

# SEPARABLE DICTIONARY LEARNING WITH GLOBAL OPTIMALITY AND APPLICATIONS TO DIFFUSION MRI

EVAN SCHWAB, BENJAMIN HAEFFELE, NICOLAS CHARON AND RENÉ VIDAL

**Abstract.** Dictionary learning is a popular class of methods for modeling complex data by learning sparse representations directly from the data. For some large-scale applications, exploiting a known structure of the signal is often essential for reducing the complexity of algorithms and representations. One such method is tensor factorization by which a large multi-dimensional dataset can be explicitly factored or separated along each dimension of the data in order to break the representation up into smaller components. Learning dictionaries for tensor structured data is called *tensor* or *separable* dictionary learning. While there have been many recent works on separable dictionary learning, typical formulations involve solving a non-convex optimization problem and guaranteeing global optimality remains a challenge. In this work, we propose a framework that uses recent developments in matrix/tensor factorization to provide theoretical and numerical guarantees of the global optimality for the separable dictionary learning problem. We will demonstrate our algorithm on diffusion magnetic resonance imaging (dMRI) data, a medical imaging modality which measures water diffusion along multiple angular directions in every voxel of an MRI volume. For this application, state-of-the-art methods learn dictionaries for the angular domain of the signals without consideration for the spatial domain. In this work, we apply the proposed separable dictionary learning method to learn spatial and angular dMRI dictionaries jointly and show results on denoising phantom and real dMRI brain data.

**Key word.** separable dictionary learning, tensor factorization, global convergence, diffusion MRI, HARDI.

**1. Introduction.** In signal processing, a well studied problem is that of reconstructing a signal from a set of noisy measurements by finding a representation of the signal in a chosen domain for which one can more easily process and analyze the data. In the most general setting, one would like to represent (or approximate) a signal  $y \in \mathbb{R}^N$  in terms of a dictionary  $D \in \mathbb{R}^{N \times r}$  with  $N_D$  dictionary atoms as

$$(1.1) \quad y = Dw,$$

where the coefficient vector  $w \in \mathbb{R}^r$  is the representation of  $y$  in terms of dictionary  $D$ . This is in general an ill-posed problem and one typically imposes additional constraints on the reconstructed signal. One usual assumption is that  $y$  is sparse in the dictionary  $D$ , i.e.  $w$  has very few non-zero entries, which then leads to the classic sparse coding problem:

$$(1.2) \quad \min_w \frac{1}{2} \|Dw - y\|_2^2 + \lambda \|w\|_1,$$

where  $\lambda > 0$  is a trade-off parameter between sparsity and data fidelity.

There are many signal processing applications of sparse coding including denoising [16], super-resolution [47] and Compressed Sensing (CS) [8]. For example, the goal of CS is to minimize the number of samples needed to accurately reconstruct a signal in order to accelerate signal acquisition. The typical number of measurements needed is directly linked to the sparsity of the representation which is obviously dependent on the choice of dictionary  $D$ . For different types of signals there may be an array of known dictionaries that produce sparse representations (e.g. Wavelets for natural or medical images [27]). However, prescribing a known dictionary for a new signal or data type may lead to suboptimal sparsity levels.

**1.1. Dictionary Learning.** To overcome this limitation, the idea of learning the dictionary from the signal itself (or a set of training examples), known as sparse dictionary learning, has been explored extensively. Although many different methodologies exist, typical formulations assume that a training set is given of  $T$  signals  $\{y_t\}_{t=1}^T$  that resemble the signals of interest. Then one seeks to approximate each signal  $y_t$  as a sparse linear combination of the atoms  $D_i$  from a dictionary  $D$ , and thus consider an optimization problem of the form:

$$(1.3) \quad \min_{D, \{w_t\}} \frac{1}{2} \sum_{t=1}^T \|Dw_t - y_t\|_2^2 + \lambda \|w_t\|_1 \quad \text{s.t. } \|D_i\|_2 \leq 1 \quad \text{for } i = 1, \dots, r,$$

where the constraints  $\|D_i\|_2 \leq 1$  are enforced to prevent an unbounded solution for  $D$ . Letting  $Y = [y_1, \dots, y_T]$  and  $W = [w_1, \dots, w_T]$  the problem can be written more compactly in matrix from as:

$$(1.4) \quad \min_{D, W} \frac{1}{2} \|DW - Y\|_F^2 + \lambda \|W\|_1 \quad \text{s.t. } \|D_i\|_2 \leq 1 \quad \text{for } i = 1, \dots, r,$$

with  $\|X\|_F = \sqrt{\sum_{i,j} |X_{i,j}|^2}$  is the Frobenius norm.

There have been many proposed algorithms to solve this dictionary learning problem (or its variants) [17, 26, 28]. One of the most popular is KSVD [1] which alternates between solving for the  $w_t$ 's while  $D$  is fixed (sparse coding update), and updating  $D_i$  one by one (dictionary learning update) via SVD decomposition. Once a dictionary is learned, it can be used in (1.2) for sparsely representing a new (test) signal. One important downside of the dictionary learning problem (1.3) is that the joint optimization over  $D$  and  $W$  results in a non-convex problem, and therefore guaranteeing a globally optimal solution is a difficult challenge. At best, optimization algorithms such as gradient descent may reach stationary points which can be either local minima or saddle points, providing sometimes suboptimal dictionary solutions.

**1.2. Separable Dictionary Learning.** While (1.3) is the classical setting for vector valued signals  $y_t \in \mathbb{R}^N$ , the problem of dictionary learning becomes increasingly more complex for signals with additional structure that we would like to preserve. For instance, for image data, instead of vectorizing an image and learning a dictionary, it may be useful to preserve the 2D structure of the image for computer vision tasks. In this case, one can attempt to learn separate 1D dictionaries for each dimension of the image to reduce complexity. Another increasingly important setting is one in which we have structured data that is very large-scale and high-dimensional (e.g. 4D time-series data) in which learning a “global” dictionary for the entire dataset may be computationally infeasible. In this case, learning smaller-scale dictionaries for each individual dimension can be a more manageable problem, reducing computational complexity and memory requirements. Formally, this is known as *separable* dictionary learning or *tensor* dictionary learning due to the tensor structure of the data, (e.g. 2D images are 2-tensors and 4D time-series are 4-tensors). In this paper we will consider as an application uniquely structured, large-scale and high-dimensional diffusion magnetic resonance imaging (dMRI) data, used for analyzing neurological diseases such as Alzheimer’s Disease.

To mathematically introduce the notation of separable dictionaries, consider a 2D signal  $S \in \mathbb{R}^{G \times V}$ . To represent  $S$  as a bilinear combination of separable dictionaries, we can write  $S = \Gamma C \Psi^\top$ , where  $\Gamma \in \mathbb{R}^{G \times r_1}$ ,  $\Psi \in \mathbb{R}^{V \times r_2}$  are dictionaries for the first and second dimensions, respectively, and  $C \in \mathbb{R}^{r_1 \times r_2}$  are the set of joint coefficients between dictionaries. Now, recall that for the case of vectored valued signals, dictionary learning utilized  $T$  training examples  $y_t$ , each with coefficients  $w_t$ , such that  $y_t = D w_t$ . For 2D signals, we will have  $T$  matrices  $S_t$ , each with coefficients  $C_t$ , such that  $S_t = \Gamma C_t \Psi^\top$ . With these expressions, the separable dictionary learning problem can be stated as follows:

$$(1.5) \quad \min_{\Gamma, \Psi, \{C_t\}} \frac{1}{2} \sum_{t=1}^T \|\Gamma C_t \Psi^\top - S_t\|_F^2 + \lambda \|C_t\|_1 \quad \text{s.t. } \|\Gamma_i\|_2 \leq 1, \|\Psi_j\|_2 \leq 1 \quad \forall (i, j).$$

Learning separable dictionaries via (1.5) and its multidimensional tensor generalization for higher-dimensional signals has been studied previously in the literature. The work of [25, 50, 51] solve variations of (1.5) using conjugate gradient methods over smooth manifolds. In terms of tensors, [37, 52], resort to solving alternatively each mode of the tensor as the usual vector dictionary learning problem after  $n$ -mode unfolding, which loses the computational gain of maintaining a tensor structure. The work of [15, 38, 43] use decompositions such as Tucker, Kruskal-Factor and tensor SVDs, while [12] considers a dictionary as the sum of Kronecker products. Finally, [4, 18] propose low-rank variations of the separable dictionary learning problem.

As we recall for (1.3), one key difficulty in dictionary learning is the lack of guarantees of global optimality due to the non-convexity of the joint optimization over the dictionary and coefficients. This issue is especially difficult for separable dictionary learning because the number of variables to jointly optimize over increases from two to three or more. To the best of our knowledge, none of the aforementioned work on separable dictionary learning come equipped with guarantees for global optimality, and so their solutions may correspond to a local minimum or saddle point and may also heavily depend on initialization. The main contribution of this work is a new framework for separable dictionary learning with guarantees of global optimality. To do this, we expand upon theoretical work on matrix factorization [3, 24] which has been applied previously to provide theoretical guarantees to the original dictionary learning problem (1.3).

Specifically, in Section 2, we recall how the dictionary learning problem (1.3) can be framed as a matrix factorization problem and make a quick summary of global optimality results obtained in [24]. Then in Section 3, we consider a fairly general class of tensor factorization problems for which we obtain similar theoretical results of global optimality. In Section 4, based on those results, we specify the analysis to the

case of separable dictionary learning in order to derive verifiable conditions for the global optimality of solutions. Additionally, we show that our proposed formulation is equivalent to a low-rank tensor factorization problem. Then, in Section 5, we derive a novel algorithm to find global optima of the separable dictionary learning problem. Finally, Section 6 provides a few preliminary results of the approach for learning separable spatial-angular dictionaries from diffusion magnetic resonance imaging data as well as some comparisons with other methods in basic denoising experiments.

## 2. Background.

**2.1. Dictionary Learning as Matrix Factorization.** The general problem of matrix factorization is concerned with finding factors  $D$  and  $W$ , such that a data matrix  $Y$  can be approximated by a matrix  $X = DW$ . Naturally, the dictionary learning problem (1.3) can be thought of in this way. In [3, 22, 23, 24] the authors develop a general matrix factorization framework for a number of applications including the dictionary learning problem. The key insight is an equivalence relation between the non-convex factorized problem with respect to the factors  $D$  and  $W$  and a convex problem with respect to  $X$ , which allows to obtain guarantees of global optimality for  $(D, W)$ .

First, the non-convex matrix factorization problem can be written as:

$$(2.1) \quad \min_{D,W} \ell(Y, DW) + \lambda \Theta(D, W),$$

where  $\ell$  is a data fidelity term or loss that measures the error between the original signal  $Y$  and the reconstruction  $X = DW$ , and  $\Theta$  is a regularizer on the factors  $D$  and  $W$  which promotes particular properties relevant to the problem. For the dictionary learning problem (1.3),  $\ell(Y, DW) = \frac{1}{2} \|DW - Y\|_F^2$ . Furthermore it can be shown that the constraints  $\|D_i\|_2 \leq 1$  can be combined with the sparsity term  $\|W\|_1$  to get  $\Theta(D, W) = \sum_{i=1}^r \|D_i\|_2 \|W_i^\top\|_1$ , where  $W_i^\top \in \mathbb{R}^T$  is the  $i^{th}$  row of  $W$ . Then (2.1) is an equivalent problem formulation of (1.3). The goal of recasting the dictionary learning problem as a matrix factorization problem is to relate (2.1), which is non-convex with respect to  $D$  and  $W$ , to a convex problem with respect to  $X$ .

**2.2. Global Optimality for Matrix Factorization.** To derive conditions for the global optimality, [24] first impose the regularizer to be of the specific form  $\Theta(D, W) = \sum_i^r \theta(D_i, W_i^\top)$  where  $\theta$  is a rank-1 regularizer that must satisfy the following properties:

DEFINITION 2.1 (from [24]). *A function  $\theta : \mathbb{R}^N \times \mathbb{R}^T \rightarrow \mathbb{R}_+ \cup \infty$  is said to be a **rank-1 regularizer** if*

1.  $\theta(u, v)$  is positively homogeneous with degree 2, i.e.  $\theta(\alpha u, \alpha v) = \alpha^2 \theta(u, v) \forall \alpha \geq 0, \forall (u, v)$ .
2.  $\theta(u, v)$  is positive semi-definite, i.e.  $\theta(0, 0) = 0$  and  $\theta(u, v) \geq 0 \forall (u, v)$ .
3. For any sequence  $(u_n, v_n)$  such that  $\|u_n v_n^\top\| \rightarrow \infty$ , we have that  $\theta(u_n, v_n) \rightarrow \infty$ .

It is easy to show that the choice of  $\theta(u, v) = \|u\|_2 \|v\|_1$  fits this definition. Another example of  $\theta$  satisfying Definition 2.1 that can be used for dictionary learning is  $\theta(u, v) = \|u\|_2 (\|v\|_2 + \alpha \|v\|_1)$  which promotes column regularization in  $u$  and  $v$  and also sparsity in  $v$  as analyzed in [3].

Now, in order to connect the non-convex (2.1) with a convex problem with respect to matrix  $X$ , we introduce a related regularizer  $\Omega_\theta(X)$  which depends on  $\theta$ :

DEFINITION 2.2 (from [24]). *Given a rank-1 regularizer  $\theta$  that satisfies the conditions of Definition 2.1, the **matrix factorization regularizer**  $\Omega_\theta : \mathbb{R}^{N \times T} \rightarrow \mathbb{R}_+ \cup \infty$  is defined as:*

$$(2.2) \quad \Omega_\theta(X) \equiv \inf_{r \in \mathbb{N}_+} \inf_{D,W} \sum_{i=1}^r \theta(D_i, W_i^\top) \text{ s.t. } DW = X.$$

If the infimum is achieved for some  $D, W$  and  $r$  then we say that  $DW$  is an optimal factorization of  $X$ .

It is important to note that the number of dictionary atoms  $r$  becomes an important variable of finding an optimal matrix factorization in this definition. As a motivating example for the origin of  $\Omega_\theta$ , when  $\theta(D_i, W_i^\top) = \|D_i\|_2 \|W_i^\top\|_2$ ,  $\Omega_\theta(X)$  becomes the variational definition of the nuclear norm:

$$(2.3) \quad \|X\|_* \equiv \inf_{r \in \mathbb{N}_+} \inf_{D,W} \sum_{i=1}^r \|D_i\|_2 \|W_i^\top\|_2 \text{ s.t. } DW = X.$$

From the results of [24],  $\Omega_\theta(X)$  is a gauge function (and even a norm if  $\theta$  is symmetric, i.e.  $\theta(-u, v) = \theta(u, v)$  or  $\theta(u, -v) = \theta(u, v)$  for all  $u, v$ ), which leads to the new convex optimization problem with respect to  $X$ :

$$(2.4) \quad \min_X \ell(Y, X) + \lambda \Omega_\theta(X).$$

Since (2.4) is convex, a local minimum  $\hat{X}$  is guaranteed to be global. The question answered in [24] is then how to relate a local minimum  $(\tilde{D}, \tilde{W})$  of the non-convex (2.1) to a global minimum of the convex (2.4) and when, if ever, can we say something about global minimum  $(\hat{D}, \hat{W})$  of (2.1). First, it is evident that (2.4) provides a global lower bound of (2.1) because  $\Omega_\theta$  is the infimum of  $\Theta$  and  $\ell(Y, X) = \ell(Y, DW)$ . The main result is then that under certain conditions local minima  $(\tilde{D}, \tilde{W})$  of the non-convex (2.1) are optimal factorizations of  $X$ , such that  $\hat{X} = \tilde{D}\tilde{W}$ . In other words, given a local solution  $(\tilde{D}, \tilde{W})$  to (2.1), we can write a matrix  $X = \tilde{D}\tilde{W}$  and under certain conditions, it turns out that the matrix  $X$  is a global minimum of (2.4), i.e.  $X \equiv \hat{X}$ . Therefore,  $(\tilde{D}, \tilde{W})$  is in fact a global minimum of (2.1),  $(\hat{D}, \hat{W})$ . We restate this main theorem of [24] here:

**THEOREM 2.3.** [from [24]] *Given a function  $\ell(S, X)$  that is convex and once differentiable w.r.t.  $X$ , a rank-1 regularizer  $\theta$  that satisfies the conditions in Definition 2.1, with constants  $r \in \mathbb{N}_+$ , and  $\lambda > 0$ , local minima  $(\tilde{D}, \tilde{W})$  of (2.1) are globally optimal if  $(\tilde{D}_i, \tilde{W}_i^\top) = (0, 0)$  for some  $i \in [r]$ . Moreover,  $\hat{X} = \tilde{D}\tilde{W}$  is a global minima of (2.4) and  $\tilde{D}\tilde{W}$  is an optimal factorization of  $\hat{X}$ .*

Since  $\theta$  is general, this matrix factorization can be applied to many problems such as low-rank, non-negative matrix factorization, sparse PCA as well as the desired dictionary learning. However, one important downside for the application of dictionary learning is that the choices of  $\theta$  stated above are not well suited to checking the criteria of Theorem 2.3 in practice. In particular verifying if a point is stationary or a local minimum remains difficult. Therefore, finding globally optimal solutions for classical dictionary learning still remains a challenging problem. In the next section we will extend the results of [24] for the more complex structured *separable* dictionary learning.

**3. Global optimality for tensor factorization.** In this section, we will first attempt to extend such guarantees of global optimality in matrix factorization to a fairly general class of *tensor* factorization problems, before connecting this result to separable dictionary learning in Section 4.

**3.1. Tensor factorization problem and notations.** Similar to matrix factorization, tensor factorization is concerned with finding factors that decompose a  $n$ -tensor  $\underline{S} \in \mathbb{R}^{N_1 \times N_2 \times \dots \times N_n}$ . There are two main types of tensor decomposition: rank-1 decomposition, with each factor  $f_i \in \mathbb{R}^{N_i}$  is a vector such that  $\underline{X} = f_1 \otimes f_2 \otimes \dots \otimes f_n$ ,  $\otimes$  denoting the tensor outer product, and the Tucker decomposition, in which there is a core  $n$ -tensor  $\underline{C} \in \mathbb{R}^{r_1 \times r_2 \times \dots \times r_n}$  and matrix factors  $F_i \in \mathbb{R}^{N_i \times r_i}$  such that  $\underline{X} = \underline{C} \times_1 F_1 \times_2 F_2 \dots \times_n F_n$ , where  $\times_n$  stands for matrix multiplication on the  $n^{\text{th}}$  dimension of the core tensor  $\underline{C}$ . (See [13] for a review of tensor decomposition.)

In this paper, as we are ultimately interested in dictionary learning and in particular for spatial-angular signals of dMRI, we simplify the rest of the presentation by restricting our discussion to decomposition of 3-tensors (although the results of this section can readily extend to general  $n$ ). Using notations consistent with (1.5), we will consider the tensor signal  $\underline{S} \in \mathbb{R}^{G \times V \times T}$  where each slice  $S_t$  corresponds to a training example. Our goal in what follows will be to decompose this tensor (at least approximately) as  $\underline{S} = \underline{C} \times_1 \Gamma \times_2 \Psi$  with a core tensor  $\underline{C} \in \mathbb{R}^{r_1 \times r_2 \times T}$  and two fixed factors  $\Gamma \in \mathbb{R}^{G \times r_1}$ ,  $\Psi \in \mathbb{R}^{V \times r_2}$ . Note that this is equivalent to writing  $S_t = \Gamma C_t \Psi^\top$  for each  $t = 1, \dots, T$  and that it can be interpreted as a Tucker decomposition of  $\underline{S}$  where we here enforce the last factor  $F_3$  to be the identity.

In all the following, tensors will be written with an underline, e.g. the 3-tensor  $\underline{C} \in \mathbb{R}^{r_1 \times r_2 \times T}$ . To index the tensor  $\underline{C}$ , all 2D slices (matrices) will be written with an upper case letter and a single index, e.g.  $C_t \in \mathbb{R}^{r_1 \times r_2}$  or  $C_i \in \mathbb{R}^{r_2 \times T}$  or  $C_j \in \mathbb{R}^{r_1 \times T}$ . Furthermore, 1D slice vectors of  $\underline{C}$  will be written with an upper case letter and two indices, e.g.  $C_{i,j} \in \mathbb{R}^T$  or  $C_{i,t} \in \mathbb{R}^{r_2}$  or  $C_{j,t} \in \mathbb{R}^{r_1}$ . Finally, single elements of  $\underline{C}$  will be simply written as  $c_{i,j,t}$ .

Similar to the matrix factorization problem (2.1) of the previous section, we will consider general tensor factorization problems formulated as:

$$(3.1) \quad \min_{\Gamma, \Psi, \underline{C}} \{f(\Gamma, \Psi, \underline{C}) \equiv \ell(S, \underline{C} \times_1 \Gamma \times_2 \Psi) + \lambda \Theta(\Gamma, \Psi, \underline{C})\}.$$

where the first term  $\ell$  is a measure of similarity to the data and  $\Theta$  is a certain regularizer that enforces some constraints on the factorization. We will make the assumption that  $\ell$  is separable in the different  $t$  slices of the tensor, namely, with a slight abuse of notation, we will write  $\ell(\underline{S}, \underline{Y}) = \sum_{t=1}^T \ell(S_t, Y_t)$ . For instance, the separable dictionary learning problem of (1.5) that we shall consider more specifically in the next section corresponds to the choice  $\ell(\underline{S}, \underline{C} \times_1 \Gamma \times_2 \Psi) = \frac{1}{2} \|\underline{C} \times_1 \Gamma \times_2 \Psi - \underline{S}\|_F^2 = \frac{1}{2} \sum_{t=1}^T \|\Gamma C_t \Psi^\top - S_t\|_F^2$  while the constraints  $\|\Gamma_i\|_2 \leq 1$ ,  $\|\Psi_i\|_2 \leq 1$  and sparse  $\underline{C}$  can be shown to be achieved by introducing a regularizer of the form:

$$(3.2) \quad \Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \|\Gamma_i\|_2 \|\Psi_j\|_2 \|C_{i,j}\|_1.$$

Now, as in the previous case of matrix factorization, we wish to link stationary points of the non-convex  $f(\Gamma, \Psi, \underline{C})$  with a global minimum of a convex function with respect to  $\underline{X}$ .

**3.2. A global optimality criterion.** To develop the theories of global optimality for separable dictionary learning we begin by extending Definitions 2.1 and 2.2 from Section 2.2. First, we will consider a regularizer in (3.1) of the form  $\Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j})$  where  $\theta$  satisfies the following conditions:

DEFINITION 3.1. A function  $\theta : \mathbb{R}^G \times \mathbb{R}^V \times \mathbb{R}^T \rightarrow \mathbb{R}_+ \cup \infty$  is said to be a **rank-1 regularizer** if

1.  $\theta$  is positively homogeneous of degree 3, i.e.  $\theta(\alpha\gamma, \alpha\psi, \alpha c) = \alpha^3 \theta(\gamma, \psi, c) \forall \alpha \geq 0, \forall (\gamma, \psi, c)$ .
2.  $\theta$  is positive semi-definite and  $\theta(\gamma, \psi, c) > 0$  if and only if  $\gamma \otimes \psi \otimes c \neq 0$ .
3. For any sequence  $(\gamma_n, \psi_n, c_n)$  such that  $\|\gamma_n \otimes \psi_n \otimes c_n\| \rightarrow \infty$ , we have  $\theta(\gamma_n, \psi_n, c_n) \rightarrow \infty$ .

Then, similarly to Definition 2.2, we define the related regularizer for tensor  $\underline{X}$ :

DEFINITION 3.2. Given a rank-1 regularizer  $\theta$  that satisfies the conditions of Definition 3.1, the **tensor factorization regularizer**  $\Omega_\theta : \mathbb{R}^{G \times V \times T} \rightarrow \mathbb{R}_+ \cup \infty$  is defined as:

$$(3.3) \quad \Omega_\theta(\underline{X}) := \inf_{r_1, r_2 \in \mathbb{N}_+} \inf_{\substack{\Gamma \in \mathbb{R}^{G \times r_1} \\ \Psi \in \mathbb{R}^{V \times r_2} \\ \underline{C} \in \mathbb{R}^{r_1 \times r_2 \times T}}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = \underline{X}.$$

If the infimum is achieved for some  $(\Gamma, \Psi, \underline{C})$  and  $r_1, r_2 \in \mathbb{N}_+$  then we say that  $\underline{C} \times_1 \Gamma \times_2 \Psi$  is an optimal factorization of  $\underline{X}$ .

PROPOSITION 3.3. Given regularizer  $\theta$  that satisfies the properties of Definition 3.1, the tensor factorization regularizer  $\Omega_\theta(\underline{X})$  satisfies the following properties:

1.  $\Omega_\theta(0) = 0$  and  $\Omega_\theta(\underline{X}) > 0 \forall \underline{X} \neq 0$ .
2.  $\Omega_\theta(\alpha \underline{X}) = \alpha \Omega_\theta(\underline{X}) \forall \alpha \geq 0 \forall \underline{X}$ .
3.  $\Omega_\theta(\underline{X} + \underline{Y}) \leq \Omega_\theta(\underline{X}) + \Omega_\theta(\underline{Y}) \forall (\underline{X}, \underline{Y})$ .
4. If  $\theta$  is symmetric in  $\gamma, \psi$  or  $c$ , then  $\Omega_\theta(-\underline{X}) = \Omega_\theta(\underline{X}) \forall \underline{X}$  and  $\Omega_\theta$  is a norm.
5. The infimum of  $\Omega_\theta(\underline{X})$  in (3.3) can be achieved with finite  $r_1$  and  $r_2$ , and  $r_1, r_2 \leq G \times V \times T$ .

The proof of Proposition 3.3 is provided in Appendix A. By definition, satisfying the first three properties show that  $\Omega_\theta$  is gauge function, and properties 2 and 3 show that  $\Omega_\theta$  is convex. Then, with respect to  $\underline{X}$ , we have the convex problem:

$$(3.4) \quad \min_{\underline{X}} \{F(\underline{X}) \equiv \ell(\underline{S}, \underline{X}) + \lambda \Omega_\theta(\underline{X})\},$$

where  $F$  is a global lower bound for  $f$ . The next theorem, an extension of Theorem 2.3, then relates global minimizers of the non-convex  $f$  (3.1) to the convex  $F$  in (3.4).

**THEOREM 3.4.** Given a function  $\ell(\underline{S}, \underline{X})$  that is convex and once differentiable w.r.t.  $\underline{X}$ , a rank-1 regularizer  $\theta$  that satisfies the conditions in Definition 3.1, and constants  $r_1, r_2 \in \mathbb{N}_+$ , and  $\lambda > 0$ , any local minima  $(\tilde{\Gamma}, \tilde{\Psi}, \tilde{\underline{C}})$  of  $f(\Gamma, \Psi, \underline{C})$  in (3.1) is globally optimal if there exists  $(i, j)$  such that  $(\tilde{\Gamma}_i, \tilde{\Psi}_j) = (0, 0)$  and for all  $t$ ,  $(\tilde{C}_{i,t}, \tilde{C}_{j,t}) = (0, 0)$ . Moreover,  $\hat{\underline{X}} = \tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$  is a global minimum of  $F(\underline{X})$  in (3.4) and  $\tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$  is an optimal factorization of  $\hat{\underline{X}}$  in (3.3).

In order to prove Theorem 3.4, we first note that  $\hat{X}$  is a global minimum of  $F(\underline{X})$  if and only if  $-\frac{1}{\lambda} \nabla_{\underline{X}} \ell(\underline{S}, \hat{X}) \in \partial \Omega_\theta(\hat{X})$  since we have a convex function. Therefore, we must first characterize the subgradient  $\partial \Omega_\theta(\underline{X})$  which is the subject of the following lemma.

LEMMA 3.5. *The subgradient  $\partial \Omega_\theta(\underline{X})$  is given by:*

$$(3.5) \quad \left\{ \underline{W} : \langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X}) \text{ and } \sum_{t=1}^T c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c) \forall (\gamma, \psi, c) \right\}.$$

*Proof.* Since  $\Omega_\theta$  is convex, by Fenchel duality,  $\underline{W} \in \partial \Omega_\theta$  if and only if  $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X}) + \Omega_\theta^*(\underline{W})$  where  $\Omega_\theta^*$  is the Fenchel dual of  $\Omega_\theta$  given by  $\Omega_\theta^*(\underline{W}) \equiv \sup_{\underline{Z}} \langle \underline{W}, \underline{Z} \rangle - \Omega_\theta(\underline{Z})$ . From the definition of  $\Omega_\theta(\underline{Z})$  we can expand the dual as

$$\begin{aligned} \Omega_\theta^*(\underline{W}) &= \sup_{r_1, r_2} \sup_{\Gamma, \Psi, \underline{C}} \langle \underline{W}, \underline{Z} \rangle - \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = \underline{Z} \\ &= \sup_{r_1, r_2} \sup_{\Gamma, \Psi, \underline{C}} \langle \underline{W}, \underline{C} \times_1 \Gamma \times_2 \Psi \rangle - \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \\ &= \sup_{r_1, r_2} \sup_{\Gamma, \Psi, \underline{C}} \sum_{t=1}^T \langle \Gamma^T W_t \Psi, C_t \rangle - \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \\ (3.6) \quad &= \sup_{r_1, r_2} \sup_{\Gamma, \Psi, \underline{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \left( \sum_{t=1}^T c_{i,j,t} \Gamma_i^\top W_t \Psi_j \right) - \theta(\Gamma_i, \Psi_j, C_{i,j}) \end{aligned}$$

If there exists  $(\gamma, \psi, c)$  such that  $\sum_{t=1}^T c_t \gamma^\top W_t \psi > \theta(\gamma, \psi, c)$ , we can see that  $\Omega_\theta^*(\underline{W}) = \infty$  by considering  $(\alpha \gamma, \alpha \psi, \alpha c)$  as  $\alpha \rightarrow \infty$  and using the positive homogeneity of  $\theta$ .

Now let  $\underline{W} \in \partial \Omega_\theta$ . Then  $\Omega_\theta^*(\underline{W}) < +\infty$  and thus, from the previous argument, we have that  $\sum_{t=1}^T c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c) \forall (\gamma, \psi, c)$ . This also implies that all the terms in the summation in (3.6) will be non-positive leaving the supremum to be 0, achieved with  $(\Gamma, \Psi, \underline{C}) = (0, 0, \underline{0})$ . It follows that  $\Omega_\theta^*(\underline{W}) = 0$  and consequently  $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X})$ .

Conversely, if  $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X})$  and  $\sum_{t=1}^T c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c) \forall (\gamma, \psi, c)$  then, reasoning as previously, we see that  $\Omega_\theta^*(\underline{W}) = 0$  which implies  $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X}) + \Omega_\theta^*(\underline{W})$  and thus  $\underline{W} \in \partial \Omega_\theta(\underline{X})$ .  $\square$

Next, using the characterization of  $\partial \Omega_\theta(\underline{X})$  in Lemma 3.5, we identify when a factorization  $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$  is optimal, i.e. when a point  $(\Gamma, \Psi, \underline{C})$  achieves the infimum of  $\Omega_\theta(\underline{X})$  in (3.3), in the following corollary.

COROLLARY 3.6. *For factorization  $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$ , if there exists  $\underline{W}$  such that  $\langle \underline{W}, \underline{X} \rangle = \Theta(\Gamma, \Psi, \underline{C})$  and  $\sum_{t=1}^T c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c) \forall (\gamma, \psi, c)$ , then  $\underline{W} \in \partial \Omega_\theta(\underline{X})$  and  $\underline{C} \times_1 \Gamma \times_2 \Psi$  is an optimal factorization of  $\underline{X}$ , i.e. it achieves the infimum of  $\Omega_\theta(\underline{X})$ .*

*Proof.* By contradiction, assume  $\underline{W} \notin \partial \Omega_\theta(\underline{X})$ . Then  $\langle \underline{W}, \underline{X} \rangle < \Omega_\theta(\underline{X}) + \Omega_\theta^*(\underline{W}) = \Omega_\theta(\underline{X})$  because  $\sum_{t=1}^T c_t \gamma^\top W_t \psi \leq \theta(\gamma, \psi, c) \forall (\gamma, \psi, c)$ , implies  $\Omega_\theta^*(\underline{W}) = 0$  as in the proof of Lemma 3.5. Then, from our assumption,  $\langle \underline{W}, \underline{X} \rangle = \Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) < \Omega_\theta(\underline{X})$  which violates the definition of  $\Omega_\theta(\underline{X})$  being the infimum, producing a contradiction. Therefore,  $\underline{W} \in \partial \Omega_\theta(\underline{X})$ . Now, since  $\underline{W} \in \partial \Omega_\theta(\underline{X})$ , by Lemma 3.5,  $\langle \underline{W}, \underline{X} \rangle = \Omega_\theta(\underline{X})$ , which implies  $\Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) = \Omega_\theta(\underline{X})$  thus showing that  $\underline{C} \times_1 \Gamma \times_2 \Psi$  achieves the infimum of  $\Omega_\theta(\underline{X})$  and is an optimal factorization of  $\underline{X}$ .  $\square$

Finally, with Lemma 3.5 and Corollary 3.6 we can now prove Theorem 3.4:

*Proof of Theorem 3.4.* From (3.4), we know  $\hat{X} = \tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$  is a global minimum of  $F(\underline{X})$  if and only if  $-\frac{1}{\lambda} \nabla_{\underline{X}} \ell(\underline{S}, \hat{X}) \in \partial \Omega_\theta(\hat{X})$ . Notice  $-\frac{1}{\lambda} \nabla_{\underline{X}} \ell(\underline{S}, \hat{X})$  can be written in terms of its slices as  $\sum_{t=1}^T -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \hat{X}_t) = \sum_{t=1}^T -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)$ . To prove that  $\hat{X} = \tilde{\underline{C}} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$  is a global minimum and an optimal factorization of  $\hat{X}$ , from Corollary 3.6, it suffices to show two conditions:

1.  $\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{t=1}^T \tilde{c}_{i,j,t} \tilde{\Gamma}_i^\top (-\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)) \tilde{\Psi}_j = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j})$
2.  $\sum_{t=1}^T c_t \gamma^\top (-\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)) \psi \leq \theta(\gamma, \psi, c) \forall (\gamma, \psi, c)$

To show condition 1, let  $\Gamma_{1\pm\epsilon} = (1 \pm \epsilon)^{1/3}\tilde{\Gamma}$  and  $\Psi_{1\pm\epsilon} = (1 \pm \epsilon)^{1/3}\tilde{\Psi}$  and  $C_{1\pm\epsilon} = (1 \pm \epsilon)^{1/3}\tilde{C}$ . Since  $(\tilde{\Gamma}, \tilde{\Psi}, \tilde{C})$  is a local minimum, there exists  $\delta > 0$  such that for all  $\epsilon \in (0, \delta)$  we have

$$(3.7) \quad \sum_{t=1}^T \ell(S_t, \Gamma_{1\pm\epsilon} C_{t\pm\epsilon} \Psi_{1\pm\epsilon}^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta((1 \pm \epsilon)^{1/3}\tilde{\Gamma}_i, (1 \pm \epsilon)^{1/3}\tilde{\Psi}_j, (1 \pm \epsilon)^{1/3}\tilde{C}_{i,j})$$

$$(3.8) \quad = \sum_{t=1}^T \ell(S_t, (1 \pm \epsilon)\tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top) + \lambda(1 \pm \epsilon) \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j})$$

$$(3.9) \quad \geq \sum_{t=1}^T \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}).$$

Rearranging the last inequality gives

$$(3.10) \quad \frac{-1}{\lambda\epsilon} \left[ \sum_{t=1}^T \ell(S_t, (1 \pm \epsilon)\tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top) - \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top) \right] \leq \pm \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}).$$

Taking the limit as  $\epsilon \searrow 0$  gives the directional derivative:

$$(3.11) \quad \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) \leq \sum_{t=1}^T \left\langle \frac{-1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top), \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top \right\rangle \leq \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j})$$

which implies equality. Rearranging the inner product gives Condition 1.

Next, to show condition 2, we use the assumption that there exists  $(i, j)$  such that  $(\tilde{\Gamma}_i, \tilde{\Psi}_j) = (0, 0)$  and for all  $t$ ,  $(\tilde{C}_{i,t}, \tilde{C}_{j,t}) = (0, 0)$ . Without loss of generality let the last column pair of  $(\tilde{\Gamma}, \tilde{\Psi})$  be zero and the last columns and rows of  $\tilde{C}$  be zero for all  $t$ . Then, given  $(\gamma, \psi, c)$ , let  $\Gamma_\epsilon = [\tilde{\Gamma}_1, \dots, \tilde{\Gamma}_{r_1-1}, \epsilon^{1/3}\gamma]$  and  $\Psi_\epsilon = [\tilde{\Psi}_1, \dots, \tilde{\Psi}_{r_2-1}, \epsilon^{1/3}\psi]$  and

$$(3.12) \quad C_{t_\epsilon} = \begin{bmatrix} \tilde{c}_{1,1,t} & \cdots & \tilde{c}_{1,r_2-1,t} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ \tilde{c}_{r_1-1,1,t} & \cdots & \tilde{c}_{r_1-1,r_2-1,t} & 0 \\ 0 & \cdots & 0 & \epsilon^{1/3}c_t \end{bmatrix} \quad \forall t.$$

Now, for all  $\epsilon > 0$  sufficiently small we have

$$(3.13) \quad \begin{aligned} & \sum_{t=1}^T \ell(S_t, \Gamma_\epsilon C_{t_\epsilon} \Psi_\epsilon^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) + \lambda\theta(\epsilon^{1/3}\gamma, \epsilon^{1/3}\psi, \epsilon^{1/3}c) = \\ & \sum_{t=1}^T \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top + \epsilon c_t \gamma \psi^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) + \epsilon\lambda\theta(\gamma, \psi, c) \geq \\ & \sum_{t=1}^T \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}), \end{aligned}$$

where the first equality follows from  $\theta$  being positively homogeneous and the last inequality from the fact that  $\tilde{C} \times_1 \tilde{\Gamma} \times_2 \tilde{\Psi}$  is assumed to be a local minimum of  $F(\Gamma, \Psi, \underline{C})$  and by choosing  $\epsilon$  small enough. Therefore, by rearranging the inequality we arrive at:

$$(3.14) \quad \frac{-1}{\lambda\epsilon} \left[ \sum_{t=1}^T \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top + \epsilon c_t \gamma \psi^\top) - \ell(S_t, \tilde{\Gamma}\tilde{C}_t\tilde{\Psi}^\top) \right] \leq \theta(\gamma, \psi, c)$$

Since  $\ell(S_t, X_t)$  is differentiable with respect to  $X_t$ , taking the limit as  $\epsilon \searrow 0$ , the directional derivative in the

direction of  $c_t \gamma \psi^\top$  gives us

$$(3.15) \quad \sum_{t=1}^T \left\langle \frac{-1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top), c_t \gamma \psi^\top \right\rangle \leq \theta(\gamma, \psi, c)$$

$$(3.16) \quad \Rightarrow \sum_{t=1}^T c_t \gamma^\top \left( \frac{-1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \right) \psi \leq \theta(\gamma, \psi, c),$$

which proves Condition 2. This together with Condition 1 proves Theorem 3.4.  $\square$

The result of Theorem 3.4 holds for any local minimum of  $f$ . Yet, in general, descent methods (e.g. gradient descent) can only be guaranteed to converge to a stationary point at best (which may only be a saddle point) and therefore even arriving at a local minimum of  $f$  may be challenging in practice. In the next section, we examine a choice of regularizer more specific to the dictionary learning problem for which we can eventually derive a more useful condition of global optimality for any  $(\Gamma, \Psi, \underline{C})$ .

**4. Global optimality for dictionary learning.** We will now apply the previous analysis to the case of the rank-1 regularizer given by  $\theta(\gamma, \psi, c) = \|\gamma\|_2 \|\psi\|_2 \|c\|_1$ , for which one easily verifies that the three conditions of Definition 3.1 are satisfied. Then, we have

$$(4.1) \quad \Theta(\Gamma, \Psi, \underline{C}) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \|\Gamma_i\|_2 \|\Psi_j\|_2 \|C_{i,j}\|_1.$$

In that case, the tensor factorization problem of the previous section becomes:

$$(4.2) \quad \min_{r_1, r_2, \Gamma, \Psi, \underline{C}} \ell(\underline{C} \times_1 \Gamma \times_2 \Psi, \underline{S}) + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \|\Gamma_i\|_2 \|\Psi_j\|_2 \|C_{i,j}\|_1.$$

When  $\ell(\underline{C} \times_1 \Gamma \times_2 \Psi, \underline{S}) = \frac{1}{2} \|\underline{C} \times_1 \Gamma \times_2 \Psi - \underline{S}\|_F^2$ , this problem is simply an unconstrained reformulation of the separable dictionary learning problem of (1.5). Yet, in contrast with state-of-the-art dictionary learning approaches, the results from the previous section will allow us to specify an explicit global optimality check for that problem.

**4.1. Necessary and sufficient condition for global optima.** As a consequence of Theorem 3.4, for our particular choice of regularizer, the following characterization holds:

**COROLLARY 4.1.** *Let  $\theta(\gamma, \psi, c) = \|\gamma\|_2 \|\psi\|_2 \|c\|_1$ . A point  $(\tilde{\Gamma}, \tilde{\Psi}, \tilde{C})$  is a global minimum of  $f(\Gamma, \Psi, \underline{C})$  if and only if it satisfies the following conditions:*

$$1. \sum_{t=1}^T \tilde{c}_{i,j,t} \tilde{\Gamma}_i^\top \left( -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \right) \tilde{\Psi}_j = \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \|\tilde{C}_{i,j}\|_1 \quad \forall (i, j)$$

$$2. \max_{1 \leq t \leq T} \sigma_{\max} \left( -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \right) \leq 1$$

where  $\sigma_{\max}$  is the maximum singular value.

*Proof.* First, we know that to be a global minimum, a point must first satisfy first-order optimality for  $f$ . Noting that  $\theta(\tilde{\Gamma}_i, \tilde{\Psi}_j, \tilde{C}_{i,j}) = \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \sum_{t=1}^T |\tilde{c}_{i,j,t}|$  and writing the first-order optimality conditions on the coefficients  $\tilde{c}_{i,j,t}$ , we obtain that:

$$(4.3) \quad 0 = \tilde{\Gamma}_i^\top \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \tilde{\Psi}_j + \lambda \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \operatorname{sign}(\tilde{c}_{i,j,t})$$

for all  $i = 1, \dots, r_1$ ,  $j = 1, \dots, r_2$  and  $t = 1, \dots, T$  if  $\tilde{c}_{i,j,t} \neq 0$ . Multiplying by  $\tilde{c}_{i,j,t}$  then leads, in all cases, to:

$$(4.4) \quad 0 = \tilde{c}_{i,j,t} \tilde{\Gamma}_i^\top \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \tilde{\Psi}_j + \lambda \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 |\tilde{c}_{i,j,t}|.$$

Now, summing over  $t$  gives for all  $i, j$ :

$$(4.5) \quad \sum_{t=1}^T \tilde{c}_{i,j,t} \tilde{\Gamma}_i^\top \left( -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \right) \tilde{\Psi}_j = \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \sum_{t=1}^T |\tilde{c}_{i,j,t}| = \|\tilde{\Gamma}_i\|_2 \|\tilde{\Psi}_j\|_2 \|\tilde{C}_{i,j}\|_1.$$

Therefore, if a point satisfies condition 1 then it is a stationary point.

Next, from Theorem 3.4, we know that for a stationary point to be a global minimum we need to check that the following condition is satisfied:

$$(4.6) \quad \sum_{t=1}^T c_t \gamma^\top \left( -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \right) \psi \leq \theta(\gamma, \psi, c) \quad \forall (\gamma, \psi, c).$$

For simplicity let  $W_t := -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top)$ . With our choice of  $\theta$ , this condition becomes:

$$(4.7) \quad \sum_{t=1}^T c_t \gamma^\top W_t \psi \leq \|\gamma\|_2 \|\psi\|_2 \|c\|_1 \quad \forall (\gamma, \psi, c).$$

Now, normalizing each variable by its respective norm, such that  $\hat{\gamma} = \gamma / \|\gamma\|_2$ ,  $\hat{\psi} = \psi / \|\psi\|_2$ ,  $\hat{c}_t = c_t / \|c\|_1$ , the previous condition becomes

$$(4.8) \quad \sum_{t=1}^T \hat{c}_t \hat{\gamma}^\top W_t \hat{\psi} \leq 1 \quad \forall \hat{\gamma} \neq 0, \hat{\psi} \neq 0, \hat{c} \neq 0,$$

which we can equivalently state as:

$$(4.9) \quad \sup_{\|\hat{\gamma}\|_2 = \|\hat{\psi}\|_2 = \|\hat{c}\|_1 = 1} \sum_{t=1}^T \hat{c}_t \hat{\gamma}^\top W_t \hat{\psi} \leq 1.$$

Now, with respect to  $\hat{c}$ , since  $\|\hat{c}\|_1 = 1$ , the supremum of a linear combination can be attained by choosing  $\hat{c}_{t_*} = 1$  and  $\hat{c}_t = 0$  for  $t \neq t_*$  with  $t_*$  such that  $\sup_{\|\hat{\gamma}\|_2 = \|\hat{\psi}\|_2 = 1} \hat{\gamma}^\top W_{t_*} \hat{\psi} = \max_t \{ \sup_{\|\hat{\gamma}\|_2 = \|\hat{\psi}\|_2 = 1} \hat{\gamma}^\top W_t \hat{\psi} \}$ . Therefore, (4.9) is equivalent to:

$$(4.10) \quad \max_{1 \leq t \leq T} \{ \sup_{\|\hat{\gamma}\|_2 = \|\hat{\psi}\|_2 = 1} \hat{\gamma}^\top W_t \hat{\psi} \} \leq 1,$$

and, with  $\sigma_{\max}$  denoting the largest singular value of the corresponding matrix, this is the same as:

$$(4.11) \quad \max_{1 \leq t \leq T} \sigma_{\max}(W_t) \leq 1,$$

which shows condition 2. Thus, if a point satisfies conditions 1 and 2 then it is a global minimum of  $f$ . Conversely, a global minimum of  $f$  satisfies first-order optimality which is equivalent to condition 1. In addition, the resulting tensor  $\underline{X}$  is then a minimum of  $F$ . It results that  $\sum_{t=1}^T c_t \gamma^\top \left( -\frac{1}{\lambda} \nabla_{X_t} \ell(S_t, \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top) \right) \psi \leq \theta(\gamma, \psi, c)$  for all  $(\gamma, \psi, c)$  which, by the previous reasoning, implies condition 2.  $\square$

Using the results of Corollary 4.1, we can devise an algorithm to find a global minimum of the separable dictionary learning problem by first finding a stationary point of (1.5) and then checking if it satisfies condition 2 in Corollary 4.1. A logical next question of this routine is what happens if the stationary point does not satisfy (2). In [24], the authors demonstrate that by appending a column of zeros to the dictionary  $\tilde{D}$  and a row of zeros to the coefficients  $\tilde{W}$ , they are guaranteed to continue in a descent direction. This is also shown here for the separable dictionary learning case. The part of the proof that gives (4.6) is essentially this argument. If (4.6) is not satisfied then that must be a descent direction. Therefore, the algorithm will consist of iterating between local descent and global optimality check and appending the resulting stationary point. In this way, the optimal size of the dictionary,  $r$ , is learned through the process.

For separable dictionaries, we have two size parameters  $r_1$  and  $r_2$ . Therefore, we have an additional option to augment one or both of the dictionaries to proceed in descent directions. Based on the application or preference of the relative dictionary sizes, we have the opportunity to schedule the increments of  $r_1$  and  $r_2$ . We will formalize and study more closely such an algorithm in Section 5.

**4.2. Connection with low-rank tensor decomposition.** Before we delve into the algorithmic side of the proposed dictionary learning approach, there are a few more important remarks to be made on the optimization problem (4.2). In particular, we give here an alternative interpretation of this problem in terms

of low-rank tensor decomposition and incidentally show a better bound for the number of dictionary elements of some global optima than the general one of Proposition 3.3.

First, we have the following statement showing that  $\Omega_\theta$  corresponds to the summation of the nuclear norms of each t-slice:

PROPOSITION 4.2. *With  $\theta(\gamma, \psi, c) = \|\gamma\|_2 \|\psi\|_2 \|c\|_1$ , we have:*

$$\Omega_\theta(\underline{X}) = \sum_{t=1}^T \|X_t\|_*.$$

*Proof.* First, with this choice of  $\theta$ , Proposition 3.3 shows that  $\Omega_\theta$  is a norm on  $\mathbb{R}^{G \times V \times T}$ . We can thus consider the dual norm that we write  $\dot{\Omega}_\theta$  defined by  $\dot{\Omega}_\theta(\underline{X}) = \sup_{\Omega_\theta(\underline{W}) \leq 1} \langle \underline{W}, \underline{X} \rangle_F$ . Now, for any  $\underline{W} \in \mathbb{R}^{G \times V \times T}$  such that  $\Omega_\theta(\underline{W}) \leq 1$ , using again Proposition 3.3, we can write  $\underline{W} = \underline{C} \times_1 \Gamma \times_2 \Psi$  with  $\Theta(\Gamma, \Psi, \underline{C}) = \Omega_\theta(\underline{W}) \leq 1$  and:

$$\langle \underline{W}, \underline{X} \rangle_F = \sum_{t=1}^T \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} c_{i,j,t} \Gamma_i^T X_t \Psi_j,$$

which shows that:

$$\dot{\Omega}_\theta(\underline{X}) = \sup_{\Gamma, \Psi, \underline{C}} \sum_{t=1}^T \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} c_{i,j,t} \Gamma_i^T X_t \Psi_j \text{ s.t. } \Theta(\Gamma, \Psi, \underline{C}) \leq 1.$$

Defining  $\Upsilon = \sup_{\gamma, \psi, c} \sum_{t=1}^T c_t \gamma^T X_t \psi$  s.t  $\theta(\gamma, \psi, c) \leq 1$ , it is then clear that  $\Upsilon \leq \dot{\Omega}_\theta(\underline{X})$ . Conversely, for any  $\Gamma, \Psi, \underline{C}$  s.t.  $\Theta(\Gamma, \Psi, \underline{C}) \leq 1$ :

$$\begin{aligned} \sum_{t=1}^T \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} c_{i,j,t} \Gamma_i^T X_t \Psi_j &= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{t=1}^T c_{i,j,t} \Gamma_i^T X_t \Psi_j \\ &\leq \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \Upsilon \\ &\leq \Theta(\Gamma, \Psi, \underline{C}) \Upsilon \leq \Upsilon \end{aligned}$$

which leads to

$$\dot{\Omega}_\theta(\underline{X}) = \sup_{\gamma, \psi, c} \sum_{t=1}^T c_t \gamma^T W_t \psi \text{ s.t. } \theta(\gamma, \psi, c) \leq 1.$$

Following the same reasoning as in the proof of Corollary 4.1, we obtain that:

$$\dot{\Omega}_\theta(\underline{X}) = \max_{t=1, \dots, T} \sigma_{\max}(W_t).$$

We have now by biduality  $\Omega_\theta(\underline{X}) = \sup_{\dot{\Omega}_\theta(\underline{W}) \leq 1} \langle \underline{X}, \underline{W} \rangle_F$ . For any  $\underline{W}$  such that  $\dot{\Omega}_\theta(\underline{W}) \leq 1$ , i.e.  $\sigma_{\max}(W_t) \leq 1$  for all  $t$ , since  $\langle \underline{X}, \underline{W} \rangle_F = \sum_{t=1}^T \langle X_t, W_t \rangle_F$ , it follows from the standard expression of the nuclear norm of matrices and its dual that:

$$\Omega_\theta(\underline{X}) = \sum_{t=1}^T \|X_t\|_* = \sum_{t=1}^T \sum_{i=1}^r \sigma_i(X_t),$$

where  $r = \min(V, G)$  and  $\sigma_i(X_t)$  are all the singular values of the slice  $X_t$ .  $\square$

Based on this alternative expression of the regularizer, we can rewrite the convex problem over  $\underline{X}$  of (3.4) as a slice by slice low rank regularization of the signal:

$$\min_{\underline{X}} \sum_{t=1}^T \ell(S_t, X_t) + \lambda \|X_t\|_*.$$

Thus, we have just shown that the solution of the dictionary learning problem with respect to the tensor  $\underline{X}$  ((1.5)) is essentially obtained by solving a set of low rank matrix approximation problems. This type of problem has been studied in many past works for instance on matrix completion [9, 7].

In the rest of this section, let's consider the simple particular fidelity term given by  $\ell(S_t, X_t) = \frac{1}{2} \|S_t - X_t\|_F^2$ . It is well-known that each optimal  $X_t$  is then given by the singular value shrinkage operator applied to  $S_t$ , i.e. (with the notations of [7]) for all  $t = 1, \dots, T$ ,  $X_t = \mathcal{D}_\lambda(S_t)$  where  $\mathcal{D}_\lambda(Y) = U\mathcal{D}_\lambda(\Sigma)V^T$  with  $U\Sigma V^T$  being the SVD of  $Y$  and  $\mathcal{D}_\lambda(\Sigma) = \text{Diag}([\sigma_i - \lambda]_+)$ . This does not yet characterize solutions for the dictionary learning problem itself as one still needs to obtain an optimal factorization for  $\underline{X}$  for the regularizer  $\Omega_\theta$ . However, in that particular case, “explicit” global minima of (4.2) can be in fact constructed as follows. Writing  $S_t = U_t \Sigma_t V_t^T$  the SVD of  $S_t$ , in which  $U_t \in O(G)$ ,  $V_t \in O(V)$  and  $\Sigma_t = \text{Diag}(\sigma_{t,1}, \dots, \sigma_{t,m})$  with  $m = \min\{G, V\}$ , one can set  $r_1 = GT$ ,  $r_2 = VT$  and  $\Gamma = [U_1, \dots, U_T]$ ,  $\Psi = [V_1, \dots, V_T]$ ,  $\underline{C} \in \mathbb{R}^{GT \times VT \times T}$  such that

$$C_t = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & \boxed{\mathcal{D}_\lambda(\Sigma_t)} & \\ & & & \ddots \\ & & & 0 \end{pmatrix}.$$

Then, by construction,  $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$  and it is a simple verification that  $(\Gamma, \Psi, \underline{C})$  satisfy the two optimality conditions of Corollary 4.1. Consequently, for that particular choice of  $\theta$  and  $\ell$ , we see that global minima in (4.2) exist with  $r_1 \leq GT$  and  $r_2 \leq VT$ . In addition, solutions can be computed based on the SVDs of the slices  $S_t$  as we just described.

There are however several remaining limitations to this approach for solving the dictionary learning problem. From a numerical point of view, computing that many complete SVDs of such potentially large matrices can prove very intensive for practical applications. But more importantly, although the resulting sizes of dictionaries  $\Gamma$  and  $\Psi$  are smaller than the upper bound of Proposition 3.3, those are still constructed by direct concatenation of one dictionary for each slice while enforcing a diagonality constraint on the  $C_t$ 's. From the perspective of dictionary learning, one is typically interested in more compact representations with total number of atoms in  $\Gamma$  and  $\Psi$  comparable to the size of the data (i.e. with  $r_1 \sim G$ ,  $r_2 \sim V$ ) and not dependent on the number  $T$  of training samples. In the following sections, we will show empirically that much more compact solutions can be found by instead introducing a more efficient algorithm that iteratively increases  $r_1$  and  $r_2$  until global optimality conditions are satisfied.

**5. Algorithm to Reach Global Minimum.** Now that Corollary 4.1 provides practical conditions to guarantee global minimality of the separable dictionary learning problem, we will outline an algorithm to reach a globally optimal solution. This involves alternating between two main sub-routines: 1) local descent to reach a stationary point with fixed number of atoms  $r_1$  and  $r_2$  in the dictionaries, and 2) a check for global optimality via Corollary 4.1. Note that since we consider the particular choice of regularizer  $\theta(\gamma, \psi, c) = \|\gamma\|_2 \|\psi\|_2 \|c\|_1$ , the global optimality check only amounts to verifying that a stationary point satisfies condition 2 in Corollary 4.1. If by the end of the local descent we have not reached a globally optimal solution, then we can find a global descent direction by adding additional atoms to the dictionaries. Algorithm 5.1 describes this general meta-algorithm in more detail and refers to each sub-routine discussed in the following sections.

**5.1. Proximal Gradient Descent to Stationary Point.** In this section, we provide an algorithm to find a stationary point of the separable dictionary learning problem with fixed sizes for the dictionaries. We again state the problem:

$$(5.1) \quad \min_{\Gamma, \Psi, \underline{C}} \frac{1}{2} \sum_{t=1}^T \|\Gamma C_t \Psi^\top - S_t\|_F^2 + \lambda \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \|\Gamma_i\|_2 \|\Psi_j\|_2 \|C_{i,j}\|_1.$$

For an optimization problem of the form  $\min_x \{\ell(x) + \lambda \Theta(x)\}$ , where  $\ell$  is differentiable and  $\Theta$  is non-differentiable, proximal gradient descent [35] is a common algorithm to arrive at a stationary points, i.e. local

---

**Algorithm 5.1** Meta-Algorithm: Local Descent and Global Optimality Check

---

```

Initialize dictionaries with set number of atoms.
while not globally optimal do
    while objective residual  $> \epsilon$  do
        descent to local minimum via Algorithm 5.2
    end while
    if Condition 2 is satisfied then
        solution is globally optimal
    else
        update dictionaries via Algorithm 5.3
    end if
end while

```

---

minima or saddle points. The general updates for proximal gradient descent follow:

$$(5.2) \quad x^{k+1} = \text{prox}_{\tau\lambda\Theta(\cdot)}(x^k - \tau\nabla\ell),$$

where  $\text{prox}_{\tau\lambda\Theta(\cdot)}(y) = \arg \min_x \left\{ \frac{1}{2\tau\lambda} \|x - y\|_2^2 + \Theta(x) \right\}$ . To solve (5.1), we apply a proximal gradient descent step to each variable while holding the rest constant. This local descent procedure is outlined in Algorithm 5.2. Recall that  $\ell(\Gamma, \Psi, \underline{C}) = \frac{1}{2} \sum_{t=1}^T \|\Gamma C_t \Psi^\top - S_t\|_F^2$ . We derive the update for each variable as:

$$(5.3) \quad \Gamma_i^{k+1} = \text{prox}_{\xi_i^k \|\cdot\|_2}(\Gamma_i^k - \xi_i^k [\nabla_{\Gamma^k} \ell]_i)$$

$$(5.4) \quad c_{i,j,t}^{k+1} = \text{prox}_{\kappa_{i,j}^k |\cdot|}(c_{i,j,t}^k - \kappa_{i,j}^k [\nabla_{C_t^k} \ell]_{i,j})$$

$$(5.5) \quad \Psi_j^{k+1} = \text{prox}_{\pi_j^k \|\cdot\|_2}(\Psi_j^k - \pi_j^k [\nabla_{\Psi^k} \ell]_j).$$

where the proximal operators for  $\|\cdot\|_2$  and  $|\cdot|$  can be written in closed form:

$$(5.6) \quad \text{prox}_{\tau\|\cdot\|_2}(x) = \begin{cases} (1 - \frac{\tau}{\|x\|_2})x & \text{for } \|x\|_2 \geq \tau \\ 0 & \text{otherwise} \end{cases},$$

$$(5.7) \quad \text{prox}_{\tau|\cdot|}(\alpha) = \max(0, \alpha - \tau) - \max(0, -\alpha - \tau),$$

for  $x \in \mathbb{R}^N$ ,  $\alpha \in \mathbb{R}$  and  $\tau \geq 0$ . Then

$$(5.8) \quad \nabla_\Gamma \ell = \sum_{t=1}^T (\Gamma C_t \Psi^\top - S_t) \Psi C_t^\top$$

$$(5.9) \quad \nabla_{C_t} \ell = \Gamma^\top (\Gamma C_t \Psi^\top - S_t) \Psi$$

$$(5.10) \quad \nabla_\Psi \ell = \sum_{t=1}^T (\Psi C_t^\top \Gamma^\top - S_t^\top) \Gamma C_t.$$

Finally,  $\xi_i$ ,  $\kappa_{i,j}$ , and  $\pi_j$  are constants composed of the other fixed variables in  $\theta$ , specifically

$$(5.11) \quad \xi_i := \lambda \sum_{t=1}^T \sum_{j=1}^{r_2} |c_{i,j,t}| \|\Psi_j\|_2 / L_\Gamma$$

$$(5.12) \quad \kappa_{i,j} := \lambda \|\Gamma_i\|_2 \|\Psi_j\|_2 / L_{C_t}$$

$$(5.13) \quad \pi_j := \lambda \sum_{t=1}^T \sum_{i=1}^{r_1} |c_{i,j,t}| \|\Gamma_i\|_2 / L_\Psi,$$

where the parameters  $1/L_\Gamma$ ,  $1/L_{C_t}$ , and  $1/L_\Psi$  correspond to the step sizes in the proximal gradient descent.

In general, to determine an appropriate step-size  $\tau$ , it has been shown that convergence is guaranteed if  $\tau \leq \frac{1}{L}$ , where  $L$  is the Lipschitz constant of  $\nabla l$ :

$$(5.14) \quad \|\nabla \ell(x^{(1)}) - \nabla \ell(x^{(2)})\|_2 \leq L \|x^{(1)} - x^{(2)}\|_2.$$

In our this setting, we can calculate (or at least bound) the Lipschitz constants with respect to,  $L_\Gamma$ ,  $L_{\underline{C}_t}$ , and  $L_\Psi$ . For  $L_\Gamma$  we have indeed:

$$\begin{aligned}
\|\nabla_\Gamma \ell(\Gamma^{(1)}) - \nabla_\Gamma \ell(\Gamma^{(2)})\|_F &= \left\| \sum_{t=1}^T (\Gamma^{(1)} C_t \Psi^\top - S_t) \Psi C_t^\top - (\Gamma^{(2)} C_t \Psi^\top - S_t) \Psi C_t^\top \right\|_F \\
&= \left\| \sum_{t=1}^T \Gamma^{(1)} C_t \Psi^\top \Psi C_t^\top - \Gamma^{(2)} C_t \Psi^\top \Psi C_t^\top \right\|_F \\
&= \left\| \sum_{t=1}^T C_t \Psi^\top \Psi C_t^\top (\Gamma^{(1)} - \Gamma^{(2)}) \right\|_F \\
&\leq \left\| \sum_{t=1}^T C_t \Psi^\top \Psi C_t^\top \right\|_F \|\Gamma^{(1)} - \Gamma^{(2)}\|_F \\
&= L_\Gamma \|\Gamma^{(1)} - \Gamma^{(2)}\|_F
\end{aligned}$$

where  $L_\Gamma = \|\sum_{t=1}^T C_t \Psi^\top \Psi C_t^\top\|_F$  is thus an upper bound for the Lipschitz constant of  $\nabla_\Gamma \ell$ . Similarly for  $\nabla_\Psi \ell$ , we can take as the Lipschitz constant  $L_\Psi = \|\sum_{t=1}^T C_t \Gamma^\top \Gamma C_t^\top\|_F$ . Then for  $\nabla_{\underline{C}_t} \ell$ ,  $L_{\underline{C}_t} = \|\Gamma^\top \Gamma\|_F \|\Psi^\top \Psi\|_F$ . Lastly, the convergence of the descent can be accelerated through the standard Nesterov scheme described as an extension of the Proximal Gradient Descent in Algorithm 7.1 in Appendix B.

---

**Algorithm 5.2** Proximal Gradient Descent

---

```

Initialize:  $k = 0, \Gamma^0, \Psi^0, \underline{C}^0, \lambda, r_1, r_2$ .
while error  $> \epsilon$  do
    Update  $\Gamma^k$  via (5.3)
    Update  $\underline{C}^k$  via (5.4)
    Update  $\Psi^k$  via (5.5)
     $k \rightarrow k + 1$ 
end while
return stationary point  $(\tilde{\Gamma}, \tilde{\Psi}, \tilde{\underline{C}})$ 

```

---

**5.2. Global Optimality Check.** Once proximal gradient descent reaches a stationary point via Algorithm 5.2, we check if the solution is a global minimum. By the result of Corollary 4.1, and since the first condition is always satisfied by a stationary point, one just needs to check if (2) holds. If so, we have reached a global minimum and the algorithm stops. If not, by adding additional atoms to the dictionaries, we can escape from the local minimum or saddle point we have reached and search for a global descent direction. Following the discussions in [24] for matrix factorization involving the nuclear norm, by appending the locally optimal dictionaries with the singular vectors of the maximum singular value in (2), we are guaranteed to move in a global descent direction. First, let  $t_* = \arg \max_t \sigma_{\max}(W_t)$ . Then with  $(\gamma_{t_*}, \psi_{t_*})$  the left and right singular vector pair corresponding to the maximum singular value of  $W_t$  over all  $t$ , we can update the locally optimal dictionaries  $\tilde{\Gamma}$  and  $\tilde{\Psi}$  by appending the last column  $\Gamma = [\tilde{\Gamma}, \gamma_{t_*}]$  and  $\Psi = [\tilde{\Psi}, \psi_{t_*}]$ . Finally,  $\underline{C}$  can be updated by appending the slice corresponding to the maximum singular value by

$$(5.15) \quad C_{t_*} = \begin{bmatrix} \tilde{C}_{t_*} & 0 \\ 0 & \tau \end{bmatrix}$$

and appending a matrix of zeros for all other slices  $C_t$  for  $t \neq t_*$ . Here  $\tau$  is the step-size and is an important parameter to select. In our formulation, the optimal  $\tau^*$  at each iteration can be found by solving:

$$(5.16) \quad \tau^* = \arg \min_{\tau} \frac{1}{2} \sum_t^T \|S_t - \hat{X}_t - \tau E_t\|_F^2 + \lambda |\tau|,$$

where  $E_{t_*} = \gamma_{t_*} \psi_{t_*}^\top$  and 0 for all  $t \neq t_*$ . By vectorizing all tensors, (5.16) reduces to the simple proximal operator of the absolute value function given in closed-form by the soft-thresholding solution.

Now, because of the separable form of this problem, we actually have the option to update just one of the two dictionaries, and not both simultaneously during each global check. In particular, if  $\gamma_{t_*} \in \text{Span}(\tilde{\Gamma})$  then it is unnecessary to add this atom to the dictionary. The same goes for  $\psi_{t_*}$ . Instead of checking these conditions *a posteriori*, we can check augmented criteria akin to (2) with the added constraint that  $\gamma \in \text{Span}(\tilde{\Gamma})$  which by definition means that there exists an  $\alpha$  such that  $\gamma = \tilde{\Gamma}\alpha$ . Using this we can make a change of variable in (2) as:

$$(5.17) \quad \sum_{t=1}^T c_t \alpha^\top \tilde{\Gamma}^\top W_t \psi \leq \|\tilde{\Gamma}\alpha\|_2 \|\psi\|_2 \|c\|_1 \quad \forall(\alpha, \psi, c).$$

By noting that  $\|\tilde{\Gamma}\alpha\|_2 \leq \|\tilde{\Gamma}\|_2 \|\alpha\|_2 = \sigma_{\max}(\tilde{\Gamma}) \|\alpha\|_2$ , we can check the looser criterion:

$$(5.18) \quad \sum_{t=1}^T c_t \alpha^\top \tilde{\Gamma}^\top W_t \psi \leq \sigma_{\max}(\tilde{\Gamma}) \|\alpha\|_2 \|\psi\|_2 \|c\|_1 \quad \forall(\alpha, \psi, c).$$

Therefore, if (5.18) is violated then so is (5.17). We prefer to check (5.18) because of its simplicity to compute, which can be done as follows. As before, normalizing  $\hat{\psi} = \psi / \|\psi\|_2$ ,  $\hat{c}_t = c_t / \|c\|_1$ , and  $\hat{\alpha} / \|\alpha\|_2$  give

$$(5.19) \quad \begin{aligned} & \frac{1}{\sigma_{\max}(\tilde{\Gamma})} \sum_{t=1}^T \hat{c}_t \hat{\alpha}^\top \tilde{\Gamma}^\top W_t \hat{\psi} \leq 1 \quad \forall(\hat{\alpha}, \hat{\psi}, \hat{c}) \\ \iff & \sup_{\hat{\alpha}, \hat{\psi}, \hat{c}} \frac{1}{\sigma_{\max}(\tilde{\Gamma})} \sum_{t=1}^T \hat{c}_t \hat{\alpha}^\top \tilde{\Gamma}^\top W_t \hat{\psi} \leq 1 \quad \text{s.t. } \|\hat{\alpha}\|_2 = \|\hat{\psi}\|_2 = \|\hat{c}\|_1 = 1. \end{aligned}$$

This is equivalent to checking that

$$(5.20) \quad \max_{1 \leq t \leq T} \frac{1}{\sigma_{\max}(\tilde{\Gamma})} \sigma_{\max}(\tilde{\Gamma}^\top W_t) \leq 1.$$

If this inequality is violated, this implies that  $\gamma_{t_*}$ , the right singular vector of  $\tilde{\Gamma}^\top W_t$  corresponding to the maximum singular value could be appended to  $\tilde{\Gamma}$  to give a global descent direction. But because  $\gamma_{t_*} \in \text{Span}(\tilde{\Gamma})$ , it is not necessary to add it to find the descent direction. Therefore, we can just update  $\Psi$  and  $C$  as  $\Psi = [\tilde{\Psi}, \psi_{t_*}]$  and  $C_{t_*} = [\tilde{C}_{t_*}, \tau \alpha_{t_*}]$  and replacing  $\alpha_{t_*}$  by 0 for all other slices. The optimal step-size  $\tau$  can again be found by solving (5.16) with  $E_{t_*} = \tilde{\Gamma} \alpha_{t_*} \psi_{t_*}^\top$ .

On the other hand, if (5.20) is satisfied, we must then check the analogous criteria for  $\Psi$  with  $\psi = \tilde{\Psi}\beta$ :

$$(5.21) \quad \max_{1 \leq t \leq T} \sigma_{\max}(W_t \tilde{\Psi}) \leq 1$$

Now, if (5.21) is violated this means we do not need to update  $\Psi$  and just update  $\Gamma = [\tilde{\Gamma}, \gamma_{t_*}]$  and  $C_{t_*} = [\tilde{C}_{t_*}; \tau \beta_{t_*}]$  with  $\beta_{t_*}$  replaced by 0 for all other slices. The optimal step size  $\tau$  is found by (5.16) with  $E_{t_*} = \gamma_{t_*} \beta_{t_*}^\top \tilde{\Psi}^\top$ . If this too is satisfied, then we must check the original criteria (2) to potentially update both dictionaries if violated. The order of these global checks can depend on knowledge of the intended sizes of each dictionary. For our purposes we propose to check which of the two violates the corresponding constraint the most, i.e. which one leads to the larger global step. Because (5.20) and (5.21) are lower bounds of (2), satisfying them will not be sufficient to guarantee that we have reached a global minimum and so (2) is still necessary to check in this case. The complete procedure for the global optimality check and dictionary size update is outlined in Algorithm 5.3.

**6. Application: Diffusion Magnetic Resonance Imaging (dMRI).** Diffusion magnetic resonance imaging (dMRI) is a medical imaging modality that has the ability to reveal the complex network of neurons in the brain, *in vivo* [46]. By measuring restrictions of water diffusion brought on by the presence of bundles of neurons called fiber tracts, one can estimate the orientation of fiber bundles in the brain, and use these local estimations to reconstruct an entire network of fiber tract connections. This allows researchers to study anatomical brain variations that are essential for understanding and predicting neurological disorders such as Alzheimer's disease or traumatic brain injury.

---

**Algorithm 5.3** Global Optimality Check and Update

---

```

for  $t = 1 \dots T$  do
     $\hat{X}_t = \tilde{\Gamma} \tilde{C}_t \tilde{\Psi}^\top$ ;
     $g_t = \sigma_{\max}(-\tilde{\Gamma}^\top (\hat{X}_t - S_t) / \lambda \sigma_{\max}(\tilde{\Gamma}))$ ;
     $p_t = \sigma_{\max}(-(\hat{X}_t - S_t) \tilde{\Psi} / \lambda \sigma_{\max}(\tilde{\Psi}))$ ;
     $c_t = \sigma_{\max}(-(\hat{X}_t - S_t) / \lambda)$ ;
end for
 $g = \max_t g_t$ ;
 $p = \max_t p_t$ ;
 $c = \max_t c_t$ ;
if  $g > 1$  and  $g > p$  then
    Compute global step-size  $\tau$  via (5.16)
    Update  $\underline{C}$  and  $\Psi$ 
else if  $p > 1$  and  $p > g$  then
    Compute global step-size  $\tau$  via (5.16)
    Update  $\Gamma$  and  $\underline{C}$ 
else if  $c > 1$  then
    Compute global step-size  $\tau$  via (5.16)
    Update  $\Gamma$ ,  $\underline{C}$  and  $\Psi$ 
else
     $\Gamma^* = \tilde{\Gamma}; \underline{C}^* = \tilde{C}; \Psi^* = \tilde{\Psi}$ ;
     $\hat{S} = \hat{X}$ ;
end if

```

---

Diffusion measurements are acquired at each voxel in a brain volume and so dMRI images are naturally defined over the product of a spatial domain and a diffusion (or angular) domain. We are interested in this application because the unique structure of dMRI is well suited for separable dictionary learning which has not been attempted before. In the next section, we briefly introduce the structure of dMRI data and review the literature of dictionary learning applied in this field.

**6.1. Dictionary Learning for dMRI.** In dMRI, a diffusion signal  $s_v \in \mathbb{R}^G$  is measured at each voxel  $v = 1, \dots, V$  in a 3D brain volume, resulting in a spatial-angular signal of total dimension  $G \times V$ . Diffusion is commonly measured angularly on a unit sphere of the diffusion domain known as  $q$ -space and therefore diffusion signals are commonly represented by an angular or spherical dictionary,  $\Gamma \in \mathbb{R}^{G \times r_1}$ , e.g. spherical harmonics, spherical wavelets, etc. Then, from these diffusion signals, researchers estimate the probability of the orientation of a fiber bundle in each voxel, known commonly as an orientation distribution function (ODF)<sup>1</sup> which is often the starting point of tractography algorithms that exploit this diffusion information to extract fiber bundles.

Learning dictionaries directly from dMRI data has been proposed for many applications such as denoising and compressed sensing and aims to solve the classical dictionary learning problem (1.3):

$$(6.1) \quad \min_{\Gamma, W} \frac{1}{2} \|\Gamma W - Y\|_F^2 + \lambda \|W\|_1 \quad \text{s.t.} \quad \|\Gamma_i\|_2 \leq 1 \quad \text{for } i = 1 \dots r_1,$$

where  $Y = [y_1, \dots, y_T] \in \mathbb{R}^{G \times T}$  are  $T$  training examples of angular signals taken, for example, from a subset of representative voxels in a brain image, and  $W \in \mathbb{R}^{r_1 \times T}$  are the associated angular coefficients. There have been a multitude of works that aim to solve (6.1) or some alternative versions by proposing different models like parametric dictionary learning [2, 10, 11, 30, 31, 48], which learn parameters of predefined diffusion models, Bayesian learning [21, 36] manifold learning [44] and dictionary learning directly from undersampled data for compressed sensing [6, 19, 20, 21, 29].

While some of these methods impose additional spatial coherence between neighboring voxels [48], training examples  $y_t$  are usually taken voxel-wise without consideration of their spatial correlations. In other words, these works have not considered learning joint spatial-angular dictionaries in the context of dMRI.

---

<sup>1</sup>In our visualizations of dMRI signals, we will display the estimated ODFs of each voxel instead of the diffusion signal itself.

To incorporate the spatial domain, we can compile the signals  $s_v$  at each voxel of the brain volume into the matrix  $S = [s_1, \dots, s_V] \in \mathbb{R}^{G \times V}$ . Then, given an angular dictionary  $\Gamma \in \mathbb{R}^{G \times r_1}$  and a spatial dictionary  $\Psi \in \mathbb{R}^{V \times r_2}$ , the entire dMRI image may be represented (or approximated) as

$$(6.2) \quad S = \Gamma C \Psi^\top,$$

where  $C \in \mathbb{R}^{r_1 \times r_2}$  stores the coefficients in the joint spatial-angular dictionary. This separable spatial-angular representation of dMRI data fits directly into our separable dictionary framework (1.5) with  $T$  training example diffusion volumes  $\{S_t\}_{t=1}^T \in \mathbb{R}^{G \times V}$ .

In prior works [39, 41], the authors demonstrated that sparse coding with separable dictionaries over the spatial and angular domain provides sparser reconstructions than the traditional angular sparse coding. Thus, we should be able to provide sparser reconstructions than the state of the art by learning both spatial and angular dictionaries. Related works also support the use of analytic spatial-angular dictionaries in order to lower subsampling rates for dMRI compressed sensing [40]. In the longer run, we hope to improve such results through spatial-angular dictionaries learned from the data. To the best of our knowledge, only the work of [42] considers both the spatial and angular components in dictionary learning for dMRI. In this paper, the authors solve the classical dictionary learning problem (1.3) with  $y_t$  as vectorization of a local four dimensional spatial-angular block used within a non-local means denoising algorithm.

**6.2. Patch-Based Training for dMRI.** In theory, our spatial-angular dictionary learning method is capable of learning global spatial and angular dictionaries,  $\Psi \in \mathbb{R}^{G \times r_1}$  and  $\Gamma \in \mathbb{R}^{V \times r_2}$ , over an entire dMRI dataset of size  $G \times V$ . However, the typical size of a HARDI brain volume is on the order of  $V = 100^3$  voxels, and  $G = 100$  diffusion measurements, i.e. of size  $G \times V = 10^8$ . Furthermore, the number of training examples  $T$  depends on the size of the training sample. This would require a very large number of training examples of entire dMRI datasets, which is largely infeasible for our algorithm. Because the spatial domain is orders of magnitude larger than the angular domain, one way to curb the computational burden is to reduce our dictionary learning to local spatial patches for all diffusion measurements (i.e. 3D patches of size  $P \times P \times P$ ). Patch-based methods are indeed very popular in image processing for tasks such as denoising, filtering, inpainting, and object detection [49]. In addition, local dictionaries are beneficial for capturing local features that are often repeated in an image, such as edges, textures or objects. Note that another possible approach could consist in learning separable dictionaries in the different coordinates  $x, y, z$  of the spatial domain as well and thus involve separable dictionaries with a larger number of factors.

For training we thus choose a random selection of spatial patches that is consistent along the diffusion domain. For computational simplicity and purposes of visualization, we limit our initial experiments to 2D spatial patches of size  $P \times P$  instead of 3D, i.e.  $S_t \in \mathbb{R}^{G \times P^2}$ . Depending on the detail and size of an image, popular patch sizes range from  $P = 5$  to  $15$ . For our data,  $P = 12$  gives a good amount of detail and is not too large to process.

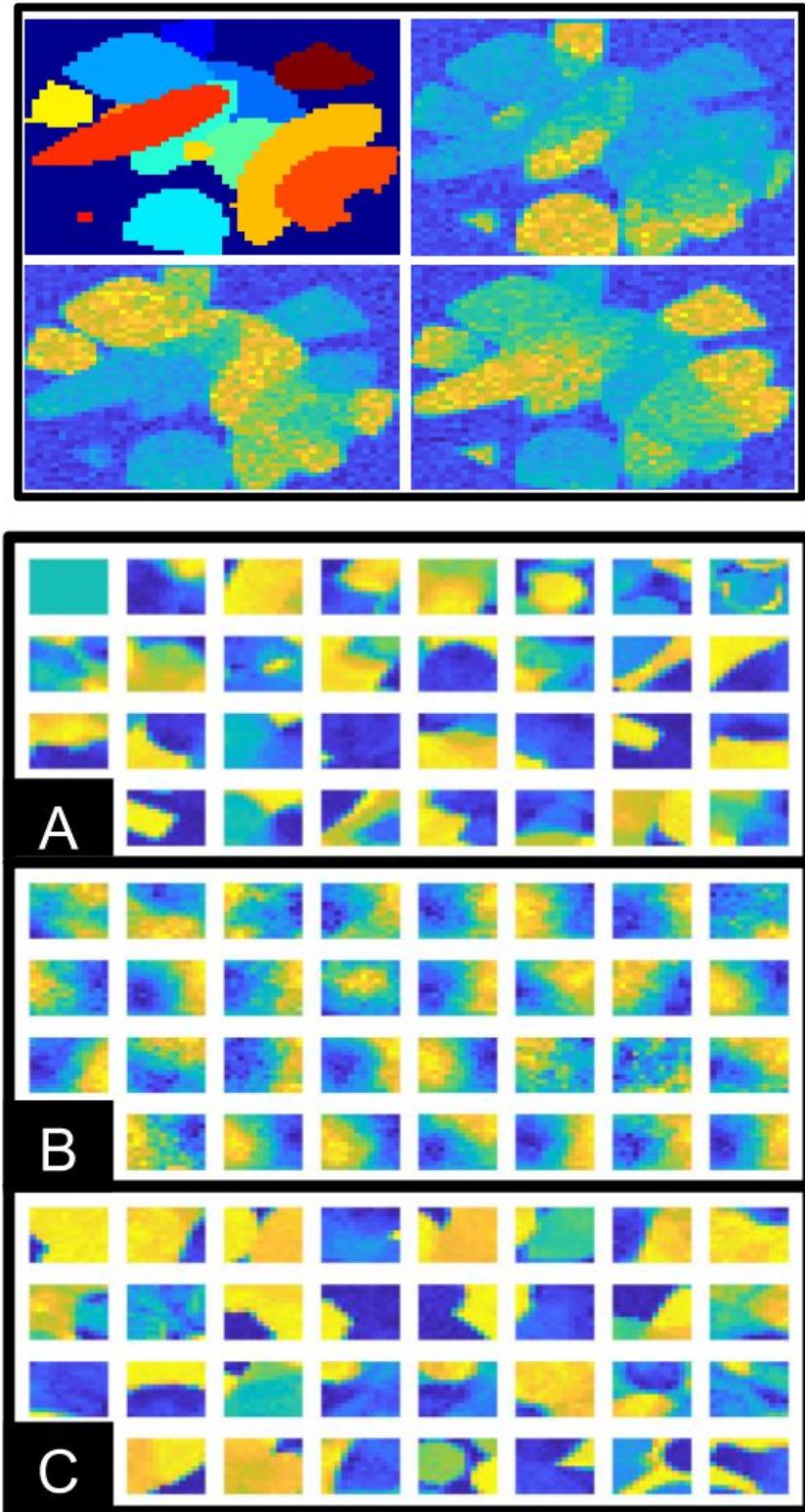
For choosing the number of training examples,  $T$ , we can consider the number of training examples typical of angular dictionary learning. For instance, the work of [30] use 5,000 angular signals to train their angular dictionary. To reach this number, we only need a relatively small number of  $G \times P \times P$  patches to provide an adequate number of spatial and angular training examples, respectively. In total, the number of angular training examples will be  $P^2 T$  and the number of spatial training examples will be  $G T$ . For a typical dMRI dataset with  $G = 100$  and  $P = 12$ , we will need on the order of  $T = 40$  training patches, to have 5,000 angular training examples and 4,000 spatial training examples. In this work, we use  $T = 100$  training patches over multiple image slices, resulting in about 14,400 angular training examples and 10,000 spatial training examples.

**6.3. Denoising Experiment.** For our application we learn our dictionaries from high angular resolution diffusion imaging (HARDI) data [14]. Specifically, we experimented on a phantom and a real HARDI brain dataset. The phantom is taken from the ISBI 2013 HARDI Reconstruction Challenge<sup>2</sup>, a  $V = 50 \times 50 \times 50$  volume consisting of 20 phantom fibers crossing intricately within an inscribed sphere, measured with  $G = 64$  diffusion measurements. Our initial experiments test on a 2D  $50 \times 50$  slice of this data for simplification.

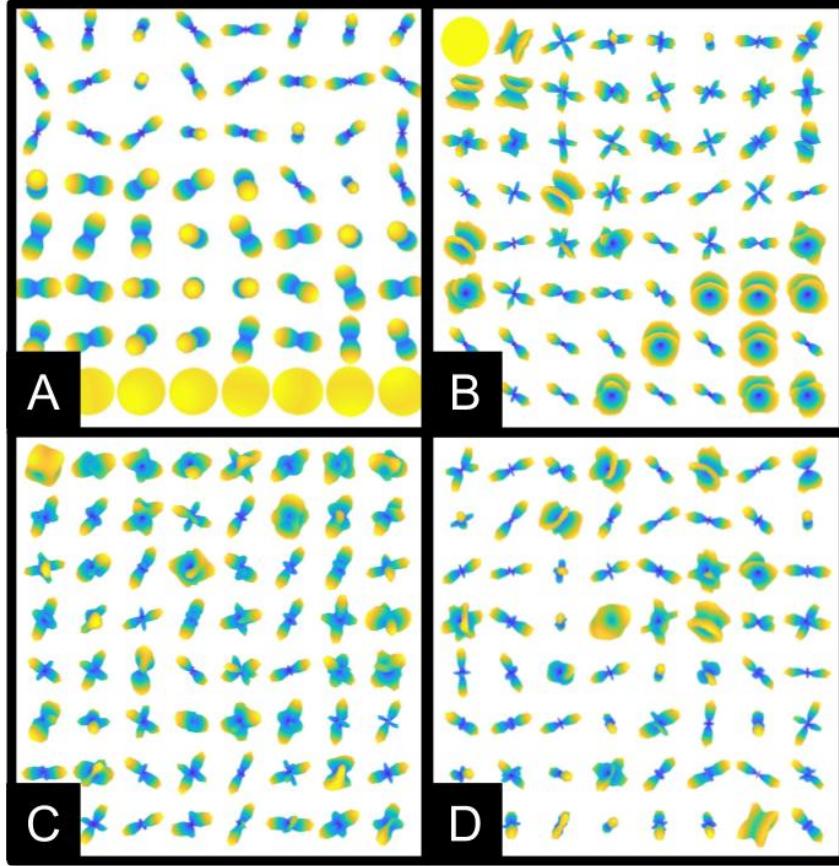
The phantom dataset includes two noise levels: a low noise level of SNR=30 dB and a high noise level of SNR=10 dB. The denoising task will be to denoise the SNR=10 dB data using dictionaries learned from the

---

<sup>2</sup>[http://hardi.epfl.ch/static/events/2013\\_ISBI/](http://hardi.epfl.ch/static/events/2013_ISBI/)



**Figure 1:** Top: Phantom HARDI ground truth fiber segmentations and three diffusion weighted images used for training on patches of size  $12 \times 12$ . Bottom: Subset of spatial patch dictionaries learned via A. KSVD independently from angular dictionary, B. KDRSDL jointly with angular dictionary, C. the proposed method jointly with angular dictionary. B. appears to have reached a spurious local minimum while A. and C. closely resemble each other and pick up sharp edges and shapes present in the training phantom.



**Figure 2:** Comparison of angular dictionaries. *A.* Fixed spherical ridglets. *B.* – *D.* Subset of angular dictionaries trained on the phantom HARDI data learned via *B.* KSVD independently from spatial dictionary, *C.* KDRSDL jointly with spatial dictionary, and *D.* the proposed method jointly with spatial dictionary. KSVD and the proposed method produce clean single fiber ODFs while KDRSDL ODFs are noisier.

SNR=30 dB data and record the error with respect to the “ground truth” SNR=30 dB data by calculating Peak SNR (PSNR):

$$(6.3) \quad PSNR = 10 \log_{10} \frac{MAX_I^2}{MSE},$$

where  $MAX_I$  indicates the maximum value in the original SNR=30 dB signal, and MSE is the mean squared error between the original SNR=30 dB signal and the reconstruction. The higher the PSNR, the more accurate the reconstruction will be. We chose a subset of slices of the SNR=30 dB to learn our 2D spatial-angular dictionaries and used a selection of the remaining slices as test data for denoising.

After validation on phantom data, we show qualitative denoising results on a real HARDI volume with  $G = 127$  diffusion measurements using our proposed spatial-angular dictionaries learned on a subset of 2D slices.

**4.4. Methods.** We validate the proposed separable dictionary learning method by showing its performance on denoising HARDI data. While there are numerous denoising methodologies in the literature, we will focus on utilizing learned dictionaries in a sparse denoising method, which has been used frequently in the dMRI literature [19]. Note that our aim here is primarily to evaluate and compare different dictionary learning strategies through sparse denoising experiments but not to compare those results with the most advanced denoising algorithms in the field which typically involve additional processing steps [42].

For our learned spatial and angular dictionaries we use the spatial-angular sparse coding approach proposed

Comparison	1		2		3		4	
Method	Angular	Spatial-Angular	Fixed	Learned	Separate	Joint	Local	Global
I-SR	✓		✓					
I-SR + TV	✓		✓					
Curve-SR		✓	✓					
I-KSVD	✓			✓	✓		✓	
KSVD-KSVD		✓		✓	✓		✓	
KDRSDL		✓		✓		✓	✓	
Proposed		✓		✓		✓		✓

**Table 1:** Checklist of properties for each dictionary type to compare each method. Purple indicates fixed dictionaries, pink indicates spatial and/or angular dictionaries learned independently, and green indicates a joint spatial-angular dictionary.

in [39, 41] which amounts in solving (1.5) only for  $C$  with  $T = 1$ . For spatial patch-based dictionaries, we will apply sparse coding for each patch and average the results across overlapping patches. For denoising, we choose a value of  $\lambda$ , consistent for all patches, that gives the highest PSNR.

To validate the results of our proposed separable dictionary learning method we consider four dictionary comparisons based on the denoising performance:

1. **Angular vs. Spatial-Angular:** will the proposed spatial-angular framework for dictionary learning and sparse coding outperform state-of-the-art framework for angular dictionary learning and sparse coding for denoising?
2. **Fixed vs. Learned:** will dictionaries learned from dMRI data outperform fixed analytic dictionaries for denoising?
3. **Separate vs. Joint:** will learning spatial and angular dictionaries jointly via separable dictionary learning better represent dMRI data than learning spatial and angular dictionaries independently each by classical methods like KSVD?
4. **Local vs. Global:** will our globally optimal separable dictionary learning outperform other locally optimal separable dictionary learning methods?

For comparison 1, we will compare against state-of-the-art angular dictionary learning and sparse coding frameworks. In particular, we will solve the angular dictionary learning problem (6.1) with the commonly used KSVD algorithm [1]. For the angular sparse denoising step, we also add a spatial regularization term based on the total-variation (TV) in the spatial domain, as is commonly done in state-of-the-art dMRI denoising [5].

For comparison 2, we will compare against two fixed angular and spatial dictionaries used in the dMRI literature: the spherical ridgelet (SR) dictionary popularly used in angular sparse coding and compressed sensing for dMRI [45, 34, 33, 32] (see Figure 2 A for visualization) and, for the spatial domain, the curvelet dictionary which has proved to be very efficient in sparsely representing classical images and was also shown to be a good choice for representing dMRI images in our recent works [39, 41].

For comparison 3, we will use the KSVD algorithm [1] to learn spatial and angular dictionaries independently. Identifying whether the proposed joint learning method is advantageous over the faster and easier approach of applying KSVD to each domain separately is indeed an important point to examine.

Finally, for comparison 4, we evaluate our approach against the Kronecker-Decomposable Robust Sparse Dictionary Learning (KDRSDL) algorithm of [4], which is also a separable dictionary learning method that does not, however, provide guarantees of global optimality. KDRSDL solves a low-rank variation of (1.5) which the authors show is useful for background subtracting and image denoising.

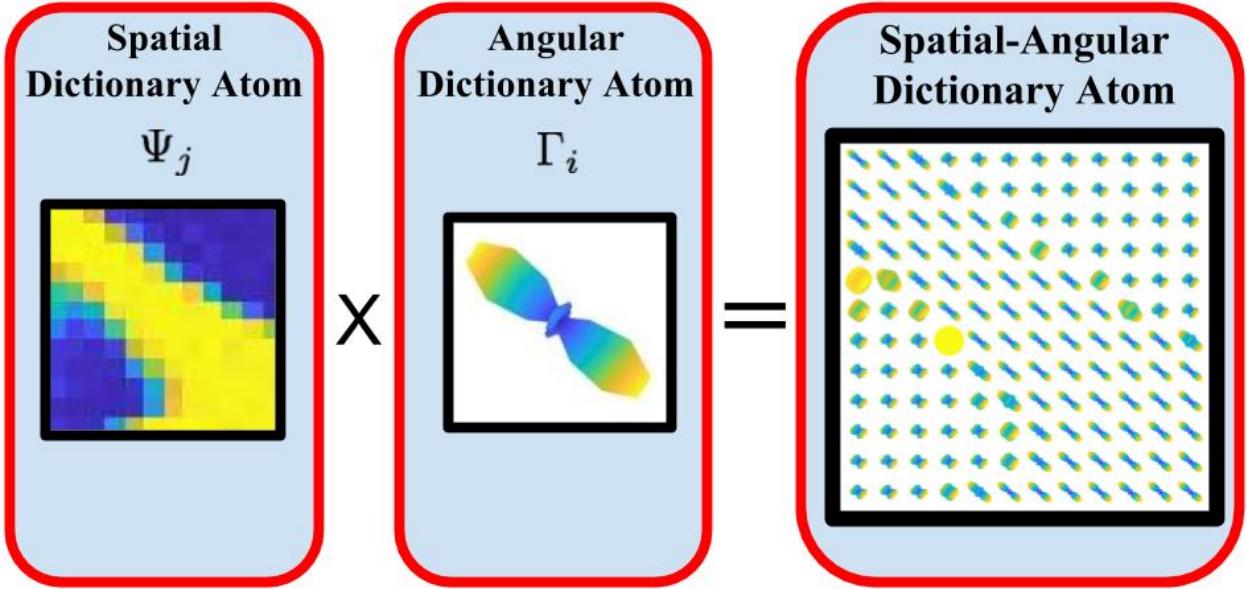
We use a “Spatial-Angular” notation to keep track of the different dictionary choices, where, for example, I-SR uses the identity for the spatial dictionary and spherical ridgelets for the angular dictionary, I-KSVD learns the angular dictionary only using KSVD, and KSVD-KVSD uses the spatial and angular dictionaries learned by KSVD independently. See Table 1 for a checklist of the different dictionary properties for each of the 4 comparisons and Table 2 for a summary of the spatial and angular domains for each method.

**6.5. Visualization.** In Figures 1 and 2 we visualize the spatial and angular dictionaries learned from each method on phantom HARDI data as well as the spherical ridgelet dictionary atoms in Figure 2 A. The learned dictionary atoms are organized left to right from top to bottom by the number of training examples that used each atom, i.e. the number of nonzero coefficients associated to each atom in training. For KSVD,

		Angular			
		SR	KSVD	KDRSDL	Proposed
I		I-SR (+ TV)	I-KSVD		
Spatial	Curve	Curve-SR	Curve-KSVD		
	KSVD	KSVD-SR	KSVD-KSVD		
	KDRSDL			KDRSDL	
	Proposed				Proposed

**Table 2:** Organization of spatial and angular dictionaries. Purple indicates fixed dictionaries, pink indicates spatial and/or angular dictionaries learned independently, and green indicates a joint spatial-angular dictionary.

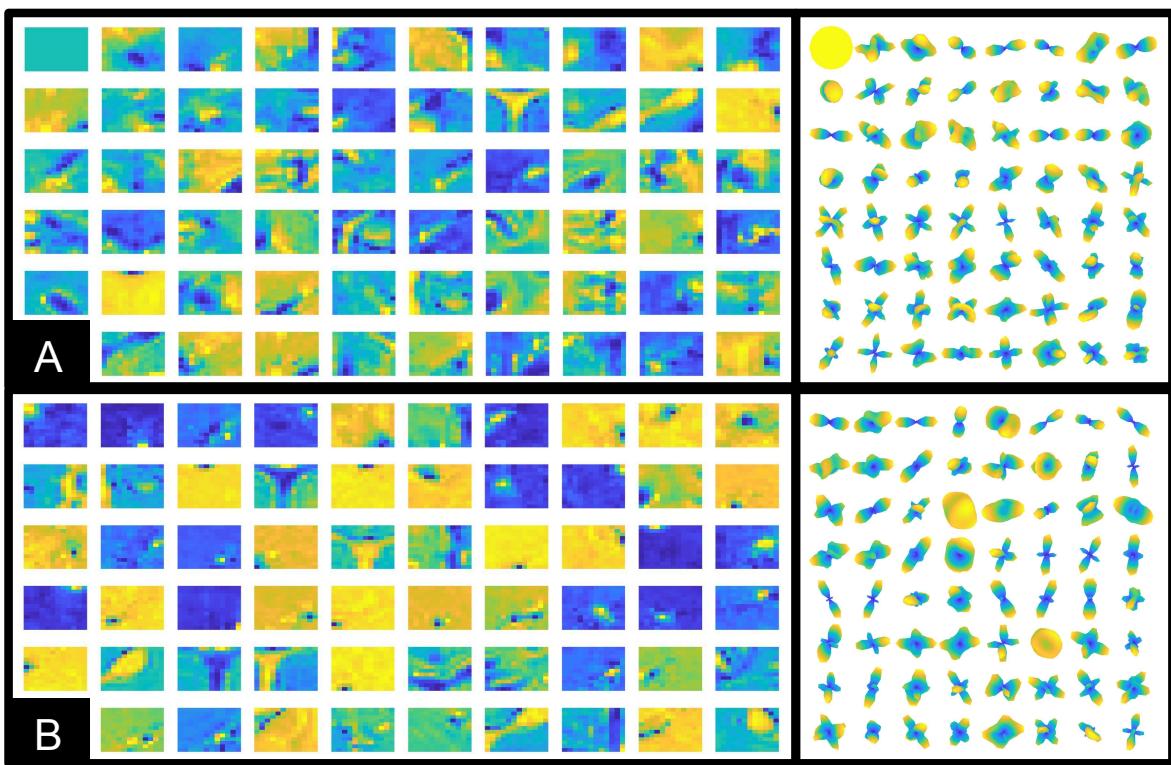
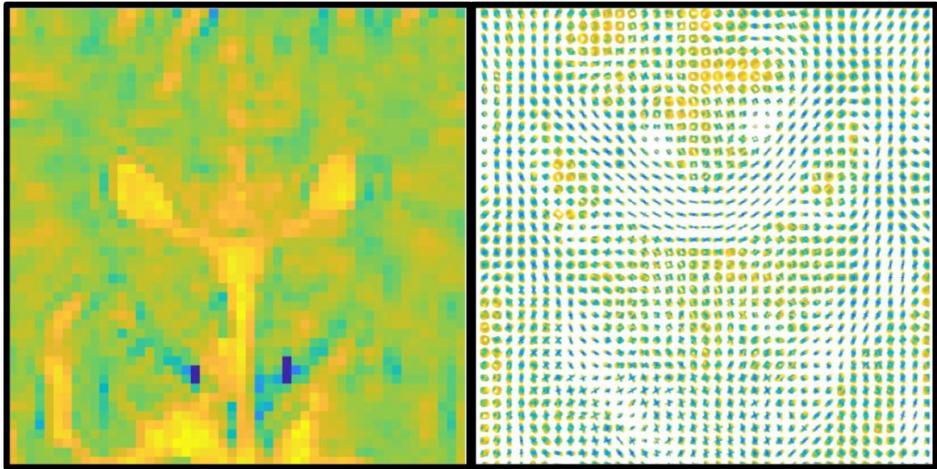
this ordering is independent for the spatial and angular dictionaries, while the atoms resulting from KDRSDL and the proposed method are ordered jointly (without repeats), i.e. the top left spatial and angular atoms combine to create the most utilized spatial-angular atom.



**Figure 3:** Spatial-Angular dictionary atom example learned jointly from phantom HARDI data with the proposed method. We can see that we have the ability to model fiber tracts with very few atoms.

For the spatial dictionaries in Figure 1, we notice clear similarities between our method and the atoms produced by KSVD. In contrast, the spatial atoms produced by KDRSDL are fuzzier, lacking the clearly defined edges and geometric shapes that are evident in the phantom dataset. These shapes resemble atoms that have landed in a local minimum or saddle point, farther from the global minimum reached by our method. This trend is similar for the angular atoms in Figure 2. We can see that the results of the proposed method has greater variation in the orientations of single fiber ODFs. The most utilized atoms in KSVD are the purely isotropic atom and the noisy isotropic atoms, whereas the atoms most frequently used with the other methods are the single fiber atoms. In Figure 3 we show an example of a single spatial-angular atom learned jointly by our proposed separable dictionary learning method the resembles a fiber tract structure.

Finally, in Figure 4 we show spatial and angular dictionaries (bottom) learned from real HARDI brain data (top) for KSVD (A.) and our proposed method (B.). We notice large structures in the spatial atoms like the CSF region as well as atoms with specific spatial patterns resembling fiber structure. Each spatial-angular atom is sorted (left to right, top to bottom) by their frequency of use in the representation of the training data. For example, the top left spatial and angular atoms are together the most frequently used joint spatial-angular atom in training. (We show only a subset of unique spatial and angular atoms for visualization.)



**Figure 4:** Top: Example of real HARDI brain training data, one of the spatial DWIs (top left) and the full field of ODFs (top right). Bottom: *A*. Spatial and angular dictionaries learned independently via KSVd. Each are sorted (left to right, top to bottom) by their individual frequencies of use in modeling the training data. *B*. Spatial and angular dictionaries learned jointly by the proposed method. Each are sorted (left to right, top to bottom), by their joint frequencies. For example, the top left spatial and angular atoms are together the most frequently used joint spatial-angular atom.

Domain	Angular			Spatial-Angular			
	Type	Fixed	Fixed	Separate	Fixed	Separate	Joint
Method	I-SR	I-SR+TV	I-KSVD	Curve-SR	KSVD-KSVD	KDRSDL	Proposed
Slice 25	16.631	16.634	18.011	17.000	19.182	18.793	<b>19.501</b>
Slice 30	16.715	16.720	16.090	17.087	17.001	16.725	<b>17.221</b>
Slice 35	17.311	17.323	16.679	17.793	17.675	17.418	<b>17.868</b>

**Table 3:** Peak Signal-to-Noise Ratio (PSNR) denoising results on three different 2D HARDI phantom image slices. We compared the domains of angular vs spatial-angular sparse coding with dictionaries that are either of type fixed (purple), learned in the spatial and angular domains separately (pink), or learned in the spatial-angular domain jointly (green). Denoising using our proposed joint spatial-angular dictionary learning method with global optimality outperforms denoising with both fixed and learned dictionaries from other methods.

**6.6. Denoising Results.** The results of the denoising experiment on the phantom HARDI data are recorded in Table 3. We repeated the experiment on three slices of the phantom HARDI data that were not used for training. For each experiment, our reconstruction using the dictionaries learned jointly from our method achieved the highest PSNR values (right-most column of Table 3), outperforming both KDRSDL which learns dictionaries jointly but with potential suboptimal local minimizers, and KSVD-KSVD which learns spatial and angular dictionaries separately, as well as fixed spatial-angular dictionaries. These results provide a preliminary validation of the importance of separable dictionary learning with global optimality guarantees. Figure 5 shows the qualitative results of our denoising experiment in comparison to the denoising results of the SR fixed dictionary with a close-up in Figure 6. Then, in Figure 7 we show denoising results on real HARDI data using our proposed dictionaries with noticeable regions of improvement highlighted in red.

**7. Conclusion.** In this work, we proposed a mathematical formulation of the separable dictionary learning problem for which we are able to derive, to the best of our knowledge, the first conditions of global optimality. To this end, we have framed this problem as a tensor factorization, extending theoretical results from two-factor matrix factorization to the more complex case of three-factor tensor factorization appearing in separable dictionary learning.

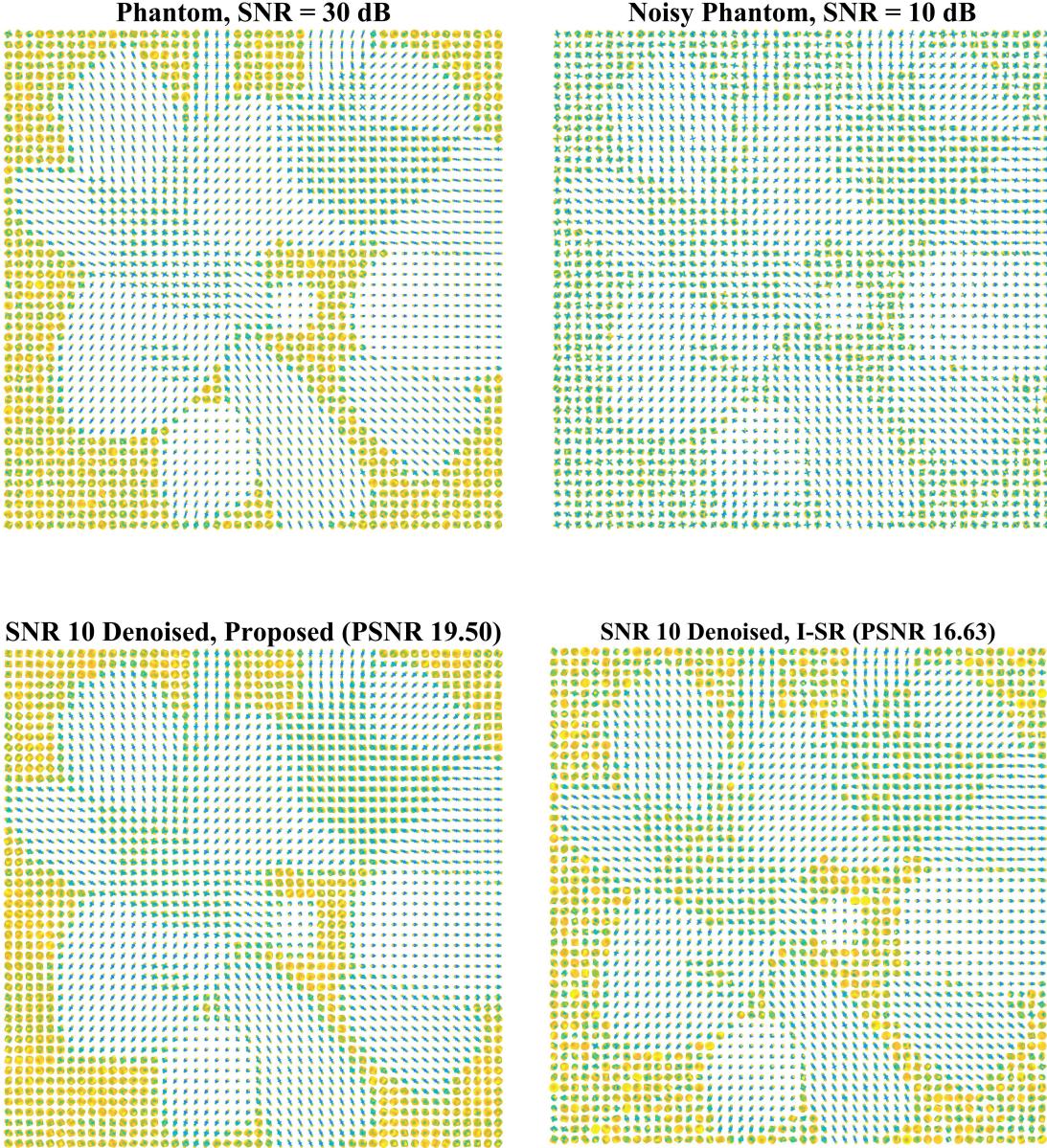
With this theoretical base, we have proposed a novel algorithm to find global minima of the separable dictionary learning problem by alternating between a local descent step to a stationary point and a check for global optimality. If the global criteria is not satisfied, the algorithm will append an additional dictionary atom and continue the descent to another stationary point. In this way, our algorithm provides a “rank-aware” methodology that could provide low-rank or overcomplete solutions, a reasonable midpoint between the low-rank solutions of KDRSDL and the overcomplete solutions of KSVD. This too depends on the initial dictionary size which may be application specific. Furthermore, the alteration of updates between each separate dictionary is flexible in our algorithm, and can be tailored to specific *a priori* knowledge of the relative dictionary sizes based on the data.

As a proof of concept, we applied the proposed algorithm to the domain of dMRI which is well suited for our framework due to the spatial-angular structure of the data. While most dictionary learning methods for dMRI restrict to learning dictionaries for the angular domain, we learn both spatial and angular dictionaries jointly in this work. We showed in a denoising task that using spatial and angular dictionaries learned jointly, outperforms dMRI denoising algorithms relying on angular dictionaries alone. Furthermore, we validated that joint learning provides better reconstructions than the alternative of learning spatial and angular dictionaries independently by simpler methods such as KSVD. Finally, our results indicate that having a globally optimal solution also outperforms methods like KDRSDL that may be subject to convergence toward local minima.

In future work we will aim to extend the theory and algorithms presented in this paper to incorporate convolutional methods that will relate local patch dictionaries to the global image for the task of global sparse coding and compressed sensing in diffusion MRI and other applications.

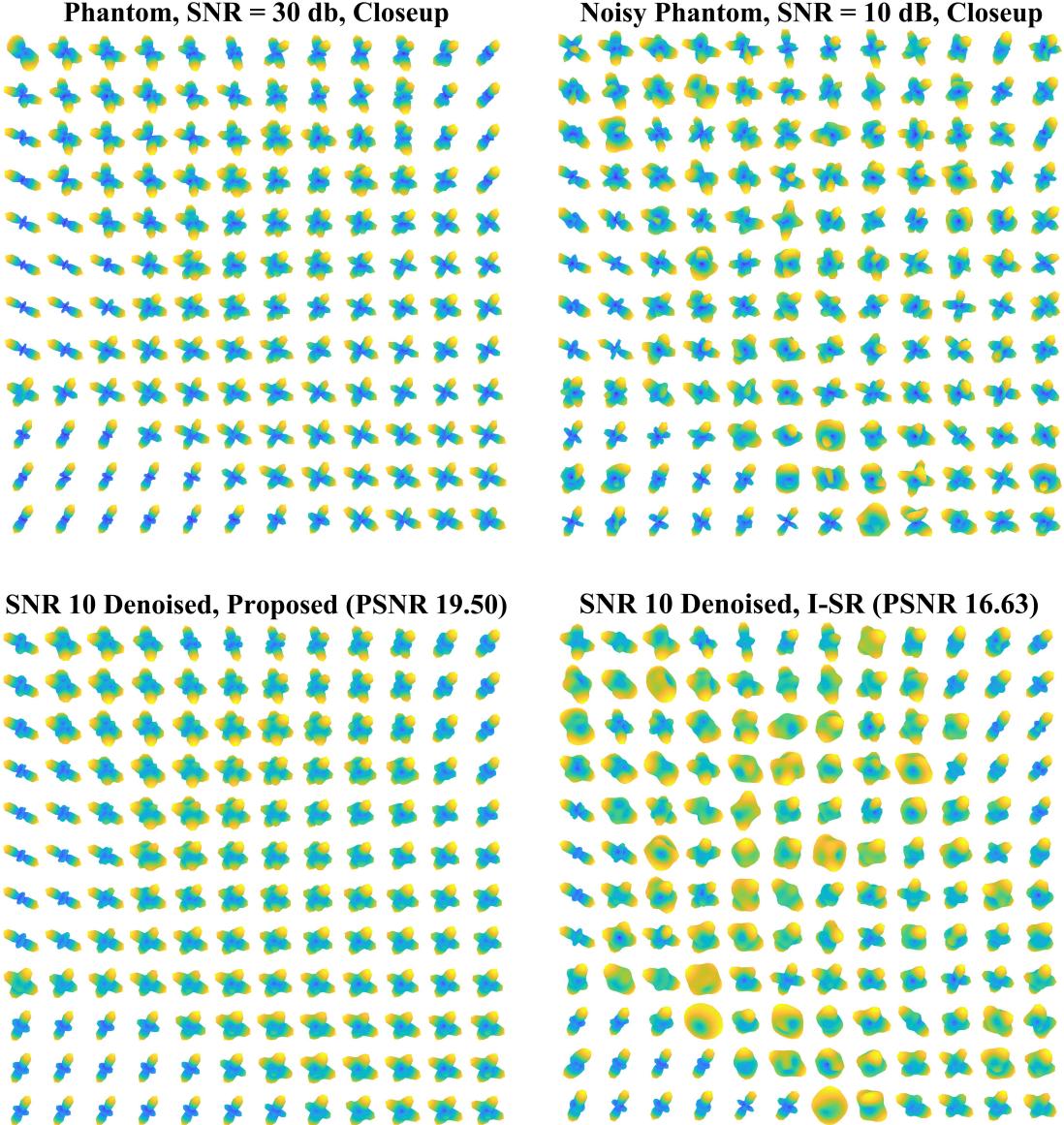
## REFERENCES

- [1] M. Aharon, M. Elad, and A. M. Bruckstein. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.



**Figure 5:** Results of HARDI phantom denoising experiment. Top left: Original Phantom data with SNR=30 dB. Top right: Noisy version with SNR=10 dB. Bottom left: Denoised reconstruction of noisy phantom using our learned spatial-angular dictionaries with spatial-angular sparse coding. Bottom right: Denoised reconstruction of noisy phantom using a fixed spherical ridgelet dictionary with angular sparse coding (I-SR). We notice our proposed method produces a more accurate reconstruction in comparison to the original SNR=30 dB. For more detailed visualization see the close-ups in Figure 6.

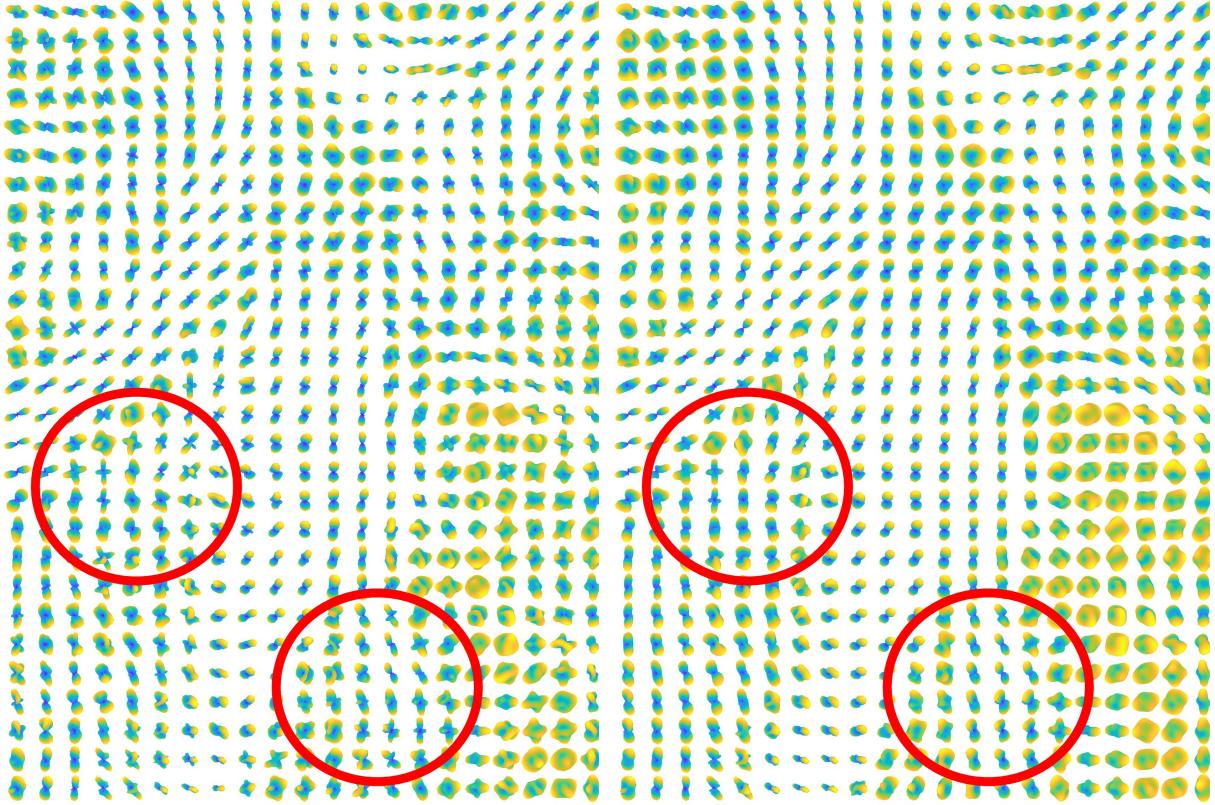
- [2] R. Aranda, A. Ramirez-Manzanares, and M. Rivera. Sparse and adaptive diffusion dictionary (SADD) for recovering intra-voxel white matter structure. *Medical Image Analysis*, 26(1):243–255, 2015.
- [3] F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. <http://arxiv.org/abs/0812.1869>, 2008.
- [4] Mehdi Bahri, Yannis Panagakis, and Stefanos Zafeiriou. Robust Kronecker-decomposable component analysis for low-rank modeling. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [5] Lijun Bao, Wanyu Liu, Yuemin Zhu, Zhaobang Pu, and Isabelle E Magnin. Sparse representation based MRI denoising with total variation. In *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, pages 2154–2157. IEEE, 2008.



**Figure 6:** Close-ups of HARDI phantom denoising results from Figure 5. The reconstruction of the noisy SNR=10 dB HARDI phantom (top right) using our proposed spatial-angular dictionary (bottom left) produces a more accurate denoised reconstruction in comparison to the original phantom with SNR=30 dB (top left), than for the fixed spherical ridgelet (SR) dictionary (bottom right).

- [6] B. Bilgic, K. Setsompop, J. Cohen-Adad, A. Yendiki, L. L Wald, and E. Adalsteinsson. Accelerated diffusion spectrum imaging with compressed sensing using adaptive dictionaries. *Magnetic Resonance in Medicine*, 68(6):1747–1754, 2012.
- [7] J-F Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- [8] E. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.
- [9] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.
- [10] J. Cheng, T. Jiang, R. Deriche, D. Shen, and P.-T. Yap. Regularized spherical polar Fourier diffusion MRI with optimal dictionary learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 639–646. Springer, 2013.

## Noisy HARDI Brain Data      Denoised HARDI, Proposed



**Figure 7:** Denoising Real HARDI brain data. Left: Original noisy HARDI brain region. Right: Denoised reconstruction using our learned spatial-angular dictionaries within our spatial-angular sparse coding.

- [11] J. Cheng, D. Shen, P.-T. Yap, and P. J. Basser. Tensorial spherical polar Fourier diffusion MRI with optimal dictionary learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2015.
- [12] Cássio Fraga Dantas, Michele N da Costa, and Renato da Rocha Lopes. Learning dictionaries as a sum of Kronecker products. *IEEE Signal Processing Letters*, 24(5):559–563, 2017.
- [13] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.
- [14] M. Descoteaux, R. Deriche, T. Knoesche, and A. Anwander. Deterministic and probabilistic tractography based on complex fiber orientation distributions. *IEEE Transactions on Medical Imaging*, 28(2):269–286, Feb. 2009.
- [15] G. Duan, H. Wang, Z. Liu, J. Deng, and Y.-W. Chen. K-CPD: Learning of overcomplete dictionaries for tensor sparse coding. In *21st International Conference on Pattern Recognition (ICPR)*, pages 493–496. IEEE, 2012.
- [16] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.
- [17] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5:2443–2446, 1999.
- [18] Mohsen Ghassemi, Zahra Shakeri, Anand D Sarwate, and Waheed U Bajwa. STARK: Structured dictionary learning through rank-one tensor recovery. *arXiv preprint arXiv:1711.04887*, 2017.
- [19] A. Gramfort, C. Poupon, and M. Descoteaux. Denoising and fast diffusion imaging with physically constrained sparse dictionary learning. *Medical Image Analysis*, 18(1):36–49, 2014.
- [20] K. Gupta, D. Adlakha, V. Agarwal, and S. P. Awate. Regularized dictionary learning with robust sparsity fitting for compressed sensing multishell HARDI. In *Computational Diffusion MRI: MICCAI Workshop*, pages 35–48. Springer, 2016.
- [21] K. Gupta and S. P. Awate. Bayesian dictionary learning and undersampled multishell HARDI reconstruction. In *Information Processing in Medical Imaging*, pages 453–465. Springer, 2017.
- [22] B. Haeffele and R. Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv*, abs/1506.07540, 2015.
- [23] B. Haeffele, E. Young, and R. Vidal. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International Conference on Machine Learning*, pages 2007–2015, 2014.

- [24] Benjamin D Haeffele and René Vidal. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *arXiv preprint arXiv:1708.07850*, 2017.
- [25] S. Hawe, M. Seibert, and M. Kleinsteuber. Separable dictionary learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 438–445, 2013.
- [26] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, and T.J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.
- [27] M. Lustig, D. Donoho, and J.M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.
- [28] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, 2010.
- [29] D. McClymont, I. Teh, H. J. Whittington, V. Grau, and J. E. Schneider. Prospective acceleration of diffusion tensor imaging with compressed sensing using adaptive dictionaries. *Magnetic Resonance in Medicine*, 2015.
- [30] S. Merlet, E. Caruyer, and R. Deriche. Parametric dictionary learning for modeling EAP and ODF in diffusion MRI. In *Medical Image Computing and Computer Assisted Intervention*, pages 10–17. Springer, 2012.
- [31] S. Merlet, E. Caruyer, A. Ghosh, and R. Deriche. A computational diffusion MRI and parametric dictionary learning framework for modeling the diffusion signal and its features. *Medical Image Analysis*, 17(7):830–843, 2013.
- [32] O. Michailovich and Y. Rathi. Fast and accurate reconstruction of HARDI data using compressed sensing. In *Medical Image Computing and Computer Assisted Intervention*, pages 607–614. Springer, 2010.
- [33] O. Michailovich and Y. Rathi. On approximation of orientation distributions by means of spherical ridgelets. *IEEE Transactions on Image Processing*, 19(2):461–477, 2010.
- [34] O. Michailovich, Y. Rathi, and S. Dolui. Spatially regularized compressed sensing for high angular resolution diffusion imaging. *IEEE Transactions on Medical Imaging*, 30(5):1100–1115, 2011.
- [35] N. Parikh and S. Boyd. Proximal Algorithms. *Found. Trends Optim.*, 1(3):127–239, 2014.
- [36] P. K. Pisharady, S. N. Sotiropoulos, J. M. Duarte-Carvajalino, G. Sapiro, and C. Lenglet. Estimation of white matter fiber parameters from compressed multiresolution diffusion MRI using sparse Bayesian learning. *NeuroImage*, 2017.
- [37] N. Qi, Y. Shi, X. Sun, and B. Yin. TenSR: Multi-dimensional tensor sparse representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5916–5925, June 2016.
- [38] F. Roemer, G. Del Galdo, and M. Haardt. Tensor-based algorithms for learning multidimensional separable dictionaries. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3963–3967. IEEE, 2014.
- [39] E. Schwab, R. Vidal, and N. Charon. Spatial-Angular Sparse Coding for HARDI. In *Medical Image Computing and Computer Assisted Intervention*, pages 475–483. Springer, 2016.
- [40] E. Schwab, R. Vidal, and N. Charon. ( $k, q$ )-Compressed Sensing for dMRI with Joint Spatial-Angular Sparsity Prior. In *MICCAI Workshop on Computational Diffusion MRI*, 2017.
- [41] E. Schwab, R. Vidal, and N. Charon. Joint spatial-angular sparse coding for dMRI with separable dictionaries. *Medical Image Analysis*, 48:25–42, 2018.
- [42] S. St-Jean, P. Coupé, and M. Descoteaux. Non local spatial and angular matching: Enabling higher spatial resolution diffusion MRI datasets through adaptive denoising. *Medical image analysis*, 32:115–130, 2016.
- [43] A. Stevens, Y. Pu, Y. Sun, G. Spell, and L. Carin. Tensor-dictionary learning with deep Kruskal-factor analysis. In *Artificial Intelligence and Statistics*, pages 121–129, 2017.
- [44] J. Sun, Y. Xie, W. Ye, J. Ho, A. Entezari, S. J. Blackband, and B. C. Vemuri. Dictionary learning on the manifold of square root densities and application to reconstruction of diffusion propagator fields. In *International Conference on Information Processing in Medical Imaging*, pages 619–631. Springer, 2013.
- [45] A. Tristán-Vega and C.-F. Westin. Probabilistic ODF estimation from reduced HARDI data with sparse regularization. In *Medical Image Computing and Computer Assisted Intervention*, pages 182–190. Springer, 2011.
- [46] D.S. Tuch, T.G. Reese, M.R. Wiegell, and V.J. Wedeen. Diffusion MRI of complex neural architecture. *Neuron*, 40:885–895, December 2003.
- [47] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, (To appear) 2010.
- [48] W. Ye, B. C. Vemuri, and A. Entezari. An over-complete dictionary based regularized reconstruction of a field of ensemble average propagators. In *IEEE International Symposium on Biomedical Imaging*, pages 940–943. Springer, 2012.
- [49] F. Yellin, B. Haeffele, and R. Vidal. Blood cell detection and counting in holographic lens-free imaging by convolutional sparse dictionary learning and coding. In *IEEE International Symposium on Biomedical Imaging*, pages 650–653, 2017.
- [50] F. Zhang, Y. Cen, R. Zhao, and H. Wang. Improved separable dictionary learning. In *IEEE 13th International Conference on Signal Processing (ICSP)*, pages 884–889. IEEE, 2016.
- [51] F. Zhang, Y. Cen, R. Zhao, H. Wang, Y. Cen, L. Cui, and S. Hu. Analytic separable dictionary learning based on oblique manifold. *Neurocomputing*, 236:32–38, 2017.
- [52] S. Zubair and W. Wang. Tensor dictionary learning with sparse tucker decomposition. In *IEEE 18th International Conference on Digital Signal Processing (DSP)*, pages 1–6, 2013.

## Appendix A.

*Proof of Proposition 3.3.* We will assume, in a first phase, that the infimum in (3.3) can be achieved for finite  $r_1$  and  $r_2$  (which is proved in the last point below) and drop the minimization in  $r_1$  and  $r_2$  to lighten the derivations.

- First, since  $\theta(\gamma, \psi, c) \geq 0 \forall (\gamma, \psi, c)$ , we have that  $\Omega_\theta(\underline{X}) \geq 0 \forall \underline{X}$ . Then, the infimum  $\Omega_\theta(0) = 0$  can be achieved by taking  $(\Gamma, \Psi, \underline{C}) = (0, 0, 0)$ . If  $\underline{X} = \underline{C} \times_1 \Gamma \times_2 \Psi$  and  $\underline{X} \neq 0$ , we can write equivalently  $\underline{X} = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \Gamma_i \otimes \Psi_j \otimes C_{i,j}$  and there exists  $i_0, j_0$  such that  $\Gamma_{i_0} \neq 0, \Psi_{j_0} \neq 0, C_{i_0, j_0} \neq 0$  and thus

$\Omega_\theta(\underline{X}) \geq \theta(\Gamma_{i_0}, \Psi_{j_0}, C_{i_0, j_0}) > 0$  thanks to the second property in Proposition ??.

2. With the substitution  $(\bar{\Gamma}, \bar{\Psi}, \bar{C}) := (\alpha^{-1/3}\Gamma, \alpha^{-1/3}\Psi, \alpha^{-1/3}\underline{C})$  and using the positive homogeneity of  $\theta$ ,

$$\begin{aligned} \Omega_\theta(\alpha\underline{X}) &= \inf_{\Gamma, \Psi, \underline{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = \alpha\underline{X} \\ &= \inf_{\bar{\Gamma}, \bar{\Psi}, \bar{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } (\alpha^{-1/3})\underline{C} \times_1 (\alpha^{-1/3})\Gamma \times_2 (\alpha^{-1/3})\Psi = \underline{X} \\ &= \inf_{\bar{\Gamma}, \bar{\Psi}, \bar{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\alpha^{1/3}\bar{\Gamma}_i, \alpha^{1/3}\bar{\Psi}_j, \alpha^{1/3}\bar{C}_{i,j}) \text{ s.t. } \bar{C} \times_1 \bar{\Gamma} \times_2 \bar{\Psi} = \underline{X} \\ &= \inf_{\bar{\Gamma}, \bar{\Psi}, \bar{C}} \alpha \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\bar{\Gamma}_i, \bar{\Psi}_j, \bar{C}_{i,j}) \text{ s.t. } \bar{C} \times_1 \bar{\Gamma} \times_2 \bar{\Psi} = \underline{X} \\ &= \alpha\Omega_\theta(\underline{X}). \end{aligned}$$

3. Let  $\underline{X} = \underline{C}^X \times_1 \Gamma^X \times_2 \Psi^X$  and  $\underline{Y} = \underline{C}^Y \times_1 \Gamma^Y \times_2 \Psi^Y$  be two  $\epsilon$ -optimal factorizations, i.e. such that  $\sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i^X, \Psi_j^X, C_{i,j}^X) \leq \Omega_\theta(\underline{X}) + \epsilon$  and a similar expression for  $\underline{Y}$ . Now we construct  $\Gamma = [\Gamma^X, \Gamma^Y]$ ,  $\Psi = [\Psi^X, \Psi^Y]$ , and  $\underline{C}$  such that for all  $t = 1, \dots, T$ ,  $C_t = \begin{bmatrix} C_t^X & 0 \\ 0 & C_t^Y \end{bmatrix}$ . Then  $\underline{X} + \underline{Y} = \underline{C} \times_1 \Gamma \times_2 \Psi$  and:

$$\begin{aligned} \Omega_\theta(\underline{X} + \underline{Y}) &\leq \sum_{i=1}^{r_1^X+r_1^Y} \sum_{j=1}^{r_2^X+r_2^Y} \theta(\Gamma_i, \Psi_j, C_{i,j}) \\ &= \sum_{i=1}^{r_1^X} \sum_{j=1}^{r_2^X} \theta(\Gamma_i^X, \Psi_j^X, C_{i,j}^X) + \sum_{i=1}^{r_1^Y} \sum_{j=1}^{r_2^Y} \theta(\Gamma_i^Y, \Psi_j^Y, C_{i,j}^Y) \\ &\leq \Omega_\theta(\underline{X}) + \Omega_\theta(\underline{Y}) + 2\epsilon \end{aligned}$$

where the second line equality results from the fact that  $\theta(\Gamma_i^X, \Psi_j^Y, C_{i,j}) = \theta(\Gamma_i^X, \Psi_j^Y, 0) = 0$  and similarly  $\theta(\Gamma_i^Y, \Psi_j^X, C_{i,j}) = 0$ . Taking  $\epsilon \rightarrow 0$  completes the proof of the triangle inequality.

4. Assuming  $\theta(-\gamma, \psi, c) = \theta(\gamma, \psi, c)$ , (as is true for  $-\psi$  or  $-c$ ) and setting  $\bar{\Gamma} := -\Gamma$ ,

$$\begin{aligned} \Omega_\theta(-\underline{X}) &= \inf_{\Gamma, \Psi, \underline{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 \Gamma \times_2 \Psi = -\underline{X} \\ &= \inf_{\Gamma, \Psi, \underline{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j}) \text{ s.t. } \underline{C} \times_1 -\Gamma \times_2 \Psi = \underline{X} \\ &= \inf_{\bar{\Gamma}, \bar{\Psi}, \bar{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(-\bar{\Gamma}_i, \bar{\Psi}_j, \bar{C}_{i,j}) \text{ s.t. } \bar{C} \times_1 \bar{\Gamma} \times_2 \bar{\Psi} = \underline{X} \\ &= \inf_{\bar{\Gamma}, \bar{\Psi}, \bar{C}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\bar{\Gamma}_i, \bar{\Psi}_j, \bar{C}_{i,j}) \text{ s.t. } \bar{C} \times_1 \bar{\Gamma} \times_2 \bar{\Psi} = \underline{X} \\ &= \Omega_\theta(\underline{X}). \end{aligned}$$

5. In order to show that there exists a global minimum with finite  $r_1$  and  $r_2$  in the definition of  $\Omega_\theta$ , we start by introducing the function defined by:

$$(7.1) \quad \tilde{\Omega}_\theta(\underline{X}) := \inf_{r \in \mathbb{N}_+} \inf_{\substack{\Gamma \in \mathbb{R}^{G \times r} \\ \Psi \in \mathbb{R}^{V \times r} \\ \Lambda \in \mathbb{R}^{T \times r}}} \sum_{i=1}^r \theta(\Gamma_i, \Psi_i, \Lambda_i) \text{ s.t. } \sum_{i=1}^r \Gamma_i \otimes \Psi_i \otimes \Lambda_i = \underline{X}.$$

where  $\Gamma_i, \Psi_i, \Lambda_i$  denote the i-th column of the respective matrices. This essentially corresponds to the same definition as  $\Omega_\theta$  but with the additional constraints that  $r_1 = r_2 = r$  and that  $\underline{C}$  is a slice by slice diagonal tensor. In fact, it turns out that the two polar functions are equal, i.e.  $\Omega_\theta(\underline{X}) = \tilde{\Omega}_\theta(\underline{X})$  for all  $\underline{X} \in \mathbb{R}^{G \times V \times T}$ , as we show below.

First, we have  $\Omega_\theta(\underline{X}) \leq \tilde{\Omega}_\theta(\underline{X})$ . Indeed, if  $\Gamma \in \mathbb{R}^{G \times r}$ ,  $\Psi \in \mathbb{R}^{V \times r}$ ,  $\Lambda \in \mathbb{R}^{r \times T}$  are such that  $\sum_{i=1}^r \Gamma_i \otimes \Psi_i \otimes \Lambda_i = \underline{X}$ , we can define the tensor  $\underline{C} \in \mathbb{R}^{r \times r \times T}$  with, for any  $t = 1, \dots, T$ ,

$$C_t = \begin{pmatrix} \Lambda_{t,1} & 0 & \cdots & 0 \\ 0 & \Lambda_{t,2} & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \Lambda_{t,r} \end{pmatrix}.$$

Then, we can see that for all  $t = 1, \dots, T$

$$\begin{aligned} (\underline{C} \times_1 \Gamma \times_2 \Psi)_t &= \Gamma C_t \Psi^T \\ &= \sum_{i=1}^r \Lambda_{t,i} \Gamma_i \Psi_i^T \\ &= \left( \sum_{i=1}^r \Gamma_i \otimes \Psi_i \otimes \Lambda_i \right)_t \end{aligned}$$

and therefore  $\underline{C} \times_1 \Gamma \times_2 \Psi = \underline{X}$ . Furthermore  $\sum_{i=1}^r \sum_{j=1}^r \theta(\Gamma_i, \Psi_j, C_{i,j}) = \sum_{i=1}^r \theta(\Gamma_i, \Psi_i, \Lambda_i)$  due to the fact that, by definition,  $C_{i,j} = 0$  for  $i \neq j$ . The inequality follows from the definition of  $\Omega_\theta$ .

Conversely, we show that  $\Omega_\theta(\underline{X}) \geq \tilde{\Omega}_\theta(\underline{X})$ . Let  $\Gamma \in \mathbb{R}^{G \times r_1}$ ,  $\Psi \in \mathbb{R}^{V \times r_2}$  and  $\underline{C} \in \mathbb{R}^{r_1 \times r_2 \times T}$  such that  $\underline{C} \times_1 \Gamma \times_2 \Psi = \underline{X}$ . Define  $r = r_1 r_2$  and the lexicographic ordering of pairs  $l : \{1, \dots, r_1\} \times \{1, \dots, r_2\} \rightarrow \{1, \dots, r\}$ . We also set  $\tilde{\Gamma} \in \mathbb{R}^{G \times r}$  such that  $\tilde{\Gamma}_{l(i,j)} = \Gamma_i$ ,  $\tilde{\Psi} \in \mathbb{R}^{V \times r}$  such that  $\tilde{\Psi}_{l(i,j)} = \Psi_j$  and  $\Lambda_{l(i,j),t} = c_{i,j,t}$ . We then obtain for all  $t = 1, \dots, T$ ,

$$\begin{aligned} X_t &= (\underline{C} \times_1 \Gamma \times_2 \Psi)_t \\ &= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} c_{i,j,t} \Gamma_i \otimes \Psi_j \\ &= \sum_{l=1}^r \Lambda_{l,t} \tilde{\Gamma}_l \otimes \tilde{\Psi}_l \\ &= \left( \sum_{l=1}^r \tilde{\Gamma}_l \otimes \tilde{\Psi}_l \otimes \Lambda_l \right)_t \end{aligned}$$

and consequently  $\underline{X} = \sum_{l=1}^r \tilde{\Gamma}_l \otimes \tilde{\Psi}_l \otimes \Lambda_l$ . Now, by construction, we also get that  $\sum_{l=1}^r \theta(\tilde{\Gamma}_l, \tilde{\Psi}_l, \Lambda_l) = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \theta(\Gamma_i, \Psi_j, C_{i,j})$ . Consequently, any value of the minimization problem (3.3) can be obtained by (7.1) thanks to the previous transformation, giving the desired inequality.

We now only need to show that a global minimum in (7.1) can be achieved with a finite  $r$ , which will give a global minimum of (3.3) with  $r_1 = r_2 = r$ . We can follow an argument similar to the one presented in [24] that we briefly recap. Let  $\Theta \subset \mathbb{R}^{G \times V \times T}$  defined by  $\Theta = \{\underline{X} : \exists(\gamma, \psi, \lambda) / \underline{X} = \gamma \otimes \psi \otimes \lambda \text{ and } \theta(\gamma, \psi, \lambda) \leq 1\}$  which is a compact subset of  $\mathbb{R}^{G \times V \times T}$  thanks to the third condition in Definition 3.1. With the same reasoning as [24], we know that  $\tilde{\Omega}_\theta$  is equivalent to the following gauge function on the convex hull of  $\Theta$ :

$$\tilde{\Omega}_\theta(\underline{X}) = \inf\{\mu : \mu \geq 0, \underline{X} \in \mu\text{conv}(\Theta)\}.$$

Now since  $\Theta$  and thus  $\text{conv}(\Theta)$  are compact sets, the previous infimum over  $\mu$  is achieved for a certain  $\mu^* \geq 0$ . Then  $\underline{X} \in \mu^*\text{conv}(\Theta)$  and from Caratheodory's theorem, we know that any point in  $\text{conv}(\Theta)$  can be written as a finite convex combination of at most  $G \times V \times T$  elements in  $\Theta$ . In other words,

there exist  $(\Gamma_i, \Psi_i, \Lambda_i)_{i=1,\dots,r}$  with  $r \leq G \times V \times T$  and  $\beta_1, \dots, \beta_r \geq 0$  with  $\beta_1 + \dots + \beta_r = 1$  such that

$$\underline{X} = \mu^* \sum_{i=1}^r \beta_i \Gamma_i \otimes \Psi_i \otimes \Lambda_i = \sum_{i=1}^r \Gamma_i^* \otimes \Psi_i^* \otimes \Lambda_i^*,$$

with  $\Gamma_i^* = \sqrt[3]{\beta_i \mu^*} \Gamma_i$ ,  $\Psi_i^* = \sqrt[3]{\beta_i \mu^*} \Psi_i$ ,  $\Lambda_i^* = \sqrt[3]{\beta_i \mu^*} \Lambda_i$  for all  $i$ . Now since  $\mu^* = \tilde{\Omega}_\theta(\underline{X})$  and by the positive homogeneity of  $\theta$ , we obtain:

$$\sum_{i=1}^r \theta(\Gamma_i^*, \Psi_i^*, \Lambda_i^*) = \mu^* \sum_{i=1}^r \beta_i \theta(\Gamma_i, \Psi_i, \Lambda_i) \leq \tilde{\Omega}_\theta(\underline{X}),$$

which implies that  $(\Gamma_i^*, \Psi_i^*, \Lambda_i^*)$  is a global minimum of (7.1) with finite  $r$ .  $\square$

## Appendix B.

**Proximal Gradient Descent with Nesterov Acceleration.** Here in Algorithm 7.1 we formalize the Proximal Gradient Descent Algorithm 5.2 with the additional process of Nesterov Acceleration to speed up the rate of convergence. We use Nesterov Acceleration within our current implementation.

---

### Algorithm 7.1 Proximal Gradient Descent with Nesterov Acceleration

---

```

Initialize:  $k = 0, \check{\Gamma}^0, \check{\Psi}^0, \check{C}^0, \lambda, r_1, r_2$ .
while error  $> \epsilon$  do
     $\Gamma_i^{k+1} = \text{prox}_{\xi_i^k \|\cdot\|_2}(\check{\Gamma}_i^k - \xi_i^k [\nabla_{\check{\Gamma}^k} \ell]_i)$ 
     $c_{i,j,t}^{k+1} = \text{prox}_{\kappa_{i,j}^k \|\cdot\|}(\check{c}_{i,j,t}^k - \kappa_{i,j}^k [\nabla_{\check{C}_t^k} \ell]_{i,j})$ 
     $\Psi_j^{k+1} = \text{prox}_{\pi_j^k \|\cdot\|_2}(\check{\Psi}_j^k - \pi_j^k [\nabla_{\check{\Psi}^k} \ell]_j)$ 
    if  $f(\Gamma^k, \Psi^k, \underline{C}^k) < f(\Gamma^{k-1}, \Psi^{k-1}, \underline{C}^{k-1})$  then
         $s_k = (1 + \sqrt{1 + 4s_{k-1}^2})/2$ 
         $\mu = (s_{k-1} - 1)/2$ 
         $\mu_\Gamma = \min(\mu, \sqrt{L_\Gamma^{k-1}/L_\Gamma^k})$ 
         $\mu_{C_t} = \min(\mu, \sqrt{L_{C_t}^{k-1}/L_{C_t}^k}) \forall t$ 
         $\mu_\Psi = \min(\mu, \sqrt{L_\Psi^{k-1}/L_\Psi^k})$ 
         $\check{\Gamma}^{k+1} = \Gamma^k + \mu_\Gamma(\Gamma^k - \Gamma^{k-1})$ 
         $\check{C}_t^{k+1} = C_t^k + \mu_{C_t}(C_t^k - C_t^{k-1})$ 
         $\check{\Psi}^{k+1} = \Psi^k + \mu_\Psi(\Psi^k - \Psi^{k-1})$ 
    else
         $s_k = s_{k-1}$ 
         $\check{\Gamma}^{k+1} = \Gamma^{k-1}$ 
         $\check{C}_t^{k+1} = C_t^{k-1} \forall t$ 
         $\check{\Psi}^{k+1} = \Psi^{k-1}$ 
    end if
     $k \rightarrow k + 1$ 
end while
return stationary point  $(\tilde{\Gamma}, \tilde{\Psi}, \tilde{C})$ 


---



```