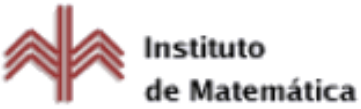

TENSOR DECOMPOSITIONS AND ALGORITHMS, WITH APPLICATIONS TO TENSOR LEARNING

Felipe Bottega Diniz

Tese de Doutorado apresentada ao Programa de Pós-graduação em Matemática, PGMAT, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Matemática.

Orientador: Gregorio Malajovich

Rio de Janeiro
Novembro de 2019

Tensor decompositions and algorithms, with applications to tensor learning

Felipe Bottega Diniz
Orientador: Gregorio Malajovich Muñoz

Tese de doutorado apresentada ao Programa de Pós-Graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de doutor em Matemática.

Aprovada por:

Presidente, Prof. Gregorio Malajovich Muñoz

Prof. Bernardo Freitas Paulo da Costa

Prof. Nick Vannieuwenhoven

Prof. André Lima Ferrer de Almeida

Prof. Amit Bhaya

Rio de Janeiro
Novembro de 2019

Bottega Diniz, Felipe

Tensor decompositions and algorithms, with applications to tensor learning/Felipe Bottega Diniz. – Rio de Janeiro: UFRJ/PGMAT, 2019.

XII, 189 p.: il.; 29, 7cm.

Orientador: Gregorio Malajovich

Tese (doutorado) – UFRJ/PGMAT/Programa de Matemática, 2019.

Bibliography: p. 184 – 189.

1. Tensors. 2. Canonical polyadic decomposition. 3. Multilinear singular value decomposition. 4. Nonlinear least squares. 5. Machine learning. I. Malajovich, Gregorio. II. Universidade Federal do Rio de Janeiro, PGMAT, Programa de Matemática. III. Título.

Resumo

Neste trabalho é apresentada uma nova implementação da canonical polyadic decomposition (CPD). Ela possui uma menor complexidade computacional e menor uso de memória do que as implementações estado da arte disponíveis.

Começamos com alguns exemplos de aplicações da CPD para problemas do mundo real. Um breve resumo das principais contribuições deste trabalho é o seguinte. No capítulo 1, revisamos a álgebra e geometria clássicas de tensores, com foco na CPD. O capítulo 2 é focado na compressão tensorial, que é considerada (neste trabalho) uma das partes mais importantes do algoritmo CPD. No capítulo 3, falamos sobre o método de Gauss-Newton, que é um método de mínimos quadrados não-lineares usado para minimizar funções não-lineares. O capítulo 4 é o mais longo deste trabalho. Neste capítulo, apresentamos o personagem principal desta tese: Tensor Fox. Basicamente, é um pacote tensorial que inclui um CPD solver. Após a introdução do Tensor Fox, realizaremos muitas experiências computacionais comparando esse solver com vários outros. No final deste capítulo, apresentamos a decomposição Tensor Train e mostramos como usá-la para calcular CPDs de ordem superior. Também discutimos alguns detalhes importantes, como regularização, pré-condicionamento, condicionamento, paralelismo, etc. No capítulo 5, consideramos a interseção entre decomposições tensoriais e machine learning. É introduzido um novo modelo, que funciona como uma versão tensorial de redes neurais. Finalmente, no capítulo 6, fazemos as conclusões finais e introduzimos nossas expectativas de trabalho futuro.

Abstract

A new algorithm of the canonical polyadic decomposition (CPD) presented here. It features lower computational complexity and memory usage than the available state of the art implementations.

We begin with some examples of CPD applications to real world problems. A short summary of the main contributions in this work follows. In chapter 1 we review classical tensor algebra and geometry, with focus on the CPD. Chapter 2 focuses on tensor compression, which is considered (in this work) to be one of the most important parts of the CPD algorithm. In chapter 3 we talk about the Gauss-Newton method, which is a nonlinear least squares method used to minimize nonlinear functions. Chapter 4 is the longest one of this thesis. In this chapter we introduce the main character of this thesis: Tensor Fox. Basically it is a tensor package which includes a CPD solver. After introducing Tensor Fox we will conduct lots of computational experiments comparing this solver with several others. At the end of this chapter we introduce the Tensor Train decomposition and show how to use it to compute higher order CPDs. We also discuss some important details such as regularization, preconditioning, conditioning, parallelism, etc. In chapter 5 we consider the intersection between tensor decompositions and machine learning. A novel model is introduced, which works as a tensor version of neural networks. Finally, in chapter 6 we reach the final conclusions and introduce our expectations for future developments.

Acknowledgments

This work is the result of a long journey, which started more than 10 years ago, when I decided to become a mathematician. At first, I was just a graduate student who liked math; however, little by little, its infinite beauty began to reveal itself to me. Today it is a big honor to be a mathematician. I owe many thanks to some wonderful people who helped me in one way or another, and I dedicate this space to them.

First I thank to my mother, Fátima Valéria, who provided me a comfortable and stimulating environment where I could focus on my studies. There is no library comparable to my home and I wouldn't be here if it weren't for her.

All my family were very supporting and believed in me. In particular I thank for my young brother, Bruno, my grandmother Dalba and grandfather Walder, my stepfather Fernando who is now gone (God bless him), and my father Nirvando.

During all my doctorate I was very motivated by my girlfriend, Eliane (Lili). She was very important to me and helped me a lot in difficult times. I admire she as a person and as a mathematician. She helped me to overcome this difficult chapter of my life. She taught me the meaning of love.

My friends suffered a little with my constant absence from social meetings, because I always said that " I have to study ", but they supported and I was very fortunate to have such good friends. Thank you very much for your patience and for believing in me, Jorge, Thiago, Rodolpho, Karina, Júlio, Jose Hugo, Bernardo, Deodato, Yuri and Taynara. You all are the best and know that your friendship means everything to me!

I would like to thank my advisor Gregorio Malajovich. In the past years we developed a healthy relation which turned out to be very productive, I learned a lot, and most importantly, thanks to him I was able to find myself in mathematics. Gregorio was the responsible for showing me the world of tensors. At the beginning I wasn't sure if I wanted to pursue this path, but as I progress it become clear that he knew better than me that this was a good path. Thank you very much for believing in me , Gregorio.

I also want to thank professor Bernardo Freitas, who accompanied my whole journey during the doctorate. He is an amazing person who is always paying attention and challenging you. His valuable observations and advice helped me a lot be where I am now.

Nilson da C. Bernardes, it was privilege to be present at your lectures of Functional Analysis and even more to have you in my qualification exam, also in Functional Analysis. Your way of teaching is inspiring, I hope to be able to teach like this in the future.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Finally, I like to thank Gregorio Malajovich, Bernardo Freitas, Nick Vannieuwenhoven, André de Almeida and Amit Bhaya for being part of my doctoral defense committee.

Contents

List of Figures

List of Tables

Introduction

A vector can be thought as data arranged in an unidimensional fashion, that is, an ordered sequence of numbers, strings, or any other kind of information. In the same way a matrix can be thought as bidimensional data, which also is a sequence of vectors with the same length. Tensors are a natural generalization of this process. One can dispose several matrices of same shape in an ordered sequence, forming a 3D-block of data, see figure 1. This is what is called a *third order tensor*. Analogously, matrices are second order tensors and vectors are first order tensors. It is useful to define scalars as zero order tensors. Recursively, one can define a *L-th order tensor* as an ordered sequence of $(L-1)$-th order tensors of same shape.

Tensors can be defined rigorously as mathematical objects but, for the moment, it will be convenient to think of tensors just as multidimensional arrays of data. Given a field $\mathbb{K}$ and positive integers $I_1, \ldots, I_L$, the set of tensors with shape $I_1 \times \ldots \times I_L$ is defined as

$$\mathbb{K}^{I_1 \times \ldots \times I_L} = \{(t_{i_1 \ldots i_L})_{1 \leq i_1 \leq I_1, \ldots, 1 \leq i_L \leq I_L} |\ t_{i_1 \ldots i_L} \in \mathbb{K}\}.$$

This is an informal definition of a tensor space since it depends on coordinates, although it is a useful way to visualize and think of tensors. Later we will give a new definition more aligned with multilinear algebra. In this work we always use an upper index to indicate a sequence of vectors and a lower index to indicate their coordinates.

Given any vectors $\mathbf{v}^{(1)} \in \mathbb{K}^{I_1}, \ldots, \mathbf{v}^{(1)} \in \mathbb{K}^{I_L}$, define the tensor $\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ by

$$(\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)})_{i_1 \ldots i_L} = v^{(1)}_{i_1} \cdot \ldots \cdot v^{(L)}_{i_L}.$$

Any tensor of this form is called a *rank one tensor*. One can also say the tensor *has rank one*. Notice that in the case of second order tensors (matrices), this definition agrees with the definition of rank one matrices. It is not hard to see that any tensor can be written as a sum of rank one tensors. It is of interest to find the minimum number of rank one terms necessary to construct such a sum, and this minimum number is called the *rank* of the tensor. The decomposition of a tensor as a sum of rank one terms is called a *canonical polyadic decomposition* (CPD). For decades, tensor decompositions have been applied to general multidimensional data with success. Today they are excel in several applications,

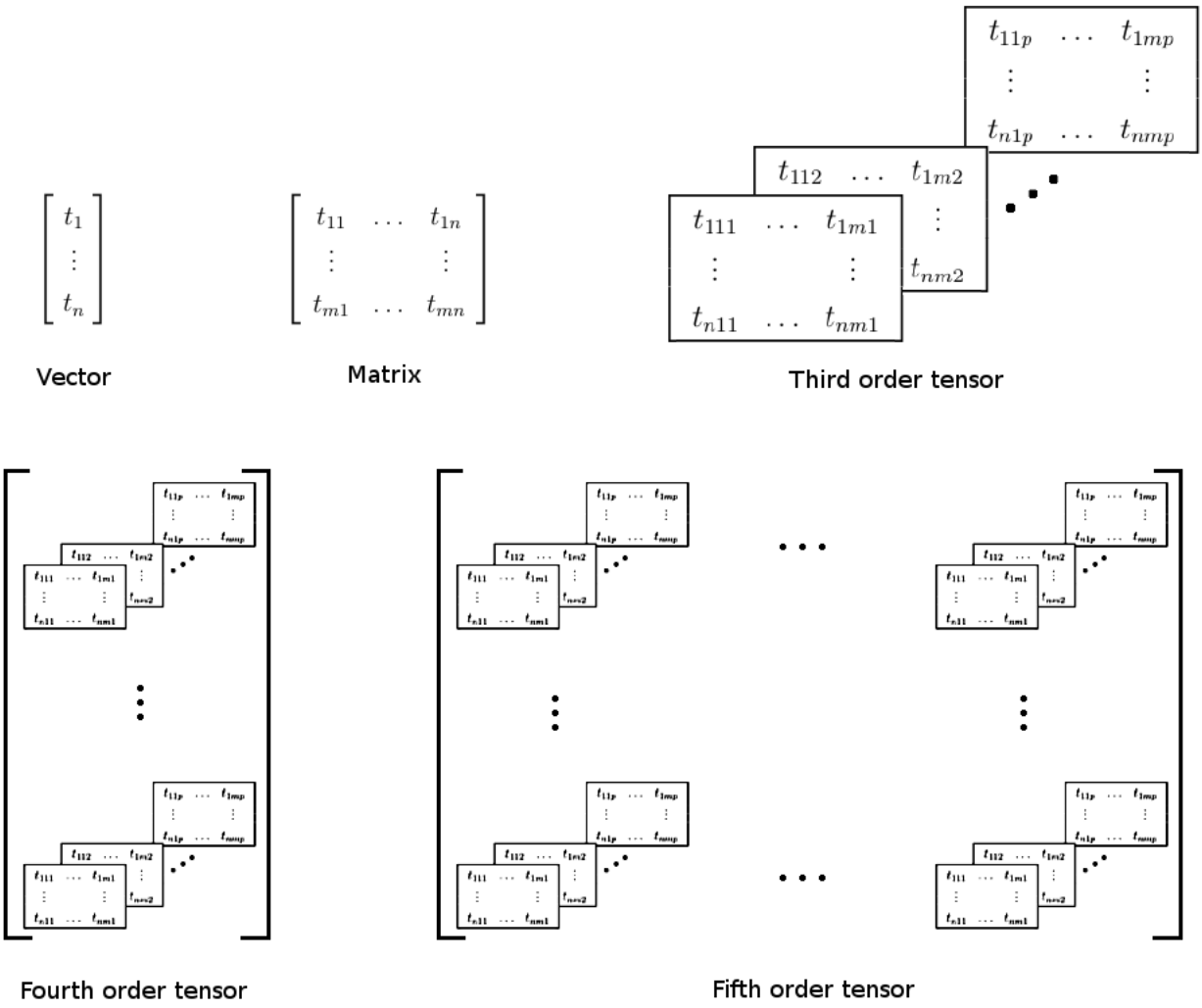

Figure 1: Shapes of tensors for the first five orders.

including blind source separation, dimensionality reduction, pattern/image recognition, machine learning and data mining [2, 3, 5, 7–9]. One of the reasons for the successfulness of tensor decompositions comes from its uniqueness, which occurs for all higher order tensors. This is a desired property which is not found by matrices.

We start giving some motivational examples which highlight the applicability of tensor decompositions, the main topic of this work. In particular, some attention will be given to machine learning applications, a theme to be explored in more details only in the last chapter, when we will have developed the necessary machinery for such.

0.1 First example: Gaussian mixtures

Consider a mixture of K Gaussian distributions with identical covariance matrices. We have lots of data with unknown averages and unknown covariance matrices. The problem at hand is to design an algorithm to *learn* these parameters from the data given. We use

$\mathbb{P}$ to denote probability and $\mathbb{E}$ to denote expectation (which we may also call the *mean* or *average*).

Let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \mathbb{R}^d$ be a set of collected data sample. Let h be a discrete random variable with values in $\{1, 2, \ldots, K\}$ such that $\mathbb{P}[h = i]$ is the probability that a sample $\mathbf{x}$ is a member of the i-th cluster. We denote $w^{(i)} = \mathbb{P}[h = i]$ and $\mathbf{w} = [w^{(1)}, \ldots, w^{(K)}]^T$, the vector of probabilities. Let $\mathbf{u}^{(i)} \in \mathbb{R}^d$ be the mean of the i-th distribution and assume that all distributions have the same covariance matrix $\sigma^2 \mathbf{I}_d$ for $\sigma > 0$. See figure 2 for an illustration of a Gaussian mixture in the case where $d = 2$ and $K = 2$.

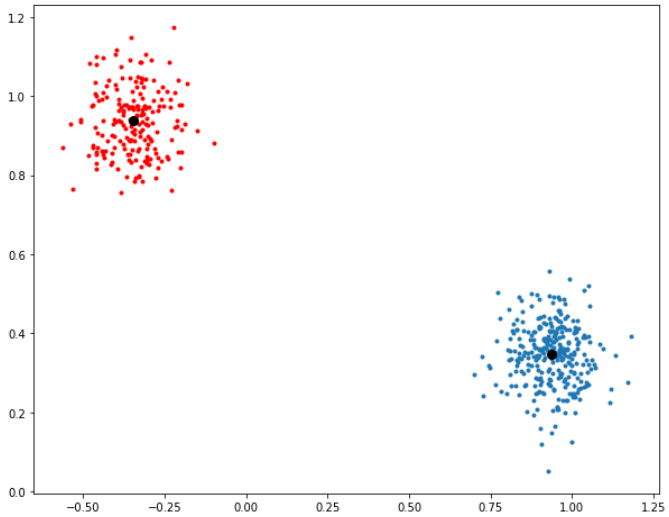

Figure 2: Gaussian mixture in the plane with 2 clusters. The first cluster has mean $\mathbf{u}^{(1)} = [-0.34,\ 0.93]^T$ and the second has mean $\mathbf{u}^{(2)} = [0.93,\ 0.34]^T$. The variance is $\sigma^2 = 0.0059$.

Given a sample point $\mathbf{x}$, note that we can write

$$\mathbf{x} = \mathbf{u}_h + \mathbf{z},$$

where $\mathbf{z}$ is a random vector with mean 0 and covariance $\sigma^2 \mathbf{I}_d$. We summarise the main results in the next theorem whose proof can be found in [9].

Theorem 0.1.1 (Hsu and Kakade, 2013)**.** *Assume $d \geq K$. The variance σ^2 is the smallest eigenvalue of the covariance matrix $\mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] - \mathbb{E}[\boldsymbol{x}] \otimes \mathbb{E}[\boldsymbol{x}]$. Furthermore, if*

$$\begin{aligned}
M_1 &= \mathbb{E}[\boldsymbol{x}], \\
M_2 &= \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x}] - \sigma^2 \boldsymbol{I}_d, \\
M_3 &= \mathbb{E}[\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}] - \sigma^2 \sum_{i=1}^{d} \left(\mathbb{E}[\boldsymbol{x}] \otimes \boldsymbol{e}_i \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \mathbb{E}[\boldsymbol{x}] \otimes \boldsymbol{e}_i + \boldsymbol{e}_i \otimes \boldsymbol{e}_i \otimes \mathbb{E}[\boldsymbol{x}] \right),
\end{aligned}$$

then

$$M_1 = \sum_{i=1}^{K} w^{(i)} \ \boldsymbol{u}^{(i)},$$
$$M_2 = \sum_{i=1}^{K} w^{(i)} \ \boldsymbol{u}^{(i)} \otimes \boldsymbol{u}^{(i)},$$
$$M_3 = \sum_{i=1}^{K} w^{(i)} \ \boldsymbol{u}^{(i)} \otimes \boldsymbol{u}^{(i)} \otimes \boldsymbol{u}^{(i)}.$$

Theorem 0.1.1 allows us to use the method of moments, which is a classical parameter estimation technique from statistics. This method consists in computing certain statistics of the data (often empirical moments) and use it to find model parameters that give rise to (nearly) the same corresponding population quantities. Now suppose that N is large enough so we have a reasonable number of sample points to make useful statistics. First we compute the empirical mean

$$\hat{\mu} := \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}^{(j)} \approx \mathbb{E}[\mathbf{x}]. \tag{1}$$

Now use this result to compute the empirical covariance matrix

$$\hat{\mathbf{S}} := \frac{1}{N} \sum_{j=1}^{N} (\mathbf{x}^{(j)} \otimes \mathbf{x}^{(j)} - \hat{\mu} \otimes \hat{\mu}) \approx \mathbb{E}[\mathbf{x} \otimes \mathbf{x}] - \mathbb{E}[\mathbf{x}] \otimes \mathbb{E}[\mathbf{x}]. \tag{2}$$

The smallest eigenvalue of $\hat{\mathbf{S}}$ is the empirical variance $\hat{\sigma}^2 \approx \sigma^2$. Now we compute the empirical third moment (empirical skewness)

$$\hat{\mathcal{S}} := \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}^{(j)} \otimes \mathbf{x}^{(j)} \otimes \mathbf{x}^{(j)} \approx \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}] \tag{3}$$

and use it to get the empirical value of M_3,

$$\hat{\mathcal{M}}_3 := \hat{\mathcal{S}} - \hat{\sigma}^2 \sum_{i=1}^{d} (\hat{\mu} \otimes \mathbf{e}_i \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \hat{\mu} \otimes \mathbf{e}_i + \mathbf{e}_i \otimes \mathbf{e}_i \otimes \hat{\mu}) \approx M_3. \tag{4}$$

By theorem 0.1.1, $M_3 = \sum_{i=1}^{K} w^{(i)} \ \mathbf{u}^{(i)} \otimes \mathbf{u}^{(i)} \otimes \mathbf{u}^{(i)}$, which is a symmetric tensor containing all parameter information we want to find. The idea is, after computing a symmetric CPD for $\hat{\mathcal{M}}_3$, normalize the factors so each vector has unit norm. By doing this we have a tensor of the form

$$\sum_{i=1}^{K} \hat{w}^{(i)} \ \hat{\mathbf{u}}^{(i)} \otimes \hat{\mathbf{u}}^{(i)} \otimes \hat{\mathbf{u}}^{(i)}$$

as a candidate to solution. Note that it is easy to make all $\hat{w}^{(i)}$ positive. If some of them is negative, just multiply it by -1 and multiply one of the associated vectors also by -1. The final tensor is unchanged but all $\hat{w}^{(i)}$ now are positive. For more on this subject we recommend reading [9].

0.2 Second example: Topic models

Consider a set of documents (texts) together with a set of K possible topics, this structured set of documents is called a *corpus*. Each topic can be represented by a number $1 \leq j \leq K$. Additionally, consider that this corpus has d distinct words in its vocabulary and that each document has $L \geq 3$ words. The words are labelled as numbers between 1 and d.

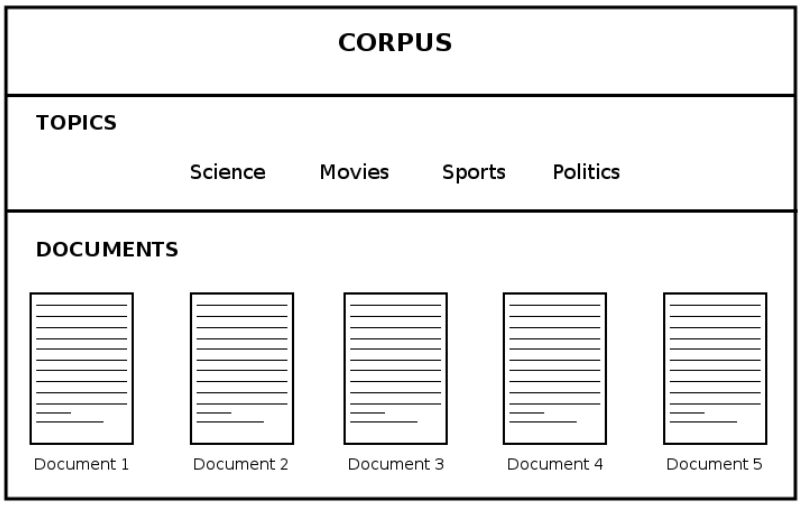

Figure 3: Example of a corpus structure.

The *bag-of-words* model [87, 88] is a system of representation used in natural language processing (NLP). In this model, any text is represented as the multiset of its words, disregarding grammar and word ordering but keeping multiplicity. This multiset is the "bag" containing the words. We consider the bag-of-words model in the problem of document classification, that is, given any text we want a method to classify it into some topic in a prescribed list of topics.

Now suppose that the topics of the documents follows a (discrete) probability distribution such that $\mathbb{P}[h = j] = w^{(j)}$ is the probability that a random document belongs to topic j (the random variable h is a latent variable, responsible for the topics assignment). Let $\mathbf{w} = [w^{(1)}, \ldots, w^{(K)}]^T$ be the vector of probabilities of the topics. Given the topic h and a random document, the words are assumed to follow a probability distribution such that $\mathbb{P}[\mathbf{x} = i | h = j] = u_j^{(i)}$. In other words, given the topic $h = j$, $u_j^{(i)}$ is the probability that a random word $\mathbf{x}$ drawn in a document is the word i. Denote $\mathbf{u}_j = [u_j^{(1)}, \ldots, u_j^{(L)}]^T$ for the vector of probabilities of the words in a document, given the topic $h = j$. Addi-

tionally, suppose that randomly drawing sample points from this distribution generates i.i.d. (independent and identically distributed) random variables.

As first observation, we have that $\sum_{j=1}^{K} w_j = 1$ and $\sum_{i=1}^{L} u_j^{(i)} = 1$ for any topic j. We also will convert the words into vectors, that is, each word i is represented by the canonical basis vector $\mathbf{e}_i \in \mathbb{R}^d$. One advantage of this encoding is that the moments of these random vectors correspond to probabilities over words. Consider a document with words $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(L)}$, then we have that

$$\mathbb{E}[\mathbf{x}^{(i')} \otimes \mathbf{x}^{(j')}] = \sum_{i,j=1}^{d} \mathbb{P}[\mathbf{x}^{(i')} = \mathbf{e}^{(i)},\ \mathbf{x}^{(j')} = \mathbf{e}^{(j)}]\ \mathbf{e}^{(i)} \otimes \mathbf{e}^{(j)} =$$

$$= \sum_{i,j=1}^{d} \mathbb{P}[i'\text{-th word} = i,\ j'\text{-th word} = j]\ \mathbf{e}^{(i)} \otimes \mathbf{e}^{(j)} =$$

$$= \begin{bmatrix} \mathbb{P}[i'\text{-th word} = 1,\ j'\text{-th word} = 1] & \ldots & \mathbb{P}[i'\text{-th word} = 1,\ j'\text{-th word} = d] \\ \vdots & & \vdots \\ \mathbb{P}[i'\text{-th word} = d,\ j'\text{-th word} = 1] & \ldots & \mathbb{P}[i'\text{-th word} = d,\ j'\text{-th word} = d] \end{bmatrix}.$$

More generally, the entry $(i_1, i_2, \ldots, i_L)$ of the tensor $\mathbb{E}[\mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \ldots \otimes \mathbf{x}^{(L)}]$ is

$$\mathbb{P}[\text{first word} = i_1,\ \text{second word} = i_2, \ldots,\ L\text{-th word} = i_L].$$

We also remark that the conditional expectation of $\mathbf{x}^{(i')}$ given h is simply $\mathbf{u}^{(j)}$. More precisely,

$$\mathbb{E}[\mathbf{x}^{(i')}|h = j] = \sum_{i=1}^{d} \mathbb{P}[i'\text{-th word} = i|h = j]\mathbf{e}^{(i)} = \sum_{i=1}^{d} u_i^{(j)}\mathbf{e}^{(i)} = \mathbf{u}^{(j)}.$$

Since the words are conditionally independent given the topic, we can use this property with conditional moments. More precisely, we have that

$$\mathbb{E}[\mathbf{x}^{(i')} \otimes \mathbf{x}^{(j')}|h = j] = \mathbb{E}[\mathbf{x}^{(i')}|h = j] \otimes \mathbb{E}[\mathbf{x}^{(j')}|h = j] = \mathbf{u}^{(j)} \otimes \mathbf{u}^{(j)}.$$

With similar calculations we obtain the following theorem [9, 10].

Theorem 0.2.1 (Anandkumar et al., 2012)**.** *Let* $1 \leq i', j', k' \leq L$. *If*

$$M_1 = \mathbb{E}[\boldsymbol{x}^{(i')}]$$
$$M_2 = \mathbb{E}[\boldsymbol{x}^{(i')} \otimes \boldsymbol{x}^{(j')}]$$
$$M_3 = \mathbb{E}[\boldsymbol{x}^{(i')} \otimes \boldsymbol{x}^{(j')} \otimes \boldsymbol{x}^{(k')}],$$

then

$$M_1 = \sum_{i=1}^{K} w^{(i)} \ \boldsymbol{u}^{(i)}$$
$$M_2 = \sum_{i=1}^{K} w^{(i)} \ \boldsymbol{u}^{(i)} \otimes \boldsymbol{u}^{(i)}$$
$$M_3 = \sum_{i=1}^{K} w^{(i)} \ \boldsymbol{u}^{(i)} \otimes \boldsymbol{u}^{(i)} \otimes \boldsymbol{u}^{(i)}.$$

We may estimate these moments using the actual data at hand. By using any method to compute a CPD for M_3 we get estimates for the latent variables $w^{(i)}$ and $\mathbf{u}^{(i)}$. In many aspects this model resembles the Gaussian mixture model, both use the method of moments to construct a third order tensor which we try to approximate with the CPD.

0.3 Third example: Approximation of functions

Consider a multivariate function $\varphi : \mathbb{R}^L \to \mathbb{R}$ which are difficult to handle analytically, but evaluating φ is feasible. Hence, it is possible to take numerical data to study this function. One may try to consider evaluating $\varphi(x^{(1)}, \ldots, x^{(L)})$ in a closed grid of points such that each direction is partitioned in I_ℓ parts. This mean we have the points $x_1^{(\ell)}, \ldots, x_{I_\ell}^{(\ell)}$ to consider in the direction of the ℓ-th coordinate. Overall we will compute $\varphi(x_{i_1}^{(1)}, \ldots, x_{i_L}^{(L)})$ for all $i_1 = 1 \ldots I_1, \ \ldots, \ i_L = 1 \ldots I_L$.

The results of the computation can be stored in a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_L}$ such that $t_{i_1 \ldots i_L} = \varphi(x_{i_1}^{(1)}, \ldots, x_{i_L}^{(L)})$. Although this is possible, note that as L increases, the *curse of dimensionality*[1] becomes apparent so storing the results this way requires too much memory. For instance, if $L = 50$ and the dimensions are small, $I_1 = I_2 = \ldots = I_{50} = 2$, storing $\mathcal{T}$ would require 9 petabytes. To overcome this problem we must to store $\mathcal{T}$ in a more economic form, and this is possible with a low rank CPD approximation. Assume that φ is *separable*, in the sense that there are functions $\varphi_r^{(\ell)} : \mathbb{R}^{I_\ell} \to \mathbb{R}$, for $\ell = 1 \ldots L$ and $r = 1 \ldots R$, such that

$$\varphi(x_{i_1}^{(1)}, \ldots, x_{i_L}^{(L)}) = \sum_{r=1}^{R} \varphi_r^{(1)}(x_{i_1}^{(1)}) \cdot \varphi_r^{(2)}(x_{i_2}^{(2)}) \cdot \ldots \cdot \varphi_r^{(1)}(x_{i_L}^{(L)}).$$

Define $\mathbf{W}^{(\ell)} = \left[\mathbf{w}_1^{(\ell)}, \ldots, \mathbf{w}_R^{(L)}\right] \in \mathbb{R}^{I_\ell \times R}$, where $\mathbf{w}_r^{(\ell)} = \left[\varphi_r^{(\ell)}(x_1^{(\ell)}), \ldots, \varphi_r^{(\ell)}(x_{I_\ell}^{(\ell)})\right]^T \in \mathbb{R}^{I_\ell}$ is the r-th column of $\mathbf{W}^{(\ell)}$. Then we have the equality $\mathcal{T} = \sum_{r=1}^{R} \mathbf{w}_r^{(1)} \otimes \ldots \otimes \mathbf{w}_r^{(L)}$, which

[1] https://en.wikipedia.org/wiki/Curse_of_dimensionality

come from a CPD for $\mathcal{T}$. Storing this CPD costs $R\sum_{\ell=1}^{L} I_\ell$ floats, which is much better than $\prod_{\ell=1}^{L} I_\ell$ floats necessary to store the tensor in coordinate-wise format. The formulation of this problem and the proposed approach to solve it is based on [80, 81]. Note that this approach still works if φ is not exactly separable but can be well approximated by a separable function. In fact this is what usually happens, one wants to approximate a function of L variables by a finite sum of products of functions of one variable.

0.4 Results

In order to obtain a CPD for the tensor it is clear that one would want to know its rank in the first place. Unfortunately this problem is known to be NP-hard [21]. Usually one already have some prior knowledge of the problem or, in the worst case, one have to compute several CPDs for different ranks to find the best fit. This second approach seems reasonable, however it is of limited use due to the *border rank* phenomenon, which will be further discussed.

In this work we always suppose the rank is known in advance, or at least a decent estimate for the rank is known. In this case all we have to worry is with the computation of the CPD. In the past years several algorithms were proposed and implemented [4, 13–18, 41, 42] so, today, we have a better understanding about how each approach performs. In particular, Gauss-Newton algorithms are proven to have better convergence properties, also verified experimentally. Algorithms based on this approach usually were much slower [70, 71], but this is not the reality today. Exploiting the structure of the approximated Hessian matrix lead to algorithms competitive in terms of speed, and with better accuracy [4, 15]. Following this path, in chapter 3 we exploit this structure towards the goal to speeding up conjugate gradient iterations of subproblems faced at each iteration of the Gauss-Newton algorithm.

While researching about previous implementations for the CPD, I tried to spot the parts overlooked by others. Aspects as damping parameter (when there is regularization), number of conjugate gradient iterations, compression - preprocessing, etc, does not always receives the due attention. With this in mind I designed a new program taking all these little details in account. The result is a tensor package called *Tensor Fox*.[2] In chapter 4 we give a detailed description of this package with respect to the CPD computation.

Let $\mathcal{T}$ be an L-order tensor with shape $\underbrace{n \times n \times \ldots \times n}_{L \text{ times}}$. Below we show the cost per iteration (in *flops* - floating point operations) of state of art implementations and Tensor Fox, to compute a rank-R CPD for $\mathcal{T}$. The constants c_1, c_2 are positive integers with

[2] This package is free for download at https://github.com/felipebottega/Tensor-Fox.

Package	Algorithm	Computational cost
Tensorlab	Gauss-Newton	$\mathcal{O}\left(2(L+1)Rn^L + c_1(\frac{5}{2}L^2R^2 + 8LR^2n + \frac{1}{3}LR^3)\right)$
Tensor Toolbox	Gradient-based optimization	$\mathcal{O}\left(LRn^L\right)$
Tensorly	Alternating Least Squares	$\mathcal{O}\left(LRn^L\right)$
Tensor Box	Gauss-Newton	$\mathcal{O}\left(L^3R^2 + LR^3 + LRn^L + L^3n + L^3R^6\right)$
Tensor Fox	Gauss-Newton	$\mathcal{O}\left((LR + L - 1)n^L + LR^2(1+n) + 3LRn + c_2(9LR + L^2R^2)n\right)$

Table 1: Cost per iteration of several CPD solvers.

little influence on the costs. At first sight it seems that Tensor Toolbox and Tensorly are better, since their cost per iteration is cheaper. In fact it is the opposite, the other solvers indeed make slower iterations, but their iterations have more quality, which leads to faster convergence. Alternating Least squares, for instance, can take thousands of iterations to converge, whereas a Gauss-Newton based algorithm may finish within less than a hundred iterations. The quality of the steps counts. Furthermore, with the exception of Tensor Fox, all the other Gauss-Newton based solvers are costly in the rank. Tensorlab has a factor of R^3 and Tensor Box has a factor of R^6, whereas Tensor Fox is quadratic on R. Computational experiments reinforces these observations.

The computational complexity to compute CPDs makes it hard to aim at really big tensors, because of the factor n^L present in all algorithms. It is the curse of dimensionality in action. In the era of Big Data it is not enough to just have good tensor models, they also need to be computable within a reasonable time. In chapter 4, section 4.6, we show that it is enough to use the Gauss-Newton approach only for third order tensors. With the ideas of [82] we are able to compute higher order CPDs while avoiding the curse of dimensionality. The *tensor train decomposition* (TTD) was recently linked to the CPD, providing a way to compute higher order CPDs much faster than any previous algorithm. These new ideas are implemented in Tensor Fox, which leads to a cost of

$$\mathcal{O}\left((3R+2)n^3 + 3R^2(1+n) + 9Rn + 9c_2(3R+R^2)n\right)$$

flops per iteration for higher order tensors. Of course this is not all. This is the cost per iteration of one third order CPD to be computed, between L of them. Additionally, before the computation of these third order CPDs we have to compute $L-$ SVDs, which adds a cost of

$$\mathcal{O}\left(2\left(n^{L+1} + R^2\frac{n^L - n^2}{n-1} + (1+R^3)n^3\right)\right) \text{ flops.}$$

Therefore we didn't avoid completely the curse of dimensionality. On the other hand, we remark that this cost has a low constant and it is added only once to the overall cost, while the costs of the previous showed algorithms are added at each iteration. The tensor train approach performed much better than any other algorithm in our tests.

We remark that only Tensorlab performs tensor compression before the iterations. In chapter 2 we take a closer look at tensor compression and show why this is crucial to alleviate the curse of dimensionality too. At the moment, most solvers consider compression as an optional action to take, but this should be default. For example, if we have a tensor of shape $n \times n \times n$ and want to compute an approximate CPD of rank $R \ll n$, then it is possible to compress it to a tensor of shape $R \times R \times R$ and use this one to find the CPD. There is virtually no loss in precision and the cost of doing that is of $\mathcal{O}\left(\sum_{\ell=1}^{3} \min\{R, n\} \prod_{\ell=1}^{3} n\right) = \mathcal{O}\left(3Rn^3\right)$ flops. Compare this to the previous costs, where we have something of $\mathcal{O}\left(3Rn^3\right)$ flops **at each iteration**. We try to stress this point here with the known results of the area and computational experiments.

Tensor decompositions are amazing tools to model multidimensional data, and that is why developing new algorithms to compute the CPD is necessary in this area. This work is an attempt to improve the state of art overall performance. The main contributions of this work are the following:

- The diagonal regularization introduced at 4.8 reduces substantially the condition number compared with the other regularization approaches used in the literature.
- The approximated Hessian of the problem has a block structure which is exploited in theorem 3.5.10 to accelerate any algorithm based on Krylov methods to solve the normal equations of the Gauss-Newton step.
- Tensor Fox performs specially better for higher order tensors. This is possible with the CPD Tensor Train technique developed at [82] and implemented in Tensor Fox.
- A development of a new algorithm combining the best parts of several state of art algorithms was implemented in tensor Fox. Improvements includes a method to exploit the approximated Hessian.
- A new tensor package software, Tensor Fox, which is competitive and freely available to the interested practitioners and researches. I remark that all routines of Tensor Fox were written from scratch, that is, not a single part of other tensor package was copied. Routines are optimized for speed.
- Several benchmarks are introduced as an attempt to obtain a fair comparison between all the state of art CPD solvers for a range of distinct problems. Usually the papers makes comparisons between the one they are introducing and just one or two outside solvers. This seems to be the first time such a broad comparison is made.
- New tensor models for machine learning problems are introduced and their potential is experimentally validated.

Chapter 1

Basic notions

We introduce the necessary notations and preliminary results of multilinear algebra. After that we will talk about tensor products, decompositions and the geometry of tensor spaces. In this chapter we also formalize the main challenge of this work, which is to compute the canonical polyadic decomposition.

1.1 Notations

Scalars will be denoted by lower case letters, including greek letters, e.g., a or λ. Sometimes we can use capital letters for natural numbers, e.g., L or C. Vectors are denoted by bold lower case letters, e.g., $\mathbf{x}$. Matrices are denoted by bold capital letters, e.g., $\mathbf{X}$. Tensors are denoted by calligraphic capital letters, e.g., $\mathcal{T}$. Capital greek letters will be more flexible, appearing sometimes as matrices, sometimes as tensors, and sometimes as sets. The i-th entry of a vector $\mathbf{x}$ is denoted by x_i, the entry (i, j) of a matrix $\mathbf{X}$ is denoted by x_{ij}, and the entry $(i_1, i_2, \ldots, i_L)$ of a tensor $\mathcal{T}$ with L indexes is denoted by $t_{i_1 i_2 \ldots i_L}$. Sometimes it will be necessary to denote the entry (i, j) of $\mathbf{X}$ by $(\mathbf{X})_{ij}$, and the same may happen to a tensor. Any kind of sequence will be indicated by superscripts. For example, we write $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ for a sequence of vectors. The $n \times n$ identity matrix will be denoted by $\mathbf{I}_n$.

In the case we have a function f with n scalar arguments, we denote them by $x_1, x_2, \ldots, x_n$ and write $f(x_1, x_2, \ldots, x_n)$. If there is just three arguments we prefer the classical $f(x, y, z)$, and similar considerations for two or just one argument, where we will use $f(x, y)$, and $f(x)$, respectively.

Vector spaces, groups and subsets in general will be denoted by blackboard bold capital letters or just capital letters, e.g., $\mathbb{V}$ or S. An important particular case is of a field, which will be denoted by $\mathbb{K}$. However, this work is limited to the cases $\mathbb{K} = \mathbb{R}$ (real numbers) and $\mathbb{K} = \mathbb{C}$ (complex numbers). In this work it will convenient to define the set of natural numbers as being the set $\mathbb{N} = \{1, 2, 3, \ldots\}$. The symbols $\mathbb{P}$ and $\mathbb{E}$ are reserved for probability and expectation, respectively. Every time we introduce a basis, we will

be using the set notation with the implicit understanding it is an ordered set. Given two tuples $(r_1, \ldots, r_L), (R_1, \ldots, R_L)$, we write $(r_1, \ldots, r_L) < (R_1, \ldots, R_L)$ if $r_i < R_i$ for all $i = 1 \ldots L$.

We also adopt the Matlab notational style when it is desired to take slices[1] or to fix a subset of indexes while varying the others. For example, if $\mathbf{X}$ is a $m \times n$ matrix, then $X_{i:}$ is the i-th row of $\mathbf{X}$, while $\mathbf{X}_{:j}$ is its j-th column.

The symbol ^{T} denotes the transpose of a vector or matrix, * denotes the conjugate transpose of a vector or matrix, $\dagger$ denotes the pseudoinverse of a matrix. If we use * for a vector space, then it means the dual space. For example, if $\mathbb{V}$ is a vector space, then $\mathbb{V}^*$ is the dual space of $\mathbb{V}$. The symbol $\cong$ will be used to denote isomorphism between vector spaces. Let $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be a basis of $\mathbb{V}$ and let $\mathbf{x} = x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n, \mathbf{y} = y_1\mathbf{e}_1 + \ldots y_n\mathbf{e}_n$ be two vectors in this space. The *Hermitian inner product* between $\mathbf{x}$ and $\mathbf{y}$ is defined by $\langle \mathbf{x}, \mathbf{y} \rangle = x_1\overline{y_1} + \ldots + x_n\overline{y_n}$. We also adopt the definition $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}} = x_1y_1 + \ldots + x_ny_n$ and call it the *Euclidean inner product*.

Finally, consider the Euclidean vector space $\mathbb{K}^n$ with basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and dual basis $\{\mathbf{f}_1^*, \ldots, \mathbf{f}_n^*\}$, where $\mathbf{f}_i \in \mathbb{K}^n$ for each $i = 1 \ldots n$. In this context, for any vector $\mathbf{x} \in \mathbb{K}^n$ we define $\mathbf{f}_i^*(\mathbf{x}) = \mathbf{f}_i^* \cdot \mathbf{x} = \langle \mathbf{x}, \mathbf{f}_i \rangle$. If the basis is orthonormal, then $\mathbf{f}_i^* = \mathbf{e}_i^*$.

1.2 Multilinear maps

Let $\mathbb{V}^{(1)}, \ldots, \mathbb{V}^{(L)}, \mathbb{V}, \mathbb{W}$ be vector spaces over the same field $\mathbb{K}$ such that $\dim(\mathbb{V}^{(\ell)}) = I_\ell$ for each $\ell = 1 \ldots L$, and $\dim(\mathbb{W}) = J$. A map $\mathcal{T} : \mathbb{V}^{(1)} \times \ldots \times \mathbb{V}^{(L)} \to \mathbb{W}$ is said to be *L-linear* if $\mathcal{T}$ is linear in each coordinate. More precisely, for all $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)}) \in \mathbb{V}^{(1)} \times \ldots \times \mathbb{V}^{(L)}$ and all $\alpha, \beta \in \mathbb{K}$ we have that

$$\mathcal{T}(\mathbf{x}^{(1)}, \ldots, \alpha\mathbf{x}^{(i)} + \beta\mathbf{x}^{(i+1)}, \ldots, \mathbf{x}^{(L)}) =$$

$$= \alpha\mathcal{T}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)}, \ldots, \mathbf{x}^{(L)}) + \beta\mathcal{T}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i+1)}, \ldots, \mathbf{x}^{(L)}).$$

Sometimes it is not relevant to mention the value L and one can just say that $\mathcal{T}$ is a *multilinear map*. Notice that for $L = 1$, $\mathcal{T}$ is just a classical linear map. We denote the space of L-linear maps $\mathbb{V}^{(1)} \times \ldots \times \mathbb{V}^{(L)} \to \mathbb{W}$ by $\mathcal{L}(\mathbb{V}^{(1)}, \ldots, \mathbb{V}^{(L)}; \mathbb{W})$. Some common abbreviations are $\mathcal{L}_L(\mathbb{V}; \mathbb{W}) = \mathcal{L}(\underbrace{\mathbb{V}, \ldots, \mathbb{V}}_{L \text{ times}}; \mathbb{W})$ and $\mathcal{L}(\mathbb{V}) = \mathcal{L}(\mathbb{V}; \mathbb{V})$.

Lemma 1.2.1. *Let $(i_1, \ldots, i_p, i'_1, \ldots, i'_{L-p})$ be any permutation of $(1, \ldots, L)$. Then*

$$\mathcal{L}(\mathbb{V}^{(1)}, \ldots, \mathbb{V}^{(L)}; \mathbb{W}) \cong \mathcal{L}\left(\mathbb{V}^{(i_1)}, \ldots, \mathbb{V}^{(i_p)}; \mathcal{L}(\mathbb{V}^{(j_1)}, \ldots, \mathbb{V}^{(j_{L-p})}; \mathbb{W})\right).$$

[1] Since the author uses much more Python/Numpy than Matlab there is a chance to appear some slight differences.

Corollary 1.2.2. $\mathcal{L}(\mathbb{V}^{(1)},\dots,\mathbb{V}^{(L)};\mathbb{W}) \cong \mathcal{L}(\mathbb{V}^{(1)},\dots,\mathbb{V}^{(L)},\mathbb{W}^*;\mathbb{K})$.

This corollary is a direct consequence of lemma 1.2.1 because

$$\mathcal{L}(\mathbb{V}^{(1)},\dots,\mathbb{V}^{(L)},\mathbb{W}^*;\mathbb{K}) \cong \mathcal{L}\left(\mathbb{V}^{(1)},\dots,\mathbb{V}^{(L)};\mathcal{L}(\mathbb{W}^*;\mathbb{K})\right) \cong \mathcal{L}(\mathbb{V}^{(1)},\dots,\mathbb{V}^{(L)};\mathbb{W}).$$

With corollary 1.2.2 we are able to concentrate our attention to multilinear maps of the form $\mathcal{T} : \mathbb{V}^{(1)} \times \ldots \times \mathbb{V}^{(L)} \to \mathbb{K}$. In the case of a linear map $\mathcal{T} : \mathbb{K}^n \to \mathbb{K}$ (linear functional), we know there is a vector $\mathbf{a} \in \mathbb{K}^n$ such that

$$\mathcal{T}(\mathbf{x}) = \mathbf{a}^T\mathbf{x} \tag{1.1}$$

for all $\mathbf{x} \in \mathbb{K}^n$. In the case of a 2-linear (bilinear) map $\mathcal{T} : \mathbb{K}^m \times \mathbb{K}^n \to \mathbb{K}$, there is a matrix $\mathbf{A} \in \mathbb{K}^{n\times m}$ such that[2]

$$\mathcal{T}(\mathbf{x},\mathbf{y}) = \mathbf{y}^T\mathbf{A}\mathbf{x} \tag{1.2}$$

for all $\mathbf{x} \in \mathbb{K}^m, \mathbf{y} \in \mathbb{K}^n$. If one want a general formula for 3-linear (trilinear) maps $\mathcal{A} : \mathbb{K}^m \times \mathbb{K}^n \times \mathbb{K}^p \to \mathbb{K}$ or more, the concept of tensors is a must. In order to have a better understanding of what is happening it is convenient to work in coordinates after fixing a basis for each $\mathbb{V}^{(\ell)}$.

Theorem 1.2.3. *Let* $\{\boldsymbol{e}_1^{(\ell)},\dots,\boldsymbol{e}_{I_\ell}^{(\ell)}\}$ *be a basis for each* $\mathbb{V}^{(\ell)}$ *and let* $\mathcal{T} \in \mathcal{L}(\mathbb{V}^{(1)},\dots,\mathbb{V}^{(L)};\mathbb{K})$ *be a L-linear map. Then there exists scalars* $t_{i_1\dots i_L} \in \mathbb{K}$ *such that*

$$\mathcal{T}(\boldsymbol{e}_{i_1}^{(1)},\dots,\boldsymbol{e}_{i_L}^{(L)}) = t_{i_1\dots i_L},$$

for $i_1 = 1,\dots,I_1,\ \ i_2 = 1,\dots,I_2,\ \ \dots,\ \ i_L = 1,\dots,I_L$.

The values $t_{i_1\dots i_L}$ are called the *coordinates of* $\mathcal{T}$ with respect to the given bases. As it happens for linear maps and matrices, once we have fixed bases it is possible to associated the multilinear map $\mathcal{T}$ with the coordinates $t_{i_1\dots i_L}$. This is the same thing we do with matrices, considering them as a static table of numbers or as a linear transformation, depending on the context. So one can identify $\mathcal{T}$ with its coordinate representation. In the case $\mathcal{T} : \mathbb{K}^n \to \mathbb{K}$ we have that

$$\mathcal{T} = \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix},$$

[2]In the complex case there is the notion of a sequilinear form, which is a map $(\mathbf{x},\mathbf{y}) \mapsto \mathbf{y}^*\mathbf{A}\mathbf{x}$. Sesquilinear forms and bilinear complex forms are not the same thing.

in the case $\mathcal{T} : \mathbb{K}^m \times \mathbb{K}^n \to \mathbb{K}$ we have

$$\mathcal{T} = \begin{bmatrix} t_{11} & \dots & t_{1n} \\ \vdots & & \vdots \\ t_{m1} & \dots & t_{mn} \end{bmatrix},$$

and in the case $\mathcal{T} : \mathbb{K}^m \times \mathbb{K}^n \times \mathbb{K}^p \to \mathbb{K}$ we have the "rectangular matrix" below.

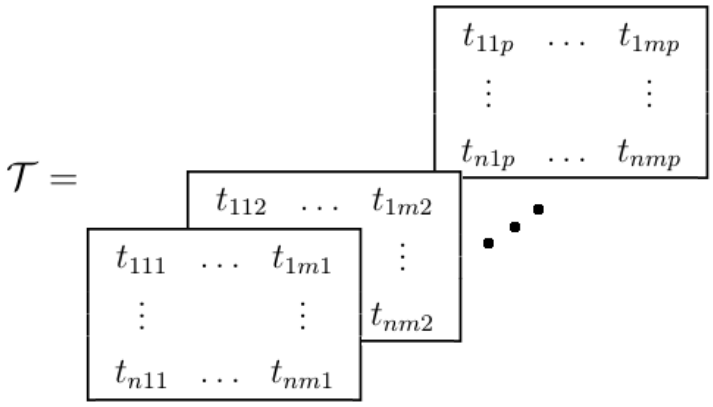

Figure 1.1: Trilinear map in coordinates.

Remark 1.2.4. *In the case $\mathcal{T} : \mathbb{K}^n \to \mathbb{K}$ the vector $\mathcal{T}$ is related with $\boldsymbol{a}$ by the identity $\mathcal{T} = \boldsymbol{a}^T$, that is, $\mathcal{T}(\boldsymbol{x}) = \boldsymbol{a}^T \cdot \boldsymbol{x}$. In the case $\mathcal{T} : \mathbb{K}^m \times \mathbb{K}^n \to \mathbb{K}$ the matrix $\mathcal{T}$ is $m \times n$ while $\boldsymbol{A}$ is $n \times m$. Both matrices are related by the identity $\mathcal{T} = \boldsymbol{A}^T$. Now we have that $\mathcal{T}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y}^T \boldsymbol{A} \boldsymbol{x}$.*

Remember that these coordinate representations are tensor of orders 1,2,3, respectively. Now let's see how one can obtain an explicit formula for $\mathcal{T}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)})$, where $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}) \in \mathbb{V}^{(1)} \times \dots \times \mathbb{V}^{(L)}$ is arbitrary and such that $\mathbf{x}^{(\ell)} = x_1^{(\ell)} \mathbf{e}_1^{(\ell)} + \dots + x_{I_\ell}^{(\ell)} \mathbf{e}_{I_\ell}^{(\ell)}$ for each $\ell = 1, \dots, L$. As consequence of theorem 1.2.3 we have that

$$\mathcal{T}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}) = \mathcal{T}(x_1^{(1)} \mathbf{e}_1^{(1)} + \dots + x_{I_1}^{(1)} \mathbf{e}_{I_1}^{(1)}, \dots, x_1^{(L)} \mathbf{e}_1^{(L)} + \dots + x_{I_L}^{(L)} \mathbf{e}_{I_L}^{(L)}) =$$

$$= \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} x_{i_1}^{(1)} \dots x_{i_L}^{(L)} \mathcal{T}(\mathbf{e}_{i_1}^{(1)}, \dots, \mathbf{e}_{i_L}^{(L)}) =$$

$$= \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} x_{i_1}^{(1)} \dots x_{i_L}^{(L)} t_{i_1 \dots i_L}. \tag{1.3}$$

The vectors obtained by fixing all dimensions except one are important and have their own name.

Definition 1.2.5. *Let $1 \leq \ell \leq L$. Then, for each choice of indexes $i_1, \ldots, i_{\ell-1}, i_{\ell+1}, \ldots, i_L$, the vector $\mathcal{T}_{i_1 \ldots i_{\ell-1} \, : \, i_{\ell+1} \ldots i_L}$ is called a mode-ℓ fiber of $\mathcal{T}$.*

In the case $L = 1$ the only fiber is $\mathcal{T}$ itself. In the case $L = 2$ the mode-1 fibers are the columns and the mode-2 fibers are the rows of $\mathcal{T}$. The case $L = 3$ is illustrated in figure 1.2.

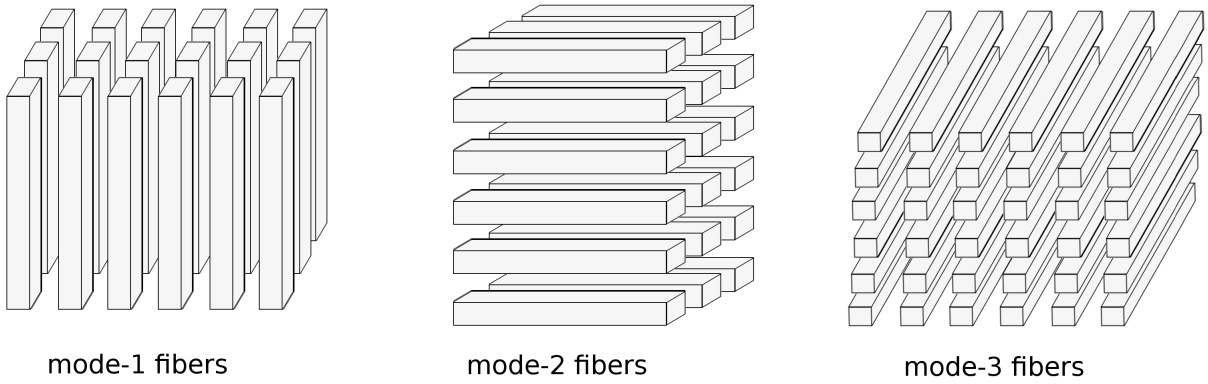

Figure 1.2: Fibers of a third order tensor.

Although we are not going to use this now, it will be important to us the subtensors we obtain by fixing all dimensions except two. This will give rise to bidimensional subtensors, that is, matrices.

Definition 1.2.6. *Let $1 \leq \ell < \ell' \leq L$. Then, for each choice of indexes*

$$i_1, \ldots, i_{\ell-1}, i_{\ell+1}, \ldots, i_{\ell'-1}, i_{\ell'+1}, \ldots, i_L,$$

the vector $\mathcal{T}_{i_1 \ldots i_{\ell-1} \, : \, i_{\ell+1} \ldots \, i_{\ell-1} \, : \, i_{\ell+1}, \ldots i_L}$ is called a slice *of $\mathcal{T}$.*

In the special case of $\mathcal{T}$ being a third order tensor, we can call the matrices $\mathcal{T}_{i::}$ the *horizontal slices*, $\mathcal{T}_{:j:}$ the *lateral slices*, and $\mathcal{T}_{::k}$ the *frontal slices*. These types of slices are illustrated in figure 1.3.

1.3 Tensor products

Definition 1.3.1. *Let $f^{(\ell)} \in (\mathbb{V}^{(\ell)})^*$ for each $\ell = 1 \ldots L$. The* tensor product *between the functionals $f^{(\ell)}$ is the map $f^{(1)} \otimes \ldots \otimes f^{(L)} : \mathbb{V}^{(1)} \times \ldots \times \mathbb{V}^{(L)} \to \mathbb{K}$ defined as*

$$f^{(1)} \otimes \ldots \otimes f^{(L)}(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(L)}) = f^{(1)}(\boldsymbol{x}^{(1)}) \cdot \ldots \cdot f^{(L)}(\boldsymbol{x}^{(L)}).$$

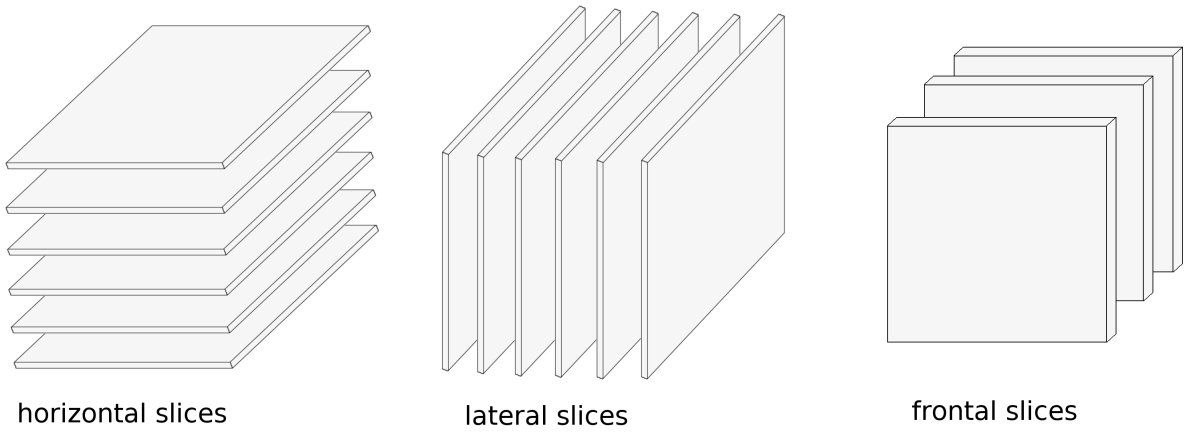

Figure 1.3: Slices of a third order tensor.

The linear space generated by all tensor products of the form $f^{(1)} \otimes \ldots \otimes f^{(L)}$ is denoted by $(\mathbb{V}^{(1)})^* \otimes \ldots \otimes (\mathbb{V}^{(L)})^*$ and called the *tensor product* between the spaces $(\mathbb{V}^{(\ell)})^*$. An element of $(\mathbb{V}^{(1)})^* \otimes \ldots \otimes (\mathbb{V}^{(L)})^*$ can be called a *covariant L-tensor* or a *covariant L-th order tensor*.

Definition 1.3.2. *Let $\boldsymbol{v}^{(\ell)} \in \mathbb{V}^{(\ell)}$ for each $\ell = 1 \ldots L$. The* tensor product *between the vectors $\boldsymbol{v}^{(\ell)}$ is the map $\boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)} : (\mathbb{V}^{(1)})^* \times \ldots \times (\mathbb{V}^{(L)})^* \to \mathbb{K}$ defined as*

$$\boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)}(f^{(1)}, \ldots, f^{(L)}) = f^{(1)}(\boldsymbol{v}^{(1)}) \cdot \ldots \cdot f^{(L)}(\boldsymbol{v}^{(L)}).$$

The linear space generated by all tensor products of the form $\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)}$ is denoted by $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ and called the *tensor product* between the spaces $\mathbb{V}^{(\ell)}$. An element of $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ can be called a *contravariant L-tensor* or a *contravariant L-th order tensor*.

One may also work with *mixed tensors*, that is, tensors which are the product of vector spaces and dual spaces. Since the ordering is not so important (lemma 1.2.1) we can define a mixed tensor to be an element of the space $(\mathbb{V}^{(1)})^* \otimes \ldots \otimes (\mathbb{V}^{(L)})^* \otimes \mathbb{W}^{(1)} \otimes \ldots \otimes \mathbb{W}^{(M)}$. These tensors are called *tensors of type* (L, M). In particular, a contravariant L-th order tensor is a tensor of type $(L, 0)$ and a covariant M-th order tensor is a tensor of type $(0, M)$. Generally, one may refer to a tensor product of vector spaces just as a *tensor space*. To finish this set of notations and terminology, when we have tensor products between the same space $\mathbb{V}$, it is common to denote $\mathbb{V}^{\otimes L} = \underbrace{\mathbb{V} \otimes \ldots \otimes \mathbb{V}}_{L \text{ times}}$. The next theorem summarizes the main properties of tensor spaces and their relation to multilinear maps. For more details about the algebra of tensor products, consult appendix B.

Theorem 1.3.3. *The following statements holds true.*

1. $(\mathbb{V}^{(1)})^* \otimes \ldots \otimes (\mathbb{V}^{(L)})^* \cong \mathcal{L}(\mathbb{V}^{(1)}, \ldots, \mathbb{V}^{(L)}; \mathbb{K})$

2. $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)} \cong \mathcal{L}\big((\mathbb{V}^{(1)})^*, \ldots, (\mathbb{V}^{(L)})^*; \mathbb{K}\big)$

3. $\dim\big((\mathbb{V}^{(1)})^* \otimes \ldots \otimes (\mathbb{V}^{(L)})^*\big) = \dim(\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}) = \prod_{\ell=1}^{L} I_\ell$

4. $\Big\{ \boldsymbol{e}_{i_1}^{(1)} \otimes \ldots \otimes \boldsymbol{e}_{i_L}^{(L)} : \ i_1 = 1 \ldots I_1, \ \ldots, \ i_L = 1 \ldots I_L \Big\}$ *is a basis for* $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$

5. *Any tensor* $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ *may be written as*

$$\mathcal{T} = \sum_{i_1=1}^{I_1} \ldots \sum_{i_L=1}^{I_L} t_{i_1 \ldots i_L} \ \boldsymbol{e}_{i_1}^{(1)} \otimes \ldots \otimes \boldsymbol{e}_{i_L}^{(L)},$$

where $t_{i_1 \ldots i_L}$ *are the coordinates given in theorem 1.2.3.*

Remark 1.3.4. *Sometimes it is useful to consider the isomorphism* $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)} \cong \mathcal{L}\big((\mathbb{V}^{(1)})^*, \ldots, (\mathbb{V}^{(L-1)})^*; \ \mathbb{V}^{(L)}\big)$ *and consider* $\boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ *as the map given by*

$$\boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)}(f^{(1)}, \ldots, f^{(L-1)}) = f^{(1)}(\boldsymbol{v}^{(1)}) \cdot \ldots \cdot f^{(L)}(\boldsymbol{v}^{(L-1)}) \cdot \boldsymbol{v}^{(L)}. \tag{1.4}$$

Example 1.3.5. *Consider the space* $\mathbb{C}^2$ *with basis* $B = \{[1,0]^T, [0,-i]^T\} = \{\boldsymbol{b}_1, \boldsymbol{b}_2\}$, *where* $i = \sqrt{-1}$ *is the imaginary unit. The dual basis associated to* B *is* $B^* = \{\boldsymbol{b}_1^*, \boldsymbol{b}_2^*\}$. *Let* $\mathcal{T} : \mathbb{C}^2 \to \mathbb{C}^2$ *such that* $\mathcal{T}(z,w) = [iz, z+w]^T$. *As a tensor, note that* $\mathcal{T} \in (\mathbb{C}^2)^* \otimes \mathbb{C}^2$ *is a mixed tensor of type* $(1,1)$.

To compute $\mathcal{T}$ *in coordinates, first note that* $\mathcal{T}(\boldsymbol{b}_1) = [i,1]^T$ *and* $\mathcal{T}(\boldsymbol{b}_2) = [0,-i]^T$. *On the other hand, by interpreting* $\mathcal{T}$ *as a tensor product we know that*

$$\mathcal{T} = \sum_{j,k=1}^{2} t_{jk} \ \boldsymbol{b}_j^* \otimes \boldsymbol{b}_k.$$

Using this formula and the identification given in 1.4 we have that

$$\mathcal{T}(\boldsymbol{b}_1) = t_{11}\boldsymbol{b}_1^*(\boldsymbol{b}_1) \cdot \boldsymbol{b}_1 + t_{12}\boldsymbol{b}_1^*(\boldsymbol{b}_1) \cdot \boldsymbol{b}_2 + t_{21}\boldsymbol{b}_1^*(\boldsymbol{b}_2) \cdot \boldsymbol{b}_1 + t_{22}\boldsymbol{b}_1^*(\boldsymbol{b}_2) \cdot \boldsymbol{b}_2 = \begin{bmatrix} t_{11} \\ -it_{12} \end{bmatrix}$$

and

$$\mathcal{T}(\boldsymbol{b}_2) = t_{11}\boldsymbol{b}_1^*(\boldsymbol{b}_2) \cdot \boldsymbol{b}_1 + t_{12}\boldsymbol{b}_2^*(\boldsymbol{b}_1) \cdot \boldsymbol{b}_2 + t_{21}\boldsymbol{b}_2^*(\boldsymbol{b}_2) \cdot \boldsymbol{b}_1 + t_{22}\boldsymbol{b}_2^*(\boldsymbol{b}_2) \cdot \boldsymbol{b}_2 = \begin{bmatrix} -t_{21} \\ -it_{22} \end{bmatrix}.$$

With this we conclude that $t_{11} = i, t_{12} = i, t_{21} = 0, t_{22} = 1$. *Therefore we have that*

$$\mathcal{T} = i\boldsymbol{b}_1^* \otimes \boldsymbol{b}_1 + i\boldsymbol{b}_1^* \otimes \boldsymbol{b}_2 + \boldsymbol{b}_2^* \otimes \boldsymbol{b}_2.$$

There is a little subtlety here. Remember that, by remark 1.2.4, it is necessary to transpose the coordinate representation of $\mathcal{T}$. *After transposing we get the matrix of* $\mathcal{T}$ *in basis* B,

$$\mathcal{T} = \begin{bmatrix} i & 0 \\ i & 1 \end{bmatrix}.$$

The procedure described here is generalizable to any kind of linear map $\mathbb{K}^m \to \mathbb{K}^n$. *It permits one to compute the associated tensor and its associated matrix.*

It is customary to use upper and lower index to distinguish between contravariant and covariant terms, but this won't be necessary here. In this work we will be only interested in studying tensors in the Euclidean space $\mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$, and for this reason we will leave aside that index convention.

Let $\mathcal{T} = \mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$. Remember we can consider $\mathcal{T}$ as the map $\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)} : (\mathbb{K}^{I_1})^* \otimes \ldots \otimes (\mathbb{K}^{I_L})^* \to \mathbb{K}$ given by

$$\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)}\big((\mathbf{x}^{(1)})^*, \ldots, (\mathbf{x}^{(L)})^*\big) = \langle \mathbf{v}^{(1)}, \mathbf{x}^{(1)} \rangle \cdot \ldots \cdot \langle \mathbf{v}^{(L)}, \mathbf{x}^{(L)} \rangle =$$

$$= \sum_{i_1=1}^{I_1} \ldots \sum_{i_L=1}^{I_L} \overline{x_{i_1}^{(1)} \ldots x_{i_L}^{(L)}} \cdot v_{i_1}^{(1)} \ldots v_{i_L}^{(L)}. \tag{1.5}$$

from which we conclude that $t_{i_1 \ldots i_L} = v_{i_1}^{(1)} \ldots v_{i_L}^{(L)}$. The values $t_{i_1 \ldots i_L}$ are the coordinates of the $\mathcal{T}$ as a multilinear map and as a tensor. It is also possible to use theorem 1.2.3 to compute these coordinates. Although we are mainly concerned with tensors in $\mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$, there are some classic examples of different types of tensors we want to show. We consider the canonical bases for all examples below.

Example 1.3.6 (Rank one matrix)**.** *Given two vectors* $\boldsymbol{v} \in \mathbb{C}^m, \boldsymbol{u} \in \mathbb{C}^n$, *consider the linear map with matrix* $\boldsymbol{u}\boldsymbol{v}^*$. *In this example we will see that the tensor associated to this map is* $\mathcal{T} = \boldsymbol{v}^* \otimes \boldsymbol{u} \in (\mathbb{C}^m)^* \otimes \mathbb{C}^n$. *First note that*

$$\mathcal{T}(\boldsymbol{x}, \boldsymbol{y}^*) = \boldsymbol{v}^*(\boldsymbol{x}) \cdot \boldsymbol{y}^*(\boldsymbol{u}) = \langle \boldsymbol{x}, \boldsymbol{v} \rangle \cdot \langle \boldsymbol{u}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i \overline{y}_j \overline{v}_i u_j$$

for all $\boldsymbol{x} \in \mathbb{C}^m, \boldsymbol{y} \in \mathbb{C}^n$. *Considering* $\mathcal{T}$ *in coordinates, we can write* $\mathcal{T} = \overline{\boldsymbol{v}}\boldsymbol{u}^T$, *where* $t_{ij} = \overline{v}_i u_j$. *The matrix of the corresponding linear map is the transpose of this matrix (see 1.2.4), that is,* $\boldsymbol{A} = (\overline{\boldsymbol{v}}\boldsymbol{u}^T)^T = \boldsymbol{u}\boldsymbol{v}^*$, *as desired. Note that now we can write* $\mathcal{T}(\boldsymbol{x}, \boldsymbol{y}^*) = \boldsymbol{y}^* \boldsymbol{A} \boldsymbol{x}$. *We may use the isomorphism* $(\mathbb{C}^m)^* \otimes \mathbb{C}^n \cong L(\mathbb{C}^m; \mathbb{C}^n)$ *from 1.4 and reinterpret*

$\mathcal{T}$ *as the map* $\mathcal{T}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{v} \rangle \cdot \boldsymbol{u} = \boldsymbol{A}\boldsymbol{x}$.

Example 1.3.7 (SVD)**.** *Again, let* $\mathcal{T} \in (\mathbb{C}^m)^* \otimes \mathbb{C}^n$*, but this time suppose there are vectors* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R \in \mathbb{C}^m$*,* $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_R \in \mathbb{C}^n$*, and scalars* $\sigma_1, \ldots, \sigma_R \in \mathbb{C}$ *such that* $\mathcal{T} = \sum_{r=1}^{R} \sigma_r \boldsymbol{v}_r^* \otimes \boldsymbol{u}_r$ *and let* $\boldsymbol{A} \in \mathbb{C}^{n \times m}$ *be the matrix of the corresponding linear map, as before. Additionally, suppose this is the least* R *with the property that such decomposition exists. As a consequence we have a SVD for* $\boldsymbol{A}$ *given by* $\boldsymbol{A} = \boldsymbol{U}\Sigma\boldsymbol{V}^* = \sum_{r=1}^{R} \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^*$*, where* $\boldsymbol{u}_r$ *is the* r*-th column of* $\boldsymbol{U}$ *(left singular vector),* $\boldsymbol{v}_r$ *is the* r*-th column* $\boldsymbol{V}$ *(right singular vector), and* $\Sigma = diag(\sigma_1, \ldots, \sigma_R)$.

Example 1.3.8 (Matrix multiplication)**.** *For this last example, consider the tensor space isomorphism* $(\mathbb{C}^{mn})^* \otimes (\mathbb{C}^{nl})^* \otimes \mathbb{C}^{ml} \cong \mathcal{L}\big(\mathbb{C}^{mn}, \mathbb{C}^{nl}; \mathbb{C}^{ml}\big)$*. Each element of* $\mathbb{C}^{mn}$ *can be thought as a* $m \times n$ *matrix or a vector of size* mn*. Given a* $m \times n$ *matrix* $\boldsymbol{X}$*, let* $vec(\boldsymbol{X})$ *be the vector obtained by vertically stacking all columns of* $\boldsymbol{X}$*. We will be identifying* $\boldsymbol{X}$ *and* $vec(\boldsymbol{X})$ *when it is convenient. The same considerations goes for the spaces* $\mathbb{C}^{nl}$ *and* $\mathbb{C}^{ml}$*. Let* $\boldsymbol{e}_{ij} \in \mathbb{C}^{m \times n}$ *be the matrix with entry* (i, j) *equal to 1 and all other entries equal to 0. Note that* $\{\boldsymbol{e}_{11}, \boldsymbol{e}_{12}, \ldots, \boldsymbol{e}_{mn}\}$ *is a basis for* $\mathbb{C}^{m \times n}$*, while* $\{vec(\boldsymbol{e}_{11}), vec(\boldsymbol{e}_{12}), \ldots, vec(\boldsymbol{e}_{mn})\}$ *is the canonical basis of* $\mathbb{C}^{mn}$*. We will commit a little abuse of notation and use the same notation for the basis vectors of* $\mathbb{C}^{nl}$ *and* $\mathbb{C}^{ml}$.

Now, define the tensor $\mathcal{T} \in (\mathbb{C}^{mn})^* \otimes (\mathbb{C}^{nl})^* \otimes \mathbb{C}^{ml}$ *by*

$$\mathcal{T} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} \boldsymbol{e}_{ij}^* \otimes \boldsymbol{e}_{jk}^* \otimes \boldsymbol{e}_{ik} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} \boldsymbol{e}_{ij}^T \otimes \boldsymbol{e}_{jk}^T \otimes \boldsymbol{e}_{ik}.$$

Given any matrices $\boldsymbol{X} \in \mathbb{C}^{m \times n}$*,* $\boldsymbol{Y} \in \mathbb{C}^{n \times l}$*, and using the isomorphism 1.4, we have that*

$$\mathcal{T}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} \langle \boldsymbol{X}, \boldsymbol{e}_{ij} \rangle \cdot \langle \boldsymbol{Y}, \boldsymbol{e}_{jk} \rangle \cdot \boldsymbol{e}_{ik} = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} x_{ij} \cdot y_{jk} \cdot \boldsymbol{e}_{ik} =$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{l} (x_{i1} y_{1k} + \ldots + x_{in} y_{nk}) \cdot \boldsymbol{e}_{ik} = \boldsymbol{X} \cdot \boldsymbol{Y}$$

In short, $\mathcal{T}$ *is a third order tensor describing the matrix multiplication. Note that we used* mnl *terms in the summation defining* $\mathcal{T}$*. It is possible to use less terms, and the problem of finding the minimum number of terms is an open problem in mathematics. For instance, see chapter 1 of [39].*

1.4 Canonical polyadic decomposition

When $\mathcal{T}$ is of the form $\mathcal{T} = \mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)}$, we saw that $t_{i_1 \ldots i_L} = v_{i_1}^{(1)} \ldots v_{i_L}^{(L)}$. However, note that this formula does not apply for all the tensor space since not all tensors are of this form. An arbitrary tensor in $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ may be written as

$$\mathcal{T} = \sum_{r=1}^{R} \mathbf{v}_r^{(1)} \otimes \ldots \otimes \mathbf{v}_r^{(L)}, \tag{1.6}$$

where each $\mathbf{v}_r^{(\ell)} \in \mathbb{V}^{(\ell)}$ is given by $\mathbf{v}_r^{(\ell)} = \left[v_{1r}^{(\ell)}, v_{2r}^{(\ell)}, \ldots, v_{I_\ell r}^{(\ell)}\right]^T$. It is of interest in applications to decompose $\mathcal{T}$ in this manner, such that R is smallest as possible [2, 3, 5, 7–9]. The generalization coordinate representation of $\mathcal{T}$ is now given by

$$t_{i_1 \ldots i_L} = \sum_{r=1}^{R} v_{i_1 r}^{(1)} \ldots v_{i_L r}^{(L)}. \tag{1.7}$$

Formula 1.6 realizes $\mathcal{T}$ as a sum of R tensor products. For each $\ell = 1 \ldots L$, the ℓ-th *factor matrix* associated to 1.6 is defined as $\mathbf{V}^{(\ell)} = [\mathbf{v}_1^{(\ell)}, \ldots, \mathbf{v}_R^{(\ell)}] \in \mathbb{K}^{I_\ell \times R}$, and each one of its columns are called *factors*. The space $\mathbb{V}^{(\ell)}$ sometimes is referred as the *ℓ-th mode*. When some definition depends on ℓ it is common to use some terminology which specify the current mode.

Definition 1.4.1. *We say a tensor $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ has* rank one *if there exists vectors $\boldsymbol{v}^{(1)} \in \mathbb{V}^{(1)}, \ldots, \boldsymbol{v}^{(L)} \in \mathbb{V}^{(L)}$ such that $\mathcal{T} = \boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)}$.*

Definition 1.4.2. *We say a tensor $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ has* rank R *if R is the smallest number such that $\mathcal{T}$ can be written as a sum of R rank one tensors. In this case we denote $rank(\mathcal{T}) = R$.*

Suppose $rank(\mathcal{T}) = R$. Then the decomposition 1.6 is called a *canonical polyadic decomposition (CPD)* for $\mathcal{T}$. Other known names for the CPD are *PARAFAC* (parallel factors), *CANDECOMP* (canonical decomposition) and *CP decomposition*. In the case R is not the rank of $\mathcal{T}$ we call this decomposition a *rank-R CPD for $\mathcal{T}$*. The first one to propose the notion of rank and this decomposition was Hitchcock [50] in a work of 1927.

In example 1.3.7 we discussed a connection between the SVD for matrices and tensors. If $\mathcal{T} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^*$ is a SVD for $\mathcal{T}$, then we can write $\mathcal{T} = \sum_{r=1}^{R} \sigma_r \mathbf{v}_r \otimes \mathbf{u}_r^*$. In particular, this implies that $rank(\mathcal{T}) = R$. Conversely, if $rank(\mathcal{T}) = R$ and $\mathcal{T} = \sum_{r=1}^{R} \sigma_r \mathbf{v}_r \otimes \mathbf{u}_r^*$, then $\mathcal{T} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^*$ is a SVD for $\mathcal{T}$ and, as a matrix, $\mathcal{T}$ has rank R. In conclusion, if $\mathcal{T}$ is a second order tensor, then its rank as a tensor equals its rank as a matrix. Furthermore, its CPD and its SVD coincide.

Given $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$, we know in advance that $rank(\mathcal{T}) \leq \prod_{\ell=1}^{L} I_\ell$, since we always can write

$$\mathcal{T} = \sum_{i_1=1}^{I_1} \ldots \sum_{i_L=1}^{I_L} t_{i_1 \ldots i_L} \, \mathbf{e}_{i_1}^{(1)} \otimes \ldots \otimes \mathbf{e}_{i_L}^{(L)}.$$

In fact, it is possible to write $\mathcal{T}$ with less terms as the next results shows. From theorem 1.2.1 we know it is possible to permute the spaces $\mathbb{V}^{(\ell)}$ without problems. Hence there is no loss of generality in considering $I_1 \geq I_2 \geq \ldots \geq I_L$.

Theorem 1.4.3 (Landsberg, [39])**.** *Let $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ be such that $I_1 \geq I_2 \geq \ldots \geq I_L$. Then $rank(\mathcal{T}) \leq \prod_{\ell=2}^{L} I_\ell$ for all $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$.*

Corollary 1.4.4. *$rank(\mathcal{T}) \leq \min\{I_1 I_2, \; I_1 I_3, \; I_2 I_3\}$ for all $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \mathbb{V}^{(2)} \otimes \mathbb{V}^{(3)}$.*

Remember the slices of third order tensors we showed in 1.3. The next result gives formulas for each one of these slices.

Theorem 1.4.5. *Let $\mathcal{T} = \sum_{r=1}^{R} \boldsymbol{x}_r \otimes \boldsymbol{y}_r \otimes \boldsymbol{z}_r \in \mathbb{K}^{I_1} \otimes \mathbb{K}^{I_2} \otimes \mathbb{K}^{I_3}$ be a third order tensor with rank $\leq R$, where each λ_r is a scalar. Then the i-th horizontal slice of $\mathcal{T}$ is given by*

$$\sum_{r=1}^{R} x_{ir} \; \boldsymbol{y}_r \otimes \boldsymbol{z}_r = \boldsymbol{Y} \cdot diag(x_{i1}, \ldots, x_{iR}) \cdot \boldsymbol{Z}^T,$$

the j-th lateral slice is given by

$$\sum_{r=1}^{R} y_{jr} \; \boldsymbol{x}_r \otimes \boldsymbol{z}_r = \boldsymbol{X} \cdot diag(y_{j1}, \ldots, y_{jR}) \cdot \boldsymbol{Z}^T.$$

and the k-th frontal slice slice is given by

$$\sum_{r=1}^{R} z_{kr} \; \boldsymbol{x}_r \otimes \boldsymbol{y}_r = \boldsymbol{X} \cdot diag(z_{k1}, \ldots, x_{kR}) \cdot \boldsymbol{Y}^T.$$

Now suppose we have a tensor $\mathcal{T}$ with rank R and we want to compute a CPD for $\mathcal{T}$. In practical applications obtaining equality as in formula 1.6 is not realistic. Usually one is content with an approximation

$$\mathcal{T} \approx \sum_{r=1}^{R} \mathbf{v}_r^{(1)} \otimes \ldots \otimes \mathbf{v}_r^{(L)}. \tag{1.8}$$

There are several algorithms to accomplish this goal and we will discuss some of them in chapter 3. For now we discuss rank properties. For instance, how should one proceed

when $rank(\mathcal{T})$ is not known? In order to obtain a CPD for $\mathcal{T}$ one would want to know its rank in the first place. Unfortunately this problem is known to be NP-hard [21]. Another possibility would be to choose a large value R, an upper bound for $rank(\mathcal{T})$, and compute a rank-R CPD for $\mathcal{T}$. As we will see soon this is not a good idea because a high rank CPD suffer from lack of uniqueness. In particular, this kind of CPD can overfit the data we are trying to model. The best choice here is to choose a low rank CPD approximation for $\mathcal{T}$. Caution is necessary to not take R too small, because in that case our model will suffer from underfitting (high bias) and accuracy is lost.

A relevant property of higher order tensors is that their CPD are often unique (in a sense we will make clear soon). This property fail for matrices. For instance, consider a matrix $\mathcal{T} \in \mathbb{K}^{n\times m}$ together with a SVD given by $\mathcal{T} = \mathbf{U}\cdot\Sigma\cdot\mathbf{V}^*$, where $\Sigma = \operatorname{diag}(\sigma_1,\ldots,\sigma_R)$. Making $\mathbf{A} = \mathbf{U}\cdot\Sigma$ and $\mathbf{B} = \mathbf{V}$ we have

$$\mathcal{T} = \mathbf{A}\mathbf{B}^* = \sum_{r=1}^{R} \mathbf{A}_{:r}\mathbf{B}^*_{:r}.$$

Notice this is a CPD for $\mathcal{T}$ since it is a sum of R rank one terms. Now let $\mathbf{W} \in \mathbb{K}^{R\times R}$ be unitary. Then we have

$$\mathcal{T} = \mathbf{A}\mathbf{W}(\mathbf{B}\mathbf{W})^* = \tilde{\mathbf{A}}\tilde{\mathbf{B}}^* = \sum_{r=1}^{R} \tilde{\mathbf{A}}_{:r}\tilde{\mathbf{B}}^*_{:r}.$$

Varying $\mathbf{W}$ we can obtain infinitely many different CPD's for $\mathcal{T}$.

Let $\mathcal{T} = \sum_{r=1}^{R} \mathcal{T}_r$ be a CPD for $\mathcal{T}$, where each $\mathcal{T}_r$ is a rank one term. Also, suppose $\mathcal{T}$ is a higher order (bigger than 2) tensor. The uniqueness of the CPD is up to the following trivial modifications:

1. Permutation of the ordering of the rank one terms. $\mathcal{T} = \sum_{r=1}^{R} \mathcal{T}_{\sigma(r)}$ is the same CPD, where $\sigma \in S_R$ is any permutation.

2. Scaling indeterminacy. $\mathcal{T} = \sum_{r=1}^{R} \frac{1}{\lambda_r}(\lambda_r\mathcal{T}_r)$ is the same CPD, where $\lambda_r \neq 0$ is arbitrary for all $r = 1\ldots R$.

Sometimes one say the CPD is *essentially unique*. Is this uniqueness what makes the CPD so attractive to applications. Most of the time the CPD is unique, but sometimes this may not be the case. For this reason we want to stablish uniqueness conditions. Additionally, we should clarify what means when we say the CPD is unique "most of the time".

The most well known result on uniqueness of tensors is due to J. B. Kruskal [45, 47] although it is limited to third order tensors. Posteriorly this result was extended to

arbitrary higher order tensors by N. D. Sidiropoulos and R. Bro [49]. Below we show this extended result.

Definition 1.4.6. *Let $\boldsymbol{X} \in \mathbb{K}^{m\times n}$ be a matrix. The* k-rank *of $\boldsymbol{X}$ is the maximum value k such that any k columns of $\boldsymbol{X}$ are linearly independent. We denote this value by $k_{\boldsymbol{X}}$.*

Theorem 1.4.7 (N. D. Sidiropoulos and R. Bro)**.** *Let $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ be a tensor of rank R with CPD given by formula 1.6 and let $\boldsymbol{V}^{(\ell)}$ be the ℓ-th factor matrix of this CPD, for $\ell = 1 \ldots L$. If*

$$\sum_{\ell=1}^{L} k_{\boldsymbol{V}^{(\ell)}} \geq 2R + L - 1,$$

then this CPD is unique.

In the matrix case ($L = 2$), suppose that both factors, $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$, have all columns linearly independent. Then we have that $k_{\mathbf{V}^{(1)}} + k_{\mathbf{V}^{(2)}} = R + R < 2R + 1$. Therefore the CPD is never unique in the matrix case, a fact we had already observed. With this we have a condition for uniqueness.

1.5 Tensor geometry

Given a tensor space $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ with a basis $\{\mathbf{e}_{i_1}^{(1)} \otimes \ldots \otimes \mathbf{e}_{i_L}^{(L)}\}$, we can consider each tensor as a element of $\mathbb{K}^{I_1\times\ldots\times I_L}$, that is, each tensor

$$\mathcal{T} = \sum_{i_1=1}^{I_1} \ldots \sum_{i_L=1}^{I_L} t_{i_1\ldots i_L}\ \mathbf{e}_{i_1}^{(1)} \otimes \ldots \otimes \mathbf{e}_{i_L}^{(L)}$$

is identified with the multidimensional array with entries $t_{i_1\ldots i_L}$. In this case we can consider $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ a space with inner product defined by

$$\langle \mathcal{T}, \mathcal{S} \rangle = \sum_{i_1=1}^{I_1} \ldots \sum_{i_L=1}^{I_L} t_{i_1\ldots i_L} \overline{s_{i_1\ldots i_L}}.$$

This induces the norm

$$\|\mathcal{T}\| = \sqrt{\sum_{i_1=1}^{I_1} \ldots \sum_{i_L=1}^{I_L} |t_{i_1\ldots i_L}|^2}.$$

This allow us to talk about proximity of tensors. This is relevant because usually one is interested is solving 1.8 in the best way possible, that is, to obtain the rank-R tensor closest to $\mathcal{T}$ between all rank-R approximations. Unfortunately, computing the best rank-

R approximation of a tensor is, in general, a *ill-posed*[3] problem [20]. In particular we have the following result, first observed in [51] and then deeply explored in [39].

Theorem 1.5.1. *The limit of a sequence of rank tensors R is not necessarily a rank-R tensor.*

Denote $\sigma_R\left(\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}\right) = \left\{\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)} : \ rank(\mathcal{T}) \leq R\right\}$ for the set of tensors with rank $\leq R$. With respect relation to the theorem above, the rank of the limit of tensors can give a "jump". Because of this, the set σ_R is not necessarily closed in the norm topology. This motivates the following definition.

Definition 1.5.2. *We say a tensor $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ has* border rank-R *if R is the smallest number such that $\mathcal{T} \in \overline{\sigma_R}$, where $\overline{\sigma_R}$ is the closure of σ_R. In this case we denote $\underline{rank}(\mathcal{T}) = R$.*

The term "border rank" first appeared in the paper [51] in the context of matrix multiplication. There is an interesting about the story of the border rank at the beginning of chapter 2 of [40]. We have the following result as a direct consequence of the definition.

Theorem 1.5.3. *If $\underline{rank}(\mathcal{T}) = R$, then there exists a sequence of rank-R tensors converging to $\mathcal{T}$ and there is not a sequence of tensors with rank $< R$ converging to $\mathcal{T}$.*

Corollary 1.5.4. $\underline{rank}(\mathcal{T}) \leq rank(\mathcal{T})$.

As we observed, computing the best rank-R approximation of a tensor is, in general, a ill-posed problem. However, this is not the case when $L = 2$, that is, the matrix case. This result is known since 1936 with Eckart and Young [53].

Theorem 1.5.5 (Eckart-Young, 1936)**.** *Let $\boldsymbol{M} = \sum_{r=1}^{R} \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^*$ be a SVD of a rank-R matrix in $\mathbb{K}^{n\times m}$. For any $1 \leq \tilde{R} \leq R$, the best rank-R approximation of $\boldsymbol{M}$ is given by $\tilde{\boldsymbol{M}} = \sum_{r=1}^{\tilde{R}} \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^*$.*

The phenomenon of border rank is the one responsible for the ill-posedness of the approximation problem. If a tensor $\mathcal{T}$ has rank R and border rank $\tilde{R} < R$, then there is a sequence of rank-$\tilde{R}$ tensors converging to $\mathcal{T}$. This implies, in particular, that it does not exists a tensor of rank $\tilde{R}$ closest to $\mathcal{T}$, so finding the best rank-$\tilde{R}$ approximation of $\mathcal{T}$ is a ill-posed problem. The first report of a such phenomenon was in [52], where they gave an explicit example of a sequence of rank-5 tensors converging to a rank 6 tensor in 1979. In [20] there is a simple example of a tensor of rank 3 and border rank 2 which we reproduce below for illustration purposes.

[3] We call a problem *well-posed* if a solution exists, is unique, and is stable in the sense it depends continuously int the input data. A problem is ill-posed if it is not well-posed.

Theorem 1.5.6 (V. de Silva and L. H. Lim, 2008)**.** *Let* $I_1, I_2, I_3 \geq 2$. *Let* $\mathcal{T} \in \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \mathbb{R}^{I_3}$ *be a tensor such that*

$$\mathcal{T} = \boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{y}^{(3)} + \boldsymbol{x}^{(1)} \otimes \boldsymbol{y}^{(2)} \otimes \boldsymbol{x}^{(3)} + \boldsymbol{y}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{x}^{(3)}$$

where each pair $\boldsymbol{x}^{(\ell)}, \boldsymbol{y}^{(\ell)} \in \mathbb{R}^{I_\ell}$ *is linearly independent. Then* $rank(\mathcal{T}) = 3$ *and*

$$\mathcal{T}^{(n)} = n\left(\boldsymbol{x}^{(1)} + \frac{1}{n}\boldsymbol{y}^{(1)}\right) \otimes \left(\boldsymbol{x}^{(2)} + \frac{1}{n}\boldsymbol{y}^{(2)}\right) \otimes \left(\boldsymbol{x}^{(3)} + \frac{1}{n}\boldsymbol{y}^{(3)}\right) - n\boldsymbol{x}^{(1)} \otimes \boldsymbol{x}^{(2)} \otimes \boldsymbol{x}^{(3)}$$

is a sequence of rank-2 tensors converging to $\mathcal{T}$. *In particular,* $\underline{rank}(\mathcal{T}) \leq 2$.

The tensor $\mathcal{T}$ of the theorem is an example of a tensor that has no best rank-2 approximation. It is interesting to note that the limit expression for $\mathcal{T}^{(n)}$ may be regarded as a derivative. In fact, define the function $f : \mathbb{R} \to \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \mathbb{R}^{I_3}$ by

$$f(t) = (\mathbf{x}^{(1)} + t\mathbf{y}^{(1)}) \otimes (\mathbf{x}^{(2)} + t\mathbf{y}^{(2)}) \otimes (\mathbf{x}^{(3)} + t\mathbf{y}^{(3)}) =$$
$$= \mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{x}^{(3)} + t\mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{y}^{(3)} + t\mathbf{x}^{(1)} \otimes \mathbf{y}^{(2)} \otimes \mathbf{x}^{(3)} + t^2\mathbf{x}^{(1)} \otimes \mathbf{y}^{(2)} \otimes \mathbf{y}^{(3)} +$$
$$+ t\mathbf{y}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{x}^{(3)} + t^2\mathbf{y}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{y}^{(3)} + t^2\mathbf{y}^{(1)} \otimes \mathbf{y}^{(2)} \otimes \mathbf{x}^{(3)} + t^3\mathbf{y}^{(1)} \otimes \mathbf{y}^{(2)} \otimes \mathbf{y}^{(3)}.$$

On one hand, using the derivative rules and making $t = 0$ we obtain

$$f'(0) = \mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{x}^{(3)} + \mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{y}^{(3)} + \mathbf{x}^{(1)} \otimes \mathbf{y}^{(2)} \otimes \mathbf{x}^{(3)}.$$

On the other hand, using the limit definition for the derivative we obtain

$$f'(0) = \lim_{t \to 0} \frac{(\mathbf{x}^{(1)} + t\mathbf{y}^{(1)}) \otimes (\mathbf{x}^{(2)} + t\mathbf{y}^{(2)}) \otimes (\mathbf{x}^{(3)} + t\mathbf{y}^{(3)}) - \mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{x}^{(3)}}{t}.$$

Making $t = 1/n$ and simplifying we obtain the expression of the theorem, that is, we have that $f'(0) = \lim_{n\to\infty} \mathcal{T}^{(n)}$. Notice that f represents a curve in σ_2. From the point $f(0) = \mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{x}^{(3)}$ we can draw secant lines in order to approximate the derivative $f'(0)$. Each secant line gives us a tensor in σ_2. At the limit we have the tangent tensor which, because of the theorem, will be outside σ_2. This is illustrated in figure 1.4.

With respect to the topology of σ_R in $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$, we have already seen that σ_R is not closed. It is also true that σ_R is not open under certain conditions, as the next result shows.

Theorem 1.5.7. *If* $R < \dim\left(\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}\right)$, *then* σ_R *is not open.*

Proof: Let $\mathcal{T} = \sum_{r=1}^{R} \mathbf{v}_r^{(1)} \otimes \ldots \otimes \mathbf{v}_r^{(L)} \in \sigma_R$ be a rank-R tensor and let $\mathbf{u}^{(1)} \otimes \ldots \otimes \mathbf{u}^{(L)}$ be a tensor which it is not a linear combination of the tensor products $\mathbf{v}_r^{(1)} \otimes \ldots \otimes \mathbf{v}_r^{(L)}$.

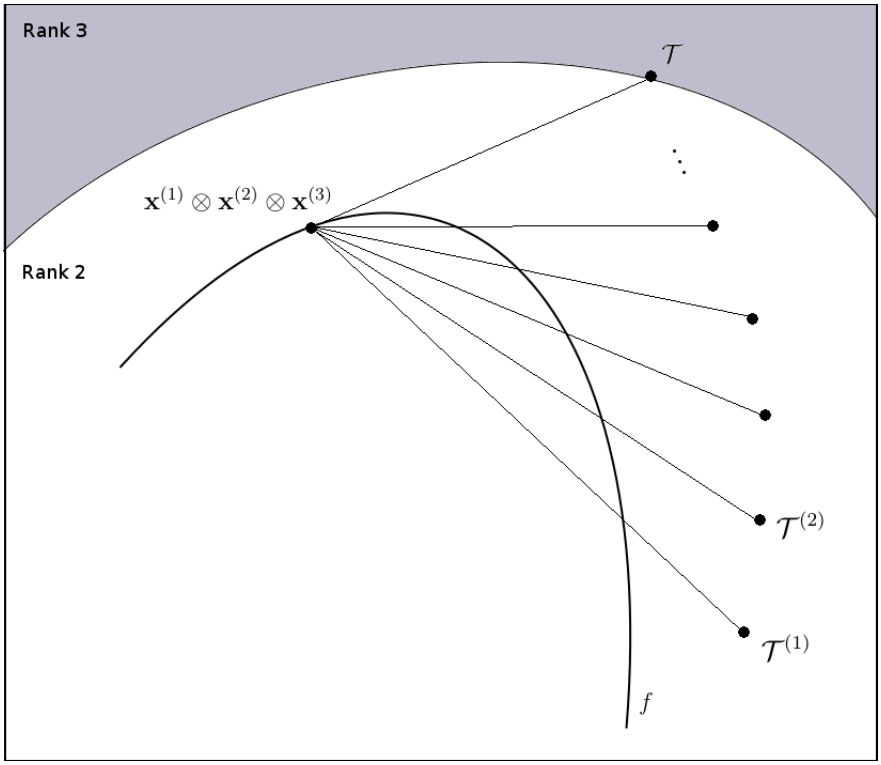

Figure 1.4: Geometry of border rank.

Then the sequence

$$\mathcal{T}^{(n)} = \frac{1}{n+1}\mathbf{u}^{(1)} \otimes \ldots \otimes \mathbf{u}^{(L)} + \mathcal{T}$$

converges to $\mathcal{T}$ and it is constituted of tensors with rank greater than R. Thus, every ball around $\mathcal{T}$ contains tensors outside σ_R. Therefore σ_R is not open. □

Note that this theorem also holds for the matrix space. More precisely, there are sequences of rank-R matrices converging to matrices with rank smaller than R. What does not occur with matrices is to have a sequence of rank-R matrices converging to a matrix with rank greater than R. This last phenomenon is unique to tensors of order > 2, and this is where we see the issue of border rank.

The argument used in the previous theorem also shows that every rank-R is an adherent point of $\left(\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}\right) \setminus \sigma_R$, hence the set of rank-$R$ tensor is contained in $\overline{\left(\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}\right) \setminus \sigma_R}$. Let $\partial(\sigma_R) = \overline{\sigma_R} \cap \overline{\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)} \setminus \sigma_R}$ be the boundary of σ_R. Then it follows that the set of rank-R tensors is contained in $\partial(\sigma_R)$. Next we give some results about norm invariance.

Theorem 1.5.8 (V. de Silva and L. H. Lim, 2008)**.** *Let* $\mathcal{T} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}, \mathcal{S} \in \mathbb{K}^{I'_1} \otimes \ldots \otimes \mathbb{K}^{I'_L}$ *and* $\boldsymbol{v}^{(1)} \in \mathbb{K}^{I_1}$, …, $\boldsymbol{v}^{(L)} \in \mathbb{K}^{I_L}$. *Then the following statements holds.*

1. $\|\boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)}\| = \|\boldsymbol{v}^{(1)}\| \cdot \ldots \cdot \|\boldsymbol{v}^{(L)}\|$
2. $\|\mathcal{T} \otimes \mathcal{S}\| = \|\mathcal{T}\| \cdot \|\mathcal{S}\|$

3. *If* $\boldsymbol{U}^{(1)} \in \mathbb{K}^{I_1 \times I_1}, \ldots, \boldsymbol{U}^{(L)} \in \mathbb{K}^{I_L \times I_L}$ *are unitary (orthogonal) matrices, then* $\|(\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(L)}) \cdot \mathcal{T}\| = \|\mathcal{T}\|$.

As already noted, σ_R is not closed (except in the case of matrices). The next result shows other equivalent statements.

Theorem 1.5.9 (V. de Silva and L. H. Lim, 2008)**.** *Consider the space* $\mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ *with* $L > 2$ *and let* $R \geq 2$*. Then the following statements are equivalent.*

1. σ_R *is not closed.*
2. *There exists a sequence of tensors in* σ_R *converging to a tensor with rank greater than* R*.*
3. *There exists a tensor* $\mathcal{S}$ *of rank greater than* R *such that* $\inf_{\mathcal{T} \in \sigma_R} \|\mathcal{T} - \mathcal{S}\| = 0$.
4. *There exists a tensor* $\mathcal{S}$ *of rank greater than* R *which does note have a best rank-*R *approximation, that is,* $\inf_{\mathcal{T} \in \sigma_R} \|\mathcal{T} - \mathcal{S}\|$ *is not attained in* σ_R*.*

It is important to emphasize that there are tensors of rank greater than R which also can't be arbitrarily approximated by rank-R tensors. Figure 1.5 illustrates the possible situations one can encounter. In the figure on the top left, the dark region represents a certain subset of tensors with rank greater than R. In the top right figure, the light region represents the set σ_R. The dotted line indicates that those border points are not in σ_R, they are part of the dark region. In the figure on the bottom left we have a sequence of points in σ_R converging to the red dot, which is at the border between the dark and the light regions. This point is at the closure of σ_R, so that it is a tensor with rank greater than R which can be approximated arbitrarily well by points in σ_R. This tensor has border rank equal to R. In the figure on the bottom right the sequence of points converges to the point of the border closest to the red point, but the limit of that convergence is not the point desired. In this case we have a tensor with rank greater than R that does not have a best rank-R approximation.

Although everything we saw up to this point may indicate that the best rank-R approximation problem is something to be avoided, there are two positive results showed below.

Theorem 1.5.10 (V. de Silva and L. H. Lim, 2008)**.** *Every tensor has a best rank 1 approximation.*

Theorem 1.5.11 (V. de Silva and L. H. Lim, 2008)**.** *Let* $\mathcal{T} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ *and integers* $R_1, \ldots, R_L > 0$*. Then* $\mathcal{T}$ *does have a best approximation of multilinear rank* $\leq (R_1, \ldots, R_L)$.

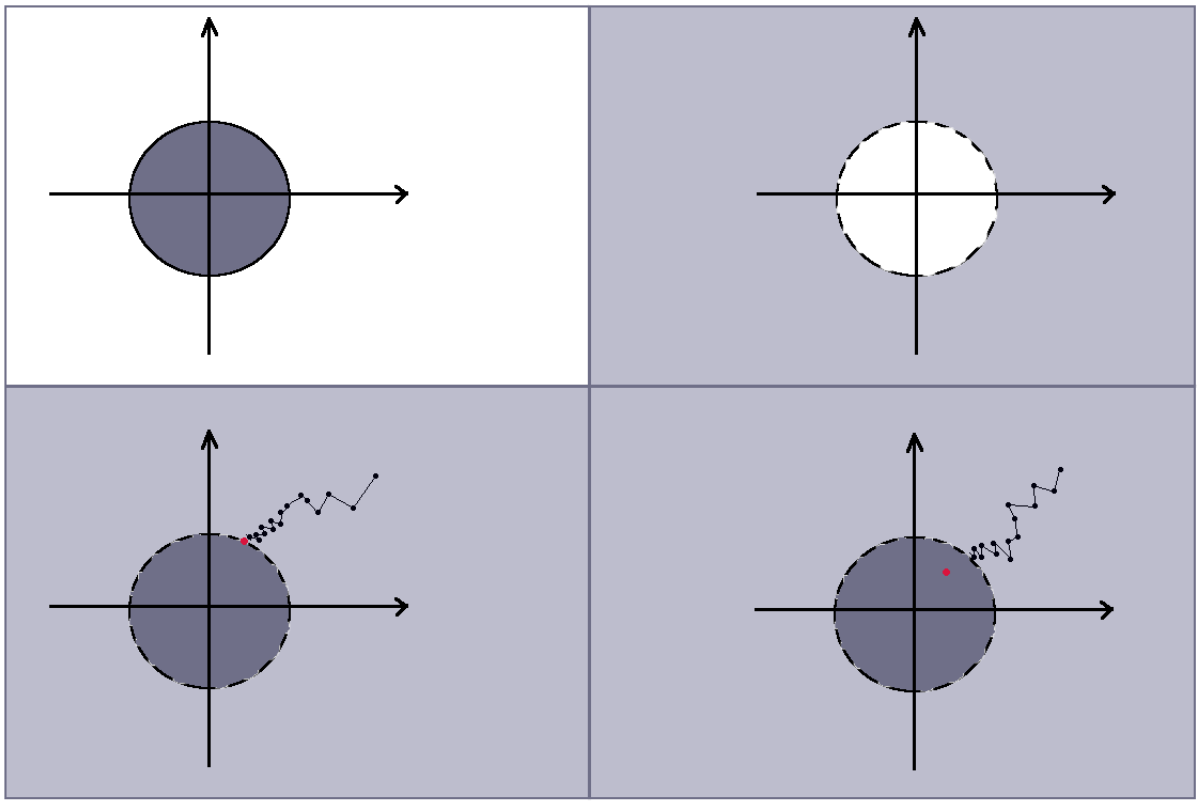

Figure 1.5: Convergence issues.

The next theorem shows one reason that can make tensor low rank approximations to fail.

Theorem 1.5.12 (V. de Silva and L. H. Lim, 2008)**.** *Let $\mathcal{T} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ a tensor with rank greater than R and let $(\mathcal{T}^{(n)})$ be a sequence in σ_R converging to $\mathcal{T}$. Furthermore, write*

$$\mathcal{T}^{(n)} = \sum_{r=1}^{R} \lambda_r^{(n)} \ \boldsymbol{v}_r^{(n,1)} \otimes \ldots \otimes \boldsymbol{v}_r^{(n,L)},$$

where each $\boldsymbol{v}_r^{(n,\ell)} \in \mathbb{K}^{I_\ell}$ is a unitary vector and $\lambda_r^{(n)} \in \mathbb{K}$ a scalar. Then there exists two distinct numbers $1 \leq r_1, r_2, \leq R$ such that $\lim_{n\to\infty} |\lambda_{r_1}^{(n)}| = \lim_{n\to\infty} |\lambda_{r_2}^{(n)}| = \infty$.

Although we have factors diverging, the sequence still converges. What happens is that these divergent factors cause cancellations as n increases. If $\mathcal{T}$ does not have a best rank-R approximation, the process of computing more approximations $\mathcal{T}^{(n)}$ can continue indefinitely, with some coefficients $\lambda_r^{(n)}$ diverging, which is a problem when making computations with finite precision. Taking away the condition of the vectors being unitary, we will have vectors diverging, which will change nothing. This phenomenon of diverging terms has been observed in practical applications of multilinear models and is referred as "degeneracy" [48, 54–57].

Chapter 2

Tensor compression

Tensor compression is an important tool to compute a CPD. It reduces the problem size, hence the computational and memory size. It relies on the computation of some SVDs of matrices, which is a familiar decomposition. Nowadays there are very fast implementations for the SVD, such as the randomized truncated SVD [31]. First we will see how to make unfoldings from tensors, then we go to multilinear rank and some related results, and finally we finish with the compression of tensors. Algorithms and their costs are shown along the way.

2.1 Multilinear multiplication

For each $\ell = 1 \dots L$, let $B^{(\ell)} = \{\mathbf{e}_1^{(\ell)}, \dots, \mathbf{e}_{I_\ell}^{(\ell)}\}$ be a basis for each space $\mathbb{K}^{I_\ell}$, and consider a tensor $\mathcal{T} \in \mathbb{K}^{I_1} \otimes \dots \otimes \mathbb{K}^{I_L}$ in coordinates $t_{i_1 \dots i_L}$. As already observed, we can use these coordinates to interpret $\mathcal{T}$ as a multidimensional array in $\mathbb{K}^{I_1 \times \dots \times I_L}$. Every time we refer to $\mathcal{T}$ as a element of $\mathbb{K}^{I_1 \times \dots \times I_L}$ it will be implicit that there are fixed bases. Matrices can act on $\mathcal{T}$ by L "distinct directions" through the usual matrix multiplication. Let $\mathbf{M}^{(1)} \in \mathbb{K}^{I'_1 \times I_1}, \dots, \mathbf{M}^{(L)} \in \mathbb{K}^{I'_L \times I_L}$ be any matrices, then we denote by $(\mathbf{M}^{(1)}, \dots, \mathbf{M}^{(L)}) \cdot \mathcal{T}$ the "multiplication" between the L-tuple $(\mathbf{M}^{(1)}, \dots, \mathbf{M}^{(L)})$ and $\mathcal{T}$. The result of this multiplication is the tensor $\mathcal{S} = (\mathbf{M}^{(1)}, \dots, \mathbf{M}^{(L)}) \cdot \mathcal{T} \in \mathbb{K}^{I'_1} \otimes \dots \otimes \mathbb{K}^{I'_L}$ defined as

$$s_{j_1 \dots j_L} = \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} m^{(1)}_{j_1 i_1} \dots m^{(L)}_{j_L i_L} t_{i_1 \dots i_L}.$$

This operation is called *multilinear multiplication*. In the particular case $\mathcal{T}$ is a matrix, we have that $(\mathbf{M}, \mathbf{N}) \cdot \mathcal{T} = \mathbf{M} \cdot \mathcal{T} \cdot \mathbf{N}^T$, and in the case $\mathcal{T}$ is a vector we have $(\mathbf{M}) \cdot \mathcal{T} = \mathbf{M} \cdot \mathcal{T}$.

As a direct consequence of the definition, the tensor $\mathcal{S}$ is given in coordinates, although we do not have defined any basis for the spaces $\mathbb{K}^{I'_1}, \dots, \mathbb{K}^{I'_L}$. The choice of these bases depends on each situation and in principle is arbitrary. The definition of the multilinear multiplication is motivated by the idea of making change of basis.

Consider the bases $B^{(\ell)} = \{\mathbf{e}_1^{(\ell)}, \ldots, \mathbf{e}_{I_\ell}^{(\ell)}\}, \tilde{B}^{(\ell)} = \{\tilde{\mathbf{e}}_1^{(\ell)}, \ldots, \tilde{\mathbf{e}}_{I_\ell}^{(\ell)}\}$ for each space $\mathbb{K}^{I_\ell}$ and let $\mathbf{M}^{(\ell)} \in \mathbb{K}^{I_\ell \times I_\ell}$ be the change of basis matrix from $\tilde{B}^{(\ell)}$ to $B^{(\ell)}$, that is, we have that

$$\mathbf{e}_i^{(\ell)} = \sum_{j=1}^{I_\ell} m_{ji}^{(\ell)} \tilde{\mathbf{e}}_j^{(\ell)}.$$

It follows that

$$\mathcal{T} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_L=1}^{I_L} t_{i_1 \ldots i_L}\ \mathbf{e}_{i_1}^{(1)} \otimes \ldots \otimes \mathbf{e}_{i_L}^{(L)} =$$

$$= \sum_{i_1=1}^{I_1} \cdots \sum_{i_L=1}^{I_L} t_{i_1 \ldots i_L} \left(\sum_{j_1=1}^{I_1} m_{j_1 i_1}^{(1)}\ \tilde{\mathbf{e}}_{j_1}^{(1)} \right) \otimes \ldots \otimes \left(\sum_{j_L=1}^{I_L} m_{j_L i_L}^{(L)}\ \tilde{\mathbf{e}}_{j_L}^{(L)} \right) =$$

$$= \sum_{i_1=1}^{I_1} \cdots \sum_{i_L=1}^{I_L} \sum_{j_1=1}^{I_1} \cdots \sum_{j_L=1}^{I_L} m_{j_1 i_1}^{(1)} \ldots m_{j_L i_L}^{(L)}\ t_{i_1 \ldots i_L}\ \tilde{\mathbf{e}}_{j_1}^{(1)} \otimes \ldots \otimes \tilde{\mathbf{e}}_{j_L}^{(L)} =$$

$$= \sum_{j_1=1}^{I_1} \cdots \sum_{j_L=1}^{I_L} s_{j_1 \ldots j_L}\ \tilde{\mathbf{e}}_{j_1}^{(1)} \otimes \ldots \otimes \tilde{\mathbf{e}}_{j_L}^{(L)},$$

where

$$s_{j_1 \ldots j_L} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_L=1}^{I_L} m_{j_1 i_1}^{(1)} \ldots m_{j_L i_L}^{(L)}\ t_{i_1 \ldots i_L}.$$

Remark 2.1.1. *The change of basis is given by the formula* $\mathcal{S} = (\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}$. *The tensor* $\mathcal{S}$ *in coordinates represents* $\mathcal{T}$ *after this change of basis. The difference between* $\mathcal{S}$ *and* $\mathcal{T}$ *is only in its representation as a multidimensional array since it depends on coordinates. But as tensors they are the same object, which we refer as the* abstract tensor. *One could be more precise and use some notation like* $\mathcal{T}_{\tilde{B}^{(1)}, \ldots, \tilde{B}^{(L)}} = (\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}_{B^{(1)}, \ldots, B^{(L)}}$. *This notation is more cumbersome and for this reason we will avoid it. Furthermore, it is not always the case that the matrices involved are change of basis matrices.*

Let $\mathcal{S} = (\mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(L)}) \cdot \mathcal{T}$. Sometimes the following equality is useful:

$$\mathcal{S} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_L=1}^{I_L} t_{i_1 \ldots i_L}\ \mathbf{M}_{:i_1}^{(1)} \otimes \ldots \otimes \mathbf{M}_{:i_L}^{(L)}$$

Denote by $GL(d, \mathbb{K})$ the linear group of matrices in $\mathbb{K}^{d \times d}$. If the field $\mathbb{K}$ is clear from the context, we just denote $GL(d)$. As we know, all change of basis matrices are invertible and every invertible matrix can be interpreted as a change of basis. This gives us a criterion of equivalence between tensors.

Definition 2.1.2. *Let two tensors $\mathcal{T}, \mathcal{S} \in \mathbb{K}^{I_1 \times \ldots \times I_L}$. We say they are* equivalent *if there are matrices $\boldsymbol{M}^{(1)} \in GL(I_1), \ldots, \boldsymbol{M}^{(L)} \in GL(I_L)$ such that $\mathcal{S} = (\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}$.*

Theorem 2.1.3. *Let two tensors $\mathcal{T}, \mathcal{S} \in \mathbb{K}^{I_1 \times \ldots \times I_L}$. They represent the same abstract tensor if, and only if, they are equivalent.*

In the case of the theorem being valid, it is possible to have $\mathcal{S} \neq \mathcal{T}$ as multidimensional arrays and $\mathcal{S} = \mathcal{T}$ as abstract tensors. This is the same situation when we have distinct matrices representing the same linear map, the difference is only due to the choice of basis. Below there are some basic properties of the multilinear multiplication.

Theorem 2.1.4 (V. de Silva and L. H. Lim, 2008)**.** *Let two tensors $\mathcal{T}, \mathcal{S} \in \mathbb{K}^{I_1 \times \ldots \times I_L}$ and the matrices $\boldsymbol{M}^{(1)} \in \mathbb{K}^{I'_1 \times I_1}$, ..., $\boldsymbol{M}^{(L)} \in \mathbb{K}^{I'_L \times I_L}$. Then*

1. *For all $\alpha, \beta \in \mathbb{K}$, we have*

$$(\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot (\alpha\mathcal{T} + \beta\mathcal{S}) = \alpha((\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}) + \beta((\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{S}).$$

2. *For all $\boldsymbol{N}^{(1)} \in \mathbb{K}^{I''_1 \times I'_1}, \ldots, \boldsymbol{N}^{(L)} \in \mathbb{K}^{I''_L \times I'_L}$, we have*

$$(\boldsymbol{N}^{(1)}, \ldots, \boldsymbol{N}^{(L)}) \cdot ((\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}) = (\boldsymbol{N}^{(1)}\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{N}^{(L)}\boldsymbol{M}^{(L)}) \cdot \mathcal{T}.$$

3. *For all $\alpha, \beta \in \mathbb{C}$ and all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{K}^{I'_j \times I_j}$, we have*

$$(\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(j-1)}, \alpha\boldsymbol{A} + \beta\boldsymbol{B}, \boldsymbol{M}^{(j+1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T} =$$

$$= \alpha((\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(j-1)}, \boldsymbol{A}, \boldsymbol{M}^{(j+1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}) + \beta((\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(j-1)}, \boldsymbol{B}, \boldsymbol{M}^{(j+1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}).$$

Theorem 2.1.5 (V. de Silva and L. H. Lim, 2008)**.** *Let $T \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_K}$ and $\boldsymbol{M}^{(1)} \in \mathbb{K}^{I'_1 \times I_1}$, ..., $\boldsymbol{M}^{(L)} \in \mathbb{K}^{I'_L \times I_L}$. Then*

1. *$rank((\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}) \leq rank(\mathcal{T})$.*
2. *If $\boldsymbol{M}^{(1)} \in GL(I'_1), \ldots, \boldsymbol{M}^{(L)} \in GL(I'_L)$, then $rank((\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}) = rank(\mathcal{T})$.*

Now let's see how the multilinear multiplication and tensor product are related.

Theorem 2.1.6 (V. de Silva and L. H. Lim, 2008)**.** *Let $\boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ be a rank one tensor and let the matrices $\boldsymbol{M}^{(1)} \in \mathbb{K}^{I'_1 \times I_1}, \ldots, \boldsymbol{M}^{(L)} \in \mathbb{K}^{I'_L \times I_L}$. Then*

$$(\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)} = (\boldsymbol{M}^{(1)}\boldsymbol{v}^{(1)}) \otimes \ldots \otimes (\boldsymbol{M}^{(L)}\boldsymbol{v}^{(L)}).$$

Corollary 2.1.7. *Let* $\mathcal{T} = \sum_{r=1}^{R} \boldsymbol{v}_r^{(1)} \otimes \ldots \otimes \boldsymbol{v}_r^{(L)} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ *be a tensor with rank* $\leq R$ *and let the matrices* $\boldsymbol{M}^{(1)} \in \mathbb{K}^{I'_1 \times I_1}, \ldots, \boldsymbol{M}^{(L)} \in \mathbb{K}^{I'_L \times I_L}$. *Then*

$$(\boldsymbol{M}^{(1)}, \ldots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T} = \sum_{r=1}^{R} (\boldsymbol{M}^{(1)} \boldsymbol{v}_r^{(1)}) \otimes \ldots \otimes (\boldsymbol{M}^{(L)} \boldsymbol{v}_r^{(L)}).$$

Theorem 2.1.8. *Let* $\mathcal{T} = \sum_{r=1}^{R} \lambda_r \ \boldsymbol{v}_r^{(1)} \otimes \ldots \otimes \boldsymbol{v}_r^{(L)} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ *be a tensor with rank* $\leq R$, *where each* λ_r *is a scalar. Then*

$$\mathcal{T} = (\ \boldsymbol{V}^{(1)}, \ldots, \ \boldsymbol{V}^{(L)}) \cdot \Lambda,$$

where $\boldsymbol{V}^{(\ell)} = [\boldsymbol{v}_1^{(\ell)}, \ldots, \boldsymbol{v}_R^{(\ell)}] \in \mathbb{K}^{I_\ell \times R}$ *for each* $\ell = 1 \ldots R$, *and* $\Lambda = diag(\lambda_r) \in \mathbb{K}^{R \times \ldots \times R}$ *is a diagonal tensor of order* L.[1]

Corollary 2.1.9. *Let* $\mathcal{T} = \sum_{r=1}^{R} \lambda_r \ \boldsymbol{x}_r \otimes \boldsymbol{y}_r \otimes \boldsymbol{z}_r \in \mathbb{K}^{I_1} \otimes \mathbb{K}^{I_2} \otimes \mathbb{K}^{I_3}$ *be a third order tensor with rank* $\leq R$, *where each* λ_r *is a scalar. Then*

$$\mathcal{T} = (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \cdot \Lambda,$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R] \in \mathbb{K}^{I_1 \times R}$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_R] \in \mathbb{K}^{I_2 \times R}$, $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_R] \in \mathbb{K}^{I_3 \times R}$ *and* $\Lambda = diag(\lambda_r) \in \mathbb{K}^{R \times R \times R}$ *is a diagonal tensor.*

2.1.1 Unfoldings

Suppose we have the mode-j fibers of a tensor $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$. These fibers are several vectors, as we've already seen. We can put them side by side and concatenate them to form a matrix. The ordering is not that important, we choose the ordering according to the order of the indexes. With this we have a matrix of shape $I_\ell \times \prod_{j \neq \ell} I_j$ called a *unfolding of* $\mathcal{T}$. We denote this unfolding by $\mathcal{T}_{(\ell)}$. Other common names are *matricization* and *flattening*. The construction of $\mathcal{T}_{(\ell)}$ with the ordering we are using can be described by the following pseudo-code. Denote by [ ] an "empty matrix" which will be filled column by column, where $\mathbf{M} \leftarrow [\mathbf{M}|\mathbf{v}]$ means to add a column vector $\mathbf{v}$ at the right of $\mathbf{M}$ and then substitute $\mathbf{M}$ by this new matrix.

Algorithm 2.1.10 (Unfolding)**.**
Input: $\mathcal{T}, \ell$

[1] These kind of tensor as sometimes called *superdiagonal*. Denote by $\lambda_{i_1 \ldots i_L}$ the entries of Λ. Then we have that $\lambda_{i_1 \ldots i_L} = \lambda_r$ if $i_1 = \ldots = i_L = r$, and $\lambda_{i_1 \ldots i_L} = 0$ otherwise.

$\mathcal{T}_{(\ell)} = [\,]$
`for` $i_L = 1 \ldots I_L$
$\quad \ddots$
$\qquad$ `for` $i_{\ell+1} = 1 \ldots I_{\ell+1}$
$\qquad\quad$ `for` $i_{\ell-1} = 1 \ldots I_{\ell-1}$
$\qquad\qquad \ddots$
$\qquad\qquad\quad$ `for` $i_1 = 1 \ldots I_1$
$\qquad\qquad\qquad \mathcal{T}_{(\ell)} \leftarrow \left[\mathcal{T}_{(\ell)} \mid \mathcal{T}_{i_1 \ldots i_{\ell-1} \,:\, i_{\ell+1} \ldots i_L}\right]$

Output: $\mathcal{T}_{(\ell)}$

Example 2.1.11. *Consider the tensor $\mathcal{T} \in \mathbb{R}^{3\times4\times2}$ given by*

$$\mathcal{T} = \left\{ \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}, \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix} \right\},$$

where this is the representation of $\mathcal{T}$ through its frontal slices. The mode-1 fibers of $\mathcal{T}$ are

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix}, \begin{bmatrix} 10 \\ 11 \\ 12 \end{bmatrix}, \begin{bmatrix} 13 \\ 14 \\ 15 \end{bmatrix}, \begin{bmatrix} 16 \\ 17 \\ 18 \end{bmatrix}, \begin{bmatrix} 19 \\ 20 \\ 21 \end{bmatrix}, \begin{bmatrix} 22 \\ 23 \\ 24 \end{bmatrix}.$$

Note that we already ordered the vectors in accord with our convention. It follows that

$$\mathcal{T}_{(1)} = \begin{bmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{bmatrix}.$$

The mode-2 fibers are

$$\begin{bmatrix} 1 \\ 4 \\ 7 \\ 10 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \\ 8 \\ 11 \end{bmatrix}, \begin{bmatrix} 3 \\ 6 \\ 9 \\ 12 \end{bmatrix}, \begin{bmatrix} 13 \\ 16 \\ 19 \\ 22 \end{bmatrix}, \begin{bmatrix} 14 \\ 17 \\ 20 \\ 23 \end{bmatrix}, \begin{bmatrix} 15 \\ 18 \\ 21 \\ 24 \end{bmatrix}.$$

It follows that

$$\mathcal{T}_{(2)} = \begin{bmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{bmatrix}.$$

Finally, the mode-3 fibers are

$$\begin{bmatrix} 1 \\ 13 \end{bmatrix}, \begin{bmatrix} 2 \\ 14 \end{bmatrix}, \begin{bmatrix} 3 \\ 15 \end{bmatrix}, \begin{bmatrix} 4 \\ 16 \end{bmatrix}, \begin{bmatrix} 5 \\ 17 \end{bmatrix}, \begin{bmatrix} 6 \\ 18 \end{bmatrix}, \begin{bmatrix} 7 \\ 19 \end{bmatrix}, \begin{bmatrix} 8 \\ 20 \end{bmatrix}, \begin{bmatrix} 9 \\ 21 \end{bmatrix}, \begin{bmatrix} 10 \\ 22 \end{bmatrix}, \begin{bmatrix} 11 \\ 23 \end{bmatrix}, \begin{bmatrix} 12 \\ 24 \end{bmatrix}.$$

It follows that

$$\mathcal{T}_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \end{bmatrix}.$$

We can note that $\mathcal{T}_{(1)} \in \mathbb{R}^{3\times 4\cdot 2}$, $\mathcal{T}_{(2)} \in \mathbb{R}^{4\times 3\cdot 2}$, $\mathcal{T}_{(3)} \in \mathbb{R}^{2\times 3\cdot 4}$, *as expected.*

Unfoldings and multilinear multiplication are highly connected. The idea is that we can multiply an unfolding by some matrix and consider the result as the unfolding of a new tensor.

Definition 2.1.12. *Let a tensor* $\mathcal{T} \in \mathbb{K}^{I_1\times\ldots\times I_L}$ *and a matrix* $\boldsymbol{M} \in \mathbb{K}^{I'_\ell \times I_\ell}$. *The* product mode-$\ell$ *between* $\mathcal{T}$ *and* $\boldsymbol{M}$ *is the tensor* $\mathcal{S} \in \mathbb{K}^{I_1\times\ldots\times I_{\ell-1}\times I'_\ell \times I_{\ell+1}\times\ldots\times I_L}$ *such that* $\mathcal{S}_{(\ell)} = \boldsymbol{M}\cdot\mathcal{T}_{(\ell)}$. *We denote* $\mathcal{S} = \mathcal{T}\times_\ell \boldsymbol{M}$.

Although the symbol $\times_\ell$ is at the right of the tensor, in the actual multiplication it comes at the left, that is, this is a left action on the tensor space. This is just a notational convention and should not cause confusion. Furthermore, we will omit parenthesis when making more than one of these products. More precisely, we will write $\mathcal{T}\times_\ell \mathbf{M}\times_{\ell'}\mathbf{N}$ instead of $(\mathcal{T}\times_\ell \mathbf{M})\times_{\ell'}\mathbf{N}$. It is possible to obtain $\mathcal{S}$ explicitly in coordinates, this gives

$$\mathcal{S}_{i_1\ldots i_{\ell-1}\, i'_\ell\, i_{\ell+1}\ldots i_L} = \sum_{k=1}^{I'_\ell} m_{i'_\ell k}\cdot t_{i_1\ldots i_{\ell-1}\, k\, i_{\ell+1}\ldots i_L}$$

for all $i'_\ell = 1\ldots I'_\ell$. This formula remind us the formula of the multilinear multiplication. Indeed there is a connection. Let a tensor $T \in \mathbb{K}^{I_1\times\ldots\times I_L}$ and matrices $\mathbf{M}^{(1)} \in \mathbb{K}^{I'_1\times I_1}$, …, $\mathbf{M}^{(L)} \in \mathbb{K}^{I'_L\times I_L}$. Then

$$(\mathbf{M}^{(1)},\ldots,\mathbf{M}^{(L)})\cdot\mathcal{T} = \mathcal{T}\times_1\mathbf{M}^{(1)}\times_2\mathbf{M}^{(2)}\ldots\times_L\mathbf{M}^{(L)}.$$

This relationship gives us a clearer picture of how the multilinear multiplication acts on tensors. Each $\mathbf{M}^{(\ell)}$ multiplies all the mode-j fibers of $\mathcal{T}$ (which in the end is equivalent to multiply $\mathbf{M}^{(\ell)}$ by $\mathcal{T}_{(\ell)}$), and thus we obtain a new tensor. The next example clarifies how works this relationship more concretely.

Example 2.1.13. *Let* $\mathcal{T}$ *be the tensor of the previous example and let* $\boldsymbol{M} = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$. *By definition, we have that* $\mathcal{T}\times_1\boldsymbol{M} \in \mathbb{R}^{2\times 4\times 2}$, *with*

$$(\mathcal{T}\times_1\boldsymbol{M})_{(1)} = \boldsymbol{M}\cdot\mathcal{T}_{(1)} =$$

$$= \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{bmatrix} =$$

$$= \begin{bmatrix} 22 & 49 & 76 & 103 & 130 & 157 & 184 & 211 \\ 28 & 64 & 100 & 136 & 172 & 208 & 244 & 280 \end{bmatrix}.$$

This is matrix of order $2 \times (4 \cdot 2) = 2 \times 8$, so that the first half of the matrix is the first frontal slice of $\mathcal{T} \times_1 \boldsymbol{M}$. From this we conclude that

$$\mathcal{T} \times_1 \boldsymbol{M} = \left\{ \begin{bmatrix} 22 & 49 & 76 & 103 \\ 28 & 64 & 100 & 136 \end{bmatrix}, \begin{bmatrix} 130 & 157 & 184 & 211 \\ 172 & 208 & 244 & 280 \end{bmatrix} \right\}.$$

To make this example simple, we will only consider the action of the matrix $\boldsymbol{M}$. Note that $\mathcal{T} \times_1 \boldsymbol{M} = \mathcal{T} \times_1 \boldsymbol{M} \times_2 \boldsymbol{I}_4 \times_3 \boldsymbol{I}_2 = (\boldsymbol{M}, \boldsymbol{I}_4, \boldsymbol{I}_2) \cdot \mathcal{T}$. Now let $\mathcal{S} = (\boldsymbol{M}, \boldsymbol{I}_4, \boldsymbol{I}_2) \cdot \mathcal{T}$. From the multilinear multiplication definition we have that

$$\mathcal{S}_{i'j'k'} = \sum_{i=1}^{3} \sum_{j=1}^{4} \sum_{k=1}^{2} m_{i'i} \cdot (\boldsymbol{I}_4)_{j'j} \cdot (\boldsymbol{I}_2)_{k'k} \cdot t_{ijk} = \sum_{i=1}^{3} m_{i'i} \cdot t_{ij'k'} = \boldsymbol{M}_{i':} \cdot T_{:j'k'}.$$

This last expression is the product of the i'-th row of $\boldsymbol{M}$ by the column vector of $\mathcal{T}$ obtained by fixing j', k' and varying the rows, that is, a mode-1 fiber. By varying i' to form a mode-1 fiber of $\mathcal{S}$ we obtain

$$\mathcal{S}_{:j'k'} = \begin{bmatrix} t_{1j'k'} \\ t_{2j'k'} \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_{1:} \cdot \mathcal{T}_{:j'k'} \\ \boldsymbol{M}_{2:} \cdot \mathcal{T}_{:j'k'} \end{bmatrix} = \begin{bmatrix} \boldsymbol{M}_{1:} \\ \boldsymbol{M}_{2:} \end{bmatrix} \cdot \mathcal{T}_{:j'k'} = \boldsymbol{M} \cdot \mathcal{T}_{:j'k'}$$

From this we conclude that

$$\mathcal{S}_{(1)} = [\mathcal{S}_{:11}, \mathcal{S}_{:21}, \mathcal{S}_{:31}, \mathcal{S}_{:41}, \mathcal{S}_{:12}, \mathcal{S}_{:22}, \mathcal{S}_{:32}, \mathcal{S}_{:42}] =$$

$$= [\boldsymbol{M}\mathcal{T}_{:11}, \boldsymbol{M}\mathcal{T}_{:21}, \boldsymbol{M}\mathcal{T}_{:31}, \boldsymbol{M}\mathcal{T}_{:41}, \boldsymbol{M}\mathcal{T}_{:12}, \boldsymbol{M}\mathcal{T}_{:22}, \boldsymbol{M}\mathcal{T}_{:32}, \boldsymbol{M}\mathcal{T}_{:42}] =$$

$$= \boldsymbol{M} \cdot [\mathcal{T}_{:11}, \mathcal{T}_{:21}, \mathcal{T}_{:31}, \mathcal{T}_{:41}, \mathcal{T}_{:12}, \mathcal{T}_{:22}, \mathcal{T}_{:32}, \mathcal{T}_{:42}] =$$

$$= \boldsymbol{M} \cdot \mathcal{T}_{(1)}.$$

2.1.2 Multilinear rank

A natural thing to do is to consider the rank of unfoldings.

Definition 2.1.14. *For each $\ell = 1 \dots L$, the* mode-ℓ rank of $\mathcal{T}$ *is the rank of $\mathcal{T}_{(\ell)}$. We denote this rank by $rank_{(\ell)}(\mathcal{T})$.*

Definition 2.1.15. *The* multilinear rank of $\mathcal{T}$ *is the L-tuple $\big(rank_{(1)}(\mathcal{T}), \dots, rank_{(L)}(\mathcal{T})\big)$. We denote this rank by $rank_{\boxplus}(\mathcal{T})$.*

The multilinear rank is also often called the *Tucker rank* of $\mathcal{T}$. Now we will see some important results regarding this rank.

Theorem 2.1.16 (V. de Silva and L. H. Lim, 2008)**.** *Let $\mathcal{T} \in \mathbb{K}^{I_1} \otimes \dots \otimes \mathbb{K}^{I_L}$ and $\boldsymbol{M}^{(1)} \in \mathbb{K}^{I'_1 \times I_1}, \dots, \boldsymbol{M}^{(L)} \in \mathbb{K}^{I'_L \times I_L}$. Then the following statements holds.*

1. *The multilinear rank doesn't depend on the field being real or complex.*
2. *$rank_{\boxplus}(\mathcal{T}) = (1, 1, \dots, 1)$ if, and only if, $rank(\mathcal{T}) = 1$.*
3. *$rank_{(\ell)}(\mathcal{T}) \leq \min\{rank(\mathcal{T}), I_1, \dots, I_L\}$ for all $\ell = 1 \dots L$.*
4. *$\|rank_{\boxplus}(\mathcal{T})\|_\infty \leq rank(T)$.*
5. *$rank_{\boxplus}((\boldsymbol{M}^{(1)}, \dots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}) \leq rank_{\boxplus}(\mathcal{T})$.*
6. *If $\boldsymbol{M}^{(1)} \in GL(I'_1), \dots, \boldsymbol{M}^{(L)} \in GL(I'_L)$, then $rank_{\boxplus}((\boldsymbol{M}^{(1)}, \dots, \boldsymbol{M}^{(L)}) \cdot \mathcal{T}) = rank_{\boxplus}(\mathcal{T})$.*

Remark 2.1.17. *Consider a tensor in coordinates $\mathcal{T} \in \mathbb{K}^{I_1 \times \dots \times I_L}$ such that $rank_{\boxplus}(\mathcal{T}) = (R_1, \dots, R_L)$. Theorem 2.1.16-6 together with theorem 2.1.5-2 makes it possible to study rank properties of $\mathcal{T}$ as a tensor in $\mathbb{K}^{R_1 \times \dots \times R_L}$.*

Theorem 2.1.18 (V. de Silva and L. H. Lim, 2008)**.** *Let $\mathcal{T} \in \mathbb{K}^{I_1 \times \dots \times I_L}$ be a tensor such that $rank_{\boxplus}(\mathcal{T}) \leq (R_1, \dots, R_L)$. Then there there exists full rank matrices $\boldsymbol{M}^{(1)} \in \mathbb{K}^{I_1 \times R_1}$, …, $\boldsymbol{M}^{(L)} \in \mathbb{K}^{I_L \times R_L}$ and a tensor $\mathcal{S} \in \mathbb{K}^{R_1 \times \dots \times R_L}$ such that $\mathcal{T} = (\boldsymbol{M}^{(1)}, \dots, \boldsymbol{M}^{(L)}) \cdot \mathcal{S}$ and $rank(\mathcal{T}) = rank(\mathcal{S})$.*

2.2 Compressing with the multilinear singular value decomposition

2.2.1 Tucker decomposition

Definition 2.2.1. *Let $\mathcal{T} \in \mathbb{K}^{I_1 \dots I_L}$ be a tensor. Then a* Tucker decomposition of $\mathcal{T}$ *is a decomposition of the form $\mathcal{T} = (\boldsymbol{M}^{(1)}, \dots, \boldsymbol{M}^{(L)}) \cdot \mathcal{S}$, where $\boldsymbol{M}^{(\ell)} \in \mathbb{K}^{I'_\ell \times I_\ell}$ for $\ell = 1 \dots L$, and $\mathcal{S} \in \mathbb{K}^{I'_1 \times \dots \times I'_L}$.*

Each matrix $\mathbf{M}^{(\ell)}$ is called a *factor matrix* and the tensor $\mathcal{S}$ is called the *core tensor*. Usually one defines this decomposition assuming all factor matrices are unitary (orthogonal), but we will prefer the more general definition. This way the CPD may be seen as a particular case of the Tucker decomposition.

The Tucker decomposition was first introduced in 1963 and refined subsequently [58, 59]. This decomposition can be considered as a form of high-order PCA (Principal Component Analysis)[27] when the core tensor has lower dimensions than the original one. In this case we consider $\mathcal{S}$ as a compressed form of $\mathcal{T}$. Usually the CPD is indicated for latent parameter estimation and the Tucker decomposition is indicated for compression and dimensionality reduction.

From what we've seen on multilinear multiplication, we can write a Tucker decomposition $\mathcal{T} = (\mathbf{M}^{(1)}, \dots, \mathbf{M}^{(L)}) \cdot \mathcal{S}$ as

$$\mathcal{T} = \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} s_{i_1 \dots i_L} \, \mathbf{M}^{(1)}_{:i_1} \otimes \dots \otimes \mathbf{M}^{(L)}_{:i_L}.$$

If $\mathcal{S} = \operatorname{diag}(s_r) \in \mathbb{K}^{R \times \dots \times R}$ is diagonal and $\mathbf{M}^{(\ell)} \in \mathbb{K}^{I_\ell \times R}$, then we have a rank-$R$ CPD of $\mathcal{T}$ and we can write

$$\mathcal{T} = \sum_{r=1}^{R} s_r \, \mathbf{M}^{(1)}_{:r} \otimes \dots \otimes \mathbf{M}^{(L)}_{:r}.$$

We can see that the factor matrices of the Tucker decomposition agrees with the factor matrices of the CPD. On the other extreme, let $\{\mathbf{e}^{(\ell)}_1, \dots, \mathbf{e}^{(\ell)}_{I_\ell}\}$ be a basis of $\mathbb{K}^{I_\ell}$ for each $\ell = 1 \dots L$. In this case we are able to write $\mathcal{T}$ as

$$\mathcal{T} = \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} t_{i_1 \dots i_L} \, \mathbf{e}^{(1)}_{i_1} \otimes \dots \otimes \mathbf{e}^{(L)}_{i_L}.$$

Denoting $\mathbf{E}^{(\ell)} = [\mathbf{e}^{(\ell)}_1, \dots, \mathbf{e}^{(\ell)}_L] \in \mathbb{K}^{I_\ell \times I_\ell}$, we can write $\mathcal{T} = (\mathbf{E}^{(1)}, \dots, \mathbf{E}^{(L)}) \cdot \mathcal{T}$. This is a trivial Tucker decomposition, whereas the CPD can be seen as the "ultimate" Tucker decomposition. Between these two there are other useful decompositions to consider.

Theorem 2.2.2 (T. G. Kolda, 2006)**.** *Let $\mathcal{T} \in \mathbb{K}^{I_1 \dots I_L}$ be a tensor with a Tucker decomposition given by $\mathcal{T} = (\boldsymbol{M}^{(1)}, \dots, \boldsymbol{M}^{(L)}) \cdot \mathcal{S}$. Then, for each $\ell = 1 \dots L$, the unfolding $\mathcal{T}_{(\ell)}$ is given by*

$$\mathcal{T}_{(\ell)} = \boldsymbol{M}^{(\ell)} \cdot \mathcal{S}_{(\ell)} \cdot (\boldsymbol{M}^{(L)} \tilde{\otimes} \dots \tilde{\otimes} \boldsymbol{M}^{(\ell+1)} \tilde{\otimes} \boldsymbol{M}^{(\ell-1)} \tilde{\otimes} \dots \tilde{\otimes} \boldsymbol{M}^{(1)})^T,$$

where $\tilde{\otimes}$ is the Kronecker product*, see appendix B.*

Corollary 2.2.3. *Let $\mathcal{T} \in \mathbb{K}^{I_1 \times I_2 \times I_3}$ be a third order tensor with a CPD given by $\mathcal{T} =$*

$\sum_{r=1}^{R} \lambda_r \ \boldsymbol{x}_r \otimes \boldsymbol{y}_r \otimes \boldsymbol{z}_r = (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \cdot \Lambda$, *where* $\Lambda = diag(\lambda_r) \in \mathbb{K}^{R \times R \times R}$. *Then*

$$\Lambda_{(1)} = \Lambda_{(2)} = \Lambda_{(3)} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \lambda_2 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ & \vdots & & & & \vdots & & & \dots & & \vdots & & \\ \underbrace{0 \ 0 \ \dots \ 0}_{R \text{ columns}} & & & & \underbrace{0 \ 0 \ \dots \ 0}_{R \text{ columns}} & & & & \dots & \underbrace{0 \ 0 \ \dots \ \lambda_R}_{R \text{ columns}} & & & \end{bmatrix}$$

and, denoting this matrix by $[\Lambda]$, *we also have that*

$$\mathcal{T}_{(1)} = \boldsymbol{X} \cdot [\Lambda] \cdot (\boldsymbol{Z} \tilde{\otimes} \boldsymbol{Y})^T,$$

$$\mathcal{T}_{(2)} = \boldsymbol{Y} \cdot [\Lambda] \cdot (\boldsymbol{Z} \tilde{\otimes} \boldsymbol{X})^T,$$

$$\mathcal{T}_{(3)} = \boldsymbol{Z} \cdot [\Lambda] \cdot (\boldsymbol{Y} \tilde{\otimes} \boldsymbol{X})^T.$$

Theorem 2.2.4 (T. G. Kolda, 2006)**.** *Let* $\mathcal{T} \in \mathbb{K}^{I_1 \dots I_L}$ *be a tensor with a Tucker decomposition given by* $\mathcal{T} = (\boldsymbol{M}^{(1)}, \dots, \boldsymbol{M}^{(L)}) \cdot \mathcal{I}_{R \times \dots \times R}$. *Then, for each* $\ell = 1 \dots L$, *the unfolding* $\mathcal{T}_{(\ell)}$ *is given by*

$$\mathcal{T}_{(\ell)} = \boldsymbol{M}^{(\ell)} \cdot (\boldsymbol{M}^{(L)} \odot \dots \odot \boldsymbol{M}^{(\ell+1)} \odot \boldsymbol{M}^{(\ell-1)} \odot \dots \odot \boldsymbol{M}^{(1)})^T.$$

2.2.2 Multilinear singular value decomposition

Now we will see how to generalize the SVD to tensors. This generalization is called *multilinear singular value decomposition (MLSVD)*, but sometimes it is also called *High order singular value decomposition (HOSVD)*. Some texts even call this as being the Tucker decomposition. This generalization has been investigated in psychometrics [59] as the *Tucker model*, which basically was a special Tucker decomposition of third order tensors. The first work to formalize this as a high order singular value decomposition is [30].

Let $\mathbf{M}$ be a matrix with SVD given by $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^*$. We can rewrite this equation using the multilinear multiplication notation, then we obtain $\mathbf{M} = (\mathbf{U}, \overline{\mathbf{V}}) \cdot \Sigma$, where $\overline{\mathbf{V}}$ is the conjugate of $\mathbf{V}$ coordinatewise. Denoting $\mathbf{U}^{(1)} = \mathbf{U}$ and $\mathbf{U}^{(2)} = \overline{\mathbf{V}}$, we can write $\mathbf{M} = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) \cdot \Sigma$. The MLSVD is a generalization of this observation.

Given a tensor $\mathcal{S} \in \mathbb{K}^{I_1 \times \dots \times I_L}$, let $\mathcal{S}_{i_\ell = k} \in \mathbb{K}^{I_1 \times \dots \times I_{\ell-1} \times I_{\ell+1} \times \dots \times I_L}$ be the subtensor of $\mathcal{S}$ obtained by fixing the ℓ-th index of $\mathcal{S}$ with value equal to k and varying all the other indexes. More precisely, $\mathcal{S}_{i_\ell = k} = \mathcal{S}_{:,\dots,k,:,\dots,:}$, where the value k is at the ℓ-th index. We call these subtensors by *hyperslices*. In the case of a third order tensors, these subtensors are the slices described in 1.3. $\mathcal{S}_{i_1 = k}$ are the horizontal slices, $\mathcal{S}_{i_2 = k}$ the lateral slices and $\mathcal{S}_{i_3 = k}$ the frontal slices.

Theorem 2.2.5 (L. De Lathauwer, B. De Moor, J. Vandewalle, 2000)**.** *Let $\mathcal{T} \in \mathbb{K}^{I_1 \times \ldots \times I_L}$ be arbitrary. Then there exists unitary (orthogonal) matrices $\boldsymbol{U}^{(1)} \in \mathbb{K}^{I_1 \times I_1}, \ldots, \boldsymbol{U}^{(L)} \in \mathbb{K}^{I_L \times I_L}$ and a tensor $\mathcal{S} \in \mathbb{K}^{I_1 \times \ldots \times I_L}$ such that*

1. $\mathcal{T} = (\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(L)}) \cdot \mathcal{S}$.
2. *For all $\ell = 1 \ldots L$, the subtensors $\mathcal{S}_{i_\ell=1}, \ldots, \mathcal{S}_{i_\ell=I_\ell}$ are orthogonal with respect to each other.*
3. *For all $\ell = 1 \ldots L$, $\|\mathcal{S}_{i_\ell=1}\| \geq \ldots \geq \|\mathcal{S}_{i_\ell=I_\ell}\|$.*

Proof: For each unfolding $\mathcal{T}_{(\ell)}$, consider the corresponding reduced SVD

$$\mathcal{T}_{(\ell)} = \mathbf{U}^{(\ell)} \cdot \Sigma^{(\ell)} \cdot (\mathbf{V}^{(\ell)})^*$$

where $\Sigma^{(\ell)} = \operatorname{diag}\big(\sigma_1^{(\ell)}, \ldots, \sigma_{I_\ell}^{(\ell)}\big) \in \mathbb{R}^{I_\ell \times I_\ell}$ and $\mathbf{U} \in \mathbb{K}^{I_\ell \times I_\ell}, \mathbf{V} \in \mathbb{K}^{\left(\prod_{j \neq \ell} I_j\right) \times I_\ell}$ are unitary (orthogonal) matrices. Let $\mathcal{S} \in \mathbb{K}^{I_1 \times \ldots I_L}$ be the tensor defined as

$$\mathcal{S} = \big(\big(\mathbf{U}^{(1)}\big)^*, \ldots, \big(\mathbf{U}^{(L)}\big)^*\big) \cdot \mathcal{T}.$$

By theorem 2.2.2 we know that

$$\mathcal{S}_{(\ell)} = (\mathbf{U}^{(\ell)})^* \cdot \mathcal{T}_{(\ell)} \cdot \big(\big(\mathbf{U}^{(L)}\big)^* \tilde{\otimes} \ldots \tilde{\otimes} \big(\mathbf{U}^{(\ell+1)}\big)^* \tilde{\otimes} \big(\mathbf{U}^{(\ell-1)}\big)^* \tilde{\otimes} \ldots \tilde{\otimes} \big(\mathbf{U}^{(1)}\big)^*\big)^T =$$

$$= (\mathbf{U}^{(\ell)})^* \cdot \mathcal{T}_{(\ell)} \cdot \left(\overline{\mathbf{U}^{(L)}} \tilde{\otimes} \ldots \tilde{\otimes} \overline{\mathbf{U}^{(\ell+1)}} \tilde{\otimes} \overline{\mathbf{U}^{(\ell-1)}} \tilde{\otimes} \ldots \tilde{\otimes} \overline{\mathbf{U}^{(1)}}\right) =$$

$$= \Sigma^{(\ell)} \cdot (\mathbf{V}^{(\ell)})^* \cdot \left(\overline{\mathbf{U}^{(L)}} \tilde{\otimes} \ldots \tilde{\otimes} \overline{\mathbf{U}^{(\ell+1)}} \tilde{\otimes} \overline{\mathbf{U}^{(\ell-1)}} \tilde{\otimes} \ldots \tilde{\otimes} \overline{\mathbf{U}^{(1)}}\right).$$

Denote by $(\mathcal{S}_{(\ell)})_{k:}$ the k-th row of $\mathcal{S}_{(\ell)}$. From the definition we have that $(\mathcal{S}_{(\ell)})_{k:} = \mathcal{S}_{i_\ell=k}$. Now let $1 \leq k, k' \leq I_\ell$ distinct, then we have $\langle \mathcal{S}_{i_\ell=k}, \mathcal{S}_{i_\ell=k'} \rangle = 0$ because $\mathcal{S}_{(\ell)}$ has orthonormal rows. This last assertion follows from the fact that $\mathcal{S}_{(\ell)}$ is the product of a diagonal matrix by two of unitary (orthogonal) matrices. Finally, note that $\|\mathcal{S}_{i_\ell=k}\| = |\sigma_k^{(\ell)}| = \sigma_k^{(\ell)}$. Hence it follows that $\|\mathcal{S}_{i_\ell=1}\| \geq \ldots \geq \|\mathcal{S}_{i_\ell=I_\ell}\|$. □

Property 2 is called *all-orthogonality*. It means that $\langle \mathcal{S}_{i_\ell=k}, \mathcal{S}_{i_\ell=k'} \rangle = 0$ for all $k, k' = 1 \ldots I_\ell$ and $k \neq k'$. The inner product considered is the one mentioned at the beginning of the tensor geometry section. The ordering given in property 3 can be considered as a convention, made to fix a particular ordering of the columns of the matrices $\mathbf{U}^{(\ell)}$. Additionally, this ordering gives us a more precise parallel to the SVD. We will denote $\sigma_k^{(\ell)} = \|\mathcal{S}_{i_\ell=k}\|$ and call this a *singular value mode-ℓ of $\mathcal{T}$*, while each column vector $\mathbf{U}_{:j}^{(\ell)}$ is called a *singular vector mode-ℓ of $\mathcal{T}$*. An algorithm for computing the MLSVD is presented below.

Algorithm 2.2.6 (MLSVD).
Input: $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_L}$

`for` $\ell = 1 \ldots L$
$\quad \mathcal{T}_{(\ell)} \leftarrow$ `Unfolding` $(\mathcal{T}, \ell)$
$\quad \boldsymbol{U}^{(\ell)}, \Sigma^{(\ell)}, \left(\boldsymbol{V}^{(\ell)}\right)^* \leftarrow$ `SVD` $\left(\mathcal{T}_{(\ell)}\right)$
$\mathcal{S} \leftarrow \left(\left(\boldsymbol{U}^{(1)}\right)^*, \ldots, \left(\boldsymbol{U}^{(L)}\right)^*\right) \cdot \mathcal{T}$

Output: $\boldsymbol{U}^{(\ell)}, \Sigma^{(\ell)} \in \mathbb{R}^{I_\ell \times I_\ell}$ *for* $\ell = 1 \ldots L$ *and* $\mathcal{S} \in \mathbb{R}^{I_1 \times \ldots \times I_L}$

The core tensor $\mathcal{S}$ of the MLSVD distributes the "energy" (i.e., the magnitude of its entries) in such a way so that it concentrates more energy around the first entry $s_{11\ldots1}$ and disperses as we move along each dimension. Figure 2.1 illustrates the energy distribution when $\mathcal{S}$ is a third order tensor. The red slices contains more energy and it changes to white when the slice contains less energy. Note that the energy of the slices are given precisely by the singular values.

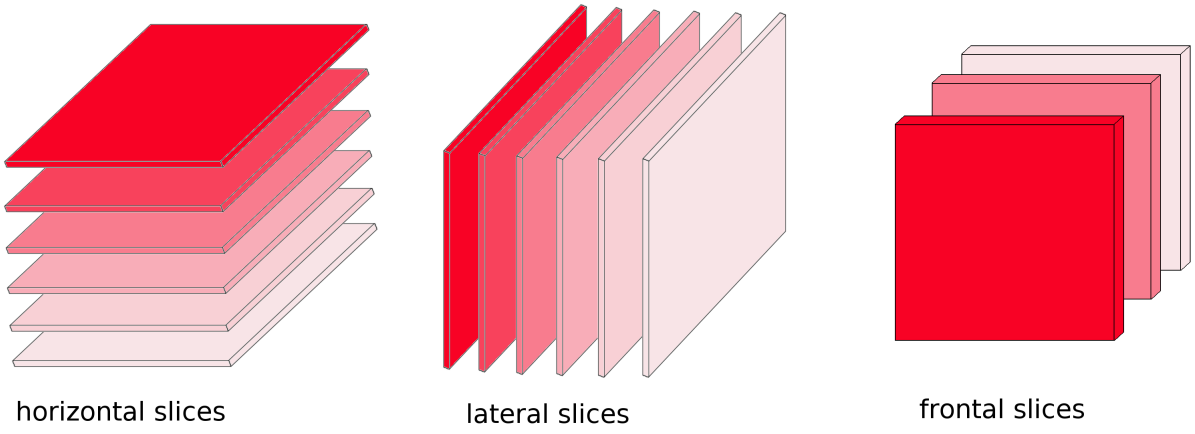

Figure 2.1: Energy of the slices of a core third order tensor $\mathcal{S}$ obtained after a MLSVD.

Theorem 2.2.7 (L. De Lathauwer, B. De Moor, J. Vandewalle, 2000). *Let* $\mathcal{S}_{(\ell)}$ *be an unfolding of* $\mathcal{S}$. *If* $(\mathcal{S}_{(\ell)})_{k:}$ *is the* k*-th row of* $\mathcal{S}_{(\ell)}$, *then* $\|(\mathcal{S}_{(\ell)})_{k:}\| = \sigma_k^{(\ell)}$.

Let $\mathcal{T} = (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(L)}) \cdot \mathcal{S}$ be the MLSVD of $\mathcal{T}$. Define the matrix $\Sigma^{(\ell)} \in \mathbb{K}^{I_\ell \times I_\ell}$ given by $\Sigma^{(\ell)} = \operatorname{diag}\left(\sigma_1^{(\ell)}, \ldots, \sigma_{I_\ell}^{(\ell)}\right)$, and let $\tilde{\mathcal{S}}_{(\ell)} \in \mathbb{K}^{I_\ell \times I_\ell}$ be defined by the relation $\mathcal{S}_{(\ell)} = \Sigma^{(\ell)} \cdot \tilde{\mathcal{S}}_{(\ell)}$. Notice that $\tilde{\mathcal{S}}_{(\ell)}$ is the normalized version of $\mathcal{S}_{(\ell)}$ (it has unit-length rows). In the case none of the $\sigma_k^{(\ell)}$ is null, we can write $\tilde{\mathcal{S}}_{(\ell)} = \left(\Sigma^{(\ell)}\right)^{-1} \cdot \mathcal{S}_{(\ell)}$. Finally, define the matrix $\mathbf{V}^{(\ell)} \in \mathbb{K}^{\left(\prod_{j \neq \ell} I_j\right) \times I_\ell}$ given by the relation

$$(\mathbf{V}^{(\ell)})^* = \tilde{\mathcal{S}}_{(\ell)} \cdot \left(\overline{\mathbf{U}^{(L)}} \tilde{\otimes} \ldots \tilde{\otimes} \overline{\mathbf{U}^{(\ell+1)}} \tilde{\otimes} \overline{\mathbf{U}^{(\ell-1)}} \tilde{\otimes} \ldots \tilde{\otimes} \overline{\mathbf{U}^{(1)}}\right).$$

With these notations and the formula for $\mathcal{T}_{(\ell)}$ (theorem 2.2.2) we can write

$$\mathcal{T}_{(\ell)} = \mathbf{U}^{(\ell)} \cdot \Sigma^{(\ell)} \cdot (\mathbf{V}^{(\ell)})^*.$$

Theorem 2.2.8 (L. De Lathauwer, B. De Moor, J. Vandewalle, 2000)**.** *With the notations above,* $\mathcal{T}_{(\ell)} = \boldsymbol{U}^{(\ell)} \cdot \Sigma^{(\ell)} \cdot (\boldsymbol{V}^{(\ell)})^*$ *is a SVD for* $\mathcal{T}_{(\ell)}$*, for each* $\ell = 1 \ldots L$*.*

With this theorem we can see that the MLSVD is a reasonable extension of the SVD, because one property desired for such an extension is to be able to use the MLSVD to obtain the SVD of each unfolding of $\mathcal{T}$. Soon we will give more reasons to consider this as a good extension of the SVD.

In the MLSVD, just as in the SVD, it is possible to have the last singular values mode-ℓ equal to zero, that is, it can exist a number $1 \leq R_\ell \leq I_\ell$ such that $\sigma_k^{(\ell)} = 0$ for $k = R_\ell + 1 \ldots I_\ell$. Therefore all hyperslices $\mathcal{S}_{i_\ell = k}$ are null (all of its entries are zero) for $k = R_\ell + 1 \ldots I_\ell$. Consequently, if we have $i_\ell > R_\ell$ for a multi-index $(i_1, \ldots, i_L)$, then $s_{i_1 \ldots i_L} = 0$. With this, we can write the tensor $\mathcal{T}$ as

$$\mathcal{T} = \sum_{i_1=1}^{R_1} \cdots \sum_{i_L=1}^{R_L} s_{i_1 \ldots i_L} \, \mathbf{U}^{(1)}_{:i_1} \otimes \ldots \otimes \mathbf{U}^{(L)}_{:i_L},$$

which shows $\mathcal{T}$ as a sum of $\prod_{\ell=1}^{L} R_\ell$ rank one terms. We illustrate the form of $\mathcal{S}$ in the figure below as a third order core tensor. The gray part corresponds to the part of nonzero values, while the white part consists only of zeros.

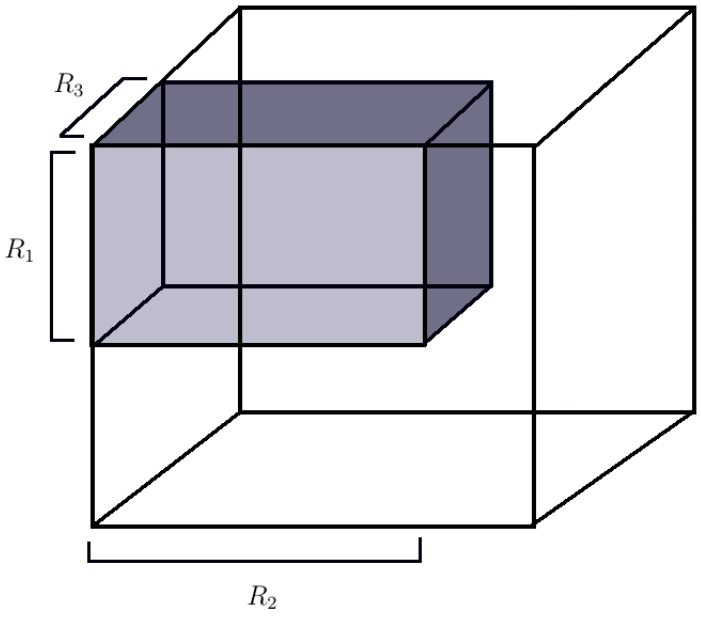

Figure 2.2: Representation of the null slices slices of core third order tensor $\mathcal{S}$ obtained after a MLSVD.

As a consequence of the above observation and theorem 2.1.16-4, we obtain

$$\max_{\ell} R_\ell \leq rank(\mathcal{T}) \leq \prod_{\ell=1}^{L} R_\ell,$$

which gives us an upper and lower bound for the rank of $\mathcal{T}$.

Instead of considering $\mathcal{S}$ as a tensor in $\mathbb{K}^{I_1 \times \ldots \times I_L}$ with lots of zeros, we can focus only on the dense part, hence we consider $\mathcal{S}$ as a tensor in $\mathbb{K}^{R_1 \times \ldots \times R_L}$. This possibility had already been discussed shortly in 2.1.17. In doing so, we can also discard the last $I_\ell - R_\ell$ columns of each $\mathbf{U}^{(\ell)}$, hence we have $\mathbf{U}^{(\ell)} \in \mathbb{K}^{I_\ell \times R_\ell}$. The equality $\mathcal{T} = (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(L)}) \cdot \mathcal{S}$ remains intact after these changes, we only removed the unnecessary terms. Note that we passed from $\prod_{\ell=1}^{L} I_\ell$ terms to just $\prod_{\ell=1}^{L} R_\ell$. The procedure of obtaining such decomposition clearly shows that $\mathcal{S}$, after deleting the unnecessary zeros, is a *compressed version of* $\mathcal{T}$. The MLSVD in compressed form can also be called a *reduced MLSVD*, in contrast to the full MLSVD of theorem 2.2.5.

Remark 2.2.9. *After computing the MLSVD of $\mathcal{T}$, notice that computing a CPD for $\mathcal{S}$ is equivalent to computing a CPD for $\mathcal{T}$. Indeed, if $\mathcal{S} = \sum_{r=1}^{R} \boldsymbol{w}_r^{(1)} \otimes \ldots \otimes \boldsymbol{w}_r^{(L)}$, then we can write $\mathcal{S} = (\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L)}) \cdot \mathcal{I}_{R \times \ldots \times R}$, which implies that*

$$\mathcal{T} = (\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(L)}) \cdot \Big((\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L)}) \cdot \mathcal{I}_{R \times \ldots \times R} \Big) =$$

$$= (\boldsymbol{U}^{(1)} \boldsymbol{W}^{(1)}, \ldots, \boldsymbol{U}^{(L)} \boldsymbol{W}^{(L)}) \cdot \mathcal{I}_{R \times \ldots \times R}.$$

Now we go back to our claim that the MLSVD is indeed a good generalization of the SVD. The next results demonstrate such claim.

Theorem 2.2.10 (L. De Lathauwer, B. De Moor, J. Vandewalle, 2000)**.** *In the case $L = 2$, the MLSVD is equal to the SVD of matrices.*

Theorem 2.2.11 (L. De Lathauwer, B. De Moor, J. Vandewalle, 2000)**.** *Let $\mathcal{T} = (\boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(L)}) \cdot \mathcal{S}$ be MLSVD of $\mathcal{T}$ and, for each $\ell = 1 \ldots L$, let R_ℓ be the largest index such that $\sigma_{R_\ell}^{(\ell)} > 0$. Then the following holds.*

1. *$rank_{\boxplus}(\mathcal{T}) = (R_1, \ldots, R_L)$.*
2. *$Im(\mathcal{T}_{(\ell)}) = span(\boldsymbol{U}_1^{(\ell)}, \ldots, \boldsymbol{U}_{R_\ell}^{(\ell)})$ for each $\ell = 1 \ldots L$.*
3. *$\ker(\mathcal{T}_{(\ell)}^*) = span(\boldsymbol{U}_{R_\ell+1}^{(\ell)}, \ldots, \boldsymbol{U}_{I_\ell}^{(\ell)})$ for each $\ell = 1 \ldots L$.*
4. *$\|\mathcal{T}\|^2 = \|\mathcal{S}\|^2 = \sum_{r=1}^{R_1} \big(\sigma_r^{(1)}\big)^2 = \ldots = \sum_{r=1}^{R_L} \big(\sigma_r^{(L)}\big)^2$.*

Item 1 of this theorem tells us that the values R_ℓ coincide with the rank mode-ℓ of $\mathcal{T}$. This is in agreement with the relation between the singular values of matrices and rank. All other items also have their immediate version for the classic SVD. A relevant SVD property that unfortunately does not extend to the MLSVD is that of the tensor closest to $\mathcal{T}$ with a fixed lower multilinear rank. In the matrix case we have the Eckart-Young theorem, which states that the rank-$\tilde{R}$ matrix closest to $\mathbf{M} = \sum_{r=1}^{R} \sigma_r \mathbf{u}_r \mathbf{v}_r^*$ is $\tilde{\mathbf{M}} = \sum_{r=1}^{\tilde{R}} \sigma_r \mathbf{u}_r \mathbf{v}_r^*$, where $\tilde{R} \leq R$ and the decomposition for $\mathbf{M}$ is a SVD. $\tilde{\mathbf{M}}$ is a low rank approximation for $\mathbf{M}$ (the best one with rank $\tilde{R}$) and in this case we have the following theorem about the error of this approximation.

Theorem 2.2.12. *With the notations above, we have that*[2]

$$\|\boldsymbol{M} - \tilde{\boldsymbol{M}}\|^2 = \sum_{r=\tilde{R}+1}^{R} \sigma_r^2.$$

We can obtain a truncation $\tilde{\mathcal{T}}$ for $\mathcal{T}$ in a totally analogous way. Let $rank_{\boxplus}(\mathcal{T}) = (R_1, \ldots, R_L)$. Given a lower multilinear rank $(\tilde{R}_1, \ldots, \tilde{R}_L) < (R_1, \ldots, R_L)$, we obtain a truncation by zeroing all coefficients $\tilde{t}_{i_1 \ldots i_L}$ when some of its indexes satisfies $i_\ell > \tilde{R}_\ell$. Then we have the a tensor of multilinear rank $(\tilde{R}_1, \ldots, \tilde{R}_L)$ given by

$$\tilde{\mathcal{T}} = \sum_{i_1=1}^{\tilde{R}_1} \ldots \sum_{i_L=1}^{\tilde{R}_L} t_{i_1 \ldots i_L} \, \mathbf{U}^{(1)}_{:i_1} \otimes \ldots \otimes \mathbf{U}^{(L)}_{:i_L}.$$

The following result first appeared in [30] and the first correct proof appeared in [44].

Theorem 2.2.13. *With the notation above, we have that*

$$\|\mathcal{T} - \tilde{\mathcal{T}}\|^2 \leq \sum_{r=\tilde{R}_1+1}^{R_1} \left(\sigma_r^{(1)}\right)^2 + \ldots + \sum_{r=\tilde{R}_L+1}^{R_L} \left(\sigma_r^{(L)}\right)^2.$$

In general, $\tilde{\mathcal{T}}$ is a very good low multilinear rank approximation for $\mathcal{T}$, but unlike the matrix case, it is not necessarily the best approximation with low multilinear rank $(\tilde{R}_1, \ldots, \tilde{R}_L)$. At the time we know that [80]

$$\|\mathcal{T} - \tilde{\mathcal{T}}\| \leq \sqrt{L} \min_{rank_{\boxplus}(\mathcal{X}) = (\tilde{R}_1, \ldots, \tilde{R}_L)} \|\mathcal{T} - \mathcal{X}\|.$$

The current algorithms to obtain the best approximation usually use the MLSVD to obtain $\tilde{\mathcal{T}}$ and then use this tensor (which is already very close to $\mathcal{T}$) as the starting point for some iterative algorithm [89].

[2] Remember that we are using the Frobenius norm.

A few words about the cost of algorithm 2.2.6 are necessary. Start noting that the construction of the unfoldings of $\mathcal{T}$ requires noncontiguous memory accesses. Although the computational time to achieve this may be noticeable, we disregard this cost because the time to compute the SVDs dominates the algorithm. In general, the unfolding $\mathcal{T}_{(\ell)}$ probably will have much more columns than rows. The procedure described in A.1.5 permits one to compute its singular values $\sigma_k^{(\ell)}$ and its left singular vectors $\mathbf{U}_{:k}^{(\ell)}$ with low cost since the left singular vectors are the eigenvectors obtained in lemma A.1.2. With this approach we don't compute the right singular vectors, but that is fine since they are not used for the MLSVD. In fact, not computing them is a good thing, we avoid unnecessary computations (note that this wouldn't be possible if we computed the SVD of $\mathcal{T}_{(\ell)}$ directly). The cost of this approach is of $\mathcal{O}\left(I_\ell \prod_{\ell'=1}^{L} I_{\ell'} + I_\ell^3\right)$ flops. The case where $\mathcal{T}_{(\ell)}$ have less columns than rows is a bit more problematic, we have a unbalanced tensor (definition B.2.10) to deal with. We could take a similar route and compute the eigenvalues of $\mathcal{T}_{(\ell)}^T \mathcal{T}_{(\ell)}$ but now the eigenvectors obtained are the right singular vectors of $\mathcal{T}_{(\ell)}$, which we don't need to use. This is a situation where we have to use the SVD and lose some performance. Computing the reduced SVD in this case has a cost of, at least, $\mathcal{O}\left(2I_\ell \prod_{\ell'\neq\ell} I_{\ell'}^2 + 2\prod_{\ell'\neq\ell} I_{\ell'}^3\right)$ flops. Instead of computing the full SVD and deal with these high costs, we can take another route and compute a truncated SVD. Since each $\mathcal{T}_{(\ell)}$ has rank not bigger than $\min\{I_\ell, R\}$, we define $P_\ell = \min\{I_\ell, R\}$ and compute the truncated SVD with P_ℓ singular values. This approach has a cost of $\mathcal{O}\left(P_\ell \prod_{\ell'=1}^{L} I_{\ell'}\right)$ flops, which is much better than the previous cost. If we are willing to lose some precision to gain speed, it is possible to use randomized algorithms for the truncated SVD [31]. This algorithm has a cost of $\mathcal{O}\left(\log(P_\ell) \prod_{\ell'=1}^{L} I_{\ell'}\right)$ flops and the loss in precision is negligible. Since we need to perform this computation for each $\ell = 1 \dots L$, the overall cost is of $\mathcal{O}\left(\sum_{\ell=1}^{L} \log(P_\ell) \prod_{\ell'=1}^{L} I_{\ell'}\right)$.

The approach just described to compute the MLSVD can be referred as the *classic truncated*, where the idea is to compute the truncated SVD of each unfolding of $\mathcal{T}$. A much faster and slightly less precise approach is the *sequentially truncated MLSVD* algorithm introduced by N. Vannieuwenhove, R. Vandebril and K. Meerberge [44], which consists of interlacing the computation of the core tensor and the factor matrices. The cost is of this algorithm is of $\mathcal{O}\left(\sum_{\ell=1}^{L} \log(P_\ell) \prod_{\ell'=1}^{\ell-1} R_{\ell'} \prod_{\ell'=\ell}^{L} I_{\ell'} + R_\ell^2 \prod_{\ell'=1}^{\ell-1} R_{\ell'} \prod_{\ell'=\ell+1}^{L} I_{\ell'}\right)$ flops. The first is cost is due to the truncated SVD and the second one is due to the matrix-matrix multiplication, see algorithm below.

Algorithm 2.2.14 (ST-MLSVD).
Input: $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_L}$

$\mathcal{S}^{(0)} = \mathcal{T}$
`for` $\ell = 1 \ldots L$
$\quad \mathcal{S}_{(\ell)}^{(\ell-1)} \leftarrow$ `Unfolding` $(\mathcal{S}^{(\ell-1)}, \ell)$
$\quad \boldsymbol{U}^{(\ell)}, \Sigma^{(\ell)}, \boldsymbol{V}^{(\ell)} \leftarrow$ `SVD` $(\mathcal{S}_{(\ell)})$
$\quad$ *Set* $\mathcal{S}^{(\ell)}$, *where* $\mathcal{S}_{(\ell)}^{(\ell)} = \Sigma^{(\ell)} \big(\boldsymbol{V}^{(\ell)} \big)^T$

Output: $\boldsymbol{U}^{(\ell)} \in \mathbb{R}^{I_\ell \times R_\ell}, \Sigma^{(\ell)} \in \mathbb{R}^{R_\ell \times R_\ell}$ *for* $\ell = 1 \ldots L$ *and* $\mathcal{S} \in \mathbb{R}^{R_1 \times \ldots \times R_L}$

The first version of Tensor Fox only used the classic truncated algorithm to compute the MLSVD, but now it has the possibility to switch between this one and the sequentially truncated algorithm.

Chapter 3

Gauss-Newton algorithm

In this chapter we really begin to touch the computational aspects of the CPD. Some of the history and known algorithms are discussed, and then we go to nonlinear least squares algorithms, which includes the Gauss-Newton method. This is the method used by Tensor Fox. The main challenge is the approximated Hessian, which is singular by construction. To overcome this issue we introduce a new kind of regularization. We also give the constructive proofs of some already known results from tensor algebra. The reason of this is to give the reader an understanding of the computational aspects related of our problems, this makes a nice parallel with actual coding necessary to make Tensor Fox. Therefore all proofs are given in coordinates. In the end of the chapter we show one of the main contributions of this work: a set of formulas to make fast matrix-vector computations with low memory cost.

3.1 Preliminaries

In 1927, Hitchcock [60, 61] proposed the idea of the CPD. This idea only became popular in 1970 inside the psychometrics community, called CANDECOMP (canonical decomposition) by Carrol and Chang [62], and PARAFAC (parallel factors) by Harshman [63]. Gradually the CPD began to be applied in more areas and today is a successful tool for multi-dimensional data analysis and prediction, such as blind source separation, food industry, dimensionality reduction, pattern/image recognition, machine learning and data mining [2, 3, 5, 7–9].

As we mentioned in section 1.4, computing the rank of a tensor is a NP-hard problem. Given a tensor $\mathcal{T}$ one may try to find its rank together with its CPD by searching for a best fit, that is, for each $R = 1, 2, \ldots$, compute a rank-R CPD for $\mathcal{T}$ until some criteria of good fit is reached. This is a possible but risky procedure because of the border rank phenomenon. More precisely, if $\underline{rank}(\mathcal{T}) < rank(\mathcal{T})$, then it is possible to obtain arbitrarily good fits with low rank approximations for $\mathcal{T}$. This may cause problems in practice [57, 64, 65].

Fix a tensor $\mathcal{T} \in \mathbb{K}^{I_1 \times \ldots \times I_L}$ and a value R. Our problem is to compute a rank-R approximation $\tilde{\mathcal{T}} = (\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}) \cdot \mathcal{I}_{R \times \ldots \times R} \in \mathbb{K}^{I_1 \times \ldots \times I_L}$. More precisely, $\tilde{\mathcal{T}}$ is given by

$$\tilde{\mathcal{T}} = \sum_{r=1}^{R} \mathbf{w}_r^{(1)} \otimes \ldots \otimes \mathbf{w}_r^{(L)}, \tag{3.1}$$

where $\mathbf{W}^{(\ell)} = [\mathbf{w}_1^{(\ell)}, \ldots, \mathbf{w}_R^{(\ell)}]^T$ are the factor matrices, for $\ell = 1, \ldots L$, and $\mathcal{I}_{R \times \ldots \times R} \in \mathbb{K}^{R \times \ldots \times R}$ is the diagonal tensor with L dimensions.

Note that we don't assume $rank(\mathcal{T}) = R$, we only assume that such R is given. Frequently the data at hand is noisy, because of this the rank of $\mathcal{T}$ may be different from the noiseless data, and this is the one we are interested in. Usually is a good idea to compute low rank approximations since the noise can cause the rank to be typical or generic, that is, the noisy tensors are not special (because they are "everywhere") but the objective tensor is. We want this approximation to be such that $\|\mathcal{T} - \tilde{\mathcal{T}}\|$ is smallest as possible. The most common approach is to formulate the problem as an unconstrained minimization problem

$$\min_{\tilde{\mathcal{T}}} \frac{1}{2} \|\mathcal{T} - \tilde{\mathcal{T}}\|^2. \tag{3.2}$$

Define the function $F : \mathbb{K}^{R \sum_{\ell=1}^{L} I_\ell} \to \mathbb{R}$ given by

$$F(\mathbf{w}) = \frac{1}{2} \|\mathcal{T} - (\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}) \cdot \mathcal{I}_{R \times \ldots \times R}\|^2,$$

where $\mathbf{w}$ is the just the vector obtained by vertically stacking[1] all columns of all factor matrices. More precisely,

$$\mathbf{w} = \begin{bmatrix} vec(\mathbf{W}^{(1)}) \\ \vdots \\ vec(\mathbf{W}^{(L)}) \end{bmatrix}.$$

Note that solving 3.2 is equivalent to minimizing F. In this work we will be considering only real tensors, but the complex case is considered in [15] for instance. The first thing we should note is that any solution of 3.2 is a critical point of F so it is of interest to derive an expression for the derivative of F and its critical points.

Lemma 3.1.1. *For any vectors $\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)} \in \mathbb{R}^{I_1}, \ldots, \boldsymbol{x}^{(L)}, \boldsymbol{y}^{(L)} \in \mathbb{R}^{I_L}$ we have that*

$$\prod_{\ell=1}^{L} \langle \boldsymbol{x}^{(\ell)}, \boldsymbol{y}^{(\ell)} \rangle = \sum_{i_1=1}^{I_1} \cdots \sum_{i_L=1}^{I_L} x_{i_1}^{(1)} \ldots x_{L}^{(L)} y_{i_1}^{(1)} \ldots y_{i_L}^{(L)}.$$

[1] Remember, from example 1.3.8, that $vec(\mathbf{W}^{(\ell)})$ is the vector obtained by vertically stacking all columns of $\mathbf{W}^{(\ell)}$.

Theorem 3.1.2. *For all $\ell' = 1 \dots L$, $i' = 1 \dots I_\ell$ and $r' = 1 \dots R$, we have*

$$\frac{\partial F}{w_{i'r'}^{(\ell')}}(\boldsymbol{w}) = -\mathcal{T}(\boldsymbol{w}_{r'}^{(1)}, \dots, \boldsymbol{e}_{i'}^{(\ell')}, \dots, \boldsymbol{w}_{r'}^{(L)}) + \sum_{r=1}^{R} w_{i'r}^{(\ell')} \prod_{\ell \neq \ell'} \langle \boldsymbol{w}_r^{(\ell)}, \boldsymbol{w}_{r'}^{(\ell)} \rangle.$$

Proof:

$$\frac{\partial F}{w_{i'r'}^{(\ell')}}(\mathbf{w}) = \frac{1}{2} \frac{\partial}{\partial w_{i'r'}^{(\ell')}} \left(\sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)} \right)^2 \right) =$$

$$= \frac{1}{2} \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} \frac{\partial}{\partial w_{i'r'}^{(\ell')}} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)} \right)^2 =$$

$$= \frac{1}{2} \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} 2 \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)} \right) \cdot \frac{\partial}{\partial w_{i'r'}^{(\ell')}} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)} \right) =$$

$$= \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)} \right) \cdot \left(- \frac{\partial}{\partial w_{i'r'}^{(\ell')}} w_{i_1 r'}^{(1)} \dots w_{i_L r'}^{(L)} \right) =$$

$$= \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)} \right) \cdot \left(- \prod_{\ell \neq \ell'} w_{i_\ell r'}^{(\ell)} \cdot \frac{\partial}{\partial w_{i'r'}^{(\ell')}} w_{i_{\ell'} r'}^{(\ell')} \right) =$$

$$= \sum_{i_1=1}^{I_1} \dots \sum_{i_{\ell'-1}=1}^{I_{\ell'-1}} \sum_{i_{\ell'+1}=1}^{I_{\ell'+1}} \dots \sum_{i_L=1}^{I_L} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i'r}^{(\ell')} \prod_{\ell \neq \ell'} w_{i_\ell r}^{(\ell)} \right) \cdot \left(- \prod_{\ell \neq \ell'} w_{i_\ell r'}^{(\ell)} \cdot \frac{\partial}{\partial w_{i'r'}^{(\ell')}} w_{i'r'}^{(\ell')} \right) =$$

$$= \sum_{i_1=1}^{I_1} \dots \sum_{i_{\ell'-1}=1}^{I_{\ell'-1}} \sum_{i_{\ell'+1}=1}^{I_{\ell'+1}} \dots \sum_{i_L=1}^{I_L} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i'r}^{(\ell')} \prod_{\ell \neq \ell'} w_{i_\ell r}^{(\ell)} \right) \cdot \left(- \prod_{\ell \neq \ell'} w_{i_\ell r'}^{(\ell)} \right) =$$

$$= - \sum_{i_1=1}^{I_1} \dots \sum_{i_{\ell'-1}=1}^{I_{\ell'-1}} \sum_{i_{\ell'+1}=1}^{I_{\ell'+1}} \dots \sum_{i_L=1}^{I_L} t_{i_1 \dots i_L} \prod_{\ell \neq \ell'} w_{i_\ell r'}^{(\ell)} + \sum_{r=1}^{R} w_{i'r}^{(\ell')} \sum_{i_1=1}^{I_1} \dots \sum_{i_{\ell'-1}=1}^{I_{\ell'-1}} \sum_{i_{\ell'+1}=1}^{I_{\ell'+1}} \dots \sum_{i_L=1}^{I_L} \prod_{\ell \neq \ell'} w_{i_\ell r}^{(\ell)} \prod_{\ell \neq \ell'} w_{i_\ell r'}^{(\ell)}.$$

$$= -\mathcal{T}(\mathbf{w}_{r'}^{(1)}, \dots, \mathbf{e}_{i'}^{(\ell')}, \dots, \mathbf{w}_{r'}^{(L)}) + \sum_{r=1}^{R} w_{i'r}^{(\ell')} \prod_{\ell \neq \ell'} \langle \mathbf{w}_r^{(\ell)}, \mathbf{w}_{r'}^{(\ell)} \rangle,$$

where the last term was calculated using the previous lemma. □

Corollary 3.1.3. *Let* $\dfrac{\partial F}{\boldsymbol{w}_{r'}^{(\ell')}}(\boldsymbol{w}) = \left[\dfrac{\partial F}{w_{1r}^{(\ell')}}, \ldots, \dfrac{\partial F}{w_{I_{\ell'}r}^{(\ell')}}\right]^T$, *then*

$$\frac{\partial F}{\boldsymbol{w}_{r'}^{(\ell')}}(\boldsymbol{w}) = -\begin{bmatrix} \mathcal{T}(\boldsymbol{w}_{r'}^{(1)}, \ldots, \boldsymbol{e}_1^{(\ell')}, \ldots, \boldsymbol{w}_{r'}^{(L)}) \\ \vdots \\ \mathcal{T}(\boldsymbol{w}_{r'}^{(1)}, \ldots, \boldsymbol{e}_{I_{\ell'}}^{(\ell')}, \ldots, \boldsymbol{w}_{r'}^{(L)}) \end{bmatrix} + \sum_{r=1}^{R} \prod_{\ell \neq \ell'} \langle \boldsymbol{w}_r^{(\ell)}, \boldsymbol{w}_{r'}^{(\ell)} \rangle \boldsymbol{w}_r^{(\ell')}.$$

Before we start talking about algorithms, it is relevant to mention what are the biggest challenges we will be facing when designing an algorithm to solve 3.2. In the following, let $\mathcal{T} = \sum_{r=1}^{R} \mathbf{v}_r^{(1)} \otimes \ldots \otimes \mathbf{v}_r^{(L)}$.

1. As mentioned in section 1.5, finding the best rank-R approximation may be a ill-posed problem.

2. Ill-posed problems are not "rare" in the sense that the set of ill-posed problems (i.e., the set of tensors such that 3.2 is ill-posed) usually have positive volume [20].

3. We say $\mathcal{T}$ has a *bottleneck* if there is a mode $1 \leq \ell \leq L$ such that at least two vectors $\mathbf{v}_{r'}^{(\ell)}$ and $\mathbf{v}_{r''}^{(\ell)}$ are almost collinear. This means $\mathcal{T}$ has two "problematic" vectors in $\mathbb{R}^{I_\ell}$, and they are considered problematic because two different rank one terms will have almost collinear vectors at the ℓ-th position.

4. We say $\mathcal{T}$ has a *swamp* if there are bottlenecks in all modes.

5. We say $\mathcal{T}$ has a *degeneracy* if some factors diverge but cancel out when the process of computing a best fit is performed. This is the same degeneracy mentioned at the end of section 1.5.

The last three items in this list constitutes a classification of possible causes of bad numerical performance in numerical algorithms to compute the CPD. This classification was proposed already in 1989 [48] and still is used today. To get a better understand of bottlenecks and swamps, we illustrate these occurrences in the next two examples.

Example 3.1.4 (Bottleneck)**.** *Consider the tensor* $\mathcal{T} \in \mathbb{R}^{2\times 2\times 2}$ *given by*

$$\mathcal{T} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1.01 \\ 1.9 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 2 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 4 \\ 4 \end{bmatrix}.$$

The vectors of the first mode are

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 1.01 \\ 1.9 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

The vectors of the second mode are

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

The vectors of the third mode are

$$\begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 4 \end{bmatrix}.$$

We can note that the vectors of the first mode introduces a possible bottleneck in $\mathcal{T}$. More precisely, consider the two vectors $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 1.01 \\ 1.9 \end{bmatrix}$ of the first mode are. Compared to the vectors of the other modes, these two are indeed very close to be equal (hence collinear) so we can consider that $\mathcal{T}$ has a bottleneck. This is a single bottleneck, but it is possible to have multiple bottlenecks. The extreme case is when we have bottlenecks at all modes, which represents a swamp.

Example 3.1.5 (Swamp)**.** *Consider the tensor $\mathcal{T} \in \mathbb{R}^{2\times 2\times 2}$ given by*

$$\mathcal{T} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1.01 \\ 1.9 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.09 \\ 2.01 \end{bmatrix} \otimes \begin{bmatrix} 10^{-6} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 2.99 \\ 3.99 \end{bmatrix}.$$

The vectors of the first mode are

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 1.01 \\ 1.9 \end{bmatrix}, \quad \begin{bmatrix} 0.09 \\ 2.01 \end{bmatrix}.$$

The vectors of the second mode are

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 10^{-6} \\ 1 \end{bmatrix}.$$

The vectors of the third mode are

$$\begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2.99 \\ 3.99 \end{bmatrix}.$$

In all modes there are at least two almost collinear vectors, therefore $\mathcal{T}$ has a swamp.

Now we briefly describe the most used approaches to compute a CPD.

3.2 Alternating least squares

Several different approaches to solving 3.2 were proposed in the past years, but before that, a single algorithm were used from decades: the *alternating least squares* (ALS). For decades this was considered to be the "workhorse" algorithm to compute the CPD. We present the algorithm for third order tensor and generalize after that. From this point, we write $m \times n \times p$ instead $I_1 \times I_2 \times I_3$ for third order tensor.

Consider a tensor $\mathcal{T} \in \mathbb{R}^{m \times n \times p}$ and suppose we want to compute a rank-R CPD for $\mathcal{T}$. Let $\tilde{\mathcal{T}} = \sum_{r=1}^{R} \mathbf{x}_r \otimes \mathbf{y}_r \times \mathbf{z}_r$ be the approximating tensor, and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_R] \in \mathbb{R}^{m \times R}$, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_R] \in \mathbb{R}^{n \times R}$, $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_R] \in \mathbb{R}^{p \times R}$ the factor matrices.

The first step of the ALS algorithm consists in generating a initial tensor $\tilde{\mathcal{T}}$ to start the iterations. The method of initialization is not part of the algorithm so we won't discuss it here. Now fix $\mathbf{Y}, \mathbf{Z}$ and solve the minimization problem

$$\min_{\mathbf{X}} \|\mathcal{T} - (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathcal{I}_{R \times \times R}\|.$$

Note that we solve it just for $\mathbf{X}$. After this is done, fix $\mathbf{X}, \mathbf{Z}$ and solve the minimization problem

$$\min_{\mathbf{Y}} \|\mathcal{T} - (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathcal{I}_{R \times \times R}\|$$

for $\mathbf{Y}$. After that, fix $\mathbf{X}, \mathbf{Y}$ and solve the minimization problem

$$\min_{\mathbf{Z}} \|\mathcal{T} - (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathcal{I}_{R \times \times R}\|$$

for $\mathbf{Z}$. Once the three factor matrices are updated we repeat this procedure all over again, updating as many times as necessary. The name comes from the fact we are alternating the factor matrices to be solved at each linear least squares problem. To see that these minimization problems indeed are linear least squares problems, note that we can rewrite the first one as

$$\min_{\mathbf{X}} \|\mathcal{T}_{(1)} - \tilde{\mathcal{T}}_{(1)}\| = \min_{\mathbf{X}} \|\mathcal{T}_{(1)} - \mathbf{X} \cdot (\mathbf{Z} \odot \mathbf{Y})^T\|,$$

where the equality $\tilde{\mathcal{T}}_{(1)} = \mathbf{X} \cdot (\mathbf{Z} \odot \mathbf{Y})^T$ comes from theorem 2.2.4. The solution of this problem is given explicitly by $\mathbf{X} = \mathcal{T}_{(1)} \cdot ((\mathbf{Z} \odot \mathbf{Y})^T)^\dagger$. Theorem B.3.5 gives a formula for the pseudoinverse of a Khatri-Rao product, so we have

$$((\mathbf{Z} \odot \mathbf{Y})^T)^\dagger = ((\mathbf{Z} \odot \mathbf{Y})^\dagger)^T = \left(((\mathbf{Z}^T\mathbf{Z}) * (\mathbf{Y}^T\mathbf{Y}))^\dagger (\mathbf{Z} \odot \mathbf{Y})^T\right)^T =$$

$$= (\mathbf{Z} \odot \mathbf{Y}) \big(\underbrace{((\mathbf{Z}^T\mathbf{Z}) * (\mathbf{Y}^T\mathbf{Y}))^T}_{\text{symmetric}} \big)^\dagger = (\mathbf{Z} \odot \mathbf{Y}) \left((\mathbf{Z}^T\mathbf{Z}) * (\mathbf{Y}^T\mathbf{Y})\right)^\dagger.$$

This formulation only requires computing the pseudoinverse of a $R \times R$ matrix rather than a $np \times R$, which is the size of $(\mathbf{Z} \odot \mathbf{Y})^T$. Below we give the algorithm for the general case.

Algorithm 3.2.1 (ALS).
Input: $\mathcal{T}, R$

`Initialize` $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L)}$
`repeat`
  `for` $\ell = 1 \ldots L$
    $\boldsymbol{W}^{(\ell)} \leftarrow \texttt{argmin}_{\boldsymbol{W}^{(\ell)}} \| \mathcal{T}_{(\ell)} - \boldsymbol{W}^{(\ell)} \cdot (\boldsymbol{W}^{(L)} \odot \ldots \odot \boldsymbol{W}^{(\ell+1)} \odot \boldsymbol{W}^{(\ell-1)} \odot \ldots \odot \boldsymbol{W}^{(1)})^T \|$
`until stopping criteria is met`

Output: $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L)}$

The ALS algorithm is simple to understand and to implement, but it has some serious drawbacks. Usually it take many iterations to converge, it is not guaranteed to converge to a global minimum or even to a stationary point, and the final solution can depend heavily on the initialization. Furthermore, the ALS algorithm is known from its convergence problems in the presence of bottlenecks and swamps. Several approaches were implemented in order to improve the ALS performance [13, 15, 27, 29, 66] and still today there are people trying improve ALS.

3.3 Optimization methods

Because of the ALS limitations, researches tried to investigate different approaches to 3.2. A natural idea is to consider it just as an unconstrained minimization problem and apply the known algorithms. Gradient-based algorithms are the main choice in this case. In the literature one can find theoretical and experimental work with *nonlinear conjugate gradient method* and *limited-memory BFGS method* [15, 17]. Improvements to these methods were tried, such as adding line search[2], regularization and dogleg trust region [22].

3.4 Nonlinear least squares

The method of *least squares* is a standard approach to approximate the solution of overdetermined systems, that is, sets of equations in which there are more equations than unknowns. The term "least squares" comes from the fact that we want to minimize the overall sum of the squares of the residuals made in the results of every single equation.

[2]It seems that the *Moré -Thuente* line search is the most popular choice of line search for tensor CPD, see [67] for more information.

Let $(\mathbf{x}^{(i)}, y^{(i)})_{i=1\dots n}$ be a dataset of n points, where $\mathbf{x}^{(i)} \in \mathbb{R}^p$ is the independent variable and $y^{(i)} \in \mathbb{R}$ is the dependent variable. The *model* is a function $h_{\mathbf{w}} : \mathbb{R}^p \to \mathbb{R}$ (also called the *hypothesis*) in the variable $\mathbf{x}$ with parameters $\mathbf{w} = [w_1, \dots, w_m]^T$ to be adjusted. Our objective is to find the parameters that best fits the data. The "best fit" is in the sense that the choice of $\mathbf{w}$ should minimize the value

$$\sum_{i=1}^{n} \left(y^{(i)} - h_{\mathbf{w}}(\mathbf{x}^{(i)})\right)^2.$$

Denote $f_i(\mathbf{w}) = y^{(i)} - h_{\mathbf{w}}(\mathbf{x}^{(i)})$. We have that each f_i is a *residual* of the model. The *error function* (or *cost function*) is the function $F : \mathbb{R}^m \to \mathbb{R}$ given by

$$F(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n} \left(y^{(i)} - h_{\mathbf{w}}(\mathbf{x}^{(i)})\right)^2 = \frac{1}{2}\sum_{i=1}^{n} f_i(\mathbf{w})^2 = \frac{1}{2}\|f(\mathbf{w})\|^2,$$

where $f : \mathbb{R}^m \to \mathbb{R}^n$ is given by $f = (f_1, \dots, f_n)$. Our objective is to minimize F.

Remark 3.4.1. *The ideal situation occurs when $\boldsymbol{w}$ is such that $F(\boldsymbol{w}) = 0$, so $h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}) = y^{(i)}$ for all $i = 1 \dots n$. This means $\boldsymbol{w}$ is a solution of the system*

$$\begin{cases} h_{\boldsymbol{w}}(\boldsymbol{x}^{(1)}) = y^{(1)} \\ \qquad\vdots \\ h_{\boldsymbol{w}}(\boldsymbol{x}^{(n)}) = y^{(n)} \end{cases}.$$

Note that this is a system of n equations and m variables. Since we always expect to have more data then parameters (i.e., $m < n$), this is a overdetermined system. In practice $\boldsymbol{w}$ will not be an exact solution but an approximated solution for this system.

We call the model *linear* when $h_{\mathbf{w}}$ is a linear combination of the parameters, that is, $h_{\mathbf{w}}$ is of the form $h_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{m} w_j\ \phi_j(\mathbf{x})$, where each ϕ_j is any function of $\mathbf{x}$. In this case one can write

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{1}{2}\min_{\mathbf{w}} \left\| \begin{bmatrix} \phi_1(\mathbf{x}^{(1)}) & \dots & \phi_m(\mathbf{x}^{(1)}) \\ \vdots & & \vdots \\ \phi_1(\mathbf{x}^{(n)}) & \dots & \phi_m(\mathbf{x}^{(n)}) \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \right\|^2 = \frac{1}{2}\|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2,$$

where $a_{ij} = \phi_j(\mathbf{x}^{(i)})$ and $\mathbf{y} = [y^{(1)}, \dots, y^{(n)}]^T$. The minimizer is the solution of the normal equation $\mathbf{A}^T\mathbf{A}\mathbf{w} = \mathbf{A}^T\mathbf{y}$, and is given explicitly by $\mathbf{w} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$.

We call the model *nonlinear* when it is not linear. In this case there is no guarantee of a closed-form solution as in the linear case.

Lemma 3.4.2. *Let $\boldsymbol{J}_f(\boldsymbol{w})$ be the Jacobian matrix of f at $\boldsymbol{w}$. Then*

$$\nabla F(\boldsymbol{w}) = \boldsymbol{J}_f^T(\boldsymbol{w}) \cdot f(\boldsymbol{w}).$$

Proof: First note that $\dfrac{\partial F}{\partial w_j}(\mathbf{w}) = \sum_{i=1}^{n} f_i(\mathbf{w}) \dfrac{\partial f_i}{\partial w_j}(\mathbf{w})$ for all $j = 1 \dots m$. In particular, we have

$$\nabla F(\mathbf{w}) = \begin{bmatrix} \sum_{i=1}^n f_i(\mathbf{w}) \frac{\partial f_i}{\partial w_1}(\mathbf{w}) \\ \vdots \\ \sum_{i=1}^n f_i(\mathbf{w}) \frac{\partial f_i}{\partial w_m}(\mathbf{w}) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1}(\mathbf{w}) & \dots & \frac{\partial f_n}{\partial w_1}(\mathbf{w}) \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial w_m}(\mathbf{w}) & \dots & \frac{\partial f_n}{\partial w_m}(\mathbf{w}) \end{bmatrix} \begin{bmatrix} f_1(\mathbf{w}) \\ \vdots \\ f_n(\mathbf{w}) \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{\partial f_1}{\partial w_1}(\mathbf{w}) & \dots & \frac{\partial f_1}{\partial w_m}(\mathbf{w}) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial w_1}(\mathbf{w}) & \dots & \frac{\partial f_n}{\partial w_m}(\mathbf{w}) \end{bmatrix}^T \begin{bmatrix} f_1(\mathbf{w}) \\ \vdots \\ f_n(\mathbf{w}) \end{bmatrix} = \mathbf{J}_f^T(\mathbf{w}) \cdot f(\mathbf{w}). \qquad \square$$

Now we bring all this to the context of tensors. Let a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times \dots \times I_L}$ and a rank R. The dataset points are indexed by multi-indexes $(i_1, \dots, i_L) \in I_1 \times \dots \times I_L$. Each observation in this case are the coordinates of $\mathcal{T}$, that is, we have $y^{(i_1,\dots,i_L)} = t_{i_1 \dots i_L}$. The independent variables are just the multi-indexes[3], to indicate what coordinate of $\mathcal{T}$ we are considering. The model $h_{\mathbf{w}} : I_1 \times \dots \times I_L \to \mathbb{R}$ tries to approximate each coordinate of $\mathcal{T}$ with the parameters in $\mathbf{w} = [vec(\mathbf{W}^{(1)})^T, \dots, vec(\mathbf{W}^{(L)})^T]^T$ through the CPD formulation given at the beginning of this chapter. More precisely, we have that

$$h_{\mathbf{w}}(i_1, \dots, i_L) = \tilde{t}_{i_1 \dots i_L} = \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)}.$$

At this point it should be clear that

$$F(\mathbf{w}) = \frac{1}{2} \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} \left(t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)} \right)^2 = \frac{1}{2} \|\mathcal{T} - \tilde{\mathcal{T}}\|^2 \tag{3.3}$$

and

$$f_{i_1 \dots i_L}(\mathbf{w}) = t_{i_1 \dots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1)} \dots w_{i_L r}^{(L)}. \tag{3.4}$$

The following result is immediate.

Lemma 3.4.3. *For all $\ell' = 1 \dots L$, $i' = 1 \dots I_\ell$ and $r' = 1 \dots R$, we have*

[3]Our "population" is composed by the indexes of the tensor, while the variables are the actual values of the tensor at such coordinates.

$$\frac{\partial f_{i_1\dots i_{\ell'}\dots i_L}}{\partial w^{(\ell')}_{i'r'}}(\boldsymbol{w}) = \begin{cases} -\prod_{\ell\neq\ell'} w^{(\ell)}_{i_\ell r'}, & \text{if } i' = i_{\ell'} \\ 0, & \text{otherwise} \end{cases}.$$

From this lemma it follows that

$$\frac{\partial F}{\partial w^{(\ell')}_{i'r'}}(\mathbf{w}) = \sum_{i_1=1}^{I_1}\cdots\sum_{i_L=1}^{I_L} f_{i_1\dots i_L}(\mathbf{w})\cdot\frac{\partial f_{i_1\dots i_L}}{\partial w^{(\ell')}_{i'r'}}(\mathbf{w}) =$$

$$= \sum_{i_1=1}^{I_1}\cdots\sum_{i_{\ell'-1}=1}^{I_{\ell'-1}}\sum_{i_{\ell'+1}=1}^{I_{\ell'+1}}\cdots\sum_{i_L=1}^{I_L} f_{i_1\dots i'\dots i_L}(\mathbf{w})\cdot\frac{\partial f_{i_1\dots i'\dots i_L}}{\partial w^{(\ell')}_{i'r'}}(\mathbf{w}) =$$

$$= \sum_{i_1=1}^{I_1}\cdots\sum_{i_{\ell'-1}=1}^{I_{\ell'-1}}\sum_{i_{\ell'+1}=1}^{I_{\ell'+1}}\cdots\sum_{i_L=1}^{I_L}\left(t_{i_1\dots i_L} - \sum_{r=1}^{R} w^{(\ell')}_{i'r}\prod_{\ell\neq\ell'} w^{(\ell)}_{i_\ell r}\right)\left(-\prod_{\ell\neq\ell'} w^{(\ell)}_{i_\ell r'}\right).$$

Observe how these manipulations leads to a faster and clean proof of theorem 3.1.2.

3.5 Gauss-Newton

There are several ways to work a nonlinear least squares problem, and our chosen method is based on the *Gauss-Newton* algorithm. This algorithm can only be used to minimize a sum of squared function values, but it has the advantage that second derivatives, which can be challenging to compute, are not required. Consider the same notations used in the previous section.

The first step of the Gauss-Newton algorithm is to consider a first order approximation of f at a point $\mathbf{w}^{(0)}$, that is,

$$f(\mathbf{w}^{(0)} + \underbrace{(\mathbf{w}-\mathbf{w}^{(0)})}_{step}) = f(\mathbf{w}) \approx f(\mathbf{w}^{(0)}) + \mathbf{J}_f(\mathbf{w}^{(0)})\cdot(\mathbf{w}-\mathbf{w}^{(0)}). \tag{3.5}$$

In order to minimize 3.5 at the neighborhood of $\mathbf{w}^{(0)}$ we can compute the minimum of $\|f(\mathbf{w}^{(0)}) + \mathbf{J}_f(\mathbf{w}^{(0)})\cdot(\mathbf{w}-\mathbf{w}^{(0)})\|$ for $\mathbf{w}$. Note that minimizing 3.5 is a least squares problem since we can rewrite this problem as $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|$ for $\mathbf{A} = \mathbf{J}_f(\mathbf{w}^{(0)})$, $\mathbf{x} = \mathbf{w}-\mathbf{w}^{(0)}$, $\mathbf{b} = -f(\mathbf{w}^{(0)})$.

The solution of 3.5 gives the next iterate $\mathbf{w}^{(1)}$. More generally, we obtain $\mathbf{w}^{(k+1)}$ from $\mathbf{w}^{(k)}$ by defining $\mathbf{w}^{(k+1)} = \mathbf{x}_* + \mathbf{w}^{(k)}$, where $\mathbf{x}_*$ is the solution of the normal equations

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b} \tag{3.6}$$

for $\mathbf{A} = \mathbf{J}_f(\mathbf{w}^{(k)})$, $\mathbf{x} = \mathbf{w} - \mathbf{w}^{(k)}$, $\mathbf{b} = -f(\mathbf{w}^{(k)})$. The explicit formula for $\mathbf{w}^{(k+1)}$ is

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \left(\mathbf{J}_f\left(\mathbf{w}^{(k)}\right)^T \mathbf{J}_f\left(\mathbf{w}^{(k)}\right)\right)^{-1} \cdot \mathbf{J}_f\left(\mathbf{w}^{(k)}\right)^T \cdot f(\mathbf{w}^{(k)})$$

although we should clarify that the inverse above never will be explicitly computed. Usually one uses an iterative algorithm to solve 3.6. This process of obtaining successive $\mathbf{w}^{(k)}$ is the Gauss-Newton algorithm, and it is guaranteed to converge to a local minima of F. For more details about this algorithm and its properties I recommend [22]. We show the most relevant results here.

Theorem 3.5.1 (K. Madsen, H. B. Nielsen, and O. Tingleff, 2004)**.** *$\boldsymbol{w}^{(k+1)} - \boldsymbol{w}^{(k)}$ is a descent direction for F at $\boldsymbol{w}^{(k)}$.*

Proof: From the formula

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \left(\mathbf{J}_f(\mathbf{w}^{(k)})^T \mathbf{J}_f(\mathbf{w}^{(k)})\right)^{-1} \cdot \mathbf{J}_f(\mathbf{w}^{(k)})^T \cdot f(\mathbf{w}^{(k)})$$

we can conclude that

$$-\mathbf{J}_f(\mathbf{w}^{(k)})^T \mathbf{J}_f(\mathbf{w}^{(k)}) \cdot \left(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right) = \mathbf{J}_f(\mathbf{w}^{(k)})^T \cdot f(\mathbf{w}^{(k)}).$$

With the above observation and lemma 3.4.2 we have that

$$\langle \nabla F(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)} \rangle = \langle \mathbf{J}_f(\mathbf{w}^{(k)})^T f(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)} \rangle =$$

$$= \langle -\mathbf{J}_f(\mathbf{w}^{(k)})^T \mathbf{J}_f(\mathbf{w}^{(k)}) \left(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)} \rangle =$$

$$= -\langle \mathbf{J}_f(\mathbf{w}^{(k)}) \left(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right), \mathbf{J}_f(\mathbf{w}^{(k)}) \left(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right) \rangle =$$

$$= -\| \mathbf{J}_f(\mathbf{w}^{(k)}) \left(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right) \|^2 \leq 0.$$

Since the derivative of F at $\mathbf{w}^{(k)}$ in the direction $\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}$ is negative, it follows that this is a descent direction. □

3.5.1 Deriving the Gauss-Newton method from the Newton's method

Denote the Hessian matrix of F at $\mathbf{w}$ by $\mathbf{H}_F(\mathbf{w})$. We can relate the derivatives of f and F through the following result.

Lemma 3.5.2. *We have that*

$$\boldsymbol{H}_F(\boldsymbol{w}) = \boldsymbol{J}_f^T(\boldsymbol{w})\boldsymbol{J}_f(\boldsymbol{w}) + \sum_{i_1=1}^{I_1} \dots \sum_{i_L=1}^{I_L} f_{i_1 \dots i_L}(\boldsymbol{w}) \cdot \boldsymbol{H}_{f_{i_1 \dots i_L}}(\boldsymbol{w}),$$

where $\boldsymbol{H}_{f_{i_1 \dots i_L}}$ *is the Hessian matrix of* $f_{i_1 \dots i_L}$.

As the algorithm converges we expect to have $f_{i_1 \dots i_L} \approx 0$ for all $i_1, \dots, i_L$. Therefore this lemma implies that $\mathbf{H}_F \approx \mathbf{J}_f^T \mathbf{J}_f$ when close to an optimal point.

We can apply Newton's method to solve the system $\nabla F = 0$. Then the iteration formula becomes

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \left(\mathbf{H}_F(\mathbf{w}^{(k)})\right)^{-1} \cdot \nabla F(\mathbf{w}^{(k)}) \approx$$

$$\approx \mathbf{w}^{(k)} - \left(\mathbf{J}_f^T(\mathbf{w}^{(k)})\mathbf{J}_f(\mathbf{w}^{(k)})\right)^{-1} \cdot \nabla F(\mathbf{w}^{(k)}) =$$

$$= \mathbf{w}^{(k)} - \left(\mathbf{J}_f^T(\mathbf{w}^{(k)})\mathbf{J}_f(\mathbf{w}^{(k)})\right)^{-1} \cdot \mathbf{J}_f^T(\mathbf{w}^{(k)}) f(\mathbf{w}^{(k)}).$$

Here we recover the Gauss-Newton iteration formula for $\mathbf{w}$. This approximation makes sense if the Hessian is nonsingular, otherwise, the approximation will be poor, which will cause the Gauss-Newton algorithm to make very small steps, slowing down convergence. We can also have problems if the function f is highly nonlinear, in which case the approximation of the Hessian will again be poor.

3.5.2 Damped Gauss-Newton

Everything of this section until this point was very general, without specifying tensors. For all this section we consider f and F in the tensor context. In this context, we don't consider f highly nonlinear but, unfortunately, the approximate Hessian converges to a singular matrix as the Gauss-Newton iterations converges. We make this statement more precise soon. For this part we will denote $\dfrac{\partial f_{i_1 \dots i_L}}{\partial w_{i'r'}^{(\ell')}} = \dfrac{\partial f_{i_1 \dots i_L}}{\partial w_{i'r'}^{(\ell')}}(\mathbf{w})$, that is, we suppress the point $\mathbf{w}$ where the derivative is being evaluated. This will make notation simpler and won't cause any confusion. Also let

$$\frac{\partial f_{i_1 \dots i_L}}{\partial \mathbf{w}_{r'}^{(\ell')}} = \left[\frac{\partial f_{i_1 \dots i_L}}{\partial w_{1r'}^{(\ell')}}, \dots, \frac{\partial f_{i_1 \dots i_L}}{\partial w_{I_{\ell'}r'}^{(\ell')}} \right] \in \mathbb{R}^{I_{\ell'}}$$

and

$$\frac{\partial f}{\partial \mathbf{w}_{r'}^{(\ell')}} = \begin{bmatrix} \dfrac{\partial f_{1\dots1}}{\partial \mathbf{w}_{r'}^{(\ell')}} \\ \\ \dfrac{\partial f_{1\dots2}}{\partial \mathbf{w}_{r'}^{(\ell')}} \\ \\ \vdots \\ \\ \dfrac{\partial f_{I_1\dots I_L-1}}{\partial \mathbf{w}_{r'}^{(\ell')}} \\ \\ \dfrac{\partial f_{I_1\dots I_L}}{\partial \mathbf{w}_{r'}^{(\ell')}} \end{bmatrix} \in \mathbb{R}^{\prod_{\ell=1}^{L} I_\ell \times I_{\ell'}}.$$

Notice that the ordering of the rows follows the indexes $i_1 \dots i_L$ as numbers in increasing order. We can consider just one more level of compact notation and define

$$\frac{\partial f}{\partial \mathbf{W}^{(\ell')}} = \left[\frac{\partial f}{\partial \mathbf{w}_1^{(\ell')}}, \dots, \frac{\partial f}{\partial \mathbf{w}_R^{(\ell')}} \right] \in \mathbb{R}^{\prod_{\ell=1}^{L} I_\ell \times I_{\ell'} R},$$

so we have that

$$J_f = \left[\frac{\partial f}{\partial \mathbf{W}^{(1)}}, \dots, \frac{\partial f}{\partial \mathbf{W}^{(L)}} \right] \in \mathbb{R}^{\prod_{\ell=1}^{L} I_\ell \times R \sum_{\ell=1}^{L} I_\ell}.$$

In order to understand the structure of $\mathbf{J}_f$ first we have to understand the structure of each block $\dfrac{\partial f}{\partial \mathbf{w}_{r'}^{(\ell')}}$. Remember that the entries of this matrix were computed in lemma 3.4.3, now we just need to find some structure on it. Given a multi-index $i_1 \dots i_{\ell'} \dots i_L$, the associated row has only one non zero entry, which is the column $i_{\ell'}$, because

$$\frac{\partial f_{i_1\dots i_{\ell'}\dots i_L}}{\partial \mathbf{w}_{r'}^{(\ell')}} = \left[\frac{\partial f_{i_1\dots i_{\ell'}\dots i_L}}{\partial w_{1r'}^{(\ell')}}, \dots, \frac{\partial f_{i_1\dots i_{\ell'}\dots i_L}}{\partial w_{i_{\ell'}r'}^{(\ell')}}, \dots, \frac{\partial f_{i_1\dots i_{\ell'}\dots i_L}}{\partial w_{I_{\ell'}r'}^{(\ell')}} \right] = [0, \dots, \underbrace{-\prod_{\ell \neq \ell'} w_{i_\ell r'}^{(\ell)}}_{i_{\ell'} \text{ column}}, \dots, 0].$$

Now we can see that $\dfrac{\partial f}{\partial \mathbf{w}_{r'}^{(\ell')}}$ will have a sparse structure, and this structure is dictated by the multi-indexes: at row $i_1 \dots i_{\ell'} \dots i_L$, the only non zero entry is given by the value $i_{\ell'}$. Since this structure does not depend on r', the block matrix $\dfrac{\partial f}{\partial \mathbf{W}^{(\ell')}}$ consists in a repetition of R blocks with the same sparse structure (but not necessarily the same values). Because of the ordering of $i_1 \dots i_{\ell'} \dots i_L$, each change of non zero column is done periodically as

the scheme in figure 3.1 illustrates.

Note that when we go to row $1 \dots 111 \dots 11$ to row $1 \dots 121 \dots 11$, the gray block moves one column to the right. When this happens, the next index, $i_{\ell'+1}$, already changed all its values exactly one time. It means all this moves to the right occurred with respect to $i_{\ell'+1}$ when it happened just one with respect to $i_{\ell'}$. This is due to the ordering we used, which functions just as natural numbers in increasing order. This explanation gets clear with a picture, so figure 3.2 illustrates better what is going on.

We finish this discussion of the sparse structure displaying the structure of a concrete tensor shape. If still there are any doubts at this point, we invite the reader to reproduce the structure of $\mathbf{J}_f$ when $\mathcal{T} \in \mathbb{R}^{4\times3\times2}$ and $R = 2$. This structure is given in figure 3.3. We separate the modes by color to facilitate understanding. At the right to the matrix we also added the respective multi-index associated to each row. The green color correspond to the right index, the blue correspond to the middle index and the red correspond to the left index.

Lemma 3.5.3 (T. G. Kolda, E. Acar, D. M. Dunlavy, 2011)**.** *For all $\ell' = 1 \dots L$ and $r' = 1 \dots R$, we have*

$$\frac{\partial f}{\partial \boldsymbol{w}_{r'}^{(\ell')}} = -\boldsymbol{w}_{r'}^{(1)} \tilde{\otimes} \dots \tilde{\otimes} \boldsymbol{w}_{r'}^{(\ell'-1)} \tilde{\otimes} \boldsymbol{I}_{I_{\ell'}} \tilde{\otimes} \boldsymbol{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \boldsymbol{w}_{r'}^{(L)}.$$

Proof: First note that

$$\mathbf{I}_{I_{\ell'}} \tilde{\otimes} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} = \begin{bmatrix} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} & & \\ & \ddots & \\ & & \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} \end{bmatrix}$$

is a block diagonal matrix, with $I_{\ell'} \times I_{\ell'}$ block shape where each entry is a vector of size $\prod_{\ell=\ell'+1}^{L} I_\ell$. With this we have

$$\mathbf{w}_{r'}^{(1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(\ell'-1)} \tilde{\otimes} \begin{bmatrix} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} & & \\ & \ddots & \\ & & \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} \end{bmatrix} =$$

$$= \begin{bmatrix} w_{1r'}^{(1)} \cdot \ldots \cdot w_{1r'}^{(\ell'-1)} \begin{bmatrix} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} & & \\ & \ddots & \\ & & \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} \end{bmatrix} \\ \vdots \\ w_{I_1 r'}^{(1)} \cdot \ldots \cdot w_{I_{\ell'-1}\, r'}^{(\ell'-1)} \begin{bmatrix} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} & & \\ & \ddots & \\ & & \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} \end{bmatrix} \end{bmatrix}.$$

By lemma B.3.2 we have that

$$\mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} = \begin{bmatrix} w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix},$$

hence we can write more explicitly the last expression as

$$\begin{bmatrix} w_{1r'}^{(1)} \cdot \ldots \cdot w_{1r'}^{(\ell'-1)} \begin{bmatrix} \begin{bmatrix} w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix} w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} \end{bmatrix} \\ \vdots \\ w_{I_1 r'}^{(1)} \cdot \ldots \cdot w_{I_{\ell'-1}\, r'}^{(\ell'-1)} \begin{bmatrix} \begin{bmatrix} w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix} w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} \end{bmatrix} \end{bmatrix} =$$

$$= \begin{bmatrix} \begin{bmatrix} \begin{bmatrix} w_{1r'}^{(1)} \cdot \ldots \cdot w_{1r'}^{(\ell'-1)} \cdot w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{1r'}^{(1)} \cdot \ldots \cdot w_{1r'}^{(\ell'-1)} \cdot w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix} w_{1r'}^{(1)} \cdot \ldots \cdot w_{1r'}^{(\ell'-1)} \cdot w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{1r'}^{(1)} \cdot \ldots \cdot w_{1r'}^{(\ell'-1)} \cdot w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \begin{bmatrix} w_{I_1 r'}^{(1)} \cdot \ldots \cdot w_{I_{\ell'-1}\, r'}^{(\ell'-1)} \cdot w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{I_1 r'}^{(1)} \cdot \ldots \cdot w_{I_{\ell'-1}\, r'}^{(\ell'-1)} \cdot w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix} w_{I_1 r'}^{(1)} \cdot \ldots \cdot w_{I_{\ell'-1}\, r'}^{(\ell'-1)} \cdot w_{1r'}^{(\ell'+1)} \cdot \ldots \cdot w_{1r'}^{(L)} \\ \vdots \\ w_{I_1 r'}^{(1)} \cdot \ldots \cdot w_{I_{\ell'-1}\, r'}^{(\ell'-1)} \cdot w_{I_{\ell'+1}\, r'}^{(\ell'+1)} \cdot \ldots \cdot w_{I_L r'}^{(L)} \end{bmatrix} \end{bmatrix} \end{bmatrix}.$$

This reveals the block structure of $\mathbf{J}_f$ already observed and illustrated in figures 3.1, 3.2, 3.3. Now, given any multi-index $i_1 \ldots i_{\ell'} \ldots i_L$ we just need to show that $\dfrac{\partial f_{i_1 \ldots i_{\ell'} \ldots i_L}}{\partial \mathbf{w}_{r'}^{(\ell')}}$ is equal to the negative of row $i_1 \ldots i_{\ell'} \ldots i_L$ of $-\mathbf{w}_{r'}^{(1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(\ell'-1)} \tilde{\otimes} \mathbf{I}_{I_{\ell'}} \tilde{\otimes} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)}$. Note that this row is the $i_{\ell'+1} \ldots i_L$ row of the ℓ'-th diagonal term of

$$-w_{i_1 r'}^{(1)} \cdot \ldots \cdot w_{i_{\ell'-1}\, r'}^{(\ell'-1)} \begin{bmatrix} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} & & \\ & \ddots & \\ & & \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \ldots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} \end{bmatrix}.$$

This row is given by

$$-w_{i_1 r'}^{(1)} \cdot \ldots \cdot w_{i_{\ell'-1}\, r'}^{(\ell'-1)} \cdot \big[0, \ldots, 0, \underbrace{w_{i_{\ell'+1} r'}^{(\ell'+1)} \cdot \ldots \cdot w_{i_L\, r'}^{(L)}}_{\ell'-\text{ column}}, 0 \ldots, 0\big] =$$

$$= \big[0, \ldots, 0, - \underbrace{\prod_{\ell \neq \ell'} w_{i_\ell r'}^{(\ell)}}_{\ell'-\text{ column}}, 0 \ldots, 0\big] = \frac{\partial f_{i_1 \ldots i_{\ell'} \ldots i_L}}{\partial \mathbf{w}_{r'}^{(\ell')}}. \qquad \square$$

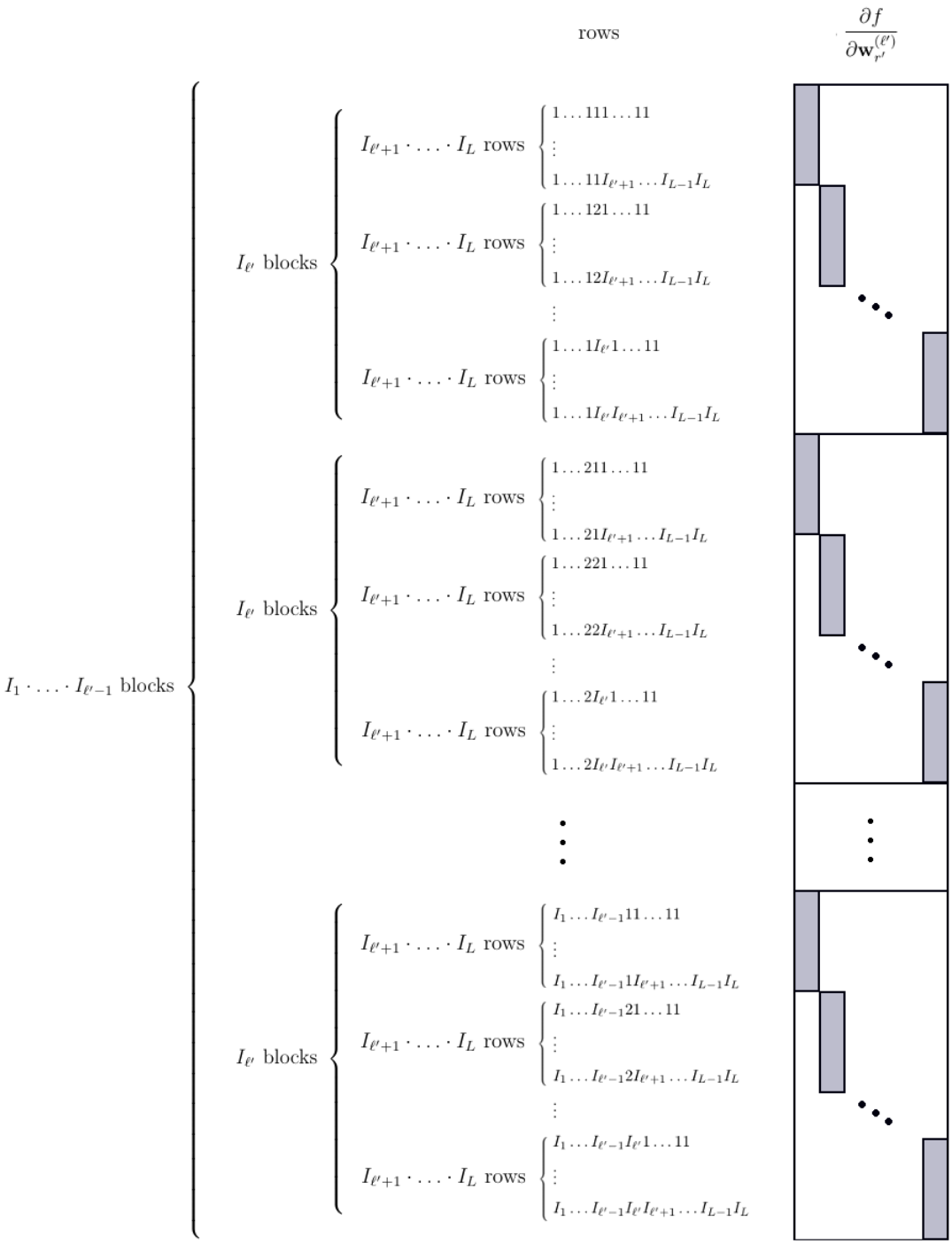

Figure 3.1: Sparse structure of $\frac{\partial f}{\partial \mathbf{w}_r^{(\ell')}}$. The gray part correspond to the non zero entries and the rest are full of zeros, and each gray column is a vector of size $\prod_{\ell=\ell'+1}^{L} I_\ell$.

Theorem 3.5.4. *$\boldsymbol{J}_f^T \boldsymbol{J}_f$ is singular.*

Proof: Notice that suffices to prove that $\mathbf{J}_f$ is not a full rank matrix. Remember that f maps possible rank-R CPDs $(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)})$ to its individual residuals, a vector with size $\prod_{\ell=1}^{L} I_\ell$. Because of the scale indeterminacy, we know that the inputs $(\lambda \mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)})$

and $(\mathbf{W}^{(1)}, \ldots, \lambda\mathbf{W}^{(\ell)}, \ldots, \mathbf{W}^{(L)})$ correspond to the same tensor for any $\ell \neq 1$. This is because the former correspond to the tensor

$$(\lambda\mathbf{w}_1^{(1)}) \otimes \ldots \otimes \mathbf{w}_1^{(L)} + \ldots + (\lambda\mathbf{w}_R^{(1)}) \otimes \ldots \otimes \mathbf{w}_R^{(L)},$$

while the latter correspond to

$$\mathbf{w}_1^{(1)} \otimes \ldots \otimes (\lambda\mathbf{w}_1^{(\ell)}) \otimes \ldots \otimes \mathbf{w}_1^{(L)} + \ldots + \mathbf{w}_R^{(1)} \otimes \ldots \otimes (\lambda\mathbf{w}_R^{(\ell)}) \otimes \ldots \otimes \mathbf{w}_R^{(L)}.$$

There are several ways of using the phenomenon of scale indeterminacy to generate different inputs to f which correspond to the same tensor. In particular, if some tensor is a local minima for F, it is also a critical point, and because of the scale indeterminacy there will be infinitely many inputs associated to the same critical point. This means the critical point is singular. Since $\nabla F = \mathbf{J}_f^T \cdot f$, we conclude that $\mathbf{J}_f$ is not of full rank at critical points. □

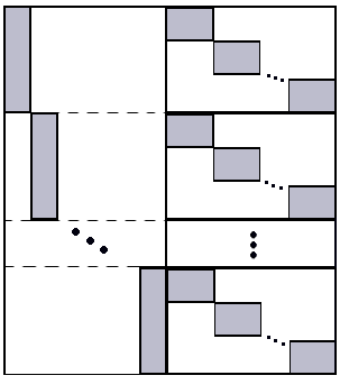

Figure 3.2: Consider the top block of the previous figure, relative to $\frac{\partial f}{\partial w_{r'}^{(\ell')}}$. While run through the rows of this block, at the same time there will be $\prod_{\ell=1}^{\ell'} I_\ell$ blocks relative to $\frac{\partial f}{\partial w_{r'}^{(\ell'+1)}}$.

Since $\mathbf{J}_f^T\mathbf{J}_f$ is singular, the Gauss-Newton update rule given by 3.6 is an ill-posed problem. We can address this issue by introducing a regularization term, so we can avoid singularity and improve convergence. A common approach is to introduce a suitable regularization matrix $\mathbf{L} \in \mathbb{R}^{R\sum_{\ell=1}^{L} I_\ell \times R\sum_{\ell=1}^{L} I_\ell}$ called *Tikhonov matrix*. Instead of solving equation 3.6 we are now solving

$$\left(\mathbf{A}^T\mathbf{A} + \mathbf{L}^T\mathbf{L}\right)\mathbf{x} = \mathbf{A}^T\mathbf{b}, \tag{3.7}$$

where $\mathbf{A} = \mathbf{J}_f(\mathbf{w}^{(k)}), \mathbf{x} = \mathbf{w} - \mathbf{w}^{(k)}, b = -f(\mathbf{w}^{(k)})$. The *damped Gauss-Newton* (dGN)[4]

[4]This algorithm is also called *Levenberg-Marquardt* in the literature. It is possible that this second

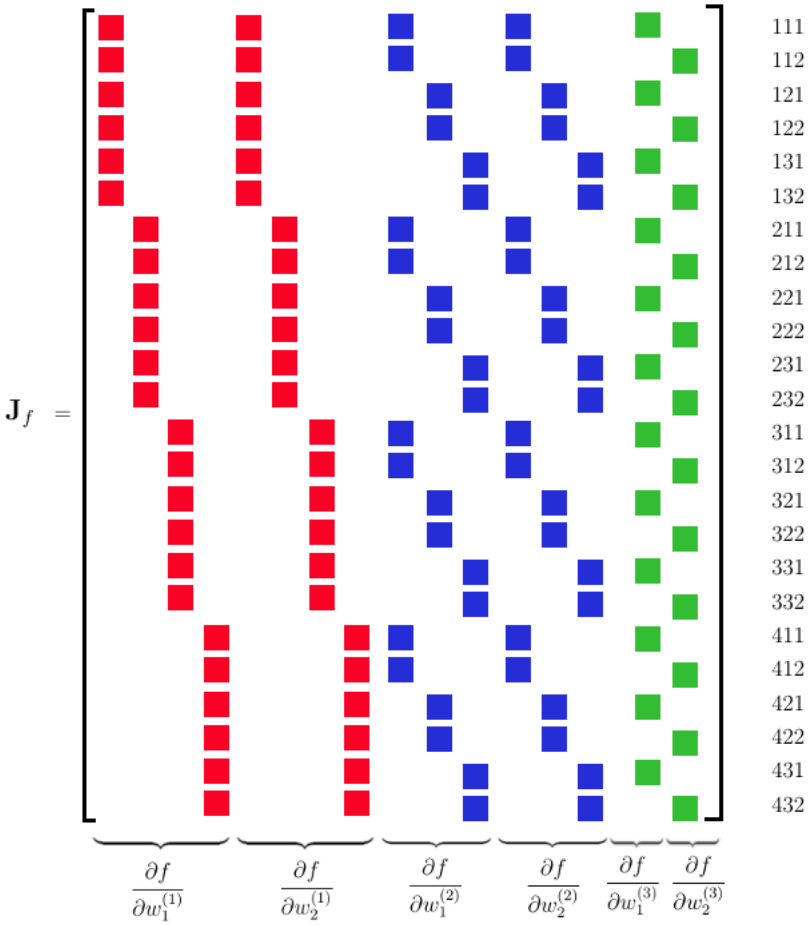

Figure 3.3: Sparse structure of $\mathbf{J}_f$ when $\mathcal{T} \in \mathbb{R}^{4\times3\times2}$ and $R = 2$.

algorithm is the case when $\mathbf{L}^T\mathbf{L} = \mu\mathbf{I}_{R\sum_{\ell=1}^{L} I_\ell}$, where $\mu > 0$ is the *damping parameter*. Usually this parameter is updated at each iteration. These updates are very important since μ influences both the direction and the size of the step at each iteration. Instead of working with the regularization matrix $\mu\mathbf{I}_{R\sum_{\ell=1}^{L} I_\ell}$ as is usual, we consider the more general matrix $\mu\mathbf{D}$, where $\mathbf{D}$ is a positive diagonal $\left(R\sum_{\ell=1}^{L} I_\ell\right) \times \left(R\sum_{\ell=1}^{L} I_\ell\right)$ matrix.

Theorem 3.5.5. *The following holds.*

1. $\boldsymbol{J}_f^T\boldsymbol{J}_f + \mu\boldsymbol{D}$ *is a positive definite matrix for all* $\mu > 0$.
2. $\boldsymbol{w}^{(k+1)} - \boldsymbol{w}^{(k)}$ *is a descent direction for* F *at* $\boldsymbol{w}^{(k)}$.
3. *If* μ *is large enough, then* $\boldsymbol{w}^{(k+1)} - \boldsymbol{w}^{(k)} \approx -\dfrac{1}{\mu}\boldsymbol{D}^{-1}\nabla F(\boldsymbol{w}^{(k)})$.
4. *If* μ *is small enough, then* $\boldsymbol{w}^{(k+1)} - \boldsymbol{w}^{(k)} \approx \boldsymbol{w}_{GN}^{(k+1)} - \boldsymbol{w}^{(k)}$, *where* $\boldsymbol{w}_{GN}^{(k+1)}$ *is the point we would obtain using classic Gauss-Newton iteration (i.e., without regularization).*

Proof: To prove 1, just take any $\mathbf{w} \in \mathbb{R}^{R\sum_{\ell=1}^{L} I_\ell}$ and note that

$$\begin{aligned}
&\left\langle \left(\mathbf{J}_f^T\mathbf{J}_f + \mu\mathbf{D}\right)\mathbf{w}, \mathbf{w}\right\rangle = \\
&= \left\langle \mathbf{J}_f^T\mathbf{J}_f\mathbf{w}, \mathbf{w}\right\rangle + \left\langle \mu\mathbf{D}\mathbf{w}, \mathbf{w}\right\rangle = \\
&= \left\langle \mathbf{J}_f\mathbf{w}, \mathbf{J}_f\mathbf{w}\right\rangle + \left\langle \sqrt{\mu}\sqrt{\mathbf{D}}\mathbf{w}, \sqrt{\mu}\sqrt{\mathbf{D}}\mathbf{w}\right\rangle = \\
&= \|\mathbf{J}_f\mathbf{w}\|^2 + \|\sqrt{\mu}\sqrt{\mathbf{D}}\mathbf{w}\|^2 > 0.
\end{aligned}$$

The proof of 2 is very similar to the previous proof in the classic Gauss-Newton. From the iteration formula

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \left(\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)}) + \mu\mathbf{D}\right)^{-1}\mathbf{J}_f(\mathbf{w}^{(k)})^T \cdot f(\mathbf{w}^{(k)})$$

we can conclude that

$$-\left(\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)}) + \mu\mathbf{D}\right) \cdot \left(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right) = \mathbf{J}_f(\mathbf{w}^{(k)})^T \cdot f(\mathbf{w}^{(k)}).$$

Now, with this identity, note that

$$\begin{aligned}
&\left\langle \nabla F(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right\rangle = \\
&= \left\langle \mathbf{J}_f(\mathbf{w}^{(k)})^T f(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right\rangle =
\end{aligned}$$

name is used more often but still we prefer the other. The name comes from the fact that K. Levenberg [68] and D. Marquardt [69] are the ones responsible for introducing the damping parameter in the Gauss-Newton method.

$$= -\left\langle \left(\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)}) + \mu\mathbf{D}\right)\cdot\left(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right\rangle < 0.$$

The inequality above follows from the fact that $\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)}) + \mu\mathbf{D}$ is positive definite.

To prove 3, take μ such that $\|\mathbf{D}^{-1}\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)})\| \ll \mu$ (this is "large enough" in this context). We know $\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)}) + \mu\mathbf{D}$ since it is positive definite. Also, by the definition of μ we have that

$$\left(\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)}) + \mu\mathbf{D}\right)^{-1} = \left(\mu\mathbf{D}\left(\frac{1}{\mu}\mathbf{D}^{-1}\mathbf{J}_f(\mathbf{w}^{(k)})^T\mathbf{J}_f(\mathbf{w}^{(k)}) + \mathbf{I}\right)\right)^{-1} \approx$$

$$\approx \left(\mu\mathbf{D}\left(\mathbf{0} + \mathbf{I}\right)\right)^{-1} = \frac{1}{\mu}\mathbf{D}^{-1}.$$

Using the iteration formula with this approximation gives

$$\mathbf{w}^{(k+1)} \approx \mathbf{w}^{(k)} - \frac{1}{\mu}\mathbf{D}^{-1}\mathbf{J}_f(\mathbf{w}^{(k)})^T f(\mathbf{w}^{(k)}) = \mathbf{w}^{(k)} - \frac{1}{\mu}\mathbf{D}^{-1}\nabla F(\mathbf{w}^{-1}).$$

Finally, to prove 4 just consider $\mu \approx 0$ and substitute in the iteration formula. Then we get the classical formula trivially. □

Remark 3.5.6. *If $\boldsymbol{D} \approx \boldsymbol{I}$, item 3 can be used when the current iteration is far from the solution, since $-\frac{1}{\mu}\nabla F(\boldsymbol{w}^{(k)})$ is a short step in the descent direction (we want to be careful when distant to the solution). This shows that dGN behaves as the gradient descent algorithm when distant to the solution. On the other hand, item 4 is to be used when the current iteration is close to the solution, since the step is closer to the classic Gauss-Newton, we may attain quadratic convergence at the final iterations.*

The damping parameter rule update is a very important issue in this algorithm and we should comment a few things about it (the interested reader may check [22] for more about this subject). Let $\mu^{(0)}$ be the initial damping parameter and $\mu^{(k)}$ be the damping parameter used at the k-th iteration. A rule of thumb for $\mu^{(0)}$ is to set $\mu^{(0)} = \tau \cdot \max_i a_{ii}$, where $\mathbf{A} = \mathbf{J}_f^T(\mathbf{w}^{(0)})\mathbf{J}_f(\mathbf{w}^{(0)})$ and $\tau \in (0, 1]$. We should choose τ smaller when $\mathbf{w}^{(0)}$ is closer to the solution. The algorithm usually is not so sensitive to the choice of τ so there is no problem in using other initial values. More important is the update rule for $\mu^{(k)}$.

After computing $\mathbf{w}^{(k+1)}$, the *gain ratio* is defined as

$$g = \frac{F(\mathbf{w}^{(k)}) - F(\mathbf{w}^{(k+1)})}{F(\mathbf{w}^{(k)}) - \frac{1}{2}\|\tilde{f}(\mathbf{w}^{(k+1)})\|^2} = \frac{\|\mathcal{T} - \tilde{\mathcal{T}}^{(k)}\|^2 - \|\mathcal{T} - \tilde{\mathcal{T}}^{(k+1)}\|^2}{\|\mathcal{T} - \tilde{\mathcal{T}}^{(k)}\|^2 - \|\tilde{f}(\mathbf{w}^{(k+1)})\|^2},$$

where $\tilde{\mathcal{T}}^{(k)}$ is the approximating tensor computed at the k-th iteration and $\tilde{f}(\mathbf{w}^{(k+1)}) = f(\mathbf{w}^{(k)}) + \mathbf{J}_f(\mathbf{w}^{(k)}) \cdot (\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)})$ is the first order approximation of f at $\mathbf{w}^{(k+1)}$. The

denominator is the predicted improvement[5] from the k-th to $(k+1)$-th iteration, whereas the numerator measures the actual improvement.

A large value of g indicates that $\|\tilde{f}(\mathbf{w}^{(k+1)})\|^2$ is a good approximation to $\|\mathcal{T}-\tilde{\mathcal{T}}^{(k+1)}\|^2$, in other words, the linear model $\tilde{f}$ is making good predictions about the actual errors. A small or negative g indicates that this approximation is poor. In the former case we decrease the damping parameter so the next iteration is more like a Gauss-Newton iteration, and in the latter case we increase the damping parameter to have more regularization and make smaller steps, this way the steps are more guaranteed to be in the steepest direction. This is the general idea, but in reality these updates depends a lot on the problem at hand and there is not a single procedure which works for all problems. We will talk more about it when introducing our own algorithm to compute the CPD.

One thing to notice is that the denominator is always positive by construction since

$$F(\mathbf{w}^{(k)}) - \|\tilde{f}(\mathbf{w}^{(k+1)})\|^2 = F(\mathbf{w}^{(k)}) - \|f(\mathbf{w}^{(k)}) + \mathbf{J}_f(\mathbf{w}^{(k)}) \cdot (\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)})\|^2 =$$

$$= F(\mathbf{w}^{(k)}) - \|f(\mathbf{w}^{(k)})\|^2 - 2\langle f(\mathbf{w}^{(k)}), \mathbf{J}_f(\mathbf{w}^{(k)})(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)})\rangle - \|\mathbf{J}_f(\mathbf{w}^{(k)})(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)})\|^2 =$$

$$= -2\langle f(\mathbf{w}^{(k)}), \mathbf{J}_f(\mathbf{w}^{(k)})(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)})\rangle - \|\mathbf{J}_f(\mathbf{w}^{(k)})(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)})\|^2 =$$

$$= -2\langle \nabla F(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rangle + \langle \nabla F(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rangle =$$

$$= -\langle \nabla F(\mathbf{w}^{(k)}), \mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\rangle =$$

$$= \|\mathbf{J}_f(\mathbf{w}^{(k)})(\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)})\|^2 > 0.$$

Notice we used a few identities from theorem 3.5.1 to make some of the manipulations above. The important thing here is to realize that g will be negative only when the numerator is negative, and this means the error of the approximation increases at iteration $k+1$. In section 4.7 we will see that it is expected that $g < 0$ in a few iterations, and we argue that this apparent drawback can actually be beneficial.

Different versions of the dGN algorithm were implemented and tested [4, 14, 15, 17, 70, 71]. It is already known that dGN have faster local convergence when close to the optimal point since it uses information from the (approximated) Hessian. In the presence of bottlenecks or swamps the ALS presents problems, and even with improvements such as regularization, line search, rotation, etc, these problems doesn't disappear completely. A severe limitation of the ALS is the fact that it updates only one factor per iteration. The dGN uses all information to make the updates, and this leads to a more robust algorithm which is insensible to bottlenecks and swamps. It is also observed that dGN

[5] Remember that we obtain $\mathbf{w}^{(k+1)}$ as a solution to the minimization problem $\min_{\mathbf{w}} \|\tilde{f}(\mathbf{w})\|^2$ so we expect it to be close to $\min_{\mathbf{w}} \|f(\mathbf{w})\|^2 = \min_{\mathbf{w}} F(\mathbf{w})$. Therefore $\|\tilde{f}(\mathbf{w}^{(k+1)})\|^2$ gives the expected error while $F(\mathbf{w}^{(k+1)})$ gives the actual error.

is less sensitive to the initialization, whereas ALS is highly sensitive, and dGN is more accurate than ALS in general [15, 71]. This means dGN is a more robust and reliable algorithm in every sense. The drawback of the dGN algorithm is the computation of a solution to 3.7 at each iteration. To solve these normal equations we have to deal with a big and dense matrix, which is computationally costly. In the next section we will present our approach to this problem.

3.5.3 Dealing with the Hessian

To work the normal equations 3.7 we need to exploit some structure of $\mathbf{J}_f^T\mathbf{J}_f$ in order to make fast computations with low memory cost. The next theorem is a first step in this direction.

Theorem 3.5.7. *Denote* $\omega_{r'r''}^{(\ell)} = \langle \boldsymbol{w}_{r'}^{(\ell)}, \boldsymbol{w}_{r''}^{(\ell)} \rangle$. *Then we have that*

$$\boldsymbol{J}_f^T\boldsymbol{J}_f = \begin{bmatrix} \boldsymbol{H}_{11} & \dots & \boldsymbol{H}_{1L} \\ \vdots & & \vdots \\ \boldsymbol{H}_{L1} & \dots & \boldsymbol{H}_{LL} \end{bmatrix},$$

where

$$\boldsymbol{H}_{\ell'\ell''} = \begin{bmatrix} \prod\limits_{\ell\neq\ell',\ell''} \omega_{11}^{(\ell)} \cdot \boldsymbol{w}_1^{(\ell')}\boldsymbol{w}_1^{(\ell'')^T} & \dots & \prod\limits_{\ell\neq\ell',\ell''} \omega_{1R}^{(\ell)} \cdot \boldsymbol{w}_R^{(\ell')}\boldsymbol{w}_1^{(\ell'')^T} \\ \vdots & & \vdots \\ \prod\limits_{\ell\neq\ell',\ell''} \omega_{R1}^{(\ell)} \cdot \boldsymbol{w}_1^{(\ell')}\boldsymbol{w}_R^{(\ell'')^T} & \dots & \prod\limits_{\ell\neq\ell',\ell''} \omega_{RR}^{(\ell)} \cdot \boldsymbol{w}_R^{(\ell')}\boldsymbol{w}_R^{(\ell'')^T} \end{bmatrix}$$

for $\ell' \neq \ell''$, *and*

$$\boldsymbol{H}_{\ell'\ell'} = \begin{bmatrix} \prod\limits_{\ell\neq\ell'} \omega_{11}^{(\ell)} \cdot \boldsymbol{I}_{I_{\ell'}} & \dots & \prod\limits_{\ell\neq\ell'} \omega_{1R}^{(\ell)} \cdot \boldsymbol{I}_{I_{\ell'}} \\ \vdots & & \vdots \\ \prod\limits_{\ell\neq\ell'} \omega_{R1}^{(\ell)} \cdot \boldsymbol{I}_{I_{\ell'}} & \dots & \prod\limits_{\ell\neq\ell'} \omega_{RR}^{(\ell)} \cdot \boldsymbol{I}_{I_{\ell'}} \end{bmatrix}.$$

Proof: First notice that

$$\mathbf{J}_f^T\mathbf{J}_f = \begin{bmatrix} \dfrac{\partial f}{\partial \mathbf{W}^{(1)}}^T \\ \vdots \\ \dfrac{\partial f}{\partial \mathbf{W}^{(L)}}^T \end{bmatrix} \begin{bmatrix} \dfrac{\partial f}{\partial \mathbf{W}^{(1)}}, \dots, \dfrac{\partial f}{\partial \mathbf{W}^{(L)}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f}{\partial \mathbf{W}^{(1)}}^T \dfrac{\partial f}{\partial \mathbf{W}^{(1)}} & \dots & \dfrac{\partial f}{\partial \mathbf{W}^{(1)}}^T \dfrac{\partial f}{\partial \mathbf{W}^{(L)}} \\ \vdots & & \vdots \\ \dfrac{\partial f}{\partial \mathbf{W}^{(L)}}^T \dfrac{\partial f}{\partial \mathbf{W}^{(1)}} & \dots & \dfrac{\partial f}{\partial \mathbf{W}^{(L)}}^T \dfrac{\partial f}{\partial \mathbf{W}^{(L)}} \end{bmatrix},$$

where

$$\frac{\partial f}{\partial \mathbf{W}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell'')}} = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{w}_1^{(\ell')}}^T \\ \vdots \\ \frac{\partial f}{\partial \mathbf{w}_R^{(\ell')}}^T \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial \mathbf{w}_1^{(\ell'')}}, \dots, \frac{\partial f}{\partial \mathbf{w}_R^{(\ell'')}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{w}_1^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{w}_1^{(\ell'')}} & \dots & \frac{\partial f}{\partial \mathbf{w}_1^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{w}_R^{(\ell'')}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial \mathbf{w}_R^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{w}_1^{(\ell'')}} & \dots & \frac{\partial f}{\partial \mathbf{w}_R^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{w}_R^{(\ell'')}} \end{bmatrix}.$$

Let $\omega_{r'r''}^{(\ell)} = \langle \mathbf{w}_{r'}^{(\ell)}, \mathbf{w}_{r''}^{(\ell)} \rangle$ and assume, without loss of generality, that $1 \leq \ell' < \ell'' \leq L$. We can use lemma 3.5.3 and theorem B.3.5 to write the entries of this matrix as

$$\frac{\partial f}{\partial \mathbf{w}_{r'}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{w}_{r''}^{(\ell'')}} =$$

$$= \left(\mathbf{w}_{r'}^{(1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(\ell'-1)} \tilde{\otimes} \mathbf{I}_{I_{\ell'}} \tilde{\otimes} \mathbf{w}_{r'}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r'}^{(L)} \right)^T \left(\mathbf{w}_{r''}^{(1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r''}^{(\ell''-1)} \tilde{\otimes} \mathbf{I}_{I_{\ell''}} \tilde{\otimes} \mathbf{w}_{r''}^{(\ell''+1)} \tilde{\otimes} \dots \tilde{\otimes} \mathbf{w}_{r''}^{(L)} \right) =$$

$$= \omega_{r'r''}^{(1)} \tilde{\otimes} \dots \tilde{\otimes} \omega_{r'r''}^{(\ell'-1)} \tilde{\otimes} \left(\mathbf{I}_{\ell'} \mathbf{w}_{r''}^{(\ell')} \right) \tilde{\otimes} \omega_{r'r''}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \omega_{r'r''}^{(\ell''-1)} \tilde{\otimes} \left(\mathbf{w}_{r'}^{(\ell'')^T} \mathbf{I}_{\ell''} \right) \tilde{\otimes} \omega_{r'r''}^{(\ell''+1)} \tilde{\otimes} \dots \tilde{\otimes} \omega_{r'r''}^{(L)} =$$

$$= \prod_{\ell \neq \ell', \ell''} \omega_{r'r''}^{(\ell)} \cdot \mathbf{w}_{r''}^{(\ell')} \mathbf{w}_{r'}^{(\ell'')^T}.$$

In the case $\ell' = \ell''$ we have

$$\frac{\partial f}{\partial \mathbf{w}_{r'}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{w}_{r''}^{(\ell')}} =$$

$$= \omega_{r'r''}^{(1)} \tilde{\otimes} \dots \tilde{\otimes} \omega_{r'r''}^{(\ell'-1)} \tilde{\otimes} \mathbf{I}_{I_{\ell'}}^2 \tilde{\otimes} \omega_{r'r''}^{(\ell'+1)} \tilde{\otimes} \dots \tilde{\otimes} \omega_{r'r''}^{(L)} =$$

$$= \prod_{\ell \neq \ell'} \omega_{r'r''}^{(\ell)} \cdot \mathbf{I}_{I_{\ell'}}.$$

Therefore,

$$\frac{\partial f}{\partial \mathbf{W}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell'')}} = \begin{bmatrix} \prod_{\ell \neq \ell', \ell''} \omega_{11}^{(\ell)} \cdot \mathbf{w}_1^{(\ell')} \mathbf{w}_1^{(\ell'')^T} & \dots & \prod_{\ell \neq \ell', \ell''} \omega_{1R}^{(\ell)} \cdot \mathbf{w}_R^{(\ell')} \mathbf{w}_1^{(\ell'')^T} \\ \vdots & & \vdots \\ \prod_{\ell \neq \ell', \ell''} \omega_{R1}^{(\ell)} \cdot \mathbf{w}_1^{(\ell')} \mathbf{w}_R^{(\ell'')^T} & \dots & \prod_{\ell \neq \ell', \ell''} \omega_{RR}^{(\ell)} \cdot \mathbf{w}_R^{(\ell')} \mathbf{w}_R^{(\ell'')^T} \end{bmatrix}$$

when $\ell' \neq \ell''$. Finally, we have that

$$\frac{\partial f}{\partial \mathbf{W}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell')}} = \begin{bmatrix} \prod_{\ell \neq \ell'} \omega_{11}^{(\ell)} \cdot \mathbf{I}_{I_{\ell'}} & \dots & \prod_{\ell \neq \ell'} \omega_{1R}^{(\ell)} \cdot \mathbf{I}_{I_{\ell'}} \\ \vdots & & \vdots \\ \prod_{\ell \neq \ell'} \omega_{R1}^{(\ell)} \cdot \mathbf{I}_{I_{\ell'}} & \dots & \prod_{\ell \neq \ell'} \omega_{RR}^{(\ell)} \cdot \mathbf{I}_{I_{\ell'}} \end{bmatrix}. \qquad \square$$

Remark 3.5.8. *Each $\boldsymbol{H}_{\ell'\ell''}$ is block $R \times R$ matrix, where each of its blocks is a $I_{\ell'} \times I_{\ell''}$ matrix, so $\boldsymbol{H}_{\ell'\ell''}$ has shape $I_{\ell'}R \times I_{\ell''}R$. Since $\boldsymbol{J}_f^T \boldsymbol{J}_f$ is a block $L \times L$ matrix, with the (ℓ', ℓ'') block being $\boldsymbol{H}_{\ell'\ell''}$, we conclude that $\boldsymbol{J}_f^T \boldsymbol{J}_f$ has shape $R\sum_{\ell=1}^{L} I_\ell \times R\sum_{\ell=1}^{L} I_\ell$. This shape is much smaller than the shape of $\boldsymbol{J}_f$, which is of $\prod_{\ell=1}^{L} I_\ell \times R\sum_{\ell=1}^{L} I_\ell$, hence we avoid the curse of dimensionality with this approach. It should be noted that this is only true if $R < \frac{\prod_{\ell=1}^{L} I_\ell}{\sum_{\ell=1}^{L} I_\ell}$. This will be almost always the case since this is the same as saying R is smaller than the generic rank, and in fact almost always we will choose R to satisfy this property. Finally, we want to remark that the notation $\boldsymbol{H}_{\ell'\ell''}$ comes from the fact that $\boldsymbol{J}_f^T \boldsymbol{J}_f \approx \boldsymbol{H}_F$ as we converges to the solution.*

It is convenient to store the products of the terms $\omega_{r'r''}^{(\ell)}$ in matrix form, so we define

$$\Pi^{(\ell',\ell'')} = \begin{bmatrix} \prod_{\ell \neq \ell', \ell''} \omega_{11}^{(\ell)} & \dots & \prod_{\ell \neq \ell', \ell''} \omega_{1R}^{(\ell)} \\ \vdots & & \vdots \\ \prod_{\ell \neq \ell', \ell''} \omega_{R1}^{(\ell)} & \dots & \prod_{\ell \neq \ell', \ell''} \omega_{RR}^{(\ell)} \end{bmatrix}$$

for $\ell' \neq \ell''$, and

$$\Pi^{(\ell')} = \begin{bmatrix} \prod_{\ell \neq \ell'} \omega_{11}^{(\ell)} & \dots & \prod_{\ell \neq \ell'} \omega_{1R}^{(\ell)} \\ \vdots & & \vdots \\ \prod_{\ell \neq \ell'} \omega_{R1}^{(\ell)} & \dots & \prod_{\ell \neq \ell'} \omega_{RR}^{(\ell)} \end{bmatrix}.$$

Notice that these (symmetric) matrices are the Hadamard product of some Gramian matrices. Define $\pi^{(\ell)} = \mathbf{W}^{(\ell)^T} \mathbf{W}^{(\ell)}$. Then we have that

$$\Pi^{(\ell',\ell'')} = \pi^{(1)} * \ldots * \pi^{(\ell'-1)} * \pi^{(\ell'+1)} * \ldots * \pi^{(\ell''-1)} * \pi^{(\ell''+1)} * \ldots * \pi^{(L)}$$

and

$$\Pi^{(\ell')} = \pi^{(1)} * \ldots * \pi^{(\ell'-1)} * \pi^{(\ell'+1)} * \ldots * \pi^{(L)}.$$

Now we can see that is possible to retrieve $\mathbf{J}_f^T\mathbf{J}_f$ from the factor matrices with few computations and low memory cost. First we compute and store all the products $\mathbf{W}^{(\ell)^T}\mathbf{W}^{(\ell)}$ and this has a total computational cost of $\mathcal{O}\left(R^2\sum_{\ell=1}^{L} I_\ell\right)$. To compute each $\Pi^{(\ell',\ell'')}$ we have to perform $L-2$ Hadamard products between $R\times R$ matrices, and this has a cost of $\mathcal{O}\left((L-2)R^2\right)$ flops. The cost to store all the matrices $\Pi^{(\ell',\ell')}$ is of L^2R^2 floats. We remark that it is possible to cut all these costs by half since $\Pi^{(\ell',\ell'')}=\Pi^{(\ell'',\ell')}$ for all $\ell'\neq\ell''$. Finally, note that $\Pi^{(\ell')}=\Pi^{(\ell',\ell'')}*\left(\mathbf{W}^{(\ell'')^T}\mathbf{W}^{(\ell'')}\right)$, so we can construct $\Pi^{(\ell')}$ with only one Hadamard product, which has a cost of $\mathcal{O}(R^2)$ flops. Doing this for all ℓ amounts to $\mathcal{O}(LR^2)$ flops. The memory cost to store each $\Pi^{(\ell')}$ is the same of $\Pi^{(\ell',\ell')}$, so we have a total memory cost of $\mathcal{O}(LR^2)$ floats. Overall, the computational cost to compute all $\Pi^{(\ell',\ell'')}$ and all $\Pi^{(\ell')}$ is of $\mathcal{O}\left(R^2\left(L+\sum_{\ell=1}^{L} I_\ell\right)\right)$ flops, and the total memory cost is of $\mathcal{O}\left(R^2(L+L^2)\right)$ floats.

By writing the matrices $\Pi^{(\ell',\ell'')}$ and $\Pi^{(\ell')}$ as the result of many Hadamard products we have a way to simplify the expressions for the blocks $\mathbf{H}_{\ell'\ell''}$.

Corollary 3.5.9. *Let $\mathbf{1}_{m\times n}$ be the $m\times n$ matrix constituted only by ones and define*

$$\boldsymbol{K}^{(\ell',\ell'')}=\begin{bmatrix}\boldsymbol{w}_1^{(\ell')}\boldsymbol{w}_1^{(\ell'')^T} & \dots & \boldsymbol{w}_R^{(\ell')}\boldsymbol{w}_1^{(\ell'')^T}\\ \vdots & & \vdots\\ \boldsymbol{w}_1^{(\ell')}\boldsymbol{w}_R^{(\ell'')^T} & \dots & \boldsymbol{w}_R^{(\ell')}\boldsymbol{w}_R^{(\ell'')^T}\end{bmatrix}$$

for $\ell'\neq\ell''$. Then

$$\boldsymbol{H}_{\ell',\ell''}=\left(\Pi^{(\ell',\ell'')}\tilde{\otimes}\mathbf{1}_{I_{\ell'}\times I_{\ell''}}\right)*\boldsymbol{K}^{(\ell',\ell'')}$$

for $\ell'\neq\ell''$, and

$$\boldsymbol{H}_{\ell',\ell'}=\Pi^{(\ell')}\tilde{\otimes}\boldsymbol{I}_{\ell'}.$$

As already mentioned, $\mathbf{J}_f^T\mathbf{J}_f$ will be used to solve the normal equations 3.7. The algorithm of choice to accomplish this is the conjugate gradient method (see appendix A). This classical algorithm is particularly efficient to solve normal equations where the matrix is positive definite, which is our case. Furthermore, our version of the conjugate gradient is *matrix-free*, that is, we are able to compute matrix-vector products $\mathbf{J}_f^T\mathbf{J}_f\cdot\mathbf{v}$ without actually constructing $\mathbf{J}_f^T\mathbf{J}_f$. By exploiting the block structure of $\mathbf{J}_f^T\mathbf{J}_f$ we can save memory and the computational cost still is lower than the naive cost of $R^2\left(\sum_{\ell=1}^{L} I_\ell\right)^2$ flops.

The next theorem is a new contribution of this work. With this result we are able to perform fast matrix-vector products with low memory cost.

Theorem 3.5.10. *Given any vector* $\boldsymbol{v} \in \mathbb{R}^{R\sum_{\ell=1}^{L} I_\ell}$*, write*

$$\boldsymbol{v} = \begin{bmatrix} vec(\boldsymbol{V}^{(1)}) \\ \vdots \\ vec(\boldsymbol{V}^{(L)}) \end{bmatrix}$$

where $\boldsymbol{V}^{(\ell)} = [\boldsymbol{v}_1^{(\ell)}, \ldots, \boldsymbol{v}_R^{(\ell)}] \in \mathbb{R}^{I_\ell \times R}$ *and each* $\boldsymbol{v}_r^{(\ell)}$ *is a column of* $\boldsymbol{V}^{(\ell)}$*. Then*

$$\boldsymbol{J}_f^T \boldsymbol{J}_f \cdot \boldsymbol{v} = \begin{bmatrix} \sum_{\ell=1}^{L} \frac{\partial f}{\partial \boldsymbol{W}^{(1)}}^T \frac{\partial f}{\partial \boldsymbol{W}^{(\ell)}} \cdot vec(\boldsymbol{V}^{(\ell)}) \\ \vdots \\ \sum_{\ell=1}^{L} \frac{\partial f}{\partial \boldsymbol{W}^{(L)}}^T \frac{\partial f}{\partial \boldsymbol{W}^{(\ell)}} \cdot vec(\boldsymbol{V}^{(\ell)}) \end{bmatrix}$$

where

$$\frac{\partial f}{\partial \boldsymbol{W}^{(\ell')}}^T \frac{\partial f}{\partial \boldsymbol{W}^{(\ell'')}} \cdot vec(\boldsymbol{V}^{(\ell'')}) = vec\left(\boldsymbol{W}^{(\ell')} \cdot \left(\Pi^{(\ell',\ell'')} * \left(\boldsymbol{V}^{(\ell'')^T} \cdot \boldsymbol{W}^{(\ell'')}\right)\right)\right)$$

for $\ell' \neq \ell''$ *and*

$$\frac{\partial f}{\partial \boldsymbol{W}^{(\ell')}}^T \frac{\partial f}{\partial \boldsymbol{W}^{(\ell')}} \cdot vec(\boldsymbol{V}^{(\ell')}) = vec\left(\boldsymbol{V}^{(\ell')} \cdot \Pi^{(\ell')}\right).$$

Proof: First notice that

$$\mathbf{J}_f^T \mathbf{J}_f \cdot \mathbf{v} = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{W}^{(1)}}^T \frac{\partial f}{\partial \mathbf{W}^{(1)}} & \cdots & \frac{\partial f}{\partial \mathbf{W}^{(1)}}^T \frac{\partial f}{\partial \mathbf{W}^{(L)}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial \mathbf{W}^{(L)}}^T \frac{\partial f}{\partial \mathbf{W}^{(1)}} & \cdots & \frac{\partial f}{\partial \mathbf{W}^{(L)}}^T \frac{\partial f}{\partial \mathbf{W}^{(L)}} \end{bmatrix} \begin{bmatrix} vec(\mathbf{V}^{(1)}) \\ \vdots \\ vec(\mathbf{V}^{(L)}) \end{bmatrix} =$$

$$= \begin{bmatrix} \sum_{\ell=1}^{L} \frac{\partial f}{\partial \mathbf{W}^{(1)}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell)}} \cdot vec(\mathbf{V}^{(\ell)}) \\ \vdots \\ \sum_{\ell=1}^{L} \frac{\partial f}{\partial \mathbf{W}^{(L)}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell)}} \cdot vec(\mathbf{V}^{(\ell)}) \end{bmatrix}.$$

Now we simplify each term in the summation above. It is necessary to consider two separate cases.

Case 1 (different modes): If $\ell' \neq \ell''$, then

$$\frac{\partial f}{\partial \mathbf{W}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell'')}} \cdot vec(\mathbf{V}^{(\ell'')}) =$$

$$= \begin{bmatrix} \prod\limits_{\ell \neq \ell', \ell''} \omega_{11}^{(\ell)} \cdot \mathbf{w}_1^{(\ell')} \mathbf{w}_1^{(\ell'')^T} & \dots & \prod\limits_{\ell \neq \ell', \ell''} \omega_{1R}^{(\ell)} \cdot \mathbf{w}_R^{(\ell')} \mathbf{w}_1^{(\ell'')^T} \\ \vdots & & \vdots \\ \prod\limits_{\ell \neq \ell', \ell''} \omega_{R1}^{(\ell)} \cdot \mathbf{w}_1^{(\ell')} \mathbf{w}_R^{(\ell'')^T} & \dots & \prod\limits_{\ell \neq \ell', \ell''} \omega_{RR}^{(\ell)} \cdot \mathbf{w}_R^{(\ell')} \mathbf{w}_R^{(\ell'')^T} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^{(\ell'')} \\ \vdots \\ \mathbf{v}_R^{(\ell'')} \end{bmatrix} =$$

$$= \begin{bmatrix} \sum\limits_{r=1}^{R} \prod\limits_{\ell \neq \ell', \ell''} \omega_{1r}^{(\ell)} \cdot \mathbf{w}_r^{(\ell')} \mathbf{w}_1^{(\ell'')^T} \cdot \mathbf{v}_r^{(\ell'')} \\ \vdots \\ \sum\limits_{r=1}^{R} \prod\limits_{\ell \neq \ell', \ell''} \omega_{Rr}^{(\ell)} \cdot \mathbf{w}_r^{(\ell')} \mathbf{w}_R^{(\ell'')^T} \cdot \mathbf{v}_r^{(\ell'')} \end{bmatrix} = \begin{bmatrix} \sum\limits_{r=1}^{R} \prod\limits_{\ell \neq \ell', \ell''} \omega_{1r}^{(\ell)} \cdot \mathbf{w}_r^{(\ell')} \langle \mathbf{w}_1^{(\ell'')}, \mathbf{v}_r^{(\ell'')} \rangle \\ \vdots \\ \sum\limits_{r=1}^{R} \prod\limits_{\ell \neq \ell', \ell''} \omega_{Rr}^{(\ell)} \cdot \mathbf{w}_r^{(\ell')} \langle \mathbf{w}_R^{(\ell'')}, \mathbf{v}_r^{(\ell'')} \rangle \end{bmatrix} =$$

$$= \begin{bmatrix} \left[\mathbf{w}_1^{(\ell')}, \dots, \mathbf{w}_R^{(\ell')} \right] \begin{bmatrix} \prod\limits_{\ell \neq \ell', \ell''} \omega_{11}^{(\ell)} \langle \mathbf{w}_1^{(\ell'')}, \mathbf{v}_1^{(\ell'')} \rangle \\ \vdots \\ \prod\limits_{\ell \neq \ell', \ell''} \omega_{1R}^{(\ell)} \langle \mathbf{w}_1^{(\ell'')}, \mathbf{v}_R^{(\ell'')} \rangle \end{bmatrix} \\ \vdots \\ \left[\mathbf{w}_1^{(\ell')}, \dots, \mathbf{w}_R^{(\ell')} \right] \begin{bmatrix} \prod\limits_{\ell \neq \ell', \ell''} \omega_{R1}^{(\ell)} \langle \mathbf{w}_R^{(\ell'')}, \mathbf{v}_1^{(\ell'')} \rangle \\ \vdots \\ \prod\limits_{\ell \neq \ell', \ell''} \omega_{RR}^{(\ell)} \langle \mathbf{w}_R^{(\ell'')}, \mathbf{v}_R^{(\ell'')} \rangle \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathbf{W}^{(\ell')} \cdot \left(\Pi_1^{(\ell', \ell'')} * \left(\mathbf{V}^{(\ell'')^T} \cdot \mathbf{w}_1^{(\ell'')} \right) \right) \\ \vdots \\ \mathbf{W}^{(\ell')} \cdot \left(\Pi_R^{(\ell', \ell'')} * \left(\mathbf{V}^{(\ell'')^T} \cdot \mathbf{w}_R^{(\ell'')} \right) \right) \end{bmatrix} =$$

$$= vec\left(\left[\mathbf{W}^{(\ell')} \cdot \left(\Pi_1^{(\ell', \ell'')} * \left(\mathbf{V}^{(\ell'')^T} \cdot \mathbf{w}_1^{(\ell'')} \right) \right), \dots, \mathbf{W}^{(\ell')} \cdot \left(\Pi_R^{(\ell', \ell'')} * \left(\mathbf{V}^{(\ell'')^T} \cdot \mathbf{w}_R^{(\ell'')} \right) \right) \right] \right) =$$

$$= vec\left(\mathbf{W}^{(\ell')} \cdot \left(\Pi^{(\ell', \ell'')} * \left(\mathbf{V}^{(\ell'')^T} \cdot \mathbf{W}^{(\ell'')} \right) \right) \right)$$

where each $\Pi_{r'}^{(\ell', \ell'')}$ is the r'-th column of $\Pi^{(\ell', \ell'')}$. Despite the notation refers to the rows of $\Pi^{(\ell', \ell'')}$, this is not a problem since this matrix is symmetric.

Case 2 (equal modes): For a mode ℓ' we have

$$\frac{\partial f}{\partial \mathbf{W}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell')}} \cdot vec(\mathbf{V}^{(\ell')}) =$$

$$= \begin{bmatrix} \prod_{\ell \neq \ell'} \omega_{11}^{(\ell)} \cdot \mathbf{I}_{I'_\ell} & \dots & \prod_{\ell \neq \ell'} \omega_{1R}^{(\ell)} \cdot \mathbf{I}_{I'_\ell} \\ \vdots & & \vdots \\ \prod_{\ell \neq \ell'} \omega_{R1}^{(\ell)} \cdot \mathbf{I}_{I'_\ell} & \dots & \prod_{\ell \neq \ell'} \omega_{RR}^{(\ell)} \cdot \mathbf{I}_{I'_\ell} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^{(\ell')} \\ \vdots \\ \mathbf{v}_R^{(\ell')} \end{bmatrix} = \begin{bmatrix} \sum_{r=1}^{R} \prod_{\ell \neq \ell'} \omega_{1r}^{(\ell)} \cdot \mathbf{v}_r^{(\ell')} \\ \vdots \\ \sum_{r=1}^{R} \prod_{\ell \neq \ell'} \omega_{Rr}^{(\ell)} \cdot \mathbf{v}_r^{(\ell')} \end{bmatrix} =$$

$$= \begin{bmatrix} \mathbf{V}^{(\ell')} \cdot \Pi_1^{(\ell')} \\ \vdots \\ \mathbf{V}^{(\ell')} \cdot \Pi_R^{(\ell')} \end{bmatrix} = vec\left(\mathbf{V}^{(\ell')} \cdot \Pi_1^{(\ell')}, \dots, \mathbf{V}^{(\ell')} \cdot \Pi_R^{(\ell')}\right) = vec\left(\mathbf{V}^{(\ell')} \cdot \Pi^{(\ell')}\right). \qquad \square$$

For $\ell' \neq \ell''$, the computation of $vec\left(\mathbf{W}^{(\ell')} \cdot \left(\Pi^{(\ell',\ell')} * \left(\mathbf{V}^{(\ell'')^T} \cdot \mathbf{W}^{(\ell'')}\right)\right)\right)$ requires two matrix-matrix multiplications and one Hadamard product. This approach benefits from the BLAS-3 efficiency while the Hadamard product can efficiently be done in parallel. Now let's we make some complexity analysis. First, since $\mathbf{V}^{(\ell'')^T}$ is of shape $R \times I_{\ell''}$ and $\mathbf{W}^{(\ell'')}$ is $I_{\ell''} \times R$, multiplying them requires $I_{\ell''}R^2$ flops. The resulting matrix is of shape $R \times R$, so the Hadamard product by $\Pi^{(\ell',\ell')}$ requires more R^2 flops. Finally, since $\mathbf{W}^{(\ell')}$ is of shape $I_{\ell'} \times R$ and $\left(\Pi^{(\ell',\ell')} * \left(\mathbf{V}^{(\ell'')^T} \cdot \mathbf{W}^{(\ell'')}\right)\right)$ is of shape $R \times R$, multiplying them requires $I_{\ell'}R^2$ flops. Overall the cost is of $(1 + I_{\ell'} + I_{\ell''})R^2$ flops.

The cost of computing $\mathbf{V}^{(\ell')} * \Pi^{(\ell')}$ is more straightforward, being of $I_{\ell'}R^2$ flops. For each $\ell' = 1 \dots L$, the compute the sum $\sum_{\ell=1}^{L} \frac{\partial f}{\partial \mathbf{W}^{(\ell')}}^T \frac{\partial f}{\partial \mathbf{W}^{(\ell)}} \cdot vec(\mathbf{V}^{(\ell)})$ requires to perform $(1 + I_{\ell'} + I_\ell)R^2$ flops for all $\ell \neq \ell'$ and them more $I_{\ell'}R^2$ flops for ℓ'. The total cost of doing this is of $I_{\ell'}R^2 + \sum_{\ell \neq \ell'}(1 + I_{\ell'} + I_\ell)R^2 = \left((L-1) + (L-1)I_{\ell'} + \left(\sum_{\ell=1}^{L} I_\ell\right)\right) R^2$ flops. Since we have to do it for each ℓ', the total cost is of $\sum_{\ell'=1}^{L} \left((L-1) + (L-1)I_{\ell'} + \left(\sum_{\ell=1}^{L} I_\ell\right)\right) R^2 = \left(L(L-1) + (2L-1)\left(\sum_{\ell=1}^{L} I_\ell\right)\right) R^2$ flops. We can summarize al this analysis by saying that the cost of the matrix-vector multiplication $\mathbf{J}_f^T\mathbf{J}_f \cdot \mathbf{v}$ is of $\mathcal{O}\left(LR^2 \sum_{\ell=1}^{L} I_\ell\right)$ flops. Recall that the naive multiplication costs $R^2\left(\sum_{\ell=1}^{L} I_\ell\right)^2$ flops. Hence the gain in performance comes from the exploitation of the BLAS-3 implementation and the low complexity cost which was possible because of the block structure of $\mathbf{J}_f^T\mathbf{J}_f$.

Table 3.1 gather the information of all costs we obtained until now. From this point we will always summarize all costs analysis in tables at the final of each section. This will

Task	**Memory**	**Computational time**
Computing all $\Pi^{(\ell',\ell'')}$ and $\Pi^{(\ell')}$	$R^2(L+L^2)$	$\mathcal{O}\left(R^2\left(L+\sum_{\ell=1}^{L} I_\ell\right)\right)$
Computing $\mathbf{J}_f^T\mathbf{J}_f\cdot\mathbf{v}$	$R^2\sum_{\ell=1}^{L} I_\ell$	$\mathcal{O}\left(LR^2\sum_{\ell=1}^{L} I_\ell\right)$

Table 3.1: Memory and computational costs - I.

help the reader to make an overview of the costs when necessary.

Chapter 4

Computational experiments

This chapter starts presenting Tensor Fox, with a detailed discussion of main subroutines and their costs. Comprehensive experiments follows, first with the introduction of the tensors used for the benchmarks. We tried to use a wide distinct choice of tensors: positive tensors, tensors from machine learning, ill-conditioned tensors, and so on. The parameters of Tensor Fox are fine tuned against this set of tensors, so they are general enough. Then we conduct the benchmarks and discuss the results. The difference between Tensor Fox and other packages are discussed too.

The last five sections discuss are a discussion of the main aspects of Tensor Fox. In section 4.7 we connect the gain ratio and the fact that Tensor Fox is not monotonic, and how this is a good thing. In sections 4.8, 4.9 and 4.10 we discuss the diagonal regularization, conditioning and parallelism, respectively, with lots of more experiments validating our claims. Finally, in section 4.11 we will see more details about the features of Tensor Fox. Which of them are more relevant, and which are less relevant.

4.1 Tensor Fox

One of the main contributions of this work is the algorithm described in this section. An implementation of this algorithm is available (open and free) for Python, by the name of *Tensor Fox*, check the link `https://github.com/felipebottega/Tensor-Fox`. In this section we present and explain it in details, together with computational and memory costs. The algorithm will be presented in parts, following the computational flow as showed below in figure 4.1. Each box represents a major part of the algorithm, and each one is an algorithm by itself.

The *Compress* box is the computation of the MLSVD of the tensor. Most of the time it is unnecessary to work in the original space since the compressed tensor is equivalent to the original one in the sense of definition 2.1.2. In theory, if we find and exactly CPD and uncompress it, then this uncompressed CPD is an exact CPD for the tensor at the original space. After compressing, we must generate an initial tensor to start the dGN

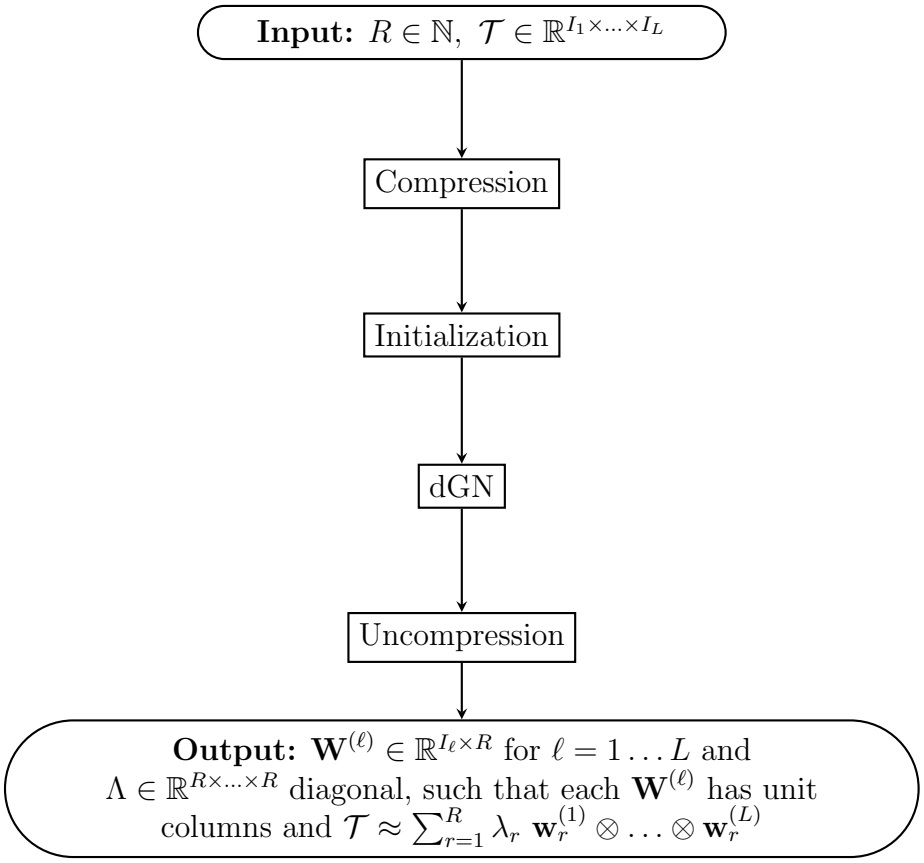

Figure 4.1: Flow chart of the main parts of Tensor Fox.

iterations. This is done at the *Initialization* part. More ahead we will show an original way to generating good initializations based on the MLSVD. The *dGN* algorithm is the most costly part and is there which lies some new ideas obtained after a lot of experimentation. These ideas are the result of much "mathematical alchemy". Finally, after computing a CPD we just need to *Uncompress* the solution.

4.1.1 Compression

Working with "raw" data is, usually, not advised because of the typical large data size. A standard approach is to compress the data before starting the actual work, and this is not different in the context of tensors. This is an efficient way of reducing the effects of the curse of dimensionality. The main tool to compress a tensor is the MLSVD, and it is interesting to use it because remark 2.2.9 guarantees one can use the core tensor of the MLSVD to compute a CPD for the original tensor. The idea is very similar to the procedure of compressing matrix data with the SVD: we compute the SVD of the matrix and truncate it to a matrix with lower rank. The choice of this rank is such that the truncated version is small enough to be tractable but close enough to the original matrix. Now we use this same idea on the tensor context.

First, compute the MLSVD of $\mathcal{T}$, obtaining a decomposition $\mathcal{T} = (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(L)}) \cdot \mathcal{S}$ where $\mathbf{U}^{(\ell)} \in \mathbb{R}^{I_\ell \times I_\ell}$ are orthogonal for $\ell = 1 \ldots L$, and $\mathcal{S} \in \mathbb{R}^{I_1 \times \ldots \times I_L}$ is the core tensor (see theorem 2.2.5). If $rank_{\boxplus}(\mathcal{T}) = (R_1, \ldots, R_L)$, then we can discard the last $I_\ell - R_\ell$ columns of each $\mathbf{U}^{(\ell)}$ and consider $\mathbf{U}^{(\ell)} \in \mathbb{R}^{I_\ell \times R_\ell}$, also we discard all hyperslices $\mathcal{S}_{i_\ell = k}$ for $k = R_\ell + 1 \ldots I_\ell$ and consider $\mathcal{S} \in \mathbb{R}^{R_1 \times \ldots \times R_L}$. After these truncations, the equality $\mathcal{T} = (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(L)}) \cdot \mathcal{S}$ remains intact. Actually, these truncations are not "real truncations" because we only transformed the decomposition from the full format to the reduced format, in the same way there is the full SVD and the reduced SVD for matrices. To get a real truncation with low multilinear rank we need to keep deleting columns of $\mathbf{U}^{(\ell)}$ and hyperslices of $\mathcal{S}$.

Let $(\tilde{R}_1, \ldots, \tilde{R}_L) \leq (R_1, \ldots, R_L)$ be a lower multilinear rank. We define $\tilde{\mathbf{U}}^{(\ell)} = \left[\mathbf{U}^{(\ell)}_{:1}, \ldots, \mathbf{U}^{(\ell)}_{:\tilde{R}_\ell}\right] \in \mathbb{R}^{I_\ell \times \tilde{R}_\ell}$ to be the matrix composed by the first columns of $\mathbf{U}^{(\ell)}$, and $\tilde{\mathcal{S}} \in \mathbb{R}^{\tilde{R}_1 \times \ldots \times \tilde{R}_L}$ is such that $\tilde{s}_{i_1 \ldots i_L} = s_{i_1 \ldots i_L}$ for $1 \leq i_1 \leq \tilde{R}_1, \ldots, 1 \leq i_L \leq \tilde{R}_L$. Figure 4.2 illustrates such a truncation in the case of a third order tensor. The white part correspond to $\mathcal{S}$ after we computed the full MLSVD (the one of theorem 2.2.5), the gray tensor is the reduced format of $\mathcal{S}$, and the red tensor is the truncated tensor $\tilde{\mathcal{S}}$. Our goal is to find the smallest $(\tilde{R}_1, \ldots, \tilde{R}_L)$ such that $\|\mathcal{S} - \tilde{\mathcal{S}}\|$ is not so large[1]. Since reducing $(\tilde{R}_1, \ldots, \tilde{R}_L)$ too much causes $\|\mathcal{S} - \tilde{\mathcal{S}}\|$ to increase, there is a trade off we have to manage in the best way possible.

Ideally, we want that each estimate $\tilde{R}_\ell$ to be exactly equal to R_ℓ, but that is not always possible. Furthermore, if $\mathcal{T}$ comes from data with noise, we are willing to truncate in order to work only with the relevant information. This relevant information is concentrated after the MLSVD is performed, and we can measure it in terms of the energy distribution, a topic already discussed in chapter 2. Because of theorem 2.2.11-1, we have that R_ℓ is the rank of the unfolding $\mathcal{T}_{(\ell)}$. Since the computation of the MLSVD requires the computation of the SVD of each unfolding, we can truncate these SVDs right after their computation. In fact, we can already start computing a truncated SVD with $P_\ell = \min(I_\ell, R)$ singular values since it is guaranteed that $R_\ell \leq \min(I_\ell, R)$. Let `mlsvd_tol` be a small positive tolerance value. For each truncated SVD of $\mathcal{T}_{(\ell)}$ with $\tilde{R}_\ell$ singular values, we want to verify its relative error with respect to $\mathcal{T}_{(\ell)}$. We start this process with $\tilde{R}_\ell = 1$ and keep increasing $\tilde{R}_\ell$ until the relative error is smaller than $\frac{1}{L} \cdot$ `mlsvd_tol`. Because of theorem 2.2.12 we don't need to actually construct these truncations and compute their respective relative errors. Denote by $\tilde{\mathcal{T}}_{(\ell)}$ the truncation with $\tilde{R}_\ell$ singular values, then we have that

$$\frac{\|\mathcal{T}_{(\ell)} - \tilde{\mathcal{T}}_{(\ell)}\|^2}{\|\mathcal{T}_{(\ell)}\|^2} = \frac{\sum_{r=\tilde{R}_\ell+1}^{P_\ell} \left(\sigma_r^{(\ell)}\right)^2}{\sum_{r=1}^{P_\ell} \left(\sigma_r^{(\ell)}\right)^2}.$$

[1] Actually, $\mathcal{S}$ and $\tilde{\mathcal{S}}$ belongs to different spaces. We committed an abuse of notation and wrote $\|\mathcal{S} - \tilde{\mathcal{S}}\|$ considering the projection of $\tilde{\mathcal{S}}$ over the space of $\mathcal{S}$, that is, enlarge $\tilde{\mathcal{S}}$ so it has the same size of $\mathcal{S}$ and consider these new entries as zeros. This is how we are projecting.

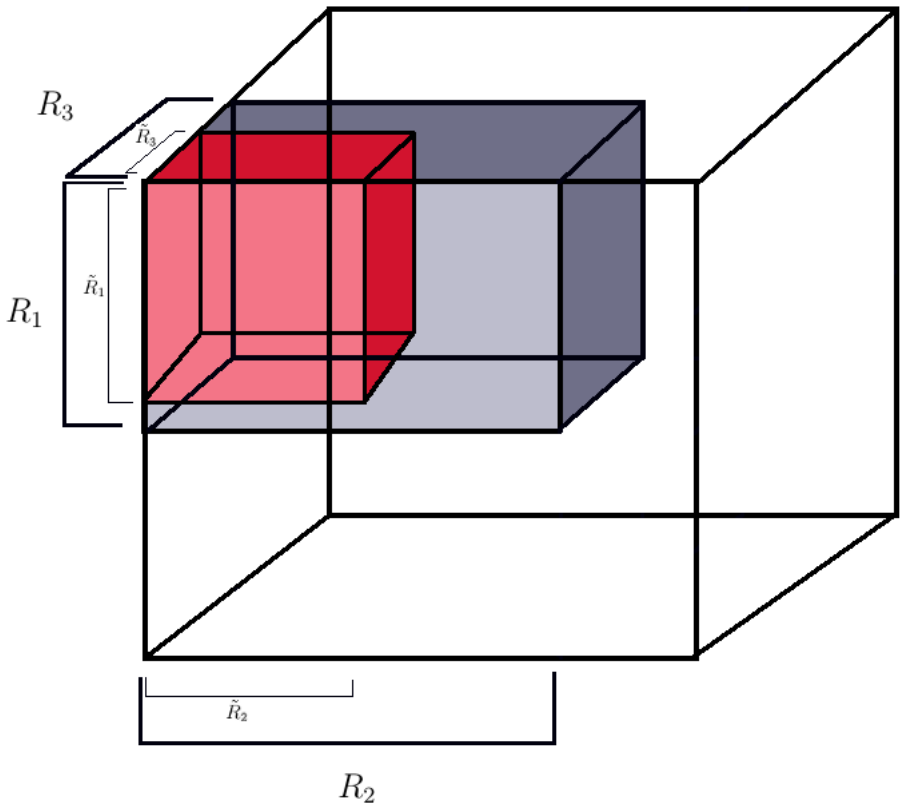

Figure 4.2: Truncated tensor $\tilde{\mathcal{S}}$.

After obtaining each $\tilde{R}_\ell$ we are able construct the truncated core tensor $\tilde{\mathcal{S}}$, which has shape $\tilde{R}_1 \times \ldots \times \tilde{R}_L$. It is not hard to see that the error between this truncation and the original core tensor is bounded as

$$\frac{\|\mathcal{S} - \tilde{\mathcal{S}}\|^2}{\|\mathcal{S}\|^2} \leq \frac{\sum_{i_1=\tilde{R}_1+1}^{P_1} \left(\sigma_{i_1}^{(1)}\right)^2 + \ldots + \sum_{i_L=\tilde{R}_L+1}^{P_L} \left(\sigma_{i_1}^{(1)}\right)^2}{\|\mathcal{S}\|^2} < L \cdot \frac{1}{L}\texttt{mlsvd\_tol} = \texttt{mlsvd\_tol},$$

where we use the fact that $\|\mathcal{S}\| = \|\mathcal{T}\| = \|\mathcal{T}_{(\ell)}\|$. Note that these computations only requires $\sum_{\ell=1}^{L}(R_\ell - \tilde{R}_\ell)$ flops. The following algorithm summarizes all we have described so far. We are using the notation $\Sigma^{(\ell)} = \{\sigma_1^{(\ell)}, \ldots, \sigma_{P_\ell}^{(\ell)}\}$ for the set of the singular values of $\mathcal{T}_{(\ell)}$.

Task	Memory	Computational cost
Computing the MLSVD	$R\sum_{\ell=1}^{L} P_\ell + \prod_{\ell=1}^{L} P_\ell$	$\mathcal{O}\left(\sum_{\ell=1}^{L} \log(P_\ell) \prod_{\ell'=1}^{L} I_{\ell'}\right)$
Truncating the MLSVD	$R\sum_{\ell=1}^{L} \tilde{R}_\ell + \prod_{\ell=1}^{L} \tilde{R}_\ell$	$\mathcal{O}\left(\sum_{\ell=1}^{L} P_\ell^2\right)$

Table 4.1: Memory and computational costs - II.

Algorithm 4.1.1 (Compression)**.**
Input: $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_L}$, `mlsvd_tol` > 0

$\left\{ \boldsymbol{U}^{(1)}, \ldots, \boldsymbol{U}^{(L)}, \Sigma^{(1)}, \ldots, \Sigma^{(L)}, \mathcal{S} \right\} \leftarrow$ `MLSVD`$(\mathcal{T})$
`for` $\ell = 1 \ldots L$
  `for` $i = 1 \ldots P_\ell$
    `rel_error` $= \dfrac{\sum_{r=\tilde{R}_\ell+1}^{P_\ell} \left(\sigma_r^{(\ell)}\right)^2}{\sum_{r=1}^{P_\ell} \left(\sigma_r^{(\ell)}\right)^2}$
    `if rel_error` $< \frac{1}{L}$ `mlsvd_tol`
      $\tilde{R}_\ell = i$
      `break`
`for` $\ell = 1 \ldots L$
  $\tilde{\boldsymbol{U}}^{(\ell)} \leftarrow$ `truncation of` $\boldsymbol{U}^{(\ell)}$ `to have` $\tilde{R}_\ell$ `columns`
$\tilde{\mathcal{S}} \leftarrow$ `truncation of` $\mathcal{S}$ `to have shape` $\tilde{R}_1 \times \ldots \times \tilde{R}_L$

Output: $\tilde{\boldsymbol{U}^{(\ell)}} \in \mathbb{R}^{I_\ell \times \tilde{R}_\ell}$ *for* $\ell = 1 \ldots L$ *and* $\tilde{\mathcal{S}} \in \mathbb{R}^{\tilde{R}_1 \times \ldots \times \tilde{R}_L}$

The function `MLSVD` is described in 2.2.6, which has a cost of $\mathcal{O}\left(\log(P_\ell) \prod_{\ell'=1}^{L} I_{\ell'}\right)$ flops if we compute the truncated SVD with P_ℓ singular values for each unfolding. This is the dominant cost in this algorithm. With respect to the memory storage, each $\Sigma^{(\ell)}$ is stored as a vector of size P_ℓ, it is necessary to store $I_\ell P_\ell$ floats for each $\mathbf{U}^{(\ell)}$, and not more than $\prod_{\ell=1}^{L} P_\ell$ floats for $\mathcal{S}$. After computing the MLSVD, the worse case of the truncation stage has a cost of $\mathcal{O}\left(\sum_{\ell=1}^{L} P_\ell^2\right)$ flops, which is cheap. Table 4.1 summarizes all main costs necessary to compress a tensor. Note that we are considering only the classic truncated MLSVD costs here. It is possible to obtain even lower costs with the sequentially truncated algorithm.

4.1.2 Initialization

As we mentioned before, the dGN method is not so sensitive to the initialization, but it still can benefits of good initializations. Furthermore, some initializations actually can lead to local minimum, and when this happens it is necessary to try again. In order to prevent local minimum it is interesting to have a method which generates initialization sufficiently close to the global minimum. Note that our objective is to compute a rank-R CPD for $\mathcal{S} \in \mathbb{R}^{R_1 \times \ldots \times R_L}$, which is already truncated at this moment.

The first approach we consider is very usual in other implementations. The program may just generate random initial random matrices $\mathbf{W}^{(\ell)} \in \mathbb{R}^{R_\ell \times R}$ with the entries draw from the standard normal distribution (that is, mean 0 and variance 1). With this we have an initial approximated CPD $\mathcal{S} = (\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}) \cdot \mathcal{I}_{R \times \ldots \times R}$ which we use to start iterating. For general purposes this approach is good enough, but sometimes one can gain performance when the initialization is close enough to the objective tensor. With this in mind we propose another approach.

Our second approach is a simple method based on the truncated MLSVD $\mathcal{T} \approx (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(L)}) \cdot \mathcal{S}$. Remember the discussion about energy we had in chapter 2. The energy gets higher as the coordinates of $\mathcal{S}$ are close to $11\ldots1$, the very first entry of $\mathcal{S}$. In this case the energy of a coordinate can be seen as its magnitude, regardless it is positive or negative. Since we want a rank-R CPD, one idea is to take R entries of $\mathcal{S}$ with high energy and construct an approximated CPD from them. Since the energy is highly concentrated around $s_{11\ldots1}$, this approximated CPD may already be very close to the objective tensor. Now, choose R multi-indexes $(j_1^{(1)}, \ldots, j_1^{(L)}), \ldots, (j_R^{(1)}, \ldots, j_R^{(L)}) \in R_1 \times \ldots \times R_L$ close to $(11\ldots1)$. These choices gives a rank-R approximation

$$\mathcal{S} \approx \sum_{r=1}^{R} s_{j_r^{(1)},\ldots,j_r^{(L)}} \ \mathbf{e}_{j_r^{(1)}}^{(1)} \otimes \ldots \otimes \mathbf{e}_{j_r^{(L)}}^{(L)},$$

where each $\mathbf{e}_j^{(\ell)}$ is the j-th basis canonical vector of $\mathbb{R}^{R_\ell}$. We denote this approximation by $\tilde{\mathcal{S}}$. We also denote $J = \{(j_1^{(1)}, \ldots, j_1^{(L)}), \ldots, (j_R^{(1)}, \ldots, j_R^{(L)})\}$ the set of multi-indexes we use to construct $\tilde{\mathcal{S}}$. Then the corresponding error is given by

$$\|\mathcal{S} - \tilde{\mathcal{S}}\|^2 = \sum_{(i_1,\ldots,i_L) \in R_1 \times \ldots \times R_L} s_{i_1\ldots i_L}^2 - \sum_{(i_1,\ldots,i_L) \in J} s_{i_1\ldots i_L}^2 = \sum_{(i_1,\ldots,i_L) \in R_1 \times \ldots \times R_L \setminus J} s_{i_1\ldots i_L}^2.$$

If our choice is adequate we can minimize this error. Furthermore, depending on the distribution energy of $\mathcal{S}$ this error may actually be very small. In my experience with tensors, that kind of initialization works well for several problems, and sometimes it works much better than the random initialization. The computational cost of the initialization

is of $\mathcal{O}\left(R\sum_{\ell=1}^{L} R_\ell\right)$ flops, the size of the starting point.

4.1.3 dGN

Let $\mathcal{T} \approx (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(L)}) \cdot \mathcal{S}$ be a truncated MLSVD for $\mathcal{T}$ which we assume to be close enough. The tensor obtained by the initialization will be denoted by $\mathcal{S}^{(0)}$. We denote by $\mathcal{S}^{(k)}$ the tensor obtained at the iteration k of the dGN algorithm. Our goal is to iteratively produce successive approximations $\mathcal{S}^{(0)}, \mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(k)}, \mathcal{S}^{(k+1)}, \ldots$ converging to $\mathcal{S}$. Associated with each $\mathcal{S}^{(k)}$ there will be the factor matrices $\mathbf{W}^{(1,k)}, \ldots, \mathbf{W}^{(L,k)}$ where $\mathbf{W}^{(\ell,k)} \in \mathbb{R}^{R_\ell \times R}$ for each $\ell = 1 \ldots R$. As have been done before, we denote $\mathbf{w}^{(k)} = \left[vec(\mathbf{W}^{(1,k)})^T, \ldots, vec(\mathbf{W}^{(L,k)})^T\right]^T$.

Let K be the last iteration of the dGN. Then we expect to have

$$\mathcal{S} \approx \underbrace{(\mathbf{W}^{(1,K)}, \ldots, \mathbf{W}^{(L,K)}) \cdot \mathcal{I}_{R\times R}}_{\mathcal{S}^{(K)}},$$

which leads to

$$\mathcal{T} \approx \left((\mathbf{U}^{(1)})^*\mathbf{W}^{(1,K)}, (\mathbf{U}^{(L)})^*\mathbf{W}^{(L,K)}\right) \cdot \mathcal{I}_{R\times\ldots\times R} =$$

$$= \left(\tilde{\mathbf{W}}^{(1)}, \ldots, \tilde{\mathbf{W}}^{(L)}\right) \cdot \mathcal{I}_{R\times\ldots\times R}.$$

$$= \sum_{r=1}^{R} \tilde{\mathbf{W}}_{:r}^{(1)} \otimes \ldots \otimes \tilde{\mathbf{W}}_{:r}^{(L)}.$$

4.1.3.1 Main parameters

First we need to establish the values of three important parameters: the initial damping parameter $\mu^{(0)}$, the maximum number of iterations `maxiter` and the tolerance `tol`. As mentioned in the "damped Gauss-Newton" section, the initial damping parameter is of the form $\mu^{(0)} = \tau \cdot \max_i a_{ii}$ (see [22] to know more about this choice), where $\mathbf{A} = \mathbf{J}_f^T(\mathbf{w}^{(0)})\mathbf{J}_f(\mathbf{w}^{(0)})$, but instead of this we use the similar definition $\mu^{(0)} = \tau \cdot \mathbb{E}(|\mathcal{S}|)$, where $\mathbb{E}(|\mathcal{S}|)$ is the average of the entries of $\mathcal{S}$ in absolute value. This initial value was observed to perform well in practice. Also, since we want to reinforce regularization at the first iterations, we use $\tau = 1$. With regard to the maximum number of iterations, nonlinear least squares methods usually converges within a few hundreds iterations, whereas ALS and general unconstrained optimization methods need thousands iterations to converge. A reasonable choice in this case is `maxiter`$= 200$. Usually there are many different tolerance parameters, one for each stopping condition. Although Tensor Fox does have the possibility to choose different values for each tolerance parameter, but for simplicity

we consider that all tolerances are equal to `tol`. At the moment we are using `tol`$= 10^{-6}$. All these choices are the default values in TensorFox. They were obtained after several tests and experiments which will be showed soon.

4.1.3.2 Computing the residual

At each iteration, the first task is to compute the residual function $f = (f_{11\ldots1}, \ldots, f_{R_1R_2\ldots R_L})$, where

$$f_{i_1\ldots i_L}(\mathbf{w}^{(k)}) = s_{i_1\ldots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1,k)} \ldots w_{i_L r}^{(L,k)}.$$

The cost of this computation is of $L-1$ flops, but since this has to be done for all the residuals, the total cost is of $(L-1)\prod_{\ell=1}^{L} R_\ell$ flops.

4.1.3.3 Computing the gradient

Recall 3.4.2 from that $\nabla F(\mathbf{w}) = \mathbf{J}_f^T(\mathbf{w}) \cdot f(\mathbf{w})$, where F is the error function defined in 3.3. In section 3.5.2 we observed that $\mathbf{J}_f(\mathbf{w})$ is sparse with a certain structure we can exploit. Usually $\mathbf{J}_f$ has LR nonzero entries in each row, these are the values $-\prod_{\ell\neq\ell'} w_{i_\ell r'}^{(\ell)}$ for $\ell' = 1\ldots L$ and $r' = 1\ldots R$. Since $\mathbf{J}_f(\mathbf{w})$ has $LR\prod_{\ell=1}^{L} R_\ell$ nonzero entries, we can expect to compute $\mathbf{J}_f^T(\mathbf{w}) \cdot f(\mathbf{w})$ with $LR\prod_{\ell=1}^{L} R_\ell$ operations, which is still much better than $R\sum_{\ell=1}^{L} R_\ell \prod_{\ell=1}^{L} R_\ell$ if we considered $\mathbf{J}_f(\mathbf{w})$ as a dense matrix. Consider the products $*$ and $\odot$ defined in B and remember that

$$\Pi^{(\ell')} = \pi^{(1)} * \ldots * \pi^{(\ell'-1)} * \pi^{(\ell'+1)} * \ldots * \pi^{(L)},$$

where $\pi^{(\ell)} = \mathbf{W}^{(\ell)^T}\mathbf{W}^{(\ell)}$. Then we have the following result from [17].

Theorem 4.1.2 (T. G. Kolda, E. Acar, D. M. Dunlavy, 2011)**.** *The the partial derivatives of F with respect to $\boldsymbol{W}^{(\ell)}$ are given by*

$$\frac{\partial F}{\partial \boldsymbol{W}^{(\ell)}}(\boldsymbol{w}) = \boldsymbol{W}^{(\ell)}\Pi^{(\ell)} - \mathcal{T}_{(\ell)}\left(\boldsymbol{W}^{(L)} \odot \ldots \odot \boldsymbol{W}^{(\ell+1)} \odot \boldsymbol{W}^{(\ell-1)} \odot \ldots \odot \boldsymbol{W}^{(1)}\right).$$

Since $\nabla F(\mathbf{w}) = \left[vec\left(\frac{\partial F}{\partial \mathbf{W}^{(1)}}(\mathbf{w})\right)^T, \ldots, vec\left(\frac{\partial F}{\partial \mathbf{W}^{(L)}}(\mathbf{w})\right)^T\right]^T$, from lemma 3.4.2 we con-

clude that

$$\mathbf{J}_f^T(\mathbf{w})\mathbf{f}(\mathbf{w}) = -\left[vec\left(\frac{\partial F}{\partial \mathbf{W}^{(1)}}(\mathbf{w})\right)^T, \ldots, vec\left(\frac{\partial F}{\partial \mathbf{W}^{(L)}}(\mathbf{w})\right)^T\right]^T.$$

The dominant cost is the computation of $\mathcal{T}_{(\ell)}\left(\mathbf{W}^{(L)} \odot \ldots \odot \mathbf{W}^{(\ell+1)} \odot \mathbf{W}^{(\ell-1)} \odot \ldots \odot \mathbf{W}^{(1)}\right)$, which is $\mathcal{O}\left(R\prod_{\ell=1}^{L} I_\ell\right)$ flops. Since we have to perform this L times, the total cost is of $\mathcal{O}\left(LR\prod_{\ell=1}^{L} I_\ell\right)$ flops. Efficient ways to deal with these Khatri-Rao products can be found in [93].

4.1.3.4 Conjugate gradient

The next stage is the computation of the step to take. Remember the description given at the beginning of section 3.5. At the point $\mathbf{w}^{(k)}$ we want to take a step in direction $\mathbf{x}$ such that the new point $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{x}$ minimizes the residual at the neighborhood of $\mathbf{w}^{(k)}$. This leads to a normal equations which we regularize (see 3.6), obtaining the system

$$(\mathbf{A}^T\mathbf{A} + \mu^{(k)}\mathbf{D})\mathbf{x} = \mathbf{A}^T\mathbf{b},$$

where $\mathbf{D}$ is a diagonal $\left(R\sum_{\ell=1}^{L} R_\ell\right) \times \left(R\sum_{\ell=1}^{L} R_\ell\right)$ matrix, $\mathbf{A} = \mathbf{J}_f(\mathbf{w}^{(k)})$, $\mathbf{x} = \mathbf{w} - \mathbf{w}^{(k)}$, $\mathbf{b} = -f(\mathbf{w}^{(k)})$. The optimal solution $\mathbf{x}_*$ gives the optimal point $\mathbf{w}^{(k+1)}$ by setting $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{x}_*$.

The regularization matrix $\mathbf{D}$ depends on $\mathbf{w}^{(k)}$ and is chosen to make $\mathbf{A}^T\mathbf{A} + \mathbf{D}$ diagonally dominant (notice the absence of the damping parameter). As the iteration goes we will have $\mu^{(k)} \to 0$ and the effect of the regularization decreases. When close to the optimal point the only effect of $\mathbf{D}$ is to guarantee that the system is well-posed. Still, the system converges to an ill-posed system as $\mu^{(k)} \to 0$ so we will be working with ill-conditioned system when close to the objective point. One way to mitigate the effects of this ill-conditioning is preconditioning the system with a Jacobi preconditioner as explained in appendix A. After the preconditioning we have the new system

$$\mathbf{M}^{-1/2}(\mathbf{A}^T\mathbf{A} + \mu^{(k)}\mathbf{D})\mathbf{M}^{-1/2}\mathbf{x} = \mathbf{M}^{-1/2}\mathbf{A}^T\mathbf{b}, \tag{4.1}$$

where $\mathbf{M} = \text{diag}(a_{ii} + \mu^{(k)}d_{ii})$. Notice this equation is the same as the one preconditioned in A.5, but here we have $\mathbf{A}^T\mathbf{A} + \mu^{(k)}\mathbf{D}$ instead of just $\mathbf{A}$, and $\mathbf{A}^T\mathbf{b}$ instead of just $\mathbf{b}$. To solve this system we rely on the conjugate gradient, see appendix A.2.1. The algorithm should be adapted to our current situation.

Algorithm 4.1.3 (Conjugate gradient - dGN)**.**
Input: $\boldsymbol{A}, \boldsymbol{M}, \boldsymbol{D}, \boldsymbol{b}, \mu^{(k)}$

$\boldsymbol{B} \leftarrow \boldsymbol{M}^{-1/2}(\boldsymbol{A}^T\boldsymbol{A} + \mu^{(k)}\boldsymbol{D})\boldsymbol{M}^{-1/2}$
$\boldsymbol{x}^{(0)} \leftarrow 0 \in \mathbb{R}^n$
$\boldsymbol{r}^{(0)} \leftarrow \boldsymbol{M}^{-1/2}\boldsymbol{A}^T\boldsymbol{b}$
$\boldsymbol{p}^{(0)} \leftarrow \boldsymbol{r}^{(0)}$
`for` $i = 1 \ldots$ `cg_maxiter`
$\quad \boldsymbol{z} \leftarrow \boldsymbol{B} \cdot p^{(i-1)}$
$\quad \alpha^{(i)} \leftarrow \frac{\|\boldsymbol{r}^{(i-1)}\|^2}{\|\boldsymbol{p}^{(i-1)}\|^2_B}$
$\quad \boldsymbol{x}^{(i)} \leftarrow \boldsymbol{x}^{(i-1)} + \alpha^{(i)}\boldsymbol{p}^{(i-1)}$
$\quad \boldsymbol{r}^{(i)} \leftarrow \boldsymbol{r}^{(i-1)} - \alpha^{(i)}\boldsymbol{z}$
$\quad \varepsilon \leftarrow \|\boldsymbol{r}^{(i)}\|^2$
$\quad \beta^{(i)} \leftarrow \frac{\|\boldsymbol{r}^{(i)}\|^2}{\|\boldsymbol{r}^{(i-1)}\|^2}$
$\quad \boldsymbol{p}^{(i)} \leftarrow \boldsymbol{r}^{(i)} + \beta^{(i)}\boldsymbol{p}^{(i-1)}$
$\quad$ `if` $\varepsilon <$ `tol`
$\quad\quad$ `break`

Output: $\boldsymbol{x}^{(i')}$*, where* i' *is the last index of the iterations*

Now we comment a few things about the algorithm, from top to bottom. The first thing we should remark is that the matrix $\mathbf{B}$ is never computed explicitly. As already mentioned, this is a matrix-free algorithm because of theorem 3.5.10. The vector $\mathbf{r}^{(0)}$ demands some computational effort as observed before. Multiplying $\mathbf{A}^T\mathbf{b}$ by $\mathbf{M}^{-1/2}$ amounts to just $R\sum_{\ell=1}^{L} R_\ell$ operations since $\mathbf{M}^{-1/2}$ is diagonal. The result is a vector of size $R\sum_{\ell=1}^{L} R_\ell$.

We call the conjugate gradient algorithm by CG for short. The parameter `cg_maxiter` is the maximum number of iterations permitted. When in situations like this, where the CG is used as a step for other iterative algorithm, it is usual to set `cg_maxiter` to a low value. Sometimes `cg_maxiter` $= 10$ is already enough to produce a reasonable step for the "bigger" algorithm (in this case, the dGN). It should be pointed that this is very case dependent and just there isn't a universal rule for how to set this parameter. We tested it for several fixed values and none was satisfactory. Low values made the dGN converge to local minimum most of the time, higher values improved the convergence but were very costly. It was noted a better performance when we increased `cg_maxiter` a little for each iteration of the dGN. However this approach forced the program to make too much CG iterations at the end of the dGN, which was wasteful. Several attempts to solve this issue were tried. The problem in all of them was the fact that they all were deterministic. As soon as the value `cg_maxiter` was set to be a random integer everything worked much better. At iteration k of the dGN, we define `cg_maxiter` as being a random integer draw from the uniform distribution in the interval $[1 + \lceil k^{0.4}\rceil, 2 + \lceil k^{0.9}\rceil]$. This strategy

to obtain `cg_maxiter` (instead of just fixing it) is the result of many experiments.

Remark 4.1.4. *Allowing the number of iterations to change dynamically allows the algorithm to make unusual steps sometimes, which proved to be a successful way to avoid local minima. This is a very specific format of interval so you can imagine how much of "mathematical alchemy" was necessary to create such solution.*

The product $\mathbf{B} \cdot \mathbf{p}^{(i-1)}$ is to be computed in three steps as showed below.

$$\begin{aligned}
\mathbf{z} &\leftarrow \mathbf{M}^{-1/2} \cdot \mathbf{p}^{(i-1)} \\
\mathbf{z} &\leftarrow \mathbf{A}^T\mathbf{A} \cdot \mathbf{z} + \mu^{(k)}\mathbf{z} \\
\mathbf{z} &\leftarrow \mathbf{M}^{-1/2} \cdot \mathbf{z}
\end{aligned}$$

However before starting these computations we need to construct all the matrices $\Pi^{(\ell',\ell'')}$ and $\Pi^{(\ell')}$. We remark that the computation of the gradient $\mathbf{A}^T \cdot \mathbf{b}$ and these matrices are performed before the CG loop, so their costs are accounted only once. We already observed that, since $\mathbf{M}^{-1/2}$ is diagonal, the first product can be computed with $R\sum_{\ell=1}^{L} R_\ell$ operations. The same goes for the multiplication by the scalar $\mu^{(k)}$. Theorem 3.5.10 and the discussion after it showed that $\mathbf{A}^T\mathbf{A} \cdot \mathbf{z}$ can be computed with $\mathcal{O}\left(LR^2\sum_{\ell=1}^{L} R_\ell\right)$ operations. Finally we have to make more $R\sum_{\ell=1}^{L} R_\ell$ operations in the last step $\mathbf{z} \leftarrow \mathbf{M}^{-1/2} \cdot \mathbf{z}$. The overall cost to obtain $\mathbf{B} \cdot \mathbf{p}^{(i-1)}$ is of

$$\mathcal{O}\left((3R + LR^2)\sum_{\ell=1}^{L} R_\ell\right)$$

flops. All the remaining lines of the algorithm together have a cost of $\mathcal{O}\left(6R\sum_{\ell=1}^{L} R_\ell\right)$ flops. Putting everything together we can see that the cost of the CG algorithm is of

$$\mathcal{O}\left(\underbrace{LR\prod_{\ell=1}^{L} R_\ell}_{\text{gradient}} + \underbrace{R^2\left(L + \sum_{\ell=1}^{L} R_\ell\right)}_{\Pi^{(\ell',\ell'')} \text{ and } \Pi^{(\ell')}} + \underbrace{i_{CG}(9R + LR^2)\sum_{\ell=1}^{L} R_\ell}_{\text{CG loop}}\right),$$

where i_{CG} is the number of iterations of the CG. We know that i_{CG} is a random integer draw from the interval $[1 + \lceil k^{0.4} \rceil, 2 + \lceil k^{0.9} \rceil]$, where k is the current iteration of the dGN. Let's consider the worst case $k = \texttt{maxiter} = 200$ just to have an idea. In this, case the expected i_{CG} is

$$\mathbb{E}(i_{CG}) = \left\lfloor \frac{2 + \lceil 200^{0.9} \rceil - 1 - \lceil 200^{0.4} \rceil}{2} \right\rfloor = 55,$$

which is a reasonable value.

4.1.3.5 Updates

After the CG is performed we have a point $\mathbf{x}_*$, which is an approximated solution of 4.1. In turn, this equation implies that $\mathbf{w}^{(k)} + \mathbf{x}_*$ minimizes the residual at the neighborhood of $\mathbf{w}^{(k)}$ (see the discussion about 3.5). For this reason, we set $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{x}_*$. This is the first update of the dGN algorithm.

Depending on the size of the dimensions $R_1, \ldots, R_L$, this second step is likely to be the most costly part of all routines inside dGN. We are talking about the evaluation of the error function F. Let $\mathcal{S}^{(k)} = \sum_{r=1}^{R} \mathbf{w}_r^{(1,k)} \otimes \ldots \otimes \mathbf{w}_r^{(L,k)}$ be the approximating tensor computed at the k-th iteration. To compute the error at iteration k we must compute

$$F(\mathbf{w}^{(k)}) = \frac{1}{2} \left\| \mathcal{S} - \mathcal{S}^k \right\|^2 = \frac{1}{2} \left\| \mathcal{S} - \sum_{r=1}^{R} \mathbf{w}_r^{(1,k)} \otimes \ldots \otimes \mathbf{w}_r^{(L,k)} \right\|^2 =$$

$$= \frac{1}{2} \sum_{i_1=1}^{R_1} \cdots \sum_{i_L=1}^{R_L} \left(s_{i_1 \ldots i_L} - \sum_{r=1}^{R} w_{i_1 r}^{(1,k)} \cdot \ldots \cdot w_{i_L r}^{(L,k)} \right)^2 .$$

we don't take in account the cost to compute this error since it is just a matter of summing the squares of the residual whose cost is already considered. Before making this evaluation is it interesting to "normalize" the factors, that is, scale them so we have $\|\mathbf{w}_r^{(1,k)}\| = \|\mathbf{w}_r^{(2,k)}\| = \ldots = \|\mathbf{w}_r^{(L,k)}\|$ for each $r = 1 \ldots R$. This is always possible, has a low cost of $R \sum_{\ell=1}^{L} R_\ell$ flops and can improve the conditioning of the problem, making it easier for the next CG iterations to find a good solution. When the factors are scaled this way they are called *norm-balanced*. In section 5 of [34] the conditioning of norm-balanced representations of tensors is discussed at some detail. In particular, the lower bound of the condition number is minimized when we use balanced-norm representations.

The last update is the damping parameter update. Let $\mu^{(k)}$ be the damping parameter at iteration k of dGN. Previously, in the "Damped Gauss-Newton" section, we saw that this update depends on the gain ratio

$$g = \frac{\|\mathcal{S} - \tilde{\mathcal{S}}^{(k-1)}\|^2 - \|\mathcal{S} - \tilde{\mathcal{S}}^{(k)}\|^2}{\|\mathcal{S} - \tilde{\mathcal{S}}^{(k-1)}\|^2 - \|\tilde{f}(\mathbf{w}^{(k)})\|^2}.$$

A large value of g indicates that $\|\tilde{f}(\mathbf{w}^{(k)})\|^2$ is a good approximation to $\|\mathcal{S} - \tilde{\mathcal{S}}^{(k)}\|^2$, in other words, the linear model $\tilde{f}$ is making good predictions about the actual errors. A small or negative g indicates that this approximation is poor (the reason for that is indicated in section 3.5.2). When the approximation is good we decrease the damping parameter so the iterations becomes more like a Gauss-Newton iteration, which converges

rapidly. Otherwise we increase the damping parameter, which adds regularization to the model and make it more alike the gradient descent method.

An update strategy widely used is the following.

$$
\begin{aligned}
&\texttt{if } g < 0.25 \\
&\quad \mu \leftarrow 2\mu \\
&\texttt{else if } g > 0.75 \\
&\quad \mu \leftarrow \mu/3
\end{aligned}
$$

This strategy was originally proposed by Marquardt in [69]. This update strategy is not so sensitive to minor changes in the thresholds $0.25, 0.75$ or the update factors $2, 1/3$. Another strategy frequently used (which was also used to compute CPDs in [4]) is the following.

$$
\begin{aligned}
&\texttt{if } g > 0 \\
&\quad \mu \leftarrow \mu \cdot \max\{1/3, 1 - (2g-1)^3\} \\
&\quad \nu \leftarrow 2 \\
&\texttt{else} \\
&\quad \mu \leftarrow \mu \cdot \nu \\
&\quad \nu \leftarrow 2\nu
\end{aligned}
$$

In general this strategy outperforms the previous one, as demonstrated in [79]. The value ν usually is initiated to $\nu = 2$, but minor changes doesn't affect the performance significantly. We tested the second strategy in our problem for a vast of different updates and none was very effective. The first one seemed to perform better for some specific choices of thresholds and factors. In the end, we decided to use the following update strategy.

$$
\begin{aligned}
&\texttt{if } g < 0.25 \\
&\quad \mu \leftarrow 3\mu \\
&\texttt{else if } g > 0.75 \\
&\quad \mu \leftarrow \mu/3
\end{aligned}
$$

These values were obtained experimentally as a result of several tests over several distinct tensors. Still, there is a reasoning to explain why these values works well. As we already know, the maximum number of iterations of the CG will be very low at the first iterations of the dGN (`cg_maxiter` can be even 2 or 3). This means the solution found by the CG has a large residual, which means $\|\tilde{f}(\mathbf{w}^{(k)})\|$ will be large when k is small. Even

in this situation the dGN is likely to decrease the errors monotonically, so we will have $\|\mathcal{S} - \mathcal{S}^{(k)}\| \leq \|\mathcal{S} - \mathcal{S}^{(k-1)}\| \leq \|\tilde{f}(\mathbf{w}^{(k)})\|$ regardless the number of iterations performed by the CG. This implies that $g < 0$. If we apply the original Marquardt strategy, then the damping parameter will increase while also increasing the regularization of the problem, which is already very regularized at the first iterations. This excessive regularization almost always leads to solutions which are some local minimum of the problem. Even if we use some good initialization (which clearly means that decreasing the damping parameter is the right move to do), the fact that `cg_maxiter` is low at the beginning will lead to more regularization, which leads to local minima. As we keep doing dGN iterations, it is expected to have $\|\tilde{f}(\mathbf{w}^{(k)})\| \ll \|\mathcal{S} - \mathcal{S}^{(k)}\| \lessapprox \|\mathcal{S} - \mathcal{S}^{(k-1)}\|$ so g is positive and close to 0. To reinforce the fact that we want to decrease the damping parameter, it is crucial to choose thresholds which favors this. Therefore, most of the time the algorithm will decrease the damping parameter and only sometimes it will increase it.

4.1.3.6 Stopping conditions

As in any iterative algorithm, good stopping conditions are crucial to improve performance. The conditions we used to stop iterating are based in other implementations, but they were tested and refined to lead to the best performance possible. We present our four stopping conditions below and in the same order they are implemented. The program stops at iteration k if

1. **Relative error**: $\dfrac{\|\mathcal{S} - \mathcal{S}^{(k)}\|}{\|\mathcal{S}\|} < \texttt{tol}$

2. **Step size**: $\|\mathbf{w}^{(k-1)} - \mathbf{w}^{(k)}\| < \texttt{tol}$

3. **Relative error improvement**: $\left| \dfrac{\|\mathcal{S} - \mathcal{S}^{(k-1)}\|}{\|\mathcal{S}\|} - \dfrac{\|\mathcal{S} - \mathcal{S}^{(k)}\|}{\|\mathcal{S}\|} \right| < \texttt{tol}$

4. **Gradient norm**: $\|\nabla F(\mathbf{w}^{(k)})\|_\infty < \texttt{tol}$

5. **Average error:**

$$\frac{1}{c} \sum_{k=k_0-2c}^{k_0-c} \frac{\|\mathcal{S} - \mathcal{S}^{(k)}\|}{\|\mathcal{S}\|} - \frac{1}{c} \sum_{k=k_0-c}^{k_0} \frac{\|\mathcal{S} - \mathcal{S}^{(k)}\|}{\|\mathcal{S}\|} < \texttt{tol}$$

where $k_0 > 2c$ and $c = 1 + \left\lceil \dfrac{\texttt{maxiter}}{10} \right\rceil$

6. **Average improvement:**

$$\frac{1}{c}\sum_{k=k_0-c}^{k_0}\left|\frac{\|\mathcal{S}-\mathcal{S}^{(k-1)}\|}{\|\mathcal{S}\|}-\frac{\|\mathcal{S}-\mathcal{S}^{(k)}\|}{\|\mathcal{S}\|}\right|<10^{-3}\cdot\frac{1}{c}\sum_{k=k_0-c}^{k_0}\frac{\|\mathcal{S}-\mathcal{S}^{(k)}\|}{\|\mathcal{S}\|}$$

where $k_0 > 2c$ and $c = 1 + \left\lceil \frac{\texttt{maxiter}}{10} \right\rceil$

7. **Divergence:** $\dfrac{\|\mathcal{S}-\mathcal{S}^{(k)}\|}{\|\mathcal{S}\|} > \dfrac{\max\{1,\ \|\mathcal{T}\|^2\}}{10^{-16}+\texttt{tol}}$

The first three conditions are standard and will be present in almost any solver in one form or another. The gradient condition can be found in Tensor Toolbox. Since the algorithm converges to a critical point, this is a natural stopping condition to include. The purpose of the last condition is clear, and the term 10^{-16} in the denominator is there to prevent division by zero in the case of $\texttt{tol} = 0$. Conditions 5 and 6 deserves a little more explanation. First of all, we remark that the constants c and 10^{-3} were obtained empirically and there is nothing special about them. Secondly, these two conditions are only verified at each c iterations so we are always comparing the average of a batch with the next one. Condition 5 avoids errors oscillating too much time without any overall decreasing, like illustrated in figure 4.3.

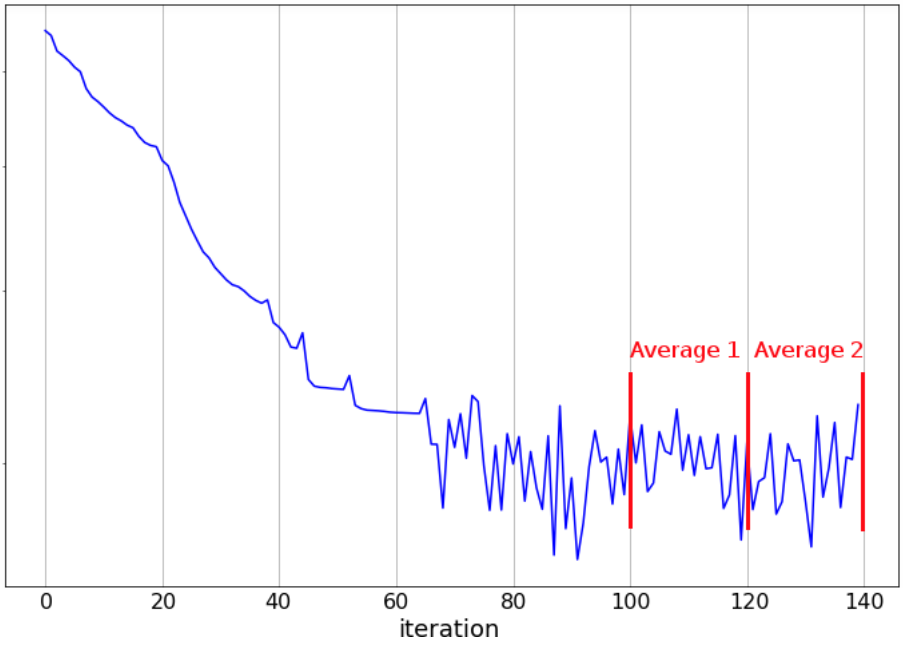

Figure 4.3: The blue line represents the evolution of the relative error in a CPD computation. The program can stop because the average 2 is bigger than average 1.

Condition 6 avoids too long periods of negligible improvements. For instance, if the error is of order $\mathcal{O}(10^{-2})$, it is not necessary the program to waste time making hundreds iterations of improvements of order $\mathcal{O}(10^{-6})$. Figure 4.4 illustrates this kind of situation.

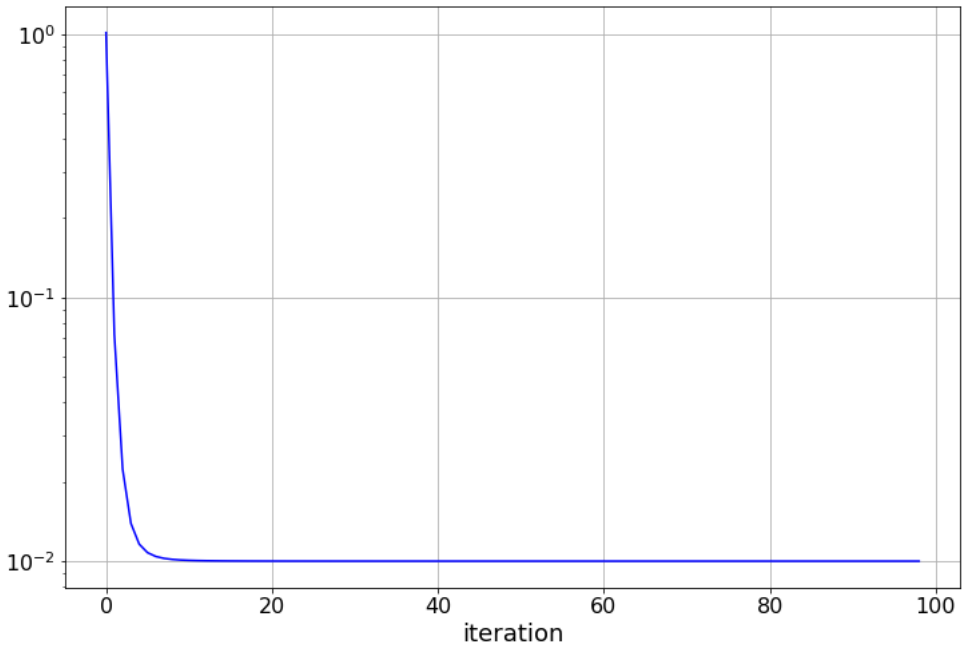

Figure 4.4: Note that the program could have stopped much earlier. Even if the errors are strictly decreasing, the additional accuracy is irrelevant compared to the final error.

The cost of the step size is of $R\sum_{\ell=1}^{L} R_\ell$ flops. Remember that $\nabla F(\mathbf{w}^{(k)}) = \mathbf{J}_f^T \cdot f(\mathbf{w}^{(k)})$ (lemma 3.4.2). With the notations used in the CG algorithm we have that $\nabla F(\mathbf{w}^{(k)}) = -\mathbf{A}^T \cdot \mathbf{b}$. Since this vector had to be computed in the CG algorithm, we can just store it be used for the stopping condition. The computation of its norm amounts to $R\sum_{\ell=1}^{L} R_\ell$ flops. The other stopping conditions are cheaper to compute so we don't count them.

4.1.3.7 Overall cost of dGN

We already have done all the complexity analysis of each algorithm and each small routine inside the dGN algorithm. Here we just put all them together in order to facilitate the presentation. We assign some memory costs with the symbol "–" when the respective cost is irrelevant.

Overall, the memory cost of one dGN iteration is of

$$\underbrace{R\sum_{\ell=1}^{L} R_\ell}_{\text{gradient}} + \underbrace{R^2(L+L^2)}_{\Pi^{(\ell',\ell'')} \text{ and } \Pi^{(\ell')}} + \underbrace{R\sum_{\ell=1}^{L} R_\ell}_{\text{vector result}} = 2R\sum_{\ell=1}^{L} R_\ell + R^2(L+L^2) \text{ floats},$$

while the computational cost is of

Task	Memory	Computational cost
Computing the residual	$\prod_{\ell=1}^{L} R_\ell$	$\mathcal{O}\left((L-1)\prod_{\ell=1}^{L} R_\ell\right)$
Computing the gradient	$R\sum_{\ell=1}^{L} R_\ell$	$\mathcal{O}\left(LR\prod_{\ell=1}^{L} R_\ell\right)$
Computing $\Pi^{(\ell',\ell'')}$ and $\Pi^{(\ell')}$	$R^2(L+L^2)$	$\mathcal{O}\left(R^2\left(L+\sum_{\ell=1}^{L} R_\ell\right)\right)$
CG method	$R\sum_{\ell=1}^{L} R_\ell$	$\mathcal{O}\left(i_{CG}(9R+LR^2)\sum_{\ell=1}^{L} R_\ell\right)$
Normalizing the factors	–	$\mathcal{O}\left(R\sum_{\ell=1}^{L} R_\ell\right)$
Damping parameter update	–	$\mathcal{O}(1)$
Stopping conditions	–	$\mathcal{O}\left(3R\sum_{\ell=1}^{L} R_\ell\right)$

Table 4.2: Memory and computational costs - III.

$$\mathcal{O}\Bigg(\underbrace{(L-1)\prod_{\ell=1}^{L} R_\ell}_{\text{residual}} + \underbrace{LR\prod_{\ell=1}^{L} R_\ell}_{\text{gradient}} + \underbrace{R^2\left(L+\sum_{\ell=1}^{L} R_\ell\right)}_{\Pi^{(\ell',\ell'')} \text{ and } \Pi^{(\ell')}} + \underbrace{i_{CG}(9R+LR^2)\sum_{\ell=1}^{L} R_\ell}_{\text{CG loop}} + \underbrace{3R\sum_{\ell=1}^{L} R_\ell}_{\text{stop conditions}}\Bigg) =$$

$$= \mathcal{O}\left((LR+L-1)\prod_{\ell=1}^{L} R_\ell + R^2\left(L+\sum_{\ell=1}^{L} R_\ell\right) + 3R\sum_{\ell=1}^{L} R_\ell + i_{CG}(9R+LR^2)\sum_{\ell=1}^{L} R_\ell\right) \text{ flops.}$$

4.1.3.8 Comparison to other algorithms

As mentioned in chapter 0, the solvers Tensorlab and Tensor Box also implements the Gauss-Newton method. It is of interest to compare the implementations and their complexities since they are based on the same method. To simplify out analysis, assume that the compression stage is already performed and limit the discussion here to cubic tensors of shape $R \times R \times \ldots \times R$.

Tensorlab is a tensor package with several algorithms to compute a more general decomposition, namely the *Block Term Decomposition* (BTD). They propose to work with the more general approximate Hessian, which have structure to be exploited, just as in this work, see theorem 4.5 [15]. Since their formulation is more general, they end with more equations to perform matrix-vector multiplication, which then can be

simplified in the particular case of the CPD. Preconditioning is employed too, but instead of a diagonal preconditioner they use a block diagonal preconditioner (also called a *block Jacobi preconditioner*). This approach may improve the conditioning substantially, but it comes with the cost of having to solve a linear system at each CG iteration. The cost to compute the error at each iteration is of $\mathcal{O}(2R^{L+1})$ flops and the cost to compute the cogradient is of $\mathcal{O}(2LR^{L+1})$ flops, see appendix of [15] for more details. The faster algorithm is the inexact Gauss-Newton, which costs $\mathcal{O}\left(c_1\left(\frac{5}{2}L^2R^2 + 8LR^2 + \frac{1}{3}LR^3\right)\right)$ flops to solve the inverse problem of each Gauss-Newton iteration. The value c_1 is the number of CG iterations performed, which is set to 15 by default. For more details we refer to the appendix of the same article.

Tensor Box starts decomposing the approximated Hessian as $\mathbf{H} = \mathbf{G} + \mathbf{Z}\mathbf{K}\mathbf{Z}^T$, where all these matrices are sparse with some structure, see theorem 4.2 of [4]. This structure is exploited so then can write the inverse of the damped approximated as $(\mathbf{H} + \mu\mathbf{I})^{-1} = \tilde{\mathbf{G}}_\mu - \mathbf{L}_\mu\mathbf{B}_\mu\mathbf{L}_\mu^T$, where all these matrices have some precise structure which can be used to accelerate the iterations. Their algorithm depends on the construction of several intermediate matrices, so we will only mention their associate costs, but the reader more interested is encouraged to read section 4.3 of [4]. Building matrix $\mathbf{K}$ costs $\mathcal{O}(L^3R^2)$ flops. Inverting all matrices $\Gamma_\mu^{(n)}$ costs $\mathcal{O}(LR^3)$ flops. The main cost of their algorithm is the computation of damped factors, which amounts to $\mathcal{O}(LR^{L+1})$ flops. These factor are used to construct a inverse linear problem with size $LR^2 \times LR^2$, which is solved with $\mathcal{O}(LR^3 + L^3R^6)$ flops.

For each solver, the costs to perform a single iteration is summarized below.

Tensor Fox:	$\mathcal{O}\left(LR^{L+1} + (L-1)R^L + 5LR^2 + LR^3 + c_2(9LR^2 + L^2R^3)\right)$ flops
Tensorlab:	$\mathcal{O}\left(2R^{L+1} + 2LR^{L+1} + c_1\left(\frac{5}{2}L^2R^2 + 8LR^2 + \frac{1}{3}LR^3\right)\right)$ flops
Tensor Box:	$\mathcal{O}\left(L^3R^2 + LR^3 + LR^{L+1} + LR^3 + L^3R^6\right)$ flops

At first, the analysis is not clear because c_2 is stochastic (it is the maximum number of CG iterations). Assuming that all 200 dGN iterations of Tensor Fox will be performed, the average number of CG iterations is $c_2 = 28$. The term R^{L+1} adds more cost to Tensorlab as the order increases. We note that Tensor Box is slower for low L and faster for bigger L (more precisely, for $L \geq 8$), with Tensor Fox being the second faster. For $L = 3$ and small R, we note that Tensorlab is faster than the others, but Tensor Fox starts to be the faster as R increases (around $R = 50$). For $L = 4, 5, 6, 7$ Tensor Fox is the faster algorithm for any $R \geq 10$. We remark that these complexity analysis depends on the constants associated to each cost.

Task	Memory
Compression	$(R+1)\sum_{\ell=1}^{L} I_\ell + \prod_{\ell=1}^{L} I_\ell + \prod_{\ell=1}^{L} R_\ell$
Initialization	$R\sum_{\ell=1}^{L} R_\ell$
dGN	$\prod_{\ell=1}^{L} R_\ell + R^2L^2 + R\sum_{\ell=1}^{L} R_\ell$
Uncompression	$R\sum_{\ell=1}^{L} I_\ell$

Table 4.3: Memory costs - IV.

4.1.4 Uncompression

After we have computed a CPD $(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)})$ for $\mathcal{S}$, we expect that

$$\mathcal{S} \approx (\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}) \cdot \mathcal{I}_{R\times\ldots\times R}.$$

On the other hand we know that $\mathcal{T} \approx (\mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(L)}) \cdot \mathcal{S}$, hence

$$\mathcal{T} \approx (\tilde{\mathbf{W}}^{(1)}, \ldots, \tilde{\mathbf{W}}^{(L)}) \cdot \mathcal{I}_{R\times\ldots\times R},$$

where $\tilde{\mathbf{W}}^{(\ell)} = \mathbf{U}^{(\ell)}\mathbf{W}^{(\ell)}$ for each $\ell = 1 \ldots R$. With this we can write

$$\mathcal{T} \approx \sum_{r=1}^{R} \tilde{\mathbf{W}}^{(1)}_{:r} \otimes \ldots \otimes \tilde{\mathbf{W}}^{(L)}_{:r},$$

an approximated CPD for $\mathcal{T}$.

Note that $\tilde{\mathbf{W}}^{(\ell)} \in \mathbb{R}^{R_\ell \times R}$, while $\mathbf{W}^{(\ell)} \in \mathbb{R}^{I_\ell \times R}$. The process of transforming smaller dimensional factor matrices $\tilde{\mathbf{W}}^{(\ell)}$ into the bigger factor matrices $\mathbf{W}^{(\ell)}$ is what we call *uncompression*. This is the last stage of all, after that we have a CPD for $\mathcal{T}$. The cost to produce each $\tilde{\mathbf{W}}^{(\ell)}$ is the cost to multiply a $I_\ell \times R_\ell$ matrix by a $R_\ell \times R$ matrix, which has a cost of $\mathcal{O}(I_\ell R_\ell R)$ flops. Therefore, the cost to produce the uncompressed CPD is of $\mathcal{O}\left(R\sum_{\ell=1}^{L} I_\ell R_\ell\right)$ flops.

We finish this section with a succinct summary of all costs to produce a CPD (tables 4.3 and 4.4) in Tensor Fox. Just as we denoted by i_{CG} the number of iterations of the CG algorithm, we denote by i_{dGN} the number of iterations of the dGN algorithm.

Task	Computational cost
Compression	$\mathcal{O}\left(\left(1+\prod_{\ell=1}^{L} I_\ell\right)\sum_{\ell=1}^{L} I_\ell + \sum_{\ell=1}^{L} I_\ell^3\right)$
Initialization	$\mathcal{O}\left(R\sum_{\ell=1}^{L} R_\ell\right)$
dGN	$\mathcal{O}\left(i_{dGN}\left((LR+R+L-1)\prod_{\ell=1}^{L} R_\ell + 3R\sum_{\ell=1}^{L} R_\ell + R^2\sum_{\ell=1}^{L} R_\ell + i_{CG}(8R+LR^2)\sum_{\ell=1}^{L} R_\ell\right)\right)$
Uncompression	$\mathcal{O}\left(R\sum_{\ell=1}^{L} I_\ell R_\ell\right)$

Table 4.4: Computational costs - IV.

4.2 Warming up

Now we start to make computational experiments. Let's start with a simple example in order to show some computational aspects we are interested in. For this suppose a multivariate function $\varphi : \mathbb{R}^L \to \mathbb{R}$ as described in example 0.3. To make everything more concrete let's consider the highly nonlinear function

$$\varphi(x, y, z) = \cos(1-x)\sin(1+y^2) - \sin(1+x^2)e^{x^2+y^2+z^2} - \cos(x)\ln(1+z).$$

Now consider some uniform grids in the interval $[0, 1]$, with six samples for x, five samples for y and four samples for z. With these points we form the tensor $\mathcal{T} \in \mathbb{R}^{6\times5\times4}$ defined as $t_{ijk} = \varphi(x_i, y_j, z_k)$, where $x_i = 0, 0.2, 0.4, 0.6, 0.8, 1$, $y_j = 0, 0.25, 0.5, 0.75, 1$ and $z_k = 0, 0.33333333, 0.66666667, 1$. We can write $\mathcal{T} = \sum_{\ell=1}^{3} \varphi_1^{(\ell)}(x)\varphi_2^{(\ell)}(y)\varphi_3^{(\ell)}(z)$, where

$$\begin{aligned}
\varphi_1^{(1)}(x) &= \cos(1-x),\\
\varphi_2^{(1)}(y) &= \sin(1+y^2),\\
\varphi_3^{(1)}(z) &= 1,\\
\varphi_1^{(2)}(x) &= -\sin(1+x^2)e^{x^2},\\
\varphi_2^{(2)}(y) &= e^{y^2},\\
\varphi_3^{(2)}(z) &= e^{z^2},\\
\varphi_1^{(3)}(x) &= -\cos(x),\\
\varphi_2^{(3)}(y) &= 1,\\
\varphi_3^{(3)}(z) &= \ln(1+z).
\end{aligned}$$

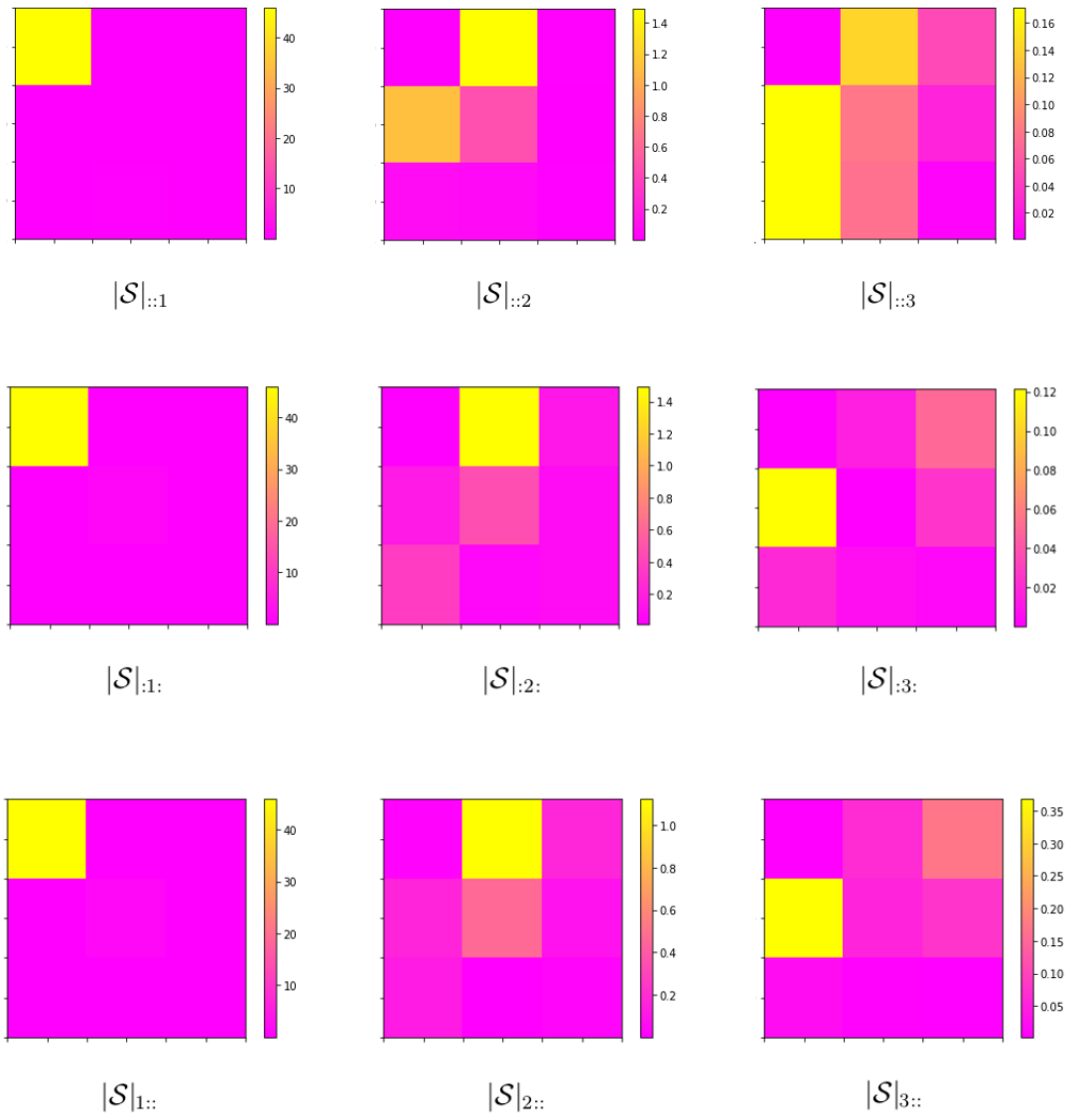

Figure 4.5: Energy distribution of truncation $\tilde{\mathcal{S}}$.

Let $\mathcal{T} = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}) \cdot \mathcal{S}$ be the MLSVD of $\mathcal{T}$. After the truncation process (described in 4.1.1) we obtain a tensor $\tilde{\mathcal{S}} \in \mathbb{R}^{3\times3\times3}$ whose respective relative error is $\frac{\|\mathcal{S}-\tilde{\mathcal{S}}\|}{\|\mathcal{S}\|} = 3.85 \cdot 10^{-14}$. Since the multilinear rank is limited by the rank, we know in advance that $rank_{\boxplus}(\mathcal{T}) \leq (3,3,3)$. Since the error showed is practically is minimal, the truncated MLSVD obtained is the actual MLSVD of $\mathcal{T}$, and we have the equality $rank_{\boxplus}(\mathcal{T}) = (3,3,3)$. We can visualize the energy distribution of $\mathcal{S}$ in figure 4.5, which plots the energy of the slices respective to each mode. Each square is an entry of a slice in absolute value. The bars on the side of each image indicates the magnitudes of the entries.

The tensor $\mathcal{S}$ can be showed explicitly, it is given by

Shape of truncation	Relative error
$1 \times 1 \times 1$	0.04345277
$1 \times 1 \times 2$	0.04345277
$1 \times 1 \times 3$	0.04345276
$1 \times 2 \times 1$	0.04345198
$1 \times 3 \times 1$	0.04345198
$2 \times 1 \times 1$	0.04345156
$3 \times 1 \times 1$	0.04345154
$2 \times 2 \times 1$	0.04330007
$2 \times 3 \times 1$	0.04321934
$3 \times 2 \times 1$	0.04255157
$3 \times 3 \times 1$	0.04246718
$2 \times 1 \times 2$	0.03591381
$3 \times 1 \times 2$	0.03588523
$2 \times 1 \times 3$	0.03572117
$3 \times 1 \times 3$	0.03549749
$1 \times 2 \times 2$	0.02888081
$1 \times 3 \times 2$	0.02887894
$1 \times 2 \times 3$	0.02871292
$1 \times 3 \times 3$	0.02869078
$2 \times 2 \times 2$	0.01093718
$2 \times 3 \times 2$	0.01060795
$2 \times 2 \times 3$	0.00964929
$2 \times 3 \times 3$	0.00919564
$3 \times 2 \times 2$	0.00720686
$3 \times 3 \times 2$	0.00668082
$3 \times 2 \times 3$	0.00296115
$3 \times 3 \times 3$	0

Table 4.5: Error of all possible truncations.

$$\mathcal{S} = \left\{ \begin{bmatrix} 45.84 & 0.012 & 0.00006 \\ 0.014 & 0.16 & 0.12 \\ 0.002 & -0.36 & -0.019 \end{bmatrix}, \begin{bmatrix} 0.0005 & 1.48 & -0.015 \\ 1.12 & -0.46 & -0.00031 \\ 0.065 & -0.053 & 0.0069 \end{bmatrix}, \begin{bmatrix} 0.00096 & -0.142 & 0.049 \\ 0.17 & -0.08 & -0.024 \\ -0.17 & 0.076 & 0.0038 \end{bmatrix} \right\}.$$

This presentation of $\mathcal{S}$ is based on its frontal slices, that is, $\mathcal{S} = \{\mathcal{S}_{::1}, \mathcal{S}_{::2}, \mathcal{S}_{::3}\}$.

Just as we did in 4.1.1, we could have obtained a smaller truncation $\tilde{\mathcal{S}}$ and then we would forget $\mathcal{S}$ by defining $\mathcal{S} = \tilde{\mathcal{S}}$. The matrices $\mathbf{U}^{(\ell)}$ would have some of its last columns deleted and we would redefine these matrices too. The original MLSVD is forgotten and we use the truncated version in its place. Note that, by limiting ourselves to the truncated MLSVD, the best possible CPD we can achieve will have relative error equal[2] to the relative error between $\tilde{\mathcal{S}}$ and $\mathcal{S}$. By varying $1 \leq i, j, k \leq 3$ and checking all possible truncations $\mathcal{S}_{:i,:j,:k}$ we can get a better picture of the relation between truncation and error. Table 4.5 below shows how the truncation and errors are related in this example.

[2]It is still possible to have smaller errors by pure luck, but we won't take this in account.

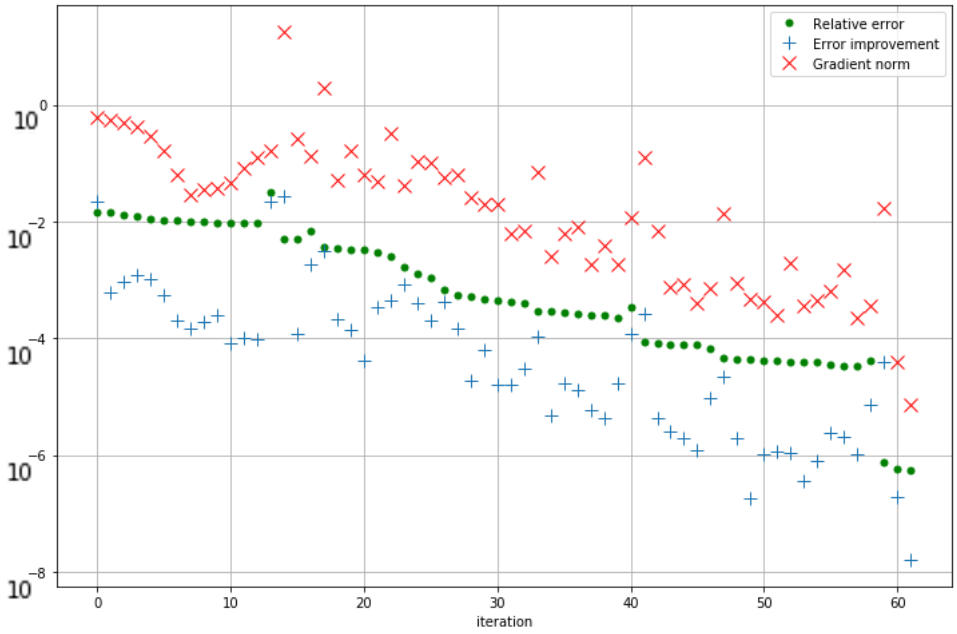

Figure 4.6: Plot of the evolution of several measures made during the computation of a CPD.

Our method of initialization based on the MLSVD chooses the three brightest yellow squares (where the energy is most concentrated) of figure 4.5 to generate an initial approximation $\mathcal{S}^{(0)}$. This initial tensor is

$$\mathcal{S}^{(0)} = s_{111}\ \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 + s_{212}\ \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + s_{122}\ \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2 =$$

$$= 45.84\ \mathbf{e}_1 \otimes \mathbf{e}_1 \otimes \mathbf{e}_1 + 1.12\ \mathbf{e}_2 \otimes \mathbf{e}_1 \otimes \mathbf{e}_2 + 1.48\ \mathbf{e}_1 \otimes \mathbf{e}_2 \otimes \mathbf{e}_2.$$

The respective error of this approximated rank-3 CPD is

$$\frac{\|\mathcal{T} - (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}) \cdot \mathcal{S}^{(0)}\|}{\|\mathcal{T}\|} = 0.0153,$$

which is already small enough for a first iteration. From this tensor we start the dGN iterations in order to compute a better rank-3 CPD for $\mathcal{T}$. The results of the computations are summarized in figure 4.6. All the plots are in $\log_{10}$ scale. We can see the error decreasing but going up sometimes. The algorithm is not necessarily monotonic, but it always auto-correct itself. Adding line-search methods at each iteration were attempted, and this indeed made the algorithm monotonic (at least in practice). However this monotonicity seems to make the iterations to be drawn by local minima. Every attempt to decrease the error a little bit at each iteration led to some kind of local minima attraction. Our current algorithm is more "erratic" sometimes but is this behavior is what makes the steps "run away" from local minima. The *error improvement* showed is the absolute value of the difference between two consecutive relative errors. Figure 4.6 shows the evolution of the error, improvement and gradient as the iterative process progresses.

After all computations are finished the program outputs the approximated CPD $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathcal{I}_{3\times3\times3}$, where

$$\mathbf{X} = \begin{bmatrix} -0.81709105 & 0.78059637 & -0.53290962 \\ -0.87159249 & 0.76495679 & -0.6872218 \\ -1.04470608 & 0.71882424 & -0.81413498 \\ -1.36099384 & 0.64403858 & -0.90858919 \\ -1.83712919 & 0.54358195 & -0.96681853 \\ -2.40011984 & 0.42145802 & -0.98650216 \end{bmatrix},$$

$$\mathbf{Y} = \begin{bmatrix} -0.97352551 & -0.72100749 & 0.83684308 \\ -1.03631189 & -0.72099582 & 0.86877435 \\ -1.25003023 & -0.72096898 & 0.94377806 \\ -1.70859268 & -0.72095345 & 0.99448294 \\ -2.64633452 & -0.72099573 & 0.90429673 \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} -1.05782625 & -0000.246843654 & -1.01914978 \\ -1.18213922 & 0.510766716 & -1.01896187 \\ -1.64980857 & 0.907135088 & -1.01880278 \\ -2.87546993 & 1.23098132 & -1.01864246 \end{bmatrix}.$$

If everything worked correctly, the first column of $\mathbf{X}$ should be some multiple[3] of the outputs of $\varphi_1^{(1)}(x) = \cos(1-x)$ for $x = 0, 0.2, 0.4, 0.6, 0.8, 1$. We can try scaling the columns of $\mathbf{X}$ to fit the data. If after several trials none of them seems to work, it may be a good idea to compute another CPD, or change the initialization method, or introduce more restrictions to the model, etc.

4.3 Benchmark tensors

As we already said sometimes, most parameters and subroutines of Tensor Fox were obtained by experience. The first attempts were based on other implementations, and from there we start to fine tuning to get the most of every single parameter. To make sure the choices are optimal and generic enough, it is necessary to test the performance for a diverse family of tensors, taking into account what are the known difficulties and what is applicable. In this section, we present all the tensors used for testing and benchmarking.

[3]This issue comes from the scaling we are using. It is possible to re-scale the factors to fit the original functions without changing the CPD.

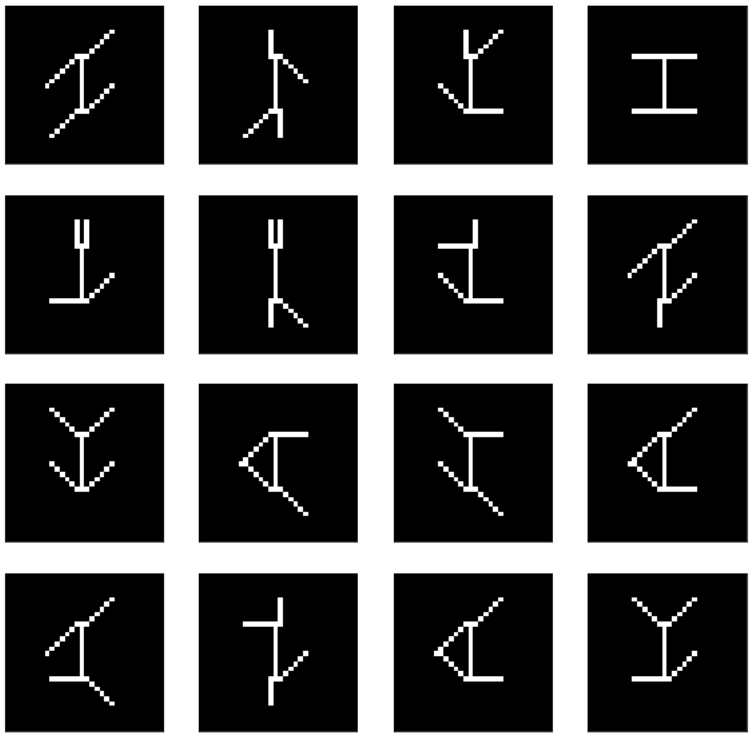

Figure 4.7: Swimmer tensor.

4.3.1 Swimmer

This tensor was constructed based on the paper [23] as an example of a nonnegative tensor. It is a set of 256 images of dimensions 32×32 representing a swimmer. Each image contains a torso (the invariant part) of 12 pixels in the center and four limbs of 6 pixels that can be in one of 4 positions. In this work they proposed to use a rank $R = 50$ tensor to approximate, and we do the same for our test. In figure 4.7 we can see some frontal slices of this tensor.

4.3.2 Handwritten digits

This is a classic tensor in machine learning, it is the MNIST[4] database of handwritten digits. Each slice is a image of dimensions 20×20 of a handwritten digit. Also, each 500 consecutive slices correspond to the same digit, so the first 500 slices correspond to the digit 0, the slices 501 to 1000 correspond to the digit 1, and so on. We choose $R = 150$ as a good rank to construct the approximating CPD to this tensor. In figure 4.8 we can see some frontal slices of this tensor.

[4] http://yann.lecun.com/exdb/mnist/

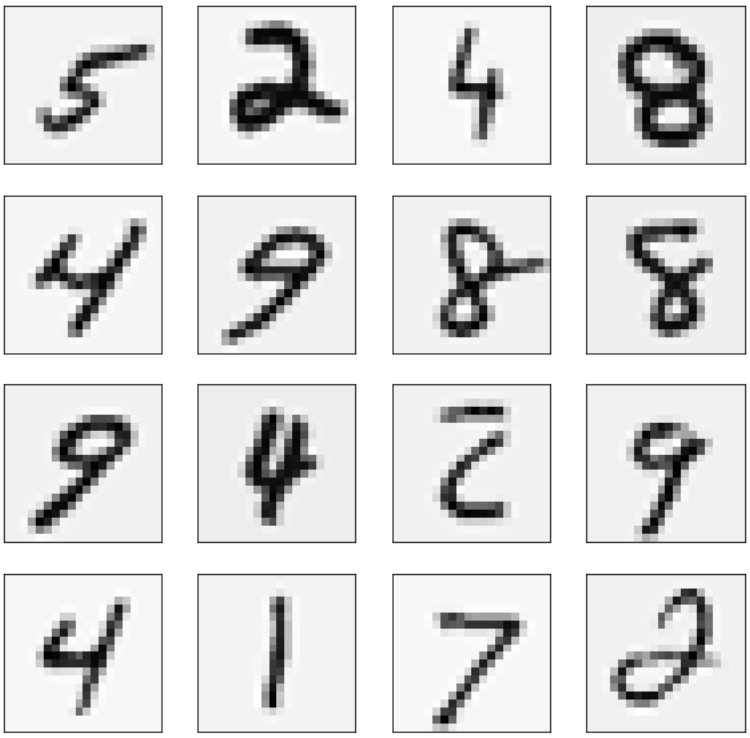

Figure 4.8: Handwritten digits tensor.

4.3.3 Border rank

Remember the definition and discussion about border rank in 1.5. We observed that the strict inequality $\underline{rank}(\mathcal{T}) < rank(\mathcal{T}) = R$ can happen, and this means that the set of rank-R tensors is not closed. This phenomenon makes the CPD computation a challenging problem. The paper [20] has a great discussion on this subject. In the same paper (and theorem 1.5.6) they show that

$$\mathcal{T}^{(n)} = n\left(\mathbf{x}^{(1)} + \frac{1}{n}\mathbf{y}^{(1)}\right) \otimes \left(\mathbf{x}^{(2)} + \frac{1}{n}\mathbf{y}^{(2)}\right) \otimes \left(\mathbf{x}^{(3)} + \frac{1}{n}\mathbf{y}^{(3)}\right) - n\mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{x}^{(3)}$$

is a sequence of rank sequence of rank 2 tensors converging to a tensor of rank 3, where each pair $\mathbf{x}^{(\ell)}, \mathbf{y}^{(\ell)} \in \mathbb{R}^{I_\ell}$ is linearly independent. Remember that the limit tensor is $\mathcal{T} = \mathbf{x}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{y}^{(3)} + \mathbf{x}^{(1)} \otimes \mathbf{y}^{(2)} \otimes \mathbf{x}^{(3)} + \mathbf{y}^{(1)} \otimes \mathbf{x}^{(2)} \otimes \mathbf{x}^{(3)}$. We choose to compute a CPD of rank $R = 2$ to see how the algorithms behaves when we try to approximate a problematic tensor by tensor with low rank. In theory it is possible to have arbitrarily good approximations.

4.3.4 Matrix multiplication

Let $\mathcal{M}_N \in \mathbb{R}^{N^2 \times N^2 \times N^2}$ be the tensor associated with the multiplication between two matrices in $\mathbb{R}^{N \times N}$. The classic form of $\mathcal{M}_N$ is given by

$$\mathcal{M}_N = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{l} vec(\mathbf{e}_{ij}) \otimes vec(\mathbf{e}_{jk}) \otimes vec(\mathbf{e}_{ik}),$$

where $\mathbf{e}_{ij}$ is the matrix $N \times N$ with entry (i, j) equal to 1 and the remaining entries equal to zero. Since Strassen [25] it is known that matrix multiplication between matrices of dimensions $N \times N$ can be made with $\mathcal{O}(N^{\log_2 7})$ operations. Many improvements were made after Strassen but we won't enter in the details here. For the purpose of testing we choose the small value $N = 5$ and the rank $R = \lceil 5^{\log_2 7} \rceil = 92$ in honor of Strassen. However note that this is probably not the exact rank of $\mathcal{M}_5$, so this test is about a strict low rank approximation of a difficult tensor.

4.3.5 Collinear factors

The phenomenon of swamps occurs when all factors in each mode are almost collinear. Their presence is a challenge for many algorithms because they can slow down convergence. Now we will create synthetic data to simulate various degrees of collinearity between the factors. We begin generating three random matrices $\mathbf{M}_X \in \mathbb{R}^{m \times R}, \mathbf{M}_Y \in \mathbb{R}^{n \times R}, \mathbf{M}_Z \in \mathbb{R}^{p \times r}$, where each entry is drawn from the normal distribution with mean 0 and variance 1. After that we perform QR decomposition of each matrix, obtaining the decompositions $\mathbf{M}_X = \mathbf{Q}_X \mathbf{R}_X, \mathbf{M}_Y = \mathbf{Q}_Y \mathbf{R}_Y, \mathbf{M}_Z = \mathbf{Q}_Z \mathbf{R}_Z$. The matrices $\mathbf{Q}_X, \mathbf{Q}_Y, \mathbf{Q}_Z$ are orthogonal. Now fix three columns $\mathbf{q}_X^{(i')}, \mathbf{q}_Y^{(j')}, \mathbf{q}_Z^{(k')}$ of each one of these matrices. The factors $\mathbf{X} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(R)}] \in \mathbb{R}^{m \times R}$, $\mathbf{Y} = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(R)}] \in \mathbb{R}^{n \times R}$, $\mathbf{Z} = [\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(R)}] \in \mathbb{R}^{p \times R}$ are generated by the equations below.

$$\begin{aligned}
\mathbf{x}^{(i)} &= \mathbf{q}_X^{(i')} + c \cdot \mathbf{q}_X^{(i)}, \quad i = 1 \ldots R \\
\mathbf{y}^{(j)} &= \mathbf{q}_Y^{(j')} + c \cdot \mathbf{q}_Y^{(j)}, \quad j = 1 \ldots R \\
\mathbf{z}^{(k)} &= \mathbf{q}_Z^{(k')} + c \cdot \mathbf{q}_Z^{(k)}, \quad k = 1 \ldots R
\end{aligned}$$

The parameter $c \geq 0$ defines the degree of collinearity between the vectors of each factor. A value of c close to 0 indicates high degree of collinearity, while a high value of c indicates low degree of collinearity.

Another phenomenon that occurs in practice is the presence of noise in the data. So we will treat these two phenomena at once in this benchmark. After generating the factors $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ we have a tensor $\mathcal{T} = (\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \cdot \mathcal{I}_{R \times R \times R}$. That is, $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are the exact CPD of $\mathcal{T}$. Now consider a noise $\mathcal{N} \in \mathbb{R}^{m \times n \times p}$ such that each entry of $\mathcal{N}$ is obtained by the

normal distribution with mean 0 and variance 1. Thus we form the tensor $\hat{\mathcal{T}} = \mathcal{T} + \nu \cdot \mathcal{N}$, where $\nu > 0$ defines the magnitude of the noises. The idea is to compute a CPD of $\hat{\mathcal{T}}$ of rank R and then evaluate the relative error between this tensor and $\mathcal{T}$. We expect the computed CPD to clear the noises and to be close to $\mathcal{T}$ (even if it is not close to $\hat{\mathcal{T}}$). We will fix $\nu = 0.01$ and generate tensors for $c = 0.1, 0.5, 0.9$. In all cases we will be using $m = n = p = 300$ and $R = 15$. This is a particularly difficult problem since we are considering swamps and noises at once. The same procedure to generate tensors were used for benchmarking in [4].

4.3.6 Double bottlenecks

We proceed almost in the same as before for swamps, we used the same procedure to generate the first two columns of each factor matrix, then the remaining columns are equal to the columns of the QR decomposition. After generating the factors $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ we consider a noise $\mathcal{N} \in \mathbb{R}^{m \times n \times p}$ such that each entry of $\mathcal{N}$ is obtained by the normal distribution with mean 0 and variance 1. Thus we form the tensor $\hat{\mathcal{T}} = \mathcal{T} + \nu \cdot \mathcal{N}$, where $\nu > 0$ defines the magnitude of the noises. The collinear parameter used for the tests are $c = 0.1, 0.5$, and we fix $\nu = 0.01$ as before. The procedure before the noise is presented below.

$$\begin{aligned}
\mathbf{x}^{(i)} &= \mathbf{q}_X^{(i')} + c \cdot \mathbf{q}_X^{(i)}, \quad i = 1, 2 \\
\mathbf{y}^{(j)} &= \mathbf{q}_Y^{(j')} + c \cdot \mathbf{q}_Y^{(j)}, \quad j = 1, 2 \\
\mathbf{z}^{(k)} &= \mathbf{q}_Z^{(k')} + c \cdot \mathbf{q}_Z^{(k)}, \quad k = 1, 2
\end{aligned}$$

$$\begin{aligned}
\mathbf{x}^{(i)} &= \mathbf{q}_X^{(i')}, \quad i = 3 \dots R \\
\mathbf{y}^{(j)} &= \mathbf{q}_Y^{(j')}, \quad j = 3 \dots R \\
\mathbf{z}^{(k)} &= \mathbf{q}_Z^{(k')}, \quad k = 3 \dots R
\end{aligned}$$

4.4 Fine tuning

We will conduct the tests using only third order tensors The dimensions will be written as m, n, p so the tensors are elements of $\mathbb{R}^{m \times n \times p}$. As already mentioned, the parameters `maxiter`, `tol`, `cg_maxiter` were found after several experimentations and tests. Here we consider varying each one of them around the default ones and test them with our test tensors. The results we are going to show here are just a fraction of the total work done in order to achieve these values. In particular, `cg_maxiter` was a result of several different models and parameter fine tuning. We will conduct an *hyperparameter* search

on the parameters. This consists in considering a high dimensional grid of values, where each dimension corresponds to a certain parameter. Each point in the grid correspond to a model (a particular choice of parameters) on which we evaluate its performance. The grid we are considering here is given by $P_1 \times P_2 \times P_3 \times P_4$, where

$$\begin{aligned} P_1 &= \{100, 200, 400\}, \\ P_2 &= \{10^{-4}, 10^{-6}, 10^{-8}\}, \\ P_3 &= \{0.2, 0.3, 0.4, 0.5\}, \\ P_4 &= \{0.6, 0.7, 0.8, 0.9\}. \end{aligned}$$

The idea is that `maxiter` $= p_1$, `tol` $= p_2 \cdot mnp$, $a = p_3, b = p_4$, where $p_i \in P_i$ for all i. Consider that `cg_maxiter` $\sim \left[1 + \lceil k^a \rceil, 2 + \lceil k^b \rceil\right]$. It is possible to use more values and larger intervals, but experience showed that good performance is attained around these values. We test each tensor 20 times and take the average time and average error to measure its performance. In the plots of figure 4.9, each small point correspond to a model (a particular choice of parameters). The better choices correspond to points closer to the origin, which translates to small error in a small time. We selected some regions of interest in order to discard the failed models. The sizes of these regions depend on the considered tensor. Sometimes the precision is more important that the time (at a certain degree) and sometimes we can afford to lose precision to gain in time. This is a subtle matter, but in general we will avoid models falling too far away from the region close to the origin. The goal is to obtain a well balanced set of parameters. The region between the red dotted lines and the axes indicates the region of interest for each test tensor. Every point inside these regions are registered. The idea is too see if there are models appearing inside these regions for different tensors. This would indicate that the model performs well in more than just one tensor. The ones appearing more frequently are more likely to be well balanced and generic. Table 4.6 shows the results of all models which appeared at least four times inside some of the defined regions. The models marked in red are the ones with five or more occurrences. These are the best ones by our criteria. We used the abbreviations Sw = Swimmer, Hw = Handwritten, Br = Border rank, Mm = Matrix multiplication, Swp 0.1 = Swamp with $c = 0.1$, and similar for the other swamp tensors, and Bn 0.1 = Bottleneck with $c = 0.1$, with the same considerations as the swamp tensors. To decide which model to choose we also plot the position of these four models. The first thing we note is that all models performs reasonably well and quite similar for the first four test tensors. The differences begin to appear when we introduce collinearity. In the case of the swamp tensors the models with tolerance 10^{-8} are able to attain the same error of the other models but in a bigger time. On the other hand, in the case of bottleneck tensors the models with tolerance 10^{-6} takes the same time of the other models but they get a noticeable bigger error for $c = 0.1$.

maxiter	tol	a	b	Sw	Hw	Br	Mm	Swp 0.1	Swp 0.5	Swp 0.9	Bn 0.1	Bn 0.5	Bn 0.9
100	10^{-8}	0.4	0.7						x	x	x	x	
200	10^{-6}	0.4	0.9	x	x			x		x	x		
200	10^{-6}	0.5	0.9		x		x			x	x		x
200	10^{-8}	0.5	0.9	x	x	x	x					x	
400	10^{-6}	0.2	0.8					x	x		x	x	
400	10^{-6}	0.2	0.9	x	x				x				x
400	10^{-6}	0.3	0.9		x			x	x		x		
400	10^{-6}	0.4	0.7					x	x		x	x	
400	10^{-6}	0.4	0.8	x	x				x	x			x
400	10^{-6}	0.4	0.9		x				x		x	x	
400	10^{-6}	0.5	0.9		x			x	x				x

Table 4.6: Hyperparameter grid search - best models.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.0011059	0.0000071	1.27 sec	0.13 sec
200, 10^{-6}, 0.5, 0.9	0.0018093	0.0000155	1.43 sec	0.19 sec
200, 10^{-8}, 0.5, 0.9	0.0013749	0.0000139	1.78 sec	0.02 sec
400, 10^{-6}, 0.4, 0.8	0.0029579	0.0000282	1.39 sec	0.25 sec

Table 4.7: Swimmer tensor - final best performances.

All these four models are good, but we have to decide between one of them. We can repeat the previous tests but with more repetitions to get more accurate performance statistics. Since there are few models to test this is a possible task. The tests are repeated with 50 repetitions instead of 20, and now we also consider the variance of the error and time.[5] The results are showed in the next tables. We start discarding the model where $\texttt{tol} = 10^{-8}$ since this one clearly is more demanding. At the end of the day, "speed shall prevail". The other three models are more well balanced so any choice we make is good enough. Our final choice is $\texttt{maxiter} = 200$, $\texttt{tol} = 10^{-6}$, $a = 0.4, b = 0.9$ since it has competitive errors measures with better timings in most of the tests.

We point out that the algorithm has a lot of room to improvements. The parameters choice may be changed, all updating strategies are wildly open to modifications, the preconditioner is surely not optimal (in fact, future work includes testing more preconditioners), and the compression strategies also can be changed.

[5] Given random variables $X_1, \dots, X_N$ i.i.d. (independent and identically distributed), the associated *empirical variance* is $\overline{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$, where $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the *empirical mean*

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.0816178	0.0000001	15.32 sec	7.49 sec
200, 10^{-6}, 0.5, 0.9	0.0815007	0.0000000	15.58 sec	4.09 sec
200, 10^{-8}, 0.5, 0.9	0.0814086	0.00000001	17.50 sec	1.74 sec
400, 10^{-6}, 0.4, 0.8	0.0817002	0.00000001	16.69 sec	7.57 sec

Table 4.8: Handwritten tensor - final best performances.

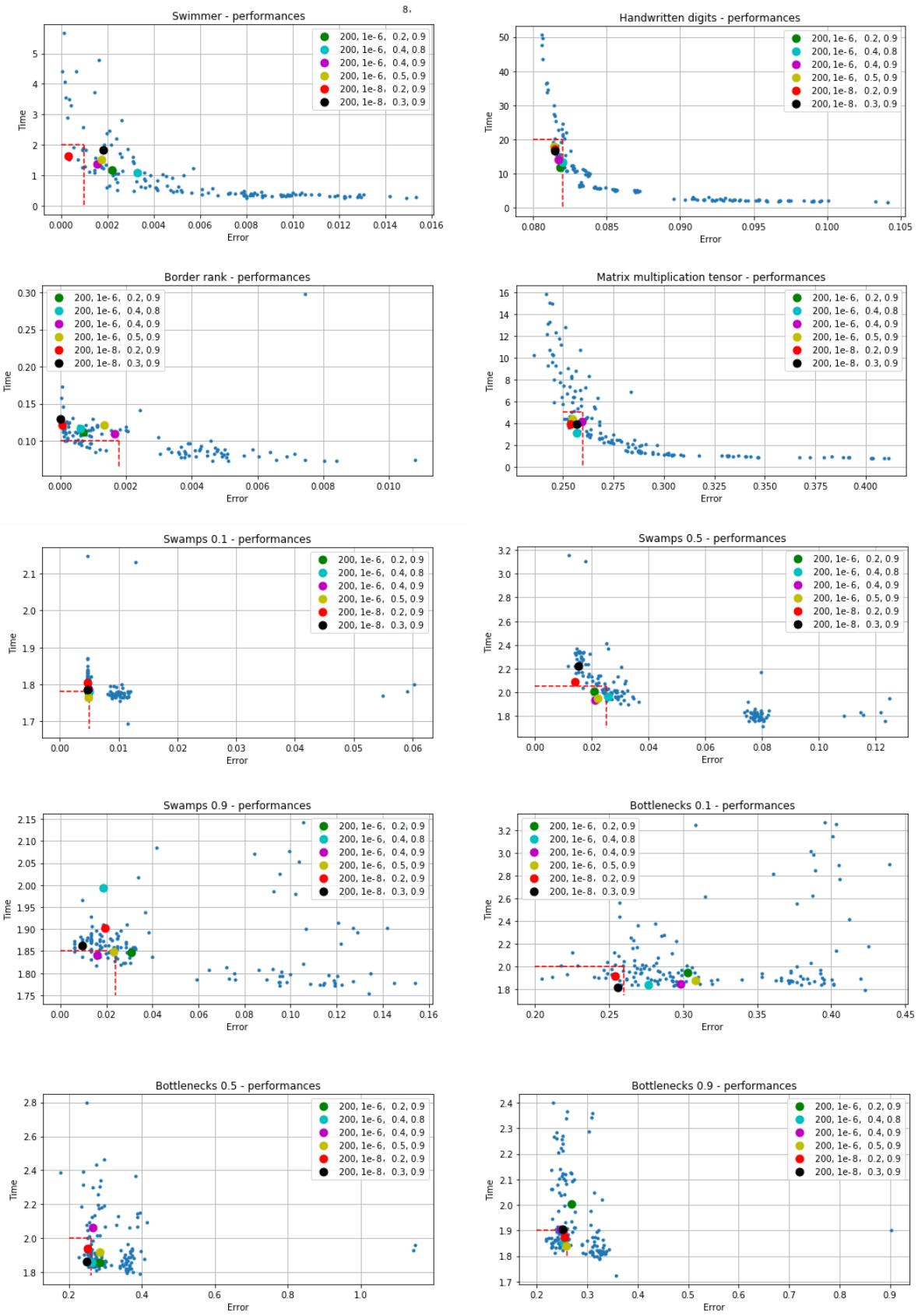

Figure 4.9: The location of the best models may help to decide which one is better.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.0016909	0.0000044	0.11 sec	0.0003 sec
200, 10^{-6}, 0.5, 0.9	0.0027261	0.0000046	0.11 sec	0.0002 sec
200, 10^{-8}, 0.5, 0.9	0.0000106	0.00000001	0.12 sec	0.0004 sec
400, 10^{-6}, 0.4, 0.8	0.0005573	0.0000018	0.13 sec	0.002 sec

Table 4.9: Border rank tensor - final best performances.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.2384032	0.0007164	4.37 sec	0.20 sec
200, 10^{-6}, 0.5, 0.9	0.2396802	0.0005141	4.34 sec	0.085 sec
200, 10^{-8}, 0.5, 0.9	0.2423333	0.0004590	4.33 sec	0.067 sec
400, 10^{-6}, 0.4, 0.8	0.2269175	0.0006688	9.35 sec	5.06 sec

Table 4.10: Matrix multiplication tensor - final best performances.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.0288477	0.0000026	1.86 sec	0.07 sec
200, 10^{-6}, 0.5, 0.9	0.0288307	0.0000027	1.83 sec	0.007 sec
200, 10^{-8}, 0.5, 0.9	0.0326175	0.0000003	2.63 sec	0.07 sec
400, 10^{-6}, 0.4, 0.8	0.0278611	0.0000026	1.79 sec	0.003 sec

Table 4.11: Swamp 0.1 tensor - final best performances.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.1000538	0.0000001	1.78 sec	0.003 sec
200, 10^{-6}, 0.5, 0.9	0.1002456	0.0000001	1.86 sec	0.013 sec
200, 10^{-8}, 0.5, 0.9	0.1010695	0.00000001	2.69 sec	0.07 sec
400, 10^{-6}, 0.4, 0.8	0.0999770	0.0000001	1.79 sec	0.007 sec

Table 4.12: Swamp 0.5 tensor - final best performances.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.1376466	0.0002145	1.96 sec	0.30 sec
200, 10^{-6}, 0.5, 0.9	0.1374688	0.0000559	1.83 sec	0.010 sec
200, 10^{-8}, 0.5, 0.9	0.1337583	0.0000065	1.83 sec	0.005 sec
400, 10^{-6}, 0.4, 0.8	0.1366625	0.0000472	1.85 sec	0.011 sec

Table 4.13: Swamp 0.9 tensor - final best performances.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.2876770	0.0090001	1.86 sec	0.016 sec
200, 10^{-6}, 0.5, 0.9	0.2738036	0.0110769	1.85 sec	0.009 sec
200, 10^{-8}, 0.5, 0.9	0.2562058	0.0129000	1.92 sec	0.019 sec
400, 10^{-6}, 0.4, 0.8	0.2624673	0.0085848	1.87 sec	0.006 sec

Table 4.14: Bottleneck 0.1 tensor - final best performances.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.2822554	0.0059065	1.87 sec	0.011 sec
200, 10^{-6}, 0.5, 0.9	0.2669849	0.0052620	1.84 sec	0.007 sec
200, 10^{-8}, 0.5, 0.9	0.2447143	0.0048144	1.87 sec	0.016 sec
400, 10^{-6}, 0.4, 0.8	0.2853812	0.0038925	1.82 sec	0.004 sec

Table 4.15: Bottleneck 0.5 tensor - final best performances.

Model	Average error	Variance error	Average time	Variance time
200, 10^{-6}, 0.4, 0.9	0.2578984	0.0018668	1.80 sec	0.003 sec
200, 10^{-6}, 0.5, 0.9	0.2654803	0.0013860	1.94 sec	0.005 sec
200, 10^{-8}, 0.5, 0.9	0.2446079	0.0022603	1.98 sec	0.004 sec
400, 10^{-6}, 0.4, 0.8	0.2568034	0.0029858	1.95 sec	0.002 sec

Table 4.16: Bottleneck 0.9 tensor - final best performances.

4.5 Tensor Fox vs. other implementations

4.5.1 Procedure

We have selected a set of very distinct tensors to test the known tensor implementations. Given a tensor $\mathcal{T}$ and a rank R, we compute the CPD of TFX (short for *Tensor Fox*) with the default maximum number of iterations[6] 100 times and retain the best result, i.e., the CPD with the smallest relative error. Let ε be this error. Now let ALG be any other algorithm implemented by some of the mentioned libraries. We set the maximum number of iterations to `maxiter`, keep the other options with their defaults, run ALG with these options 100 times. The only accepted solutions are the ones with relative error smaller that $\varepsilon + \varepsilon/100$. Between all accepted solutions we return the one with the smallest running time. If none solution is accepted, we increase it to `maxiter` by a certain amount and repeat.

We try the values `maxiter` $= 5, 10, 50, 100, 150, \ldots, 900, 950, 1000$, until there is an accepted solution. The running time associated with the accepted solution is the accepted time. Otherwise we consider that ALG failed. We also consider a fail if the computational time is exceedingly high, in which case we can stop increasing `maxiter` earlier. These procedures favour all implementations against TFX since we are trying a solution close to the solution of TFX with the minimum number of iterations. This benchmark measures the computational effort that each program makes to achieve the precision of TFX's precision. We consider it to be a fair measure of performance, although other kind of benchmarks might be considered. We remark that the iteration process is always initiated with a random point. The option to generate a random initial point is offered by all libraries, and we use each one they offer (sometimes random initialization was already the default option). There is no much difference in their initializations, which basically amounts to draw random factor matrices from the standard Gaussian distribution. The time to perform the MLSVD, or any kind of preprocessing, is included in the time measurements. If one want to reproduce the tests presented here, they can be found at https://github.com/felipebottega/Tensor-Fox/tree/master/tests.

4.5.2 State of art implementations

Now we briefly describe what algorithms will be used in our tests together with their corresponding implementations source. Some general algorithms are repeated but with variations, which are particular of each implementation. The ALS algorithm, for instance, is the one with more variations because people are trying to improve it for decades. For all Tensorlab algorithm implementation we recommend reading [15], for the Tensor Toolbox we recommend [17], for TensorLy we recommend [18], and for TensorBox we recommend

[6] Remember that the default for the maximum number of iterations of TFX is `maxiter`= 200.

[4, 6]. In these benchmarks we used Tensorlab version 3.0 and Tensor Toolbox version 3.1.

4.5.2.1 TFX

The algorithm used in TFX's implementation is the nonlinear least squares scheme described in the previous section. There are more implementation details to be discussed, but the interested reader can check [1] for more information.

4.5.2.2 ALS

This is the Tensorlab's implementation of ALS algorithm. Although ALS is remarkably fast and easy to implement, it is not very accurate specially in the presence of bottlenecks or swamps. It seems (see [15]) that this implementation is very robust while still fast.

4.5.2.3 NLS

This is the Tensorlab's implementation of NLS algorithm. This is the one we described in the previous section. We should remark that this implementation is similar to TFX's implementation at some points, but there are big differences when we look in more details. In particular the compression procedure, the preconditioner, the damping parameter update rule and the number of iterations of the conjugate gradient are very different.

4.5.2.4 MINF

This is the Tensorlab's implementation of the problem as an optimization problem. They use a quasi-Newton method, the limited-memory BFGS, and consider equation 3.2 just as a minimization of a function.

4.5.2.5 OPT

Just as the MINF approach, the OPT algorithm is a implementation of Tensor Toolbox, which considers 3.2 as a minimization of a function. They claim that using the algorithm option "lbfgs" is the preferred option[7], so we used this way.

4.5.2.6 Tly-ALS

TensorLy has only one way to compute the CPD, which is a implementation of the ALS algorithm. We denote it by Tly-ALS, do not confuse with ALS, the latter is the Tensorlab's implementation.

[7] https://www.tensortoolbox.org/cp_opt_doc.html

Solver	Shape of factor matrices
Tensor Toolbox	$m \times R,\ \ n \times R,\ \ p \times R$
Tensorbox	$m \times R,\ \ n \times R,\ \ p \times R$
Tensorly	$m \times R,\ \ n \times R,\ \ p \times R$
Tensorlab	$\min(m, R) \times R,\ \ \min(n, R) \times R,\ \ \min(p, R) \times R$
Tensor Fox	$R_1 \times R,\ \ R_2 \times R,\ \ R_3 \times R$

Table 4.17: Shapes of the factor matrices of each implementation for rank-R CPD computation of a tensor with shape $m \times n \times p$.

4.5.2.7 fLMa

fLMA stands for *fast Levenberg-Marquardt algorithm*, and it is a different version of the damped Gauss-Newton.

In Tensorlab it is possible to disable or not the option of refinement. The first action in Tensorlab is to compress the tensor and work with the compressed version. If we want to use refinement then the program uses the compressed solution to compute a refined solution in the original space. This can be more demanding but can improve the result considerably. In our experience working in the original space is not a good idea because the computational cost increases drastically and the gain in accuracy is generally very small. Still we tried all Tensorlab algorithms with and without refinement. Table 4.17 shows compares the shapes of the factor matrices for each implementation, and there we can see that most of them them doesn't compress the tensor. We will write ALSr, NLSr and MINFr for the algorithms ALS, NLS and MINF with refinement, respectively.

4.5.3 Computational results

We used Linux Mint operational system in our tests. All tests were conducted using a processor Intel Core i7-4510U - 2.00GHz (2 physical cores and 4 threads) and 8GB of memory. The packages mentioned run in Python (Tensorly and TensorFox) or Matlab (Tensorlab, Tensor Toolbox and TensorBox). We use Python - version 3.6.5 and Matlab - version 2017a. In both platforms we used BLAS MKL-11.3.1. Finally, we want to mention that TFX is implemented in Python with Numba, a specialized library to make "just-in-time compilation" (JIT).[8] We used the version 0.41 of Numba in these tests.

In figure 4.10 there are some charts, each one shows the best running time of the algorithms with respect to each one of the tensors describe previously. If some algorithm is not included in a chart, it means that the algorithm was unable to achieve an acceptable error within the conditions described at the beginning if this section.

The first thing we should note is that there isn't a single algorithm which was able to match the speed of TFX in all tests. In the first four tests NLS algorithm could

[8] http://numba.pydata.org/

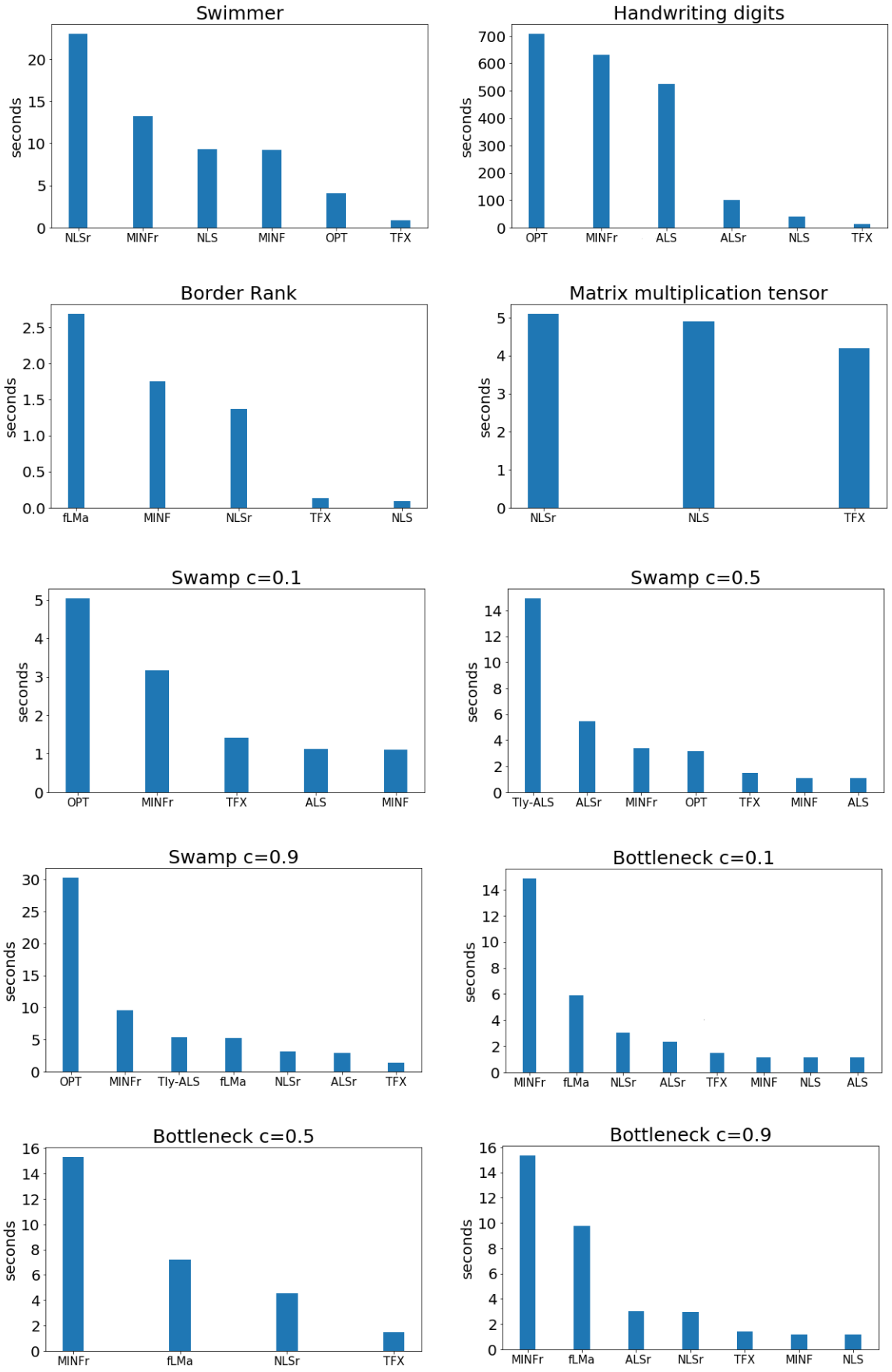

Figure 4.10: Benchmarks of all tensors and all implementations.

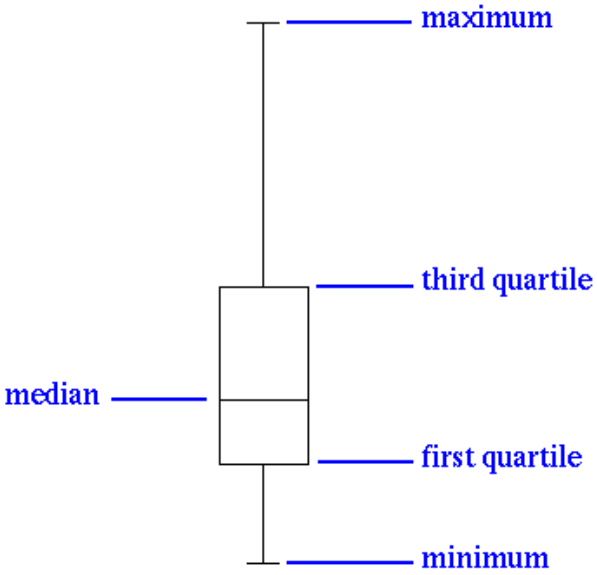

Figure 4.11: The box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.

deliver an accepted solution in a reasonable time, but when collinearity was introduced this algorithm faced problems. When it produced a solution it was within a small time, but the algorithm can't be regarded as robust since it can fail to produce a solution in these cases. For this MINF and ALS looks more reliable while also being very fast. Still, these two algorithms couldn't produce an acceptable solution in some cases (for instance ALS only produced 4 solutions in the total of 12 tensors) and the existence of a solution can depend on the refinement option being active. Because of that their robustness is really only partial. About the timings, it is noticeable that TFX is the faster in most of the tests, and when it is not the faster it is close enough to the faster (in fact, every time an algorithm was faster than TFX it was by a difference of a fraction of a second). The algorithms NLS, ALS and MINF also are very fast (their version without refinement). NLS was faster than TFX two times, ALS was faster three times, and MINF was faster four times. Apart from that, none of the other algorithms seemed to stand out for some test.

These results may lead the reader to the conclusion that Tensor Fox is the fastest implementation of all. This is partially true as we will see in the next experiment. This time, instead of making all algorithms to match Tensor Fox's accuracy, we run all with default options and show their box plots for accuracy and running time. With this we have a better understanding about the usual accuracy and usual running time of each algorithm. A box plot is used to summarize the main statistics of a random variable in a single image as 4.21.

In the next figures we have the box plots of all algorithms, 20 runs for each tensor.

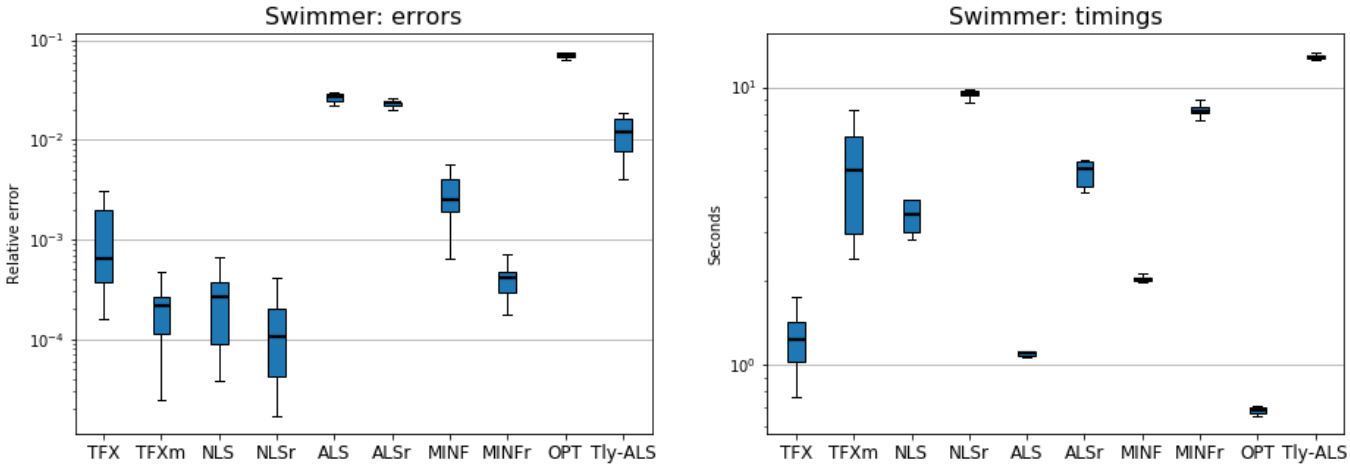

Figure 4.12: Box plots of the swimmer tensor.

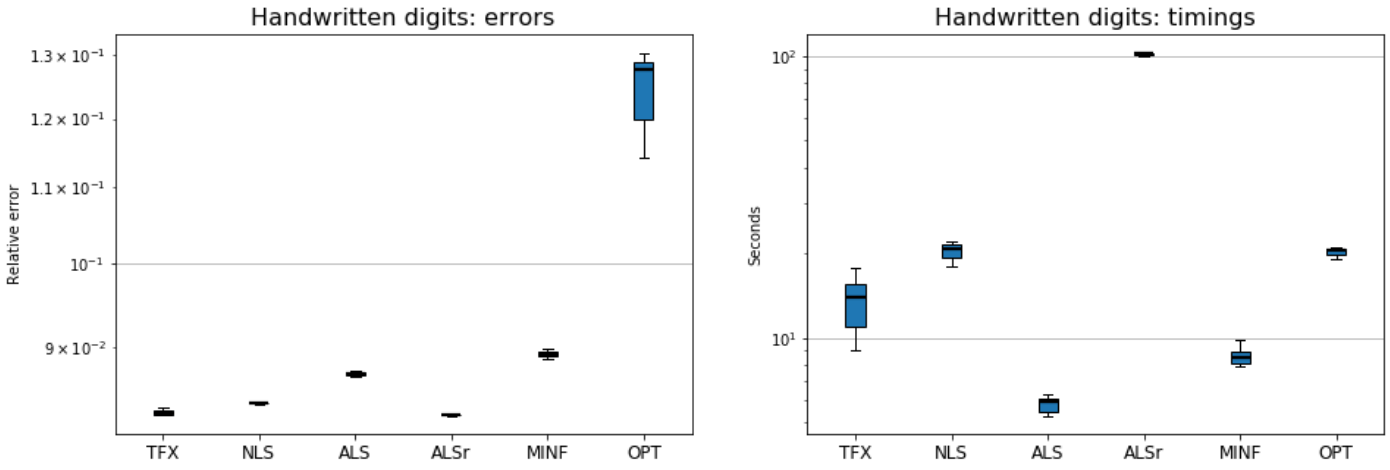

Figure 4.13: Box plots of the handwritten digits tensor.

An algorithm is omitted only when its running time is too big compared to the other or when the memory size to compute the CPD is too large (more than 7 GB in this context). Remember that in the previous benchmarks we manipulated the iteration of the algorithms in order to match Tensor Fox's accuracy with minimal time. Besides the Tensor Fox default, were we make the inverse procedure and manipulate Tensor Fox iterations[9] to match the best solution (minimal error) of all other algorithms. This modified algorithm is labelled as TFXm.

Tensor Fox default is competitive for all tensor except the double bottlenecks, where its accuracy is not optimal. In this case the modified versions are competitive, being as accurate and faster as the best algorithms. With these box plots we can see that no

[9]More precisely, we change the number of CG iterations and activate refinement sometimes, so the changes are similar to Tensorlab's.

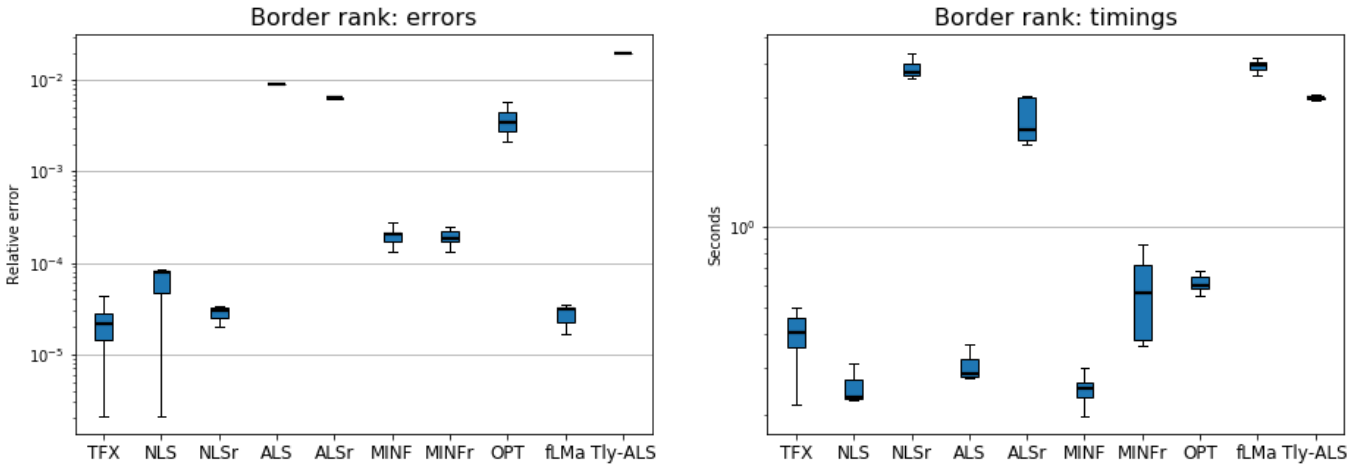

Figure 4.14: Box plots of the border rank tensor.

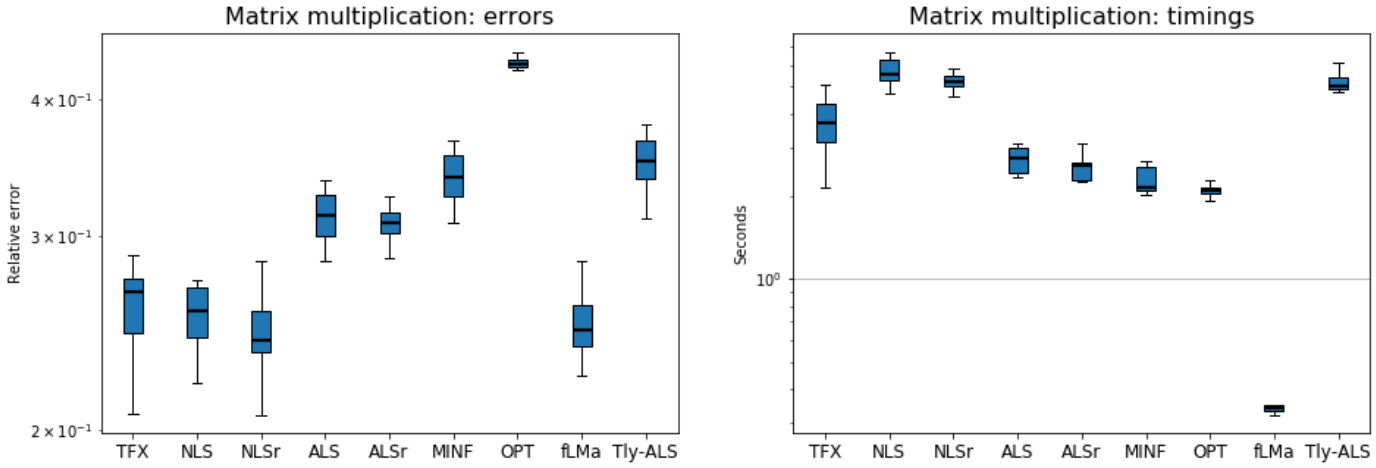

Figure 4.15: Box plots of the 5×5 matrix multiplication tensor.

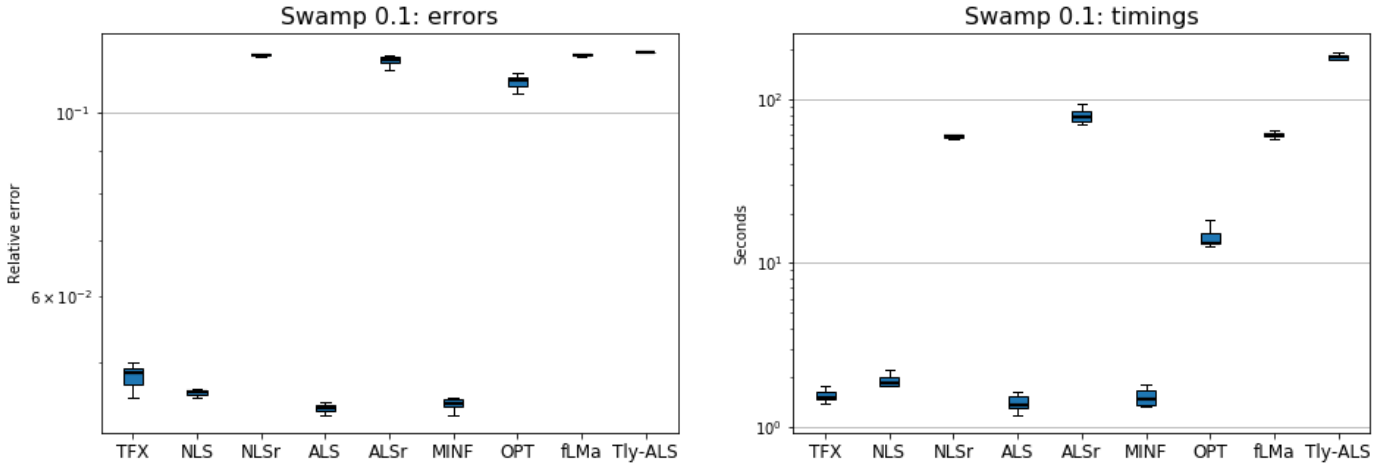

Figure 4.16: Box plots of the swamp tensor with $c = 0.1$.

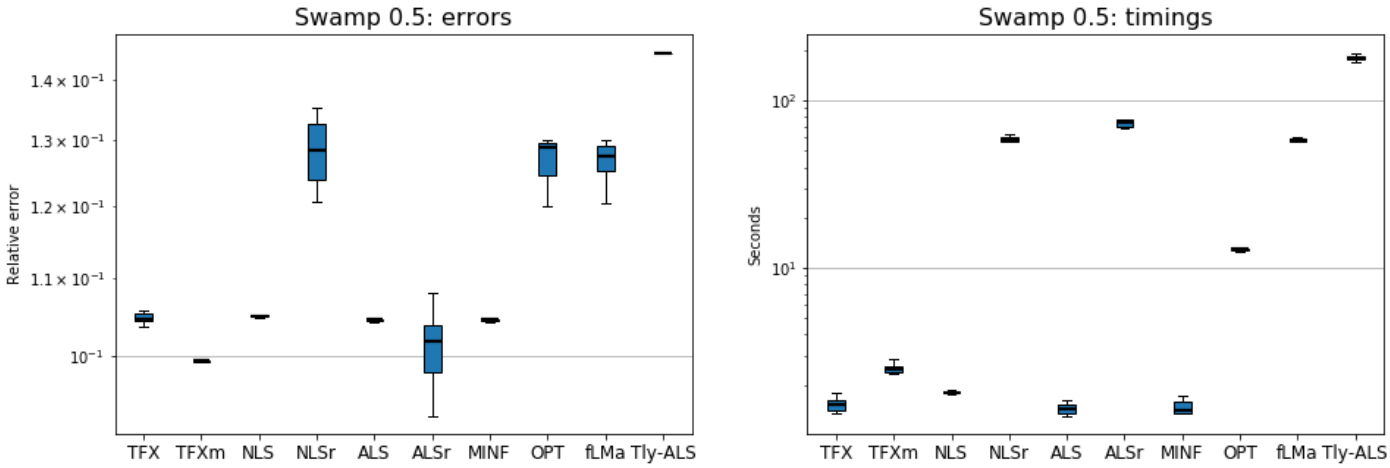

Figure 4.17: Box plots of the swamp tensor with $c = 0.5$.

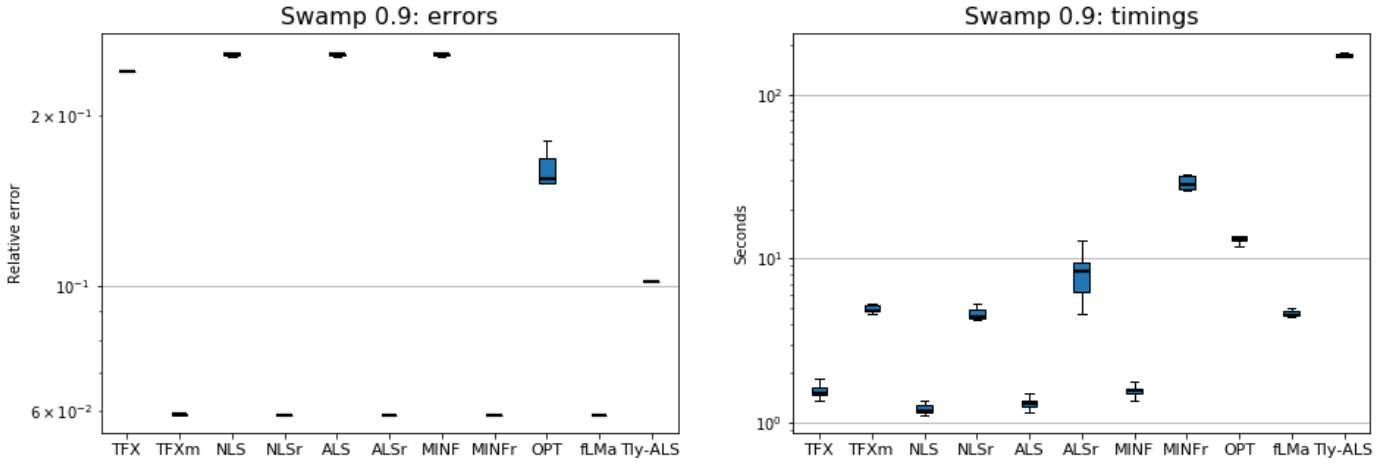

Figure 4.18: Box plots of the swamp tensor with $c = 0.9$.

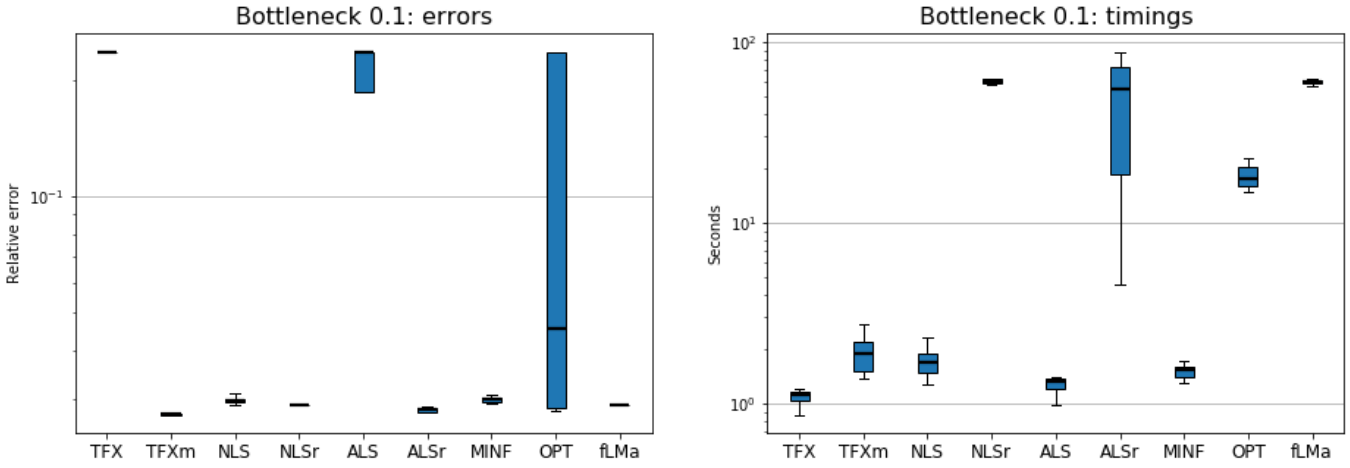

Figure 4.19: Box plots of the double bottleneck tensor with $c = 0.1$.

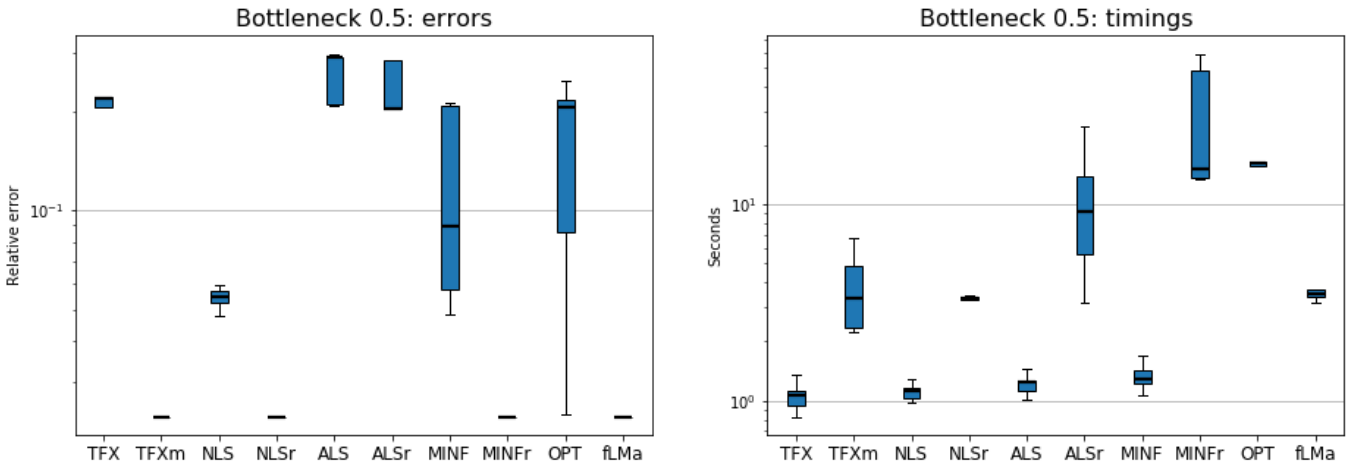

Figure 4.20: Box plots of the double bottleneck tensor with $c = 0.5$.

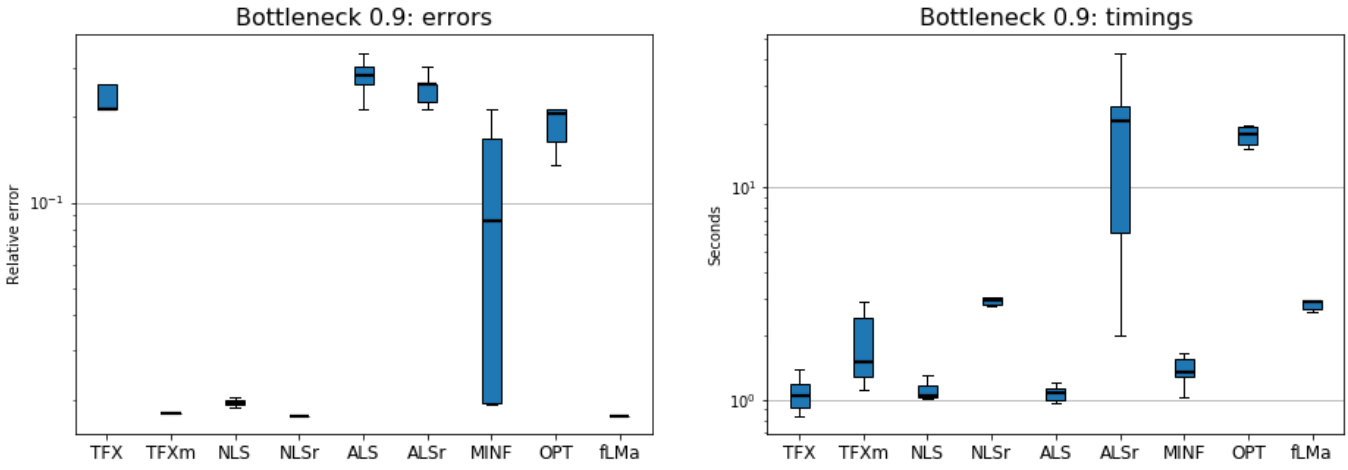

Figure 4.21: Box plots of the double bottleneck tensor with $c = 0.9$.

algorithm is optimal for all tensors, but the combinations TFX + TFXm or NLS + NLSr gives optimal solutions as fast as possible.

4.6 Tensor train and the CPD

In table 4.4 we gave a summary of all costs necessary to compute a rank-R CPD. The dominant cost comes from the dGN algorithm, which is dominated by a factor of $\mathcal{O}\left(LR\prod_{\ell=1}^{L} R_\ell\right)$ floats. With this cost we can see that the CPD computation suffers from the curse of dimensionality. More precisely, the cost increases exponentially with the tensor order. This limitation can be seen in every CPD implementation, and this is the main reason why this decomposition still is not widely used on large-scale problems. In the article [82] they propose a way around this limitation, based on the *tensor train decomposition* (TTD) [83]. As a result, we get an algorithm capable of handling higher order tensors with low computational cost. First we will briefly introduce the TTD, following [83]. After that we show how to connect the CPD with this new decomposition. This connection makes it possible to retrieve the CPD from the TTD.

4.6.1 Tensor train decomposition

Let $\mathcal{T} \in \mathbb{R}^{I_1 \times \dots I_L}$ be any order-L tensor. The main idea of the TTD is that we can approximate $\mathcal{T}$ by a tensor $\tilde{\mathcal{T}}$ such that the entries of this tensor are given by

$$\tilde{t}_{i_1 i_2 \dots i_L} = \mathcal{G}^{(1)}(i_1)\mathcal{G}^{(2)}(i_2)\dots\mathcal{G}^{(L)}(i_L) \tag{4.2}$$

where each $\mathcal{G}^{(\ell)}(i_\ell)$ is a $r_{\ell-1} \times r_\ell$ full rank matrix for $\ell = 2 \dots L$, with $r_1 = r_L = 1$.

Note that this definition generalizes the definition of rank one tensor. Instead of the coordinates of the tensor being given by a product of scalars, they are given by a product of matrices, where the first has one row and the last has one column so the result of the product still is a scalar. We can take one more step and consider each $\mathcal{G}^{(\ell)}(i_\ell)$ as the slice of a third order tensor with shape $r_{\ell-1} \times I_\ell \times r_\ell$. Denoting by $\mathcal{G}^{(\ell)}$ this tensor, we define $\mathcal{G}^{(\ell)}(i_\ell) = \mathcal{G}^{(\ell)}_{:\ i_\ell\ :}$ (the i_ℓ-th vertical slice of $\mathcal{G}^{(\ell)}$). In coordinates we have that

$$\tilde{t}_{i_1 i_2 \dots i_L} = \sum_{j_0, j_1, \dots, j_L} \mathcal{G}^{(1)}_{j_0 i_1 j_1}\ \mathcal{G}^{(2)}_{j_1 i_2 j_2}\ \dots\ \mathcal{G}^{(L)}_{j_{L-1} i_L j_L}.$$

Since $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(L)}$ are tensors of order $1 \times I_1 \times r_1$ and $r_{L-1} \times I_L \times 1$, respectively, they can be regarded as matrices (in particular, $j_0 = j_L = 1$). The ranks r_ℓ are called *TT-ranks* and the three dimensional tensors $\mathcal{G}^{(\ell)}$ are called the *cores* of the TTD. Note that the last index of each factor is the first of the next factor. This relation is illustrated in figure 4.22. The rectangles contain spatial indexes (the indexes $j_{\ell-1}, i_\ell, j_\ell$), and the

circles contain only the auxiliary nodes j_ℓ, the ones representing the link between the cores. Quoting the original author of the TTD (Ivan Oseledets):

"This picture looks like a train with carriages and links between them, and that justifies the name *tensor train decomposition*, or simply *TT-decomposition*."

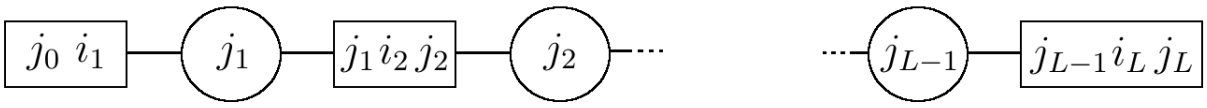

Figure 4.22: Tensor-train network representation.

The computation of the cores are made through successive SVDs and reshaping. The method used to reshape the matrices in the original article is the MATLAB reshape function, but it can be different as long as it is consistent with the other reshapes of the algorithm. In particular, note that in the first iteration of the TT-SVD loop we have $\mathbf{M} = \mathcal{T}_{(1)}$, the first unfolding of $\mathcal{T}$. Thus the reshape function must be consistent with the unfolding function. The algorithm is presented without truncation in this context but the interested reader can refer to the mentioned article to see the proofs and the more general algorithm. The functions `Unfolding` and `SVD` were already introduce, while `Reshape`$(\mathbf{M}, dims)$ reshapes a certain array $\mathbf{M}$ in order that the resulting array has the dimensions described by the tuple $dims$.

Algorithm 4.6.1 (TT-SVD)**.**
Input: $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_L}$

$r_0 \leftarrow 1$
$\boldsymbol{M} \leftarrow \mathcal{T}$
`for` $\ell = 1 \ldots L-1$
  $n_\ell \leftarrow \prod_{\ell' > \ell} I_{\ell'}$
  $\boldsymbol{M} \leftarrow$ `Reshape` $(\boldsymbol{M}, (r_{\ell-1} I_\ell, n_\ell))$
  $\boldsymbol{U}, \Sigma, \boldsymbol{V} \leftarrow$ `SVD`$(\boldsymbol{M})$
  $r_\ell \leftarrow rank(\boldsymbol{M})$
  $\mathcal{G}^{(\ell)} \leftarrow$ `Reshape`$(\boldsymbol{U}, (r_{\ell-1}, I_\ell, r_\ell))$
  $\boldsymbol{M} \leftarrow \Sigma \boldsymbol{V}^T$
$\mathcal{G}^{(L)} \leftarrow \boldsymbol{M}$

Output: $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(L)}$

4.6.2 CPD tensor train

Given two tensors $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_\ell \times \ldots \times I_L}$ and $\mathcal{U} \in \mathbb{R}^{J_1 \times \ldots \times J_m \times \ldots \times J_M}$ such that $I_\ell = J_m$, where the modes ℓ, m may be different, the $\times_\ell^m$-contraction between $\mathcal{T}$ and $\mathcal{U}$ is the tensor $\mathcal{T} \times_m^\ell \mathcal{U} \in \mathbb{R}^{I_1 \times \ldots \times I_{\ell-1} \times I_{\ell+1} \times \ldots \times I_L \times J_1 \times \ldots \times J_{m-1} \times J_{m+1} \times \ldots \times J_M}$ given by

$$(\mathcal{T} \times_\ell^m \mathcal{U})_{i_1 \ldots i_{\ell-1} i_{\ell+1} \ldots i_L \; j_1 \ldots j_{m-1} j_{m+1} \ldots j_M} = \sum_{k=1}^{I_\ell} t_{i_1 \ldots i_{\ell-1} \; k \; i_{\ell+1} \ldots i_L} \cdot u_{j_1 \ldots j_{M-1} \; k \; j_{M+1} \ldots j_M}.$$

Note that with the notion of contraction one can write the TTD in 4.2 more compactly as

$$\mathcal{T} = \mathcal{G}^{(1)} \times_2^1 \mathcal{G}^{(2)} \times_3^1 \ldots \times_{L-1}^1 \mathcal{G}^{(L-1)} \times_L^1 \mathcal{G}^{(L)},$$

where the operations are made from left to right but we suppressed the parenthesis to maintain a clean notation. Now we are able to prove our first theorem connecting the CPD and the TTD. Since the TTD assumes that all cores are of full rank, from this point we need to assume that all factor matrices of any CPD also are of full rank. This assumption is implicit in each of the next results. As we did in chapter 3, the objective of this part is to give the reader not only an understanding of the theoretical aspects, but also the computational aspects related of our problems, hence all proofs are given in coordinates.

Theorem 4.6.2 (I. Oseledets, 2010)**.** *Let $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_\ell \times \ldots \times I_L}$ be a rank-R tensor with CPD given by*

$$\mathcal{T} = \left(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L)} \right) \cdot \mathcal{I}_{R \times \ldots \times R}.$$

Then $\mathcal{T}$ admits a TTD of the form $\mathcal{T} = \mathcal{G}^{(1)} \times_2^1 \mathcal{G}^{(2)} \times_3^1 \ldots \times_{L-1}^1 \mathcal{G}^{(L-1)} \times_L^1 \mathcal{G}^{(L)}$, where

$$\begin{aligned}
\mathcal{G}^{(1)} &= \boldsymbol{W}^{(1)}, \\
\mathcal{G}^{(\ell)} &= \left(\boldsymbol{I}_R, \boldsymbol{W}^{(\ell)}, \boldsymbol{I}_R \right) \cdot \mathcal{I}_{R \times R \times R}, \quad \ell = 2 \ldots L-1, \\
\mathcal{G}^{(L)} &= \left(\boldsymbol{W}^{(L)} \right)^T.
\end{aligned}$$

Proof: Define the tensor $\mathcal{U} = \mathcal{G}^{(1)} \times_2^1 \mathcal{G}^{(2)} \times_3^1 \ldots \times_{L-1}^1 \mathcal{G}^{(L-1)} \times_L^1 \mathcal{G}^{(L)}$ with cores as showed above. We will prove that $\mathcal{T} = \mathcal{U}$. First note that $\mathcal{G}_{ij}^{(1)} = \mathbf{w}_{ij}^{(1)}$ and $\mathcal{G}_{ij}^{(L)} = \mathbf{w}_{ji}^{(L)}$. For the other cores we have that

$$\mathcal{G}^{(\ell)} = \left(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)} \right) \cdot \mathcal{I}_{R \times \ldots \times R} = \sum_{r=1}^{R} \mathbf{e}_r \otimes \mathbf{w}_r^{(\ell)} \otimes \mathbf{e}_r =$$

$$= \sum_{r=1}^{R} \mathbf{e}_r \otimes \left(\sum_{i_\ell=1}^{I_\ell} w_{i_\ell r}^{(\ell)} \mathbf{e}_{i_\ell} \right) \otimes \mathbf{e}_r = \sum_{r=1}^{R} \sum_{i_\ell=1}^{I_\ell} w_{i_\ell r}^{(\ell)} \ \mathbf{e}_r \otimes \mathbf{e}_{i_\ell} \otimes \mathbf{e}_r.$$

In particular, this expansion of $\mathcal{G}^{(\ell)}$ shows that this tensor can be written as

$$\mathcal{G}^{(\ell)} = \left\{ \begin{bmatrix} w_{11}^{(\ell)} & \dots & w_{I_\ell 1}^{(\ell)} \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \begin{bmatrix} 0 & \dots & 0 \\ w_{12}^{(\ell)} & \dots & w_{I_\ell 2}^{(\ell)} \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ w_{1r}^{(\ell)} & \dots & w_{I_\ell r}^{(\ell)} \end{bmatrix} \right\},$$

where each matrix is a frontal slice. Now we can just compute the coordinates of $\mathcal{U}$ and realize that they are equal to the $\mathcal{T}$ coordinates.

$$u_{i_1 i_2 \dots i_L} = \sum_{j_0, j_1, \dots, j_L} \mathcal{G}^{(1)}_{j_0 i_1 j_1} \ \mathcal{G}^{(2)}_{j_1 i_2 j_2} \ \dots \ \mathcal{G}^{(L-1)}_{j_{L-2} i_{L-1} j_{L-1}} \ \mathcal{G}^{(L)}_{j_{L-1} i_L j_L} =$$

$$= \sum_{j_1, \dots, j_L} \left(\mathbf{W}^{(1)} \right)_{i_1 j_1} \left(\sum_{r=1}^{R} \left(\mathbf{e}_r \otimes \mathbf{W}^{(2)} \otimes \mathbf{e}_r \right)_{j_1 i_2 j_2} \right) \dots \left(\sum_{r=1}^{R} \left(\mathbf{e}_r \otimes \mathbf{W}^{(L-1)} \otimes \mathbf{e}_r \right)_{j_{L-2} i_{L-1} j_{L-1}} \right) \left(\mathbf{w}_r^{(L)} \right)_{i_L j_{L-1}}.$$

Note that only the terms with $j_1 = j_2 = \dots = j_{L_1} = r$, for $r = 1 \dots R$, are not null. Therefore the expression above reduces to

$$\sum_{r=1}^{R} w_{i_1 r}^{(1)} \ w_{i_2 r}^{(2)} \ \dots \ w_{i_{L-1} r}^{(L-1)} \ w_{i_L r}^{(L)},$$

from which we can conclude the equality, as desired. □

The TTD based on the CPD given in this theorem will be called *CPD-train*. In particular, note that all TT-ranks of this decomposition are equal to R, the rank of the tensor. Now suppose we don't have the factors $\mathbf{W}^{(\ell)}$ but the CPD-train holds. If we apply the TT-SVD algorithm to obtain the cores of the TTD we can expect to retrieve the factors of the CPD. It should be noted that the cores obtained are note necessarily in the form $\mathcal{G}^{(\ell)} = \left(\mathbf{I}_R, \mathbf{W}^{(\ell)}, \mathbf{I}_R \right) \cdot \mathcal{I}_{R \times R \times R}$ since their representation are not unique (we can post multiply one core by an invertible matrix and pre multiply the next core for its inverse, therefore changing the cores but maintaining the TTD).

The next lemma says that if we have the CPDs of two tensors $\mathcal{T}$,$\mathcal{U}$, and the last factor matrix of $\mathcal{T}$ is the transpose of the inverse of the first factor matrix of $\mathcal{U}$, then these factors are cancelled when contracted. This lemma is a preparation to prove next theorem, whose proof was only hinted in [82]. Here we give a detailed proof with a computational flavor. To prove that lemma we rely on the fact that the $\times_L^1$-contraction of rank-1 terms (definition B.1.3) of the form $\mathbf{a}^{(1)} \otimes \dots \otimes \mathbf{a}^{(L)}, \mathbf{b}^{(1)} \otimes \dots \otimes \mathbf{b}^{(M)}$ (we are

assuming that $\mathbf{a}^{(L)}$ and $\mathbf{b}^{(1)}$ have the same shape) is a tensor $\mathcal{T}$ given by

$$\mathcal{T} = \langle \mathbf{a}^{(L)}, \mathbf{b}^{(1)} \rangle \ \mathbf{a}^{(1)} \otimes \ldots \otimes \mathbf{a}^{(L-1)} \otimes \mathbf{b}^{(2)} \otimes \ldots \otimes \mathbf{b}^{(M)}.$$

Lemma 4.6.3. *Let two tensors $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_\ell \times \ldots \times I_L}$ and $\mathcal{U} \in \mathbb{R}^{J_1 \times \ldots \times J_m \times \ldots \times J_M}$ such that*

$$\mathcal{T} = \left(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(L)} \right) \cdot \mathcal{I}_{R \times \ldots \times R}$$

for matrices $\boldsymbol{A}^{(\ell)} \in \mathbb{R}^{I_\ell \times R}$, and

$$\mathcal{U} = \left(\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(M)} \right) \cdot \mathcal{I}_{R \times \ldots \times R}$$

for matrices $\boldsymbol{B}^{(m)} \in \mathbb{R}^{I_m \times R}$. If $\boldsymbol{A}^{(L)} = \left(\boldsymbol{B}^{(1)} \right)^{-T}$ (which means $I_L = J_1 = R$), then

$$\mathcal{T} \times_L^1 \mathcal{U} = \left(\left(\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(L-1)}, \boldsymbol{I}_R \right) \cdot \mathcal{I}_{R \times \ldots \times R} \right) \times_L^1 \left(\left(\boldsymbol{I}_R, \boldsymbol{B}^{(2)}, \ldots, \boldsymbol{B}^{(M)} \right) \cdot \mathcal{I}_{R \times \ldots \times R} \right).$$

Proof: First we expand the expression for the contraction, obtaining

$$\mathcal{T} \times_L^1 \mathcal{U} = \left(\sum_{r=1}^{R} \mathbf{a}_r^{(1)} \otimes \ldots \otimes \mathbf{a}_r^{(L)} \right) \times_L^1 \left(\sum_{r'=1}^{R} \mathbf{b}_{r'}^{(1)} \otimes \ldots \otimes \mathbf{b}_{r'}^{(M)} \right) =$$

$$= \sum_{r,r'=1}^{R} \left(\mathbf{a}_r^{(1)} \otimes \ldots \otimes \mathbf{a}_r^{(L)} \right) \times_L^1 \left(\mathbf{b}_{r'}^{(1)} \otimes \ldots \otimes \mathbf{b}_{r'}^{(M)} \right) =$$

$$= \sum_{r,r'=1}^{R} \langle \mathbf{a}^{(L)}, \mathbf{b}^{(1)} \rangle \ \mathbf{a}_r^{(1)} \otimes \ldots \otimes \mathbf{a}_r^{(L-1)} \otimes \mathbf{b}_{r'}^{(2)} \otimes \ldots \otimes \mathbf{b}_{r'}^{(M)}.$$

For the moment, instead of writing $\mathbf{a}_r^{(L)}$ for the r-th column of $\mathbf{A}^{(L)}$, it will be convenient to use the Matlab convention and write $\mathbf{a}_{:r}^{(L)}$. Furthermore, since $\mathbf{a}_{:r}^{(L)}$ is the r-th row of $\left(\mathbf{B}^{(L)} \right)^{-1}$ transposed, we have that

$$\langle \mathbf{a}_{:r}^{(L)}, \mathbf{b}_{:r'}^{(1)} \rangle = \left(\mathbf{a}_{:r}^{(L)} \right)^T \mathbf{b}_{:r'}^{(1)} =$$

$$= \left(\mathbf{b}_{r:}^{(1)} \right)^{-1} \mathbf{b}_{:r'}^{(1)} = \delta_{rr'},$$

where $\mathbf{b}_{r:}^{(1)}$ denotes the r-th row of $\mathbf{B}^{(1)}$ and $\delta_{rr'}$ is the Kronecker delta of r and r'. Thus the summation for $\mathcal{T} \times_L^1 \mathcal{U}$ reduces to

$$\sum_{r=1}^{R} \langle \mathbf{a}^{(L)}, \mathbf{b}^{(1)} \rangle \ \mathbf{a}_r^{(1)} \otimes \ldots \otimes \mathbf{a}_r^{(L-1)} \otimes \mathbf{b}_r^{(2)} \otimes \ldots \otimes \mathbf{b}_r^{(M)}.$$

It is not hard to see that

$$\Big(\Big(\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(L-1)},\mathbf{I}_R\Big)\cdot\mathcal{I}_{R\times\ldots\times R}\Big)\times_L^1\Big(\Big(\mathbf{I}_R,\mathbf{B}^{(2)},\ldots,\mathbf{B}^{(M)}\Big)\cdot\mathcal{I}_{R\times\ldots\times R}\Big)$$

leads to the same expression, therefore they are the same tensor. □

Theorem 4.6.4 (Y. Zniyed, R. Boyer, Andre L.F. de Almeida, G. Favier, 2018)**.** *Let $\mathcal{T}\in\mathbb{R}^{I_1\times\ldots\times I_L}$ be a rank-R tensor with TTD given by*

$$\mathcal{T}=\mathcal{G}^{(1)}\times_2^1\mathcal{G}^{(2)}\times_3^1\ldots\times_{L-1}^1\mathcal{G}^{(L-1)}\times_L^1\mathcal{G}^{(L)}$$

such that

$$\begin{aligned}
\mathcal{G}^{(1)}&=\boldsymbol{W}^{(1)}\big(\boldsymbol{M}^{(1)}\big)^{-T}\\
\mathcal{G}^{(2)}&=\Big(\boldsymbol{M}^{(1)},\boldsymbol{W}^{(2)},\boldsymbol{M}^{(2)}\Big)\cdot\mathcal{I}_{R\times R\times R}\\
\mathcal{G}^{(3)}&=\Big(\big(\boldsymbol{M}^{(2)}\big)^{-T},\boldsymbol{W}^{(3)},\boldsymbol{M}^{(3)}\Big)\cdot\mathcal{I}_{R\times R\times R}\\
&\vdots\\
\mathcal{G}^{(L-1)}&=\Big(\big(\boldsymbol{M}^{(L-2)}\big)^{-T},\boldsymbol{W}^{(L-1)},\boldsymbol{M}^{(L-1)}\Big)\cdot\mathcal{I}_{R\times R\times R}\\
\mathcal{G}^{(L)}&=\big(\boldsymbol{M}^{(L-1)}\big)^{-T}\big(\boldsymbol{W}^{(L)}\big)^{T}.
\end{aligned}$$

Then $\mathcal{T}=\big(\boldsymbol{W}^{(1)},\ldots,\boldsymbol{W}^{(L)}\big)\cdot\mathcal{I}_{R\times\ldots\times R}$ is a CPD for $\mathcal{T}$. All matrices $\boldsymbol{M}^{(\ell)}$ are square $R\times R$ and invertible.

Proof: The prove is based on successive applications of the last lemma. First note that

$$\mathcal{G}^{(1)}\times_2^1\mathcal{G}^{(2)}=\mathbf{W}^{(1)}\big(\mathbf{M}^{(1)}\big)^{-T}\times_2^1\Big(\Big(\mathbf{M}^{(1)},\mathbf{W}^{(2)},\mathbf{M}^{(2)}\Big)\cdot\mathcal{I}_{R\times R\times R}\Big)=$$

$$=\mathbf{W}^{(1)}\times_2^1\Big(\Big(\mathbf{I}_R,\mathbf{W}^{(2)},\mathbf{M}^{(2)}\Big)\cdot\mathcal{I}_{R\times R\times R}\Big).$$

Going to the next term gives

$$\mathcal{G}^{(1)}\times_2^1\mathcal{G}^{(2)}\times_3^1\mathcal{G}^{(3)}=$$

$$=\mathbf{W}^{(1)}\times_2^1\Big(\Big(\mathbf{I}_R,\mathbf{W}^{(2)},\mathbf{M}^{(2)}\Big)\cdot\mathcal{I}_{R\times R\times R}\Big)\times_3^1\Big(\Big(\big(\mathbf{M}^{(2)}\big)^{-T},\mathbf{W}^{(3)},\mathbf{M}^{(3)}\Big)\cdot\mathcal{I}_{R\times R\times R}\Big)=$$

$$=\mathbf{W}^{(1)}\times_2^1\Big(\Big(\mathbf{I}_R,\mathbf{W}^{(2)},\mathbf{I}_R\Big)\cdot\mathcal{I}_{R\times R\times R}\Big)\times_3^1\Big(\Big(\big(\mathbf{I}_R,\mathbf{W}^{(3)},\mathbf{M}^{(3)}\Big)\cdot\mathcal{I}_{R\times R\times R}\Big).$$

This pattern keeps going until we have

$$\mathcal{G}^{(1)}\times_2^1\mathcal{G}^{(2)}\times_3^1\ldots\times_{L-1}^1\mathcal{G}^{(L-1)}\times_L^1\mathcal{G}^{(L)}=$$

$$=\mathbf{W}^{(1)}\times_2^1\Big(\Big(\mathbf{I}_R,\mathbf{W}^{(2)},\mathbf{I}_R\Big)\cdot\mathcal{I}_{R\times R\times R}\Big)\times_3^1\ldots\times_{L-1}^1\Big(\Big(\mathbf{I}_R,\mathbf{W}^{(L-1)},\mathbf{M}^{(L-1)}\Big)\cdot\mathcal{I}_{R\times R\times R}\Big)\times_L^1\Big(\big(\mathbf{M}^{(L-1)}\big)^{-T}\big(\mathbf{W}^{(L)}\big)^{T}\Big)=$$

$$= \mathbf{W}^{(1)} \times_2^1 \Big(\Big(\mathbf{I}_R, \mathbf{W}^{(2)}, \mathbf{I}_R \Big) \cdot \mathcal{I}_{R\times R\times R} \Big) \times_3^1 \ldots \times_{L-1}^1 \Big(\Big(\mathbf{I}_R, \mathbf{W}^{(L-1)}, \mathbf{I}_R \Big) \cdot \mathcal{I}_{R\times R\times R} \Big) \times_L^1 \big(\mathbf{W}^{(L)} \big)^T.$$

This is the decomposition showed in 4.6.2, which proves the theorem. □

We stated the theorem in a way it already gives insights about the computations to be made. More precisely, we should first compute a rank-R CPD for the third order tensor $\mathcal{G}^{(2)}$. Since $\mathcal{G}^{(1)}$ is already know, we can obtain the first factor through the equality $\mathbf{W}^{(1)} = \mathcal{G}^{(1)} \big(\mathbf{M}^{(1)} \big)^T$. To compute $\mathbf{W}^{(3)}$ we must compute a rank-R CPD for $\mathcal{G}^{(3)}$ fixing the first factor to $\big(\mathbf{M}^{(2)} \big)^{-T}$. We can keep going sequentially with this procedure, using the third factor of the previous CPD to construct the first factor of the next CPD. The last factor $\mathbf{W}^{(L)}$ is then computed through the equality $\mathbf{W}^{(L)} = \big(\mathcal{G}^{(L)} \big)^T \mathbf{M}^{(L-1)}$ since $\mathbf{M}^{(L-1)}$ will be already known at this point. This gives rise to the following algorithm.

Algorithm 4.6.5 (CPD-TTD).
Input: $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_L}, R$

$\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(L)} \leftarrow$ `TT-SVD`$(\mathcal{T})$
$\boldsymbol{M}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{M}^{(2)} \leftarrow$ `CPD`$(\mathcal{G}^{(2)}, R)$
`for` $\ell = 3 \ldots L-1$
  $\boldsymbol{W}^{(\ell)}, \boldsymbol{M}^{(\ell)} \leftarrow$ `Bi-CPD` $\Big(\big(\boldsymbol{M}^{(\ell-1)} \big)^{-T}, \mathcal{G}^{(\ell)}, R \Big)$
$\boldsymbol{W}^{(1)} \leftarrow \mathcal{G}^{(1)} \big(\boldsymbol{M}^{(1)} \big)^T$
$\boldsymbol{W}^{(L)} \leftarrow \big(\mathcal{G}^{(L)} \big)^T \boldsymbol{M}^{(L-1)}$

Output: $\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L)}$

`CPD`$(\mathcal{T}, R)$ is any algorithm to compute a rank-R CPD to $\mathcal{T}$ and `Bi-CPD`$(\mathbf{M}, \mathcal{T}, R)$ does the same job but with the first factor, $\mathbf{M}$, fixed. Since we are talking about third order tensors, the latter algorithm just computes two factors, hence the name "Bi-CPD". Notice we can't use the conjugate descent gradient algorithm to compute each iteration of the dGN for the `Bi-CPD` function. Suppose we are given the triple $\Big(\big(\mathbf{M}^{(\ell-1)} \big)^{-T}, \mathbf{W}^{(\ell)}, \mathbf{M}^{(\ell)} \Big)$ respective to the CPD of $\mathcal{G}^{(\ell)}$ and now we have to fix $\big(\mathbf{M}^{(\ell)} \big)^{-T}$ for the next CPD. At each iteration of the dGN we will be solving the system

$$\big(\mathbf{A}_{:,R^2+1:}^T \mathbf{A}_{:,R^2+1:} + \mu \mathbf{D}_{:,R^2+1:} \big) \mathbf{x}_{R^2+1:} = \mathbf{A}^T \mathbf{b} - \big(\mathbf{A}_{:,:R^2}^T \mathbf{A}_{:,:R^2} + \mu \mathbf{D}_{:,:R^2} \big) \mathbf{x}_{:R^2},$$

where $\mathbf{x}_{:R^2} = vec \Big(\big(\mathbf{M}^{(\ell)} \big)^{-T} \Big)$ is the fixed part. Notice we have to solve a system of $R^2 + RI_{\ell+1} + R^2$ equations and $RI_\ell + R^2$ variables, which can be solved by more than one algorithm.

The most relevant aspect of this algorithm is its cost, which is a great improvement compared to the costs showed in table 4.4. The higher costs comes from the TT-SVD algorithm, which amounts to computing $L-1$ SVDs of shapes $I_1 \times \prod_{\ell=2}^{L}$, $RI_2 \times \prod_{\ell=3}^{L} I_\ell$, ...,

	TT-SVD	**Compression**
First SVD	$2I_1\prod_{\ell=2}^{L} I_\ell + 2I_1^3$	$I_1\prod_{\ell=1}^{L} I_\ell + I_1^3$
Second SVD	$2I_1^2I_2\prod_{\ell=3}^{L} I_\ell + 2I_1^3I_2^3$	$I_2\prod_{\ell=1}^{L} I_\ell + I_2^3$
$\vdots$	$\vdots$	$\vdots$
$(L-1)$ -th SVD	$2I_1^2I_{L-1}I_L + 2I_1^3I_{L-1}^3$	$I_{L-1}\prod_{\ell=1}^{L} I_\ell + I_{L-1}^3$
L -th SVD	0	$I_L\prod_{\ell=1}^{L} I_\ell + I_L^3$

Table 4.18: Costs of TT-SVD vs. Compression

$RI_{L-1} \times I_L$, respectively. This has a total cost of

$$\mathcal{O}\left(2I_1^2\prod_{\ell=2}^{L} I_\ell + 2I_1^3\right) + \mathcal{O}\left(2R^2I_2^2\prod_{\ell=3}^{L} I_\ell + 2R^3I_2^3\right) + \ldots + \mathcal{O}\left(2R^2I_{L-1}^2I_L + 2R^3I_{L-1}^3\right) =$$

$$= \mathcal{O}\left(2\left(I_1^2\prod_{\ell=2}^{L} I_\ell + I_1^3 + R^2\left(\sum_{\ell=2}^{L-1} I_\ell^2\prod_{\ell'=\ell+1}^{L} I_{\ell'} + RI_\ell^3\right)\right)\right) =$$

$$= \mathcal{O}\left(2\left(I_1\prod_{\ell=1}^{L} I_\ell + I_1^3 + R^2\left(\sum_{\ell=2}^{L-1} I_\ell\prod_{\ell'=\ell}^{L} I_{\ell'} + RI_\ell^3\right)\right)\right) \text{ flops.}$$

Since R is used as the rank for all cores of the TT-SVD, we must have $R \leq rank(T_{(\ell)}) \leq I_\ell$ for all ℓ. We can put side by side the costs of computing these SVDs and the MLSVD computations of the previous algorithm, showed in table 4.18. At this point it should be clear the difference between both approaches. While the TT-SVD is always cutting a dimensions for each computed SVD, the MLSVD doesn't do this, it computes SVDs of unfolding with the size of the whole tensor L times. The MLSVD suffers from the curse of dimensionality, whereas the TT-SVD doesn't (actually it still suffers, but very little and nothing compared to the MLSVD cost).

The other costs consists of the computations of third order CPDs of shapes $R \times I_\ell \times R$, for $\ell = 2 \ldots L-1$. With this we can see that the CPD-TTD approach is much faster than our previous algorithm, by orders of magnitude. To reinforce this claim we show some computational experiments with random tensors. We generate random tensors of shape $n \times n \times \ldots \times n$ and rank $R = 5$, where the entries of each factor matrix are drawn from the normal distribution (mean 0 and variance 1). First we consider fourth order tensors with shape $n \times n \times n \times n$, for $n = 10, 20, 30, 40, 50, 60, 70, 80$. Since the Tensorlab's NLS performed very well in the previous tests, we use only this one for Tensorlab and

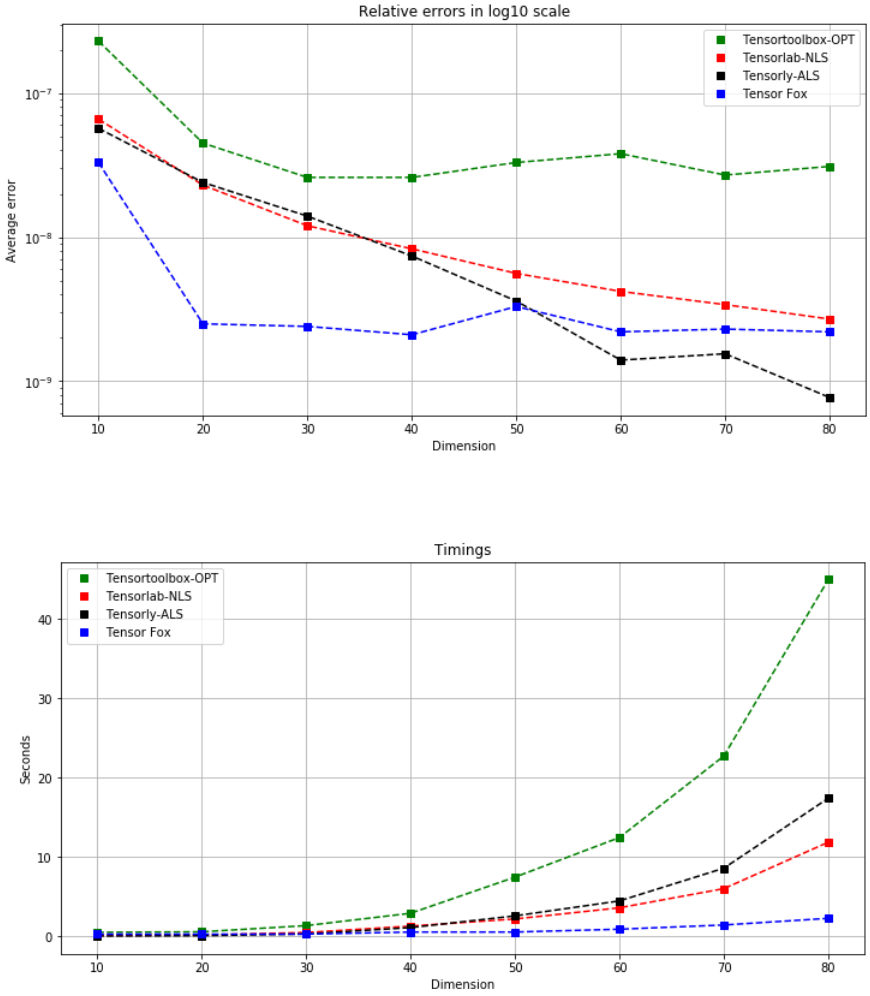

Figure 4.23: Rank-5 CPD errors and timings of tensors with shape $n \times n \times n \times n$, for $n = 10, 20, \ldots, 70, 80$.

start with this algorithm, making 20 computations for each dimension n and averaging the errors and time.[10] After that we run the other algorithms adjusting their tolerance in order to match the NLS results.

In all tests we tried to choose the parameters in order to speed up the algorithms without losing accuracy. For example, we noticed that it was unnecessary to use compression, detection of structure and refinement for the NLS algorithm. These routines are very expensive and didn't bring much extra precision, so they were disabled in order to make the NLS computations faster. Similarly we used the initialization 'svd' for Tensorly because it proved to be faster than 'random', and we used the algorithm 'lbfgs' for Tensor Toolbox OPT. Finally, for Tensor Fox we just decreased its tolerance in order to match the precision given by the NLS algorithm. The results are showed in figure 4.23.

Next, we make the same procedure but this time we fixed n to $n = 10$ and increased

[10] In order to achieve the least possible difference between the errors, we accepted to discard 1 to 10 tests with bad approximations.

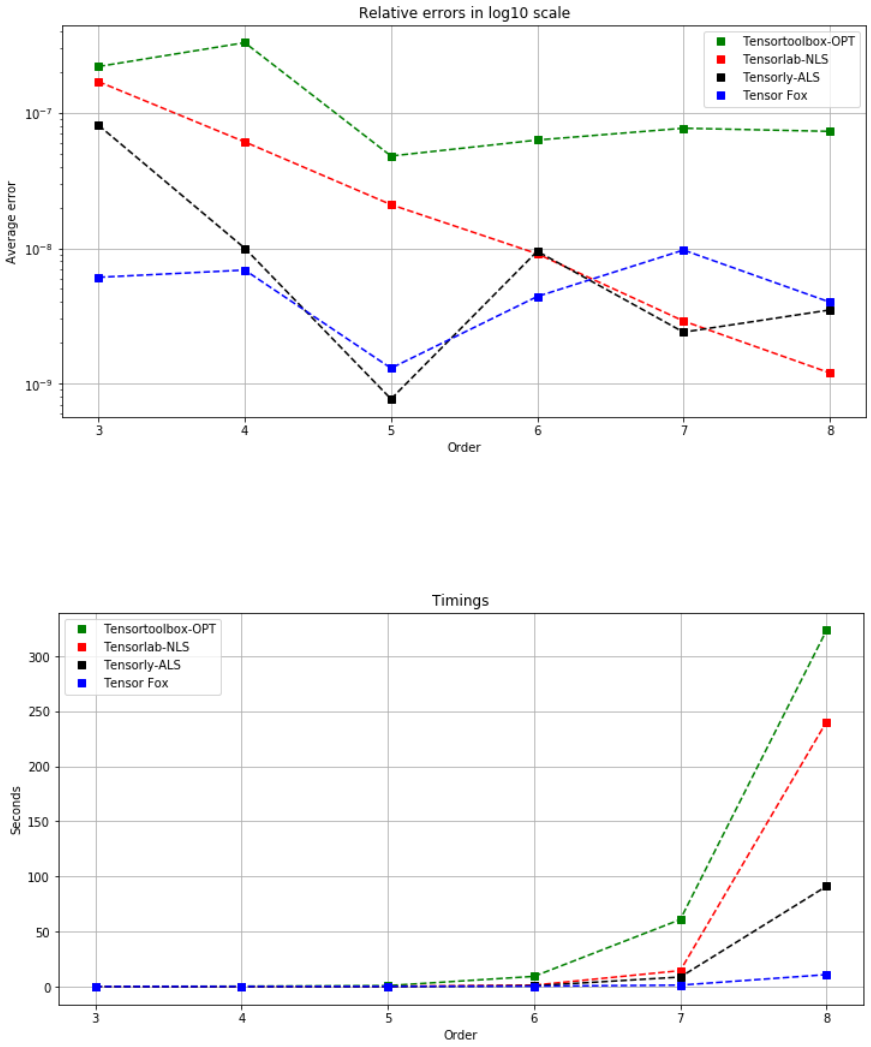

Figure 4.24: Rank-5 CPD errors and timings of tensors with shape $10 \times 10 \times \ldots \times 10$ (L times), for $L = 3, 4, \ldots, 8$.

the order, from order 3 to 8. These last tests shows an important aspect of the CPD-TTD: it avoids the curse of dimensionality, whereas the other algorithms still suffers from that. We consider random rank-5 tensors of shape $10 \times 10 \times 10$, them $10 \times 10 \times 10 \times 10$, up to tensors of order 8, i.e., with shape $\underbrace{10 \times 10 \times \ldots \times 10}_{8 \text{ times}}$, with the same distribution as before.[11]

One limitation of the CPD-TTD is the rank itself. When constructing the cores of the tensor train we noted that we must have $R \leq \min_\ell I_\ell$, so this approach doesn't work for higher rank tensors. Since the CPD-TTD already is very fast compared to the other algorithms, we can afford to increase the cost a little when dealing with higher ranks. In the case $R > I_\ell$ for some ℓ, we just increase this dimension to be of size R. The new values

[11] For anyone interested in reproducing these tests, the routine to generate these tensors can be found in https://github.com/felipebottega/Tensor-Fox/blob/master/tests/gen_rand_tensor.py.

added to the tensor are random noise close to zero. For example, if we have a tensor $\mathcal{T}$ of shape $4\times5\times6\times7$ and $R = 6$, we increase the dimensions of $\mathcal{T}$ so we have a new tensor $\mathcal{T}'$ of shape $6\times6\times6\times7$ such that $\mathcal{T}'_{ijkl} = \mathcal{T}_{ijkl}$ for all $i = 1\ldots4, j = 1\ldots5, k = 1\ldots6, l = 1\ldots7$. The extra entries of $\mathcal{T}'$ won't affect too much the precision since they are very small, and since the CPD-TTD algorithm is already very fast, the CPD computation of this bigger tensor still is much faster than any other algorithm. We should point out that these new entries should be added after the MLSVD is computed, otherwise we are just introducing more complexity for nothing. After the CPD is computed we can truncate the CPD to its original dimensions. This approach works very well as long as R is not too large or bigger than all dimensions (in practice we observed that we should have $R \leq I_\ell$ for at least one ℓ).

4.7 Tensor Fox is not monotonic

Usually one expects that any minimization program produces a sequence of steps such that the corresponding sequence of errors decreases monotonically. In the "warming up" example, figure 4.6 shows a sequence of errors which is not monotonic, sometimes the error increases, then it decreases back. This behavior was hinted at 4.1.4 where we remarked that the way Tensor Fox handles the maximum number of CG iterations can introduce unusual steps sometimes.[12] Now we make this statement clear.

By allowing the CG method to perform a random number of steps, we observed that local minima are consistently avoided. Usually a poor step is followed by a very good step. The idea is that these steps go a little away from the local minimum, just enough so that the next iteration can search for better directions. If no new direction is found, the program goes back to the previous local minimum and probably will stay there. We say "probably" because there is always a little chance that an unusual step is taken, but most of the time the program are making usual steps. At the end of the day we conclude that this stochastic factor is successful when dealing with the high nonlinearity of the problem. This phenomenon is illustrated in figure 4.25.

[12] These steps happen when `cg_maxiter` is large, since the CG algorithm become unstable.

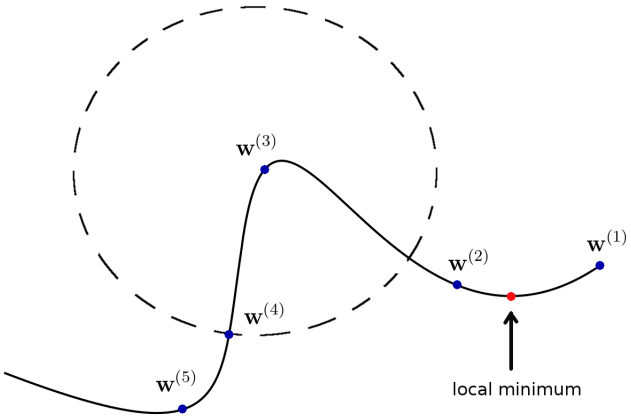

Figure 4.25: After 2 iterations, the third one causes an increase in the error. However this allows the program to search in different regions (the circled region), which can lead the next iteration to be much better. In the worse case the next iteration is much likely to go back to the already located minimum.

In 4.1.3.5 we discussed the gain ratio g and what it represents. Basically, large values of g means that the error is being reduced substantially, $g = 0$ means that the error didn't change, and $g < 0$ means that the error increased. After the observations made here it is now clear that the unusual steps are responsible for making g to be negative, and that the strategy to produce `cg_maxiter` (described in 4.1.3.4) is what causes this behavior.

Let's take a concrete example to observe this phenomenon happening in more details. In figure 4.26 we show the evolution of the error and gain ratio corresponding to a CPD computation of the swimmer tensor. We note that $g < 0$ precisely when the error increases, as expected. The interesting phenomenon here is the fact that the error always decreases substantially after these points. As explained, the program found a better direction to follow, which lead to better steps. Tensor Fox was tested with a fixed small number of CG iterations but this always lead to less accurate solutions, because the program got stuck at local minima.

Still with this same example, we can take iterations 93 and 94 to look in more detail. Iteration 93 is normal, with the error decreasing, but in iteration 94 the error increases. Let $F(\mathbf{w}^{(93)})$ and $F(\mathbf{w}^{(94)})$ be the errors at iterations 93 and 94, respectively. Remember that this notation was introduced in 3.2. We can consider the line $\mathbf{w}_t = \mathbf{w}^{(93)} + t\left(\mathbf{w}^{(94)} - \mathbf{w}^{(93)}\right)$ between the approximated CPDs and analyze how is the curve $F(\mathbf{w}_t)$, see figure 4.27

4.8 Regularization and preconditioning

4.8.1 Diagonal regularization

As suggested in A, our implementation of the preconditioner for Tensor Fox is the diagonal preconditioner. Remember we are using the following normal equations to compute the

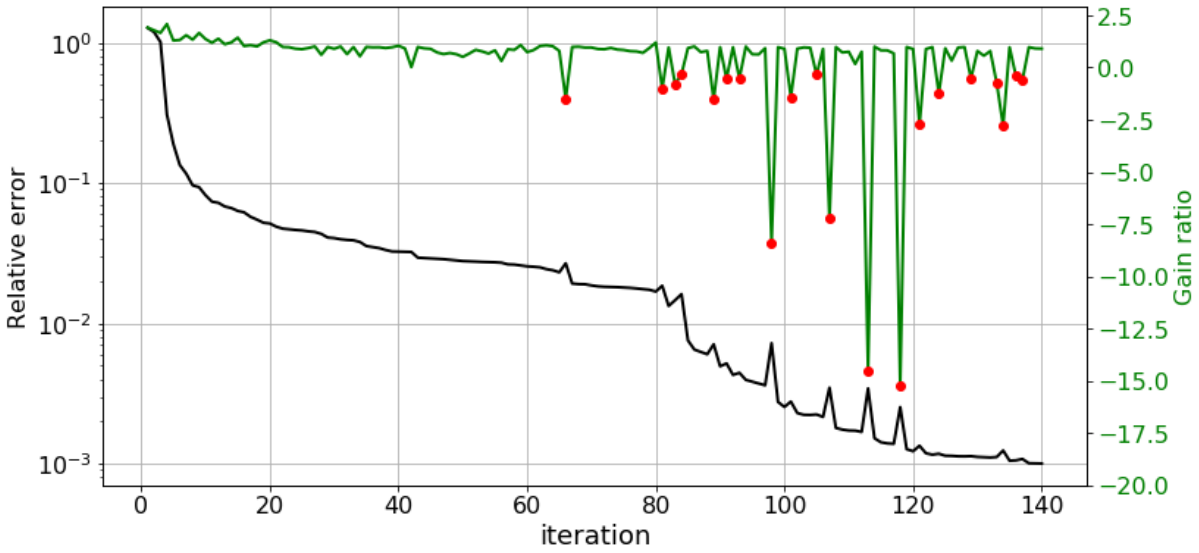

Figure 4.26: The black curve is the represents the error, the green curve represents the gain ratio and the red dots are the points when the gain ratio became negative. Note that there is always a peak in the error when this happens. This is to be expected, but more remarkable is the fact that the error always decreases substantially after these points.

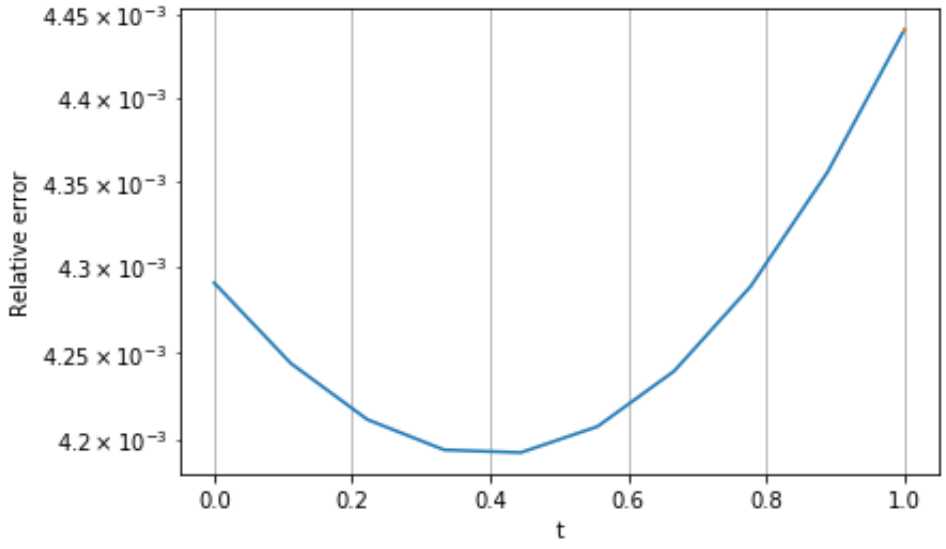

Figure 4.27: Error curve $F(\mathbf{w}_t)$. The error is minimal for $t \approx 0.4$ but we can see that the actual error (for $t = 1$) is bigger. At iteration 93 the gain ratio is $g = 0.974$ and the number it the program performed 19 CG iterations. At iteration 94 the gain ratio is $g = -0.518$ and the program performed 61 CG iterations. In both iterations the predicted error is $\mathcal{O}(10^{-8})$.

Gauss-Newton steps.

$$(\mathbf{J}_f^T\mathbf{J}_f + \mu\mathbf{D})\mathbf{x} = \mathbf{J}_f^T\mathbf{b}$$

where $\mathbf{J}_f = \mathbf{J}_f(\mathbf{w}^{(k)})$, $\mathbf{x} = \mathbf{w} - \mathbf{w}^{(k)}$, $\mathbf{b} = -f(\mathbf{w}^{(k)})$. For simplicity we start with the case of third order tensors, so $\mathbf{w} = [vec(\mathbf{X})^T, vec(\mathbf{Y})^T, vec(\mathbf{Z})^T]^T$, where $\mathbf{X} \in \mathbb{R}^{m\times R}, \mathbf{Y} \in \mathbb{R}^{n\times R}, \mathbf{Z}^{p\times R}$ are the corresponding factor matrices at the current step.

This preconditioner is ideal to use when the approximated Hessian $\mathbf{A}^T\mathbf{A}$ is diagonally dominant. However this may not always be the case. We can overcome this problem by choosing a suitable regularization matrix $\mathbf{D}$. The idea is to force $\mathbf{J}_f^T\mathbf{J}_f + \mu\mathbf{D}$ to be diagonally dominant, then the preconditioner matrix is applied. The first row of $\mathbf{J}_f^T\mathbf{J}_f$ is the first row of the matrix

$$\left[\langle \mathbf{Y}_1, \mathbf{Y}_1\rangle\langle \mathbf{Z}_1, \mathbf{Z}_1\rangle \mathbf{I}_m \ \ \ldots \ \ \langle \mathbf{Y}_1, \mathbf{Y}_R\rangle\langle \mathbf{Z}_1, \mathbf{Z}_R\rangle \mathbf{I}_m, \ \ \langle \mathbf{Z}_1, \mathbf{Z}_1\rangle \mathbf{X}_1\mathbf{Y}_1^T, \ \ \ldots \ \ \langle \mathbf{Z}_1, \mathbf{Z}_R\rangle \mathbf{X}_R\mathbf{Y}_1^T, \ \ \langle \mathbf{Y}_1, \mathbf{Y}_1\rangle \mathbf{X}_1\mathbf{Z}_1^T, \ \ \ldots \ \ \langle \mathbf{Y}_1, \mathbf{Y}_R\rangle \mathbf{X}_R\mathbf{Z}_1^T\right].$$

where $\mathbf{X}_r, \mathbf{Y}_r, \mathbf{Z}_r$ denotes the r-th column of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, respectively. We want to add a new term γ to the first entry $\langle \mathbf{Y}_1, \mathbf{Y}_1\rangle\langle \mathbf{Z}_1, \mathbf{Z}_1\rangle$ such that

$$\langle \mathbf{Y}_1, \mathbf{Y}_1\rangle\langle \mathbf{Z}_1, \mathbf{Z}_1\rangle + \gamma \geq |\langle \mathbf{Y}_1, \mathbf{Y}_r\rangle\langle \mathbf{Z}_1, \mathbf{Z}_r\rangle|, \quad |\langle \mathbf{Z}_1, \mathbf{Z}_r\rangle x_{1r}y_{j1}|, \quad |\langle \mathbf{Y}_1, \mathbf{Y}_r\rangle x_{1r}z_{k1}|$$

for all r, j, k. For this it is sufficient to define

$$\gamma = \|\mathbf{Y}_1\|\|\mathbf{Z}_1\| \max\left\{\max_r\{\|\mathbf{Y}_r\|\|\mathbf{Z}_r\|\}, \max_r\{\|\mathbf{X}_r\|\|\mathbf{Z}_r\|\}, \max_r\{\|\mathbf{X}_r\|\|\mathbf{Y}_r\|\}\right\}.$$

This choice is cheap to compute and it also works for all first m rows of $\mathbf{J}_f^T\mathbf{J}_f$. For the next m rows we can apply a similar idea to obtain the term

$$\gamma = \|\mathbf{Y}_2\|\|\mathbf{Z}_2\| \max\left\{\max_r\{\|\mathbf{Y}_r\|\|\mathbf{Z}_r\|\}, \max_r\{\|\mathbf{X}_r\|\|\mathbf{Z}_r\|\}, \max_r\{\|\mathbf{X}_r\|\|\mathbf{Y}_r\|\}\right\}.$$

Now we generalize this idea and introduce the notations

$$\gamma_{\mathbf{X}}^{(r)} = \|Y_r\|\|Z_r\| \max\left\{\max_{r'}\|Y_{r'}\|\|Z_{r'}\|, \ \max_{r'}\|X_{r'}\|\|Z_{r'}\|, \ \max_{r'}\|X_{r'}\|\|Y_{r'}\|\right\},$$

$$\gamma_{\mathbf{Y}}^{(r)} = \|X_r\|\|Z_r\| \max\left\{\max_{r'}\|Y_{r'}\|\|Z_{r'}\|, \ \max_{r'}\|X_{r'}\|\|Z_{r'}\|, \ \max_{r'}\|X_{r'}\|\|Y_{r'}\|\right\},$$

$$\gamma_{\mathbf{Z}}^{(r)} = \|X_r\|\|Y_r\| \max\left\{\max_{r'}\|Y_{r'}\|\|Z_{r'}\|, \ \max_{r'}\|X_{r'}\|\|Z_{r'}\|, \ \max_{r'}\|X_{r'}\|\|Y_{r'}\|\right\}.$$

From these values we finally define the regularization matrix

$$\mathbf{D} = \begin{bmatrix} \gamma_{\mathbf{X}}^{(1)}\mathbf{I}_m & & & & & & & & \\ & \ddots & & & & & & & \\ & & \gamma_{\mathbf{X}}^{(R)}\mathbf{I}_m & & & & & & \\ & & & \gamma_{\mathbf{Y}}^{(1)}\mathbf{I}_n & & & & & \\ & & & & \ddots & & & & \\ & & & & & \gamma_{\mathbf{Y}}^{(R)}\mathbf{I}_n & & & \\ & & & & & & \gamma_{\mathbf{Z}}^{(1)}\mathbf{I}_p & & \\ & & & & & & & \ddots & \\ & & & & & & & & \gamma_{\mathbf{Z}}^{(R)}\mathbf{I}_p \end{bmatrix}.$$

Note that $\mathbf{J}_f^T\mathbf{J}_f + \mathbf{D}$ is diagonally dominant, as desired, but now we are solving the regularized problem

$$\min_{\mathbf{w}} \|\mathbf{J}_f^T\mathbf{J}_f\mathbf{w} - \mathbf{J}_f^T\mathbf{b}\|^2 + \|\mathbf{D}\mathbf{w}\|^2.$$

This introduces a penalization over the solution, which can lead to local minimum, and this will have influence on the Gauss-Newton iterations. That is the reason for introducing the damping factor $\mu > 0$. Usually in the first iterations $\mu > 0$ has a moderate size, thus $\mathbf{J}_f^T\mathbf{J}_f + \mu\mathbf{D}$ still is very regularized. On the other hand, the solution obtained with this system will be closer to the objective point than the initial guess anyway. Besides that, this will be done very fast since the preconditioner is being taken full advantage of. As the iteration continues, we have $\mu \to 0$, thus the regularization has less effect, and the solutions computed starts to get closer to the real objective solution. This also implies that at this stage the preconditioner will have less effect, which means we need to perform more CG iterations in order to obtain meaningful solutions. Note how this is in accordance with the number of CG iterations describe in 4.1.3.4, since the strategy adopted foresee the necessity to increase the number of CG iterations as we keep making Gauss-Newton iterations.

To finish, we conclude that the preconditioner $\mathbf{M}$ is given by

$$\mathbf{M} = \begin{bmatrix} \mathbf{D}_{\mathbf{X}} & & \\ & \mathbf{D}_{\mathbf{Y}} & \\ & & \mathbf{D}_{\mathbf{Z}} \end{bmatrix},$$

where

$$\mathbf{D}_{\mathbf{X}} = \begin{bmatrix} \left(\|\mathbf{Y}_1\|^2\|\mathbf{Z}_1\|^2 + \mu\gamma_{\mathbf{X}}^{(1)}\right)\mathbf{I}_m & & \\ & \ddots & \\ & & \left(\|\mathbf{Y}_R\|^2\|\mathbf{Z}_R\|^2 + \mu\gamma_{\mathbf{X}}^{(R)}\right)\mathbf{I}_m \end{bmatrix},$$

$$\mathbf{D}_{\mathbf{Y}} = \begin{bmatrix} \left(\|\mathbf{X}_1\|^2\|\mathbf{Z}_1\|^2 + \mu\gamma_{\mathbf{Y}}^{(1)}\right)\mathbf{I}_n & & \\ & \ddots & \\ & & \left(\|\mathbf{X}_R\|^2\|\mathbf{Z}_R\|^2 + \mu\gamma_{\mathbf{Y}}^{(R)}\right)\mathbf{I}_n \end{bmatrix},$$

$$\mathbf{D}_{\mathbf{Z}} = \begin{bmatrix} \left(\|\mathbf{X}_1\|^2\|\mathbf{Y}_1\|^2 + \mu\gamma_{\mathbf{Z}}^{(1)}\right)\mathbf{I}_p & & \\ & \ddots & \\ & & \left(\|\mathbf{X}_R\|^2\|\mathbf{Y}_R\|^2 + \mu\gamma_{\mathbf{Z}}^{(R)}\right)\mathbf{I}_p \end{bmatrix}.$$

For the reader interested in the general case, first we need to know what is the diagonal of $\mathbf{J}_f^T\mathbf{J}_f$. From theorem 3.5.7 we know that

$$\mathbf{J}_f^T\mathbf{J}_f = \begin{bmatrix} \mathbf{H}_{11} & \dots & \mathbf{H}_{1L} \\ \vdots & & \vdots \\ \mathbf{H}_{L1} & \dots & \mathbf{H}_{LL} \end{bmatrix},$$

where

$$\mathbf{H}_{\ell'\ell'} = \begin{bmatrix} \prod_{\ell\neq\ell'}\omega_{11}^{(\ell)}\cdot\mathbf{I}_{I_{\ell'}} & \dots & \prod_{\ell\neq\ell'}\omega_{1R}^{(\ell)}\cdot\mathbf{I}_{I_{\ell'}} \\ \vdots & & \vdots \\ \prod_{\ell\neq\ell'}\omega_{R1}^{(\ell)}\cdot\mathbf{I}_{I_{\ell'}} & \dots & \prod_{\ell\neq\ell'}\omega_{RR}^{(\ell)}\cdot\mathbf{I}_{I_{\ell'}} \end{bmatrix}$$

are the diagonal blocks of $\mathbf{J}_f^T\mathbf{J}_f$. The diagonal of $\mathbf{H}_{\ell'\ell'}$ is

$$\mathbf{D}_{\ell'\ell'} = \begin{bmatrix} \prod_{\ell\neq\ell'}\omega_{11}^{(\ell)}\cdot\mathbf{I}_{I_{\ell'}} & & \\ & \ddots & \\ & & \prod_{\ell\neq\ell'}\omega_{RR}^{(\ell)}\cdot\mathbf{I}_{I_{\ell'}} \end{bmatrix} = \begin{bmatrix} \prod_{\ell\neq\ell'}\|\mathbf{w}_1^{(\ell)}\|^2\cdot\mathbf{I}_{I_{\ell'}} & & \\ & \ddots & \\ & & \prod_{\ell\neq\ell'}\|\mathbf{w}_R^{(\ell)}\|^2\cdot\mathbf{I}_{I_{\ell'}} \end{bmatrix},$$

hence the diagonal of $\mathbf{J}_f^T\mathbf{J}_f$ is the matrix

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & & \\ & \ddots & \\ & & \mathbf{D}_{LL} \end{bmatrix}.$$

In the case the factors are norm-balanced, we know that $\|\mathbf{w}_r^{(1)}\| = \ldots = \|\mathbf{w}_r^{(L)}\|$ for all $r = 1 \ldots R$. Denote this norm by α_r. Then we have that $\prod_{\ell\neq\ell'}\|\mathbf{w}_r^{(\ell)}\|^2 = \alpha_r^{2(L-1)}$. Therefore all matrices $\mathbf{D}_{\ell'\ell'}$ are equal and

$$\mathbf{D} = \begin{bmatrix} \begin{bmatrix} \alpha_1^{2(L-1)} \cdot \mathbf{I}_{I_1} & & \\ & \ddots & \\ & & \alpha_R^{2(L-1)} \cdot \mathbf{I}_{I_1} \end{bmatrix} & & \\ & \ddots & \\ & & \begin{bmatrix} \alpha_1^{2(L-1)} \cdot \mathbf{I}_{I_L} & & \\ & \ddots & \\ & & \alpha_R^{2(L-1)} \cdot \mathbf{I}_{I_L} \end{bmatrix} \end{bmatrix}.$$

This may be useful if one want to use a diagonal preconditioner $\mathbf{M}$ such that the matrix $\mathbf{M}^{-\frac{1}{2}}(\mathbf{J}_f^T\mathbf{J}_f + \mu\mathbf{D})\mathbf{M}^{-\frac{1}{2}}$ is unit diagonal (but not necessarily diagonally dominant). We remark that this kind of preconditioner may be enough, since the balancing of the factor matrices (see section 4.1.3.5) already helps $\mathbf{J}_f^T\mathbf{J}_f$ to be more diagonally dominant.

4.8.2 Computational experiments

To simplify notation let $\mathbf{A} = \mathbf{J}_f, \mathbf{H} = \mathbf{A}^T\mathbf{A}, \mathbf{b} = -f$, where we suppress the argument $\mathbf{w}^{(k)}$. Tensor Fox use equation 4.1 for the CG algorithm, however it is more usual to use the Levenberg-Marquardt equation,

$$\text{diag}(\mathbf{H} + \mu\mathbf{I})^{-1/2} \cdot (\mathbf{H} + \mu\mathbf{I}) \cdot \text{diag}(\mathbf{H} + \mu\mathbf{I})^{-1/2}\mathbf{x} = \text{diag}(\mathbf{H} + \mu\mathbf{I})^{-1/2}\mathbf{A}^T\mathbf{b}. \tag{4.3}$$

The only difference is that Tensor Fox uses $\mathbf{D}$ instead of the identity matrix. We know that the condition number of $\text{diag}(\mathbf{H} + \mu\mathbf{I})^{-1/2} \cdot (\mathbf{H} + \mu\mathbf{I}) \cdot \text{diag}(\mathbf{H} + \mu\mathbf{I})^{-1/2}$ must be small to solve equation 4.3 efficiently. Although the Levenberg-Marquardt formulation is used in many implementations of the dGN, we will see that our diagonal matrix produces better results.

Besides that formulation, we also have the Tensorlab's formulation

$$\mathbf{M}_{bd} \cdot (\mathbf{H} + \mu\mathbf{I})\mathbf{x} = \mathbf{M}_{bd} \cdot \mathbf{A}^T\mathbf{b}, \tag{4.4}$$

where $\mathbf{M}_{bd}$ is the block diagonal preconditioner as explained in 4.1.3.8. In figures 4.28, 4.29, 4.30, 4.31 we show the evolution of the condition number of the approximated (with and without preconditioning) Hessian for some of the test tensors used before.

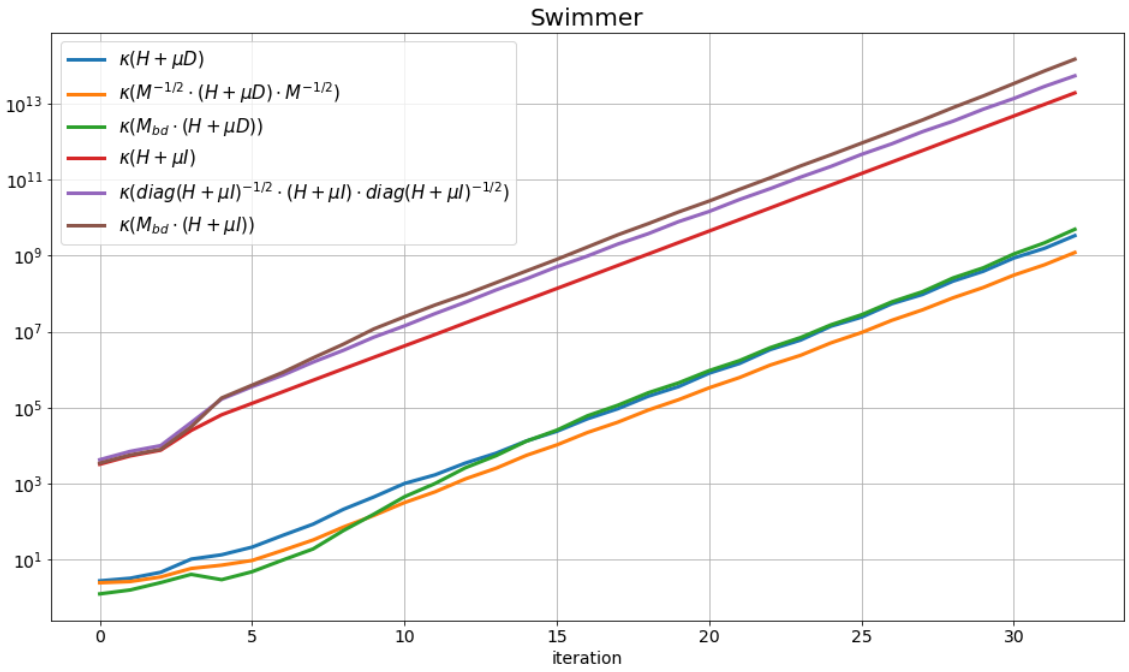

Figure 4.28: Condition number for each iteration of several approaches to compute a CPD for the swimmer tensor. Note that it is showed the condition number of the approximated Hessian (always regularized) with and without regularization. The condition number of Tensor Fox is the orange one.

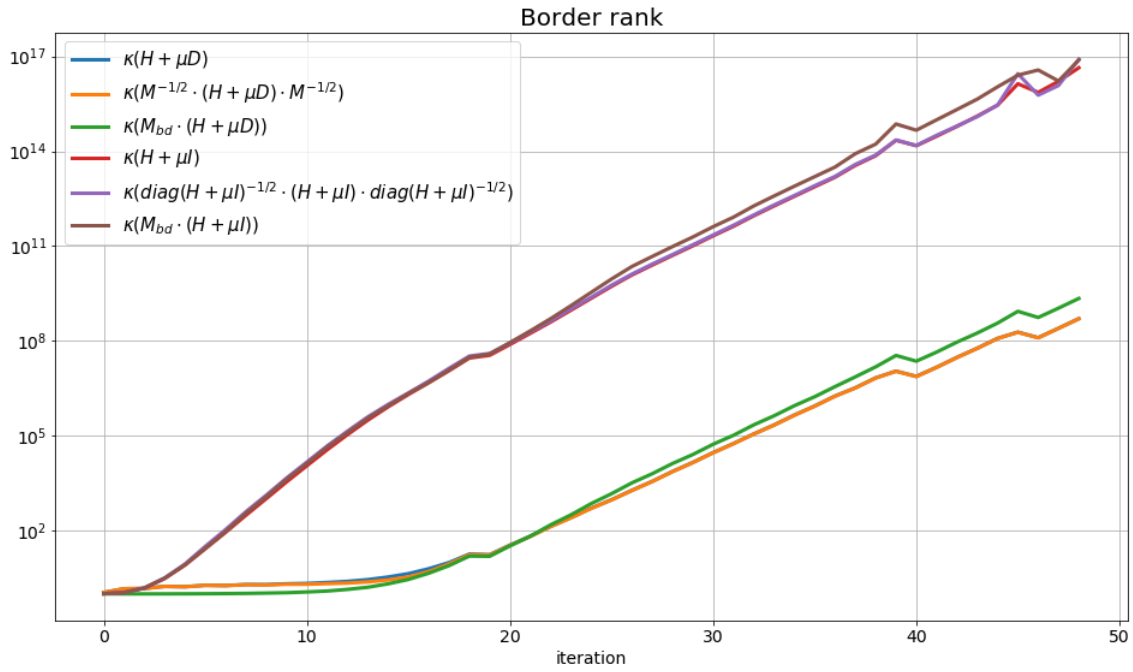

Figure 4.29: Condition number for each iteration of several approaches to compute a CPD for the border rank tensor.

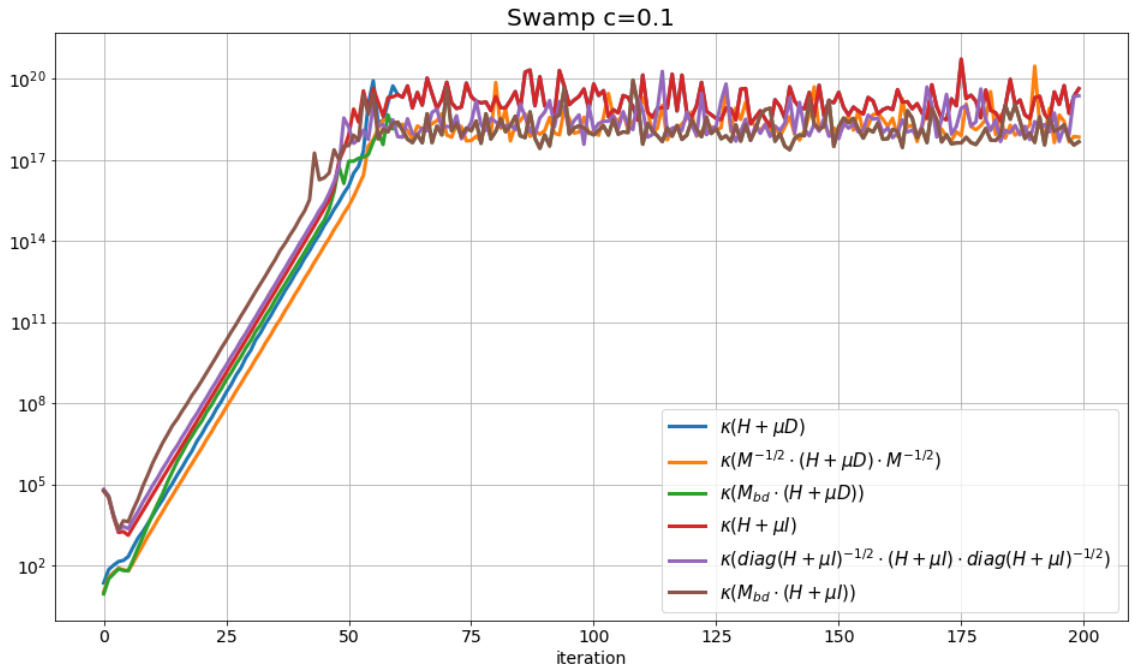

Figure 4.30: Condition number for each iteration of several approaches to compute a CPD for the swamp tensor with $c = 0.1$.

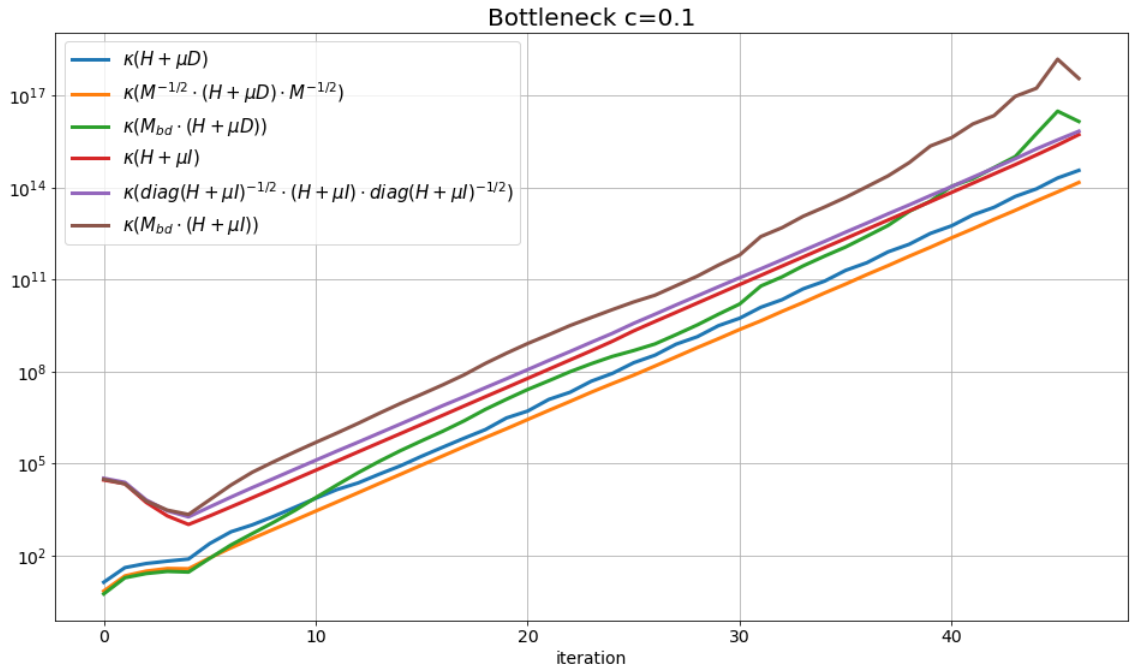

Figure 4.31: Condition number for each iteration of several approaches to compute a CPD for the double bottleneck tensor with $c = 0.1$.

Note that we tried the block diagonal preconditioner $\mathbf{M}_{bd}$ both on the Levenberg-Marquardt and Tensor Fox formulations. It is clear that the diagonal preconditioner we are using here brings the best results in terms of lowering the condition number. Maybe the only exception are the swamp tensors, where after some iterations all approaches are as large as possible.

In the next section we talk about the condition number of tensors and introduce a family of difficult tensors to handle. These tensors are such that most close CPDs are ill-conditioned even when the original CPD is well-conditioned. Figure 4.32 shows the evolution of the condition number of the approximated Hessian for the parameters $r = 25, c = 0.75, s = 3$ (they are explained in the next section). In this case, again, we observe that the approach of Tensor Fox is better than the others.

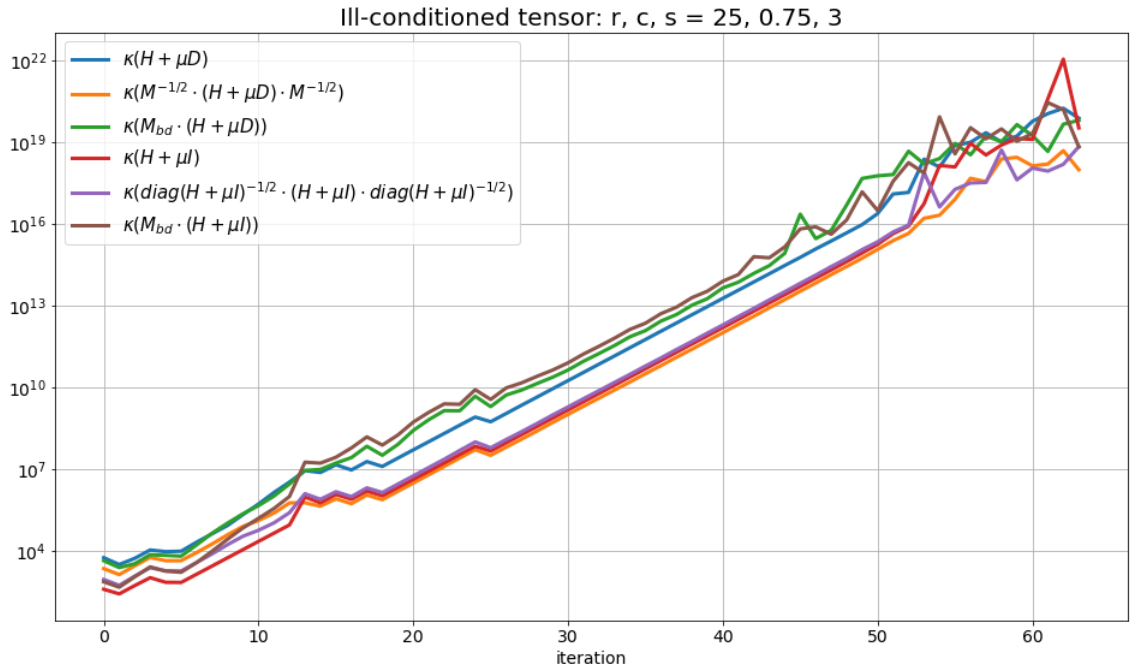

Figure 4.32: Condition number for each iteration of several approaches to compute a CPD for an ill-conditioned tensor with $r = 25, c = 0.75, s = 3$.

4.9 Conditioning

In this section we briefly mention the main concepts and results of conditioning of tensors. For more about the subject the reader is invited to check the references [34, 41–43].

4.9.1 Definitions and results

Definition 4.9.1. *The* Segre map *is the map given by*

$$\begin{aligned} Seg: \ & \mathbb{R}^{I_1} \times \ldots \times \mathbb{R}^{I_L} \to \mathbb{R}^{I_1 \times \ldots \times I_L} \\ & (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(L)}) \mapsto \boldsymbol{x}^{(1)} \otimes \ldots \otimes \boldsymbol{x}^{(L)} \end{aligned}.$$

The image of this map without its origin is called the *Segre variety*. It will be denoted by $S = \{\mathbf{x}^{(1)} \otimes \ldots \otimes \mathbf{x}^{(L)} : \ \mathbf{x}^{(\ell)} \neq 0 \text{ for all } \ell = 1 \ldots L\}$. Now consider the additive map given by

$$\begin{aligned} \Phi_R: \ & S \times \ldots \times S \to \mathbb{R}^{I_1 \times \ldots \times I_L} \\ & (\mathcal{T}_1, \ldots, \mathcal{T}_R) \mapsto \mathcal{T}_1 + \ldots + \mathcal{T}_R \end{aligned}.$$

Let σ_R be the set of tensors with rank $\leq R$. From the definition we have that $\mathrm{Im}(\Phi_R) = \sigma_R$ and $rank(T) = \min\{r : \ \mathcal{T} \in \sigma_r\}$ for all $T \in \mathbb{R}^{I_1 \times \ldots \times I_L}$. The derivative of Φ_R at

$(\mathcal{T}_1, \ldots, \mathcal{T}_R)$ is the map

$$\mathrm{d}_{(\mathcal{T}_1,\ldots,\mathcal{T}_R)}\Phi_R : \mathrm{T}_{\mathcal{T}_1}S \times \ldots \times \mathrm{T}_{\mathcal{T}_R}S \to \mathrm{T}_{\Phi_R(\mathcal{T}_1,\ldots,\mathcal{T}_R)}\mathbb{R}^{I_1\times\ldots\times I_L}$$

given by

$$\mathrm{d}_{(\mathcal{T}_1,\ldots,\mathcal{T}_R)}\Phi_R(\dot{\mathcal{T}}_1, \ldots, \dot{\mathcal{T}}_R) = \dot{\mathcal{T}}_1 + \ldots + \dot{\mathcal{T}}_R,$$

where $\mathrm{T}_{\mathcal{T}_r}S$ is the tangent space of S at $\mathcal{T}_r$. Then the condition number [92] of the CPD at $(\mathcal{T}_1, \ldots, \mathcal{T}_R)$ is

$$\kappa\left((\mathcal{T}_1, \ldots, \mathcal{T}_R), \mathcal{T}\right) = \lim_{\varepsilon\to 0} \max_{\mathcal{T}'\in B_\varepsilon(\mathcal{T})\cap\sigma_R} \frac{\|\tilde{\Phi}^{-1}(\mathcal{T}) - \tilde{\Phi}^{-1}(\mathcal{T}')\|}{\|\mathcal{T} - \mathcal{T}'\|},$$

where $B_\varepsilon(\mathcal{T}')$ is an ε-ball centered at $\mathcal{T}$ and $\tilde{\Phi}^{-1}$ is a local inverse of Φ at $(\mathcal{T}_1, \ldots, \mathcal{T}_R)$. If this local inverse doesn't exist then we define $\kappa\left((\mathcal{T}_1, \ldots, \mathcal{T}_R), \mathcal{T}\right) = \infty$. Both norm are the Frobenius norm as it is the case in all this work. Note that $\kappa\left((\mathcal{T}_1, \ldots, \mathcal{T}_R), \mathcal{T}\right)$ is completely determined by the choice of $\mathcal{T}_1, \ldots, \mathcal{T}_R$ since $\mathcal{T} = \Phi_R(\mathcal{T}_1, \ldots, \mathcal{T}_R)$, therefore we set $\kappa\left(\mathcal{T}_1, \ldots, \mathcal{T}_R\right) = \kappa\left((\mathcal{T}_1, \ldots, \mathcal{T}_R), \mathcal{T}\right)$. In [43] it is showed that the condition number is the inverse of the smallest singular value of $\mathrm{d}_{(\mathcal{T}_1,\ldots,\mathcal{T}_R)}\Phi_R$.

Remember that the condition number measures the sensitivity of $(\mathcal{T}_1, \ldots, \mathcal{T}_R)$ to perturbations of $\Phi_R(\mathcal{T}_1, \ldots, \mathcal{T}_R)$. A rule of thumb of numerical analysis is the inequality

$$\text{forward error} \lessapprox \text{condition number} \cdot \text{backward error},$$

which in this context translates to

$$\|(\mathcal{T}'_1, \ldots, \mathcal{T}'_R) - (\mathcal{T}_1, \ldots, \mathcal{T}_R)\| \lessapprox \kappa(\mathcal{T}_1, \ldots, \mathcal{T}_R) \cdot \|\mathcal{T}' - \mathcal{T}\|, \tag{4.5}$$

where

$$\|(\mathcal{T}'_1, \ldots, \mathcal{T}'_R) - (\mathcal{T}_1, \ldots, \mathcal{T}_R)\| = \min_{\sigma\in S_R} \sqrt{\sum_{r=1}^{R} \|\mathcal{T}'_r - \mathcal{T}_{\sigma(r)}\|^2}$$

and S_R be the group of permutations of R elements. From this inequality we can see that all algorithms showed so far are minimizing the backward error instead of the forward error. Relying on the backward error alone can be a dangerous practice in the presence of ill-conditioned CPDs.

4.9.2 A special family of tensors

In [42] this aspect of conditioning is explored and a family of $15 \times 15 \times 15$ tensors is proposed for testing. This family is such that the best algorithm of Tensorlab (NLS) failed to produce well conditioned approximations even when the original tensor was well

RGN-HR	RGN-Reg	Tensorlab
23.6%	42.3%	51.0%

Table 4.19: Fraction of ill-conditioned CPDs.

conditioned. The parameters $(r, c, s) \in \{15, 20, 25, 30\} \times \{0, 0.25, 0.5, 0.75\} \times \{1, 2, 3, 4\}$ are used to generate each tensor of the family. Let $\mathbf{R}_c$ be the upper triangular factor in the Cholesky decomposition $\mathbf{R}_c^T \mathbf{R}_c = c\mathbf{1}\mathbf{1}^T + (1-c)\mathbf{I}_r$, where $\mathbf{1} \in \mathbb{R}^r$ is the vector of ones and $\mathbf{I}_r$ is the 15×15 identity matrix. Then each CPD of this family with parameters r, c, s is given by $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \in \mathbb{R}^{15 \times r}$, where

$$\mathbf{A}_i = \mathbf{N}_i \cdot \mathbf{R}_c \cdot \operatorname{diag}\left(10^{\frac{s}{3r}}, 10^{\frac{2s}{3r}}, \ldots, 10^{\frac{rs}{3r}}\right),$$

where $\mathbf{N}_i$ is a $15 \times r$ random matrix with its entries draw from the normal distribution $\mathcal{N}(0, 1)$ with mean 0 and variance 1.

Their experiment proceed as follow:

1. Randomly sample $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ as described above, and let $\mathcal{T}' = \Phi_r(\mathcal{T}'_1, \ldots, \mathcal{T}'_r)$, where the rank one tensor are obtained from the factors $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$;

2. create a perturbed tensor $\mathcal{T} = \frac{\mathcal{T}'}{\|\mathcal{T}'\|} + 10^{-3}\frac{\mathcal{E}}{\|\mathcal{E}\|}$, where $\mathcal{E}$ is a $15 \times 15 \times 15$ random tensor with its entries draw from the normal distribution $\mathcal{N}(0, 1)$;

3. randomly samples factor matrices $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3 \in \mathbb{R}^{15 \times r}$ to be used as initialization;

4. compute an approximated CPD for $\mathcal{T}$ from the initialization $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$.

In the original experiment they used some specific parameter configuration for Tensorlab NLS, but we won't go into such details here. The reader may read section 7 of [42] for information. For each triple (r, c, s) one tensor is generated by the procedure above, then 25 random initializations are generated and 25 CPDs are computed, on for each different initialization. For a very significant fraction of initializations, the state-of-the-art method halt at extremely ill-conditioned CPDs. Below we show a piece of their results, where Tensorlab had more trouble to produce well-conditioned solutions. The percentage showed is the fraction of cases where the condition number of the approximated CPD is bigger than 10^3 among those CPDs whose backward error is very small (less than $1.1 \cdot 10^{-3}$). From 4.5 we can conclude that these CPDs are completely uninterpretable, that is, no correct significant digits are present in the individual rank one terms.

RGN-HR stands for *Riemannian Gauss-Newton with hot restarts* and RGN-Reg stands for *Riemannian Gauss-Newton with regularization*. The former is their main method, while the latter was included only to illustrate that the proposed hot restarts mechanism

provides a superior way of handling ill-conditioning. RGN-HR is not just more well-conditioned than Tensorlab's NLS, but it is also faster in most cases as showed in the paper.

4.9.3 Results

Here we tried to make the parameter setting as close as possible to the choices made for Tensorlab. First it was observed that better results were obtained with fixed damping parameter $\mu = 10^{-8}$. The most relevant parameter is the number of CG iterations. Increasing this value consistently lead to more well-conditioned solutions, which makes sense because the ill-conditioning comes from the approximated Hessian formulation, which is singular. In the paper they used a maximum of `cg_maxiter` $= 75$ number of iterations for the CG. Also, although Tensor Fox normally does not use the dogleg method[13], for this problem in particular this method showed to be handful.

The dogleg feature is used when the error of some iteration is too big compared to the previous error. This is a case when the unusual step is not considered unusual but some kind of divergence. This rarely is used in Tensor Fox because the program only recognizes divergence when the error is 100 times bigger than the previous error, and this almost never happens. In any case, when this does happen, the program draws back to a "version" of the previous step and perform the dogleg method to produce the new step. We explain what is this version of the previous step. Let $\mathbf{w}^{(k)}$ be the k-th dGN step. Remember that before computing the next iteration, first we make the factor norm balanced, let $N(\mathbf{w}^{(k)})$ be this norm-balanced representation. Then the next dGN iteration is computed with $N(\mathbf{w}^{(k)})$ instead of $\mathbf{w}^{(k)}$. We already observed in 4.1.3.5 that this procedure always improves the conditioning of the iteration. In the case the next step, $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \Delta\mathbf{w}$, has a large error, we take its norm-balanced representation and shift it back with the dGN step $\Delta\mathbf{w}$. The new point obtained, $N(\mathbf{w}^{(k+1)}) - \Delta\mathbf{w}$ should be close to the norm-balanced representation of $\mathbf{w}^{(k)}$, but not quite the same. Then the dogleg method takes action and improves the step. Figure 4.33 illustrates the process of drawing back a step. We remark that $N(\mathbf{w}^{(k+1)}) - \Delta\mathbf{w}$ usually is different than $N(\mathbf{w}^{(k)})$.

[13] See [22] for more about this method.

RGN-HR	**RGN-Reg**	**Tensorlab**	**Tensor Fox** (75)	**Tensor Fox** (1600)
23.6%	42.3%	51.0%	65.5%	21.4%

Table 4.20: Fraction of ill-conditioned CPDs, including Tensor Fox.

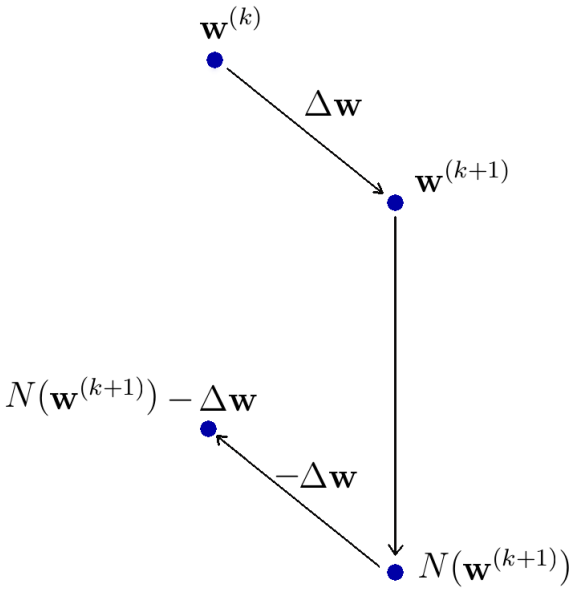

Figure 4.33: After making a bad step, the program draws back in a way that the new point is close to norm-balanced, then it applies the dogleg method.

For this particular family of tensors we could verify that it is better perform draw back and apply the dogleg method in every iteration. In the table below we repeat the previous results plus the Tensor Fox results in two cases. As we can see, Tensor Fox is no better than Tensorlab for `cg_maxiter` $= 75$ and it is better than RGN-HR for `cg_maxiter` $= 1600$. The situation changes as we increase `cg_maxiter`. We tested the values `cg_maxiter` $= 75, 200, 400, 600, 800, 1000, 1200, 1400, 1600$, see figure 4.34.

Of course the running time increases as we increase `cg_maxiter`, however the expected time to get a well-conditioned CPD decreases. The same does not occurs with Tensorlab. We increased `cg_maxiter` for Tensorlab and observed an improvement with regard the conditioning but the running time in this case increase substantially. For instance, when `cg_maxiter` $= 400$, the slowest expected time for Tensor Fox obtained was 33.3 seconds, with $r = 30, c = 0.75, s = 1$, and for Tensorlab it was 101.2 seconds, with $r = 30, c = 0.25, s = 1$. Even more, in all tests with Tensor Fox this expect time is always less than 40 seconds. This happens because the CG iterations of Tensor Fox are lighter than Tensorlab's, which is something already observed in 4.1.3.8.

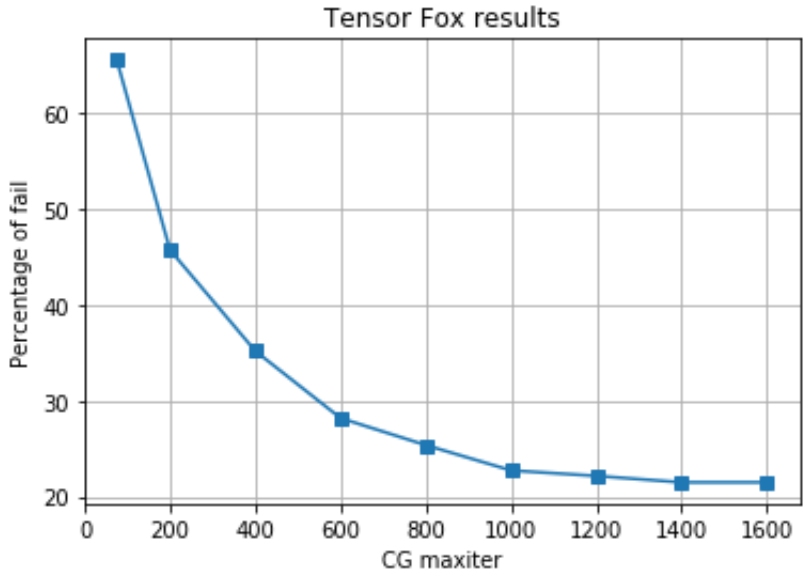

Figure 4.34: Percentage of fails as we increase the maximum number of CG iterations.

We finish this section observing that the algorithm RGN-HR is able to obtain many well-conditioned solutions in less time than Tensor Fox and Tensorlab. Their approach is very different in some aspects and it is worth reading about it. The speed of RGN-HR tends to be slower for larger tensors, which is a major drawback. However if the rank or the multilinear rank is small enough, using tolerance-based compression may reduced the problem substantially, and then RGN-HR can be efficiently used.

4.10 Parallelism

Several parts of Tensor Fox are open to parallelism: unfoldings, SVDs, matrix multiplications, Khatri-Rao products, and so on. With regard to the dGN algorithm, the main costs comes from the Khatri-Rao products to obtain the gradient, and the matrix-matrix multiplications. Both routines run in parallel in Tensor Fox. In particular, the matrix-matrix multiplication uses the BLAS parallelism, which is as good as one can get.

Still, the dGN algorithm, by design is not highly parallel, that is, each iteration must be computed sequentially. Besides that, within each iteration we must run the CG algorithm, which run each iteration sequentially again. Figure 4.35 shows the speed-up obtained by the addition of threads to compute a rank-15 CPD of a $2000 \times 2000 \times 2000$ random tensor with rank 15. We are not compressing in this example. The machine used in this experiment was the AWS instance m5d.24xlarge, consisting of a Intel Xeon Platinum 8000 series (Skylake-SP) processor with a sustained all core Turbo CPU clock speed of up to 3.1 GHz, and 384 GB of memory. This computational and memory power was necessary due to the size of the tensor. The reason for using such a big tensor is because the parallelism increases as we increase the dimensions and the rank.

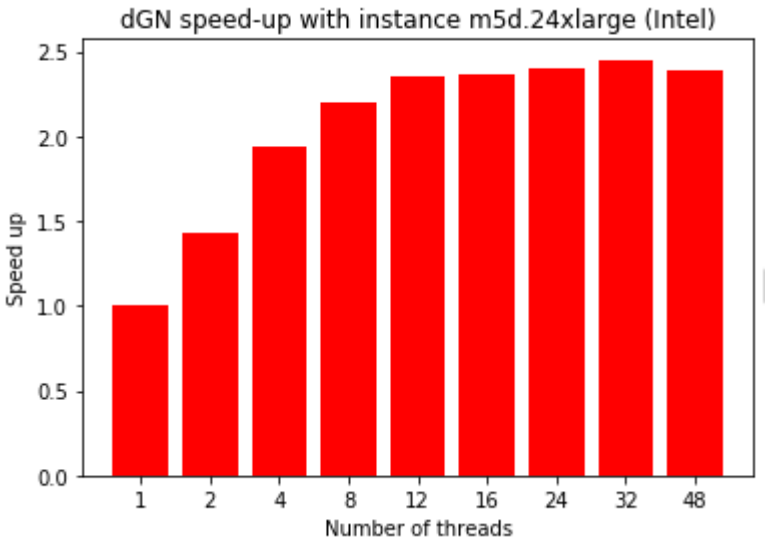

Figure 4.35: Speed-up with the number of threads when computing a rank-15 CPD of a $2000 \times 2000 \times 2000$ random tensor with rank 15 without compression.

Even with a big tensor as this one, the speed-up curve gets flat already for 16 threads. This shows how the dGN algorithm, the way it is presented in this thesis, is not highly parallel. For smaller tensors as the swimmer tensor (see figure 4.36) one can expect the speed-up to drop as the number of threads increases due to the overhead communication. There is just too much data being passed for a small number of operations.

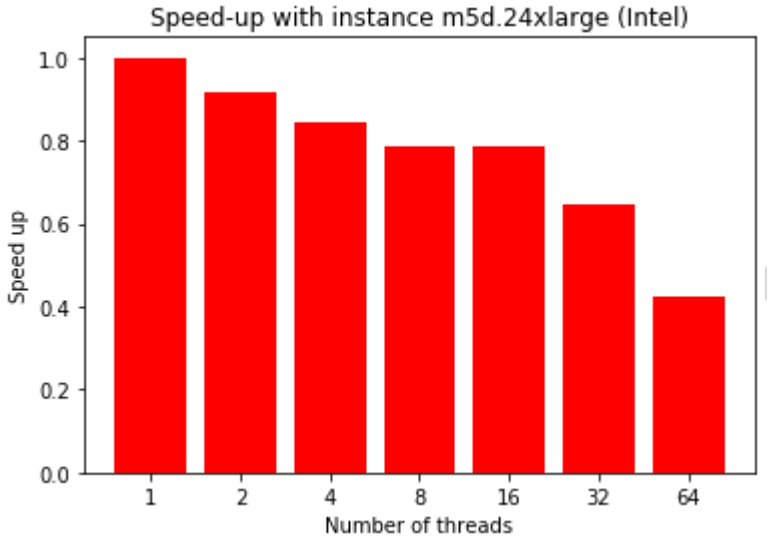

Figure 4.36: Speed-up with the number of threads when computing a rank-50 CPD of the swimmer tensor.

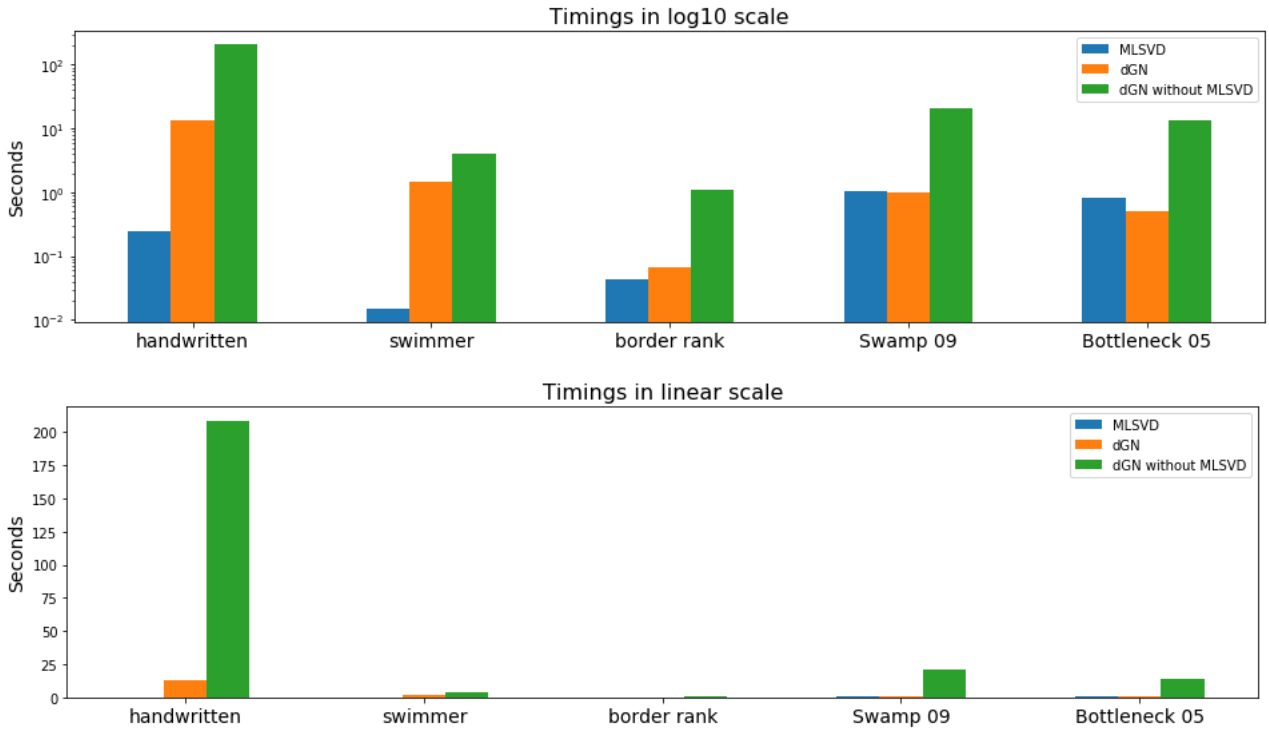

Figure 4.37: Timings to compute parts of the CPDs of known tensors. Note that summing the bars of the MLSVD and dGN timings is not the correct comparison in log scale. That is why we also showed the linear scale, so the reader have a notion of the real difference in time when there is no compression.

The main source of parallelism comes from the MLSVD, which relies on several randomized SVDs [31], and this method is highly parallel. [14] As we stressed before, compressing before running the dGN is a procedure we should always do. Figure 4.37 reinforce this claim by showing that the compression time plus the dGN time is less than the dGN time without compression.

4.11 What are the main features of Tensor Fox?

In the previous sections we introduced Tensor Fox in details, and after that we conduct a lot of experiments proving that Tensor Fox is competitive in a wide range of examples. After that the reader is left with the impression that all new features are relevant and indispensable. In this section we investigate the main features, turning them off and running the test tensors. The features to be investigated are:

- Stopping conditions
- Regularization
- Preconditioning

[14] In fact, the randomized method for SVDs are also very efficient if running in a GPU.

- Number of CG iterations

Before anything, we want to remark that one of the most relevant features of Tensor Fox is the tolerance based compression, but this feature is already investigated enough.

One thing the reader may notice is the large number of stopping conditions introduced in 4.38. Below we show their frequency when running 20 CPDs for each tensor. As we can see, most of the time the stopping condition used is the error improvement, which usually the most common in all solvers. Sometimes none condition is triggered and the program just reach the maximum number of iterations. The novelty here are the conditions with averages, which really helps to stop the iterations earlier.

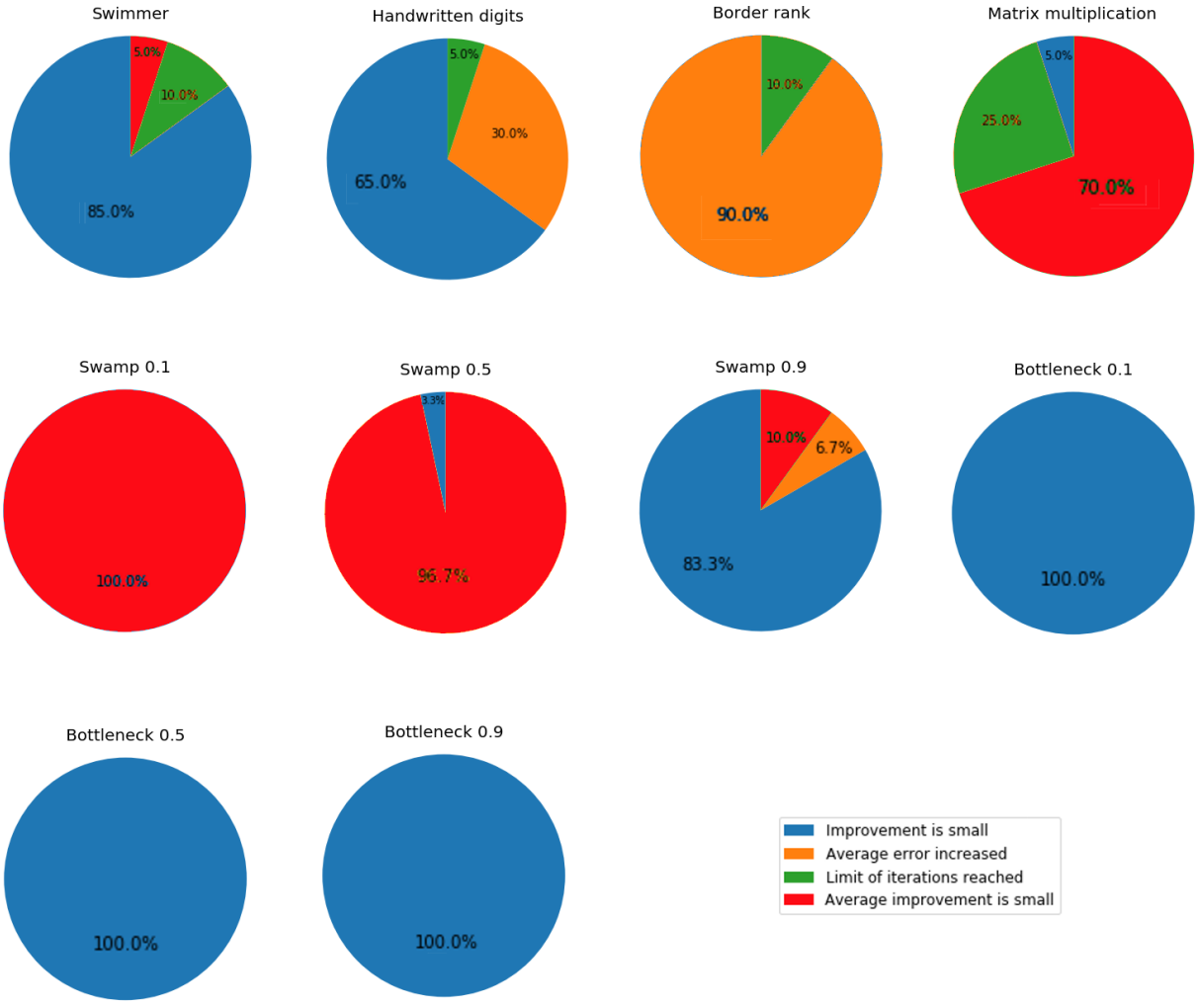

Figure 4.38: Pie charts with the stopping condition frequencies.

For instance, we can note that the averages of the improvements are the unique stopping condition triggered for the swamp tensor with $c = 0.1$. In figure 4.39 we show the error evolution to compute a CPD for this tensor with this stopping condition disabled. The red dot is the iteration where the program would stop if the condition were enabled.

We can see that there is almost no reason to continue iterating, therefore the stopping condition is appropriate.

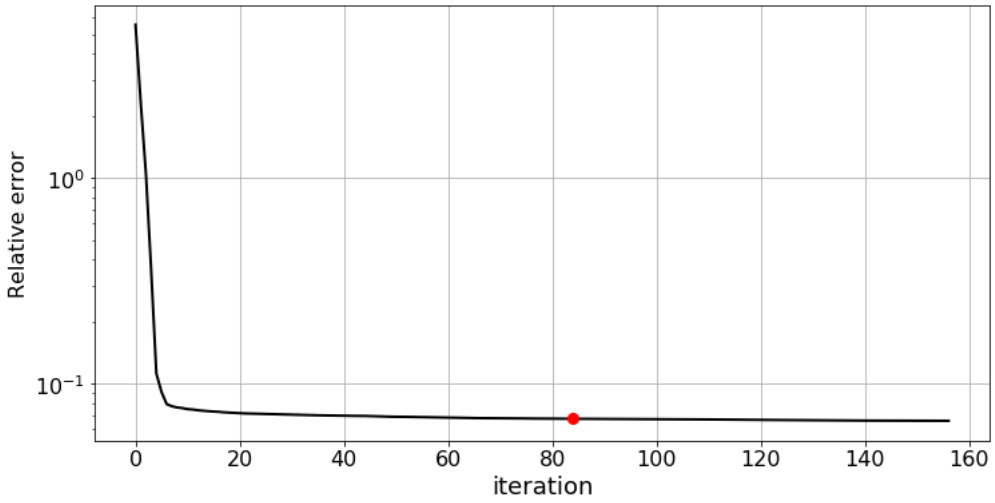

Figure 4.39: Disabling stopping condition for the swamp tensor with $c = 0.1$.

The other features of Tensor Fox we analyze at the same time. Basically we will repeat the experiments as we did to produce the box plots before, but the models compared are variations of Tensor Fox, with the relevant features turned on or off. We denote by **RND** the Tensor Fox default model, using random `cg_maxiter` as described in 4.1.3.4, and **CG100** the model where `cg_maxiter`= 100. Additionally, we write "no reg" when regularization is removed from the model and "no prec" when preconditioning is removed from the model. All results are summarized in figures 4.40 to 4.49.

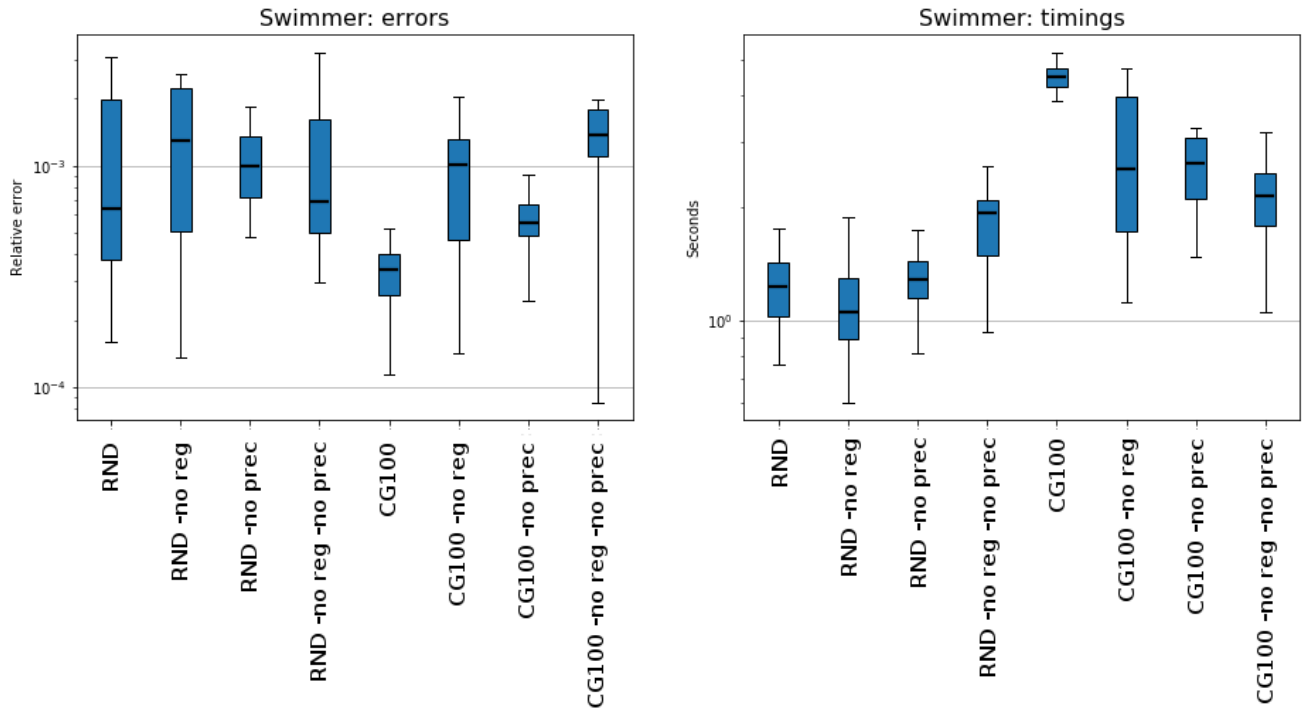

Figure 4.40: Box plots with errors and timings for the swimmer tensor.

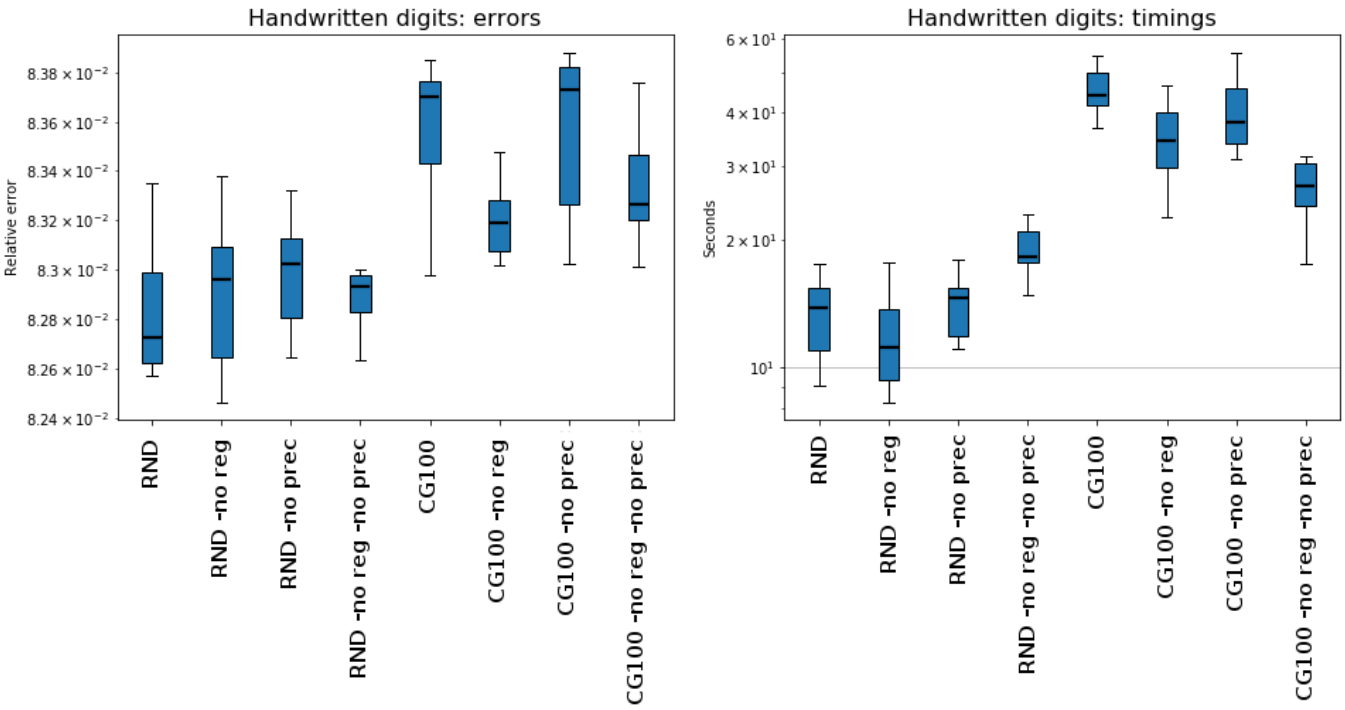

Figure 4.41: Box plots with errors and timings for the handwritten digits tensor.

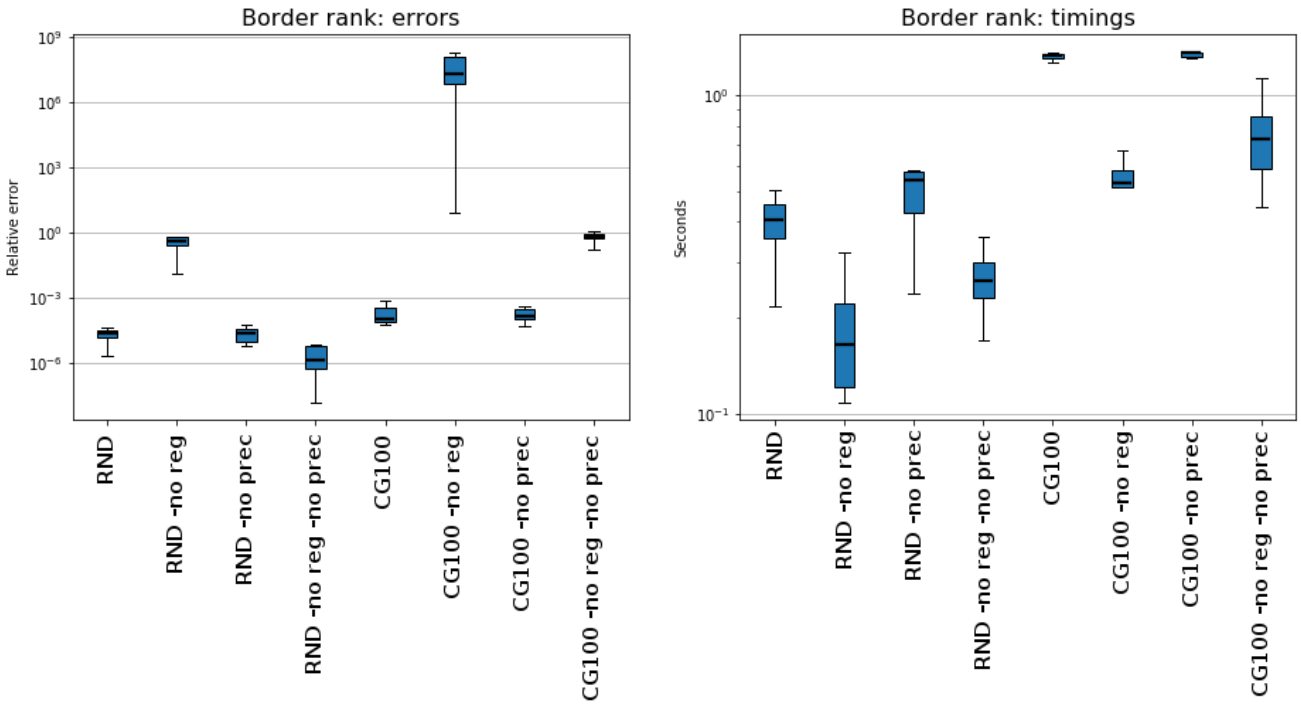

Figure 4.42: Box plots with errors and timings for the border rank tensor.

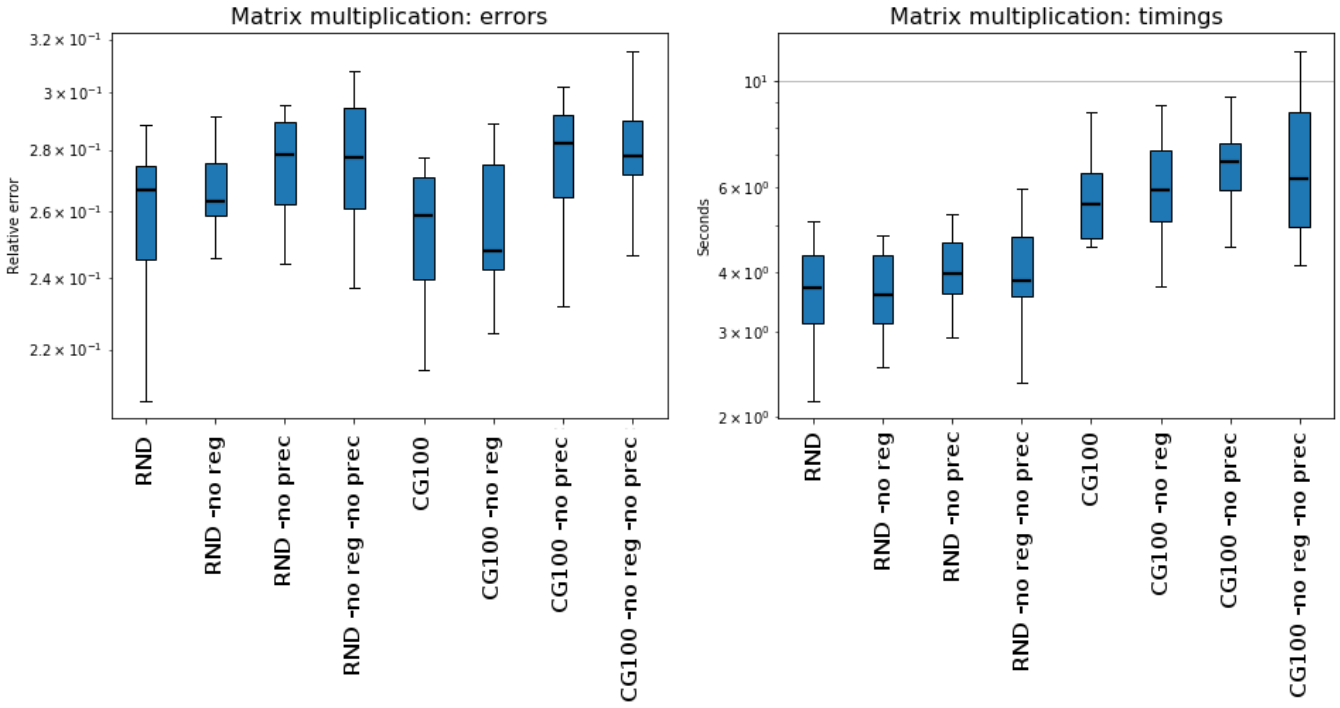

Figure 4.43: Box plots with errors and timings for the matrix multiplication tensor.

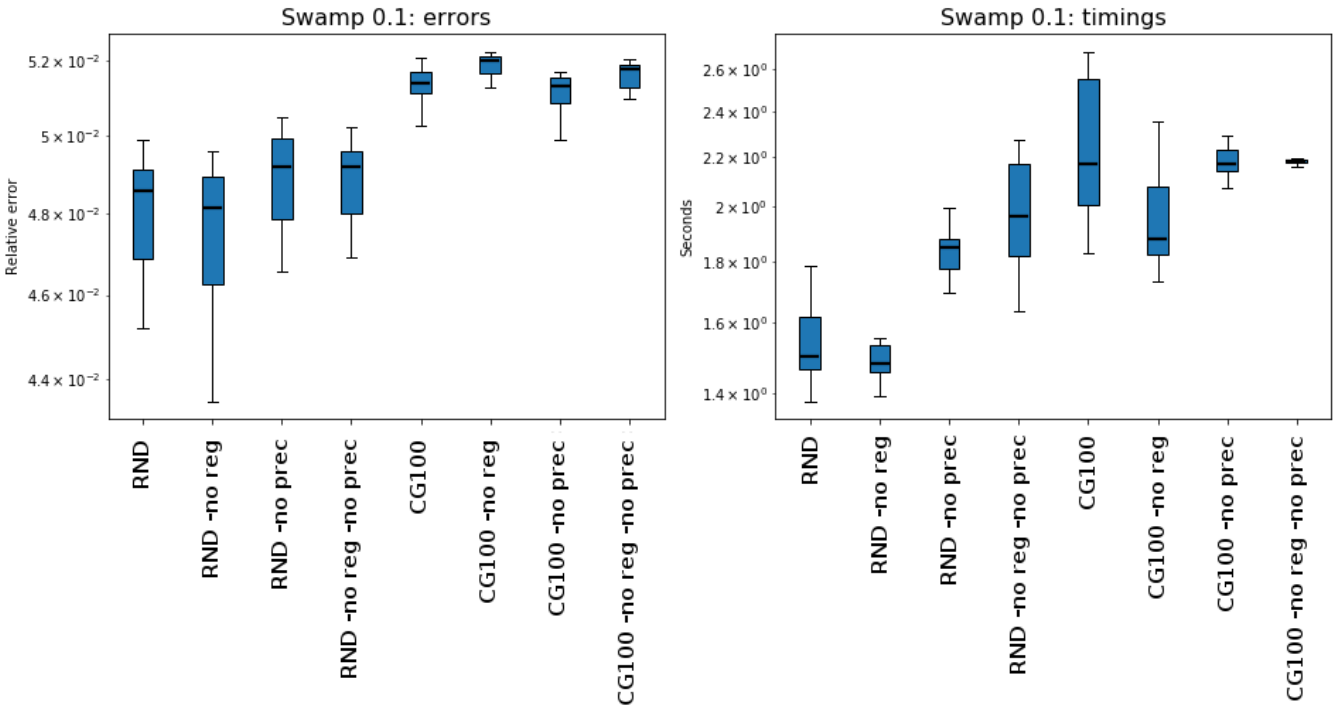

Figure 4.44: Box plots with errors and timings for the swamp $c = 0.1$ tensor.

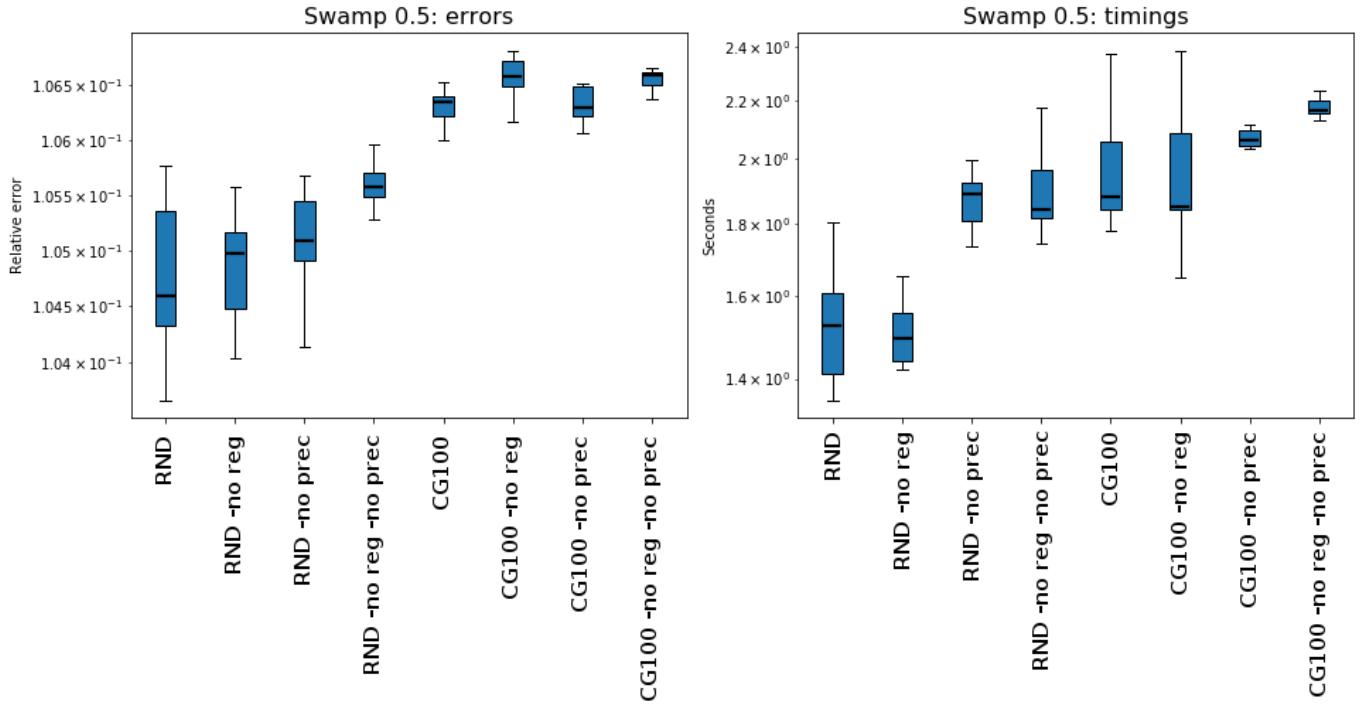

Figure 4.45: Box plots with errors and timings for the swamp $c = 0.5$ tensor.

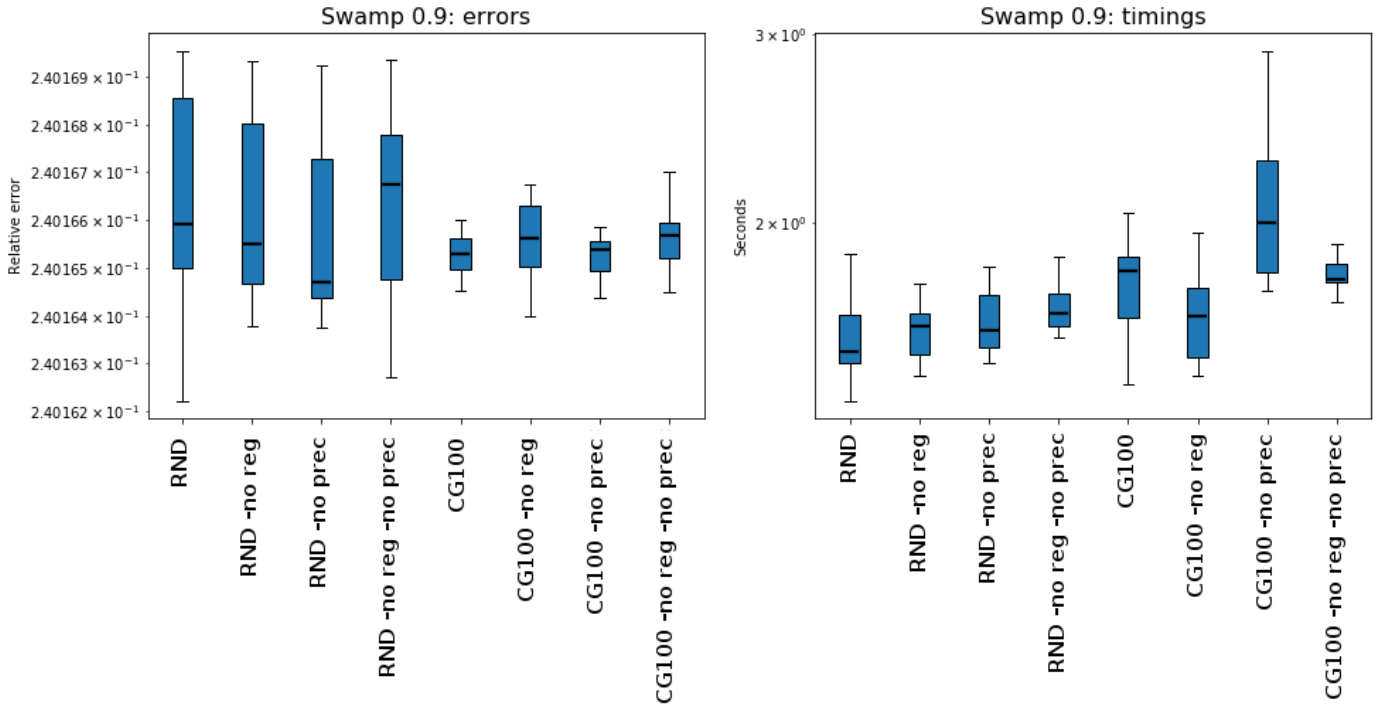

Figure 4.46: Box plots with errors and timings for the swamp $c = 0.9$ tensor.

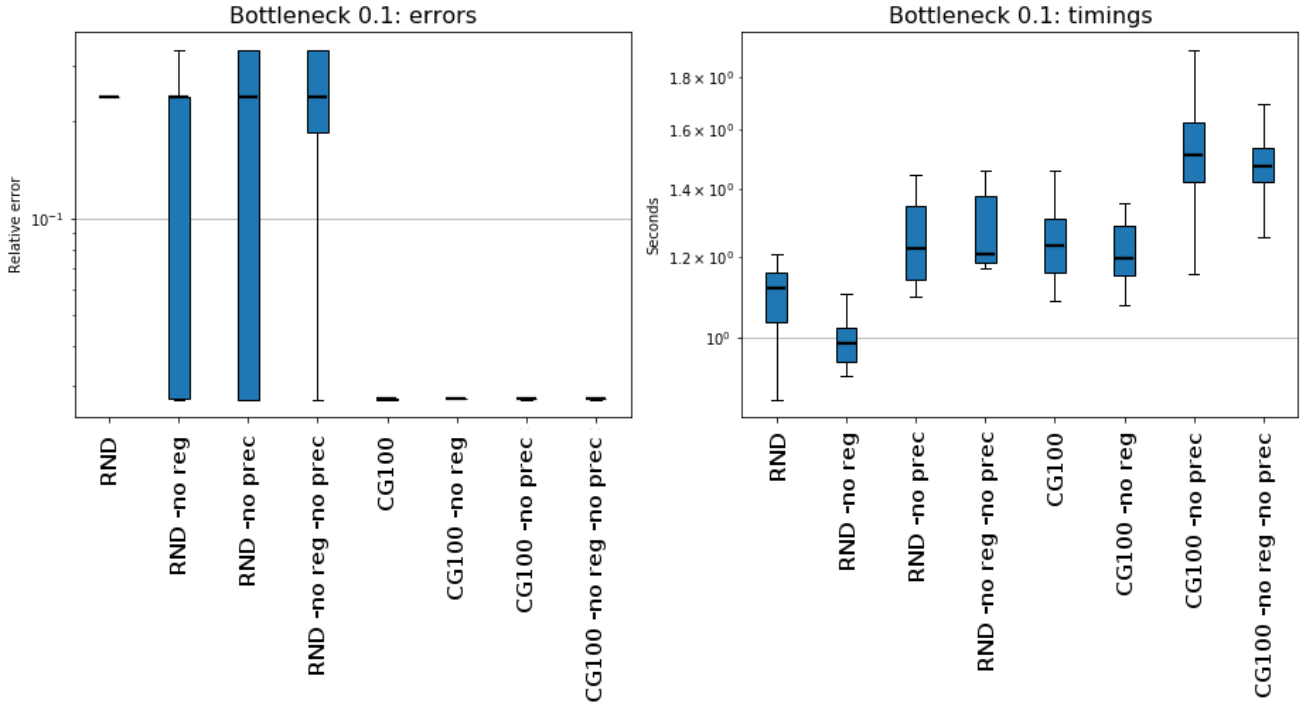

Figure 4.47: Box plots with errors and timings for the bottleneck $c = 0.1$ tensor.

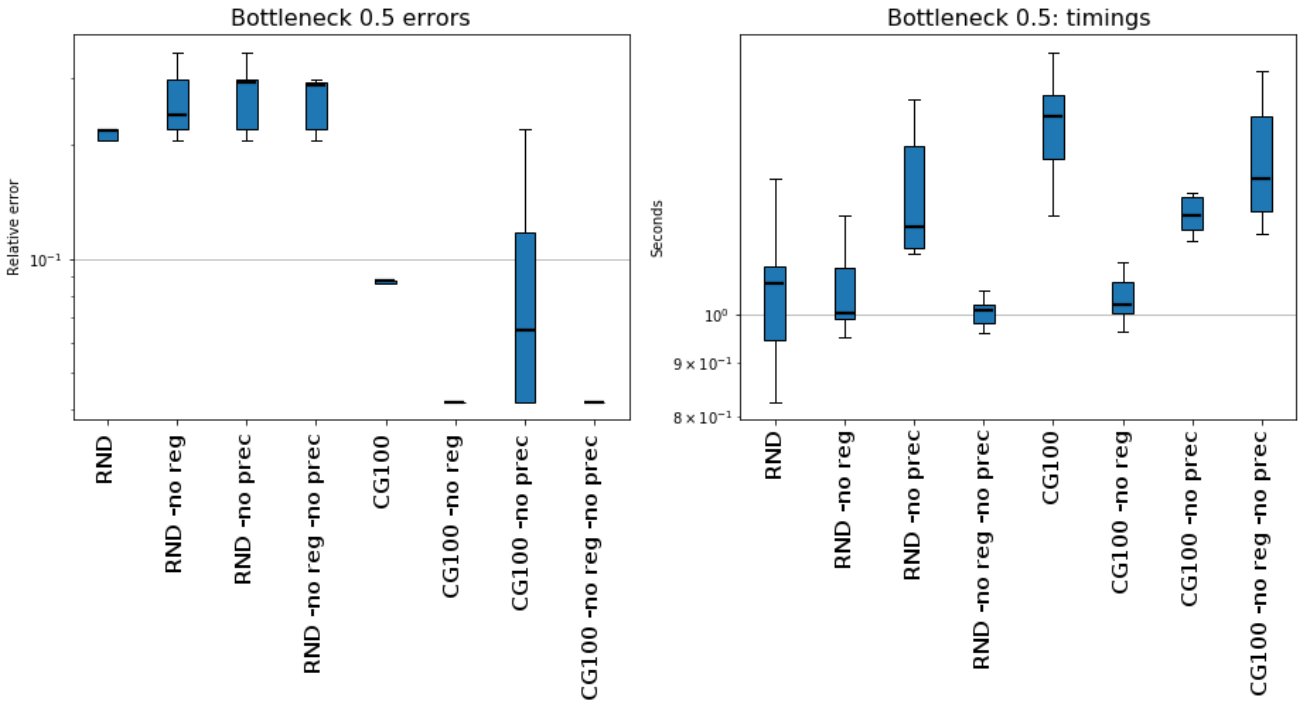

Figure 4.48: Box plots with errors and timings for the bottleneck $c = 0.5$ tensor.

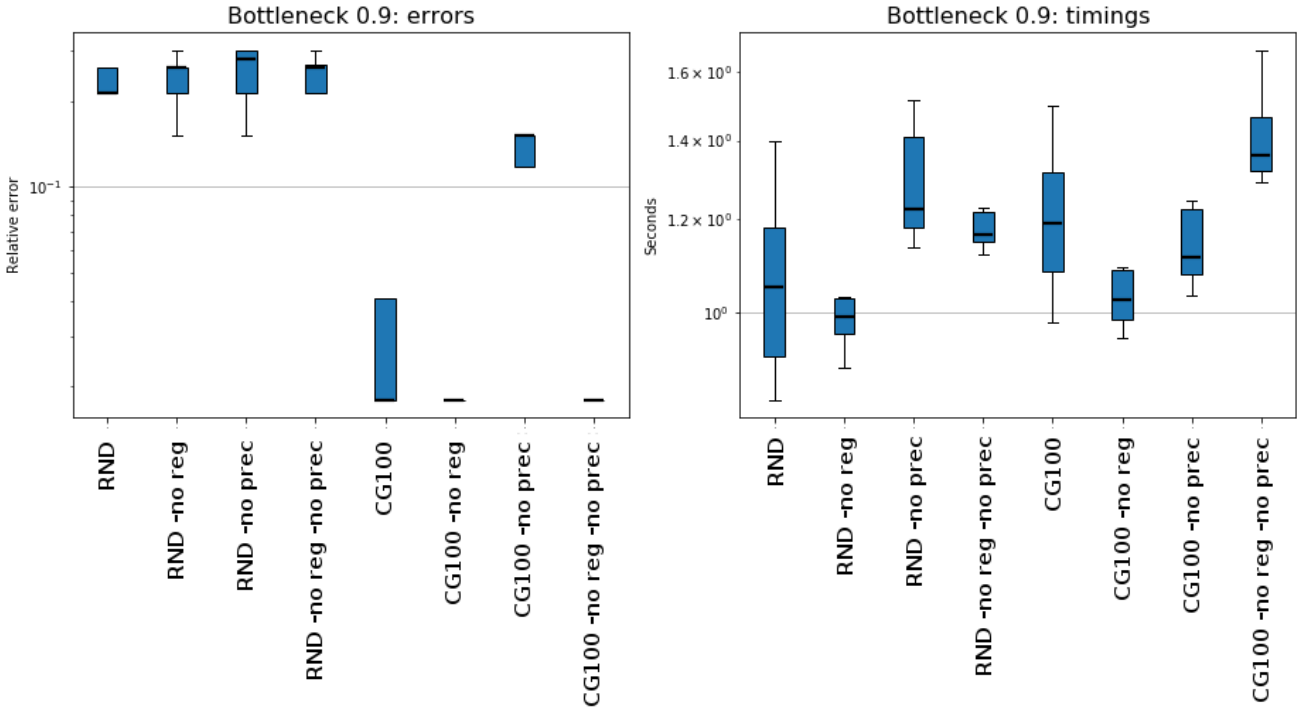

Figure 4.49: Box plots with errors and timings for the bottleneck $c = 0.9$ tensor.

The first thing to note is that **CG100** is not always more accurate than **RND**, that is, making more CG iterations does not necessarily improve the results. Also it is evident that all **CG100** based models are slower than their corresponding **RND** models. We observe

that the use of regularization is not always necessary, in most cases the performance is the same. The exception are the border rank tensor, where removing regularization made the accuracy much worse, and in all bottleneck tensors, where removing regularization increased the variance of the solutions. In fact, there are several other cases where the removal of regularization increases the variance of the errors. This is true for the **RND** and **CG100** models. Removing the preconditioning has the expected result, that is, the increase of the running time. Because the main role of the preconditioning is to decrease the number of CG iterations. In some cases removing the preconditioning even increased the error. We conclude that removing regularization is feasible sometimes depending on the tensor, and removing the preconditioning is never good. For this reason Tensor Fox has the option `init_damp`, which can be initialized to zero, thus removing the effect of regularization, but the preconditioning is always performed.

Chapter 5

Tensor learning

We call *tensor learning* any tensor technique applied to a machine learning problem. Here we will some of these techniques and their performance will be evaluated. The main tools are the MLSVD and the CPD. Usually the MLSVD is used to reduce dimensionality and the CPD is used to find latent parameters. In order to solve machine learning problems with tensor techniques, the biggest problem challenge is the modelling part, that is, transforming a machine learning problem into a tensor problem.

The main contribution is in 5.2, where a model of neural network based on the CPD is introduced. Although this neural network architecture still acts like a black box, the outputs consists of multilinear maps, which are more understandable than classical neural networks. The work here only takes in account dense neural networks, so the comparison is limited. There is still a lot of research to be done before discarding or not this new model, but for the moment this model seems to be reasonable.

5.1 Classification with the MLSVD

The MLSVD can be considered as a high order PCA. In this section we present a method to solve classification problems using the MLSVD, following [7] and [90]. The problem of handwritten digit classification will be used to test the performance of the method. We start with the same tensor described in 4.3.2 and apply some transformations to it in order to take the classes (the digits) in account. Each frontal slice (a 20×20 matrix) is vectorized to to form a column vector of size 400. Then we gather all these vectors side by side to form a 400×500 matrix corresponding to a certain class. In total we have 10 classes, each one corresponds to a frontal slices of the tensor $\mathcal{T} \in \mathbb{R}^{400\times 500\times 10}$. The first slice correspond to the digit 0, the second correspond to the digit 1, and so on. Figure 5.1 describes the procedure to form each frontal slice of $\mathcal{T}$.

Let $\mathcal{T} = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}) \cdot \mathcal{S}$ be a truncated MLSVD of $\mathcal{T}$, where $\mathbf{U}^{(1)} \in \mathbb{R}^{400\times R_1}$, $\mathbf{U}^{(2)} \in \mathbb{R}^{500\times R_2}$, $\mathbf{U}^{(3)} \in \mathbb{R}^{10\times R_3}$ and $\mathcal{S} \in \mathbb{R}^{R_1\times R_2\times R_3}$. Since we don't want to truncate the number of classes, it is necessary that $R_3 = 10$. Remember that $\mathbf{U}^{(1)}$ acts on the

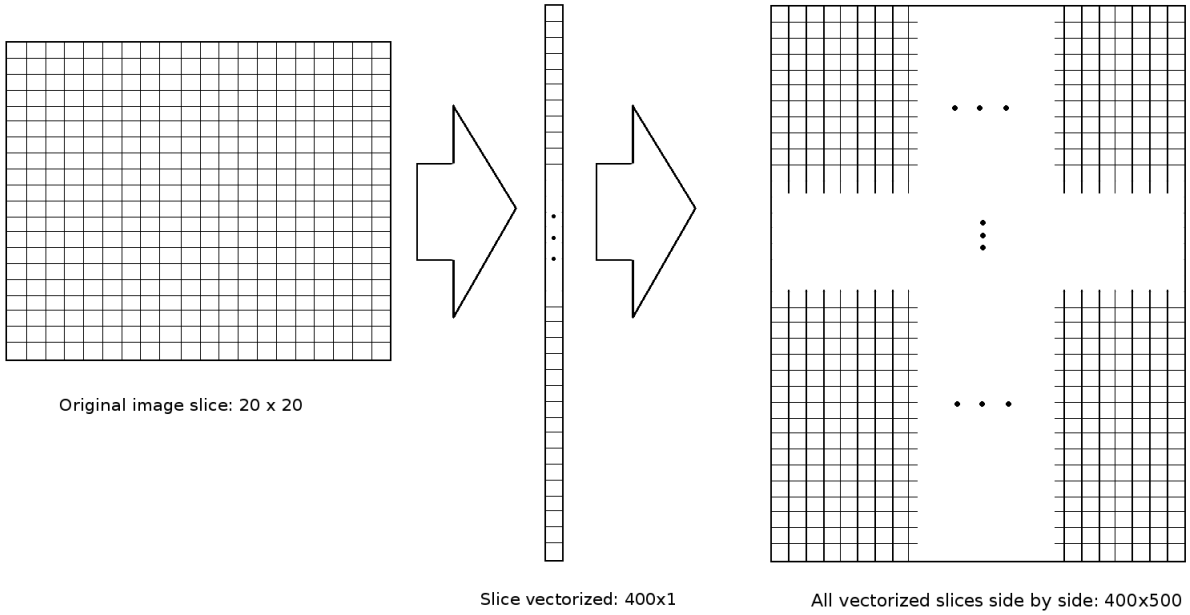

Figure 5.1: Transformation of a frontal slice into a vector and attaching dimension responsible to the class of the vector (the image represents the class $\mathbf{e}_2$, which corresponds to the digit 2).

mode-1 fibers of $\mathcal{T}$, which are the vectors $\mathcal{T}_{:jk}$. Each one of these vectors correspond to a vectorized image, so we can regard $\mathbf{U}^{(1)}$ as a compression matrix, which compress images from 400 pixels to R_1 pixels. The matrix $\mathbf{U}^{(2)}$ acts of the fibers $\mathcal{T}_{i:k}$, which correspond to the pixel i of all images in class k. This matrix "decides" if all images are really necessary to describe this pixel, so we can regard $\mathbf{U}^{(2)}$ as a matrix of redundancy detection, which reduces the number of images from 500 to R_2, for all classes. The matrix $\mathbf{U}^{(3)}$ is a square nonsingular so it doesn't change the dimension. It just makes some rotations of minor importance. We want to work with $\mathcal{S}$ but without the rotated classes. For this we just disconsider $\mathbf{U}^{(3)}$ from the definition of $\mathcal{S} = \left(\mathbf{U}^{(1)^T}, \mathbf{U}^{(2)^T}, \mathbf{U}^{(3)^T}\right) \cdot \mathcal{T}$. Thus we define the tensor $\mathcal{F} = \left(\mathbf{U}^{(1)^T}, \mathbf{U}^{(2)^T}, \mathbf{I}_{10}\right) \cdot \mathcal{T} \in \mathbb{R}^{R_1 \times R_2 \times 10}$. The training stage basically consists in computing the MLSVD and constructing the tensor $\mathcal{F}$.

The computation of a truncated MLSVD in Tensor Fox needs some value R for the rank as parameter, even if we are not interested in computing the CPD (as is the case here). The reason for this was mentioned in 4.1.1, we can't have $R_\ell > R$ for $\ell = 1, 2, 3$, and in order to assure this it is necessary to give a rank estimate. Furthermore, if the tensor has enough noise, it will be hard to truncate and the program will choose the truncation $R \times R \times 10$. After some tests we choose to use $R = 40$.

Now let's talk about the test stage. For each new image $\mathbf{z} \in \mathbb{R}^{400}$, convert it to the R_1 dimensional space using the transformation $\mathbf{z}_{new} = \mathbf{U}^{(1)^T} \cdot \mathbf{z}$. In this space we compare $\mathbf{z}_{new}$ to the space generated by the columns of each frontal slice of $\mathcal{F}$ (remember that each frontal slice correspond to a class). The best match will be the class of $\mathbf{z}_{new}$. This stage

of comparison is not made using the slice directly, but instead we compute the SVD of the slice and truncate it to another with lower rank. The overall procedure (training + test) is summarized in the steps below.

1. Compute truncated MLSVD $\mathcal{T} = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)}) \cdot \mathcal{S}$
2. Define $\mathcal{F} = \left(\mathbf{U}^{(1)^T}, \mathbf{U}^{(2)^T}, \mathbf{I}_{10}\right) \cdot \mathcal{T}$
3. Input: $\mathbf{z} \in \mathbb{R}^{400}$
4. Compress input: $\mathbf{z}_{new} = \mathbf{U}^{(1)^T} \cdot \mathbf{z} \in \mathbb{R}^{R_1}$
5. Compute SVD of slices: $\mathcal{F}_{::k} = \mathbf{W}^{(k)} \Sigma^{(k)} \mathbf{V}^{(k)^T}$ for $k = 1, \ldots, 10$
6. Truncate $\mathbf{W}^{(k)}$ to have $R_1' < R_1$ columns: $\tilde{\mathbf{W}}^{(k)} = \left[\mathbf{W}_1^{(k)}, \ldots, \mathbf{W}_{R_1'}^{(k)}\right]$
7. Solve least squares problems: $\min_{\mathbf{x}} \|\tilde{\mathbf{W}}^{(k)} \cdot \mathbf{x} - \mathbf{z}_{new}\|$ for $k = 1, \ldots, 10$
8. The class k associated with the smallest error (of all least squares problems solved) is chosen to be the class of $\mathbf{z}$

Note that, since the columns of $\tilde{\mathbf{W}}^{(k)}$ are orthonormal, the solution of the least squares problem is given by

$$\mathbf{x}_* = \tilde{\mathbf{W}}^{(k)^T} \cdot \mathbf{z}_{new},$$

which is easy to compute at this point. It should be noted that the matrices $\tilde{\mathbf{W}}^{(k)}$ doesn't have to be computed for each new input. Compute them just once and use them for all new incoming inputs. Finally, the choice of the SVD truncations R_1', R_2' depends on each case. There are some known procedures to guess good values but we won't address this issue here.

We used 2000 images for training and 3000 for testing. The dimensions of $\mathcal{S}$ are not bigger than $40 \times 40 \times 10$. Since the training data is $400 \times 200 \times 10$, we see that $100\left(1 - \frac{40 \cdot 40 \cdot 10}{400 \cdot 200 \cdot 10}\right) = 98\%$ of the data is compressed. Working with only 2% of the information may sound too few, but only with this much of data we could obtain more than 95% of accuracy in just a few seconds. Furthermore, this technique clearly can be generalized for practically any classification problem with minor changes. This example can be reproduced with the code in the following link: https://github.com/felipebottega/Tensor-Fox/blob/master/examples/handwritten_digit_mlsvd.ipynb.

5.2 Tensor learning vs. neural network

Given an input data $\mathbf{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, the usual neural network uses a certain number of layers with weights and activation functions in order to compute the respective

output $\mathbf{y} \in \mathbb{R}^m$. Each layer of weights forms is a matrix, usually bigger than $n \times n$ (except in the last layer). As an alternative to the neural network model, we propose to use m tensors $\mathcal{T}_1, \ldots, \mathcal{T}_m \in \mathbb{R}^{I_1 \times \ldots \times I_L}$, given by

$$\mathcal{T}_k = \sum_{r=1}^{R} \mathbf{w}_r^{(k,1)} \otimes \ldots \otimes \mathbf{w}_r^{(k,L)},$$

for $I_1 = I_2 = \ldots = I_L = n$ and $k = 1 \ldots m$.

Each $\mathbf{w}_r^{(k,\ell)} \in \mathbb{R}^n$ is a vector of weights. These weights are computed in the same way we compute the weights of a neural network, that is, by minimizing a cost function. In this model, the input-output function, also called the *hypothesis*, is the function $h : \mathbb{R}^n \to \mathbb{R}^m$ defined as

$$h(\mathbf{x}) = \mathbf{f} \begin{bmatrix} \mathcal{T}_1(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{L \text{ times}}) \\ \vdots \\ \mathcal{T}_m(\underbrace{\mathbf{x}, \ldots, \mathbf{x}}_{L \text{ times}}) \end{bmatrix} \tag{5.1}$$

where $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}$ is defined as $\mathbf{f} = (\underbrace{f, \ldots, f}_{m \text{ times}})$ for a nonlinear function $f : \mathbb{R} \to \mathbb{R}$. If one wants to write 5.1 in explicit form, then we have that

$$h(\mathbf{x}) = \mathbf{f} \begin{bmatrix} \sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \mathbf{w}_r^{(1,\ell)}, \mathbf{x} \rangle \\ \vdots \\ \sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \mathbf{w}_r^{(m,\ell)}, \mathbf{x} \rangle \end{bmatrix} =$$

$$= \begin{bmatrix} f\left(\sum_{r=1}^{R} \sum_{i_1,\ldots,i_L=1}^{n} w_{i_1 r}^{(1,1)} \ldots w_{i_\ell r}^{(1,L)} x_{i_1} \ldots x_{i_L} \right) \\ \vdots \\ f\left(\sum_{r=1}^{R} \sum_{i_1,\ldots,i_L=1}^{n} w_{i_1 r}^{(m,1)} \ldots w_{i_\ell r}^{(m,L)} x_{i_1} \ldots x_{i_L} \right) \end{bmatrix}.$$

Notice that $rank(\mathcal{T}_k) \leq R$ by construction. As is usual in machine learning algorithms, we use the convention $\mathbf{x} = [1, x_2, \ldots, x_n]^T$, that is, $x_1 = 1$ is the bias of the model. Now let $S = \{(i_1, \ldots, i_L) : 1 \leq i_1, \ldots, i_L \leq n,\ (i_1, \ldots, i_L) \neq (1, \ldots, 1)\}$, then we can write the above expression as

$$\begin{bmatrix} f\left(b_1 + \sum_{r=1}^{R} \sum_{(i_1,\ldots,i_L) \in S} w_{i_1 r}^{(1,1)} \ldots w_{i_L r}^{(1,L)} x_{i_1} \ldots x_{i_L} \right) \\ \vdots \\ f\left(b_m + \sum_{r=1}^{R} \sum_{(i_1,\ldots,i_L) \in S} w_{i_1 r}^{(m,1)} \ldots w_{i_L r}^{(m,L)} x_{i_1} \ldots x_{i_L} \right) \end{bmatrix}$$

where each $b_k = \sum_{r=1}^{R} (w_{1r}^{(k,1)})^L$ is the bias parameter of the model.

In a neural network, each activation function receives a weighted sum of the form

$$\sum_{j=1}^{n} w_j x_j,$$

whereas in the tensor learning we have multilinear weighted sums of the form

$$\sum_{i_1,\ldots,i_L=1}^{n} w_{i_1 r}^{(k,1)} \ldots w_{i_L r}^{(k,L)} x_{i_1} \ldots x_{i_L}$$

and just one activation per output. Contrary to the neural network model where we have lots of activations and few direct interactions between the weights[1], in the tensor learning approach we have few activations and lots of direct interactions between the weights, through the multilinearity. Our hope is that this multilinearity compensates the lack of activation functions.

5.2.1 Tensor learning as a special neural network

For a fixed input, note that each multilinear weighted sum shown above is just a polynomial of degree L with some weights as variables. This polynomial is such that all monomials have degree L. From this point of view one can interpret this model as a neural network (with no hidden layers) where we just changed from dot products to polynomials, a specific class of polynomials. Although this is a totally valid interpretation, it misses the importance of what these polynomials represent. They are tensors evaluated at the input, and their structure is such that all weights are directly mixed during the learning, whereas in the classic neural network[2] the weights are not so much mixed. This lack of interaction between the weights in the classic neural network is addressed by the tensor learning.

It should be noted that this model probably wouldn't be noticed by the usual machine learning practitioner even if it is desired to use polynomial weights as input to the activation functions. The usual way to model weights with polynomials is by introducing a weight per monomial. Therefore, one used to work with machine learning models would probably come with a polynomial of the form

$$\sum_{i_1,\ldots,i_L=1}^{n} w_{i_1 \ldots i_L} x_{i_1} \ldots x_{i_L},$$

for $w_{i_1 \ldots i_L} \in \mathbb{R}$. This would cause a lost of interactions between the weights. The tensor approach is not one usually would think, unless one is working with tensors, as it is the

[1] By *direct interactions* we mean multiplications between the weights.

[2] By *classic neural network* I mean the neural network model in which only dot products are passed to the activation functions.

case of this thesis.

It is possible to interpret tensor learning as a neural network (with 2 hidden layers) which relies almost entirely on the classic dot product plus activation function recipe. This interpretation shades more light on what is really happening and why tensor learning may be a valid alternative to the classic neural network. Figure 5.2 shows the architecture of such interpretation for $n = 3, m = 2, L = 2, R = 2$. The red lines represents classic neural network connections, that is, each line has a weight which is multiplied by the input and summed with all other lines reaching the same neuron. Blue lines means to multiply the inputs by the inputs of all other lines reaching the same neuron. The first layer of lines contains all the weights in our model. The second layer of lines makes the multiplications, and is in this part where we the model is no more the classic one. The third layer of lines are classic neural network connections with all weights equal to 1. Except for the last layer of neurons (the output layer), all other neurons have activation function equals the identity function. In the last layer the activation function is the nonlinear function f. We can note that the blue lines are the great novelty in this approach compared to the classic neural network. They represent the part where the weights are mixed together. Also note that this special neural network has $mRL + mR + m$ neurons. This usually will be much smaller than any neural network applied to the same problem, because a neural network uses $\mathcal{O}(n)$ neurons at the first layers, and we almost always have that $m, R, L \ll n$.

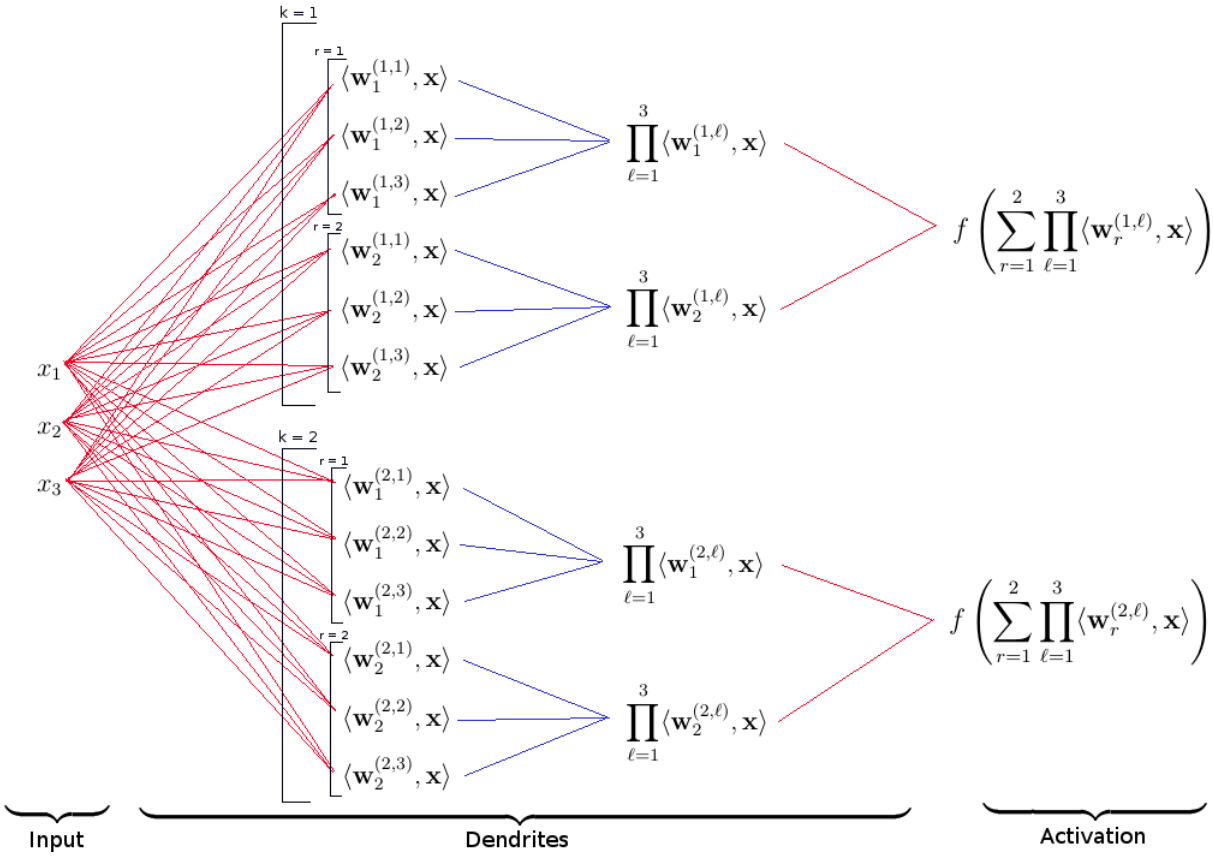

Figure 5.2: Tensor learning as a special neural network.

One could argue that these activations are too artificial and doesn't match any notion of a neuron. A new, and very different interpretation, is given in figure 5.2. The idea is that the dendrites intercept sometimes before the layer of activations, which is the last one. Each node of the dendrite layers correspond to a point of intersection between the dendrites. The information travels through a dendrite while also touching other dendrites, and this causes the interaction between the weights and the data. When it finally gets to the activation, we will have a multilinear combination to process. This interpretation correspond to a scenario where the dendrites path are not isolated, they touch each other several times and the information between one neuron and the other is affected by these interactions.

5.2.2 Cost function

Denote $\mathbf{W} = \left(\mathbf{W}^{(1,1)}, \dots, \mathbf{W}^{(1,L)}, \dots, \mathbf{W}^{(m,1)}, \dots, \mathbf{W}^{(m,L)}\right)$, where we have that $\mathcal{T}_k = \left(\mathbf{W}^{(k,1)}, \dots, \mathbf{W}^{(k,L)}\right) \cdot \mathcal{I}_{R\times\dots\times R}$, using the multilinear multiplication notation. Let $(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})$, $j = 1 \dots N$, be the training dataset, where we expect to have $n, m, L, R \ll N$. We will use the classic cost function $J : \mathbb{R}^{LmnR} \to \mathbb{R}$ defined as

$$J(\mathbf{W}) = \frac{1}{2N} \sum_{j=1}^{N} \left\| h(\mathbf{x}^{(j)}) - \mathbf{y}^{(j)} \right\|^2 =$$

$$= \frac{1}{2N} \sum_{j=1}^{N} \sum_{k=1}^{m} \left(h_k(\mathbf{x}^{(j)}) - y_k^{(j)} \right)^2 =$$

$$= \frac{1}{2N} \sum_{j=1}^{N} \sum_{k=1}^{m} \left(f\left(\sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \mathbf{w}_r^{(k,\ell)}, \mathbf{x}^{(j)} \rangle \right) - y_k^{(j)} \right)^2 .$$

Theorem 5.2.1. *For $k' = 1 \dots m$, $\ell' = 1 \dots L$, $i' = 1 \dots n$ and $r' = 1 \dots R$, we have that*

$$\frac{\partial J}{\partial w_{i'r'}^{(k',\ell')}} = \frac{1}{N} \sum_{j=1}^{N} \nabla J_{k',\ell',i',r'}^{(j)},$$

where

$$\nabla J_{k',\ell',i',r'}^{(j)} = \left(f\left(\sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \boldsymbol{w}_r^{(k',\ell)}, \boldsymbol{x}^{(j)} \rangle \right) - y_{k'}^{(j)} \right) \cdot f'\left(\sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \boldsymbol{w}_r^{(k',\ell)}, \boldsymbol{x}^{(j)} \rangle \right) \cdot \left(x_{i'}^{(j)} \prod_{\ell=1, \ell\neq\ell'}^{L} \langle \boldsymbol{w}_{r'}^{(k',\ell)}, \boldsymbol{x}^{(j)} \rangle \right).$$

Proof:

$$\frac{\partial J}{\partial w_{i'r'}^{(k',\ell')}} = \frac{1}{2N} \sum_{j=1}^{N} \sum_{k=1}^{m} \frac{\partial}{\partial w_{i'r'}^{(k',\ell')}} \left(h_k(\mathbf{x}^{(j)}) - y_k^{(j)} \right)^2 =$$

$$= \frac{1}{2N}\sum_{j=1}^{N}\sum_{k=1}^{m}\left[2\left(h_k(\mathbf{x}^{(j)}) - y_k^{(j)}\right)\cdot\frac{\partial}{\partial w_{i'r'}^{(k',\ell')}}h_k(\mathbf{x}^{(j)})\right] =$$

$$= \frac{1}{2N}\sum_{j=1}^{N}\sum_{k=1}^{m}\left[\left(h_k(\mathbf{x}^{(j)}) - y_k^{(j)}\right)\cdot f'\left(\sum_{r=1}^{R}\prod_{l=1}^{\ell}\langle\mathbf{w}_r^{(k,l)},\mathbf{x}^{(j)}\rangle\right)\cdot\frac{\partial}{\partial w_{i'r'}^{(k',\ell')}}\left(\sum_{r=1}^{R}\prod_{l=1}^{\ell}\langle\mathbf{w}_r^{(k,l)},\mathbf{x}^{(j)}\rangle\right)\right] =$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left[\left(h_{k'}(\mathbf{x}^{(j)}) - y_{k'}^{(j)}\right)\cdot f'\left(\sum_{r=1}^{R}\prod_{l=1}^{\ell}\langle\mathbf{w}_r^{(k',l)},\mathbf{x}^{(j)}\rangle\right)\cdot\left(\frac{\partial}{\partial w_{i'r'}^{(k',\ell')}}\prod_{l=1}^{\ell}\langle\mathbf{w}_{r'}^{(k',l)},\mathbf{x}^{(j)}\rangle\right)\right] =$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left[\left(h_{k'}(\mathbf{x}^{(j)}) - y_{k'}^{(j)}\right)\cdot f'\left(\sum_{r=1}^{R}\prod_{l=1}^{\ell}\langle\mathbf{w}_r^{(k',l)},\mathbf{x}^{(j)}\rangle\right)\cdot\left(\frac{\partial}{\partial w_{i'r'}^{(k',\ell')}}\langle\mathbf{w}_{r'}^{(k',\ell')},\mathbf{x}^{(j)}\rangle\prod_{l=1,l\neq\ell'}^{\ell}\langle\mathbf{w}_{r'}^{(k',l)},\mathbf{x}^{(j)}\rangle\right)\right] =$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left[\left(h_{k'}(\mathbf{x}^{(j)}) - y_{k'}^{(j)}\right)\cdot f'\left(\sum_{r=1}^{R}\prod_{l=1}^{\ell}\langle\mathbf{w}_r^{(k',l)},\mathbf{x}^{(j)}\rangle\right)\cdot\left(x_{i'}^{(j)}\prod_{l=1,l\neq\ell'}^{\ell}\langle\mathbf{w}_{r'}^{(k',l)},\mathbf{x}^{(j)}\rangle\right)\right] =$$

$$= \frac{1}{N}\sum_{j=1}^{N}\left(f\left(\sum_{r=1}^{R}\prod_{l=1}^{\ell}\langle\mathbf{w}_r^{(k',l)},\mathbf{x}^{(j)}\rangle\right) - y_{k'}^{(j)}\right)\cdot f'\left(\sum_{r=1}^{R}\prod_{l=1}^{\ell}\langle\mathbf{w}_r^{(k',l)},\mathbf{x}^{(j)}\rangle\right)\cdot\left(x_{i'}^{(j)}\prod_{l=1,l\neq\ell'}^{\ell}\langle\mathbf{w}_{r'}^{(k',l)},\mathbf{x}^{(j)}\rangle\right) =$$

$$= \frac{1}{N}\sum_{j=1}^{N}\nabla J_{k',\ell',i',r'}^{(j)}. \qquad \square$$

5.2.3 Regularization

If one want to use regularization, just introduce a regularization parameter $\lambda > 0$ and work with the cost function

$$J(\mathbf{W}) = \frac{\lambda}{2N}\sum_{j=1}^{N}\|\mathbf{W}\|^2 + \frac{1}{2N}\sum_{j=1}^{N}\left\|h(\mathbf{x}^{(j)}) - \mathbf{y}^{(j)}\right\|^2.$$

Now we have that

$$\nabla J_{k',\ell',i',r'}^{(j)} =$$

$$= \lambda w_{i'r'}^{(k',\ell')} + \left(f\left(\sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \mathbf{w}_r^{(k',\ell)}, \mathbf{x}^{(j)} \rangle \right) - y_{k'}^{(j)} \right) \cdot f'\left(\sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \mathbf{w}_r^{(k',\ell)}, \mathbf{x}^{(j)} \rangle \right) \cdot \left(x_{i'}^{(j)} \prod_{\ell=1, \ell \neq \ell'}^{L} \langle \mathbf{w}_{r'}^{(k',\ell)}, \mathbf{x}^{(j)} \rangle \right).$$

5.2.4 Stochastic gradient

After some initial guess $\mathbf{W} = \left(\mathbf{W}^{(1,1)}, \ldots, \mathbf{W}^{(1,L)}, \ldots, \mathbf{W}^{(m,1)}, \ldots, \mathbf{W}^{(m,L)} \right)$, we could use the gradient descent iteration

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha \nabla J(\mathbf{W})$$

to obtain sequential improvements for the weights. This procedure is costly and we will prefer to use the stochastic gradient iteration

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha \nabla J^{(j)}(\mathbf{W}),$$

where $\nabla J^{(j)} = \left[\nabla J_{k,\ell,i,r}^{(j)} \right]_{k,\ell,i,r}$. After making this iteration for each $j = 1 \ldots N$, we probably will have a good set of weights. Otherwise we may repeat the procedure above sometimes. It should be that this model of learning, by construction, avoids the problem of vanishing gradients, which is a common problem in neural networks.

5.2.5 Computational cost

Now we analyze the cost of the stochastic gradient descent algorithm. Fix a sample $\mathbf{x}^{(j)}$. First we compute and save each $\langle \mathbf{w}_r^{(k,\ell)}, \mathbf{x}^{(j)} \rangle$. Since each one costs n flops and we have to compute it LmR times, we have a total cost of $LmnR$ flops. The computational cost of $\nabla J_{k,\ell,i,r}^{(j)}$ is summarized below.

$$\begin{aligned}
&\lambda w_{i'r'}^{(k',\ell')} : \quad 1 \text{ flop} \\
&f\left(\sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \mathbf{w}_r^{(k',\ell)}, \mathbf{x}^{(j)} \rangle \right) - y_{k'}^{(j)} : \quad \ell R + 1 \text{ flops} \\
&f'\left(\sum_{r=1}^{R} \prod_{\ell=1}^{L} \langle \mathbf{w}_r^{(k',\ell)}, \mathbf{x}^{(j)} \rangle \right) : \quad \ell R + 1 \text{ flops} \\
&\left(x_{i'}^{(j)} \prod_{\ell=1, \ell \neq \ell'}^{L} \langle \mathbf{w}_{r'}^{(k',\ell)}, \mathbf{x}^{(j)} \rangle \right) : \quad L \text{ flops} \\
&\text{-----------------------} \\
&\textbf{Total: } 1 + 2(LR + 1) + L \text{ flops}
\end{aligned}$$

To compute all $\nabla J^{(j)}_{k,\ell,i,r}$ we have make these computations $LmnR$ times, so we have a total of $LmnR + 2(LmnR)(LR+1) + L^2mnR$ flops. Taking in account also the computations of the dot products, we have a total of $3LmnR + 2L^2mnR^2 + L^2mnR$ flops. This is for one iteration of the stochastic gradient. Running the program over all the samples has a cost of $(3LmnR + 2L^2mnR^2 + L^2mnR)N$ flops. Assuming we have to repeat the whole process E times (number of epochs), we have a final cost of

$$(3LmnR + 2L^2mnR^2 + L^2mnR)NE \text{ flops.}$$

Now assume a neural network with $\tilde{L}$ layers and Cn neurons in each layer (except the last one), where $C > 1$. For each layer we have to compute Cn dot products (one per neuron), except the last layer which we have to compute only m dot products. This gives us a total of $Cn(\tilde{L}-1) + m$ dot products. Since each dot product costs n flops, this amounts to a total of $Cn^2(\tilde{L}-1) + mn\tilde{L}$ flops. We also have to evaluate as many function evaluations as neurons, which gives us a total of $Cn(\tilde{L}-1) + m$ flops. Therefore we need to perform

$$Cn^2(\tilde{L}-1) + mn\tilde{L} + Cn(\tilde{L}-1) + m = (n^2+n)C(\tilde{L}-1) + mn\tilde{L} + m \text{ flops}$$

in order to make the forward propagation of one sample. Running this over all the samples has a cost of $((n^2+n)C(\tilde{L}-1) + mn\tilde{L} + m)N$ flops. Assuming we have to repeat the whole process E times, we have a final cost of

$$((n^2+n)C(\tilde{L}-1) + mn\tilde{L} + m)NE \text{ flops.}$$

Below we put both costs together for comparison.

$$\textbf{Tensor learning: } (3LmnR + 2L^2mnR^2 + L^2mnR)NE \text{ flops}$$

$$\textbf{Neural network: } ((n^2+n)C(\tilde{L}-1) + mn\tilde{L} + m)NE \text{ flops}$$

We expect to have $m, R \ll n$ and $L < \tilde{L}$ so the cost of the tensor learning is, in general, much lower than the neural network. There is no rule of thumb about the number of layers and the number of neurons at each layer, but some reasonable values, for example, would be $R = 3, C = 2$ and $\tilde{L} = L = 3$. In this case we have the following costs.

$$\textbf{Tensor learning: } 3^3mn + 2 \cdot 3^4mn + 3^3mn = 216mn$$

$$\textbf{Neural network: } (n^2+n)2^2 + 3mn + m = 4n^2 + 4n + 3mn + m$$

After the weights are trained we have a CPD format $(\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}) \cdot \mathcal{I}_{R\times\ldots\times R}$ to makes predictions. As already observed, $rank(\mathcal{T}_k) \leq R$ by construction. Therefore it

is interesting to compute low rank CPD approximations for all the tensors $\mathcal{T}_k$, hence simplifying the model. We tested this formulation over the MNIST handwritten digits, the same tensor described in the previous section. Making $L = 3$ (third order tensors) and $R = 20$ we could find a set of weights which predicts the entire dataset with 97.8% of accuracy. After that we start computing low rank approximations and the new rank $R' = 17$ worked very well, producing a smaller set of weights which models the entire dataset with the practically the same accuracy (it achieved 97.628%).

Chapter 6

Conclusions

Chapters 1 and 2 were devoted to the study of tensor decompositions and some geometrical aspects of the tensor rank. In chapter 3 we introduced the Gauss-Newton algorithm and showed how it can be used to the problem of finding an approximated CPD. The block structure of the approximated Hessian is of big importance. The study of this structure allowed us to make fast matrix-vector products for the CG algorithm. Chapter 4 was devoted to the development, in details, of the algorithms used within Tensor Fox. These algorithms now are part of a tensor package which includes not only a CPD solver, but also many routines of multilinear algebra and machine learning tools. The CPD algorithm developed in this work was compared to the state of art and it proved to be competitive, both in terms of speed and accuracy. Finally, in chapter 5 we introduced the concept of tensor learning and showed how tensor techniques can be used to machine learning problems.

This thesis now is part of an undergoing project. The solver we develop is fast, accurate and scalable, but it still has room to improvements. Below we list the points to be addressed in the future.

- The diagonal preconditioner is certainly not the best preconditioner one could use to solve the subproblems of the dGN algorithm. For instance Tensorlab uses a block diagonal preconditioner, thus reducing the number of CG iterations much more times than Tensor Fox. The drawback of this approach is the necessity of solving more linear systems in order to obtain these preconditioners. It would be ideal to exploit the block structure of the approximated Hessian somehow to obtain a better preconditioner with few computations.

- The rule of the current damping parameter update doesn't seem to be optimal. Designing and testing new rule updates could bring more improvements.

- In recent work [42] it was verified that Tensorlab can produce very accurate solutions that are extremely ill-conditioned, even when the objective solution is well-

conditioned. In 4.9 we could verify that Tensor Fox suffers from the same problem. It is not a secret that problems in high-dimensional spaces usually are full of *degenerate saddle points* [11]. There is the possibility that both solvers are converging to these points, which happen to be close to the objective but are ill-conditioned. Since they are "looking" only for the forward error of the approximation, nothing are being made to avoid ill-conditioned solutions. Increasing the number of CG iterations could address this problem but this is costly.

- Currently Tensor Fox is able to handle large sparse tensors with low memory cost. It is of interest to make use of GPUs to perform some computations for these kind of tensors.

- Expanding the technique of tensor learning present in 5.2 can be fruitful for machine learning problems. Although the neural network architecture still acts like a black box, the outputs consists of multilinear maps, which are understandable. Additionally, comparing the tensor model with other kind of neural networks is something to be addressed. It is also of interest to test the tensor model with different orders, different datasets, a wider range of ranks and so on. In short, there are a lot to be done in this subject, and it is the hope of the author that this first step is a step in a rich direction

Appendix A

Numerical linear algebra

A.1 Singular value decomposition

In this section $\mathbb{K}$ can be $\mathbb{R}$ or $\mathbb{C}$. We start with the singular value decomposition (SVD). The condition $m \geq n$ won't be assumed.

Theorem A.1.1 (Reduced SVD)**.** *Let $\boldsymbol{M} \in \mathbb{K}^{m\times n}$. Then there exists unitary (orthogonal) matrices*[1] *$\boldsymbol{U} \in \mathbb{K}^{m\times n}$, $\boldsymbol{V} \in \mathbb{K}^{n\times n}$ and a diagonal matrix $\Sigma = diag(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$, with $\sigma_1 \geq \ldots \geq \sigma_n$, such that $\boldsymbol{M} = \boldsymbol{U}\Sigma\,\boldsymbol{V}^*$.*

The columns of $\mathbf{U}$ are called *left singular vector*, the columns of $\mathbf{V}$ are called *right singular vectors*, and the values σ_i are called *singular values*. The decomposition given in the theorem is the *reduced SVD of $\boldsymbol{M}$*. In the *full SVD of $\boldsymbol{M}$* we can write $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^*$, where $\mathbf{U} \in \mathbb{K}^{m\times m}$ and $\mathbf{V} \in \mathbb{K}^{n\times n}$ are unitary (orthogonal), and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}$, where $p = \min\{m, n\}$. Note that Σ has the same dimensions of $\mathbf{M}$ so we are calling it "diagonal" in accord with the structures below. There are three possible cases.

$\underline{m = n}$

$$\mathbf{M} = \mathbf{U} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \mathbf{V}^*$$

$\underline{m > n}$

$$\mathbf{M} = \mathbf{U} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & \\ & & \end{bmatrix} \mathbf{V}^*$$

[1] Here we are talking about rectangular unitary (orthogonal) matrices. There are two definitions for a rectangular matrix $\mathbf{U} \in \mathbb{K}^{m\times n}$. In the case $m \geq n$ we say $\mathbf{U}$ is unitary (orthogonal) if its columns are orthonormal (which is equivalent to $\mathbf{U}^*\mathbf{U} = \mathbf{I}_n$). In the case $m \leq n$ we say $\mathbf{U}$ is unitary (orthogonal) if its rows are orthonormal (which is equivalent to $\mathbf{U}\mathbf{U}^* = \mathbf{I}_m$).

$\underline{m < n}$

$$\mathbf{M} = \mathbf{U} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{bmatrix} \mathbf{V}^*$$

The following results will be useful.

Lemma A.1.2. *Let $\boldsymbol{M} \in \mathbb{K}^{m\times m}$ be symmetric with eigendecomposition $\boldsymbol{M} = \boldsymbol{U}\Lambda\boldsymbol{U}^*$ where $\boldsymbol{U} \in \mathbb{K}^{m\times m}$ is unitary (orthogonal) and $\Lambda = diag(\lambda_1, \ldots, \lambda_m) \in \mathbb{K}^{m\times m}$. Furthermore, assume the eigenvalues λ_i are such that $|\lambda_1| \geq \ldots \geq |\lambda_m|$. Then an SVD of $\boldsymbol{M}$ is $\boldsymbol{M} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$, where $\sigma_i = |\lambda_i|$ and $\boldsymbol{V}_{:i} = sign(\lambda_i)\boldsymbol{U}_{:i}$.*

Lemma A.1.3. *Let $\boldsymbol{M} \in \mathbb{K}^{m\times n}$ such that $m \leq n$ and let $\sigma_1 \geq \ldots \geq \sigma_m$ be the singular values of $\boldsymbol{M}$. Then the singular values of $\boldsymbol{M}\boldsymbol{M}^T$ are $\sigma_1^2, \ldots, \sigma_m^2$.*

Theorem A.1.4. *Let $\boldsymbol{M} \in \mathbb{K}^{m\times n}$ such that $m \leq n$ and let $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $\boldsymbol{M}\boldsymbol{M}^T$, where $|\lambda_1| \geq \ldots \geq |\lambda_m|$. Then the singular values of $\boldsymbol{M}$ are $\sqrt{|\lambda_1|} \geq \ldots \geq \sqrt{|\lambda_m|}$.*

Remark A.1.5. *This theorem is particularly useful when $m < n$. Suppose we want to compute a SVD of $\boldsymbol{M}$ in this situation. Then the usual cost is at least of $2m^2n + 2m^3$ flops (see lecture 31 of [73]). Instead of this we can compute $\boldsymbol{M}\boldsymbol{M}^T$ first, at the cost of m^2n flops. Then the cost of computing the eigenvalues of a $m \times m$ symmetric matrix is of $\mathcal{O}(m^3)$ flops with a low constant, giving a total of $\mathcal{O}(m^2n + m^3)$ flops.*

A.2 Conjugate gradient

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric positive definite and $\mathbf{b} \in \mathbb{R}^n$. Suppose we want to solve a linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{A.1}$$

for $\mathbf{x}$. The solution of this system is $\mathbf{x}_* = \mathbf{A}^{-1}\mathbf{b}$. The *i-th Krylov space generated by $\boldsymbol{A}$ and $\boldsymbol{b}$* is defined as

$$\mathcal{K}_i = \text{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{i-1}\mathbf{b}).$$

Note that $\mathcal{K}_i$ is a subspace of $\mathbb{R}^{n\times n}$. Moreover. the vectors generating the space are orthonormal so they form a orthonormal basis for $\mathcal{K}_i$. Any algorithm that constructs these vectors sequentially is called a *Krylov method*. One of the advantages of Krylov methods is that sometimes they can be carried on without actually storing the matrix $\mathbf{A}$. Depending on the structure of $\mathbf{A}$ it is possible to accomplish this. When this is possible we way the method is *matrix-free*. For example, one can see that theorem 3.5.10 allow us to compute the product $\mathbf{J}_f^T\mathbf{J}_f \cdot \mathbf{x}$ without needing to construct $\mathbf{J}_f^T\mathbf{J}_f$ explicitly.

The conjugate gradient method is a Krylov method made specifically to deal with linear systems with symmetric positive definite matrices. Given the system A.1, this method iteratively constructs a sequence of approximations $\mathbf{x}^{(i)}$ which minimizes $\varepsilon^{(i)} = \|\mathbf{x}_* - \mathbf{x}^{(i)}\|_{\mathbf{A}}$, the $\boldsymbol{A}$-*residual of* $\boldsymbol{x}^{(i)}$, where $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle}$. We won't go into the details of this method. This is just a brief introduction of the algorithm and some of its properties.

Algorithm A.2.1 (Conjugate gradient)**.**
Input: $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *symmetric positive definite,* $\boldsymbol{b} \in \mathbb{R}^n$

$\boldsymbol{x}^{(0)} \leftarrow 0 \in \mathbb{R}^n$
$\boldsymbol{r}^{(0)} \leftarrow \boldsymbol{b}$
$\boldsymbol{p}^{(0)} \leftarrow \boldsymbol{r}^{(0)}$
$i = 1$
`repeat`
$\alpha^{(i)} \leftarrow \frac{\|\boldsymbol{r}^{(i-1)}\|^2}{\|\boldsymbol{p}^{(i-1)}\|_{\boldsymbol{A}}^2}$
$\boldsymbol{x}^{(i)} \leftarrow \boldsymbol{x}^{(i-1)} + \alpha^{(i)}\boldsymbol{p}^{(i-1)}$
$\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{r}^{(i-1)} - \alpha^{(i)}\boldsymbol{A}\boldsymbol{p}^{(i-1)}$
$\beta^{(i)} \leftarrow \frac{\|\boldsymbol{r}^{(i)}\|^2}{\|\boldsymbol{r}^{(i-1)}\|^2}$
$\boldsymbol{p}^{(i)} \leftarrow \boldsymbol{r}^{(i)} + \beta^{(i)}\boldsymbol{p}^{(i-1)}$
$i \leftarrow i + 1$ `until stopping criteria is met`

Output: $\boldsymbol{x}^{(i')}$, *where* i' *is the last index of the iterations*

Each iteration involves several vector manipulations and one matrix-vector multiplication, which appears twice in the algorithm but need to be computed only once. In general the matrix-vector multiplication dominates the computational cost so we can consider that each iteration costs $\mathcal{O}(n^2)$ flops. If $\mathbf{A}$ has some structure to be exploited this cost may be lower. The term "conjugate" in the name of the algorithm comes from the fact that the "search direction" vectors $\mathbf{p}^{(i)}$ are $\mathbf{A}$-conjugate with respect to each other, that is, $\langle \mathbf{p}^{(i)}, \mathbf{p}^{(j)}\rangle_{\mathbf{A}} = 0$ whenever $i \neq j$.

Usually one can use as stopping condition the distance between the current approximation, $\mathbf{x}^{(i)}$, and the objective solution, $\mathbf{b}$. The error is measured by $\varepsilon = \|\mathbf{b} - \mathbf{A}\mathbf{x}^{(i)}\|^2$, but a faster way to compute this error is through the identity $\varepsilon = \|\mathbf{r}^{(i)}\|^2$. To see this identity holds just note that

$$\varepsilon = \|\mathbf{r}^{(i)}\|^2 = \|\mathbf{r}^{(i-1)} - \alpha^{(i)}\mathbf{A}\mathbf{p}^{(i-1)}\|^2 =$$

$$= \|\mathbf{r}^{(i-2)} - \alpha^{(i-1)}\mathbf{A}\mathbf{p}^{(i-2)} - \alpha^{(i)}\mathbf{A}\mathbf{p}^{(i-1)}\|^2 = \dots$$

$$\ldots = \|\mathbf{r}^{(0)} - \alpha^{(1)}\mathbf{A}\mathbf{p}^{(0)} - \alpha^{(2)}\mathbf{A}\mathbf{p}^{(1)} - \ldots - \alpha^{(i)}\mathbf{A}\mathbf{p}^{(i-1)}\|^2 =$$

$$= \left\| \mathbf{b} - \mathbf{A}\sum_{j=1}^{i} \alpha^{(j)}\mathbf{p}^{(j-1)} \right\|^2 = \left\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(i)}\right\|^2.$$

Theorem A.2.2. *Consider conjugate gradient algorihtm applied to the system A.1. If the iteration i has not already converged ($\boldsymbol{r}^{(i-1)} \neq 0$), then $\boldsymbol{x}^{(i)}$ is the unique point in $\mathcal{K}_i$ that minimizes $\|\varepsilon^{(i)}\|_{\boldsymbol{A}}$. The convergence is monotonic (that is, $\|\varepsilon^{(i)}\|_{\boldsymbol{A}} \leq \|\varepsilon^{(i-1)}\|_{\boldsymbol{A}}$) and $\varepsilon^{(i)} = 0$ is achieved for some $i \leq n$.*

This theorem highlights some of the most known properties of the conjugate gradient. The last property, in particular, assures the algorithm will find the exact solution so it can be regarded as a direct algorithm, not an iterative one. We say it is iterative because in practice (working with finite precision) this exact solution is not necessarily found in less than n iterations. Usually one just compute some iterations of the algorithm, stopping before if the residual increases, because the monotonic property also can be lost when working in finite precision.

As we will see now, the rate of convergence depends a lot on the distribution of the eigenvalues of $\mathbf{A}$. For the next theorem denote P_i as the set of univariate polynomials p with degree $\leq i$ and $p(0) = 1$. More precisely P_i is the set of polynomials of the form $a_i x^i + a_{i-1}x^{i-1} + \ldots + a_1 x + 1$. We also use the notation $\Lambda(\mathbf{A})$ to be the spectrum of $\mathbf{A}$.

Theorem A.2.3. *At iteration i of the conjugate gradient algorithm, we have*

$$\frac{\|\varepsilon^{(i)}\|_{\boldsymbol{A}}}{\|\varepsilon^{(0)}\|_{\boldsymbol{A}}} \leq \inf_{p \in P_i} \max_{\lambda \in \Lambda(\boldsymbol{A})} |p(\lambda)|.$$

Corollary A.2.4. *If $\boldsymbol{A}$ has at most i distinct eigenvalues, then the conjugate gradient converges in at most i iterations.*

This corollary basically says that matrices with repeated eigenvalues may perform better. the general idea is that if the eigenvalues of $\mathbf{A}$ are group in i small clusters, then we can expect to have convergence in at most i iterations. The corollary is the extreme case where these clusters collapse in just points.

Let $\kappa(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2$ be the condition number of $\mathbf{A}$, where $\|\cdot\|_2$ is the spectral norm. The next result shows that the error is also bounded by a function of the condition number.

Corollary A.2.5. *At iteration i of the conjugate gradient algorithm, we have*

$$\frac{\|\varepsilon^{(i)}\|_{\boldsymbol{A}}}{\|\varepsilon^{(0)}\|_{\boldsymbol{A}}} \leq 2\left(\frac{\sqrt{\kappa(\boldsymbol{A})} - 1}{\sqrt{\kappa(\boldsymbol{A})} + 1}\right)^i.$$

Since $\kappa(\mathbf{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}$, this theorem implies that the largest eigenvalue shouldn't be too far from the smallest. Again, this reinforce the idea that the eigenvalues should be grouped together. Finally, since $\frac{\sqrt{\kappa(\mathbf{A})}-1}{\sqrt{\kappa(\mathbf{A})}+1} \approx 1 - \frac{2}{\sqrt{\kappa(\mathbf{A})}}$, we have that the convergence to a specified tolerance is expected to be done in $\mathcal{O}(\sqrt{\kappa(\mathbf{A})})$ iterations. This is just an upper bound, convergence may be faster if the spectrum is clustered. For more information about this subject check lecture 38 of [73] and 6.6.3 of [74].

A.3 Preconditioning

We assume the reader is familiar with the subject of condition numbers in linear algebra. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b} \in \mathbb{R}^n$. Given a linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{A.2}$$

to be solved for $\mathbf{x}$, it is possible to have numerical problems depending on the conditioning of $\mathbf{A}$. If $\mathbf{A}$ is ill-conditioned there will be problems to compute an accurate solution regardless of the algorithm used. When facing this situation we can use the technique of *preconditioning* to transform the ill-conditioned problem A.2 into a well-conditioned one. First, suppose we have at our disposal an invertible matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ such that $\mathbf{M}$ is well-conditioned and $\mathbf{M}^{-1}\mathbf{A}$ is easily computable and well-conditioned. $\mathbf{M}$ is called the *preconditioner*. Note that the system

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b} \tag{A.3}$$

has the same solution of A.2. Therefore, we can solve this well-conditioned system instead of an ill-conditioned and obtain the desired solution.

We don't actually need to construct the inverse $\mathbf{M}^{-1}$. First we solve the system $\mathbf{M}\mathbf{y} = \mathbf{b}$ which should be easy since $\mathbf{M}$ is well-conditioned. The solution is $\mathbf{y}_* = \mathbf{M}^{-1}\mathbf{b}$. With this solution now we solve the system

$$(\mathbf{M}^{-1}\mathbf{A})\mathbf{x} = \mathbf{y}_*. \tag{A.4}$$

We assume that $\mathbf{M}^{-1}\mathbf{A}$ is already computed and that it is a well behaved matrix so we can use it to solve the system without any problems. The respective solution is $\mathbf{x}_* = \mathbf{M}^{-1}\mathbf{A})^{-1}\mathbf{y}_* = \mathbf{A}^{-1}\mathbf{M}\mathbf{M}^{-1}\mathbf{b} = \mathbf{A}^{-1}\mathbf{b}$, which is the desired solution. This is the procedure of transforming and solving an ill-conditioned problem into a well-conditioned one. In the case $\mathbf{M}$ is diagonal we don't need to worry since $\mathbf{M}^{-1}$ is very easy to obtain.

This approach has to be changed when $\mathbf{A}$ is symmetric positive definite and we want to use the conjugate gradient algorithm. The reason for this is because $\mathbf{M}^{-1}\mathbf{A}$ is generally not symmetric even if $\mathbf{M}$ is symmetric. To derive the correct preconditioning assume $\mathbf{M}^{-1}$ is

symmetric positive definite. We can rewrite the preconditioned system $\mathbf{M}^{-1}\mathbf{A}\mathbf{x} = \mathbf{M}^{-1}\mathbf{b}$ as

$$\underbrace{(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2})}_{\hat{\mathbf{A}}}\underbrace{(\mathbf{M}^{1/2}\mathbf{x})}_{\hat{\mathbf{x}}} = \underbrace{\mathbf{M}^{-1/2}\mathbf{b}}_{\hat{\mathbf{b}}}. \tag{A.5}$$

Note that $\hat{\mathbf{A}}$ is symmetric positive definite so we can apply the conjugate gradient algorithm. Moreover $\hat{\mathbf{A}}$ and $\mathbf{M}^{-1}\mathbf{A}$ have the same eigenvalues since they are similar, therefore the performance of the gradient conjugate algorithm is not altered.

One of the simplest preconditioners is the *Jacobi preconditioner*, also called *diagonal preconditioner*. Its defined as $\mathbf{M} = \text{diag}(a_{ii})$, where we suppose $\mathbf{A}$ has positive diagonal. The result of this preconditioning is the matrix

$$\mathbf{M}^{-1}\mathbf{A} = \begin{bmatrix} 1 & \frac{a_{12}}{a_{11}} & \dots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & 1 & \dots & \frac{a_{2n}}{a_{22}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{a_{nn}} & \frac{a_{n2}}{a_{nn}} & \dots & 1 \end{bmatrix}.$$

This is particularly useful if $\mathbf{A}$ is diagonally dominant. Now suppose $\mathbf{A}$ is symmetric positive definite. Then we have

$$\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2} = \begin{bmatrix} 1 & \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}} & \dots & \frac{a_{1n}}{\sqrt{a_{11}}\sqrt{a_{nn}}} \\ \frac{a_{21}}{\sqrt{a_{22}}\sqrt{a_{11}}} & 1 & \dots & \frac{a_{2n}}{\sqrt{a_{22}}\sqrt{a_{nn}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{a_{n1}}{\sqrt{a_{nn}}\sqrt{a_{11}}} & \frac{a_{n2}}{\sqrt{a_{nn}}\sqrt{a_{22}}} & \dots & 1 \end{bmatrix}.$$

Even if $\mathbf{A}$ is not diagonally dominant this preconditioner may improve the conditioning by lowering the condition number. It can be shown [76, 77] that

$$\kappa(\mathbf{A}) \leq n \cdot \kappa(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2})$$

and

$$\kappa(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}) \leq n \cdot \min_{\mathbf{D}\in\mathbf{D}_n} \kappa(\mathbf{D}\mathbf{A}\mathbf{D}),$$

where $\kappa(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2$ and $\mathbf{D}_n$ is the set of nonsingular diagonal matrices. With this result we have a lower and an upper bound for $\kappa(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2})$, namely,

$$\frac{1}{n}\kappa(\mathbf{A}) \leq \kappa(\mathbf{M}^{-1/2}\mathbf{A}\mathbf{M}^{-1/2}) \leq \kappa(\mathbf{M}^{-1/2})^2 \cdot \kappa(\mathbf{A}) = \frac{\max_i a_{ii}}{\min_i a_{ii}} \cdot \kappa(\mathbf{A}).$$

A good survey on preconditioning I would recommend is [78], also part 6.6.5 of [74] is worth reading.

Appendix B

Tensor algebra

This appendix section can be viewed as a complement of chapter 1.

B.1 Tensor product properties

The main algebraic properties about tensor product are summarized below.

Theorem B.1.1. *Consider the vector spaces* $\mathbb{V}^{(1)}, \mathbb{V}^{(2)}, \mathbb{V}^{(1)}$*, the vectors* $\boldsymbol{v}^{(1)}, \boldsymbol{u}^{(1)} \in \mathbb{V}^{(1)}$*,* $\boldsymbol{v}^{(2)}, \boldsymbol{u}^{(2)} \in \mathbb{V}^{(2)}$*,* $\boldsymbol{v}^{(3)} \in \mathbb{V}^{(3)}$ *and a scalar* $\alpha \in \mathbb{K}$*. Then*

1. $\boldsymbol{v}^{(1)} \otimes \boldsymbol{v}^{(2)} = 0 \implies \boldsymbol{v}^{(1)} = 0$ *or* $\boldsymbol{v}^{(2)} = 0$*.*
2. $(\alpha \boldsymbol{v}^{(1)}) \otimes \boldsymbol{v}^{(2)} = \boldsymbol{v}^{(1)} \otimes (\alpha \boldsymbol{v}^{(2)}) = \alpha(\boldsymbol{v}^{(1)} \otimes \boldsymbol{v}^{(2)})$
3. $\boldsymbol{v}^{(1)} \otimes (\boldsymbol{v}^{(2)} + \boldsymbol{u}^{(2)}) = \boldsymbol{v}^{(1)} \otimes \boldsymbol{v}^{(2)} + \boldsymbol{v}^{(1)} \otimes \boldsymbol{u}^{(2)}$
4. $(\boldsymbol{v}^{(1)} + \boldsymbol{u}^{(1)}) \otimes \boldsymbol{v}^{(2)} = \boldsymbol{v}^{(1)} \otimes \boldsymbol{v}^{(2)} + \boldsymbol{u}^{(1)} \otimes \boldsymbol{v}^{(2)}$
5. $\mathbb{V}^{(1)} \otimes \mathbb{K} \cong \mathbb{V}^{(1)}$
6. $(\mathbb{V}^{(1)} \otimes \mathbb{V}^{(2)})^* \cong (\mathbb{V}^{(1)})^* \otimes (\mathbb{V}^{(1)})^*$
7. $\mathbb{V}^{(1)} \otimes \mathbb{V}^{(2)} \cong \mathbb{V}^{(2)} \otimes \mathbb{V}^{(1)}$
8. $(\mathbb{V}^{(1)} \otimes \mathbb{V}^{(2)}) \otimes \mathbb{V}^{(3)} \cong \mathbb{V}^{(1)} \otimes (\mathbb{V}^{(2)} \otimes \mathbb{V}^{(3)})$
9. $\boldsymbol{v}^{(1)} \otimes (\boldsymbol{v}^{(2)} \otimes \boldsymbol{v}^{(3)}) = (\boldsymbol{v}^{(1)} \otimes \boldsymbol{v}^{(2)}) \otimes \boldsymbol{v}^{(3)}$

Remark B.1.2. *The generalization of these results to more products instead 3 is immediate. With respect to property 7, it is important to note that it does not imply commutativity. In general tensor spaces are associative but not commutative. Finally, these*

properties implies that tensor spaces are also an algebra[1]*, and for this reason it is common to call them a* tensor algebra.

For $L = 1$ one can derive formula 1.3 directly from the computation of $\langle \mathbf{x}, \mathcal{T} \rangle_{\mathbb{R}}$ (identifying $\mathcal{T}$ with a vector), and in the case $L = 2$ the formula is derived from the computation of $\langle \mathcal{T}\mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}}$ (identifying $\mathcal{T}$ with a matrix). For $L = 3$ we want to use the coordinate representation of $\mathcal{T}$ (that is, identify $\mathcal{T}$ with a 3-D matrix) to obtain some kind of generalization. To accomplish this, let's take a step back and look at the case $L = 2$ again. Observe that it is possible to reduce the expression to a particular case of $L = 1$. More precisely, we can write

$$\langle \mathcal{T}\mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}} = \mathbf{y}^T \left(\sum_{j=1}^{m} x_j \mathcal{T}_{:j} \right) = \sum_{j=1}^{m} x_j (\mathbf{y}^T \mathcal{T}_{:j}) = \langle \mathbf{x}, \mathbf{w} \rangle_{\mathbb{R}},$$

where $\mathbf{w} = \left[\mathbf{y}^T \mathcal{T}_{:1}, \ldots, \mathbf{y}^T \mathcal{T}_{:m} \right]^T$. The idea is to compute the Euclidean inner product between $\mathbf{y}$ and the columns of $\mathcal{T}$ and use the results to construct a vector, which have the same size of $\mathbf{x}$. Then we compute the Euclidean inner product between $\mathbf{x}$ and this vector. We can obtain the same result when starting with $\mathbf{x}$ instead of $\mathbf{y}$, but in this case we compute the inner product between the rows of $\mathcal{T}$. In the case $L = 3$ we proceed similarly, first we fix all dimensions of $\mathcal{T}$, except the one associated to $\mathbf{x}$ (the first dimension), which will vary to create a vector $\mathcal{T}_{:jk}$. Then we compute the inner product between $\mathbf{x}$ and this vector. After doing this for all possible vectors, we use the results of these computations to produce a matrix such that the entry (j, k) is $\langle \mathbf{x}, \mathcal{T}_{:jk} \rangle_{\mathbb{R}}$, in which we repeat the process using the previous procedure ($L = 2$) to obtain a scalar. The ordering in which we use the vectors doesn't affect the final result, so we will work first with $\mathbf{x}$, then $\mathbf{y}$ and then $\mathbf{z}$. After computing all inner products $\langle \mathbf{x}, \mathcal{T}_{:jk} \rangle_{\mathbb{R}}$, we construct the matrix

$$\mathbf{W} = \begin{bmatrix} \langle \mathbf{x}, \mathcal{T}_{:11} \rangle_{\mathbb{R}} & \ldots & \langle \mathbf{x}, \mathcal{T}_{:1p} \rangle_{\mathbb{R}} \\ \vdots & & \vdots \\ \langle \mathbf{x}, \mathcal{T}_{:n1} \rangle_{\mathbb{R}} & \ldots & \langle \mathbf{x}, \mathcal{T}_{:np} \rangle_{\mathbb{R}} \end{bmatrix}$$

respecting the ordering of the indexes. Now we repeat the process for $\mathbf{y}$, that is, compute

[1] An *algebra* is a vector spaces V with a multiplication $\cdot$ satisfying the following properties:

$$(\mathbf{x} + \mathbf{y}) \cdot \mathbf{z} = \mathbf{xz} + \mathbf{yz} \quad \text{(right distributivity)}$$

$$\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{xy} + \mathbf{xz} \quad \text{(left distributivity)}$$

$$(\alpha \mathbf{x}) \cdot (\beta \mathbf{y}) = (\alpha\beta)(\mathbf{x} \cdot \mathbf{y}) \quad \text{(compatibility with scalars)}$$

all inner products $\langle \mathbf{y}, \mathbf{W}_{:k} \rangle_{\mathbb{R}}$ and form the vector

$$\mathbf{w} = \begin{bmatrix} \langle \mathbf{y}, \mathbf{W}_{:1} \rangle_{\mathbb{R}} \\ \vdots \\ \langle \mathbf{y}, \mathbf{W}_{:p} \rangle_{\mathbb{R}} \end{bmatrix}.$$

Finally, compute the inner product $\langle \mathbf{z}, \mathbf{w} \rangle_{\mathbb{R}}$. It is not difficult to verify that this value is equal to $\mathcal{T}(\mathbf{x}, \mathbf{y}, \mathbf{z})$. This procedure can be generalized for any L, the idea is always to compute the inner products with respect to one dimension, thus cutting the order by one. This is repeated until order one, when we compute a single inner product. We can formalize this procedure with the following definition.

Definition B.1.3. *Let two tensors $\mathcal{T} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_L}$ and $\mathcal{W} \in \mathbb{K}^{J_1} \otimes \ldots \otimes \mathbb{K}^{J_M}$ such that $\mathcal{T} = \boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)}$, $\mathcal{W} = \boldsymbol{w}^{(1)} \otimes \ldots \otimes \boldsymbol{w}^{(M)}$ and $\dim(\mathbb{K}^{I_\ell}) = \dim(\mathbb{K}^{I_m})$ for some ℓ, m. Then the $\times_\ell^m$-contraction between $\mathcal{T}$ and $\mathcal{W}$ is the tensor $\mathcal{T} \times_m^\ell \mathcal{W} \in \mathbb{K}^{I_1} \otimes \ldots \otimes \mathbb{K}^{I_{\ell-1}} \otimes \mathbb{K}^{I_{\ell+1}} \otimes \ldots \otimes \mathbb{K}^{I_L} \otimes \mathbb{K}^{J_1} \otimes \ldots \otimes \mathbb{K}^{J_{m-1}} \otimes \mathbb{K}^{J_{m+1}} \otimes \ldots \otimes \mathbb{K}^{J_M}$ defined as*

$$\mathcal{T} \times_\ell^m \mathcal{W} = \langle \boldsymbol{v}^{(\ell)}, \boldsymbol{w}^{(m)} \rangle_{\mathbb{R}} \, \boldsymbol{v}^{(1)} \otimes \ldots \otimes \boldsymbol{v}^{(\ell-1)} \otimes \boldsymbol{v}^{(\ell+1)} \otimes \ldots \otimes \boldsymbol{v}^{(L)} \otimes \boldsymbol{w}^{(1)} \otimes \ldots \otimes \boldsymbol{w}^{(m-1)} \otimes \boldsymbol{w}^{(m+1)} \otimes \ldots \otimes \boldsymbol{w}^{(L)}.$$

Consider the tensor $\mathcal{T}$ of the definition above. Given L vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)}$ we can use equation 1.5 and the definition above to write

$$\begin{aligned} \mathcal{T}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)}) &= \mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)}) = \\ &= \langle \mathbf{x}^{(1)}, \mathbf{v}^{(1)} \rangle_{\mathbb{R}} \cdot \ldots \cdot \langle \mathbf{x}^{(L)}, \mathbf{v}^{(L)} \rangle_{\mathbb{R}} = \\ &= \mathcal{T} \times_1^1 \mathbf{x}^{(1)} \times_2^1 \mathbf{x}^{(2)} \times_3^1 \ldots \times_L^1 \mathbf{x}^{(L)}. \end{aligned} \tag{B.1}$$

For a generic tensor this gives an immediate expression for $\mathcal{T}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)})$ since it is a linear combination of tensors of the form $\mathbf{e}_{i_1}^{(1)} \otimes \ldots \otimes \mathbf{e}_{i_L}^{(L)}$, whose contraction is described by the formula B.1.

Another useful definition is the *mode ℓ tensor-vector* product, which is a contraction along one dimension. Given a vector $\mathbf{x} \in \mathbb{R}^{I_\ell}$, we define $\mathcal{T} \times_\ell \mathbf{x} \in \mathbb{R}^{I_1 \times \ldots \times I_{\ell-1} \times I_{\ell+1} \times \ldots \times I_L}$ by the relation

$$\mathcal{T} \times_\ell \mathbf{x} = \left(\mathbf{I}_1, \ldots, \mathbf{I}_{\ell-1}, \mathbf{x}^T, \mathbf{I}_{\ell+1}, \ldots, \mathbf{I}_L\right) \cdot \mathcal{T}.$$

B.2 Rank properties

We begin with some basic result regard to the tensor rank.

Theorem B.2.1. *Let $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ and $\mathcal{S} \in \mathbb{W}^{(1)} \otimes \ldots \otimes \mathbb{W}^{(M)}$. Then the following holds.*

1. *If* $rank(\mathcal{S}) = 1$, *then* $rank(\mathcal{T} \otimes \mathcal{S}) = rank(\mathcal{T})$.

2. *If* $rank(\mathcal{T}), rank(\mathcal{S}) > 1$, *then* $rank(\mathcal{T} \otimes \mathcal{S}) \leq rank(\mathcal{T}) \cdot rank(\mathcal{S})$.

3. *Let* $\mathcal{T} = \sum_{r=1}^{R} \boldsymbol{v}_r^{(1)} \otimes \ldots \otimes \boldsymbol{v}_r^{(L)}$ *be such that* $\boldsymbol{v}_1^{(\ell)}, \ldots, \boldsymbol{v}_R^{(\ell)} \in \mathbb{V}^{(\ell)}$ *are linearly independent vectors for each* $\ell = 1 \ldots L$. *Then* $rank(\mathcal{T}) = R$.

Note that in the last item of the theorem the hypothesis about the vectors is equivalent to saying that each factor matrix $\mathbf{V}^{(\ell)} = [\mathbf{v}_1^{(\ell)}, \ldots, \mathbf{v}_R^{(\ell)}]$ is of full rank.

It is of interest to understand how often the CPD is unique, and for this we will need to introduce other types of rank.

Definition B.2.2. *We say* R *is a* typical rank of $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ *if the set of rank-*R *tensors in this space has positive probability (positive Lebesgue measure).*

Definition B.2.3. *We say* R *is a* generic rank of $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ *if the set of rank-*R *tensors in this space has probability 1.*

These definitions assume that the tensors are being drawn according to a continuous probability distribution. Note that R is the generic rank if, and only if there is only one typical rank in $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$. In this case all other ranks occur with zero probability. It is important to call attention to the fact that either type of rank depends on the field $\mathbb{K}$. In [46] there is a example of a real third order tensor which has rank 3 over $\mathbb{R}$ and rank 2 over $\mathbb{C}$. Another big difference between the real case and the complex case is given by the following theorem.

Theorem B.2.4. *If* $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ *is a* $\mathbb{C}$*-vector space, then it has a single typical rank.*

As this result shows, complex tensor spaces only have one typical rank, which is the generic rank, whereas real tensor spaces usually have more than one typical rank, so they does not have a generic rank. In the real case, some authors calls the least typical rank as the generic rank, but we won't use this terminology here. There is only one more type of rank we need to introduce.

Example B.2.5. *Consider the space* $\mathbb{K}^m \otimes \mathbb{K}^n \otimes \mathbb{K}^p$. *Let* $\boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z}$ *be a rank one tensor such that* x_1, y_1, z_1 *are not zero. Then we can rewrite this tensor as*

$$\boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z} = [x_1, x_2, \ldots, x_m]^T \otimes [y_1, y_2, \ldots, y_n]^T \otimes [z_1, z_2, \ldots, z_p]^T =$$

$$= \left(x_1 \left[1, \frac{x_2}{x_1}, \ldots, \frac{x_m}{x_1} \right]^T \right) \otimes \left(y_1 \left[1, \frac{y_2}{y_1}, \ldots, \frac{y_n}{y_1} \right]^T \right) \otimes \left(z_1 \left[1, \frac{z_2}{z_1}, \ldots, \frac{z_p}{z_1} \right]^T \right) =$$

$$= (x_1 y_1 z_1) \cdot [1, \tilde{x}_1, \ldots, \tilde{x}_{m-1}]^T \otimes [1, \tilde{y}_1, \ldots, \tilde{y}_{n-1}]^T \otimes [1, \tilde{z}_1, \ldots, \tilde{z}_{p-1}]^T =$$

$$= \lambda \cdot [1, \tilde{x}_1, \ldots, \tilde{x}_{m-1}]^T \otimes [1, \tilde{y}_1, \ldots, \tilde{y}_{n-1}]^T \otimes [1, \tilde{z}_1, \ldots, \tilde{z}_{p-1}]^T .$$

With this we can conclude that any tensor of this form can be described using only $1 + (m-1) + (n-1) + (p-1)$ *parameters instead of* $m + n + p$ *if we used the original form. Now let* M_{ijk} *be the set of tensors* $\boldsymbol{x} \otimes \boldsymbol{y} \otimes \boldsymbol{z}$ *such that* $x_i \neq 0, y_j \neq 0, z_k \neq 0$*. We have that* M_{ijk} *has dimension* $1 + (m-1) + (n-1) + (p-1)$ *since we can parametrize it with* $1 + (m-1) + (n-1) + (p-1)$ *parameters and no less than that. Note that the set of all rank one tensors also has dimension* $1 + (m-1) + (n-1) + (p-1)$ *since this set is the (finite) union*

$$\bigcup_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n \\ 1 \leq k \leq p}} M_{ijk}$$

of sets with dimension $1 + (m-1) + (n-1) + (p-1)$.

It is not hard to generalize this example to the space $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ and then conclude that the set of rank one tensors has dimension $1 + (I_1 - 1) + \ldots + (I_L - 1) = 1 + \sum_{\ell=1}^{L} (I_\ell - 1)$. Given a value $1 \leq R \leq \prod_{\ell=1}^{L} I_\ell$ for the rank, the set of rank-R tensors can be written as the sum of R copies of the set of rank one tensors. Therefore we could conclude that the dimension of this set is $R \left(1 + \sum_{\ell=1}^{L} (I_\ell - 1) \right)$. The parametrization given is not necessarily optimal, it may be possible to obtain other with less parameters. The best we can say is that this parametrization may be close to optimal, so we concluded that the dimension is close to the actual one. Furthermore, note that this set actually contains not only the rank-R tensors, it also contains all tensors of rank $\leq R$.

Denote $\sigma_R \left(\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)} \right) = \left\{ \mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)} : \ rank(\mathcal{T}) \leq R \right\}$ for the set of tensors with rank $\leq R$. If the space is understood from the context we may just denote σ_R. With this notation and the previous observation about the parametrization one can write

$$\dim(\sigma_R) \lessapprox R \left(1 + \sum_{\ell=1}^{L} (I_\ell - 1) \right).$$

We are talking about dimension of sets in a informal manner to refer to the number of parameters necessary to parameterize the set but in general it is not true that σ_R is a manifold. Now suppose R is such that our parametrization is optimal and it generates

the entire space. Then we have

$$\dim(\sigma_R) = \dim\left(\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}\right) \implies$$

$$\implies R\left(1 + \sum_{\ell=1}^{L}(I_\ell - 1)\right) = \prod_{\ell=1}^{L} I_\ell \implies R = \left\lceil \frac{\prod_{\ell=1}^{L} I_\ell}{1 + \sum_{\ell=1}^{L}(I_\ell - 1)} \right\rceil.$$

By definition, this R is a good candidate to be the generic rank of $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$. This motivates the following definition.

Definition B.2.6. *The* expected generic rank of $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ *is defined as being the value*

$$R_E = \left\lceil \frac{\prod_{\ell=1}^{L} I_\ell}{1 + \sum_{\ell=1}^{L}(I_\ell - 1)} \right\rceil.$$

Now we are ready to use these definitions to state some useful theorems about tensor rank.

Theorem B.2.7. *Let $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ be such that $I_1 \geq I_2 \geq \ldots \geq I_L$ and let R be its least typical rank. Then $rank(\mathcal{T}) \leq \min\left\{\prod_{\ell=2}^{L} I_\ell, 2R\right\}$ for all $\mathcal{T} \in \mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$.*

Theorem B.2.8. *Let $\mathbb{C}^n \otimes \mathbb{C}^n \otimes \mathbb{C}^n$ be such that $n \neq 3$, let R be its generic rank and R_E its expected generic rank. Then $R = R_E = \left\lceil \frac{n^3-1}{3n-2} \right\rceil$.*

Theorem B.2.9. *Let $\mathbb{C}^n \otimes \mathbb{C}^n \otimes \mathbb{C}^n$ be such that $n \neq 3$ and let R_E be its expected generic rank. If $\mathcal{T} \in \mathbb{C}^n \otimes \mathbb{C}^n \otimes \mathbb{C}^n$ is such that $rank(\mathcal{T}) < R_E$, then, with probability 1, $\mathcal{T}$ has finite CPD's.*

Definition B.2.10. *Let $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ such that $I_1 \geq I_2 \geq \ldots \geq I_L$. We say this space is* unbalanced *if*

$$\sum_{\ell=1}^{L}(I_\ell - 1) > \prod_{\ell=2}^{L} I_\ell.$$

Otherwise we say it is balanced.

Intuitively, a tensor space is unbalanced when the largest dimension is much larger than the other dimensions of the tensor product. In this case the tensor space essentially behaves as a matrix space. As the next result shows, the notion of expected generic rank is not necessary for unbalanced spaces because now we have an explicit formula for the generic rank.

Theorem B.2.11. *Let $\mathbb{V}^{(1)} \otimes \ldots \otimes \mathbb{V}^{(L)}$ be unbalanced and such that $I_1 \geq I_2 \geq \ldots \geq I_L$. If R is its generic rank, then $R = \min\left\{I_1, \prod_{\ell=2}^{L} I_\ell\right\}$.*

B.3 Special products

In addition to the tensor product and multilinear multiplication, there are a few other products that will be relevant to us.

Definition B.3.1. *Let two matrices* $\boldsymbol{A} \in \mathbb{K}^{k\times\ell}, \boldsymbol{B} \in \mathbb{K}^{m\times n}$. *The* Kronecker product *between* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *is defined by*

$$\boldsymbol{A}\tilde{\otimes}\boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1\ell}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2\ell}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1}\boldsymbol{B} & a_{k2}\boldsymbol{B} & \dots & a_{k\ell}\boldsymbol{B} \end{bmatrix}.$$

The matrix given in the definition is a block matrix such that each block is a $m \times n$ matrix, so $\mathbf{A}\tilde{\otimes}\mathbf{B}$ is a $km \times \ell n$ matrix. We would like to point out that some texts uses $\otimes$ for the Kronecker product and $\circ$ for the tensor product.

Lemma B.3.2. *For any vectors* $\boldsymbol{x}^{(1)} \in \mathbb{R}^{I_1}, \dots, \boldsymbol{x}^{(L)} \in \mathbb{R}^{I_L}$ *we have that*

$$\boldsymbol{x}^{(1)}\tilde{\otimes}\dots\tilde{\otimes}\boldsymbol{x}^{(L)} = \begin{bmatrix} x_1^{(1)}\cdot\ldots\cdot x_1^{(L-1)}x_1^{(L)} \\ x_1^{(1)}\cdot\ldots\cdot x_1^{(L-1)}x_2^{(L)} \\ \vdots \\ x_1^{(1)}\cdot\ldots\cdot x_1^{(L-1)}x_{I_L}^{(L)} \\ x_1^{(1)}\cdot\ldots\cdot x_2^{(L-1)}x_1^{(L)} \\ \vdots \\ x_{I_1}^{(1)}\cdot\ldots\cdot x_{I_L-1}^{(L-1)}x_1^{(L)} \\ \vdots \\ x_{I_1}^{(1)}\cdot\ldots\cdot x_{I_L-1}^{(L-1)}x_{I_L}^{(L)} \end{bmatrix}.$$

The important thing to notice in this lemma is the ordering of the multi-index from top to bottom. It goes to $1\dots11$ to $I_L\dots I_{L-1}I_L$ following the numbering increasing order. In particular this lemma implies that $\mathbf{x}^{(1)}\tilde{\otimes}\dots\tilde{\otimes}\mathbf{x}^{(L)} = vec(\mathbf{x}^{(1)}\otimes\dots\otimes\mathbf{x}^{(L)})$, assuming we are using the same ordering of the multi-indexes. This explains why the notation $\otimes$ is commonly used for the Kronecker product.

Definition B.3.3. *Let two matrices* $\boldsymbol{A} \in \mathbb{K}^{k\times n}, \boldsymbol{B} \in \mathbb{K}^{m\times n}$. *The* Khatri-Rao product *between* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *is defined by*

$$\boldsymbol{A}\odot\boldsymbol{B} = \left[\boldsymbol{A}_{:1}\tilde{\otimes}\boldsymbol{B}_{:1},\ \boldsymbol{A}_{:2}\tilde{\otimes}\boldsymbol{B}_{:2},\ \dots,\ \boldsymbol{A}_{:n}\tilde{\otimes}\boldsymbol{B}_{:n}\right].$$

In the above definition, each product $\mathbf{A}_{:i}\tilde{\otimes}\mathbf{B}_{:i}$ is a Kronecker product of two column vectors. Thus, each $\mathbf{A}_{:i}\tilde{\otimes}\mathbf{B}_{:i}$ is a $km \times 1$ vector, so that $\mathbf{A} \odot \mathbf{B}$ is $km \times n$ matrix. In particular, if $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}$ are the factor matrices of a CPD, then $\mathbf{W}^{(1)} \odot \ldots \odot \mathbf{W}^{(L)}$ is a matrix of shape $\prod_{\ell=1}^{L} I_\ell \times R$, where its r-th column is the coordinate representation of the r-th rank one term of the CPD.

Definition B.3.4. *Let two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{K}^{m\times n}$. The* Hadamard product *between $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined by*

$$\boldsymbol{A} * \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \ldots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \ldots & a_{2n}b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \ldots & a_{mn}b_{mn} \end{bmatrix}.$$

Note that the Hadamard product is nothing more than the coordinate-wise product between matrices. The next theorem summarizes some of the key properties of these three products.

Theorem B.3.5. *Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ be matrices with the sizes necessary to have well defined operations below. The the following holds.*

1. $\boldsymbol{A}\tilde{\otimes}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{A}\tilde{\otimes}\boldsymbol{B} + \boldsymbol{A}\tilde{\otimes}\boldsymbol{C}$.
2. $(\boldsymbol{A}+\boldsymbol{B})\tilde{\otimes}\boldsymbol{C} = \boldsymbol{A}\tilde{\otimes}\boldsymbol{C} + \boldsymbol{B}\tilde{\otimes}\boldsymbol{C}$.
3. *For all* $\alpha \in \mathbb{K}$, $(\alpha\boldsymbol{A})\tilde{\otimes}\boldsymbol{B} = \boldsymbol{A}\tilde{\otimes}(\alpha\boldsymbol{B}) = \alpha(\boldsymbol{A}\tilde{\otimes}\boldsymbol{B})$.
4. $(\boldsymbol{A}\tilde{\otimes}\boldsymbol{B})\tilde{\otimes}\boldsymbol{C} = \boldsymbol{A}\tilde{\otimes}(\boldsymbol{B}\tilde{\otimes}\boldsymbol{C})$.
5. $(\boldsymbol{A}\tilde{\otimes}\boldsymbol{B})(\boldsymbol{C}\tilde{\otimes}\boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C})\tilde{\otimes}(\boldsymbol{B}\boldsymbol{D})$.
6. $\boldsymbol{A}^{-1}\tilde{\otimes}\boldsymbol{B}^{-1} = (\boldsymbol{A}\tilde{\otimes}\boldsymbol{B})^{-1}$.
7. $\boldsymbol{A}^{\dagger}\tilde{\otimes}\boldsymbol{B}^{\dagger} = (\boldsymbol{A}\tilde{\otimes}\boldsymbol{B})^{\dagger}$.
8. $\boldsymbol{A}^{T}\tilde{\otimes}\boldsymbol{B}^{T} = (\boldsymbol{A}\tilde{\otimes}\boldsymbol{B})^{T}$.
9. $(\boldsymbol{A}\tilde{\otimes}\boldsymbol{B})^{\dagger} = \boldsymbol{A}^{\dagger}\tilde{\otimes}\boldsymbol{B}^{\dagger}$.
10. $(\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{C} = \boldsymbol{A} \odot (\boldsymbol{B} \odot \boldsymbol{C})$.
11. $(\boldsymbol{A} \odot \boldsymbol{B})^{T}(\boldsymbol{A} \odot \boldsymbol{B}) = (\boldsymbol{A}^{T}\boldsymbol{A}) * (\boldsymbol{B}^{T}\boldsymbol{B})$.
12. $(\boldsymbol{A} \odot \boldsymbol{B})^{\dagger} = ((\boldsymbol{A}^{T}\boldsymbol{A}) * (\boldsymbol{B}^{T}\boldsymbol{B}))^{\dagger}(\boldsymbol{A} \odot \boldsymbol{B})^{T}$.
13. *If $\boldsymbol{a}, \boldsymbol{b}$ are vectors, then* $\boldsymbol{a}\tilde{\otimes}\boldsymbol{b} = \boldsymbol{a} \odot \boldsymbol{b}$.
14. $(\boldsymbol{A}\tilde{\otimes}\boldsymbol{B}) * (\boldsymbol{C}\tilde{\otimes}\boldsymbol{D}) = (\boldsymbol{A} * \boldsymbol{C})\tilde{\otimes}(\boldsymbol{B} * \boldsymbol{D})$.

B.4 Symmetric tensors

Let S_L be the group of permutations of L elements. Each element of S_L will be denoted by σ. We can see σ simply as a L-tuple of integers between 1 and L, without repetitions. Thus, we denote each entry of σ by $\sigma(\ell)$. For example, suppose $L = 3$. A possible permutation $\sigma \in S_3$ is $\sigma = (3, 2, 1)$. In this case, σ permutes the 3-tuple $(1, 2, 3)$ to $(3, 2, 1)$. Thus we have that $\sigma(1) = 3$, $\sigma(2) = 2$, $\sigma(3) = 1$. Now, define the map $\pi_S : \mathbb{V}^{\otimes L} \to \mathbb{V}^{\otimes L}$ by

$$\pi_S(\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)}) = \frac{1}{L!} \sum_{\sigma \in S_L} \mathbf{v}^{(\sigma(1))} \otimes \ldots \otimes \mathbf{v}^{(\sigma(L))}.$$

We may interpret $\pi_S(\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)})$ as the average of the tensor products between the vectors $\mathbf{v}^{(\ell)}$ considering all possible permutations of the indexes. Usually one denote $\mathbf{v}^{(1)} \cdot \ldots \cdot \mathbf{v}^{(L)} = \pi_S(\mathbf{v}^{(1)} \otimes \ldots \otimes \mathbf{v}^{(L)})$ and call this the *symmetric tensor product* between the vectors $\mathbf{v}^{(\ell)}$. In the case all vectors are equal, say equal to $\mathbf{v}$, we denote $\mathbf{v}^L = \pi_S(\underbrace{\mathbf{v} \otimes \ldots \otimes \mathbf{v}}_{L \text{ times}})$. The image of π_S is called the space of *symmetric tensors*[2] of $\mathbb{V}$ and it is denoted by $S^L(\mathbb{V})$. Another way to characterize symmetric tensors is given by the following theorem.

Theorem B.4.1. *A tensor $\mathcal{T} \in \mathbb{V}^{\otimes L}$ is symmetric if, and only if,*

$$\mathcal{T}(\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(L)}) = \mathcal{T}(\boldsymbol{v}^{(\sigma(1))}, \ldots, \boldsymbol{v}^{(\sigma(L))})$$

for all $\sigma \in S_L$.

We finish our discussion about symmetric tensors giving a few more relevant results.

Theorem B.4.2. *Let $\dim(\mathbb{V}) = n$ and $\{\boldsymbol{e}^{(1)}, \ldots, \boldsymbol{e}^{(n)}\}$ be a basis for $\mathbb{V}$. Then*

1. *$\{e_{i_1} \cdot \ldots \cdot e_{i_L} : 1 \leq i_1 \leq \ldots \leq i_L \leq n\}$ is a basis of $S^L(\mathbb{V})$.*
2. *$\dim S^L(\mathbb{V}) = \binom{L - 1 + n}{L}$.*

Theorem B.4.3. *Let $\mathcal{T} \in \mathbb{V}^{\otimes L}$ and consider a choice of basis for $\mathbb{V}$. If $\mathcal{T}$ is symmetric, then $\mathcal{T}_{i_1 \ldots i_L} = \mathcal{T}_{i_{\sigma(1)} \ldots i_{\sigma(L)}}$ for all $\sigma \in S_L$.*

Example B.4.4 (Derivatives)**.** *Let $f : \mathbb{C}^m \to \mathbb{C}^n$ be a function L times differentiable at $\boldsymbol{a} \in \mathbb{C}^m$. The L-th derivative of f at $\boldsymbol{a}$ is a L-linear map $Df^{(L)}(\boldsymbol{a}) : \underbrace{\mathbb{C}^m \times \ldots \times \mathbb{C}^m}_{L \text{ times}} \to \mathbb{C}^n$.*

Because of the isomorphism

[2] If one want to specify the tensor order it is possible to refer to this space as the space of *symmetric L-tensors* or *symmetric L-order tensors*

$$\mathcal{L}_L(\mathbb{C}^m;\mathbb{C}^n) \cong (\mathbb{C}^m)^* \otimes \ldots \otimes (\mathbb{C}^m)^* \otimes \mathbb{C}^n,$$

we can write $Df^{(L)}(\boldsymbol{a}) \in (\mathbb{C}^m)^* \otimes \ldots \otimes (\mathbb{C}^m)^* \otimes \mathbb{C}^n$ *and consider this map as a mixed tensor of type* $(L,1)$.

Let $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_m\}$ *be the canonical basis of* $\mathbb{C}^m$. *Given any vectors* $\dot{\boldsymbol{x}}^{(1)},\ldots,\dot{\boldsymbol{x}}^{(L)} \in \mathbb{C}^m$, *we have that*

$$Df^{(L)}(\boldsymbol{a})(\dot{\boldsymbol{x}}^{(1)},\ldots,\dot{\boldsymbol{x}}^{(L)}) = \sum_{i_1,\ldots,i_L=1}^{m} \dot{\boldsymbol{x}}_{i_1}^{(1)} \ldots \dot{\boldsymbol{x}}_{i_L}^{(L)} \cdot Df^{(L)}(\boldsymbol{a})(\boldsymbol{e}_{i_1},\ldots,\boldsymbol{e}_{i_k}) =$$

$$= \sum_{i_1,\ldots,i_L=1}^{m} \dot{\boldsymbol{x}}_{i_1}^{(1)} \ldots \dot{\boldsymbol{x}}_{i_L}^{(L)} \cdot \begin{bmatrix} \dfrac{\partial^L f_1}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a}) \\ \vdots \\ \dfrac{\partial^L f_n}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a}) \end{bmatrix}.$$

Using the notation

$$\frac{\partial^L f}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a}) = \begin{bmatrix} \dfrac{\partial^L f_1}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a}) \\ \vdots \\ \dfrac{\partial^L f_n}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a}) \end{bmatrix}$$

and considering the dual basis $dx_i : \mathbb{C}^m \to \mathbb{C}$ *such that* $dx_i(\boldsymbol{x}) = dx_i(x_1,\ldots,x_m) = x_i$, *we have that*

$$Df^{(L)}(\boldsymbol{a})(\dot{\boldsymbol{x}}^{(1)},\ldots,\dot{\boldsymbol{x}}^{(L)}) = \sum_{i_1,\ldots,i_L=1}^{m} dx_{i_1} \otimes \ldots \otimes dx_{i_L}(\dot{\boldsymbol{x}}^{(1)},\ldots,\dot{\boldsymbol{x}}^{(L)}) \cdot \frac{\partial^L f}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a}).$$

Notice that $Df^{(L)}(\boldsymbol{a})$ *a tensor is such that each coordinate* $(i_1,\ldots,i_L,j)$ *is given by* $\dfrac{\partial^L f_j}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a})$, *for* $1 \leq i_1,\ldots,i_L \leq m$ *and* $1 \leq j \leq n$. *In a more simplified way, we can write*

$$Df^{(L)}(\boldsymbol{a}) = \sum_{i_1,\ldots,i_L=1}^{m} dx_{i_1} \otimes \ldots \otimes dx_{i_L} \otimes \frac{\partial^L f}{\partial x_{i_1}\ldots\partial x_{i_L}}(\boldsymbol{a}),$$

with the understanding that we are using the interpretation $Df^{(L)}(\boldsymbol{a}) \in \mathcal{L}_L(\mathbb{C}^m;\mathbb{C}^n)$ *and not* $Df^{(L)}(\boldsymbol{a}) \in \mathcal{L}(\mathbb{C}^m,\ldots,\mathbb{C}^m,\mathbb{C}^{n*};\mathbb{C})$. *This is a nice characterization of the derivative. Another elegant one is to expand*

$$\frac{\partial^L f}{\partial x_{i_1}\dots\partial x_{i_L}}(\boldsymbol{a}) = \begin{bmatrix} \dfrac{\partial^L f_1}{\partial x_{i_1}\dots\partial x_{i_L}}(\boldsymbol{a}) \\ \vdots \\ \dfrac{\partial^L f_n}{\partial x_{i_1}\dots\partial x_{i_L}}(\boldsymbol{a}) \end{bmatrix} = \sum_{j=1}^{n} \frac{\partial^L f_j}{\partial x_{i_1}\dots\partial x_{i_L}}(\boldsymbol{a})\cdot \boldsymbol{e}_j$$

so that the previous formula becomes

$$Df^{(L)}(\boldsymbol{a}) = \sum_{i_1,\dots,i_L=1}^{m} dx_{i_1}\otimes\dots\otimes dx_{i_L}\otimes\left(\sum_{j=1}^{n}\frac{\partial^L f_j}{\partial x_{i_1}\dots\partial x_{i_L}}(\boldsymbol{a})\cdot\boldsymbol{e}_j\right) =$$

$$= \sum_{i_1,\dots,i_L=1}^{m}\sum_{j=1}^{n}\frac{\partial^L f_j}{\partial x_{i_1}\dots\partial x_{i_L}}(\boldsymbol{a})\cdot dx_{i_1}\otimes\dots\otimes dx_{i_L}\otimes\boldsymbol{e}_j.$$

In particular, with this expression we can see that the coordinates of $Df^{(L)}(\boldsymbol{a})$*, as a* $(L+1)$*-th order tensor, are given by*

$$Df^{(L)}(\boldsymbol{a})_{i_1\dots i_L j} = \frac{\partial^L f_j}{\partial x_{i_1}\dots\partial x_{i_L}}(\boldsymbol{a}).$$

It is also possible to consider the identification we are using so far and consider $Df^{(L)}(\boldsymbol{a})$ *as a tensor in* $\mathbb{C}^m\otimes\dots\otimes\mathbb{C}^m\otimes\mathbb{C}^n$*. Then, instead of* dx_i *we use the canonical basis vector* $\boldsymbol{e}_i$*.*

B.5 Antisymmetric tensors

Given $\sigma\in S_L$, the *sign* of σ is denoted by $\operatorname{sgn}(\sigma)$. We define $\operatorname{sgn}(\sigma)=1$ if σ makes an even quantity of permutation and $\operatorname{sgn}(\sigma)=-1$ if σ makes an odd quantity of permutations. For example, let $L=3$ and $\sigma=(3,2,1)$. Note that $\sigma(2)=2$, then this number is not permuted, but the numbers 1 and 3 are permuted. From this it follows that σ makes 2 permutations. Since 2 is even, we conclude that $\operatorname{sgn}(\sigma)=1$. Now let $\tilde{\sigma}=(3,1,2)$. In this case all three numbers are permuted. Since 3 is odd, we conclude that $\operatorname{sgn}(\tilde{\sigma})=-1$.

Now define the map $\pi_\Lambda:\mathbb{V}^{\otimes L}\to\mathbb{V}^{\otimes L}$ by

$$\pi_\Lambda(\mathbf{v}^{(1)}\otimes\dots\otimes\mathbf{v}^{(L)}) = \frac{1}{L!}\sum_{\sigma\in S_L}\operatorname{sgn}(\sigma)\mathbf{v}^{(\sigma(1))}\otimes\dots\otimes\mathbf{v}^{(\sigma(L))}.$$

We may interpret $\pi_\Lambda(\mathbf{v}^{(1)}\otimes\dots\otimes\mathbf{v}^{(L)})$ as the signed average of the tensor products between the vectors $\mathbf{v}^{(\ell)}$ considering all possible permutations of the indexes. Usually one denote $\mathbf{v}^{(1)}\wedge\dots\wedge\mathbf{v}^{(L)}=\pi_\Lambda(\mathbf{v}^{(1)}\otimes\dots\otimes\mathbf{v}^{(L)})$ and can this the *exterior product* or *antisymmetric product* between the vectors $\mathbf{v}^{(\ell)}$. The image of π_Λ is called the space of *antisymmetric tensors* or *alternating tensors* of V and it is denoted by $\Lambda^L(\mathbb{V})$.

Theorem B.5.1. $\boldsymbol{v}^{(1)}\wedge\dots\wedge\boldsymbol{v}^{(L)}=0 \iff$ *the vectors* $\boldsymbol{v}^{(1)},\dots,\boldsymbol{v}^{(L)}$ *are linearly dependent.*

Theorem B.5.2. $\boldsymbol{v}^{(1)} \wedge \ldots \wedge \boldsymbol{v}^{(\ell)} \wedge \ldots \wedge \boldsymbol{v}^{(j)} \wedge \ldots \wedge \boldsymbol{v}^{(L)} = -\boldsymbol{v}^{(1)} \wedge \ldots \wedge \boldsymbol{v}^{(j)} \wedge \ldots \wedge \boldsymbol{v}^{(\ell)} \wedge \ldots \wedge \boldsymbol{v}^{(L)}$ *for all* $i \neq j$.

Theorem B.5.3. *Let* $\dim(\mathbb{V}) = n$ *and* $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ *be a basis for* $\mathbb{V}$. *Then*

1. $\{\boldsymbol{e}_{i_1} \wedge \ldots \wedge \boldsymbol{e}_{i_L} : 1 \leq i_1 < \ldots < i_L \leq n\}$ *is a basis of* $\Lambda^L(V)$.
2. $\dim \Lambda^L(\mathbb{V}) = \binom{n}{L}$.

Theorem B.5.4. *Let* $\mathcal{T} \in \mathbb{V}^{\otimes L}$ *and consider a choice of basis for* $\mathbb{V}$. *If* $\mathcal{T}$ *is antisymmetric, then* $\mathcal{T}_{i_1 \ldots i_p \ldots i_q \ldots i_L} = -\mathcal{T}_{i_1 \ldots i_q \ldots i_p \ldots i_L}$ *for all* $p \neq q$.

Bibliography

[1] F. B. Diniz, https://github.com/felipebottega/Tensor-Fox.

[2] R. Bro, *Multi-way Analysis in the Food Industry: Models, Algorithms and Applications*, PhD thesis, University of Amsterdam, 1998.

[3] A. Cichoki, R. Zdunek, A. H. Phan and A. I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley, 2009.

[4] A. H. Phan, P. Tichavsky, and A. Cichoki, *Low Complexity Damped Gauss-Newton Algorithm for CANDECOMP/PARAFAC*, SIAM Journal on Matrix Analysis and Applications, 34 (1), 126-147 (2013).

[5] A. Cichoki, *Era of Big Data Processing: A New Approach via Tensor Networks and Tensor Decompositions*, arXiv:1403.2048v4 (2014).

[6] Tensorbox package: https://github.com/phananhhuy/TensorBox.

[7] B. Savas, and Lars Eldén, *Handwritten Digit Classification Using Higher Order Singular Value Decomposition*, Pattern Recognition Society, vol. 40, no. 3, pp. 993-1003, 2007.

[8] A. Smilde, R. Bro, and P. Geladi, *Multi-way Analysis with Applications in the Chemical Sciences*, Wiley, 2004.

[9] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, *Tensor Decompositions for Learning Latent Variable Models*, Journal of Machine Learning Research, vo. 15, pp. 2773-2832, 2014.

[10] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, *A Method of Moments for Mixture Models and Hidden Markov models*, In *Twenty-Fifth Annual Conference on Learning Theory*, volume 23, pages 33.1-33.34, 2012c.

[11] A. Anandkumar,R. Ge, *Efficient Approaches for Escaping Higher Order Saddle Points in Non-Convex Optimization*, JMLR: Workshop and Conference Proceedings vol 49:1-22, 2016.

[12] P. Comon, X. Luciani, and A. L. F. de Almeida, *Tensor Decompositions, Alternating Least Squares and other Tales*, Journal of Chemometrics, Wiley, 2009.

[13] M. Rajih and P. Comon, *Enhanced line search: A novel method to accelerate PARAFAC*, in EUSIPCO-05: Proceedings of the 13th European Signal Processing Conference, 2005.

[14] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, *Tensorlab 3.0*, Available online, Mar. 2016. https://www.tensorlab.net/.

[15] L. Sorber, M. Van Barel, and L. De Lathauwer, *Optimization-based Algorithms for Tensor Decompositions: Canonical Polyadic Decomposition, Decomposition in Rank-$(L_r, L_r, 1)$ Terms, and a New Generalization*, SIAM Journal on Optimization, 2013.

[16] B. W. Bader, T. G. Kolda and others. *MATLAB Tensor Toolbox Version 3.1*, Available online, February 2019, https://www.tensortoolbox.org/.

[17] E. Acar, D. M. Dunlavy and T. G. Kolda. *A Scalable Optimization Approach for Fitting Canonical Tensor Decompositions*, Journal of Chemometrics 25(2):67-86, February 2011.

[18] J. Kossaifi, Y. Panagakis, and A. Anandkumar. 2017. *TensorLy - Simple and Fast Tensor Learning in Python (2017)*, https://github.com/tensorly/.

[19] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar and Maja Pantic, *TensorLy: Tensor Learning in Python*, Journal of Machine Learning Research (JMLR), 2019, volume 20, number 26.

[20] V. de Silva, and L.-H. Lim, *Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem*, SIAM Journal on Matrix Analysis and Applications, 30 (2008), pp. 1084-1127.

[21] C. J. Hillar, and L.-H. Lim. *Most tensor problems are NP-hard*, Journal of the ACM, 60(6):45:1-45:39, November 2013. ISSN 0004-5411. doi: 10.1145/2512329.

[22] K. Madsen, H. B. Nielsen, and O. Tingleff, *Methods for Non-Linear Least Squares Problems, 2nd edition*, Informatics and Mathematical Modelling, Technical University of Denmark, 2004.

[23] A. Shashua, and T. Hazan, *Non-negative Tensor Factorization with Applications to Statistics and Computer Vision*, Proceedings of the 22nd International Conference on Machine Learning (ICML), 22 (2005), pp. 792-799.

[24] D. Donoho and V. Stodden, *When does non-negative matrix factorization give a correct decomposition into parts*. Proceedings of the conference on Neural Information Processing Systems (NIPS) (2003).

[25] V. Strassen, *Gaussian Elimination is not Optimal*, Numerische Mathematik, 13:354-356, 1969.

[26] S. Rabanser, O. Shchur, and S. Günnemann, *Introduction to Tensor Decompositions and their Applications in Machine Learning*, arXiv:1711.10781v1 (2017).

[27] T. G. Kolda and B. W. Bader, *Tensor Decompositions and Applications*, SIAM Review, 51:3, in press (2009).

[28] T. G. Kolda, *Multilinear Operators for Higher-Order Decompositions*, Tech. Report SAND2006-2081, Sandia National Laboratories, Albuquerque, NM, Livermore, CA, 2006.

[29] C. Battaglino, G. Ballard, and T. G. Kolda. *A Practical Randomized CP Tensor Decomposition*, ArXiv, 1701.06600:1-26, 2017.

[30] L. De Lathauwer, B. De Moor, and J. Vandewalle, *A Multilinear Singular Value Decomposition*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1253-1278.

[31] N. Halko, P.G. Martinsson, J.A. Tropp, *Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions*, SIAM Rev, Vol.53, No.2, 2011, pp. 217-288.

[32] Computational methods to approximate the MLSVD: https://en.wikipedia.org/wiki/Higher-order_singular_value_decomposition#Computation

[33] N. Vannieuwenhoven, R. Vandebril, and K. Meerbergen, *A New Truncation Strategy for the Higher-Order Singular Value Decomposition*, SIAM J. Sci. Comput. 34 (2012), no. 2, A1027-A1052.

[34] N. Vannieuwenhoven, *Condition Numbers for the Tensor Rank Decomposition*, Linear Algebra Appl., 535 (2017), pp. 35-86.

[35] M. Benzi, *Preconditioning Techniques for Large Linear Systems: A Survey*, Journal of Computational Physics 182, 418-477 (2002).

[36] G. Blekherman, and Z. Teitler, *On Maximum, Typical, and Generic Ranks*, arXiv:1402.2371v3 (2014).

[37] P. Comon, M. Sorensen, *Tensor Diagonalization by Orthogonal Transforms*, Report ISRN I3S/RR-2007-06-FR, Universite Nice, Sophia Antipolis.

[38] L. De Lathauwer, B. De Moor, and J. Vandewalle, *On the Best Rank-1 and Rank-*$(R_1, R_2, \ldots, R_N)$ *Approximation of Higher-Order Tensors*, SIAM Journal on Matrix Analysis and Applications., 21 (2000), pp. 1324-1342.

[39] J. M. Landsberg, *Tensors: Geometry and Applications*, AMS, Providence, RI, 2012.

[40] J. M. Landsberg, *Geometry and Complexity Theory*, Cambridge, 2017.

[41] P. Breiding, *Numerical and Statistical Aspects of Tensor Decompositions*, PhD thesis, TU Berlin, 2017.

[42] P. Breiding and N. Vannieuwenhoven, *A Riemannian Trust Region Method for the Canonical Tensor Rank Approximation Problem*, arXiv:1709.00033v2 (2018).

[43] P. Breiding and N. Vannieuwenhoven, *The Condition Number of Join Decompositions*, arXiv:1611.08117v3 (2018).

[44] N. Vannieuwenhoven, R. Vandebril, and K. Meerbergen, *A new truncation strategy for the higher-order singular value decomposition*, SIAM J. Sci. Comput. 34 (2012), no. 2, A1027-A1052.

[45] J. B. Kruskal, *Three-way arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics*, Linear Algebra Appl., 18 (1977), pp. 95-138.

[46] J. B. Kruskal, *Statement of Some Current Results about Three-Way Arrays*, manuscript, AT&T Bell Laboratories, Murray Hill, NJ. Available at http://three-mode.leidenuniv.nl/pdf/k/kruskal1983.pdf, 1983.

[47] J. B. Kruskal, *Rank, Decomposition, and Uniqueness for 3-way and N -way arrays, in Mul- tiway Data Analysis*, R. Coppi and S. Bolasco, eds., North-Holland, Amsterdam, 1989, pp. 7-18.

[48] J. B. Kruskal, R. A. Harshman, and M. E. Lundy, *How 3-MFA Data can Cause Degenerate Parafac Solutions, Among Other Relationships*, in Multiway Data Analysis, R. Coppi and S. Bolasco, eds., Elsevier Science, Amsterdam, 1989, pp. 115-122.

[49] N. D. Sidiropoulos and R. Bro, *On the Uniqueness of Multilinear Decomposition of N-way arrays*, J. Chemometrics, 14 (2000), pp. 229-239.

[50] F. L. Hitchcock, *The Expression of a Tensor or a Polyadic as a Sum of Products*, J. Math. Phys., 6 (1927), pp. 164-189.

[51] D. Bini, G. Lotti, and F. Romani, *Approximate Solutions for the Bilinear Form Computational Problem*, SIAM J. Comput. 9 (1980), no. 4, 692-697.MR592760 (82a:68065).

[52] D. Bini, M. Capovani, G. Lotti, and F. Romani, $O(n^{2.7799})$ *Complexity for* $n \times n$ *Approximate Matrix Multiplication*, Inform. Process. Lett., 8 (1979), pp. 234-235.

[53] C. Eckart and G. Young, *The Approximation of One Matrix by Another of Lower Rank*, Psychometrika, 1 (1936), pp. 211-218.

[54] W. P. Krijnen, T. K. Dijkstra, and A. Stegeman, *On the Non-Existence of Optimal Solutions and the Occurrence of "Degeneracy" in the Candecomp/Parafac Model*, preprint, 2007.

[55] A. Stegeman, *Degeneracy in Candecomp/Parafac Explained for* $p \times p \times 2$ *Arrays of Rank* $p + 1$ *or Higher*, Psychometrika, 71 (2006), pp. 483-501.

[56] A. Stegeman, *Low-rank Approximation of Generic* $p \times q \times 2$ *Arrays and Diverging Components in the Candecomp/Parafac Model*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 988-1007.

[57] P. Paatero, *Construction and Analysis of Degenerate Parafac Models*, J. Chemometrics, 14 (2000), pp. 285-299.

[58] L. R. Tucker, *Implications of Factor Analysis of Three-way Matrices for Measurement of Change*, in Problems in Measuring Change, C. W. Harris, ed., University of Wisconsin Press, 1963, pp. 122-137.

[59] L. R. Tucker, *Some Mathematical Notes on Three-mode Factor Analysis*, Psychometrika, 31 (1966), pp. 279-311.

[60] F. L. Hitchcock, *The Expression of a Tensor or a Polyadic as a Sum of Products*, J. Math. Phys., 6 (1927), pp. 164-189.

[61] F. L. Hitchcock, *Multilple Invariants and Generalized Rank of a p-way Matrix or Tensor*, J. Math. Phys., 7 (1927), pp. 39-79.

[62] J. D. Carroll and J. J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition*, Psychometrika, 35 (1970), pp. 283-319.

[63] R. A. Harshman, *Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multi-Modal Factor Analysis*, UCLA Working Papers in Phonetics, 16 (1970), pp. 1-84.

[64] B. C. Mitchell and D. S. Burdick, *Slowly Converging PARAFAC Sequences: Swamps and Two-Factor Degeneracies*, J. Chemometrics, 8 (1994), pp. 155-168.

[65] W. S. Rayens and B. C. Mitchell, *Two-Factor Degeneracies and a Stabilization of PARAFAC*, Chemometrics and Intelligent Laboratory Systems, 38 (1997), p. 173.

[66] D. Nion and L. De Lathauwer, *An Enhanced Line Search Scheme for Complex-Valued Tensor Decompositions*, application in DS-CDMA, Signal Process., 88 (2008), pp. 749-755.

[67] J. J. Moré and D. J. Thuente, *Line Search Algorithms with Guaranteed Sufficient Decrease*, ACM Trans. Math. Softw., 20 (1994), pp. 286-307.

[68] K. Levenberg, *A Method for the Solution of Certain Problems in Least Squares*. Quart. Appl. Math. 2 (1944), pp 164-168.

[69] D. Marquardt, *An Algorithm for Least Squares Estimation on Nonlinear Parameters*. SIAM J. A PPL . M ATH .11 (1963), pp 431-441.

[70] G. Tomasi and R. Bro, *INDAFAC and PARAFAC3W*, http://www.models.life.ku.dk/indafac.

[71] G. Tomasi and R. Bro, *A comparison of algorithms for fitting the PARAFAC model*, Computational Statistics and Data Analysis, 50 (2006), pp. 1700-1734.

[72] G. Tomasi and R. Bro, *PARAFAC and Missing Values*, Chemometrics Intell. Lab. Systems (2004), in press.

[73] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia (1997).

[74] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, (1997).

[75] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, (2002).

[76] A. van der Sluis, *Condition Numbers and Equilibration of Matrices*, Numer. Math., 14 (1969), pp. 14-23.

[77] A. M. Bradley and W. Murray, *Matrix-Free Approximate Equilibration*, arXiv:1110.2805v2 (2012).

[78] M. Benzi, *Preconditioning Techniques for Large Linear Systems: A Survey*, J. Comput. Phys. 182 (2002) 418-477.

[79] H. B. Nielsen, *Damping Parameter in Marquardt's Method*, IMM, DTU, Report IMM-REP-1999-05.

[80] L. Grasedyck, D. Kessner, and C. Tobler, *A Literature Survey of Low-Rank Tensor Approximation Techniques*, CGAMM-Mitteilungen, vol. 36, pp. 53-78, 2013.

[81] W. Hackbusch and B. N. Khoromskij, *Tensor-Product Approximation to Multidimensional Integral Operators and Green's Functions*, SIAM J. Matrix Anal. Appl. 30 (2008), no. 3, 1233-1253. MR2447450 (2009g:45023).

[82] Y. Zniyed, R. Boyer, André L.F. de Almeida, G. Favier, *High-Order CPD Estimation with Dimensionality Reduction Using a Tensor Train Model*, 26th European Signal Processing Conference (EUSIPCO - 2018).

[83] I. Oseledets, *Tensor-Train Decomposition*, SIAM Journal on Scientific Computing, vol. 33, no. 5, pp. 2295-2317, 2011.

[84] I. Oseledets and E. Tyrtyshnikov, *TT-Cross Approximation for Multidimensional Arrays*, Linear Algebra Appl., 432 (2010), pp. 70-88.

[85] Y. Song and L. Qi, *Infinite Dimensional Hilbert Tensors on spaces of Analytic Functions*, arXiv, Nov. 2016.

[86] D. C.-L. Fong and M.A. Saunders, *LSMR: An Iterative Algorithm for Sparse Least-Squares Problems*, SIAM Journal on Scientific Computing, 33(5):2950-2971, January 2011.

[87] Weinberger, K. Q., Dasgupta A., Langford J., Smola A. and Attenberg, J. (2009). *Feature Hashing for Large Scale Multitask Learning*. Proceedings of the 26th Annual International Conference on Machine Learning: 1113-1120. (2009) arXiv:0902.2206

[88] M. McTear, Z. Callejas and D. Griol. *The Conversational Interface*. Springer International Publishing, 2016.

[89] B. Savas and L.-H. Lim, *Quasi-Newton Methods on Grassmannians and Multilinear Approximations of Tensors*, SIAM J. Scientific Computing, vol. 32, no. 6, pp. 3352-3393, 2010.

[90] G. Bergqvist and E.G. Larsson, *The Higher-Order Singular Value Decomposition: Theory and an Application*, IEEE Signal Processing Magazine, vol. 27, no. 3, pp. 151-154, 2010.

[91] R. Bro, *PARAFAC: Tutorial and Applications*, Chemometrics and Intelligent Laboratory Systems, 38:149-171, 1997.

[92] P. Burgisser and F. Cucker, *Condition: The Geometry of Numerical Algorithms*, vol. 349 of Grundlehren der mathematischen Wissenschaften, Springer, Heidelberg, 2013.

[93] A. H. Phan, P. Tichavsky, and A. Cichocki, *Fast alternating LS algorithms for high order CANDECOMP/PARAFAC tensor factorizations*, IEEE Trans. Signal Process., vol. 61, no. 19, pp. 4834-4846, Oct. 2013.