Accepted version of the article published in IEEE Signal Processing Letters. DOI:10.1109/LSP.2024.3396660. Available at: http://ieeexplore.ieee.org

On the convergence of Block Majorization-Minimization algorithms on the Grassmann Manifold

Carlos Alejandro Lopez, *Student Member, IEEE,* and Jaume Riba, *Senior Member, IEEE*

***Abstract*—The Majorization-Minimization (MM) framework is widely used to derive efficient algorithms for specific problems that require the optimization of a cost function (which can be convex or not). It is based on a sequential optimization of a surrogate function over closed convex sets. A natural extension of this framework incorporates ideas of Block Coordinate Descent (BCD) algorithms into the MM framework, also known as block MM. The rationale behind the block extension is to partition the optimization variables into several independent blocks, to obtain a surrogate for each block, and to optimize the surrogate of each block cyclically. However, known convergence proofs of the block MM are only valid under the assumption that the constraint sets are closed and convex. Hence, the global convergence of the block MM is not ensured for non-convex sets by classical proofs, which is needed in iterative schemes that naturally emerge in a wide range of subspace-based signal processing applications. For this purpose, the aim of this letter is to review the convergence proof of the block MM and extend it for blocks constrained in the Grassmann manifold.**

***Index Terms*—Non-convex Optimization, Majorization-Minimization, Riemannian Optimization, Grassmann Manifold, Geodesically Convex Optimization, Convergence.**

I. Introduction

The Majorization-Minimization (MM) framework is an optimization paradigm that is capable to meet the requirements of emerging applications [1], which often necessitate that the algorithms are capable to handle multimodal datasets, provide integrity to the final solution, be computationally fast, and scale efficiently. This framework is based on the iterative minimization of a surrogate function (also referred to as the majorant [2]) of the original cost. This surrogate is a tight upperbound of the cost function at the current iterate and it is often selected such that the variables are separable (for parallel computing) and that it is easy to find an optimal solution of the surrogate problem [3]. The MM framework is often considered to be a generalization of other well-known algorithmic frameworks such as the Expectation-Maximization [4], [5], the Concave-Convex Procedure [6], and the Proximal algorithms [7], among others.

This work was supported by project MAYTE (PID2022-136512OB-C21 financed by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", EU), by project RODIN (PID2019-105717RB-C22/AEI/10.13039/501100011033), by the grant 2021 SGR 01033, and the fellowship 2023 FI-3 00155 by Generalitat de Catalunya and the European Social Fund.

The authors are with the Signal Processing and Communications Group (SPCOM), Dept. de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya (UPC), 08034 Barcelona, Spain (e-mail: carlos.alejandro.lopez@upc.edu; jaume.riba@upc.edu).

In relation to the separability of the optimization variables, a common approach is to combine the Block Coordinate Descent (BCD) procedure [8] with the MM framework. This approach, also termed as block MM in the literature, consists on the partition of the variables into independent blocks with the aim of employing the MM framework in each block. The motivation behind the block MM framework is that, in some cases, the original cost function can be better approximated using several blocks [3]. However, classical convergence theorems of block MM algorithms assume that each block is constrained in a closed convex set [2], [9] and thus they do not ensure the global convergence for non-convex sets.

Examples of widely used non-convex sets in optimization are the Stiefel and Grassmann manifolds [10]. Those manifolds appear in problems with orthogonality constraints, which are commonly faced in signal processing applications, such as Principal Component Analysis (PCA) [11], [12], Subspace Learning [13]–[15], Matrix Factorization [16], [17] and Non-Coherent Communications [18], to name a few.

Although the convergence of the MM framework on Stiefel (tightly related to the Grassmann manifold) constrained variables is proven in [19], no rigorous proofs exist for the block MM case. For this reason, the main goal of this letter is to revisit the convergence proof of the block MM in [9] and to extend it for blocks constrained in the Grassmann manifold. Despite the non-convexity of the Grassmanian, we show that the compactness, the geodesic convexity of subsets and the geodesic quasiconvexity [20] of the surrogate functions are sufficient conditions to ensure the convergence of a block MM algorithm involving the Grassmannian. Examples of the theoretical need for this convergence proof are found in any algorithm that combines PCA with the estimation of additional parameters or in the blind deconvolution problem [21]–[26]. Additionally, the theoretical study in this paper serves as an alternative way to proof the convergence of the Proximal Gradient method on the Grassmann manifold (see [27] for more insights).

II. Mathematical Preliminaries and Definitions

We next review some of the necessary concepts involved in the convergence proof without the necessity to delve into differential geometry. We refer to [10] for an in-depth treatment of the Grassmann manifold geometry for optimization algorithms and to [28], [29] for an overview on geodesically convex optimization.

A. Grassmann manifold definitions and concepts

1) Grassmann manifold: The Grassmann manifold, $\text{Gr}(N, D)$, is the set of D-dimensional subspaces in $\mathbb{R}^N$. We consider that each point is represented by the following equivalence class:

$$[\mathbf{X}] = \{\mathbf{X}\mathbf{R} \in \mathbb{R}^{N\times D} : \mathbf{X}^T\mathbf{X} = \mathbf{I}_D, \mathbf{R} \in \text{O}(D)\}, \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{N\times D}$ and $\text{O}(D)$ denotes the set of matrices that satisfy $\mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = \mathbf{I}_D$. We identify the entire equivalent class with a particular representative, $\mathbf{X}$, while a subindex is used to represent any related matrix that spans the same subspace, e.g. $\mathbf{X}_r = \mathbf{X}\mathbf{R}$. This approach of representing points in the Grassmann manifold is often referred to as the Orthonormal Basis (ONB) perspective [30]. In other words, we represent a point in the Grassmann manifold with a variable constrained in the Stiefel manifold [10].

2) Tangent Space of the Grassmann manifold: Let $\mathbf{X} \in \text{Gr}(N, D)$, then:

$$\mathcal{T}_{\mathbf{X}}\text{Gr}(N, D) = \{\mathbf{\Delta} \in \mathbb{R}^{N\times D} : \mathbf{X}^T\mathbf{\Delta} = \mathbf{0}\}, \quad (2)$$

denotes the tangent space at $\mathbf{X}$.

3) Principal angles between two subspaces: In $\text{Gr}(N, D)$, the principal angles between any two points $[\mathbf{X}]$ and $[\mathbf{Y}]$ are defined as the minimal angles between all possible subspace basis of those two points [31]. They can be computed efficiently using the Singular Value Decomposition (SVD) of $\mathbf{X}^T\mathbf{Y}$. Indeed, the SVD of $\mathbf{X}^T\mathbf{Y}$ can be expressed as $\mathbf{U}\cos(\mathbf{\Theta})\mathbf{V}^T$ where $\mathbf{\Theta}$ is the diagonal matrix containing the D principal angles and $\cos(\cdot)$ is the cosine function applied element-wise on the main diagonal of its input matrix. Provided that the singular values are all positive, the matrix containing the principal angles is such that $[\mathbf{\Theta}]_{d,d} \in [0, \frac{\pi}{2}]$ for all $d = 1, ..., D$ where $[\cdot]_{i,j}$ is the (i, j)-th entry of its input matrix. From the aforementioned SVD, one can obtain clean expressions using the following lemma (see [17, Proposition 1] for more details):

Lemma 1. *For any two points $[\mathbf{X}], [\mathbf{Y}] \in Gr(N, D)$, one can find two aligned representatives $\mathbf{X}_a$ and $\mathbf{Y}_a$ such that $\mathbf{X}_a^T\mathbf{Y}_a = \cos(\mathbf{\Theta})$.*

4) Grassmann geodesics: A geodesic in the Grassmann manifold is defined as the shortest path between two subspaces. In fact, it is only when the principal angles between $[\mathbf{X}]$ and $[\mathbf{Y}]$ are all (strictly) smaller than $\frac{\pi}{2}$ that the geodesic that joins them is unique [32]. Consider the direction vector $\mathbf{H} \in \mathcal{T}_{\mathbf{X}}\text{Gr}(N, D)$ pointing to $\mathbf{Y}$ and its compact SVD, $\mathbf{H} = \mathbf{U}_h\mathbf{\Theta}\mathbf{V}_h^T$, such that $\mathbf{\Theta} \preceq \frac{\pi}{2}\mathbf{I}$ where $\mathbf{\Theta}$ is the matrix containing the principal angles between $[\mathbf{X}]$ and $[\mathbf{Y}]$. Then, the expression of the geodesic that connects $\mathbf{X}$ with $\mathbf{Y}$ is:

$$\mathbf{\Gamma}(t) = \mathbf{X}\mathbf{V}_h\cos(\mathbf{\Theta}t)\mathbf{V}_h^T + \mathbf{U}_h\sin(\mathbf{\Theta}t)\mathbf{V}_h^T, \quad (3)$$

where $\sin(\cdot)$ is applied element-wise on its input matrix main diagonal and $t \in [0, 1]$ parameterizes the geodesic. Note that the previous geodesic must be such that $\mathbf{\Gamma}(0) = \mathbf{X}$, $\mathbf{\Gamma}(1) = \mathbf{Y}$ and $\frac{\text{d}\mathbf{\Gamma}(t)}{\text{d}t}\big|_{t=0} = \mathbf{H}$. Given Lemma 1, there exists an orthonormal matrix $\mathbf{\Delta}_a$ such that the geodesic connecting the aligned representatives, $\mathbf{X}_a$ and $\mathbf{Y}_a$, yields:

$$\mathbf{\Gamma}_a(t) = \mathbf{X}_a\cos(\mathbf{\Theta}t) + \mathbf{\Delta}_a\sin(\mathbf{\Theta}t), \quad (4)$$

where $\mathbf{\Delta}_a$ is such that $\mathbf{\Gamma}_a(1) = \mathbf{Y}_a$. The intuition behind (4) can be grasped from the following expression:

$$\mathbf{X}_a^T\mathbf{Y}_a = \mathbf{X}_a^T\mathbf{\Gamma}_a(1) = \cos(\mathbf{\Theta}) + \mathbf{X}_a^T\mathbf{\Delta}_a\sin(\mathbf{\Theta}), \quad (5)$$

from where $\mathbf{\Delta}_a$ must comply with the following constraint so that Lemma 1 holds:

$$\mathbf{X}_a^T\mathbf{\Delta}_a = \mathbf{0}, \quad (6)$$

which means that $\mathbf{\Delta}_a$ belongs to $\mathcal{T}_{\mathbf{X}}\text{Gr}(N, D)$.

5) Canonical distance of the Grassmann manifold: The canonical distance between any two points $[\mathbf{X}], [\mathbf{Y}] \in \text{Gr}(N, D)$, also known as the arclength, is given by the following expression:

$$d_c(\mathbf{X}, \mathbf{Y}) = ||\mathbf{\Theta}||_F, \quad (7)$$

where $||\cdot||_F$ is the Frobenius norm and $\mathbf{\Theta} \preceq \frac{\pi}{2}\mathbf{I}$ are the principal angles between $[\mathbf{X}]$ and $[\mathbf{Y}]$.

B. Geodesically convex optimization concepts

1) Geodesically convex set: A set $\mathcal{G} \subseteq \text{Gr}(N, D)$ is geodesically convex if there exists a geodesic joining any two points of $\mathcal{G}$ that lies entirely on $\mathcal{G}$. An example of a geodesically convex subset of the Grassmann manifold can be found in [17]. This concept is a particularization of totally convex sets [28].

2) Geodesic quasi-convexity: A function $f(\mathbf{X})$ is geodesically quasiconvex in the geodesically convex set $\mathcal{G} \subseteq \text{Gr}(N, D)$ with respect to the geodesic $\mathbf{\Gamma}(t)$ if [20]:

$$f(\mathbf{\Gamma}(t)) \leq \max(f(\mathbf{X}), f(\mathbf{Y})) \ \forall \mathbf{X}, \mathbf{Y} \in \mathcal{G}, \quad (8)$$

for $t \in [0, 1]$. Note that (8) is a particularization on Riemannian Manifolds of the definition of connected functions in [2]. Convexity implies quasiconvexity:

$$(1-t)f(\mathbf{X}) + tf(\mathbf{Y}) \leq \max(f(\mathbf{X}), f(\mathbf{Y})) \ \forall t \in [0, 1]. \quad (9)$$

3) Lower directional derivative: For any function $f(\mathbf{X})$, the lower directional derivative at a point $\mathbf{X}$ in the direction $\mathbf{\Delta}$ is defined as:

$$f'(\mathbf{X}; \mathbf{\Delta}) = \lim_{h\to 0} \inf \frac{f(\mathbf{X} + h\mathbf{\Delta}) - f(\mathbf{X})}{h}, \quad (10)$$

which is an alternative definition of a derivative that encompasses a wider range of functions than the classical definition of derivative. This definition is independent of whether $\mathbf{X}$ is constrained to the Grassmann manifold or not.

4) Regularity of a function: Let the partition of $\mathbf{X}$ into N blocks of variables be $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N)$ where the dimensions of the n-th block of variables are denoted by N_n for $n = 1, ..., N$. In addition, we denote $\mathbf{\Delta}_n$ as the vector that contains all zeroes except in the n-th block containing the direction $\mathbf{D}_n$ (of dimensions N_n), i.e. $\mathbf{\Delta}_n = (\mathbf{0}, ..., \mathbf{D}_n, ..., \mathbf{0})$. Then, $f(\mathbf{X})$ is regular in $\mathcal{X}$ if:

$$\begin{aligned} &f'(\mathbf{X}; \mathbf{\Delta}) \geq 0 \ \forall \mathbf{X} \in \mathcal{X}, \forall \mathbf{\Delta} \\ \text{and} \ \ &f'(\mathbf{X}; \mathbf{\Delta}_n) \geq 0 \ \forall \mathbf{X} \in \mathcal{X}, \forall \mathbf{\Delta}_n, \ n = 1, ..., N, \end{aligned} \quad (11)$$

where $\mathbf{\Delta}$ and $\mathbf{\Delta}_n$ are such that any path emanating from $\mathbf{X}$ in those directions remains on $\mathcal{X}$. In other words, every

coordinate-wise local minimum of $f(\mathbf{X})$ is also a stationary point. A simple case in which a function is regular is a differentiable function, but the regularity of a function can also be proven for some non-differentiable functions as well. We refer to [33, Lemma 3.1] for more details on this definition of regularity.

III. Block MM convergence proof for variables constrained in the Grassmann Manifold

Consider the following optimization problem:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} f(\mathbf{X}) \quad \text{s.t. } \mathbf{X} \in \mathcal{X}, \tag{12}$$

where $f(\mathbf{X})$ is any continuous function, $\mathcal{X} = \mathcal{G} \times \mathcal{C}$, $\mathcal{C}$ is any closed convex set and $\mathcal{G} \subseteq \text{Gr}(N, D)$ is any geodesically convex subset of the Grassmann manifold. The reason of choosing only two blocks, one constrained in the Grassmannian and the other being a classical convex constraint set, is to emphasize the novelty of our convergence proof (block MM with Grassmann blocks) without any loss of generality.

Given the structure of the constraint set, $\mathcal{X}$, the optimization variables are split into two independent blocks, $\mathbf{X} = (\mathbf{G}, \mathbf{c})$ where $\mathbf{G} \in \mathcal{G}$ and $\mathbf{c} \in \mathcal{C}$. Accordingly, we rewrite $f(\mathbf{X})$ as $f(\mathbf{G}, \mathbf{c})$ from now on. Note that since the optimization variable is a function of a Grassmann variable block, the cost function in (12) must satisfy the following homogeneity condition [10]:

$$f(\mathbf{G}, \mathbf{c}) = f(\mathbf{G}\mathbf{R}, \mathbf{c}), \tag{13}$$

for any $D \times D$ orthonormal matrix $\mathbf{R}$. The block MM rationale applied to (12) consists on finding the solution (global minimum or stationary point) of (12) by the successive minimization of the following problems:

$$\mathbf{G}_{i+1} = \arg\min_{\mathbf{G}} g_G(\mathbf{G}|\mathbf{G}_i, \mathbf{c}_i) \quad \mathbf{G} \in \mathcal{G}, \tag{14a}$$

$$\mathbf{c}_{i+1} = \arg\min_{\mathbf{c}} g_c(\mathbf{c}|\mathbf{G}_{i+1}, \mathbf{c}_i) \quad \mathbf{c} \in \mathcal{C}, \tag{14b}$$

where $g_G(\mathbf{G}|\mathbf{G}_i, \mathbf{c}_i)$ and $g_c(\mathbf{c}|\mathbf{G}_i, \mathbf{c}_i)$ are the majorant functions of $f(\mathbf{G}, \mathbf{c})$ for each respective block constructed using the i-th iterates. As a result of the update equations in (14), a sequence of iterates is generated, denoted as $\{\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$. We assume the following properties of the majorant functions:

(A1) The surrogates must have the same value as the original cost at the current iterate:

$$g_G(\mathbf{G}|\mathbf{G}, \mathbf{c}) = f(\mathbf{G}, \mathbf{c}) \ \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C},$$
$$g_c(\mathbf{c}|\mathbf{G}, \mathbf{c}) = f(\mathbf{G}, \mathbf{c}) \ \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C}.$$

(A2) The surrogates must majorize the original cost function:

$$g_G(\mathbf{H}|\mathbf{G}, \mathbf{c}) \geq f(\mathbf{G}, \mathbf{c}) \ \forall \mathbf{G}, \mathbf{H} \in \mathcal{G}, \forall \mathbf{c} \in \mathcal{C},$$
$$g_c(\mathbf{d}|\mathbf{G}, \mathbf{c}) \geq f(\mathbf{G}, \mathbf{c}) \ \forall \mathbf{G} \in \mathcal{G}, \forall \mathbf{c}, \mathbf{d} \in \mathcal{C}.$$

(A3) The first (lower) directional derivatives of the surrogates and of the original cost must agree:

$$g'_G(\mathbf{G}|\mathbf{G}, \mathbf{c}; \boldsymbol{\Delta}) = f'(\mathbf{G}, \mathbf{c}; \boldsymbol{\Delta}, \mathbf{0}),$$

for all tangent directions $\boldsymbol{\Delta} \in \mathcal{T}_{\mathbf{G}}\text{Gr}(N, D)$ whose resulting geodesic remains on $\mathcal{G}$ and:

$$g'_c(\mathbf{c}|\mathbf{G}, \mathbf{c}; \boldsymbol{\delta}) = f'(\mathbf{G}, \mathbf{c}; \mathbf{0}, \boldsymbol{\delta}) \quad \text{s.t. } \mathbf{c} + \boldsymbol{\delta} \in \mathcal{C}.$$

Note that $\mathbf{c} + \boldsymbol{\delta}$ is a geodesic in the Euclidean space.

(A4) $g_G(\cdot|\cdot)$ and $g_c(\cdot|\cdot)$ must be continuous on its input arguments.

(A5) $g_G(\mathbf{G}|\mathbf{G}, \mathbf{c})$ must be geodesically quasiconvex on $\mathcal{G}$ and $g_c(\mathbf{c}|\mathbf{G}, \mathbf{c})$ must be quasiconvex on $\mathcal{C}$.

Notice that (A1) and (A2) ensure that the majorants are tight upperbounds of the original cost while the remaining assumptions enforce that the majorants resemble the original cost. In fact, we generalized the assumptions presented in [9] to equivalent expressions in terms of the geodesically convex optimization paradigm. We generalize Theorem 2 in [9] for blocks of variables constrained in the Grassmann manifold in the following theorem:

Theorem 1. *Supose that a sequence is generated by (14) and that the majorant functions satisfy (A1)-(A5). In addition, assume that (14a) and (14b) have unique minimizers, that the sublevel set, i.e. $S = \{\mathbf{X} \in \mathcal{X} : f(\mathbf{X}) \leq f(\mathbf{X}_0)\}$ for some initialization $\mathbf{X}_0$, is compact and that $f(\mathbf{X})$ is regular and continuous in $\mathcal{X}$. Then, this sequence converges to a stationary point of (12).*

Proof. The rationale of the proof consists on two steps. Firstly, we prove that the sequence $\{\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$ converges to the limit point $\{\bar{\mathbf{G}}, \bar{\mathbf{c}}\}$ and, secondly, we prove that this limit point is a stationary point of the problem depicted in (12).

From assumptions (A1)-(A2) and (14), we have the following series of inequalities:

$$f(\mathbf{G}_i, \mathbf{c}_i) \underbrace{=}_{(A1)} g_G(\mathbf{G}_i|\mathbf{G}_i, \mathbf{c}_i) \underbrace{\geq}_{\text{Eq. (14a)}} g_G(\mathbf{G}_{i+1}|\mathbf{G}_i, \mathbf{c}_i) \underbrace{\geq}_{(A2)} \tag{15a}$$

$$f(\mathbf{G}_{i+1}, \mathbf{c}_i) \underbrace{=}_{(A1)} g_c(\mathbf{c}_i|\mathbf{G}_{i+1}, \mathbf{c}_i) \underbrace{\geq}_{\text{Eq. (14b)}} \tag{15b}$$

$$g_c(\mathbf{c}_{i+1}|\mathbf{G}_{i+1}, \mathbf{c}_i) \underbrace{\geq}_{(A2)} f(\mathbf{G}_{i+1}, \mathbf{c}_{i+1}), \tag{15c}$$

where each underbrace describes the reason why each equality/inequality holds. The inequalities in (15) yield $f(\mathbf{G}_i, \mathbf{c}_i) \geq f(\mathbf{G}_{i+1}, \mathbf{c}_{i+1}) \ \forall i$ which, in addition to the continuity of $f(\mathbf{G}, \mathbf{c})$ and to the compactness of the sublevel set, implies that the sequence $\{f(\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$ is non-increasing and thus has at least one limit point, $f(\bar{\mathbf{G}}, \bar{\mathbf{c}})$. Since the iterates also belong to a compact set, the generated sequence also admits some limit point, denoted as $(\bar{\mathbf{G}}, \bar{\mathbf{c}})$.

Now, we prove that $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$ converges to the aforementioned limit point. Consider a subsequence $\{\mathbf{G}_{i_k}, \mathbf{c}_{i_k}\}_{k\in\mathbb{N}}$ that converges to the limit point $(\bar{\mathbf{G}}, \bar{\mathbf{c}})$. Without loss of generality, we can assume that the Grassmann variable, $\mathbf{G}_{i_k}$, is updated infinitely often in $\{\mathbf{G}_{i_k}, \mathbf{c}_{i_k}\}_{k\in\mathbb{N}}$, so its convergence can be proved by contradiction. Let us assume that the Grassmann variable does not converge and hence there exists a positive value $\bar{\gamma}$ such that:

$$d_c(\mathbf{G}_{i_{k+1}}, \mathbf{G}_{i_k}) = \gamma_{i_{k+1}} \geq \bar{\gamma} > 0. \tag{16}$$

Also, consider that the aligned geodesic joining $\mathbf{G}_{i_{k+1}}$ and $\mathbf{G}_{i_k}$ with arclength $\gamma_{i_{k+1}}$ is given by:

$$\boldsymbol{\Gamma}_{i_{k+1}}(t) = \mathbf{G}_{a,i_k} \cos(\boldsymbol{\Theta}_{i_k} t) + \boldsymbol{\Delta}_{a,i_k} \sin(\boldsymbol{\Theta}_{i_k} t), \tag{17}$$

where $\boldsymbol{\Delta}_{a,i_k}$ and $\boldsymbol{\Theta}_{i_k}$ are such that $\boldsymbol{\Gamma}_{i_{k+1}}(1) = \mathbf{G}_{a,i_{k+1}}$. With this geodesic in mind, we obtain:

$$f(\mathbf{G}_{i_{k+1}}, \mathbf{c}_{i_{k+1}}) \underbrace{\leq}_{(A2)} g_G(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{=}_{\mathbf{G}_{i_{k+1}}=\boldsymbol{\Gamma}_{i_{k+1}}(1)} \tag{18a}$$

$$g_G(\boldsymbol{\Gamma}_{i_{k+1}}(1)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{\leq}_{\text{Eq. (14}a\text{)}} g_G(\boldsymbol{\Gamma}_{i_{k+1}}(t)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{=}_{\text{Eq. (17)}} \tag{18b}$$

$$g_G(\mathbf{G}_{a,i_k}\cos(\boldsymbol{\Theta}_{i_k}t) + \boldsymbol{\Delta}_{a,i_k}\sin(\boldsymbol{\Theta}_{i_k}t)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{\leq}_{(A5)} \tag{18c}$$

$$g_G(\mathbf{G}_{i_k}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \underbrace{=}_{(A1)} f(\mathbf{G}_{i_k}, \mathbf{c}_{i_k}), \tag{18d}$$

from where we get:

$$f(\mathbf{G}_{i_{k+1}}, \mathbf{c}_{i_{k+1}}) \leq g_G(\boldsymbol{\Gamma}_{i_{k+1}}(t)|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \leq f(\mathbf{G}_{i_k}, \mathbf{c}_{i_k}). \tag{19}$$

Considering that $\boldsymbol{\Theta}_{i_k}$ (because the principal angles belong to $[0, \frac{\pi}{2}]$) and $\boldsymbol{\Delta}_{a,i_k}$ (due to the orthonormality constraints, see (4)) belong to closed and bounded sets, the sequences that they generate have limit points $\bar{\boldsymbol{\Theta}}$ and $\bar{\boldsymbol{\Delta}}_a$, respectively. As a consequence, the geodesic defined by those limit points is denoted as $\bar{\boldsymbol{\Gamma}}(t)$ and, with some abuse of notation, can be referred to as the *limit point* geodesic. By further restricting to a subsequence that has limit points $\bar{\boldsymbol{\Theta}}$ and $\bar{\boldsymbol{\Delta}}_a$, invoking (A4) and letting $k \to \infty$, (19) yields:

$$f(\bar{\mathbf{G}}, \bar{\mathbf{c}}) \leq g_G(\bar{\boldsymbol{\Gamma}}(t)|\bar{\mathbf{G}}, \bar{\mathbf{c}}) \leq f(\bar{\mathbf{G}}, \bar{\mathbf{c}}), \tag{20}$$

which is equivalent to:

$$\begin{aligned} f(\bar{\mathbf{G}}, \bar{\mathbf{c}}) = g_G(\bar{\boldsymbol{\Gamma}}(t)|\bar{\mathbf{G}}, \bar{\mathbf{c}}) = \\ g_G(\bar{\mathbf{G}}_a\cos(\bar{\boldsymbol{\Theta}}t) + \bar{\boldsymbol{\Delta}}_a\sin(\bar{\boldsymbol{\Theta}}t)|\bar{\mathbf{G}}, \bar{\mathbf{c}}). \end{aligned} \tag{21}$$

However, (21) is contradictory with the unique minimizer assumption when the arclength of the geodesic, $\bar{\gamma}$, is different from 0. From (A2), we know that:

$$g_G(\mathbf{G}_{i_{k+1}}|\mathbf{G}_{i_{k+1}}, \mathbf{c}_{i_{k+1}}) \leq g_G(\mathbf{G}|\mathbf{G}_{i_k}, \mathbf{c}_{i_k}) \ \ \forall\ \mathbf{G} \in \mathcal{G}, \tag{22}$$

whose limit for $k \to \infty$ is:

$$g_G(\bar{\mathbf{G}}|\bar{\mathbf{G}}, \bar{\mathbf{c}}) \leq g_G(\mathbf{G}|\bar{\mathbf{G}}, \bar{\mathbf{c}}) \ \ \forall\ \mathbf{G} \in \mathcal{G}, \tag{23}$$

so $\bar{\mathbf{G}}$ is a minimizer of $g_G(\cdot|\bar{\mathbf{G}}, \bar{\mathbf{c}})$ and so is $\bar{\mathbf{G}}_a$. Since we assumed that the minimizers of the majorants are unique, this means that the limit point geodesic in (21) must remain in the same point. As a result, (21) is only true for $\bar{\boldsymbol{\Gamma}}(t) = \bar{\mathbf{G}}_a\ \forall t$ and, given that $\mathcal{G}$ is a geodesically convex subset (there is only a unique path joining two points), it means that:

$$\lim_{k\to\infty} d_c(\mathbf{G}_{i_{k+1}}, \mathbf{G}_{i_k}) = 0. \tag{24}$$

Considering that the above argument must be satisfied by every subsequence, the sequence $\{\mathbf{G}_i\}_{i\in\mathbb{N}}$ converges to $\bar{\mathbf{G}}$. Note that this limit point depends on the initialization points, $\mathbf{G}_0$ and $\mathbf{c}_0$. Likewise, the methodology that proves the convergence of the sequence $\{\mathbf{c}_i\}_{i\in\mathbb{N}}$ is shown in [9, Theorem 2] and thus the joint sequence, $\{\mathbf{G}_i, \mathbf{c}_i\}_{i\in\mathbb{N}}$, also converges.

Finally, we prove that the limit point of the Grassmann variable iterates is a stationary point of the original problem. Notice that (23) implies, after taking the lower directional derivative from both sides at $\bar{\mathbf{G}}$, that:

$$g'_G(\bar{\mathbf{G}}|\bar{\mathbf{G}}, \bar{\mathbf{c}}; \boldsymbol{\Delta}) \geq 0, \tag{25}$$

for all directions $\boldsymbol{\Delta} \in \mathcal{T}_{\bar{\mathbf{G}}}\mathrm{Gr}(N, D)$ whose respective geodesic emanating from $\bar{\mathbf{G}}$ stays in $\mathcal{G}$. Due to assumption (A3), $g'_G(\bar{\mathbf{G}}|\bar{\mathbf{G}}, \bar{\mathbf{c}}; \boldsymbol{\Delta}) = f'(\bar{\mathbf{G}}, \bar{\mathbf{c}}; \boldsymbol{\Delta}, \mathbf{0})$, and thus:

$$f'(\bar{\mathbf{G}}, \bar{\mathbf{c}}; \boldsymbol{\Delta}, \mathbf{0}) \geq 0. \tag{26}$$

In other words, $\bar{\mathbf{G}}$ is a coordinate-wise minimum of $f(\mathbf{G}, \mathbf{c})$. A similar argument is derived for the remaining convex block, $\mathbf{c}$, in [9, Theorem 2] with the given assumptions. Since $\bar{\mathbf{G}}$ and $\bar{\mathbf{c}}$ are coordinate-wise minimums, this limit point is also a stationary point of $f(\mathbf{G}, \mathbf{c})$ thanks to the regularity of $f(\mathbf{G}, \mathbf{c})$. □

Corollary 1. The optimality of the convergent points generated by the update equations in (14) depends on the cardinality of the set of stationary points of (12). If the original problem has only one stationary point, then the previous theorem ensures that the convergence point is its (unique) global optimum. Otherwise, it converges to a locally optimal point.

Example 1. An example of a block MM-like algorithm is found in the blind sparse deconvolution problem, which is based on the following optimization [25, Eq. (6)] [26, Eq. (2.6)]:

$$\min_{\mathbf{a},\mathbf{x}} ||\mathbf{y} - \mathbf{a} \circledast \mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1 \ \ \text{s.t. } \mathbf{a} \in \mathrm{Gr}(N, 1), \tag{27}$$

where $\mathbf{a}$ is the convolution kernel, $\mathbf{x}$ is the estimated signal, $\circledast$ denotes the circular convolution and λ is a regularizing parameter that is related to the sparsity of the solution. Although the approaches proposed in [25], [26] consider the Stiefel manifold, its convergence can be verified from Theorem 1 since there is a sign ambiguity [25, Section 2.1] and thus, (27) satisfies the homogeneity condition of the Grassmannian [10], implying that it can also be constrained in terms of $\mathrm{Gr}(N, 1)$. An intuitive approach is to optimize with respect to each block of variables in (27) cyclically [25], [26]. In fact, the approach proposed in [26] solves (27) by proximal gradient methods and Riemannian gradient descent algorithms for $\mathbf{x}$ and $\mathbf{a}$, respectively. Both algorithms are known to be particular instances of the MM framework [3] and, as a by-product, the convergence of this approach is certified by Theorem 1. For simulation results on this approach, see [26, Sections 5 and 6].

IV. Conclusions

In this letter, we have assessed the conditions in which a block MM algorithm converges when one of its blocks is constrained in the Grassmann manifold. Indeed, some of the classical assumptions about the majorant functions that are needed for the convergence of block MM algorithms are generalized within the geodesically convex optimization paradigm. The convergence theorem opens new geometric insights on known signal processing problems such as the one shown in Example 1. Future lines of work aim to generalize Theorem 1 to other Riemannian manifolds and to study rigorously the convergence rate of block MM algorithms of this kind.

REFERENCES

[1] J. Singh, G. Singh, and A. Verma, "The Anatomy of Big Data: Concepts, Principles and Challenges," in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2022, pp. 986–990.

[2] M. W. Jacobson and J. A. Fessler, "An Expanded Theoretical Treatment of Iteration-Dependent Majorize-Minimize Algorithms," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2411–2422, 2007.

[3] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[4] T. T. Wu and K. Lange, "The MM Alternative to EM," *Statistical Science*, vol. 25, no. 4, pp. 492 – 505, 2010. [Online]. Available: https://doi.org/10.1214/08-STS264

[5] H. Zhou and Y. Zhang, "EM vs MM: A case study," *Computational Statistics & Data Analysis*, vol. 56, no. 12, pp. 3909–3920, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947312002174

[6] A. L. Yuille and A. Rangarajan, "The Concave-Convex Procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[7] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014. [Online]. Available: https://doi.org/10.1561/2400000003

[8] S. Rabanser, L. Neumann, and M. Haltmeier, "Analysis of the Block Coordinate Descent Method for Linear Ill-Posed Problems," *SIAM Journal on Imaging Sciences*, vol. 12, no. 4, pp. 1808–1832, 2019. [Online]. Available: https://doi.org/10.1137/19M1243956

[9] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013. [Online]. Available: https://doi.org/10.1137/120891009

[10] A. Edelman, T. A. Arias, and S. Smith, "The Geometry of Algorithms with Orthogonality Constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, pp. 303–353, 1998.

[11] I. T. Jolliffe, *Principal Component Analysis and Factor Analysis*. New York, NY: Springer New York, 1986, pp. 115–128. [Online]. Available: https://doi.org/10.1007/978-1-4757-1904-8_7

[12] J. Zhan and N. Vaswani, "Robust PCA With Partial Subspace Knowledge," *IEEE Transactions on Signal Processing*, vol. 63, no. 13, pp. 3332–3347, 2015.

[13] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010, pp. 704–711.

[14] M. Vilà, C. A. López, and J. Riba, "Affine projection subspace tracking," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3705–3709.

[15] X. G. Doukopoulos and G. V. Moustakides, "Fast and stable subspace tracking," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1452–1465, 2008.

[16] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1828–1832.

[17] K. Ahn and F. Suarez, "Riemannian Perspective on Matrix Factorization," *ArXiv*, vol. abs/2102.00937, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231740389

[18] L. Zheng and D. Tse, "Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel," *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, 2002.

[19] A. Breloy, S. Kumar, Y. Sun, and D. P. Palomar, "Majorization-Minimization on the Stiefel Manifold With Application to Robust Sparse PCA," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1507–1520, 2021.

[20] T. Yamaguchi, "Locally Geodesically Quasiconvex Functions on Complete Riemannian Manifolds," *Transactions of the American Mathematical Society*, vol. 298, no. 1, pp. 307–330, 1986. [Online]. Available: http://www.jstor.org/stable/2000622

[21] A. J. Landgraf and Y. Lee, "Dimensionality reduction for binary data through the projection of natural parameters," *Journal of Multivariate Analysis*, vol. 180, p. 104668, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X20302499

[22] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *J. Mach. Learn. Res.*, vol. 14, pp. 899–925, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:938396

[23] Q. Qu, X. Li, and Z. Zhu, "A Nonconvex Approach for Exact and Efficient Multichannel Sparse Blind Deconvolution," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/02e656adee09f8394b402d9958389b7d-Paper.pdf

[24] Z. Zhu, Y. Wang, D. Robinson, D. Naiman, R. Vidal, and M. Tsakiris, "Dual Principal Component Pursuit: Improved Analysis and Efficient Algorithms," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf

[25] Y. Zhang, Y. Lau, H.-W. Kuo, S. Cheung, A. Pasupathy, and J. Wright, "On the Global Geometry of Sphere-Constrained Sparse Blind Deconvolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4381–4389.

[26] Y. Lau, Q. Qu, H. Kuo, P. Zhou, Y. Zhang, and J. Wright, "Short and sparse deconvolution - A geometric approach," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[27] S. Chen, S. Ma, A. So, and T. Zhang, "Proximal Gradient Method for Nonsmooth Optimization over the Stiefel Manifold," *SIAM Journal on Optimization*, vol. 30, pp. 210–239, 01 2020.

[28] N. K. Vishnoi, "Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity," *ArXiv*, vol. abs/1806.06373, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49300656

[29] Y. Liu, F. Shang, J. Cheng, H. Cheng, and L. Jiao, "Accelerated First-order Methods for Geodesically Convex Optimization on Riemannian Manifolds," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[30] T. Bendokat, R. Zimmermann, and P.-A. Absil, "A Grassmann Manifold Handbook: Basic Geometry and Computational Aspects," *ArXiv*, vol. abs/2011.13699, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227209168

[31] J. Zhang, G. Zhu, R. W. Heath, and K. Huang, "Grassmannian Learning: Embedding Geometry Awareness in Shallow and Deep Learning," *ArXiv*, vol. abs/1808.02229, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:51936810

[32] Y.-C. Wong, "Sectional Curvatures of Grassmann Manifolds," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 60, no. 1, pp. 75–79, 1968. [Online]. Available: http://www.jstor.org/stable/58432

[33] P. Tseng, "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, Jun. 2001. [Online]. Available: https://doi.org/10.1023/a:1017501703105