Online multimodal dictionary learning

Abraham Traoré, Maxime Berar, Alain Rakotomamonjy

▶ **To cite this version:**

Abraham Traoré, Maxime Berar, Alain Rakotomamonjy. Online multimodal dictionary learning. Neurocomputing, 2019, 368 (7), pp.163-179. 10.1016/j.neucom.2019.08.053 . hal-01850923v5

HAL Id: hal-01850923
https://hal.science/hal-01850923v5

Submitted on 15 Nov 2019

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

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

Online Multimodal Dictionary Learning

Abraham Traoré[a] , Maxime Berar[b], Alain Rakotomamonjy[c]

[a,b,c] *LITIS, University of Rouen Normandie 76800 Saint-Etienne du Rouvray, FRANCE*
[a] *abraham.traore@etu.univ-rouen.fr*
[b,c] *maxime.berar,alain.Rakotomamonjy@univ-rouen.fr*

Abstract

We propose a new online approach for multimodal dictionary learning. The method developed in this work addresses the great challenges posed by the computational resource constraints in dynamic environment when dealing with large scale tensor sequences. Given a sequence of tensors, i.e. a set composed of equal-size tensors, the approach proposed in this paper allows to infer a basis of latent factors that generate these tensors by sequentially processing a small number of data samples instead of using the whole sequence at once. Our technique is based on block coordinate descent, gradient descent and recursive computations of the gradient. A theoretical result is provided and numerical experiments on both real and synthetic data sets are performed.

Keywords: tensor, block coordinate descent, gradient descent, recursive computations, online dictionary learning

1. Introduction

Multimodal dictionary learning is the task of constructing succinct representations of multiway data (also called tensor) using dictionary atoms per mode (a mode corresponding to one dimension of the data) learned from this data. The recent interest in this problem, motivated by the fact that the processing of multiway data with separate matrices extracted from the data block may lead to the loss of the covariance information among various modes [1], enhances the importance of developing efficient learning algorithms able to extract information of interest from such data. The two most common decompositions used for tensor analysis are the *Tucker* decomposition, introduced by Tucker in 1963 [2] and the Canonical Polyadic Decomposition also named *CPD* introduced independently by Hitchcock in [3] and by Cattell in [4]. These decompositions, which embody emerging tools for exploratory multiway data analysis, have been successfully applied to numerous applications such as: cluster analysis [5], image denoising [6], pattern recognition [7], face recognition [8]. Their links with dictionary learning have already been established in the literature [9], [10].

In a wide variety of applications, multiway datasets are naturally represented by sequences (i.e. set of several samples representing themselves tensors), especially when we are dealing with applications where new data samples keep coming over time (e.g. climate data [11]). Although the application of a batch-based decomposition (i.e. using the whole

dataset) is always possible through a recomputation from scratch, this approach can swiftly lead to a computational bottleneck for two main reasons. Firstly, we have to deal with the storage cost of the data samples. Secondly, the intermediate problems of standard techniques may lead to high space complexity (e.g. some standard approaches for nonnegative *Tucker* decomposition compute the Kronecker product of the $N-1$ loading matrices [12], N being the tensor order). Hence, memory and computationally efficient methods that are able to infer the latent factors using only a small number of samples at a time is of primary importance.
This paper is focused on multimodal dictionary learning through online *Tucker* decomposition. In the literature, the online techniques for *Tucker* decomposition can be split into two categories. For the first class of methods, the principle relies on inferring latent factors by sequentially processing streaming data samples with no need to resort to the past data [13], [14]. The second class of approaches lies in the idea of stacking both the past and the newly acquired data and performing a recursive update of the factors [15], [16]. The difference between the two classes of methods is that for the first one, only a small number of data samples is processed at the same time.

In this paper, the class of methods considered is the first one. A plethora of excellent techniques have already been proposed to tackle this problem. An approach presented in [13] infers multidimensional separable dictionaries through recursive updates in compressive sensing framework, the drawback being that the core tensor is assumed to be known in advance, which is not the case in general even though it is a natural assumption in compressive sensing. A Riemannian approach proposed in [14] deals with a fixed-rank tensor completion problem by turning the problem into an optimization problem on a Stiefel manifold. This method imposes orthogonality constraints, which may not be relevant for a specific task at hand (e.g. the nonnegativity constraint for inherent positive data is necessary to keep physical interpretability [17]). A *Tucker*-based method in the framework of multivariate spatio-temporal prediction has also been presented in [18] and whose core principle is based on projection of randomly generated subspaces. The main limitation that suffers this method is the so-called low rank assumption, which is not verified in practice for each dataset. The remaining approaches suffer similar limitations: they are not generic enough to take into account some characteristics of the data, impose constraints on the latent factors sizes (which makes them unsuitable for some specific tasks, e.g. overcomplete multimodal dictionary learning) or assume prior knowledge of some factors.

In this paper, we introduce a new online tensor decomposition approach based on block coordinate descent, gradient descent and recursive computations. Our contributions are the following ones:

1. We propose an online multimodal dictionary learning algorithm which is flexible enough to incorporate common constraints such as nonnegativity, sparsity, orthogonality with no assumption on the latent factors sizes, which makes it more general than the existing online approaches.
2. A theoretical result stating the convergence to a stationary point of the sequence of dictionary matrices generated by our algorithm is provided.

3. Comparison with state-of-the-art techniques through numerical experiments on both synthetic and real datasets is performed.

2. Notations

An N−order tensor is denoted by a boldface Euler script letter $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times ... \times I_N}$. The matrices are denoted by bold capital letters (e.g. $\mathbf{A}$). Matricization is the process of reordering all the elements of a tensor into a matrix. The mode-n matricization of a tensor $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{I_1 \times ... \times I_N}$ arranges the mode-n fibers to be the columns of the resulting matrix $\mathbf{Y}^{(n)} \in \mathbb{R}^{I_n \times (\prod_{k \neq n} I_k)}$. The mode-n product of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times ... \times I_N}$ with a matrix $\mathbf{B} \in \mathbb{R}^{J_n \times I_n}$ denoted by $\boldsymbol{\mathcal{X}} \times_n \mathbf{B}$ yields a tensor of the same order $\widehat{\boldsymbol{\mathcal{Y}}} \in \mathbb{R}^{I_1 \times .. \times J_n \times .. \times I_N}$ whose mode-n matricized form is defined by: $\widehat{\mathbf{Y}}^{(n)} = \mathbf{B}\mathbf{Y}^{(n)}$. The i^{th} slice of an order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times ... \times I_N}$, denoted by $\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^{I_2 \times ... \times I_N}$, is the tensor derived from $\boldsymbol{\mathcal{X}}$ by fixing the first index to i. The Kronecker product of two matrices is denoted by $\otimes$. For writing simplicity, we introduce the following notations:

- The set of integers from n to N is denoted by $I_N^n = \{n, .., N\}$. If n=1, the set is simply denoted by I_N
- The set of integers from 1 to N with n excluded is defined by $I_{N \neq n} = \{1, .., n-1, n+1, .., N\}$.
- Let $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{J_1 \times .. \times J_N}$ be a tensor and N matrices $\left\{\mathbf{A}^{(m)}\right\}_{1 \leq m \leq N}, \mathbf{A}^{(m)} \in \mathbb{R}^{I_m \times J_m}$. The contracted form of the product of $\boldsymbol{\mathcal{G}}$ with the matrices $\left\{\mathbf{A}^{(m)}\right\}_{1 \leq m \leq N}$ is denoted by:

$$\boldsymbol{\mathcal{G}} \times_{p \in I_{n-1}} \mathbf{A}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q \in I_N^{n+1}} \mathbf{A}^{(q)} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} .. \times_N \mathbf{A}^{(N)}$$

- The Kronecker product of the N matrices $\mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}$ is denoted by:

$$\otimes_{m \in \mathbf{I}_N} \mathbf{A}^{(m)} = \mathbf{A}^{(1)} \otimes \mathbf{A}^{(2)} ... \otimes \mathbf{A}^{(N)}$$

- The Kronecker product of the $N-1$ matrices $\mathbf{A}^{(1)}, .., \mathbf{A}^{(n-1)}, \mathbf{A}^{(n+1)}, .., \mathbf{A}^{(N)}$:

$$\otimes_{m \in \mathbf{I}_{N \neq n}} \mathbf{A}^{(m)} = \mathbf{A}^{(1)} \otimes \mathbf{A}^{(2)} ... \otimes \mathbf{A}^{(n-1)} \otimes \mathbf{A}^{(n+1)} ..... \otimes \mathbf{A}^{(N)}$$

The absolute value is denoted by $|\cdot|$. The Frobenius and ℓ_1 norms of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times .. \times I_N}$ denoted respectively by $\|\boldsymbol{\mathcal{X}}\|_F$ and $\|\boldsymbol{\mathcal{X}}\|_1$ are defined by:

$$\|\boldsymbol{\mathcal{X}}\|_F = \left(\sum_{1 \leq i_n \leq I_n, 1 \leq n \leq N} \boldsymbol{\mathcal{X}}_{i_1,..,i_N}^2 \right)^{\frac{1}{2}}, \|\boldsymbol{\mathcal{X}}\|_1 = \sum_{1 \leq i_n \leq I_n, 1 \leq n \leq N} | \boldsymbol{\mathcal{X}}_{i_1,..,i_N} |$$

3. Multimodal dictionary learning

The purpose of this section is to introduce briefly *Tucker* decomposition, present our algorithm and our theoretical result (as well as the underlying assumptions).

3.1. Brief overview of Tucker decomposition

Tucker decomposition is one of the most common decompositions used in tensor framework. Given a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times ... \times I_N}$, the standard *Tucker* decomposition consists of approximating $\boldsymbol{\mathcal{X}}$ by the product of a tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{J_1 \times ... \times J_N}$ with N matrices $\left\{\mathbf{A}^{(n)}\right\}_{1 \leq n \leq N}, \mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times J_n}$, i.e.:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{G}} \times_{n \in I_N} \mathbf{A}^{(n)}$$

The tensor $\boldsymbol{\mathcal{G}}$ is generally named the core tensor and the matrices $\mathbf{A}^{(n)}$ the loading matrices. This decomposition associated with orthogonality constraints can be thought of as the multidimensional counterpart of the singular value decomposition [19]. It can also be interpreted in terms of dictionary where the matrices $\mathbf{A}^{(n)}$ embody the dictionary matrices and the entries of $\boldsymbol{\mathcal{G}}$, the activation coefficients [20]. For the remainder of the paper, the latent factors $\mathbf{A}^{(n)}$ and $\boldsymbol{\mathcal{G}}$ will respectively be referred to as dictionary matrices and activation tensor.

The problem we are interested in is a particular case of *Tucker* decomposition where a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{T \times I_1 \times .. \times I_N}$ represents a set of T observations of size $I_1 \times ... \times I_N$ drawn from an unknown probability distribution, the objective being to perform the decomposition:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{G}} \times_1 \mathbf{I} \times_2 \mathbf{A}^{(1)} \times_2 ... \times_{N+1} \mathbf{A}^{(N)}, \boldsymbol{\mathcal{G}} \in \mathbb{R}^{T \times J_1 \times .. \times J_N}, \mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times J_n}$$

This problem can be specifically expressed via the following optimization problem:

$$\min_{\mathbf{A}^{(n)}, \{\boldsymbol{\mathcal{G}}_t\}_{1 \leq t \leq T}} \sum_{i=1}^{T} \|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}}_i \times_{n \in I_N} \mathbf{A}^{(n)}\|_F^2 \tag{1}$$

with $\boldsymbol{\mathcal{X}}_i \in \mathbb{R}^{I_1 \times ... \times I_N}$ and $\boldsymbol{\mathcal{G}}_i \in \mathbb{R}^{J_1 \times .. \times J_N}$ being respectively the i^{th} slices (tensor derived by fixing the first index to i) of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{G}}$. It is worth to notice that $\boldsymbol{\mathcal{X}}_i$ (respectively $\boldsymbol{\mathcal{G}}_i$) is an N-order tensor since it corresponds to a slice of a tensor of order $N+1$.

In the classical dictionary learning framework, it is common to seek to infer sparse activation coefficients while preventing dictionary atoms from diverging by introducing penalty functions [21]. This observation paves the way to the problem:

$$\min_{\mathbf{A}^{(n)}, \{\boldsymbol{\mathcal{G}}_t\}_{1 \leq t \leq T}} \sum_{i=1}^{T} \left(\|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}}_i \times_{n \in I_N} \mathbf{A}^{(n)}\|_F^2 + \Omega_1(\boldsymbol{\mathcal{G}}_i)\right) + \Omega_2(\mathbf{A}^{(1)}, .., \mathbf{A}^{(N)}) \tag{2}$$

The sparsity and boundedness are respectively enforced via general penalty functions Ω_1 and Ω_2.

3.2. Online multimodal dictionary learning OTL

The problem addressed in this work is the inference of multimodal dictionary atoms (dictionary atoms per tensor mode) while bypassing the computational bottleneck induced by the processing of all of the data samples at once.

To solve this problem for a large tensor sequence (i.e. a large number of tensors), we propose a probabilistic approach, which has already been used for online matrix decomposition

(e.g. dictionary learning [21]) and has proven its effectiveness for the numerical resolution of non-convex minimization problems [22].
Let's assume we are dealing with samples sequentially acquired over time and drawn from an unknown probability distribution $\mathbb{P}$ on the set of tensors of size $I_1 \times ... \times I_N$. Let's denote by $\boldsymbol{\mathcal{X}}_t \in \mathbb{R}^{I_1 \times .. I_N}$ the tensor acquired at the time step t. Let's also consider l the discrepancy between $\boldsymbol{\mathcal{X}}_t$ and its approximation given by the dictionary matrices defined by:

$$l(\boldsymbol{\mathcal{X}}, \left\{\mathbf{A}^{(n)}\right\}) = \min_{\boldsymbol{\mathcal{G}}} \frac{1}{2} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{G}} \times_{n \in I_N} \mathbf{A}^{(n)}\|_F^2 + \Omega_1(\boldsymbol{\mathcal{G}}) + \Omega_2(\mathbf{A}^{(1)}, .., \mathbf{A}^{(N)})$$

This expression is justified by the fact that our objective is to approximate each data sample by the product of an activation tensor with dictionary matrices.
The underlying idea of our approach is to update the dictionary matrices in such a way that the discrepancy between $\{\boldsymbol{\mathcal{X}}_t\}_{t \geq 1}$ and their approximation given by these matrices has a low expectation. Hence, a relevant problem is:

$$\min_{\mathbf{A}^{(1)}, .., \mathbf{A}^{(N)}} \left\{ f(\mathbf{A}^{(1)}, .., \mathbf{A}^{(N)}) \triangleq \mathbb{E}_{\boldsymbol{\mathcal{X}} \backsim \mathbb{P}} \left(l \left(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(1)}, .., \mathbf{A}^{(N)} \right) \right) \right\} \tag{3}$$

In the sequel, we assume Ω_2 is differentiable and Ω_1 admits a proximal operator [23].

3.3. Algorithm

Given that the probability distribution $\mathbb{P}$ is unknown, the problem (3) does not admit any analytical solution, i.e. the solution cannot be expressed in terms of a formula. To circumvent this deadlock, we replace the objective function by its estimator given by the mean sample, leading to the problem:

$$\min_{\mathbf{A}^{(1)}, .. \mathbf{A}^{(N)}} \left\{ \widehat{f}_t \left(\mathbf{A}^{(1)}, .. \mathbf{A}^{(N)} \right) \triangleq \frac{1}{t} \sum_{i=1}^{t} l \left(\boldsymbol{\mathcal{X}}_t, \mathbf{A}^{(1)}, .., \mathbf{A}^{(N)} \right) \right\} \tag{4}$$

The idea of our approach is to update the dictionary matrices $\left\{\mathbf{A}^{(n)}\right\}_{1 \leq n \leq N}$ by minimizing $\widehat{f}_t(\mathbf{A}^{(1)}, .. \mathbf{A}^{(N)})$ using only the sample unit $\boldsymbol{\mathcal{X}}_t$, the factors inferred from the processing of $\boldsymbol{\mathcal{X}}_{t-1}$ and a certain number of statistics (functions of the data samples). Let's denote $\left\{\mathbf{A}_t^{(n)}\right\}_{1 \leq n \leq N}$ the factors inferred by minimizing the function $\widehat{f}_t$.
Let's assume the acquisition of a new sample unit $\boldsymbol{\mathcal{X}}_t$ at a given time step t. The updates of the dictionary matrices (i.e. the determination of $\mathbf{A}_t^{(n)}$), inspired from previous works undertaken in matrix-based dictionary learning framework [24] (the difference is about the number of dictionary matrices), is performed via a two-step strategy:

- The computation of the activation tensor $\boldsymbol{\mathcal{G}}_t$ by projecting $\boldsymbol{\mathcal{X}}_t$ on the latent factors inferred at the time step $t-1$. In the sequel, this stage will be referred to as *Sparse coding*;
- The resolution of the problem (4) by a block coordinate descent, that is the minimization of the objective function with respect to one variable while others are fixed in the order $1 \rightarrow .. \rightarrow N$. This step will be referred to as *Block Coordinate Descent*.

In the sequel, our dictionary learning approach will be referred to as *OTL*.

3.3.1. Sparse coding

For this step, we consider the following minimization problem:

$$\min_{\boldsymbol{\mathcal{G}}} \underbrace{\frac{1}{2}\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{G}} \times_{n\in I_N} \mathbf{A}_{t-1}^{(n)}\|_F^2}_{\mathcal{O}(\boldsymbol{\mathcal{G}})} + \Omega_1(\boldsymbol{\mathcal{G}})$$

Instead of vectorizing this expression as it is the case for some *Tucker* decomposition techniques [12], which can induce a high space complexity due to the computation of the Kronecker product of the dictionary matrices, we propose to apply directly a proximal minimization approach [25] in the tensor domain.
A proximal minimization is a technique used to minimize of the sum of two functions: one differentiable and the other admitting a proximal operator [23]. In our case, the differentiable function is $\mathcal{O}$ whose derivative is given by:

$$\frac{\partial \mathcal{O}}{\partial \boldsymbol{\mathcal{G}}}(\boldsymbol{\mathcal{G}}) = -\boldsymbol{\mathcal{X}}_t \times_{n\in I_N} \mathbf{A}_{t-1}^{(n)T} + \boldsymbol{\mathcal{G}} \times_{n\in I_N} \mathbf{A}_{t-1}^{(n)T}\mathbf{A}_{t-1}^{(n)} \tag{5}$$

Given that the penalty Ω_1 is assumed to have a proximal operator denoted $\mathbf{prox}_{\Omega_1}$, the sparse coding step can be resolved numerically through **Algorithm** 1.

Algorithm 1 : *Sparse coding*

Inputs: new observation $\boldsymbol{\mathcal{X}}_t$, the dictionary matrices $\left\{\mathbf{A}_{t-1}^{(n)}\right\}_{1\leq n\leq N}$, the gradient descent step η, initial value $\boldsymbol{\mathcal{G}}_0$
Initialization: iter=0, $\boldsymbol{\mathcal{G}}_{iter} = \boldsymbol{\mathcal{G}}_0$

1: **while** a stopping criterion is not met **do**
2: $\quad\boldsymbol{\mathcal{G}}_{iter+1}$=$\mathbf{prox}_{\eta\Omega_1}(\boldsymbol{\mathcal{G}}_{iter} - \eta\frac{\partial\mathcal{O}}{\partial\boldsymbol{\mathcal{G}}}(\boldsymbol{\mathcal{G}}_{iter}))$, $\frac{\partial\mathcal{O}}{\partial\boldsymbol{\mathcal{G}}}$ defined according to (5)
3: $\quad iter \leftarrow iter + 1$
4: **end while**
5: **return** $\boldsymbol{\mathcal{G}}$

3.3.2. Block coordinate descent

In this section, we consider the updates problem of the dictionary matrices. Let's assume the computation of the activation tensor $\boldsymbol{\mathcal{G}}_t$ is already performed and the $n-1$ first matrices have already been updated, i.e. the matrices $\left\{\mathbf{A}_{k+1}^{(p)}\right\}_{1\leq p\leq n-1}$ are determined (k referring to the iteration number, an iteration corresponding to a round of updates of the variables $\mathbf{A}^{(1)}, .., \mathbf{A}^{(N)}$). The update of $\mathbf{A}^{(n)}$ is performed by solving the problem:

$$\mathbf{A}_{k+1}^{(n)} \leftarrow \arg\min_{\mathbf{A}^{(n)}} \widehat{f}_{n,t}(\mathbf{A}^{(n)}) \tag{6}$$

with:

$$\widehat{f}_{n,t}(\mathbf{A}^{(n)}) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}}_i \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \Omega_1(\boldsymbol{\mathcal{G}}_i) + \Omega_2(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}) \tag{7}$$

The problem (6) has no analytical solution, i.e. the solution cannot be expressed in terms of a formula. In view of the differentiability of the objective function $\widehat{f}_{n,t}$ with respect to $\mathbf{A}^{(n)}$ and since the solution cannot be expressed in terms of a formula, we propose a numerical resolution scheme based on gradient descent. With this choice, the dictionary matrix $\mathbf{A}^{(n)}$ can be updated using only $\boldsymbol{\mathcal{X}}_t$ and the matrices $\left\{\mathbf{A}_{k+1}^{(p)}\right\}_{1\leq p\leq n-1}, \left\{\mathbf{A}_k^{(q)}\right\}_{n+1\leq q\leq N}$ provided some statistics are updated over time. Indeed, the derivative of $\widehat{f}_{n,t}$ is given by:

$$\begin{aligned}\frac{\partial \widehat{f}_{n,t}}{\partial \mathbf{A}^{(n)}}\left(\mathbf{A}^{(n)}\right) = -\frac{1}{t}\sum_{i=1}^{t}\left(\widehat{\mathbf{X}}_i^{(n)}\mathbf{G}_i^{(n)T} - \mathbf{A}^{(n)}\mathbf{B}_i^{(n)}\mathbf{B}_i^{(n)T}\right) \\ + \frac{\partial \Omega_2}{\partial \mathbf{A}^{(n)}}(..,\mathbf{A}_{k+1}^{(n-1)},\mathbf{A}^{(n)},\mathbf{A}_k^{(n+1)},..)\end{aligned} \tag{8}$$

with $\mathbf{G}_i^{(n)}$ being the mode-n matricized form of the tensor $\boldsymbol{\mathcal{G}}_i$, $\mathbf{B}_i^{(n)}$ and $\widehat{\mathbf{X}}_i^{(n)}$ the mode-n matricized forms of the tensors $\boldsymbol{\mathcal{B}}_i$ and $\boldsymbol{\mathcal{X}}_i$ defined by:

$$\boldsymbol{\mathcal{B}}_i = \boldsymbol{\mathcal{G}}_i \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{I} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)} \tag{9}$$

$$\widehat{\boldsymbol{\mathcal{X}}}_i = \boldsymbol{\mathcal{X}}_i \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)T} \times_n \mathbf{I} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)T} \tag{10}$$

with $\mathbf{I}$ representing the identity matrix ($\mathbf{I}\in\mathbb{R}^{J_n\times J_n}$ for (9) and $\mathbf{I}\in\mathbb{R}^{I_n\times I_n}$ for (10))
The derivative can then be rewritten in the following form:

$$\frac{\partial \widehat{f}_{n,t}}{\partial \mathbf{A}^{(n)}}(\mathbf{A}^{(n)}) = -\frac{\mathbf{P}_t^{(n)}}{t} + \frac{\mathbf{A}^{(n)}\mathbf{Q}_t^{(n)}}{t} + \frac{\partial \Omega_2}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_{k+1}^{(1)},.,\mathbf{A}_{k+1}^{(n-1)},\mathbf{A}^{(n)},\mathbf{A}_k^{(n+1)},.,\mathbf{A}_k^{(N)}) \tag{11}$$

with: $\mathbf{P}_t^{(n)} = \sum_{i=1}^{t}\widehat{\mathbf{X}}_i^{(n)}\mathbf{G}_i^{(n)T}$, $\mathbf{Q}_t^{(n)} = \sum_{i=1}^{t}\mathbf{B}_i^{(n)}\mathbf{B}_i^{(n)T}$.
It is straightforward to notice that:

$$\mathbf{P}_t^{(n)} = \mathbf{P}_{t-1}^{(n)} + \widehat{\mathbf{X}}_t^{(n)}\mathbf{G}_t^{(n)T} \tag{12}$$

$$\mathbf{Q}_t^{(n)} = \mathbf{Q}_{t-1}^{(n)} + \mathbf{B}_t^{(n)}\mathbf{B}_t^{(n)T} \tag{13}$$

Since the sequences $\mathbf{P}_t^{(n)}$ and $\mathbf{Q}_t^{(n)}$ verify the equations (12) and (13), the computation of the gradient of $\widehat{f}_{n,t}$ only requires the newly acquired tensor $\boldsymbol{\mathcal{X}}_t$ and the loading matrices $\left\{\mathbf{A}_{k+1}^{(p)}\right\}_{1\leq p\leq n-1}, \left\{\mathbf{A}_k^{(q)}\right\}_{n+1\leq q\leq N}$. Concretely, for a fixed t, the derivative of $\widehat{f}_{n,t+1}$ can be derived from the statistics $\mathbf{P}_t^{(n)}, \mathbf{Q}_t^{(n)}$, the dictionary matrices $\left\{\mathbf{A}^{(n)}\right\}_{1\leq n\leq N}$ and the processing of the sample $\boldsymbol{\mathcal{X}}_{t+1}$. This is the key of our approach since it means that the inference of the latent factors $\left\{\mathbf{A}^{(n)}\right\}_{1\leq n\leq N}$ can be performed by processing a single subtensor at a time. The update of a single matrix $\mathbf{A}^{(m)}$ for the *block coordinate descent* step is summarized by **Algorithm 2**.

Algorithm 2 *Update of* $\mathbf{A}^{(n)}$

Inputs: $\boldsymbol{\mathcal{X}}_t$: newly added tensor, $\left\{\mathbf{A}_{k+1}^{(p)}\right\}_{1\leq p\leq n-1}, \left\{\mathbf{A}_{k}^{(q)}\right\}_{n+1\leq q\leq N}$: dictionary matrices, $\mathbf{A}_0^{(n)}$: initial value, $\boldsymbol{\mathcal{G}}_t$: sparse code associated to $\boldsymbol{\mathcal{X}}_t$, $\mathbf{P}_{t-1}^{(n)}$,$\mathbf{Q}_{t-1}^{(n)}$: statistics at the time step $t-1$.

Statistics update:

- Compute $\mathbf{P}_t^{(n)}$ via the recursive equation (12)
- Compute $\mathbf{Q}_t^{(n)}$ via the recursive equation (13)

Initialization: iter=0,$\mathbf{A}_{iter}^{(n)} = \mathbf{A}_0^{(n)}$

1: **while** a stopping criterion is not met **do**
2: $\quad\mathbf{A}_{iter+1}^{(n)} = \left(\mathbf{A}_{iter}^{(n)} - \eta\frac{\partial \widehat{f}_{n,t}}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_{iter}^{(n)})\right), \frac{\partial \widehat{f}_{n,t}}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_{iter}^{(n)})$ defined according to (11), (12) and (13)
3: $\quad iter \leftarrow iter + 1$
4: **end while**
5: **return** $\mathbf{A}^{(n)}$

3.3.3. Summary of the inference process

The whole inference process is summarised by **Algorithm** 3. One can notice that our decomposition technique is quite simple and does not require any constraint on the latent factors sizes, which makes it more general compared to the existing approaches in the sense that it is more suitable for a larger class of problem (e.g. overcomplete multimodal dictionary learning [9], [20]).

4. Theoretical result

Our theoretical result regarding the behavior of **Algorithm 3** is about exploring a double asymptotic behavior: the increasing of the number of samples over time and the increasing of the number of iterations in block coordinate descent. Thus, we consider two layers of loops: a first layer from block coordinate descent and a second one from the streaming samples. Let's consider the assumptions described below:

- The data samples are drawn from a probability distribution with compact support: this a natural assumption since we work only with finite values;
- We choose ℓ_1 penalty for activation coefficients and ℓ_2 penalty for dictionary matrices, i.e. we assume $\Omega_1(\boldsymbol{\mathcal{G}}) = \alpha\theta\|\boldsymbol{\mathcal{G}}\|_1$ and $\Omega_2(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}) = \frac{\alpha(1-\theta)}{2}\sum_{n=1}^{N}\|\mathbf{A}^{(n)}\|_F^2$ with $\alpha > 0, 0 < \theta < 1$: this choice is simply motivated by the fact that these functions embody two of the most used penalty functions in dictionary learning.
- For all the minimization problems considered, the activation tensor $\boldsymbol{\mathcal{G}}$ and the dictionary matrices $\mathbf{A}^{(n)}$ are assumed to belong to some compact sets, i.e. $\boldsymbol{\mathcal{G}} \in \mathbb{K}_{\boldsymbol{\mathcal{G}}}, \mathbf{A}^{(n)} \in \mathbb{K}_{I_n,J_n}$: this is a natural interpretation of the $\ell_1 - \ell_2$ penalty;

Algorithm 3 : *OTL*

Inputs: set of samples $\{\boldsymbol{\mathcal{X}}_1, .., \boldsymbol{\mathcal{X}}_t, ...\}$, Initial dictionary matrices $\left\{\mathbf{A}_0^{(n)}\right\}$, initial core tensor $\boldsymbol{\mathcal{G}}_0$

while a stopping criterion is not met **do**
  Draw $\boldsymbol{\mathcal{X}}_t$ according to a probability distribution $\mathbb{P}$
  Sparse coding

$$\boldsymbol{\mathcal{G}}_t \leftarrow \arg\min \frac{1}{2}\|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{G}} \times_{n \in I_N} \mathbf{A}_{t-1}^{(n)}\|_F^2 + \Omega_1(\boldsymbol{\mathcal{G}})$$

  Block Coordinate Descent
  for n from 1 to N **do**
    Update of the statistics $\mathbf{P}_t^{(n)}$ and $\mathbf{Q}_t^{(n)}$

$$\mathbf{P}_t^{(n)} = \mathbf{P}_{t-1}^{(n)} + \widehat{\mathbf{X}}_t^{(n)} \mathbf{G}_t^{(n)T}$$
$$\mathbf{Q}_t^{(n)} = \mathbf{Q}_{t-1}^{(n)} + \mathbf{B}_t^{(n)} \mathbf{B}_t^{(n)T}$$

    with $\mathbf{B}_i^{(n)}$ and $\widehat{\mathbf{X}}_i^{(n)}$ the mode-n matricized forms of the tensors $\boldsymbol{\mathcal{B}}_i$ and $\boldsymbol{\mathcal{X}}_i$ defined by the equations (9) and (10)
    Update of $\mathbf{A}^{(n)}$

$$\mathbf{A}_{k+1}^{(n)} \leftarrow \arg\min_{\mathbf{A}^{(n)}} \widehat{f}_{n,t}(\mathbf{A}^{(n)}) \text{with } \widehat{f}_{n,t} \text{ defined by the equation (6)}$$

    $\mathbf{A}_t^{(n)} \leftarrow \mathbf{A}_{k+1}^{(n)}$
  end for
  $k \leftarrow k + 1$
end while
return $\mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}$

- The minimization problem related to the *Sparse coding* step admits a unique solution: this assumption can be enforced by considering a $\ell_1 - \ell_2$ penalty on the activation tensor. We drop the ℓ_2 penalty term for sake of simplicity and it does not require a great effort to see that it does not change anything to the reasoning related to the analysis;
- During the block coordinate descent, we seek $\mathbf{A}_{k+1}^{(n)}$ in a ball centered around $\mathbf{A}_k^{(n)}$ of radius $\frac{1}{k^{\frac{1}{2}}}$, i.e. $\|\mathbf{A}_{k+1}^{(n)} - \mathbf{A}_k^{(n)}\|_F^2 \leq \frac{1}{k}$. A natural way to incite the enforcing of this assumption is to increase the objective function associated to the update problem of $\mathbf{A}^{(n)}$ by $\rho\|\mathbf{A}^{(n)} - \mathbf{A}_k^{(n)}\|_F^2$;
- For $\forall k \in \mathbb{N}, \mathbf{A}_k^{(n)}$ is an interior point of $\mathbb{K}_{I_n,J_n}$;

Under these assumptions, the sequence $\left\{\mathbf{A}_k^{(1)}, .., \mathbf{A}_k^{(N)}\right\}$ generated from **Algorithm 3** (applied to the update problems of all dictionary matrices $\left\{\mathbf{A}^{(n)}\right\}_{1\leq n\leq N}$) converges to the set of stationary points of the objective function of the problem (3) when k tends to infinity with:

$$\mathbf{A}_{k+1}^{(n)} = \begin{cases} \lim_{t\to\infty} \mathbf{A}_{k+1,t}^{(n)} & \text{if } \mathbf{A}_{k+1,t}^{(n)} \text{ converges} \\ \arg\min_{\mathbf{A}^{(n)}} \widehat{f}_\infty(\mathbf{A}^{(n)}) & \text{otherwise} \end{cases}$$

$$\mathbf{A}_{k+1,t}^{(n)} = \underset{\mathbf{A}^{(n)}\in\mathbb{K}_{I_n,J_n}}{\arg\min}\ \widehat{f}_{n,t}(\mathbf{A}^{(n)})$$

$$\widehat{f}_\infty(\mathbf{A}^{(n)}) = \frac{1}{2}Trace\left(\left(\tilde{\xi}_t + \frac{\alpha(1-\theta)}{2}\mathbf{I}\right)\mathbf{A}^{(n)T}\mathbf{A}^{(n)}\right) - Trace\left(\mathbf{A}^{(n)}\tilde{\eta}_t\right)$$

$$\widehat{f}_{n,t}(\mathbf{A}^{(n)}) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}}_i \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \Omega_1(\boldsymbol{\mathcal{G}}_i) + \Omega_2(\mathbf{A}^{(1)}, .., \mathbf{A}^{(N)})$$

In this expression, $Trace$ represents the trace of a matrix. The sequences $\tilde{\xi}_t$ and $\tilde{\eta}_t$ embody some accumulation points of the bounded sequences ξ_t and η_t defined by:

$$\xi_t = \frac{1}{t}\sum_{i=1}^{t}\Gamma_i\Gamma_i^T, \eta_t = \frac{1}{t}\sum_{i=1}^{t}\Gamma_i\mathbf{X}_i^{(n)T}, \Gamma_i = \mathbf{G}_i^{(n)} \otimes_{q\in I_{n-1}} \mathbf{A}_{k+1}^{(q)T} \otimes_{p\in I_N^{n+1}} \mathbf{A}^{(p)T}$$

Sketch of the proof

To establish our convergence result, we setup a two-step strategy:

First step: firstly, we prove that the sequence $\left\{\mathbf{A}_{k+1,t}^{(n)}\right\}_t$ for a fixed k converges to a stationary point of the function $\mathbf{A}^{(n)} \to f(\mathbf{A}_{k+1}^{(1)}, .., \mathbf{A}_{k+1}^{(n-1)}, \mathbf{A}^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)})$, f being defined by the equation (3). This is mainly inspired from the work performed in [21], which cannot be applied directly to our problem since tha objective functions are different (for our problem, we have several dictionary matrices while the work described in [21] considers a single dictionary matrix). To circumvent this deadlock, we rely on some simple algebraic

inequalities and some well known properties on the Kronecker product.
Second step: from the first step as well as the assumptions laid out in this section, we prove that $\left\{\mathbf{A}_k^{(1)}, ...\mathbf{A}_k^{(N)}\right\}$ converges to a stationary point of the optimization problem given the equation (3).

5. Extensions

The purpose of this section is to propose some extensions of *OTL* through the incorporation of classical constraints encountered in tensor decomposition such as positivity or orthogonality as well as an extension allowing to process more than one sample at a time.

5.1. Minibatch extension

So far, we devise an algorithm by assuming that the newly incoming data is a single tensor $\boldsymbol{\mathcal{X}}_t \in \mathbb{R}^{I_1 \times ... \times I_N}$. Now, let's assume that the newly acquired data samples represent a batch of ρ tensors $\{\boldsymbol{\mathcal{X}}_1, .., \boldsymbol{\mathcal{X}}_\rho\}, \boldsymbol{\mathcal{X}}_i \in \mathbb{R}^{I_1 \times ...; \times I_N}$. Instead of applying the natural strategy, that is about processing one tensor at a time, we propose the following framework:

- Run the *Sparse coding* algorithm (**Algorithm** 1) ρ times in order to infer the activation tensors associated to the samples $\boldsymbol{\mathcal{X}}_j, 1 \leq j \leq \rho$ denoted $\boldsymbol{\mathcal{G}}_r$.
- Run the **Algorithm 2** by replacing (12) and (13) by:
  $\mathbf{P}_t^{(n)} = \mathbf{P}_{t-1}^{(n)} + \sum_{r=1}^{\rho} \widehat{\mathbf{X}}_r^{(n)} \mathbf{G}_r^{(n)T}$, $\widehat{\mathbf{X}}_r^{(n)}$ is the mode-n matricized form of the tensor $\widehat{\boldsymbol{\mathcal{X}}}_r$ defined by (10)
  $\mathbf{Q}_t^{(n)} = \mathbf{Q}_{t-1}^{(n)} + \sum_{r=1}^{\rho} \mathbf{B}_r^{(n)} \mathbf{B}_r^{(n)T}$, $\mathbf{B}_r^{(n)}$ is the mode-n matricized form of the tensor $\boldsymbol{\mathcal{B}}_r$ defined by (9)

The idea of this extension is to reduce the complexity in time of the approach developed in section 3 by incorporating simultaneously the information carried by several tensors while accelerating the decrease of the objective function.

5.2. Decomposition with positivity constraints

In several applications, the data we deal with is inherently positive. For the tensor decomposition of nonnegative data, it is convenient to impose nonnegativity constraint on all the latent factors (activation tensor and the dictionary matrices) in order to keep the physical interpretability. To perform a nonnegative tensor factorization, we simply replace all the updates in **Algorithm 1** and **Algorithm 2** by projected gradient descent [26]. This choice is motivated by the efficiency of projected gradient descent to yield good local minima for convex minimization problems subject to linear constraints [26].

5.3. Orthogonality constraints

The orthogonality constraint play an important role in *Tucker* decomposition since it ensures essential uniqueness [27]. To infer orthogonal factors (i.e. the dictionary matrices are such that $\mathbf{A}^{(n)T}\mathbf{A}^{(n)} = I$), we apply the gradient descent method on the Stiefel manifold defined by $St(I_n, J_n) = \left\{\mathbf{A} \in \mathbb{R}^{I_n \times J_n}, \mathbf{A}^T\mathbf{A} = I\right\}$. Let's consider the update problem of $\mathbf{A}^{(n)}$ subject to orthogonality constraints and let's denote $\mathbf{A}^{(n)}_{iter}$ the value at the iteration iter. The difference with the classical gradient descent lies in the update scheme given by:

$$\mathbf{A}^{(n)}_{iter+1} = \Psi_{\mathbf{A}^{(n)}_{iter}}\left(-\eta \times \Pi_{\tau_{\mathbf{A}^{(n)}_{iter}}(St(I_n,J_n))}\left(\frac{\partial \widehat{f}_{n,t}}{\partial \mathbf{A}^{(n)}}(\mathbf{A}^{(n)}_{iter})\right)\right)$$

with $\frac{\partial \widehat{f}_{n,t}}{\partial \mathbf{A}^{(n)}}\left(\mathbf{A}^{(n)}_{iter}\right)$ being the derivative of $\frac{\partial \widehat{f}_{n,t}}{\partial \mathbf{A}^{(n)}}(\cdots)$ evaluated at $\mathbf{A}^{(n)}_{iter}$. The operator $\Pi_{\tau_{\mathbf{A}^{(n)}_{iter}}(St(I_n,J_n))}$ corresponds to the projection on the tangent space to $St(I_n, J_n))$ at $\mathbf{A}^{(n)}_{iter}$ and $\Psi_{\mathbf{A}^{(n)}_{iter}}$ the retraction on $St(I_n, J_n)$ at $\mathbf{A}^{(n)}_{iter}$. There are several ways to compute the retraction on a Stiefel manifold such as polar decomposition [28], QR decomposition [28]. The projection of $\mathbf{A}$ on the tangent space to $St(p, q)$ at $\mathbf{X}$ is given by [29]:

$$(I_p - \mathbf{X}\mathbf{X}^T)\mathbf{A} + \frac{1}{2}\mathbf{X}(\mathbf{X}^T\mathbf{A} - \mathbf{A}^T\mathbf{X})$$

The update scheme of the activation tensor $\boldsymbol{\mathcal{G}}_t$ remains unchanged.

6. Experiments

In this section, we evaluate the efficiency of our approach on synthetic and real data sets. The methods considered are the following ones:

- *TuckerBatch*: this approach infers the latent factors by processing all the samples at once. The resolution scheme is alternate minimization: the update problems of the dictionary matrices and the activation tensor are respectively solved by gradient descent and proximal gradient descent for the unconstrained case. We highlight the fact that we do not impose orthogonality constraint as it is the case for most works on *Tucker* decomposition.
- *OTLsingle*: this corresponds to the online approach presented in this paper;
- *OTLminibatch*: this method is the proposed extension of *OTLsingle*;
- *ALTO*: this method, proposed in [18], is a state-of-the-art online tensor decomposition technique. Its difference with *OTLsingle* is based on the update strategy of the latent factors: projection on randomly generated subspaces for *ALTO*, numerical resolution of minimization problems for *OTLsingle*. The *ALTO* method yields convergence property if the data samples verify the so-called low rank property [18], that is: the rank of each mode-n matricized form is no greater than R, with (R,R,R) being the size of the core tensor.

The objectives of these experiments are two-fold.

- Firstly, we prove that our approach achieves similar results compared to the batch-method with much less computation time, i.e. the gain at this point is mainly about running time.
- Secondly, we prove the genericity of our approach: we demonstrate numerically that the proposed method yields similar results compared to *ALTO* in applications framework "favorable" to *ALTO* while significantly outperforming it when we are dealing with data with specificities that do not match *ALTO* assumptions. Contrary to the first point, the gain at this point is about accuracy.

We also consider our *OTL* approach with orthogonal and positivity constraints. These constrained decompositions will be referred to as *OTL* followed by the constraint name. The penalty functions Ω_1 and Ω_2 are respectively defined by the ℓ_1 norm and the square of Frobenius norm, i.e.:

$$\Omega_1(\boldsymbol{\mathcal{G}}) = \alpha\theta\|\boldsymbol{\mathcal{G}}\|_1, \Omega_2(\mathbf{A}^{(1)}, ..., \mathbf{A}^{(N)}) = \frac{\alpha(1-\theta)}{2}\sum_{n=1}^{N}\|\mathbf{A}^{(n)}\|_F^2, \alpha > 0, 0 \leq \theta < 1$$

.

The penalty functions verify the assumptions on Ω_1 and Ω_2 since Ω_1 admits a proximal operator, that is the *Soft thresholding* [23] and Ω_2 is differentiable. The retraction chosen to incorporate orthogonal constraint is the QR-based one [29], which enforces the orthogonality by retaining the Q matrix of the QR decomposition.

6.1. Spatio-temporal data prediction

Spatio-temporal forecasting is the task of predicting the future values of multivariate time series given historical observations. For this task, we use the Var(L) model [18]. Given an three-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{P\times T\times M}$ (whose orders represent respectively the locations, the time and the variables), this model seeks to determine a parameter tensor $\boldsymbol{\mathcal{W}}$ [18] through the numerical resolution of the minimization problem given by:

$$\min_{\boldsymbol{\mathcal{W}}}\left\{f(\boldsymbol{\mathcal{W}}) \triangleq \|\boldsymbol{\mathcal{X}} - \widehat{\boldsymbol{\mathcal{X}}}\|_F^2 + \mu\sum_{m=1}^{M}\text{Trace}\left(\widehat{\mathbf{X}}_{:,:,m}^T\mathbf{S}\widehat{\mathbf{X}}_{:,:,m}\right)\right\} \tag{14}$$

subject to:

$$\widehat{\mathbf{X}}_{:,t,m} = \mathbf{W}_{:,:,m}\mathbf{X}_{:,t,m}^L$$
$$\mathbf{X}_{:,t,m}^L = [\mathbf{X}_{:,t-1,m}^T, .., \mathbf{X}_{:,t-L,m}^T]^T, \text{rank}(\mathbf{W}^{(n)}) \leq R$$

The matrices $\mathbf{W}_{:,:,m}$ and $\mathbf{W}^{(n)}$ respectively correspond to the subtensor derived from $\boldsymbol{\mathcal{W}}$ fixing its third index to m and to the mode-n matricized forms of $\boldsymbol{\mathcal{W}}$. The functions *Trace* and *rank* represent the trace and the rank of a matrix. The role of the similarity matrix $\mathbf{S}$, predefined by user, is to ensure local consistency (i.e. the variables of interest are not

significantly different for two close locations). The parameter L is referred to as the number of lags. It is worth to notice that this problem cannot be split into independent minimization problems (with respect to the variables $\{\mathbf{W}_{:,:,m}\}_{1\leq m\leq M}$) due to the constraint on the rank of the matricized forms of $\boldsymbol{\mathcal{W}}$.

6.1.1. Resolution scheme

An alternative expression of $f(\boldsymbol{\mathcal{W}})$ is given by:

$$f(\boldsymbol{\mathcal{W}}) = \sum_{t=1}^{T}\sum_{m=1}^{M} \|\mathbf{X}_{:,t,m} - \mathbf{W}_{:,:,m}\mathbf{X}^{L}_{:,t,m}\|_F^2 + \sum_{t=1}^{T}\sum_{m=1}^{M} \mathbf{z}_t \left(\mathbf{W}^{T}_{:,:,m}\mathbf{S}\mathbf{W}_{:,:,m}\right)\mathbf{z}_t^T \tag{15}$$

with $\mathbf{z}_t = (\mathbf{X}^{L}_{:,:,m})^{T}_{t,:}$ representing t^{th} row of the matrix $(\mathbf{X}^{L}_{:,:,m})^T$.
A resolution scheme named *ALTO* has been proposed in [18] for the problem (14). The principle is to infer the parameter tensor $\boldsymbol{\mathcal{W}}$ through a two-step approach. The first step is to split the tensor $\boldsymbol{\mathcal{X}}$ into subtensors along the second mode representing the time steps and sequentially update the parameter tensor $\boldsymbol{\mathcal{W}}$ (this is possible because of the equation (15)). Let's denote $\boldsymbol{\mathcal{W}}_t$ the t^{th} parameter tensor of the sequence. The second step consists to project $\boldsymbol{\mathcal{W}}_t$ in a low rank subspace via random projections using factors inferred from $\boldsymbol{\mathcal{W}}_{t-1}$. The second step of *ALTO* follows the same principle as *OTL* in the sense that the two methods update the factors of $\boldsymbol{\mathcal{W}}_t$ using the latent factors inferred from $\boldsymbol{\mathcal{W}}_{t-1}$. Hence, to perform a fair comparison between our method and *ALTO*, we simply choose a two-step approach: the first is identical to the resolution scheme of *ALTO* and the second is replaced by *OTLsingle*. This resolution strategy will be referred to as *OTLsingle*. We consider a third approach referred to as *TuckerBatch* following the same principle as *ALTO* with the second step replaced by standard *Tucker* decomposition.

6.1.2. Data sets

Synthetic data set: we consider a synthetic data set $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{P\times T\times M}$ defined by:
$\boldsymbol{\mathcal{X}}_{:,t,m} = \boldsymbol{\mathcal{W}}_{:,:,m}\boldsymbol{\mathcal{Y}}_{:,t,m} + \boldsymbol{\mathcal{E}}_{:,t,m}$ with $\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times_3 \mathbf{A}^{(3)}$, $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{R\times R\times R}$ being a diagonal tensor (i.e. all its extra-diagonal elements are equal to 0) whose r^{th} diagonal element is equal to r. The entries of the matrices $\left\{\mathbf{A}^{(n)}\right\}_{1\leq n\leq 3}$ are drawn from a uniform distribution on $[0, 1]$ and the noise $\boldsymbol{\mathcal{E}}$ is a tensor whose entries are drawn from a centered Gaussian distribution with standard deviation $\frac{1}{2}$.

The similarity matrix $\mathbf{S}$ is defined by: $\mathbf{S} = \mathbf{B}\mathbf{B}^T + 10^{-1} \times \mathbf{I}$ ($\mathbf{I}$ being the identity matrix and $\mathbf{B}$ a random matrix whose entries are drawn from a centered Gaussian distribution with standard deviation $\frac{1}{2}$). This choice is motivated by the fact that the problem (14) is easy to solve when the similarity matrix is positive definite [18]. For fairness purpose, all the parameters that the methods have in common have been chosen identically. The parameters α and θ are respectively fixed to 10^2 and 10^{-2}. The initial activation tensor as well as the initial dictionary matrices entries are drawn from a standard Gaussian distribution. The number of epochs is fixed to 1.

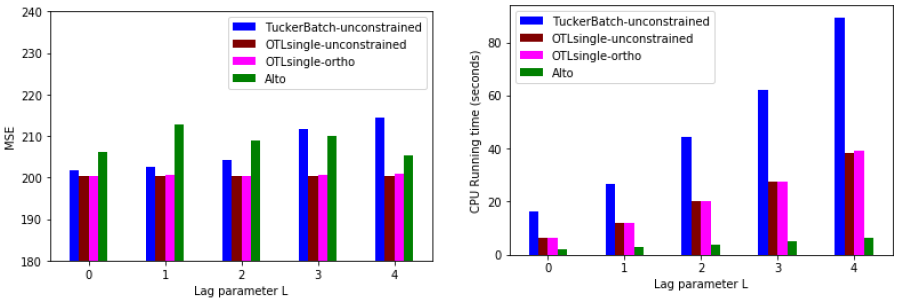

Figure 1: Synthetic spatio-temporal prediction problem: left (MSE over 10 different runs), right (CPU running time)

Real data set: we use a dataset containing 121 users ratings in the Pittsburgh area for 1200 intervals of time back on 4 months for different types of venue.
The value of the rank is determined by 5-fold cross-validation among 10 values ranging from 3 to 12. The hyperparameters α and θ are fixed to 10^2 and 10^{-2} and the number of epochs to 1.

6.1.3. Evaluation criterion

The evaluation criterion is the Mean Square Error defined by:

$$MSE = \frac{1}{MT}\sum_{m=1}^{M}\sum_{t=L+1}^{T} \|\mathbf{X}_{:,t,m} - \mathbf{W}_{:,:,m}\mathbf{X}^{L}_{:,t,m}\|_F^2$$

The parameter tensor $\boldsymbol{\mathcal{W}}$ is determined by processing a training tensor $\boldsymbol{\mathcal{X}}_{train}$ and the evaluation criterion evaluated using a test tensor $\boldsymbol{\mathcal{X}}_{test}$ different from $\boldsymbol{\mathcal{X}}_{train}$.

6.1.4. Results

Synthetic data set
The Mean Square Error values over 10 runs (the tensor $\boldsymbol{\mathcal{X}}$ is defined by 10 different random seeds) and the CPU running time with respect to the parameter L are reported on the figure 1. Our approach yields results very close to those obtained by *Tucker* decomposition with less running time and outperforms *ALTO* in term of Mean Square Error. This is explained by the fact that the bias induced by solving minimization problem is less important [30] than the one induced via random projections. The *ALTO* approach is faster than the two other methods because it simply performs projections on randomly generated subspaces while the two other approaches numerically solve minimization problems.
Real data set
The figure 2 provide the Mean Square Errors and the CPU running time. Our resolution scheme outperforms *Tucker* and *ALTO* methods with much less running time compared to *Tucker* and much important running time compared to *ALTO*.

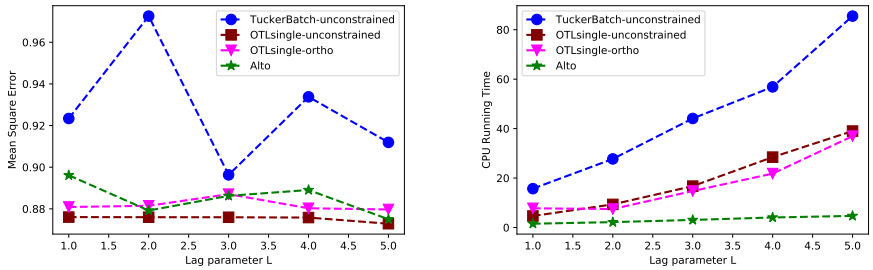

Figure 2: Real spatio-temporal prediction problem: left (MSE), right (CPU running time)

6.2. Multimodal dictionary learning

6.2.1. Dictionary learning on synthetic data set

The synthetic data set considered is a sequence of 20000 three-order tensors of size $30\times30\times30$ split into two sets $\mathbf{S}_{train}$ and $\mathbf{S}_{test}$ whose sizes are respectively equal to 15000 and 5000. The set $\mathbf{S}_{train}$ is used for the training stage, i.e. the inference of the three dictionary matrices and the evaluation criterion is determined on the test set $\mathbf{S}_{test}$. Each tensor $\boldsymbol{\mathcal{X}}_t$ (for both the training and the test sets) is defined by: $\boldsymbol{\mathcal{X}}_t = \boldsymbol{\mathcal{G}}_t \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times_3 \mathbf{A}^{(3)}$.
The tensor $\boldsymbol{\mathcal{G}}_t \in \mathbb{R}^{R\times R\times R}$ is a three-order tensor whose entries are drawn from a centered Gaussian distribution with standard deviation $\frac{1}{5}$. The entries of the matrices $\mathbf{A}^{(1)} \in \mathbb{R}^{30\times R}$, $\mathbf{A}^{(2)} \in \mathbb{R}^{30\times R}$, $\mathbf{A}^{(3)} \in \mathbb{R}^{30\times R}$ are drawn from a centered Gaussian distribution with standard deviation $\frac{1}{10}$. For the dictionary matrices inference, the value of the activation tensor rank is fixed to (R,R,R), i.e. the value used for the sample definition in order to prevent biased comparison. For fairness purpose, we choose the same initial points for the four competitors. The activation tensors are initialized by drawing random number from a centered Gaussian distribution with standard deviation $\frac{1}{10}$ and the dictionary matrices from the same distribution with standard deviation $\frac{1}{100}$. The gradient descent step is fixed to $\eta = 10^{-5}$. The hyperparameters α and θ are respectively fixed to 10^2 and 10^{-2}. The block coordinate descent is stopped when the fitting error is inferior to a threshold fixed to 10^{-3} or when 20 iterations is reached (an iteration being the updates of all the dictionary matrices $\left\{\mathbf{A}^{(n)}\right\}_{1\leq n\leq 3}$). The "*Sparse coding*" (**Algorithm 1**) and the "*Update of* $\mathbf{A}^{(n)}$" (**Algorithm 2**) algorithms are stopped when the fitting error is inferior to 10^{-3} or 20 iterations (an iteration being an update during the gradient descent) are reached. The number of epochs is fixed to 1. The evaluation criterion is the Root Mean Square Error RMSE on the test set defined by:

$$RMSE = \left(\frac{1}{T}\sum_{\boldsymbol{\mathcal{X}}_t\in\mathbf{S}_{test}} \|\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{G}}_t \times_1 \mathbf{A}_s^{(1)} \times_2 \mathbf{A}_s^{(2)} \times_3 \mathbf{A}_s^{(3)}\|_F^2\right)^{\frac{1}{2}}$$

with T being the cardinality of $\mathbf{S}_{test}$, $\mathbf{A}_s^{(n)}$ the factors inferred from the processing of $\mathbf{S}_{train}$ and $\boldsymbol{\mathcal{G}}_t$ the projection of $\boldsymbol{\mathcal{X}}_t$ on $\left\{\mathbf{A}_s^{(n)}\right\}_{1\leq n\leq 3}$.

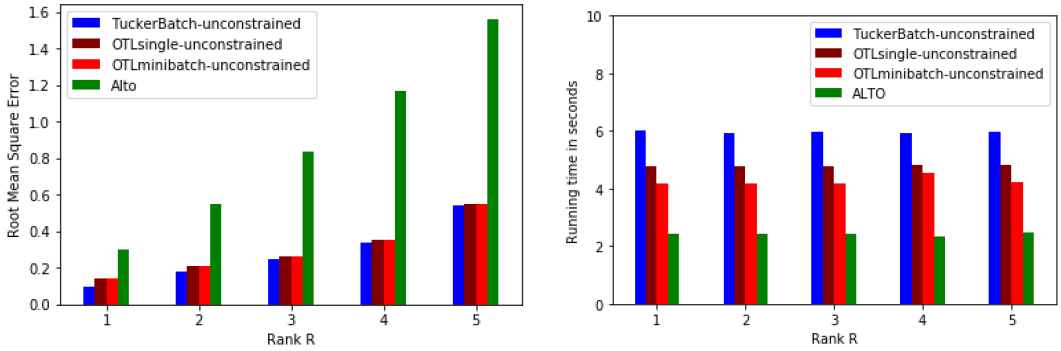

Figure 3: Left: Root Mean Square Error over three runs, Right: CPU running time on log_{10} scale

The results are reported on the figure 3. The predictive error of our approach is three times less important than the error of *ALTO*. This is due to the fact that the synthetic data does not verify the low rank assumption. Besides, it performs as well as the Batch method with less running time.

6.2.2. Nonnegative multimodal dictionary learning

For this experiment, we compare our approach to *ALTO* for the dictionary inference from a data set subject to positivity constraint (i.e. all the data samples have positive entries). The evaluation is performed on the inpainting task, which aims to infer missing pixels in an image (2D in our case). For this application, the training data set $\mathbf{S}_{train}$ is constructed by selecting overlapping patches $\boldsymbol{\mathcal{P}}_t$ of size 8×8 from the input image and the dictionary matrices (length, width dictionary matrices respectively denoted by $\mathbf{A}_l \in \mathbb{R}^{8\times R}$, $\mathbf{A}_w \in \mathbb{R}^{8\times R}$, R being an integer whose value has to be defined) are learned from $\mathbf{S}_{train}$ through a nonnegative decomposition (motivated by the positivity of the pixel values in an image) . Each patch is then inpainted by estimating the sparse coefficients through the projection of the non-missing pixels on the learned dictionaries, i.e.:

$$\widehat{\boldsymbol{\mathcal{P}}}_t = \boldsymbol{\mathcal{G}}_t \times_1 \mathbf{A}_l \times_2 \mathbf{A}_w, \boldsymbol{\mathcal{G}}_t \leftarrow \underset{\boldsymbol{\mathcal{G}}\geq 0}{\arg\min} \|\mathbf{M} \odot (\boldsymbol{\mathcal{X}}_t - \boldsymbol{\mathcal{G}} \times_1 \mathbf{A}_l \times_2 \mathbf{A}_w)\|_F^2 + \lambda \|\boldsymbol{\mathcal{G}}\|_1$$

$\widehat{\boldsymbol{\mathcal{P}}_t}$ is the reconstructed patch, $\mathbf{M}$ is the matrix defining the mask and $\odot$ the element-wise multiplication. The sparsity level λ is fixed to 10^{-3} and the number of epochs to 5. The evaluation criterion is the PSNR defined by:

$$PSNR = 10 \log_{10} \left(\frac{m \times n \times 255^2}{\|\widehat{\mathbf{I}} - \mathbf{I}\|_F^2} \right)$$

The variables $m, n, \mathbf{I}, \widehat{\mathbf{I}}$ respectively represent the image sizes, the real and the reconstructed images.
The inpainting task results are illustrated on the figures 4, 5 and 6.

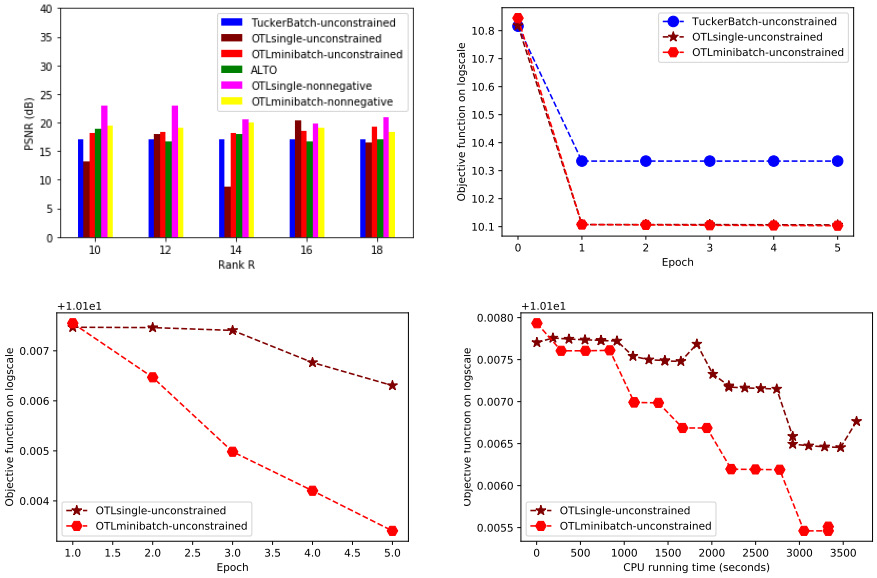

Figure 4: Top left: average PSNR over three different generations of the random pixels, Top right: objective function on logarithmic scale with respect to the number of epochs (one epoch is one pass over data samples) for Rank=16, Low left: objective function on logarithmic scale with respect to the number of epochs (one epoch is one pass over data samples) for Rank=16, Low right: Objective function with respect to time

As expected, we notice that the nonnegativity constraint improves the image quality in terms of PSNR, i.e. the nonnegative *OTLsingle* method yields greater *PSNR* than the unconstrained approach and the image quality is better visually. The same remark holds for the *OTLminibatch* method. It is worth to notice that all of our constrained approaches yield better result compared to *ALTO*. This justifies empirically the importance of setting up decomposition methods that take into account the characteristics of the data we are dealing with (positivity in this experiment).
A part of figure 4 presents the convergence of the objective value with respect to the number of epochs and confirms our intuition that is the faster convergence of *OTLminibatch* (see figure 4: Low left and Low right) compared to *OTLsingle* as well as the faster convergence of our approach with respect to *TuckerBatch* (see figure 4: Top left). The figure 4 proves that for a fixed number of epochs, our objetcive function decreases and for a fixed number of iteration, it learns better with the number of epochs 7: this prouves empirically the convergence of our approach.

7. Conclusion

In this paper, we propose an online tensor decomposition approach named *OTL* inspired from matrix-based online dictionary learning. The approach proposed in this work is used to perform a multimodal dictionary learning task via online *Tucker* decomposition while being enough flexible to incorporate common constraints that are frequently encountered in signal processing, to name but a few *sparsity*, *nonnegativity*, *orthogonality*. A theoretical

Figure 5: inpainting results from left to right: image with 50% of missing pixels, real image, *Nonnegative OTLsingle*, *Unconstrained OTLsingle*

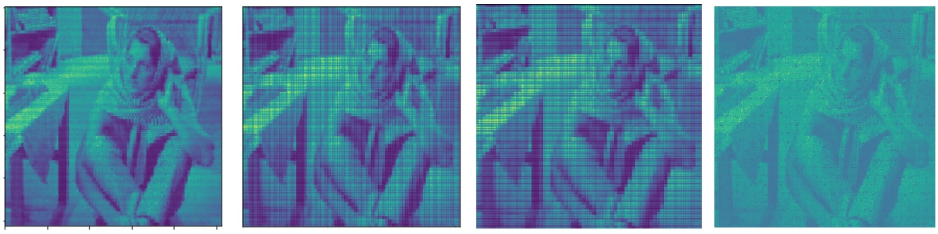

Figure 6: inpainting results from left to right : *OTLminibatch-nonnegative*, *OTLminibatch-unconstrained*, *Tucker-unconstrained*, *ALTO*

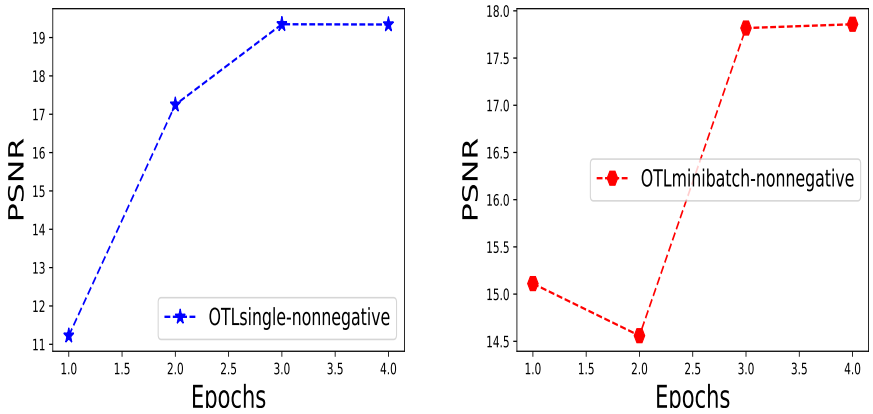

Figure 7: PSNR over the number of epochs for OTLsingle and OTLminibatch

result is provided along with numerical experiments. In our numerical experiments, we prove the promising character of our approach with respect to standard techniques. Precisely, we demonstrate that our approach produces a prediction error similar to the one obtained with a batch-based technique with less running time. Besides, the comparison with a state-of the-art-approach *ALTO* yields promising results under much milder convergence conditions (i.e. with no constraints on the latent factors sizes). In future work, we aim to investigate a supervised extension of this approach as well as an extension for a tensor which simultaneously grows in every mode.

References

[1] A. H. Phan, A. Cichocki, Extended hals algorithm for nonnegative tucker decomposition and its applications for multiway analysis and classification, Neurocomput. 74 (11) (2011) 1956–1969.

[2] L. R. Tucker, Implications of factor analysis of three-way matrices for measurement of change, C.W. Harris (Ed.), Problems in Measuring Change, University of Wisconsin Press (1963) 122–137.

[3] F. L. Hitchcock, The expression of a tensor or a polyadic as a sum of products, J. Math.Phys. 6 (1) (1927) 164–189.

[4] R. B. Cattell, Parallel proportional profiles and other principles for determining the choice of factors by rotation, Psychometrika 9 (4) (1944) 267–283.

[5] A. Cichocki, R. Zdunek, A. H. Phan, S.-I. Amari, Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation, Chichester, UK: John Wiley and Sons, Ltd, (2009) .

[6] A. Rajwade, A. Rangarajan, A. Banerjee, Image denoising using the higher order singular value decomposition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (4) (2013) 849–862.

[7] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, H.-J. Zhang, Multilinear discriminant analysis for face recognition, IEEE Transactions on Image Processing 16 (1) (2006) 212–220.

[8] M. A. O. Vasilescu, D. Terzopoulos, Multilinear analysis of image ensembles: Tensorfaces, in: Proceedings of the 7th European Conference on Computer Vision-Part I, 2002, pp. 447–460.

[9] S. Zubair, W. Wang, Tensor dictionary learning with sparse tucker decomposition, International conference on digital signal processing(DSP) (2013) 1–6.

[10] F. Huang, A. Anandkumar, Convolutional dictionary learning through tensor factorization, The 1st International Workshop Feature Extraction: Modern Questions and Challenges (2015) 116–129.

[11] J. Xu, J. Zhou, P.-N. Tan, X. Liu, L. Luo, Wisdom: Weighted incremental spatio-temporal multi-task learning via tensor decomposition, International Conference on Big Data (2016) 522–531.

[12] Q. Shi, Y. ming Cheung, Q. Zhao, Feature extraction for incomplete data via low-rank tucker decomposition, ECML PKDD (2017) 564–581.

[13] T. Variddhisa, D. P. Mandic, Online multilinear dictionary learning for sequential compressive sensing, CoRR, abs/1703.02492, (2017) .

[14] H. Kasai, B. Mishra, Low-rank tensor completion:a riemannian manifold preconditioning approach, ICML 48 (2016) 1012–1021.

[15] X. Li, W. Hu, Z. Zhang, X. Zhang, G. Luo, Robust visual tracking based on incremental tensor subspace learning, ICCV (2007) 1–8.

[16] S. Zhou, N. X. Vinh, J. Bailey, Y. Jia, I. Davidson, Accelerating online cp decompositions for higher order tensors, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1375–1384.

[17] G. Zhou, A. Cichocki, Q. Zhao, S. Xie, Efficient nonnegative tucker decompositions: Algorithms and uniqueness, IEEE Transactions on Image Processing 24 (12) (2015) 4990–5003.

[18] R. Yu, D. Cheng, Y. Liu, Accelerated online low-rank tensor learning for multivariate spatio-temporal streams, in: Proceedings of the 32Nd International Conference on International Conference on Machine Learning, 2015, pp. 238–247.

[19] L. D. Lathauwer, B. D. Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.
[20] A. Traoré, M. Berar, A. Rakotomamonjy, Non-negative tensor dictionary learning, ESANN (2018) .
[21] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, ICML '09 (2009) 689–696.
[22] K. Slavakis, G. B. Giannakis, Online dictionary learning from big data using accelerated stochastic approximation algorithms, ICASSP (2014) 16–20.
[23] N. G. Polson, J. G. Scott, B. T. Willard, Proximal algorithms in statistics and machine learning, Statistical Science 30 (4) (2015) 559–581.
[24] S.-J. Kim, Online kernel dictionary learning, 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (2015) 103–107.
[25] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (1-2) (2014) 459–494.
[26] R. Zdunek, A. Cichocki, Fast nonnegative matrix factorization algorithms using projected gradient approaches for large-scale problems, Intell. Neuroscience 2008 (2008) 3:1–3:13.
[27] A. Cichocki, D. Mandic, L. D. Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, H. A. PHAN, Tensor Decompositions for Signal Processing Applications: From two-way to multiway component analysis, IEEE Signal Processing Magazine 32 (2) (2015) 145–163.
[28] P.-A. Absil, R. Mahony, R. Sepulchre, Optimization algorithms on matrix manifolds, Princeton University Press, Princeton, NJ, USA.
[29] G. Olikier, P.-A. Absil, L. D. Lathauwer, Variable projection applied to block term decomposition of higher-order tensors, LVA/ICA (2018) 139–148.
[30] P. Zhao, T. Zhang, Accelerating minibatch stochastic gradient descent using stratified sampling, CoRR (2014) .
[31] W. GANDER, Algorithms for the qr-decomposition, RESEARCH REPORT (1980) .

Appendix A. Complexity (in time) analysis

In this section, we present the gain of complexities of the mini-batch extension with respect to the standard approach in the constrained (positivity and orthogonality) and the unconstrained cases. The complexities (in time) per update for the *Sparse coding* and the dictionary update are given by the table A.1. From this table, we notice that the mini-batch extension requires less computations than the standard online approach. Indeed, for $\rho \geq 1$ sample units newly acquired $\{\boldsymbol{\mathcal{X}}_i\}_{1\leq i\leq \rho}$, let's consider two update strategies **StrategyI** and **StrategyII**, which respectively are about updating the dictionary matrices by applying the standard approach *OTLsingle* and its extension *OTLminibatch*. For the *Sparse coding*, the complexity per update is $\mathcal{O}(2\rho N \prod_{k=1}^{N} I_k J_k^2)$ for the two strategies. For the dictionary update stage, we have the following observations:

- Unconstrained decomposition: **StrategyII** requires $(\rho-1)\left(I_n \prod_{k=1}^{N} J_k + 13 I_n J_n\right)$ operations less than **StrategyI** per update;
- Nonnegative decomposition: the gain in terms of complexity of **StrategyII** with respect to **StrategyI** is equal to $(\rho-1)\left(I_n \prod_{k=1}^{N} J_k + 14 I_n J_n\right)$ per update;
- Orthogonal decomposition: **StrategyII** requires $(\rho-1)\left(I_n J_n + J_n^2 + I_n \prod_{k=1}^{N} J_k\right) + (\rho-1) J_n^2 \prod_{k=1,k\neq n}^{N} I_k$ less computations than **StrategyI** per update.

The proposed extension can then be a good alternative since it does not induce a substantial lost of information with respect to *OTLsingle* (proved through numerical experiments) while requiring less computations.

OTLsingle		
Steps / Constraints	Sparse coding	Update of $\mathbf{A}^{(n)}$
Unconstrained	$N\prod_{k=1}^{N} I_k J_k + 2N\prod_{k=1}^{N} I_k J_k^2 + 10\prod_{k=1}^{N} J_k$	$6I_nJ_n + I_nJ_n^2 + J_n^2 + I_n\prod_{k=1}^{N} J_k + J_n^2\prod_{k\neq n} I_k$
Nonnegativity	$N\prod_{k=1}^{N} I_k J_k + 2N\prod_{k=1}^{N} I_k J_k^2 + 6\prod_{k=1}^{N} J_k$	$7I_nJ_n + I_nJ_n^2 + J_n^2 + I_n\prod_{k=1}^{N} J_k + J_n^2\prod_{k\neq n} I_k$
Orthogonality	$N\prod_{k=1}^{N} I_k J_k + 2N\prod_{k=1}^{N} I_k J_k^2 + 10\prod_{k=1}^{N} J_k$	$7I_nJ_n + 6I_nJ_n^2 + 2J_n^2 + I_n^2(1+2J_n) + I_n\prod_{k=1}^{N} J_k + J_n^2\prod_{k\neq n} J_k$
OTLminibatch		
Unconstrained	$\rho N\prod_{k=1}^{N} I_k J_k + 2\rho N\prod_{k=1}^{N} I_k J_k^2 + 10\rho\prod_{k=1}^{N} J_k$	$5I_nJ_n + I_nJ_n^2 + \rho(I_nJ_n + I_n\prod_{k=1}^{N} J_k + J_n^2 + J_n^2\prod_{k\neq n} I_k)$
Nonnegativity	$\rho N\prod_{k=1}^{N} I_k J_k + 2\rho N\prod_{k=1}^{N} I_k J_k^2 + 6\rho\prod_{k=1}^{N} J_k$	$6I_nJ_n + I_nJ_n^2 + \rho(I_nJ_n + I_n\prod_{k=1}^{N} J_k + J_n^2 + J_n^2\prod_{k\neq n} I_k)$
Orthogonality	$\rho N\prod_{k=1}^{N} I_k J_k + 2\rho N\prod_{k=1}^{N} I_k J_k^2 + 10\rho\prod_{k=1}^{N} J_k$	$I_n^2(1+2J_n) + 6I_nJ_n + 6I_nJ_n^2 + J_n^2 + \rho(I_nJ_n + I_n\prod_{k=1}^{N} J_k + J_n^2 + J_n^2\prod_{k\neq n} I_k)$

Table A.1: Complexity in time per update: gradient descent for $\mathbf{A}^{(n)}$ and proximal gradient descent for the activation tensor $\boldsymbol{\mathcal{G}}$

NB: these complexities have been computed by considering Ω_1 and Ω_2 as defined in the numerical experiments. The orthogonal constraint is imposed only on the dictionary matrices. The QR decomposition algorithm considered is the **Modified Gram-Schmidt** [31].

Convergence analysis

In the convergence analysis is about exploring a "double" asymptotic behavior in the sense we consider the behavior of the approach when the both the number of samples and the number of iterations in the block coordinate go to infinity.

Appendix B. Recall of the definitions and supplemental notations

Appendix B.1. ***Recall of definitions***

$$\min_{\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}} f(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)})$$

with

$$\begin{aligned} f(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}) &= \mathbb{E}_{\mathbb{P}}(l(\boldsymbol{\mathcal{X}},\mathbf{A}^{(1)},..,\mathbf{A}^{(N)})) \\ &= \lim_{t\to\infty} f_t(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}) \\ &= \lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} l(\boldsymbol{\mathcal{X}}_i,\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}) \end{aligned}$$

We denote by $\widehat{f}_{n,t}$ the function $\mathbf{A} \to \widehat{f}_t(\mathbf{A}_{k+1}^{(1)},..,\mathbf{A}_{k+1}^{(n-1)},\mathbf{A},\mathbf{A}_k^{(n+1)},..\mathbf{A}_k^{(N)})$. We prove that $\mathbf{A}_{k+1,t}^{(n)}$ converges to the set of stationary points of $\mathbf{A} \to f(\mathbf{A}_{k+1}^{(1)},..,\mathbf{A}_{k+1}^{(n-1)},\mathbf{A},\mathbf{A}_k^{(n+1)},..\mathbf{A}_k^{(N)})$. From this result, we prove that $\left\{\mathbf{A}_k^{(1)},..,\mathbf{A}_k^{(N)}\right\}$ converges to the set of stationary points of $\left(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}\right) \to f(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)})$. For the minimization $\widehat{f}_{n,t}$ with respect to $\mathbf{A}^{(n)}$ (defining the sequence $\mathbf{A}_{k+1,t}^{(n)}$: see section 4), we replace $\widehat{f}_{n,t}$ by $\widehat{f}_{n,t} - \frac{\alpha(1-\theta)}{2}(\sum_{p=1}^{n-1}\|\mathbf{A}_{k+1}^{(p)}\|_F^2 + \sum_{q=n+1}^{N}\|\mathbf{A}_k^{(q)}\|_F^2)$ since the terms dropped does not change anything to the minimization problem. Hence, in the sequel, we consider the following notations:

$$\mathbf{A}_{k+1,t}^{(n)} = \underset{\mathbf{A}^{(n)}\in\mathbb{K}_{I_n,J_n}}{\arg\min}\ \widehat{f}_{n,t}(\mathbf{A}^{(n)}) \tag{B.1}$$

$$l(\boldsymbol{\mathcal{X}},\mathbf{A}^{(n)}) = \min_{\boldsymbol{\mathcal{G}}} \frac{1}{2}\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{G}} \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}\|_1 + \frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2$$

$$f_n(\mathbf{A}^{(n)}) = \mathbb{E}_{\mathbb{P}}(l(\boldsymbol{\mathcal{X}},\mathbf{A}^{(n)}))$$

$$f_{n,t}(\mathbf{A}^{(n)}) = \frac{1}{t}\sum_{i=1}^{t} l(\boldsymbol{\mathcal{X}}_t,\mathbf{A}^{(n)})$$

$$\begin{aligned} \widehat{f}_{n,t}(\mathbf{A}^{(n)}) &= \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}}_i \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}_i\|_1 + \frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2 \\ &= \frac{1}{2}Trace(\xi_t\mathbf{A}^{(n)T}\mathbf{A}^{(n)}) - Trace(\mathbf{A}^{(n)}\eta_t) + \frac{1}{t}\sum_{i=1}^{t}(\frac{1}{2}\|\boldsymbol{\mathcal{X}}_i\|_F^2 + \Omega_1(\boldsymbol{\mathcal{G}}_i)) + \Omega_2(\mathbf{A}^{(1)},..,\mathbf{A}^{(N)}) \end{aligned}$$

The sequences ξ_t and η_t are defined by:

$$\xi_t = \frac{1}{t}\sum_{i=1}^{t}\Gamma_i\Gamma_i^T, \eta_t = \frac{1}{t}\sum_{i=1}^{t}\Gamma_i\mathbf{X}_i^{(n)T}, \Gamma_i = \mathbf{G}_i^{(n)} \otimes_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)T} \otimes_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)T} \tag{B.2}$$

with:

$$\boldsymbol{\mathcal{G}}_i = \arg\min_{\boldsymbol{\mathcal{G}}} \|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}} \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}_{i-1}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}\|_1 + \frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2$$

The derivatives of f_n and l with respect to $\mathbf{A}^{(n)}$ are simply referred to as ∂f_n and $\partial_2 l$.

Appendix B.2. ***Supplemental notations***

The norm on $\mathbb{R}^{m_1\times n_1} \times .. \times \mathbb{R}^{m_K\times n_K}$ ($\mathbb{R}^{m_k\times n_k}$ being the set of matrices of size $m_k \times n_k$), denoted by $\|\cdot\|_{\prod_{k\in I_K}\mathbb{R}^{m_k\times n_k}}$ is defined by: $\|(\mathbf{A}^{(1)}, .., \mathbf{A}^{(K)})\|_{\prod_{k\in I_K}\mathbb{R}^{m_k\times n_k}} = \sum_{k=1}^{K}\|\mathbf{A}^{(k)}\|_F$. For two matrices $\mathbf{A} \in \mathbb{R}^{M\times N}$ and $\mathbf{B} \in \mathbb{R}^{M\times N}$, the dot product is defined by: $\langle \mathbf{A}, \mathbf{B}\rangle = \sum_{1\le m\le M, 1\le n\le N} \mathbf{A}_{m,n}\mathbf{B}_{m,n}$. The supremum of a continuous function f defined on a compact domain $\mathbb{D}$ will be referred to as $\|f\|_\infty$: $\|f\|_\infty = \sup_{\mathbf{x}\in\mathbb{D}} \mid f(\mathbf{x}) \mid$. The sign function is denoted by *sign*.

Appendix C. Assumptions

- *Assumption I*: the data samples are drawn from a probability distribution with compact support: this a natural assumption since we work only with finite values.
- *Assumption II*: for all the minimization problems considered, the activation tensor $\boldsymbol{\mathcal{G}}$ and and the dictionary matrices $\mathbf{A}^{(n)}$ are assumed to belong to some compact sets, i.e. $\boldsymbol{\mathcal{G}} \in \mathbb{K}_{\boldsymbol{\mathcal{G}}}, \mathbf{A}^{(n)} \in \mathbb{K}_{I_n,J_n}$: this corresponds to the interpretation of $\ell_1 - \ell_2$ penalty.
- *Assumption III*: $\mathbf{C}_{:,j}\left(\text{vec}(\boldsymbol{\mathcal{X}}) - \mathbf{C}\,\text{vec}(\boldsymbol{\mathcal{G}})\right) = \theta\alpha\,\text{vec}(\boldsymbol{\mathcal{G}})_j)$ if $sign(\text{vec}(\boldsymbol{\mathcal{G}})_j) \neq 0$
  $|\mathbf{C}_{:,j}\left(\text{vec}(\boldsymbol{\mathcal{X}}) - \mathbf{C}\,\text{vec}(\boldsymbol{\mathcal{G}})\right)| \le \alpha\theta$ otherwise, with: $\mathbf{C} = \left(\otimes_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)}\right)\otimes\mathbf{A}\otimes\left(\otimes_{q\in I_N^{n+1}}\mathbf{A}_k^{(q)}\right)$
  This condition ensures the uniqueness of the *Sparse Coding* problem.
- *Assumption IV*: during the block coordinate descent, we seek $\mathbf{A}_{k+1}^{(n)}$ in a ball centered around $\mathbf{A}_k^{(n)}$ of radius $\frac{1}{k^{\frac{1}{2}}}$, i.e. $\|\mathbf{A}_{k+1}^{(n)} - \mathbf{A}_k^{(n)}\|_F^2 \le \frac{1}{k}$.

  A natural way to incite the enforcing of this assumption is to increase the objective function associated to the update problem of $\mathbf{A}^{(n)}$ by $\rho\|\mathbf{A}^{(n)} - \mathbf{A}_{k+1}^{(n)}\|_F^2$. Again, it is straightforward to see that this term does not change anything to the reasoning related to the analysis. Thus, it is dropped for writing simplicity;
- *Assumption V*: $\forall k \in \mathbb{N}, \mathbf{A}_k^{(n)}$ is an interior point of $\mathbb{K}_{I_n,J_n}$.

Algorithm 4 OTL-infinite

Inputs: $\left\{\mathbf{A}_0^{(n)}\right\}_{1\leq n\leq N}$**, the hyperparameters** $\alpha > 0$**,** $0 \leq \theta < 1$

for n from 1 to N **do**
for k from 1 to ∞ **do**

$$\mathbf{A}_{k+1}^{(n)} = \begin{cases} \lim_{t\to\infty} \mathbf{A}_{k+1,t}^{(n)} & \text{if } \mathbf{A}_{k+1,t}^{(n)} \text{ converges} \\ \arg\min_{\mathbf{A}^{(n)}} \widehat{f}_\infty(\mathbf{A}^{(n)}) & \text{otherwise} \end{cases}$$

with $\mathbf{A}_{k+1,t}^{(n)}$ defined by the equation (B.1)
end for
end for
return $\left\{\mathbf{A}^{(n)}\right\}_{1\leq n\leq N}$

Appendix D. Algorithm for infinite number of samples and infinite number of iterations

In the theoretical framework, we assume that the number of iterations as well as the number of samples are infinite. Then, the rewriting of the algorithm yields Algorithm 4.

In the sequel, the penalty functions considered are the ℓ_1 norm for Ω_1 (ℓ_1- penalty) and the square of the Frobenius norm for Ω_2 (l_2- penalty) as in the numerical experiments. The function $\widehat{f}_\infty$ is defined by:

$$\widehat{f}_\infty(\mathbf{A}^{(n)}) = \frac{1}{2} Trace\left(\left(\tilde{\xi}_t + \frac{\alpha(1-\theta)}{2}\mathbf{I}\right)\mathbf{A}^{(n)T}\mathbf{A}^{(n)}\right) - Trace\left(\mathbf{A}^{(n)}\tilde{\eta}_t\right)$$

where: $\tilde{\xi}_n$ and $\tilde{\eta}_t$ represent accumulation points of the sequences $\tilde{\xi}_t$ and $\tilde{\eta}_t$ defined by (B.2) whose existences are ensured by the boundedness of ξ_t and η_t.

Remark 1. *The differences between the algorithms* ***Algorithm ??*** *and* ***Algorithm 3*** *result from the gap between theory and numerical analysis: firstly, all the data samples are available at the same time and secondly, the sparse code of a sample* $\boldsymbol{\mathcal{X}}_t$ *is computed via projection on* $\left\{\mathbf{A}_{t-1}^{(1)}, .., \mathbf{A}_{t-1}^{(N)}\right\}$ *instead of* $\left\{\mathbf{A}_{k+1}^{(1)}, .., \mathbf{A}_{k+1}^{(n-1)}, \mathbf{A}_{t-1}^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)}\right\}$ *in order to alleviate the computational burden of the approach.*

Appendix E. Useful properties

Property 1. *For two matrices* $\mathbf{A}, \mathbf{B}$ *such that* $\|\mathbf{A}\|_F \leq \delta_1, \|\mathbf{B}\|_F \leq \delta_2$*. There exists* $M > 0$ *such that:*

$$\|\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2\| \leq M_1\|\mathbf{A} - \mathbf{B}\|_F$$

Proof:
$\|\mathbf{A}\|_F^2 = \|\mathbf{A} - \mathbf{B} + \mathbf{B}\|_F^2 = \|\mathbf{A} - \mathbf{B}\|_F^2 + 2 < \mathbf{A} - \mathbf{B}, \mathbf{B} > +\|\mathbf{B}\|_F^2$

$\Rightarrow \|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2 \leq \|\mathbf{A}-\mathbf{B}\|_F^2 + 2\|\mathbf{A}-\mathbf{B}\|_F\|\mathbf{B}\|_F$
$\Rightarrow \|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2 \leq \|\mathbf{A}-\mathbf{B}\|_F\|\mathbf{A}-\mathbf{B}\|_F + 2\delta_2\|\mathbf{A}-\mathbf{B}\|_F$
$\Rightarrow \|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2 \leq (\delta_1+\delta_2)\|\mathbf{A}-\mathbf{B}\|_F + 2\delta_2\|\mathbf{A}-\mathbf{B}\|_F$
Thus, we have:

$$\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2 \leq (\delta_1+3\delta_2)\|\mathbf{A}-\mathbf{B}\|_F \leq max(\delta_1+3\delta_2, \delta_2+3\delta_1)\|\mathbf{A}-\mathbf{B}\|_F \tag{E.1}$$

By permuting the role of $\mathbf{A}$ and $\mathbf{B}$, we have:

$$\|\mathbf{B}\|_F^2 - \|\mathbf{A}\|_F^2 \leq (\delta_2+3\delta_1)\|\mathbf{A}-\mathbf{B}\|_F \leq max(\delta_1+3\delta_2, \delta_2+3\delta_1)\|\mathbf{A}-\mathbf{B}\|_F \tag{E.2}$$

By combining (E.1) and (E.2), we have:
$\|\|\mathbf{A}\|_F^2 - \|\mathbf{B}\|_F^2\| \leq max(\delta_1+3\delta_2, \delta_2+3\delta_1)\|\mathbf{A}-\mathbf{B}\|_F$

Property 2. *For four matrices $\mathbf{A}_1, \mathbf{B}_1 \in \mathbb{R}^{N\times M}$ and $\mathbf{A}_2, \mathbf{B}_2 \in \mathbb{R}^{M\times L}$ such that $\|\mathbf{A}_1\|_F \leq \delta_1, \|\mathbf{B}_2\|_F \leq \delta_2$, the following inequality holds:*
$\|\mathbf{A}_1\mathbf{A}_2 - \mathbf{B}_1\mathbf{B}_2\|_F \leq \delta_1\|\mathbf{A}_2-\mathbf{B}_2\|_F + \delta_2\|\mathbf{A}_1-\mathbf{B}_1\|_F$

Proof:
$\|\mathbf{A}_1\mathbf{A}_2 - \mathbf{B}_1\mathbf{B}_2\|_F = \|\mathbf{A}_1(\mathbf{A}_2-\mathbf{B}_2+\mathbf{B}_2) - \mathbf{B}_1\mathbf{B}_2\|_F = \|\mathbf{A}_1(\mathbf{A}_2-\mathbf{B}_2) + \mathbf{A}_1\mathbf{B}_2 - \mathbf{B}_1\mathbf{B}_2\|_F$
$\Rightarrow \|\mathbf{A}_1\mathbf{A}_2 - \mathbf{B}_1\mathbf{B}_2\|_F \leq \delta_1\|\mathbf{A}_2-\mathbf{B}_2\|_F + \delta_2\|\mathbf{A}_1-\mathbf{B}_1\|_F$

Property 3. *If $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{G}} \times_1 \mathbf{A}^{(1)} .. \times_N \mathbf{A}^{(N)} = \boldsymbol{\mathcal{G}} \times_{n\in I_N} \mathbf{A}^{(n)}$, the mode-n matricized form of $\boldsymbol{\mathcal{X}}$ denoted $\mathbf{X}^{(n)}$ is defined by:*

$$\mathbf{X}^{(n)} = \mathbf{A}^{(n)}\mathbf{G}^{(n)}\left(\mathbf{A}^{(1)} \otimes .. \otimes \mathbf{A}^{(n-1)} \otimes \mathbf{A}^{(n+1)} ... \otimes \mathbf{A}^{(N)}\right)^T = \mathbf{A}^{(n)}\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(n)T}$$

This is the matricization property of the *Tucker* decomposition.

Property 4. *Let's denote $g(\mathbf{A}^{(n)}) = \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{G}} \times_1 \mathbf{A}^{(1)} .. \times_n \mathbf{A}^{(n)} .. \times_N \mathbf{A}^{(N)}\|_F^2$. The derivative of f is given by:*

$$\begin{aligned}\partial g(\mathbf{A}^{(n)}) &= -2\left(\mathbf{X}^{(n)} - \mathbf{A}^{(n)}\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}\right)\left(\otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)}\right)\mathbf{G}^{(n)T} \\ &= -2\left(\widehat{\mathbf{X}}^{(n)}\mathbf{G}^{(n)T} - \mathbf{A}^{(n)}\mathbf{B}^{(n)}\mathbf{B}^{(n)T}\right)\end{aligned}$$

with $\mathbf{X}^{(n)}$ and $\mathbf{B}^{(n)}$ being respectively the mode-n matricized forms of the tensors $\widehat{\boldsymbol{\mathcal{X}}}$ and $\boldsymbol{\mathcal{B}}$ defined by:
$\widehat{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}} \times_{p\in I_{n-1}} \mathbf{A}^{(p)T} \times_n \mathbf{I} \times_{q\in I_N^{n+1}} \mathbf{A}^{(q)T}, \mathbf{I} \in \mathbb{R}^{I_n\times I_n}$*: identity matrix*
$\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{G}} \times_{p\in I_{n-1}} \mathbf{A}^{(p)} \times_n \mathbf{I} \times_{q\in I_N^{n+1}} \mathbf{A}^{(q)}, \mathbf{I} \in \mathbb{R}^{J_n\times J_n}$*: identity matrix*

Proof:
$g(\mathbf{A}^{(n)}+h) = \|\mathbf{X}^{(n)} - (\mathbf{A}^{(n)}+h)\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}\|_F^2$
$\Rightarrow g(\mathbf{A}^{(n)}+h) = \|\mathbf{X}^{(n)} - \mathbf{A}^{(n)}\mathbf{G}^{(n)} \otimes_{m\in I_{\neq n}} \mathbf{A}^{(m)T} + h\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}\|_F^2$
$\Rightarrow g(\mathbf{A}^{(n)}+h) = \|\mathbf{X}^{(n)} - \mathbf{A}^{(n)}\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}\|_F^2 + \|h\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}\|_F^2 - 2\langle \mathbf{X}^{(n)} - \mathbf{A}^{(n)}\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}, h\mathbf{G}^{(n)} \otimes_{m\neq n} \mathbf{A}^{(m)T}\rangle$

$\Rightarrow g(\mathbf{A}^{(n)}+h) = g(\mathbf{A}^{(n)}) - 2\langle\left(\mathbf{X}^{(n)} - \mathbf{A}^{(n)}\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}\right) \otimes_{m\in I_{m\neq n}} \mathbf{A}^{(m)}\mathbf{G}^{(n)T}, h\rangle +$
$O\left(\|h\|_F^2\right)$.
$\Rightarrow \partial g(\mathbf{A}^{(n)}) = -2\left(\mathbf{X}^{(n)} - \mathbf{A}^{(n)}\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}\right) \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)}\mathbf{G}^{(n)T}$
$\Rightarrow \partial g(\mathbf{A}^{(n)}) = -2\mathbf{X}^{(n)}\left(\otimes_{m\in I_{N\neq n}}\mathbf{A}^{(m)}\right)\mathbf{G}^{(n)T} + 2\mathbf{A}^{(n)}\left(\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(n)T}\right)\left(\otimes_{m\in I_{N\neq n}}\mathbf{A}^{(n)}\mathbf{G}^{(n)T}\right)$
$\Rightarrow \partial g(\mathbf{A}^{(n)}) = -2\mathbf{X}^{(n)}\left(\otimes_{m\in I_{N\neq n}}\mathbf{A}^{(m)}\right)\mathbf{G}^{(n)T} + 2\mathbf{A}^{(n)}\left(\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(n)T}\right)\left(\mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(n)T}\right)^T$
(property of transpose of Kronecker product).
By the **Property** 3, the mode-n matricization of $\widehat{\boldsymbol{\mathcal{X}}}$ and $\boldsymbol{\mathcal{B}}$ are given by:
$\widehat{\mathbf{X}}^{(n)} = \mathbf{X}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(n)}$ and $\mathbf{B}^{(n)} = \mathbf{G}^{(n)} \otimes_{m\in I_{N\neq n}} \mathbf{A}^{(m)T}$
From this observation, it is straightforward to see that:
$\partial g(\mathbf{A}^{(n)}) = -2\widehat{\mathbf{X}}^{(n)}\mathbf{G}^{(n)T} + 2\mathbf{A}^{(n)}\mathbf{B}^{(n)}\mathbf{B}^{(n)T}$

Property 5. *If the loading factors and the core tensor are bounded, $\widehat{f}_{n,t}$ is bounded by a constant independent from t*

Proof:
This is straightforward by triangular inequality along with *Assumption II*

Property 6. *The function $\widehat{f}_{n,t}$ is strictly convex and Hessian lower bounded.*

Proof:
$\widehat{f}_{n,t}(\mathbf{A}) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}}_i \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}_i\|_1 + \frac{(1-\theta)\alpha}{2}\|\mathbf{A}\|_F^2$,
$0 \leq \theta < 1, \alpha > 0$.

Appendix E.0.1. Strict convexity

$\widehat{f}_{n,t}(\mathbf{A}) - \frac{(1-\theta)\alpha}{2}\|\mathbf{A}\|_F^2$ is a convex function. Hence $\widehat{f}_{n,t}(\mathbf{A})$ is strongly convex, which implies strict convexity.

Appendix E.0.2. Hessian (lower) boundedness

$\widehat{f}_{n,t}(\mathbf{A}^{(n)}) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{\mathcal{X}}_i - \boldsymbol{\mathcal{G}}_i \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}_i\|_1 + \frac{(1-\theta)\alpha}{2}\|\mathbf{A}^{(n)}\|_F^2$
By the **Property** 4, we have:
$\partial\widehat{f}_{n,t}(\mathbf{A}^{(n)}) = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\left(-2\widehat{\mathbf{X}}_i^{(n)}\mathbf{G}_i^{(n)T} + 2\mathbf{A}^{(n)}\mathbf{B}_i^{(n)}\mathbf{B}_i^{(n)T}\right) + \alpha(1-\theta)\mathbf{A}^{(n)}$ with:
$\mathbf{B}_i^{(n)} = \mathbf{G}_i^{(n)}\left(\otimes_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)T} \otimes_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)T}\right)$, $\widehat{\mathbf{X}}_i^{(n)} = \mathbf{X}_i^{(n)}\left(\otimes_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)} \otimes_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\right)$
Hence, we have:
$\partial\widehat{f}_{n,t}(\mathbf{A}^{(n)}) = \mathbf{A}^{(n)}\mathbf{S}_t - \mathbf{M}_t + \alpha(1-\theta)\mathbf{A}^{(n)}$ with $\mathbf{S}_t = \frac{1}{t}\sum_{i=1}^{t}\mathbf{B}_i^{(n)}\mathbf{B}_i^{(n)T}$, $\mathbf{M}_t = -\frac{1}{t}\sum_{i=1}^{t}\widehat{\mathbf{X}}_i^{(n)}\mathbf{G}_i^{(n)T}$
Thus, the second derivative with respect to $\mathbf{A}^{(n)}$ yields:
$\partial^2\widehat{f}_{n,t}(\mathbf{A}^{(n)}) = \mathbf{S}_t^T \otimes I_{I_n} + (1-\theta)\alpha I_{I_nJ_n}$, $I_{I_nJ_n} \in \mathbb{R}^{I_nJ_n \times I_nJ_n}$ and $I_{I_n} \in \mathbb{R}^{I_n \times I_n}$ being the identity matrices
$\Rightarrow \partial^2 f_{n,t}(\mathbf{A}^{(n)}) = \mathbf{S}_t \otimes I_{I_n} + (1-\theta)\alpha I_{I_nJ_n}$ (symmetry of $\mathbf{S}_t$)
Simple computations along with the mixed product property on the Kronecker operator yield:
$\mathbf{S}_t \otimes \mathbf{I}_{I_n} = \frac{1}{t}\sum_{i=1}^{t}(\mathbf{B}_i^{(n)} \otimes I_{I_n})(\mathbf{B}_i^{(n)} \otimes I_{I_n})^T$
Hence, $\partial^2 f_{n,t}(\mathbf{A}^{(n)})$ is a symmetric positive definite matrix with all eigenvalues lower bounded by $(1-\theta)\alpha$.

Property 7. *The lost function l defined in the section Appendix B is a Lipschitz function with respect to $\mathbf{A}^{(n)}$ and uniformly bounded.*

Proof:

A. Boundedness

This is straightforward by triangular inequality and the fact that the loading matrices and the core tensor are bounded.

B. l est Lipschitz with respect to $\mathbf{A}^{(n)}$

The function $\mathbf{A}^{(n)} \to \frac{1}{2}\|\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{G}}\times_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)}\times_n\mathbf{A}^{(n)}\times_{q\in I_N^{n+1}}\mathbf{A}_k^{(q)}\|_F^2+\alpha\theta\|\boldsymbol{\mathcal{G}}\|_1+\frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2$ is differentiable and its derivative is continuous on $\mathbb{R}^{J_1\times J_2\ldots\times J_N}\times\mathbb{R}^{I_n\times J_n}$ (since it is a polynomial in the entries of $\boldsymbol{\mathcal{G}}$ and $\mathbf{A}^{(n)}$). Given that the minimization problem with respect to $\boldsymbol{\mathcal{G}}$ is a minimization problem on a compact set of a continuous function, it admits a solution that is unique (by *Assumption III*). Hence, by the theorem of Bonnan and Shapiro, the function lost function $\mathbf{A}^{(n)} \to l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) = \min_{\boldsymbol{\mathcal{G}}} \frac{1}{2}\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{G}} \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}\|_1 + \frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2$ is directionally differentiable and we have by **Property** 4

$$\partial_2 l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) = -\left(\widehat{\mathbf{X}}^{(n)}\mathbf{G}_0^{(n)T}(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) - \mathbf{A}^{(n)}\mathbf{B}^{(n)}\mathbf{B}^{(n)T}\right) + \alpha(1-\theta)\mathbf{A}^{(n)}$$

with $\widehat{\mathbf{X}}^{(n)}$ and $\mathbf{B}^{(n)}$ being the matricized forms of the tensors:
$\widehat{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}} \times_{p\in I_{n-1}} \mathbf{A}^{(p)T} \times_n \mathbf{I} \times_{q\in I_N^{n+1}} \mathbf{A}^{(q)T}, \boldsymbol{\mathcal{B}} = \mathbf{G}_0 \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{I} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}$
$\boldsymbol{\mathcal{G}}_0(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) = \arg\min_{\boldsymbol{\mathcal{G}}} \|\boldsymbol{\mathcal{X}}-\boldsymbol{\mathcal{G}}\times_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)}\times_n\mathbf{A}^{(n)}\times_{q\in I_N^{n+1}}\mathbf{A}_k^{(q)}\|_F^2+\alpha\theta\|\boldsymbol{\mathcal{G}}\|_1+\frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2$.
Since the derivative is bounded, $l(\boldsymbol{\mathcal{X}}, .)$ is Lipschitz by the Mean Value theorem.

Property 8. *$\widehat{f}_{n,t}$ is Lipschitz and bounded with constant independent of t.*

Proof:
$\widehat{f}_{n,t}\left(\mathbf{A}^{(n)}\right) = \frac{1}{t}\sum_{i=1}^t \frac{1}{2}\|\mathbf{X}_i^{(n)} - \mathbf{A}^{(n)}\mathbf{B}^{(n)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}_i\|_1 + \frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2$ with:
$\mathbf{B}^{(n)} = \mathbf{G}_i^{(n)}\left(\otimes_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)T} \otimes_{q\in I_N^{n+1}} \mathbf{A}_{k+1}^{(q)T}\right)$, $\mathbf{X}_i^{(n)}$ and $\mathbf{G}_i^{(n)}$ being the mode-n matricized forms of $\boldsymbol{\mathcal{X}}_i$ and $\boldsymbol{\mathcal{G}}_i$.
$\widehat{f}_{n,t}\left(\mathbf{A}_1^{(n)}\right) - \widehat{f}_{n,t}\left(\mathbf{A}_2^{(n)}\right) = \frac{1}{t}\sum_{i=1}^t \frac{1}{2}\left(\|\mathbf{X}_i^{(n)} - \mathbf{A}_1^{(n)}\mathbf{B}^{(n)}\|_F^2 - \|\mathbf{X}_i^{(n)} - \mathbf{A}_2^{(n)}\mathbf{B}^{(n)}\|_F^2\right)$
$+\frac{\alpha(1-\theta)}{2}\left(\|\mathbf{A}_1^{(n)}\|_F^2 - \|\mathbf{A}_2^{(n)}\|_F^2\right)$
$\Rightarrow| \widehat{f}_{n,t}\left(\mathbf{A}_1^{(n)}\right) - \widehat{f}_{n,t}\left(\mathbf{A}_2^{(n)}\right) | \leq \frac{1}{t}\sum_{i=1}^t M_1\| \left(\mathbf{A}_1^{(n)} - \mathbf{A}_2^{(n)}\right)\mathbf{B}^{(n)}\|_F + M_2\| \left(\mathbf{A}_1^{(n)} - \mathbf{A}_2^{(n)}\right)\|_F$
(**Property** 1)
$\Rightarrow| \widehat{f}_{n,t}\left(\mathbf{A}_1^{(n)}\right) - \widehat{f}_{n,t}\left(\mathbf{A}_2^{(n)}\right) | \leq \frac{1}{t}\sum_{i=1}^t M_3\|\mathbf{A}_1^{(n)} - \mathbf{A}_2^{(n)}\|_F + M_2\|\mathbf{A}_1^{(n)} - \mathbf{A}_2^{(n)}\|_F$ (the latent factors are bounded)
$\Rightarrow| \widehat{f}_{n,t}\left(\mathbf{A}_1^{(n)}\right) - \widehat{f}_{n,t}\left(\mathbf{A}_2^{(n)}\right) | \leq (M_3 + M_2)\,\|\mathbf{A}_1^{(n)} - \mathbf{A}_2^{(n)}\|_F$

Property 9. *$f_{n,t}$ is Lipschitz and bounded with constant independent of t.*

Proof: since the lost function l is Lipschitz with respect to $\mathbf{A}^{(n)}$, $f_{n,t}(\mathbf{A}^{(n)}) = \frac{1}{t}\sum_{i=1}^t l(\boldsymbol{\mathcal{X}}_i, \mathbf{A}^{(n)})$ is Lipschitz as linear combination with positive coefficients of Lipschitz functions and boundedness result from the boundedness of the loss function.

Property 10. *Let's consider the function* $\xi(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}, \boldsymbol{\mathcal{G}})$*:*
$\xi(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}, \boldsymbol{\mathcal{G}}) = \frac{1}{2}\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{G}} \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}\|_1 + \frac{\alpha(1-\theta)}{2}\|\mathbf{A}^{(n)}\|_F^2$.
Let's denote $\Delta(\boldsymbol{\mathcal{G}})$ *the function defined by:* $\Delta(\boldsymbol{\mathcal{G}}) = \xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}) - \xi(\boldsymbol{\mathcal{X}}_2, \mathbf{A}_2^{(n)}, \boldsymbol{\mathcal{G}})$*. The function* Δ *is Lipschitz.*

Proof:
$\Delta(\boldsymbol{\mathcal{G}}_\alpha) - \Delta(\boldsymbol{\mathcal{G}}_\beta) = \xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\alpha) - \xi(\boldsymbol{\mathcal{X}}_2, \mathbf{A}_2^{(n)}, \boldsymbol{\mathcal{G}}_\alpha) - \Big(\xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\beta) - \xi(\boldsymbol{\mathcal{X}}_2, \mathbf{A}_2^{(n)}, \boldsymbol{\mathcal{G}}_\beta)\Big)$
$\Rightarrow \Delta(\boldsymbol{\mathcal{G}}_\alpha) - \Delta(\boldsymbol{\mathcal{G}}_\beta) = \xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\alpha) - \xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\beta) - \Big(\xi(\boldsymbol{\mathcal{X}}_2, \mathbf{A}_2^{(n)}, \boldsymbol{\mathcal{G}}_\alpha) - \xi(\boldsymbol{\mathcal{X}}_2, \mathbf{A}_2^{(n)}, \boldsymbol{\mathcal{G}}_\beta)\Big)$
$\Rightarrow \mid \Delta(\boldsymbol{\mathcal{G}}_\alpha) - \Delta(\boldsymbol{\mathcal{G}}_\beta) \mid \leq \underbrace{\mid \xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\alpha) - \xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\beta) \mid}_{FT}$
$+ \underbrace{\mid \xi(\boldsymbol{\mathcal{X}}_2, \mathbf{A}_2^{(n)}, \boldsymbol{\mathcal{G}}_\alpha) - \xi(\boldsymbol{\mathcal{X}}_2, \mathbf{A}_2^{(n)}, \boldsymbol{\mathcal{G}}_\beta) \mid}_{ST}$
$\xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\alpha) - \xi(\boldsymbol{\mathcal{X}}_1, \mathbf{A}_1^{(n)}, \boldsymbol{\mathcal{G}}_\beta) = \frac{1}{2}\|\mathbf{X}_1^{(n)} - \mathbf{A}_1^{(n)}\mathbf{G}_\alpha^{(n)}\mathbf{M}^{(n)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}_\alpha\|_1 + \frac{\alpha(1-\theta)}{2}\|\mathbf{A}_1^{(n)}\|_F^2 - \frac{1}{2}\|\mathbf{X}_1^{(n)} - \mathbf{A}_1^{(n)}\mathbf{G}_\beta^{(n)}\mathbf{M}^{(n)}\|_F^2 - \alpha\theta\|\boldsymbol{\mathcal{G}}_\beta\|_1 - \frac{\alpha(1-\theta)}{2}\|\mathbf{A}_1^{(1)}\|_F$
$FT \leq M_1\|\mathbf{G}_\alpha^{(n)} - \mathbf{G}_\beta^{(n)}\|_F + \alpha\theta\|\|\boldsymbol{\mathcal{G}}_\alpha\| - \|\boldsymbol{\mathcal{G}}_\beta\|_1\|$ (boundedness+**Property** 1)
$\Rightarrow FT \leq M_1\|\boldsymbol{\mathcal{G}}_\alpha - \boldsymbol{\mathcal{G}}_\beta\|_F + M_2\|\boldsymbol{\mathcal{G}}_\alpha - \mathbf{G}_\beta\|_F$ (second triangle inequality+norms equivalence).
An identical reasoning on ST concludes the proof.

Property 11. *Let* $\mathbf{A}$ *be a matrix whose entries represent random variables,* $\exists\beta > 0, \|\mathbb{E}(\mathbf{A})\|_F \leq \beta\mathbb{E}(\|\mathbf{A}\|_F)$ *with* $(\mathbb{E}(\mathbf{A}))_{i,j} = \mathbb{E}(\mathbf{A}_{i,j})$

Proof:
Since the set of matrices of a given size represent finite dimensional space and all norms in a finite dimensional space are equivalent,$\beta_1, \beta_2 > 0$ deterministic such that:
$\beta_1\|\mathbf{A}\|_F \leq \|\mathbf{A}\|_1 \leq \beta_2\|\mathbf{A}\|_F$. and $\beta_1\|\mathbb{E}(\mathbf{A})\|_F \leq \|\mathbb{E}(\mathbf{A})\|_1 \leq \beta_2\|\mathbb{E}(\mathbf{A})\|_F$
Thus, we have:
$\|\mathbb{E}(\mathbf{A})\|_F \leq \frac{1}{\beta_1}\|\mathbb{E}(\mathbf{A})\|_1 \leq \frac{1}{\beta_1}\mathbb{E}(\|\mathbf{A}\|_1) \leq \frac{\beta_2}{\beta_1}\mathbb{E}(\|\mathbf{A}\|_F)$
The first and third inequalities are from norms equivalence in a finite dimensional vector space and the second one is from triangular inequality.

Appendix F. Block-wise convergence

The reasoning for this section is inspired from the analysis provided by [21]. The objective of this section is to prove that $\mathbf{A}_{k+1,t}^{(n)}$ converges to the set of stationary points of $\mathbf{A}^{(n)} \rightarrow f(\mathbf{A}_{k+1}^{(1)}, .., \mathbf{A}_{k+1}^{(n-1)}, \mathbf{A}^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(n)})$. For notations simplicity, we denote $\mathbf{A}_{k+1,t}^{(n)}$ by $\mathbf{A}_t^{(n)}$.

Appendix F.1. $\exists M > 0, \|\mathbf{A}_t^{(n)} - \mathbf{A}_{t+1}^{(n)}\|_F \leq \frac{M}{t}$

Proof:
$\widehat{f}_{n,t}(\mathbf{A}^{(n)})$ is strictly convex with Hessian lower bounded (**Property** 6)

$$\widehat{f}_{n,t}(\mathbf{A}_{t+1}^{(n)}) \geq \widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) + M\|\mathbf{A}_{t+1}^{(n)} - \mathbf{A}_t^{(n)}\|_F^2 \tag{F.1}$$

Moreover:
$\widehat{f}_{n,t}(\mathbf{A}_{t+1}^{(n)}) = \widehat{f}_{n,t}(\mathbf{A}_{t+1}^{(n)}) - \widehat{f}_{n,t+1}(\mathbf{A}_{t}^{(n)}) + \widehat{f}_{n,t+1}(\mathbf{A}_{t}^{(n)})$
$\leq \widehat{f}_{n,t}(\mathbf{A}_{t+1}^{(n)}) - \widehat{f}_{n,t+1}(\mathbf{A}_{t+1}^{(n)}) + \widehat{f}_{n,t+1}(\mathbf{A}_{t}^{(n)})$: :definition of$\mathbf{A}_{t+1}^{(n)}$
$\Rightarrow \widehat{f}_{n,t}(\mathbf{A}_{t+1}^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_{t}^{(n)}) \leq \widehat{f}_{n,t}(\mathbf{A}_{t+1}^{(n)}) - \widehat{f}_{n,t+1}(\mathbf{A}_{t+1}^{(n)}) + \widehat{f}_{n,t+1}(\mathbf{A}_{t+1}^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_{t}^{(n)})$
This implies:

$$\widehat{f}_{n,t}(\mathbf{A}_{t+1}^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_{t}^{(n)}) \leq Q_t(\mathbf{A}_{t+1}^{(n)}) - Q_t(\mathbf{A}_{t}^{(n)}), Q_t(\mathbf{A}^{(n)}) = \widehat{f}_{n,t}(\mathbf{A}^{(n)}) - \widehat{f}_{n,t+1}(\mathbf{A}^{(n)}) \quad \text{(F.2)}$$

It is worth to notice (by definition of $\mathbf{A}_{t}^{(n)}$) that:

$$\mathbf{Q}_t(\mathbf{A}_{t+1}^{(n)}) - \mathbf{Q}_t(\mathbf{A}_{t}^{(n)}) \geq 0 \quad \text{(F.3)}$$

$\widehat{f}_{n,t+1}(\mathbf{A}^{(n)}) = \frac{t}{t+1}\widehat{f}_{n,t}(\mathbf{A}^{(n)}) + \frac{\alpha(1-\theta)}{2(t+1)}\|\mathbf{A}^{(n)}\|_F^2 + \frac{1}{2(t+1)}\|\boldsymbol{\mathcal{X}}_{t+1} - \boldsymbol{\mathcal{G}}_{t+1} \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}^{(q)}\|_F^2$
Hence, we have:
$Q_t(\mathbf{A}^{(n)}) = -\frac{1}{t}\widehat{f}_{n,t}(\mathbf{A}^{(n)}) + \frac{1}{t+1}\left(\frac{1}{2}\|\boldsymbol{\mathcal{X}}_{t+1} - \boldsymbol{\mathcal{G}}_{t+1} \times_{p\in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A}^{(n)} \times_{q\in I_N^{n+1}} \mathbf{A}_{k}^{(q)}\|_F^2 + \alpha\theta\|\boldsymbol{\mathcal{G}}_{t+1}\|_1\right) + \frac{1}{2(t+1)}\|\mathbf{A}^{(n)}\|_F^2$
$\Rightarrow Q_t(\mathbf{A}_1^{(n)}) - Q_t(\mathbf{A}_2^{(n)}) = \frac{1}{t}\left(\widehat{f}_{n,t}(\mathbf{A}_2^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_1^{(n)})\right) + \frac{1}{2(t+1)}\left(\|\mathbf{X}_{t+1}^{(n)} - \mathbf{A}_1^{(n)}\mathbf{B}^n\|_F^2 - \|\mathbf{X}_{t+1}^{(n)} - \mathbf{A}_2^{(n)}\mathbf{B}^n\|_F^2\right) + \frac{\alpha(1-\theta)}{2(t+1)}(\|\mathbf{A}_1^{(n)}\|_F^2 - \|\mathbf{A}_2^{(n)}\|_F^2), \mathbf{B}^{(n)} = \mathbf{G}_{t+1}^{(n)}\left(\otimes_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)T} \otimes_{q\in I_N^{n+1}} \mathbf{A}_k^{(q)T}\right)$
$\Rightarrow |\ Q_t(\mathbf{A}_1^{(n)}) - Q_t(\mathbf{A}_2^{(n)})\ | \leq \frac{1}{t}\|\widehat{f}_{n,t}(\mathbf{A}_1^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_2^{(n)})\| + \frac{M}{2(t+1)}\|\mathbf{A}_1^{(n)} - \mathbf{A}_2^{(n)}\|_F$ (boundedness+**Property** 1)
$\Rightarrow Q_t(\mathbf{A}_{t+1}^{(n)}) - Q_t(\mathbf{A}_t^{(n)}) \leq \frac{M_2}{t}\|\mathbf{A}_{t+1}^{(n)} - \mathbf{A}_t^{(n)}\|_F + \frac{M}{2(t+1)}\|\mathbf{A}_{t+1}^{(n)} - \mathbf{A}_t^{(n)}\|_F \leq \frac{M_3}{t}\|\mathbf{A}_{t+1}^{(n)} - \mathbf{A}_t^{(n)}\|_F$ (**Property** 8+equation (F.3))
Hence, we have with (F.1) and (F.2):
$M\|\mathbf{A}_{t+1}^{(n)} - \mathbf{A}_t^{(n)}\|_F^2 \leq \frac{M_3}{t}\|\mathbf{A}_{t+1}^{(n)} - \mathbf{A}_t^{(n)}\|_F$. By assuming $\mathbf{A}_t^{(n)} \neq \mathbf{A}_{t+1}^{(n)}$, we have:
$\|\mathbf{A}_{t+1}^{(n)} - \mathbf{A}_t^{(n)}\|_F = O(\frac{1}{t})$

Appendix F.2. The gradient of the function f_n is Lipschitz

Proof:
$f_n(\mathbf{A}^{(n)}) = \mathbb{E}_{\mathbb{P}}(l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}))$
The function $\mathbf{A} \to l(\boldsymbol{\mathcal{X}}, \mathbf{A})$ is differentiable and its derivative is bounded (**property** 7) by a constant K, which is integrable. By the theorem of differentiation under integral, we have:
$\partial f_n(\mathbf{A}) = \mathbb{E}_{\mathbb{P}}\left(\partial_2 l(\boldsymbol{\mathcal{X}}, \mathbf{A})\right)$

$$\Rightarrow \partial f_n(\mathbf{A}) = \mathbb{E}_{\mathbb{P}}\left(\underbrace{-\left(\widehat{\mathbf{X}}^{(n)}\mathbf{G}_0^{(n)T}(\boldsymbol{\mathcal{X}}, \mathbf{A}) - \mathbf{A}\mathbf{B}^{(n)}\mathbf{B}^{(n)T}\right) + (1-\theta)\alpha\mathbf{A}}_{\mathbf{V}(\mathbf{A})}\right)$$

with: $\mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}) = \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A})\left(\otimes_{p\in I_{n-1}}\mathbf{A}_{k+1}^{(p)T} \otimes_{q\in I_N^{n+1}} \mathbf{A}_k^{(p)T}\right)$

$\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2) = -2\widehat{\mathbf{X}}^{(n)} \left(\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2) \right) + \alpha(1-\theta)(\mathbf{A}_1 - \mathbf{A}_2)$
$+2 \left(\mathbf{A}_1 \mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) \mathbf{B}^{(n)T}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{A}_2 \mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2) \mathbf{B}^{(n)T}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2) \right)$
$\Rightarrow \|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F \leq \alpha_1 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + 2\|\mathbf{A}_1 \mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) \mathbf{B}^{(n)T}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{A}_2 \mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2) \mathbf{B}^{(n)T}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + \alpha(1-\theta)\|\mathbf{A}_1 - \mathbf{A}_2\|_F$
Since all the matrices are bounded, we have:
$\Rightarrow \|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F \leq \alpha_1 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + 2\alpha_2 \|\mathbf{A}_1 - \mathbf{A}_2\|_F$
$+2\alpha_3 \|\mathbf{B}_n(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) \mathbf{B}_n^T(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{B}_n(\boldsymbol{\mathcal{X}}, \mathbf{A}_2) \mathbf{B}_n^T(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + \alpha(1-\theta)\|\mathbf{A}_1 - \mathbf{A}_2\|_F$ (**Property** 2) $\Rightarrow \|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F \leq \alpha_1 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1)\|_F + (2\alpha_2 + \alpha(1-\theta))\|\mathbf{A}_1 - \mathbf{A}_2\|_F + 2\alpha_3(\alpha_4 \|\mathbf{B}_n(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{B}_n(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + \alpha_5 \|\mathbf{B}_n^T(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{B}_n^T(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F)$(**Property** 2)
$\Rightarrow \|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F \leq \alpha_1 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + (2\alpha_2 + \alpha(1-\theta))\|\mathbf{A}_1 - \mathbf{A}_2\|_F + \alpha_6 \|\mathbf{B}_n(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{B}_n(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F$ (because the term $\|\mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F$ appears two times)
$\Rightarrow \|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F \leq \alpha_1 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + (2\alpha_2 + \alpha(1-\theta))\|\mathbf{A}_1 - \mathbf{A}_2\|_F + \alpha_6 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) \left(\otimes_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)T} \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(p)T} \right) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2) \left(\otimes_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)T} \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(p)T} \right) \|_F$
(by replacing the expression of $\mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A})$)
$\Rightarrow \|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F \leq \alpha_1 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + (2\alpha_2 + \alpha(1-\theta))\|\mathbf{A}_1 - \mathbf{A}_2\|_F$
$+ \alpha_6 \|(\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)) \left(\otimes_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)T} \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(p)T} \right) \|_F$ (distributivity of the Kronecker product)
By the boundedness of the matrices, we have:

$$\|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F \leq \alpha_8 \|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F + (2\alpha_2 + \alpha(1-\theta))\|\mathbf{A}_1 - \mathbf{A}_2\|_F \quad \text{(F.4)}$$

By definition of ∂f_n, we have:
$\partial f_n(\mathbf{A}_1) - \partial f_n(\mathbf{A}_2) = \mathbb{E}_{\mathbb{P}}(\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2))$
$\Rightarrow \|\partial f_n(\mathbf{A}_1) - \partial f_n(\mathbf{A}_2)\|_F = \|\mathbb{E}_{\mathbb{P}}(\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2))\|_F$
By **Property** 11,$\exists \beta > 0$ such that:
$\|\partial f_n(\mathbf{A}_1) - \partial f_n(\mathbf{A}_2)\|_F \leq \beta \mathbb{E}_{\mathbb{P}}(\|\mathbf{V}(\mathbf{A}_1) - \mathbf{V}(\mathbf{A}_2)\|_F$
By the inequality (F.4), we have:
$\|\partial f_n(\mathbf{A}_1) - \partial f_n(\mathbf{A}_2)\|_F \leq M_1 \mathbb{E}_{\mathbb{P}}(\|\mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F) + M_2 \|\mathbf{A}_1 - \mathbf{A}_2\|_F$
It is sufficient to prove $\mathbf{G}_0$ is Lipschitz to prove the result
By definition,

$$\begin{aligned}
\mathbf{G}_0(\boldsymbol{\mathcal{X}}, \mathbf{A}) &\leftarrow \arg\min \frac{1}{2} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{G}} \times_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \times_n \mathbf{A} \times_{q \in I_N^{n+1}} \mathbf{A}_k^{(q)}\|_F^2 + \alpha\theta \|\boldsymbol{\mathcal{G}}\|_1 + \frac{\alpha(1-\theta)}{2} \|\mathbf{A}\|_F^2 \\
&= \arg\min_{\boldsymbol{\mathcal{G}}} \frac{1}{2} \operatorname{vec}(\boldsymbol{\mathcal{X}}) - \left(\otimes_{p \in I_{n+1}} \mathbf{A}_{k+1}^{(p)} \otimes \mathbf{A} \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(q)} \right) \operatorname{vec}(\boldsymbol{\mathcal{G}})\|_F + \alpha\theta \|\operatorname{vec}(\boldsymbol{\mathcal{G}})\|_1
\end{aligned}$$

Given the *Assumption III* (uniqueness of the *Sparse coding* problem), **Property** 10 along with the fact that the *Sparse coding* problem amounts to a classical least squares problem with L_1- penalty by vectorization of the *Tucker* decomposition, the same reasoning as the one conducted in [21]: proposition 2 yields:
$\|\mathbf{G}_0(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F$

$\leq M_1 \| \left(\otimes_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \otimes \mathbf{A}_1 \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(q)} \right) - \left(\otimes_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \otimes \mathbf{A}_2 \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(q)} \right) \|_F$
$= M_1 \| \left(\otimes_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)} \otimes (\mathbf{A}_1 - \mathbf{A}_2) \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(q)} \right) \|_F$ (distributivity of the Kronecker product)
$\Rightarrow \|\mathbf{G}_0(\boldsymbol{\mathcal{X}}, \mathbf{A}_1) - \mathbf{G}_0(\boldsymbol{\mathcal{X}}, \mathbf{A}_2)\|_F \leq M_3 \|\mathbf{A}_1 - \mathbf{A}_2\|_F$, which concludes the proof.

Appendix F.3. $f_n(\mathbf{A}_t^{(n)})$ converges

First step: $\widehat{f}_{n,t}(\mathbf{A}_t^{(n)})$ converges almost surely.

Proof

$u_t = \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})$
$u_{t+1} - u_t = \widehat{f}_{n,t+1}(\mathbf{A}_{t+1}^{(n)}) - \widehat{f}_{n,t+1}(\mathbf{A}_t^{(n)}) + \widehat{f}_{n,t+1}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})$
$\Rightarrow u_{t+1} - u_t \leq \widehat{f}_{n,t+1}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})$ (by definition of $\mathbf{A}_{t+1}^{(n)}$)
$\widehat{f}_{n,t+1}(\mathbf{A}_t^{(n)}) = \frac{t}{t+1}\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) + \frac{1}{t+1} l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)})$ (by definition of $l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)})$ and $\boldsymbol{\mathcal{G}}_{t+1}$)
Hence, we have:

$$u_{t+1} - u_t \leq \frac{l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})}{t+1} = \frac{l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)}) + f_{n,t}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})}{t+1} \tag{F.5}$$

By noticing $f_{n,t} \leq \widehat{f}_{n,t}$, we have:

$$u_{t+1} - u_t \leq \frac{l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1} \tag{F.6}$$

By conditioning the expression (F.5) with respect to the filtration $\boldsymbol{\mathcal{F}}_t$ (filtration determined by the past information at time t), we have:
$\mathbb{E}(u_{t+1} - u_t | \boldsymbol{\mathcal{F}}_t) \leq \frac{\mathbb{E}(l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)}) | \boldsymbol{\mathcal{F}}_t) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1} = \frac{\mathbb{E}(l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)})) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1}$ (since $l(\boldsymbol{\mathcal{X}}_{t+1}, \mathbf{A}_t^{(n)})$ is independent from the filtration $\boldsymbol{\mathcal{F}}_t$)
$\Rightarrow \mathbb{E}_{\mathbb{P}}(u_{t+1} - u_t | \boldsymbol{\mathcal{F}}_t) \leq \frac{f_n(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1}$(definition of $f_n(\mathbf{A}_t^{(n)})$)
$\Rightarrow \mathbb{E}_{\mathbb{P}}(u_{t+1} - u_t | \boldsymbol{\mathcal{F}}_t) \leq \frac{\|f_n - f_{n,t}\|_\infty}{t+1}$,(the functions are bounded)
$\Rightarrow \mathbb{E}_{\mathbb{P}}(u_{t+1} - u_t | \boldsymbol{\mathcal{F}}_t)^+ \leq \frac{\|f_n - f_{n,t}\|_\infty}{t+1}$(because $\frac{\|f_n - f_{n,t}\|_\infty}{t+1} \geq 0$)
Hence, we have:

$$\mathbb{E}_{\mathbb{P}}(\mathbb{E}_{\mathbb{P}}(u_{t+1} - u_t | \boldsymbol{\mathcal{F}}_t)^+) \leq \frac{\mathbb{E}_{\mathbb{P}}\left(\|f_n - f_{n,t}\|_\infty\right)}{t+1} \tag{F.7}$$

The lost function $l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)})$ is uniformly bounded on a bounded set and $\mathbb{E}_{\mathbb{P}}(l^2(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}))$ exists and is bounded (because the lost function is uniformly bounded). Hence, according to the *Theorem IV*, $\exists M > 0$:
$\mathbb{E}_{\mathbb{P}}\left(\sqrt{t} \mid \frac{1}{t}\sum_{i=1}^t l(\boldsymbol{\mathcal{X}}_i, \mathbf{A}^{(n)}) - \mathbb{E}_{\mathbb{P}}(l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) \mid)\right) \leq M$,$\boldsymbol{\mathcal{X}}$ drawn from $\mathbb{P}, \forall \mathbf{A}^{(n)}$
$\Rightarrow \mathbb{E}_{\mathbb{P}}\left(\sqrt{t} \mid f_{n,t}(\mathbf{A}^{(n)}) - f_n(\mathbf{A}^{(n)}) \mid\right) \leq M$,$\forall \mathbf{A}^{(n)}$
$\Rightarrow \mathbb{E}_{\mathbb{P}}\left(\|f_{n,t} - f_n\|_\infty\right) \leq \frac{M}{\sqrt{t}}$
This inequality associated to (F.7) yields:
$\mathbb{E}_{\mathbb{P}}(\mathbb{E}_{\mathbb{P}}(u_{t+1} - u_t | \boldsymbol{\mathcal{F}}_t)^+) \leq \frac{M}{\sqrt{t}(t+1)} \leq \frac{M}{t^{\frac{3}{2}}}$.

$\Rightarrow \sum_{t=1}^{\infty} \mathbb{E}_{\mathbb{P}}(\delta_t(u_{t+1}-u_t)) = \sum_{t=1}^{\infty} \mathbb{E}_{\mathbb{P}}(\mathbb{E}_{\mathbb{P}}(u_{t+1}-u_t|\boldsymbol{\mathcal{F}}_t)^+) < \infty$($\delta_t$ being defined in *Theorem III*). By the *Theorem III*, u_t converges almost surely to u_∞.
Second step: $f_{n,t}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) \to 0$ a.s.
By the equation (F.5) we have:
$u_{t+1} - u_t \leq \frac{l(\boldsymbol{\mathcal{X}}_{t+1},\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1} + \frac{f_{n,t}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})}{t+1}$
$\Rightarrow \mathbb{E}(u_{t+1} - u_t) \leq \mathbb{E}(\frac{f_{n,t}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})}{t+1})$(the expectation of the first term is zero since the data samples are assumed to be identically distributed)
$\Rightarrow \mathbb{E}_{\mathbb{P}}(\frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1}) \leq \mathbb{E}_{\mathbb{P}}(u_t - u_{t+1})(*)$
$\Rightarrow \sum_{t=1}^{T-1} \mathbb{E}_{\mathbb{P}}(\frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1}) \leq \sum_{t=1}^{T-1} \mathbb{E}_{\mathbb{P}}(u_t - u_{t+1}) = \mathbb{E}_{\mathbb{P}}(u_1 - u_T)(**)$
The sequence $\{u_1 - u_T\}_T$ converges (**First step**). Since it is bounded, by the dominated convergence theorem, the sequence $\{\mathbb{E}_{\mathbb{P}}(u_1 - u_T)\}_T$ converges. Therefore, we deduce from the equality (**) the convergence of the series $\mathbb{E}_{\mathbb{P}}(u_t - u_{t+1})$.
Since the series $\mathbb{E}_{\mathbb{P}}(\frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1})$ and $\mathbb{E}_{\mathbb{P}}(u_t - u_{t+1})$ are positive (positivity of the first and the second terms result respectively from the definition and (*)) and the series $\mathbb{E}_{\mathbb{P}}(u_t - u_{t+1})$ converges , by the inequality (*), the series $\mathbb{E}_{\mathbb{P}}(\frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1})$ converges. Hence, we have:
$\sum_{t=1}^{\infty} \mathbb{E}_{\mathbb{P}}(\frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1}) < \infty$
$\Rightarrow \mathbb{E}_{\mathbb{P}}(\sum_{t=1}^{\infty} \frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1}) < \infty$ (by Fubini Theorem for positive random variables)
Given that $\sum_{t=1}^{\infty} \frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1}$ is a positive random variable and its expectation is finite, this implies (see Processus Aléatoires:page 9 by Delmas and al.):

$$\sum_{t=1}^{\infty} \frac{\widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) - f_{n,t}(\mathbf{A}_t^{(n)})}{t+1} < \infty a.s. \tag{F.8}$$

Let's consider two sequences a_t, b_t defined by: $a_t = \frac{1}{t+1}, b_t = f_{n,t}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)})$
It does not require great effort to see that the assumptions of *Theorem II* are verified. Thus, we have:
$f_{n,t}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) \to 0$ a.s. when $t \to \infty$
Third step: $f_n(\mathbf{A}_t^{(n)})$ converges almost surely.
By triangular inequality and the definition of $|.|_\infty$, we have:

$$| f_n(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) | \leq \|f_n - f_{n,t}\|_\infty + | f_{n,t}(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) | \tag{F.9}$$

A class of function $\mathcal{F} = \left\{ f : \Omega \times \mathbb{K}_{I_n,J_n} \to \mathbb{R}, f(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) = l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) \right\}, \Omega$ representing the set of tensors of compact support. Since $\mathbb{K}_{I_n,J_n}$ is compact and the class of functions $\mathcal{F}$ verifies *Theorem V* , we have:
$\lim_{t\to\infty} | \frac{1}{t} \sum_{i=1}^{t} l(\boldsymbol{\mathcal{X}}_i, \mathbf{A}^{(n)}) - \mathbb{E}_{\mathbb{P}}(l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)})) | = 0, \forall \mathbf{A}^{(n)}$
$\Rightarrow \lim_{t\to\infty} | \frac{1}{t} \sum_{i=1}^{t} l(\boldsymbol{\mathcal{X}}_i, \mathbf{A}^{(n)}) - \mathbb{E}_{\mathbb{P}}(l(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)})) | = 0, \forall \mathbf{A}^{(n)}$
$\Rightarrow \lim_{t\to\infty} | f_{n,t}(\mathbf{A}^{(n)}) - f_n(\mathbf{A}^{(n)}) | = 0, \forall \mathbf{A}^{(n)}$
$\Rightarrow \lim_{t\to\infty} \|f_{n,t} - f_n\|_\infty = 0$
By Glivenko-Cantelli and the **Second step**, the inequality (F.9) yields:

$\mid f_n(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) \mid \to 0$ when $t \to \infty$
$\Rightarrow f_n(\mathbf{A}_t^{(n)}) - \widehat{f}_{n,t}(\mathbf{A}_t^{(n)}) \to 0$ a.s. $t \to \infty$
$\Rightarrow f_n(\mathbf{A}_t^{(n)})$ converges a.s when $t \to \infty$ (since $\widehat{f}_{n,t}(\mathbf{A}_t^{(n)})$ converges a.s. by the **First step**)

Appendix F.4. $\mathbf{A}_t^{(n)}$ *converges to a stationary point of* f_n*: asymptotic behavior with respect to* t

Proof:
With the assumptions I, II, III, a reasoning identical to the one presented in [21] (*Proposition 4*) yields this convergence.

Appendix G. Global convergence: asymptotic behavior with respect to k

In this section, we extend the result of the last section to the set of stationary point of f considered as a function of all of the dictionary matrices.
From the section Appendix F.4, $\forall n, \mathbf{A}_t^{(n)}$ converges to $\mathbf{A}_\infty^{(n)} = \mathbf{A}_{k+1}^{(n)}$, a stationary point of f_n on $\mathbb{K}_{I_n,J_n}$. This implies the derivative of $-\frac{\partial f_n}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_{k+1}^{(n)})$ belongs to the normal cone of $\mathbb{K}_{I_n,J_n}$ at $\mathbf{A}_{k+1}^{(n)}$. Since $\mathbf{A}_{k+1}^{(n)}$ is an interior point of $\mathbb{K}_{I_n,J_n}$ (*Assumption V*), it is well known that this normal cone is reduced to the singleton $\{0\}$. Hence:
$\frac{\partial f}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_{k+1}^{(1)}, .., \mathbf{A}_{k+1}^{(n-1)}, \mathbf{A}_{k+1}^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)}) = 0$ since $\mathbf{A}^{(n)} \to f(.., \mathbf{A}_{k+1}^{(n-1)}, \mathbf{A}^{(n)}, \mathbf{A}_k^{(n+1)}, ..)$ and $\mathbf{A}^{(n)} \to f_n(\mathbf{A}_{k+1}^{(1)}, .., \mathbf{A}_{k+1}^{(n-1)}, \mathbf{A}^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)})$ are proportional up to an additive constant.
We proved in Appendix F.2 that:
$\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}^{(n)}}(\mathbf{A}^{(1)}, .., \mathbf{A}^{(n-1)}, \mathbf{A}^{(n)}, \mathbf{A}^{(n+1)}, .., \mathbf{A}^{(N)}) = \mathbb{E}_{\mathbb{P}}\left(\mathbf{V}(\mathbf{A}^{(1)}, .., \mathbf{A}^{(n-1)}, \mathbf{A}^{(n)}, \mathbf{A}^{(n+1)}, .., \mathbf{A}^{(N)})\right)$
$\mathbf{V}(\mathbf{A}^{(1)}, .., \mathbf{A}^{(n-1)}, \mathbf{A}^{(n)}, \mathbf{A}^{(n+1)}, .., \mathbf{A}^{(N)}) = -\left(\widehat{\mathbf{X}}^{(n)}\mathbf{G}_0^{(n)T}(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) - \mathbf{A}^{(n)}\mathbf{B}^{(n)}\mathbf{B}^{(n)T}\right) + (1 - \theta)\alpha\mathbf{A}^{(n)}$, $\mathbf{B}^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) = \mathbf{G}_0^{(n)}(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)}) \otimes_{p \in I_{n-1}} \mathbf{A}_{k+1}^{(p)T} \otimes_{q \in I_N^{n+1}} \mathbf{A}_k^{(p)T}$
Since the function $\mathbf{A}^{(n)} \to \mathbf{G}_0(\boldsymbol{\mathcal{X}}, \mathbf{A}^{(n)})$ is uniformly continuous because it is Lipschitz (proved in section Appendix E), the function $\mathbf{V}$ is uniformly continuous. Thus, we have:
$\forall \epsilon > 0, \exists \eta > 0, \forall x, y, \|x - y\|_{\prod_{n \in I_N} \mathbb{R}^{I_n \times J_n}} \leq \eta \Rightarrow \|\mathbf{V}(x) - \mathbf{V}(y)\|_F \leq \epsilon$
Let's choose $y = (\mathbf{A}_k^{(1)}, .., \mathbf{A}_k^{(n-1)}, \mathbf{A}_k^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)})$ and $x = (\mathbf{A}_{k+1}^{(1)}, .., \mathbf{A}_{k+1}^{(n-1)}, \mathbf{A}_{k+1}^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)})$
By *Assumption IV*, we have:
$\|x - y\|_{\prod_{n \in I_N} \mathbb{R}^{I_n \times J_n}} \leq \frac{n}{k^{\frac{1}{2}}}$
By convergence of $\frac{n}{k^{\frac{1}{2}}}$ to 0 when k goes to infinity (n being fixed):
$\exists k_0 \in \mathbb{N}, \forall k \geq k_0, \frac{n}{k^{\frac{1}{2}}} \leq \eta$
Hence, for all $k \geq k_0$, we have:
$\|x - y\|_{\prod_{n \in I_N} \mathbb{R}^{I_n \times J_n}} \leq \eta$
$\Rightarrow \|\mathbf{V}(x) - \mathbf{V}(y)\|_F \leq \epsilon$
$\Rightarrow \mathbb{E}_{\mathbb{P}}\left(\|\mathbf{V}(x) - \mathbf{V}(y)\|_F\right) \leq \epsilon$
$\Rightarrow \|\mathbb{E}_{\mathbb{P}}(\mathbf{V}(x)) - \mathbb{E}(\mathbf{V}(y))\|_F \leq \beta\epsilon$ (**Property** 11)
By definition, $\mathbb{E}_{\mathbb{P}}(\mathbf{V}(x)) = 0$ and $\frac{\partial f}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_k^{(1)}, .., \mathbf{A}_k^{(n-1)}, \mathbf{A}_k^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)}) = \mathbb{E}_{\mathbb{P}}(\mathbf{V}(y))$.
So far, we have proven that:

$\forall \epsilon > 0, \exists k_0 \in \mathbb{N}, \forall k \geq k_0, \| \frac{\partial f}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_k^{(1)}, .., \mathbf{A}_k^{(n-1)}, \mathbf{A}_k^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)}) \|_F \leq \beta\epsilon, \forall n$
$\Rightarrow \lim_{k\to\infty} \| \frac{\partial f}{\partial \mathbf{A}^{(n)}}(\mathbf{A}_k^{(1)}, .., \mathbf{A}_k^{(n-1)}, \mathbf{A}_k^{(n)}, \mathbf{A}_k^{(n+1)}, .., \mathbf{A}_k^{(N)}) \|_F = 0.$
This means that the more k increases, the more the $N - uplet \left\{ \mathbf{A}_k^{(1)}, .., \mathbf{A}_k^{(N)} \right\}$ gets closer to the set of stationary point of the function $\left\{ \mathbf{A}^{(1)}, .., \mathbf{A}^{(N)} \right\} \in \mathbb{R}^{I_1 \times J_1} .. \times \mathbb{R}^{I_N \times J_N} \rightarrow f(\mathbf{A}^{(1)}, .., \mathbf{A}^{(N)})$, which concludes the proof.

Miscellaneous results used for the analysis.

Theorem I: *from Bonnan and Shapiro*
Let f : $\mathbb{R}^p \times \mathbb{R}^q \leftarrow \mathbb{R}$. Suppose that for all $\mathbf{x} \in \mathbb{R}^p$ the function $\mathbf{u} \leftarrow f(x, u)$ is differentiable, and that f and $\partial_2 f(\mathbf{x}, u)$ the derivative of $\mathbf{u} \rightarrow f(\mathbf{x}, \mathbf{u})$ are continuous on $\mathbb{R}^p \times \mathbb{R}^q$. Let $\nu(u)$ be the optimal value function $\nu(\mathbf{u}) = \min_{\mathbf{x}\in C} f(\mathbf{x}, \mathbf{u})$, where C is a compact subset of $\mathbb{R}^p$. Then $\nu(\mathbf{u})$ is directionally differentiable.
Furthermore, if for $\mathbf{u}_0 \in \mathbb{R}^q$, $f(., \mathbf{u}_0)$ has a unique minimizer $\mathbf{x}_0$ then $\nu(u)$ is differentiable in $\mathbf{u}_0$ and $\partial\nu(\mathbf{u}_0) = \partial_2 f(\mathbf{x}_0, \mathbf{u}_0)$
Theorem II (see *Lemma 8* in [21])
Let a_n, b_n be two real sequences such that for all n, $a_n \geq 0$,$b_n \geq 0$ $\sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n b_n < \infty, \exists K > 0$ s.t $|b_{n+1} - b_n| \leq K a_n$. Then, $\lim_{n\to\infty} b_n = 0$.
Theorem III (see *Theorem 6* in [21])
Let $(\mathcal{W}, \mathcal{F}, \mathcal{P})$ be a measurable probability space, u_t , for $t \geq 0$, be the realization of a stochastic process and $\mathcal{F}_t$ be the filtration determined by the past information at time t. Let δ_t defined by: $\delta_t = 1$ if $\mathbb{E}(u_{t+1} - u_t|\mathcal{F}_t) > 0$ otherwise. If for all t, $u_t \geq 0$ and $\sum_{t=1}^{\infty} \mathbb{E}(\delta_t(u_{t+1} - u_t)) < \infty$. Then u_t is a quasi-martingale and converges almost surely. Moreover, $\sum_{t=1}^{\infty} \|\mathbb{E}(u_{t+1} - u_t|\mathcal{F}_t)\| < \infty$ a.s
Theorem IV(A corollary of Donsker theorem see Van der Vaart, 1998, chap. 19.2, lemma 19.36 and example 19.7)
Let $F = \{f_\theta : \beta \rightarrow \mathbb{R}, \theta \in \Theta\}$ be a set of measurable functions indexed by a bounded subset $\Theta \subset \mathbb{R}^d$. Suppose that there exists a constant K such that:
$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq K|\theta_1 - \theta_2|$ for every $\theta_1, \theta_2 \in \Theta, x \in \beta$.Then, F is P-Donsker (see Van der Vaart, 1998, chap. 19.2).
For any f in F, let us define $\mathbb{P}_n f, \mathbb{P}_f$ and $\mathbb{G}_n(f)$ as: $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i)$,
$\mathbb{P}f = \mathbb{E}_{\mathbf{X}}(f(\mathbf{X})), \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)$
Let us also suppose that for all f, $\mathbf{P}f^2 \leq \delta^2$ and $\|f\|_\infty \leq M$ and that the variables $\mathbf{X}_1, ..$ are Borel-measurable. Then, we have: $\mathbb{E}_P(\|\mathbb{G}_n\|_F) = O(1)$,$\|\mathbb{G}_n f\|_F = sup_{f\in F}|\mathbb{G}_n f|$
Theorem V: from Wald (1949) and Le Cam (1953)
Let $\mathcal{F} = \{f : \Omega \times \mathbb{T} \rightarrow \mathbb{R}\}$ where the functions $f : \Omega \times \mathbb{T} \rightarrow \mathbb{R}$, are continuous in t for $\mathbb{P}$ almost all x $\in \Omega$ ($\mathbb{P}$: probability on Ω). Suppose that $\mathbb{T}$ is compact and that the envelope function F defined by: $F(x) = sup_{t\in\mathbb{T}} \mid f(x, t) \mid$ satisfying $\mathbb{P}(F) = \mathbb{E}_{\mathbb{P}}(F(\boldsymbol{\mathcal{X}})) < \infty, \boldsymbol{\mathcal{X}} \sim \mathbb{P}$. Then, $N_{[]}(\epsilon, \mathcal{F}, L_1(\mathbb{P})) < \infty, \forall \epsilon > 0$ (hence, $\mathcal{F}$ is $\mathbb{P}-$Glivenko-Cantelli.)