\Big)^N\Big)^{\frac{2}{N}} + \varepsilon_1(t),$$

where $\varepsilon_1(t) := \max_{\substack{r\in\{1\dots,R\}\\(n,m)\in[\![1,N]\!]^2}} \Big|\|\boldsymbol{w}_r^n(t)\|^2 - \|\boldsymbol{w}_r^m(t)\|^2\Big|$ denotes the unbalancedness magnitude of the weight vectors $\boldsymbol{w}_r^n(t)$.

$$\implies \|\boldsymbol{w}_r^n(t)\|^2 \leq \Big(\prod_{n'=1}^{N} \|\boldsymbol{w}_r^{n'}(t)\|\Big)^{\frac{2}{N}} + \varepsilon_1(t) = \Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^{\frac{2}{N}} + \varepsilon_1(t)$$

$$\overset{\gamma_r(t)\geq 0}{\implies} \prod_{n'\neq n} \|\boldsymbol{w}_r^{n'}(t)\|^2\gamma_r(t) \leq \Big(\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^{\frac{2}{N}} + \varepsilon_1(t)\Big)^{N-1}\gamma_r(t).$$ This gives the following upper bound:

$$\frac{d}{dt}\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\| \leq N\gamma_r(t)\Big(\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^{\frac{2}{N}} + \varepsilon_1(t)\Big)^{N-1}.$$

On the other hand, we can write

$$\frac{d}{dt}\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\| = \Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^2 \sum_{n=1}^{N}\frac{1}{\|\boldsymbol{w}_r^n(t)\|^2}\gamma_r(t).$$

Since,

$$\frac{1}{\|\boldsymbol{w}_r^n(t)\|^2} \geq \frac{1}{\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\|^{\frac{2}{N}} + \varepsilon_1(t)},$$

We obtain the following lower bound:

$$\frac{d}{dt}\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\| \geq N\gamma_r(t)\frac{\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^2}{\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^{\frac{2}{N}} + \varepsilon_1(t)}.$$

Thus

$$N\gamma_r(t)\frac{\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^2}{\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^{\frac{2}{N}} + \varepsilon_1(t)} \leq \frac{d}{dt}\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\| \leq N\gamma_r(t)\Big(\Big\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\Big\|^{\frac{2}{N}} + \varepsilon_1(t)\Big)^{N-1}.$$

When $\gamma_r(t) \leq 0$, this inequality is reversed.

It is easy to see that when $\varepsilon(0)=0$, $\varepsilon_1(0)$ will be also equal to zero. Moreover by Lemma 4.1, $\varepsilon_1(t)$ stays constant over time, and thus

$$\frac{d}{dt}\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\| = N\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\|^{2-\frac{2}{N}}\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\widehat{\boldsymbol{w}}_r^{n'}(t)\right\rangle.$$
$$= N\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\|^{2-\frac{2}{N}}\prod_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\|\boldsymbol{A}_i^{n',r}(t)\|\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\widehat{\boldsymbol{A}}_i^{n',r}(t)\widehat{\boldsymbol{w}}_r^{n'}(t)\right\rangle,$$

where $\widehat{\boldsymbol{A}}_i^{n',r}(t)=\dfrac{\boldsymbol{A}_i^{n',r}(t)}{\|\boldsymbol{A}_i^{n',r}(t)\|}$.

Assuming that $\varepsilon(0)=0$, which implies that $\varepsilon_1(0)=0$, and using Lemma 4.1 we also obtain that

$$\|\boldsymbol{A}_i^{n,r}(t)\| = \|\boldsymbol{w}_r^m(t)\| = \left\|\bigotimes_{m=1}^{N}\boldsymbol{w}_r^m(t)\right\|^{\frac{1}{N}}.$$

Plugging this into the equality just above gives

$$\frac{d}{dt}\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\| = N\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\|^{2-\frac{2}{N}}\prod_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\|^{\frac{1}{N}}\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\widehat{\boldsymbol{A}}_i^{n',r}(t)\widehat{\boldsymbol{w}}_r^{n'}(t)\right\rangle$$
$$= N\left\|\bigotimes_{n=1}^{N}\boldsymbol{w}_r^n(t)\right\|^{2-\frac{2}{N}+\frac{k_1+\ldots+k_N}{N}}\left\langle -\nabla\mathcal{L}(\mathcal{W}(t)), \bigotimes_{n'=1}^{N}\prod_{i=1}^{k_{n'}}\widehat{\boldsymbol{A}}_i^{n',r}(t)\widehat{\boldsymbol{w}}_r^{n'}(t)\right\rangle,$$

which completes the proof.

A.2.2. Proof of (ii)

The proof is based on the following lemma which characterizes the evolution of the singular values of the matrix $\boldsymbol{A}^{n,r}(t):=\prod_{i=1}^{k_n}\boldsymbol{A}_i^{n,r}(t)$. We denote by $(\sigma_l^{n,r}(t))_{1\le l\le d_n}$ the singular values of $\boldsymbol{A}^{n,r}(t)$. The singular value decomposition of $\boldsymbol{A}^{n,r}(t)$ is $\boldsymbol{A}^{n,r}(t)=\boldsymbol{U}^{n,r}(t)\boldsymbol{S}^{n,r}(t)\boldsymbol{V}^{n,r}(t)^\top$, with $\boldsymbol{S}^{n,r}(t)=\mathrm{diag}\{\sigma_l^{n,r}(t), 1\le l\le d_n\}$ and $\boldsymbol{U}^{n,r}(t)\in\mathbb{R}^{d_n\times d_n}$ and $\boldsymbol{V}^{n,r}(t)\in\mathbb{R}^{d_n\times d_n}$ are two orthogonal matrices.

Lemma A.2. *If $\{\boldsymbol{A}_i^{n,r}(0)\}_{i=1}^{k_n}$ are matrices satisfying $\boldsymbol{A}_i^{n,r}(0)^\top\boldsymbol{A}_i^{n,r}(0)=\boldsymbol{A}_{i+1}^{n,r}(0)\boldsymbol{A}_{i+1}^{n,r}(0)^\top$, for all $i\in\{1,\ldots,k_n-1\}$, then the singular values of the product matrix $\boldsymbol{A}^{n,r}(t)$ evolve by:*

$$\frac{d}{dt}\big(\sigma_l^{n,r}(t)\big) = k_n\big(\sigma_l^{n,r}(t)\big)^{2(1-\frac{1}{k_n})}\beta_l^{n,r}(t) \quad , l=1,\ldots,d_n,$$

where $\beta_l^{n,r}(t) = \left\langle -\nabla\mathcal{L}(\boldsymbol{\mathcal{W}}(t)), \bigotimes_{n'=1}^{n-1}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\otimes\boldsymbol{u}_l^{n,r}(t)\boldsymbol{v}_l^{n,r}(t)^\top\boldsymbol{w}_r^n(t)\otimes\bigotimes_{n'=n+1}^{N}\prod_{i=1}^{k_{n'}}\boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right\rangle$,
and $\boldsymbol{u}_l^{n,r}(t)$ and $\boldsymbol{v}_l^{n,r}(t)$ are the l-th columns of $\boldsymbol{U}^{n,r}(t)$ and $\boldsymbol{V}^{n,r}(t)$, respectively.

Proof. First, we use the same arguments of the proof of Theorem 1 in Arora et al. (2018) to prove that:

$$\boldsymbol{A}_i^{n,r}(t)^\top\boldsymbol{A}_i^{n,r}(t)=\boldsymbol{A}_{i+1}^{n,r}(t)\boldsymbol{A}_{i+1}^{n,r}(t)^\top \quad , \forall t\ge 0, \forall i\in\{1,\ldots,k_n-1\}.$$

Indeed, since by (9) we have

$$\frac{d}{dt}\boldsymbol{A}_i^{n,r}(t) = -\Big(\prod_{j=1}^{i-1}\boldsymbol{A}_j^{n,r}(t)\Big)^\top[\nabla\mathcal{L}(\mathcal{W})]_{(n)}\Big(\bigodot_{n'\neq n}\prod_{j=1}^{k_{n'}}\boldsymbol{A}_j^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\Big)\boldsymbol{w}_r^n(t)^\top\Big(\prod_{j=i+1}^{k_n}\boldsymbol{A}_j^{n,r}(t)\Big)^\top,$$

then $\frac{d}{dt}\left(\boldsymbol{A}_i^{n,r}(t)^\top \boldsymbol{A}_i^{n,r}(t)\right) = \frac{d}{dt}\left(\boldsymbol{A}_{i+1}^{n,r}(t)\boldsymbol{A}_{i+1}^{n,r}(t)^\top\right)$. Having that $\boldsymbol{A}_i^{n,r}(0)^\top \boldsymbol{A}_i^{n,r}(0) = \boldsymbol{A}_{i+1}^{n,r}(0)\boldsymbol{A}_{i+1}^{n,r}(0)^\top$, we obtain

$$\boldsymbol{A}_i^{n,r}(t)^\top \boldsymbol{A}_i^{n,r}(t) = \boldsymbol{A}_{i+1}^{n,r}(t)\boldsymbol{A}_{i+1}^{n,r}(t)^\top, \forall t \geq 0. \tag{10}$$

Using the singular value decomposition of $\boldsymbol{A}_i^{n,r}(t)$ and $\boldsymbol{A}_{i+1}^{n,r}(t)$, (10) implies that $\boldsymbol{A}_i^{n,r}(t)$ and $\boldsymbol{A}_{i+1}^{n,r}(t)$ have the same singular values. So, the two matrices can then be written as $\boldsymbol{A}_i^{n,r}(t) = \boldsymbol{U}_i^{n,r}(t)\boldsymbol{\Sigma}^{n,r}(t)\boldsymbol{V}_i^{n,r}(t)^\top$ and $\boldsymbol{A}_{i+1}^{n,r}(t) = \boldsymbol{U}_{i+1}^{n,r}(t)\boldsymbol{\Sigma}^{n,r}(t)\boldsymbol{V}_{i+1}^{n,r}(t)^\top$, where $\boldsymbol{\Sigma}^{n,r}(t) = \operatorname{diag}\left(\lambda_1^{n,r}(t)\boldsymbol{I}_{\alpha_1}, \ldots, \lambda_m^{n,r}(t)\boldsymbol{I}_{\alpha_m}\right)$, where, $\forall s \in \{1, \ldots, m\}$, α_s is the multiplicity of the singular value $\lambda_s^{n,r}(t)$ and $\boldsymbol{I}_{\alpha_s}$ is the $\alpha_s \times \alpha_s$ identity matrix. Moreover, (10) also implies that

$$\boldsymbol{U}_{i+1}^{n,r}(t) = \boldsymbol{V}_i^{n,r}(t)\boldsymbol{O}_i^{n,r}(t),$$

where $\boldsymbol{O}_i^{n,r}(t) = \operatorname{diag}\left(\boldsymbol{O}_{i,1}^{n,r}(t), \ldots, \boldsymbol{O}_{i,m}^{n,r}(t)\right)$ and $\boldsymbol{O}_{i,s}^{n,r}(t) \in \mathbb{R}^{\alpha_s \times \alpha_s}$ is an orthogonal matrix, $\forall s \in \{1, \ldots, m\}$ (Arora et al., 2018, Section A.1).

Using the fact that, $\forall i \in \{1, \ldots, k_n - 1\}$, $\boldsymbol{O}_i^{n,r}(t)$ and $\boldsymbol{\Sigma}^{n,r}(t)$ commute, we obtain that

$$\boldsymbol{A}^{n,r}(t) := \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) = \boldsymbol{U}_1^{n,r}(t) \prod_{i=1}^{k_n - 1} \boldsymbol{O}_i^{n,r}(t)\left(\boldsymbol{\Sigma}^{n,r}(t)\right)^{k_n} \boldsymbol{V}_{k_n}^{n,r}(t)^\top, \tag{11}$$

and

$$\boldsymbol{S}^{n,r}(t) = \left(\boldsymbol{\Sigma}^{n,r}(t)\right)^{k_n}. \tag{12}$$

We now characterize the evolution of the singular values of $\boldsymbol{A}^{n,r}(t)$ using the same arguments as in the proof of Theorem 3 in Arora et al. (2019).

$$\begin{aligned}
\frac{d}{dt}\boldsymbol{A}^{n,r}(t) &= \sum_{j=1}^{k_n} \prod_{i=1}^{j-1} \boldsymbol{A}_i^{n,r}(t)\left(\frac{d}{dt}\boldsymbol{A}_j^{n,r}(t)\right) \prod_{i=j+1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \\
&= -\sum_{j=1}^{k_n} \prod_{i=1}^{j-1} \boldsymbol{A}_i^{n,r}(t)\left(\prod_{i=1}^{j-1} \boldsymbol{A}_i^{n,r}(t)\right)^\top [\nabla \mathcal{L}(\mathcal{W})]_{(n)} \left(\bigodot_{n' \neq n} \prod_{i=1}^{k_{n'}} \boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right)\boldsymbol{w}_r^n(t)^\top \left(\prod_{i=j+1}^{k_n} \boldsymbol{A}_i^{n,r}(t)\right)^\top \prod_{i=j+1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \\
&\overset{(*)}{=} -\sum_{j=1}^{k_n} \left[\boldsymbol{A}^{n,r}(t)\boldsymbol{A}^{n,r}(t)^\top\right]^{\frac{j-1}{k_n}} [\nabla \mathcal{L}(\mathcal{W})]_{(n)} \left(\bigodot_{n' \neq n} \prod_{i=1}^{k_{n'}} \boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right)\boldsymbol{w}_r^n(t)^\top \left[\boldsymbol{A}^{n,r}(t)^\top \boldsymbol{A}^{n,r}(t)\right]^{\frac{k_n - j}{k_n}}.
\end{aligned} \tag{13}$$

(*) follows from (11) as in Arora et al. (2018, Proof of Thm. 1).

Let $\boldsymbol{u}_l^{n,r}(t)$ and $\boldsymbol{v}_l^{n,r}(t)$ be the l-th columns of $\boldsymbol{U}^{n,r}(t)$ and $\boldsymbol{V}^{n,r}(t)$, respectively. Using Eq. 25 in Arora et al. (2019), we have

$$\begin{aligned}
\frac{d}{dt}\sigma_l^{n,r}(t) &= \boldsymbol{u}_l^{n,r}(t)^\top \left[\frac{d}{dt}\boldsymbol{A}^{n,r}(t)\right] \boldsymbol{v}_l^{n,r}(t) \\
&= -k_n \left(\sigma_l^{n,r}(t)\right)^{2\frac{k_n - 1}{k_n}} \boldsymbol{u}_l^{n,r}(t)^\top [\nabla \mathcal{L}(\mathcal{W})]_{(n)} \left(\bigodot_{n' \neq n} \prod_{i=1}^{k_{n'}} \boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right)\boldsymbol{w}_r^n(t)^\top \boldsymbol{v}_l^{n,r}(t) \quad \text{(using Eq. 13 and Eq. 11)} \\
&= -k_n \left(\sigma_l^{n,r}(t)\right)^{2(1-\frac{1}{k_n})} \left\langle \boldsymbol{u}_l^{n,r}(t)\boldsymbol{v}_l^{n,r}(t)^\top \boldsymbol{w}_r^n(t), [\nabla \mathcal{L}(\mathcal{W})]_{(n)} \left(\bigodot_{n' \neq n} \prod_{i=1}^{k_{n'}} \boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t)\right) \right\rangle \\
&\overset{(**)}{=} -k_n \left(\sigma_l^{n,r}(t)\right)^{2(1-\frac{1}{k_n})} \left\langle \nabla \mathcal{L}(\mathcal{W}), \bigotimes_{n'=1}^{n-1} \prod_{i=1}^{k_{n'}} \boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t) \otimes \boldsymbol{u}_l^{n,r}(t)\boldsymbol{v}_l^{n,r}(t)^\top \boldsymbol{w}_r^n(t) \otimes \bigotimes_{n'=n+1}^{N} \prod_{i=1}^{k_{n'}} \boldsymbol{A}_i^{n',r}(t)\boldsymbol{w}_r^{n'}(t) \right\rangle.
\end{aligned}$$

(**) follows from Lemma A.1. □

Proof of Theorem 3.2 (ii) As shown in Lemma A.2, if $\boldsymbol{A}_i^{n,r}(0)^\top \boldsymbol{A}_i^{n,r}(0) = \boldsymbol{A}_{i+1}^{n,r}(0)\boldsymbol{A}_{i+1}^{n,r}(0)^\top$, then all the matrices $\{\boldsymbol{A}_i^{n,r}(t)\}_{i=1}^{k_n}$ have the same singular values, $\forall t \geq 0$. Let $\boldsymbol{\Sigma}^{n,r}(t) := \operatorname{diag}\{\rho_l^{n,r}(t), 1 \leq l \leq d_n\}$ be their diagonal singular value matrix. Using (12), we then have

$$\sigma_l^{n,r}(t) = \left(\rho_l^{n,r}(t)\right)^{k_n}, \tag{14}$$

where $\{\sigma_l^{n,r}(t)\}_{l=1}^{d_n}$ are the singular values of $\boldsymbol{A}^{n,r}(t) := \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t)$.

Using Lemma A.2 above and Lemma 4 in Arora et al. (2019) and since $1 - \frac{1}{k_n} \geq \frac{1}{2}$, we obtain that $\sigma_l^{n,r}(t) = 0, \forall t \geq 0$, if $\sigma_l^{n,r}(0) = 0$. Thus by (14), we have $\rho_l^{n,r}(t) = 0, \forall t \geq 0$, if $\rho_l^{n,r}(0) = 0$. Since $\{\boldsymbol{A}_i^{n,r}(0)\}_{i=1}^{k_n}$ are rank-one matrices, then $\rho_l^{n,r}(0) = 0, \forall l \in \{2, \ldots, d_n\}$. This implies that $\rho_l^{n,r}(t) = 0, \forall l \in \{2, \ldots, d_n\}, \forall t \geq 0$.

$$\begin{aligned}
\left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\|^2 &= \left\| \boldsymbol{U}_1^{n,r}(t) \prod_{i=1}^{k_n-1} \boldsymbol{O}_i^{n,r}(t) \left(\boldsymbol{\Sigma}^{n,r}(t)\right)^{k_n} \boldsymbol{V}_{k_n}^{n,r}(t)^\top \boldsymbol{w}_r^n(t) \right\|^2 \text{ (by Eq. 11)} \\
&= \left\| \left(\boldsymbol{\Sigma}^{n,r}(t)\right)^{k_n} \boldsymbol{V}_{k_n}^{n,r}(t)^\top \boldsymbol{w}_r^n(t) \right\|^2 \\
&= \left(\rho_1^{n,r}(t)\right)^{2k_n} \left\langle \tilde{\boldsymbol{v}}_r^n(t), \boldsymbol{w}_r^n(t) \right\rangle^2,
\end{aligned}$$

with $\tilde{\boldsymbol{v}}_r^n(t)$ is the first column of $\boldsymbol{V}_{k_n}^{n,r}(t)$.

Now, we have

$$\begin{aligned}
\delta_r(t) &:= \left\langle -\nabla \mathcal{L}(\mathcal{W}), \bigotimes_{n=1}^{N} \prod_{i=1}^{k_n} \widehat{\boldsymbol{A}}_i^{n,r}(t) \widehat{\boldsymbol{w}}_r^n(t) \right\rangle \\
&= \left\langle -\nabla \mathcal{L}(\mathcal{W}), \bigotimes_{n=1}^{N} \frac{\prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t)}{\left(\prod_{i=1}^{k_n} \|\boldsymbol{A}_i^{n,r}(t)\|\right) \|\boldsymbol{w}_r^n(t)\|} \right\rangle \\
&= \left\langle -\nabla \mathcal{L}(\mathcal{W}), \bigotimes_{n=1}^{N} \frac{\prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t)}{\left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\|} \right\rangle \prod_{n=1}^{N} \frac{\left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\|}{\left(\prod_{i=1}^{k_n} \|\boldsymbol{A}_i^{n,r}(t)\|\right) \|\boldsymbol{w}_r^n(t)\|}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\prod_{n=1}^{N} \frac{\left\| \prod_{i=1}^{k_n} \boldsymbol{A}_i^{n,r}(t) \boldsymbol{w}_r^n(t) \right\|}{\left(\prod_{i=1}^{k_n} \|\boldsymbol{A}_i^{n,r}(t)\|\right) \|\boldsymbol{w}_r^n(t)\|} &= \prod_{n=1}^{N} \frac{\left(\rho_1^{n,r}(t)\right)^{k_n} |\left\langle \tilde{\boldsymbol{v}}_r^n(t), \boldsymbol{w}_r^n(t) \right\rangle|}{\left(\rho_1^{n,r}(t)\right)^{k_n} \|\boldsymbol{w}_r^n(t)\|} \\
&= \prod_{n=1}^{N} |\left\langle \tilde{\boldsymbol{v}}_r^n(t), \widehat{\boldsymbol{w}}_r^n(t) \right\rangle| \\
&= \left| \left\langle \bigotimes_{n=1}^{N} \tilde{\boldsymbol{v}}_r^n(t), \bigotimes_{n=1}^{N} \widehat{\boldsymbol{w}}_r^n(t) \right\rangle \right|,
\end{aligned}$$

which concludes the proof.

B. An Experiment with Rank-One Matrix Initialization

We conduct the same experiment as in Section 5.5 but with matrices $\boldsymbol{A}_i^{n,r}$ initialized by random rank-one matrices. We used Meteo-UK data set and launched more than 125 learning experiments using various initialization parameters and seeds, for

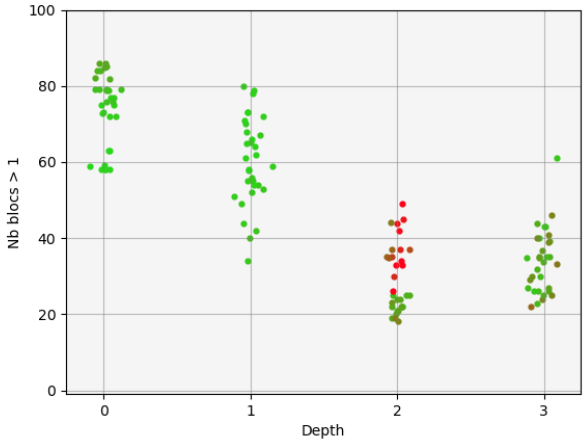


Figure 7. Tensor completion using Meteo-UK data sets with rank-one initialization of the matrix parameters: analysis of the impact of the depth on the rank of the learned tensor. The figure shows for shallow (on the left, depth=0) to deep models (up to depth=5, on the right) the effective rank of the learned tensors for a number of runs that differ from initialization setting. Every single point stands for a learning experiment. Points are plotted with a small random displacement in x and y coordinates to better see point clouds. The color represents the test loss of the model, from green to red respectively for small to large loss

depth ranging from 0 to 5. We report for each experiment the effective CP-rank of the model, i.e. the number of blocks that emerged at convergence (blocks that have a norm greater than 1). The percentage of observed and unobserved inputs in the tensor are 20% and 80%, respectively. Figure 7 shows the same behavior as Figures 4 and 6. For shallower architectures, the impact of initialization is huge and the solutions are mostly of high rank. For deeper architectures, the learned tensor generally has a lower rank.