Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn

Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

I.V. Oseledets. M.V. Rakhuba, A. Uschmajew

Alternating least squares as moving subspace correction

INS Preprint No. 1712

September 2017

Alternating least squares as moving subspace correction

Ivan V. Oseledets* Maxim V. Rakhuba* André Uschmajew†

Abstract

In this note we take a new look at the local convergence of alternating optimization for low-rank matrices and tensors. Our abstract interpretation as sequential optimization on moving subspaces yields insightful reformulations of some known convergence conditions that focus on the interplay between the contractivity of classical multiplicative Schwarz methods with overlapping subspaces and the curvature of low-rank matrix and tensor manifolds. While the verification of the abstract conditions in concrete scenarios remains open in the most cases, we are able to provide an alternative and conceptually simple derivation of the convergence of the two-sided block power method of numerical algebra for computing the dominant singular subspaces of a rectangular matrix. This method is equivalent to an alternating least squares (ALS) method applied to a distance function. The theoretical results are illustrated and validated by numerical experiments.

Keywords: ALS, nonlinear Gauss-Seidel method, low-rank approximation, local convergence

1 Introduction

Consider a real-valued function $F(x)$, where $x = (\xi_1, \dots, \xi_N)$ is a tuple of vectors $\xi_i \in \mathbb{R}^{n_i}$. The *alternating optimization* (AO) or *block coordinate descent* (BCD) methods try to solve the problem

$$\min F(x) = \min F(\xi_1, \dots, \xi_N)$$

by alternating between updates of single (block) variables ξ_i while fixing all the other ξ_j, $j \neq i$:

$$\xi_i \leftarrow \operatorname*{argmin}_{\xi \in \mathbb{R}^{n_i}} F(\xi_1, \dots, \xi_{i-1}, \xi, \xi_{i+1}, \dots, \xi_N).$$

In other words, such an update is a minimization of f on the affine linear manifold $x + T_i(x)$ with the linear subspaces

$$T_i(x) = \{0\} \times \cdots \times \{0\} \times \mathbb{R}^{n_i} \times \{0\} \times \cdots \times \{0\}.$$

Such an approach is effective if optimization on the hyperplanes $x + T_i(x)$ is easy, for instance because it is of lower dimension, or because F takes a simple form on it. Obviously, the hyperplanes $x + T_i(x)$ are changing during this process, as they depend on x.

There are too many areas of application of AO to mention here. In this paper, we wish to focus on multilinear optimization. This includes low-rank matrix and tensor approximation. Here the scenario is slightly more structured. Let us explain this using the example of low-rank matrix optimization. Assume we are given a function $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ on the space of real $m \times n$ matrices,

*Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, 143026 Moscow, Russia (i.oseldets@skoltech.ru, m.rakhuba@skoltech.ru)

†Hausdorff Center for Mathematics & Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany (uschmajew@ins.uni-bonn.de)

and we wish to minimize it subject to the constraint $\operatorname{rank}(X) \leq k$. Then it is natural to use the parametrization $X = UV^T$ with $U \in \mathbb{R}^{m\times k}$, $V \in \mathbb{R}^{n\times k}$ and attempt solving

$$\min F(U,V) \coloneqq f(UV^T)$$

via AO between U and V:

$$U \leftarrow \operatorname*{argmin}_{\hat{U}\in\mathbb{R}^{m\times k}} f(\hat{U}V^T), \qquad V \leftarrow \operatorname*{argmin}_{\hat{V}\in\mathbb{R}^{n\times k}} f(U\hat{V}^T). \tag{1.1}$$

An alternative viewpoint, however, which is the starting point for the present work, is that in terms of the initial function f, the AO procedure (1.1) amounts in a sequence of optimization problems

$$X \leftarrow \operatorname*{argmin}_{X\in T_1(X)} f(X), \qquad X \leftarrow \operatorname*{argmin}_{X\in T_2(X)} f(X), \tag{1.2}$$

on *varying linear subspaces*

$$T_1(X) = \{Y \in \mathbb{R}^{m\times n} \colon \operatorname{row}(Y) \subseteq \operatorname{row}(X)\}, \tag{1.3}$$

resp.

$$T_2(X) = \{Y \in \mathbb{R}^{m\times n} \colon \operatorname{col}(Y) \subseteq \operatorname{col}(X)\}. \tag{1.4}$$

Here row and col denote the row and column space of a matrix. To be precise, one should emphasize that the update rules (1.1) and (1.2) are only equivalent as long as all constructed matrices retain full possible rank k. Also note that $X \in T_1(X)$ and $X \in T_2(X)$, hence we can formally see (1.2) as minimizations on $X + T_1(X)$ and $X + T_2(X)$ as for the BCD method.

The point we wish to make is that the formulation (1.2) is a more appropriate viewpoint on the AO method (1.1), since it is intrinsically invariant under different choices of U and V in the bilinear parametrization $X = UV^T$, which, at least formally, is highly non-unique. To see this more clearly, let us compare the two following pseudocodes:

Algorithm 1: Low-rank AO, vanilla

Input: $U_0 \in \mathbb{R}^{m\times k}$, $V_0 \in \mathbb{R}^{n\times k}$.
for $\ell = 0, 1, 2, \dots$ **do**
  $U_{\ell+1} \coloneqq \operatorname*{argmin}_{\hat{U}\in\mathbb{R}^{m\times k}} f(\hat{U}V_\ell^T)$
  $V_{\ell+1} \coloneqq \operatorname*{argmin}_{\hat{V}\in\mathbb{R}^{n\times k}} f(U_{\ell+1}\hat{V}^T)$
end

Algorithm 2: Low-rank AO with QR

Input: $U_0 \in \mathbb{R}^{m\times k}$, $V_0 \in \mathbb{R}^{n\times k}$.
for $\ell = 0, 1, 2, \dots$ **do**
  $U \leftarrow \operatorname*{argmin}_{\hat{U}\in\mathbb{R}^{m\times k}} f(\hat{U}V_\ell^T), \quad U = Q_1R_1$
  $V \leftarrow \operatorname*{argmin}_{\hat{V}\in\mathbb{R}^{n\times k}} f(Q_1\hat{V}^T), \quad V = Q_2R_2$
  $U_{\ell+1} \coloneqq Q_1R_2^{-T}, \quad V_{\ell+1} \coloneqq Q_2$
end

The right algorithm uses QR decompositions of the factors U and V in order to keep the low-rank representations stable, which is generally advised in practice. At first, Algorithm 2 appears considerably harder to analyze than Algorithm 1, which is a plain BCD method. A closer inspection, however, reveals that this is not true in the case that the solutions to the minimizations problems are unique (for instance if all matrices retain rank k and f is strictly convex), since then in both algorithms the same sequences of low-rank matrices $X_\ell = U_\ell V_\ell^T$ are constructed when starting from the same initialization. The underlying reason is that replacing U by its qr-factor Q does not change the column space, and replacing V by its qr-factor does not change the row space. Hence the subspaces of $\mathbb{R}^{m\times n}$ over which the argmin's are taken are the same in both algorithms. The superiority of the 'subspace viewpoint' compared to the 'representation

viewpoint' lies in realizing this equivalence of both algorithms, although numerically they may still behave quite differently.

The above example of low-rank matrix optimization via alternating optimization generalizes to the scenario, where we are given a multilinear map

$$\tau\colon V_1 \times \cdots \times V_d \to \mathbf{V}$$

mapping from d linear spaces $V_1, \dots, V_d$ to a space $\mathbf{V}$, and wish to optimize a function

$$F(\xi_1, \dots, \xi_d) = f(\tau(\xi_1, \dots, \xi_d)). \tag{1.5}$$

For instance the tasks of computing approximations to tensors in low-rank CP, tensor train or (hierarchical) Tucker formats are of this type; see [14, 11, 9, 4].

The aim of this paper is to subsume previous local convergence analysis of alternating optimization for multilinear optimization [14, 11] (see sec. 5 for an overview on related work) into a transparent theorem that reduces to the subspaces correction method for the linearized problem at a fixed point. Furthermore, in sec. 3 we apply our framework to derive in a new way the (known) convergence rate of a two-sided block power method for computing dominant the k-dimensional singular subspaces of a matrix, by relating this power method to alternating optimization for the distance function.

2 Abstract setup

To generalize our two motivating examples, we consider a C^1 function $f\colon \mathbf{V} \to \mathbf{V}$ on a Hilbert space $\mathbf{V}$. To every $x \in \mathbf{V}$ we attach a closed subspace $T(x)$ of $\mathbf{V}$. Further, we assume that we are given a possibly overlapping partition

$$T(x) = T_1(x) + \cdots + T_d(x)$$

into d closed subspaces $T_i(x)$. Then we define d maps

$$\mathbf{P}_i\colon \mathbf{V} \to \mathcal{L}(\mathbf{V}), \quad i = 1, \dots, d,$$

such that for every $x \in \mathbf{V}$ the linear operator $\mathbf{P}_i(x)$ is the orthogonal projection onto the space $T_i(x)$. Correspondingly, we let $\mathbf{P}(x)$ be the orthogonal projector on $T(x)$.

Next, let $\mathbf{S}_i$, $i = 1, \dots, d$, be (nonlinear) operators on $\mathbf{V}$ such that $y = \mathbf{S}_i(x)$ satisfies

$$y \in x + T_i(x), \qquad \mathbf{P}_i(x)\nabla f(y) = 0. \tag{2.1}$$

It means that $\mathbf{S}_i$ maps x to a relative critical point of f on the hyperplane $x + T_i(x)$. If, for instance, f is strictly convex and coercive, than such an operator $\mathbf{S}_i$ is uniquely defined and corresponds to minimizing f on $x + T_i(x)$.

Alternating optimization on moving hyperplanes corresponds to an iteration of the form

$$x_{\ell+1} = \mathbf{S}(x_\ell) \coloneqq (\mathbf{S}_d \circ \cdots \circ \mathbf{S}_1)(x_\ell).$$

We wish to investigate its local convergence properties under suitable differentiability assumptions. For this we a consider a fixed point

$$\bar{x} = \mathbf{S}_i(\bar{x})$$

of all $\mathbf{S}_i$ in whose neighborhood all $\mathbf{P}_i$, all $\mathbf{S}_i$ and also $\mathbf{P}$ are continuously (Fréchet) differentiable mappings. Then $\bar{x}$ is obviously a fixed point of $\mathbf{S}$.

The local contractivity around $\bar{x}$ is governed by the spectral properties of the derivatives $\mathbf{S}_i'(\bar{x})$, which are computed in the next section. Some preliminary properties of $\mathbf{S}_i'(\bar{x})$ are obtained by differentiating the equation

$$\mathbf{P}_i(x)(\mathbf{S}_i(x) - x) = \mathbf{S}_i(x) - x.$$

It gives the relation

$$\mathbf{P}_i'(x;h)(\mathbf{S}_i(x) - x) + \mathbf{P}_i(x)(\mathbf{S}_i'(x)h - h) = \mathbf{S}_i'(x)h - h \tag{2.2}$$

for all $h \in \mathbf{V}$. Here, $\mathbf{P}_i'(x;h) \in \mathcal{L}(\mathbf{V})$ denotes the application of the derivative of $\mathbf{P}_i(x)$ at x to h. Hence, in a fixed-point $\bar{x} = \mathbf{S}_i(\bar{x})$, it holds

$$(\mathbf{I} - \mathbf{P}_i(\bar{x}))\mathbf{S}_i'(\bar{x})h = (\mathbf{I} - \mathbf{P}_i(\bar{x}))h. \tag{2.3}$$

This equation is interesting as it shows the following.

Proposition 2.1. *Assume $\mathbf{P}_i$ and $\mathbf{S}_i$ are continuously differentiable around a fixed point $\bar{x}$ as considered. Then*

(i) *the subspace $T_i(\bar{x})$ is an invariant subspace of $\mathbf{S}_i'(\bar{x})$,*

(ii) *the restriction of $\mathbf{S}_i'(\bar{x})$ to the orthogonal complement $T_i(\bar{x})^\perp$ has spectral radius at least one, and equals one if and only if $T_i(\bar{x})^\perp$ is also an invariant subspace of $\mathbf{S}_i'(\bar{x})$.*

2.1 Computation of derivatives

By $\mathbf{A}(x) = \nabla^2 f(x)$ we denote the Hessian of f at x. For brevity the following shorthand notation will be used for the rest of the paper

$$\bar{\mathbf{P}}_i \coloneqq \mathbf{P}_i(\bar{x}), \qquad \bar{\mathbf{P}} \coloneqq \mathbf{P}(\bar{x}), \qquad \bar{\mathbf{A}} \coloneqq \mathbf{A}(\bar{x}), \qquad \bar{\mathbf{B}}_i \coloneqq (\bar{\mathbf{P}}_i\bar{\mathbf{A}}\bar{\mathbf{P}}_i)^{-1}.$$

To obtain a formula for $\mathbf{S}'(\bar{x})$, we differentiate each $\mathbf{S}_i$ separately. The derivatives $\mathbf{S}_i'(\bar{x})$ are given as follows.

Proposition 2.2. *Assume that $\mathbf{P}_i$ and $\mathbf{S}_i$ are continuously differentiable in a neighborhood of a fixed point $\bar{x}$, and that f is twice continuously differentiable around $\bar{x}$. If the linear operator $\bar{\mathbf{P}}_i\bar{\mathbf{A}}\bar{\mathbf{P}}_i$ is invertible on $T_i(\bar{x})$, then*

$$\mathbf{S}_i'(\bar{x})h = h - \bar{\mathbf{B}}_i\bar{\mathbf{P}}_i\bar{\mathbf{A}}h - \bar{\mathbf{B}}_i\mathbf{P}_i'(\bar{x};h)\nabla f(\bar{x}). \tag{2.4}$$

In particular,

$$\mathbf{S}_i'(\bar{x}) = \bar{\mathbf{B}}_i\mathbf{P}_i'(\bar{x};h)\nabla f(\bar{x}) \quad \text{on } T_i(\bar{x}).$$

Note that it follows from (2.2) and (2.3), that for any $h \in \mathbf{V}$ the linear operator $\mathbf{P}_i'(\bar{x};h)$ maps into the space $T_i(\bar{x})$. Hence the composition of $\bar{\mathbf{B}}_i$ with this operator is well defined.

Proof. Differentiating the equation $\mathbf{P}_i(x)\nabla f(\mathbf{S}_i(x)) = 0$ yields

$$\mathbf{P}_i'(x;h) \cdot \nabla f(x) + \mathbf{P}_i(x)\mathbf{A}(x)\mathbf{S}_i'(x)h = 0 \tag{2.5}$$

for all variations $h \in \mathbf{V}$. Splitting the term of interest $\mathbf{S}_i'(x)h$ in (2.5) into its parts on $T_i(\bar{x})$ and the orthogonal complement, we get

$$\mathbf{P}_i(x)\mathbf{A}(x)\mathbf{P}_i(x)\mathbf{S}_i'(x)h = -\mathbf{P}_i(x)\mathbf{A}(x)(\mathbf{I} - \mathbf{P}_i(x))\mathbf{S}_i'(x)h - \mathbf{P}_i'(x;h)\nabla f(x).$$

At a fixed point, we can use (2.3). Therefore

$$\bar{\mathbf{P}}_i\bar{\mathbf{A}}\bar{\mathbf{P}}_i\mathbf{S}_i'(\bar{x})h = -\bar{\mathbf{P}}_i\bar{\mathbf{A}}(\mathbf{I}-\bar{\mathbf{P}}_i)h - \mathbf{P}_i'(\bar{x};h)\cdot\nabla f(\bar{x}).$$

Assuming $\bar{\mathbf{P}}_i\bar{\mathbf{A}}\bar{\mathbf{P}}_i$ has an inverse $\bar{\mathbf{B}}_i$ on $T_i(\bar{x})$, this gives

$$\bar{\mathbf{P}}_i\mathbf{S}_i'(\bar{x})h = \bar{\mathbf{P}}_i h - \bar{\mathbf{B}}_i\bar{\mathbf{P}}_i\bar{\mathbf{A}}h - \bar{\mathbf{B}}_i\mathbf{P}_i'(\bar{x};h)\nabla f(\bar{x}).$$

Using (2.3) once more, we arrive at

$$\mathbf{S}_i'(\bar{x})h = (\mathbf{I}-\bar{\mathbf{P}}_i)\mathbf{S}_i'(\bar{x})h + \bar{\mathbf{P}}_i\mathbf{S}_i'(\bar{x})h = (\mathbf{I}-\bar{\mathbf{P}}_i)h + \bar{\mathbf{P}}_i h - \bar{\mathbf{B}}_i\bar{\mathbf{P}}_i\bar{\mathbf{A}}h - \bar{\mathbf{B}}_i\mathbf{P}_i'(\bar{x};h)\nabla f(\bar{x}),$$

which is (2.4). □

It will be useful to simplify notation. We denote

$$\bar{\mathbf{P}}_i^{\bar{\mathbf{A}}} \coloneqq \bar{\mathbf{B}}_i\bar{\mathbf{P}}_i\bar{\mathbf{A}}.$$

If $\bar{\mathbf{A}}$ is a positive definite operator, then $\bar{\mathbf{B}}_i$ is always well defined and $\bar{\mathbf{P}}_i^{\bar{\mathbf{A}}}$ allows an interpretation as the $\bar{\mathbf{A}}$-orthogonal projection onto subspace $T_i(\bar{x})$, that is, an orthogonal projection with respect to the inner product $(x,y)\mapsto\langle x,\bar{\mathbf{A}}y\rangle$.[1]

Further, we define the linear operator $\bar{\mathbf{N}}_i$ on $\mathbf{V}$ such that

$$\bar{\mathbf{N}}_i h \coloneqq \mathbf{P}_i'(\bar{x};h)\nabla f(\bar{x}) \tag{2.6}$$

for all h. With this notation, and under the assumptions of Proposition 2.2, $\mathbf{S}_i'(\bar{x})$ can be conveniently written as

$$\mathbf{S}_i'(\bar{x}) = (\mathbf{I}-\bar{\mathbf{P}}_i^{\bar{\mathbf{A}}}) - \bar{\mathbf{B}}_i\bar{\mathbf{N}}_i.$$

The formula for $\mathbf{S}'(\bar{x})$ is now obtained by the chain rule. For later reference we formulate it as a theorem.

Theorem 2.3. *Assume that all $\mathbf{P}_i$ and $\mathbf{S}_i$ are continuously differentiable in a neighborhood of a fixed point $\bar{x}$, and that f is twice continuously differentiable around $\bar{x}$. Assume all $\bar{\mathbf{B}}_i = (\bar{\mathbf{P}}_i\bar{\mathbf{A}}\bar{\mathbf{P}}_i)^{-1}$ exist on $T_i(\bar{x})$. Then*

$$\mathbf{S}'(\bar{x}) = \prod_{i=d}^{1}\mathbf{S}_i'(\bar{x}) = \prod_{i=d}^{1}[(\mathbf{I}-\bar{\mathbf{P}}_i^{\bar{\mathbf{A}}}) - \bar{\mathbf{B}}_i\bar{\mathbf{N}}_i]. \tag{2.7}$$

2.2 Curvature free cases ($\bar{\mathbf{N}}_i = 0$)

An easy case to investigate is when all $\bar{\mathbf{N}}_i = 0$, since in this case we obtain the formula

$$\mathbf{S}'(\bar{x}) = \prod_{i=d}^{1}(\mathbf{I}-\bar{\mathbf{P}}_i^{\bar{\mathbf{A}}}),$$

which is well known from the theory of subspace correction methods for a fixed space partition, specifically the *multiplicative Schwarz method*. The following statement is obtained from the standard results on multiplicative Schwarz method, e.g. [5, Theorem 3.7], by restricting to the subspace $\bar{T}(x)$ and replacing $\bar{\mathbf{A}}$ with $\bar{\mathbf{P}}\bar{\mathbf{A}}\bar{\mathbf{P}}$.

Theorem 2.4. *Assume all $\bar{\mathbf{N}}_i = 0$ and $\bar{\mathbf{A}}$ is positive definite on $T(\bar{x})$. Then $\rho(\bar{\mathbf{P}}\mathbf{S}'(\bar{x})\bar{\mathbf{P}}) < 1$. Specifically, $\|\bar{\mathbf{P}}\mathbf{S}'(\bar{x})h\|_{\bar{\mathbf{A}}} < \|h\|_{\bar{\mathbf{A}}}$ for all $h\in T(\bar{x})$, where $\|x\|_{\bar{\mathbf{A}}} = (x^T\bar{\mathbf{A}}x)^{1/2}$ is a norm on $T(\bar{x})$.*

The case $\bar{\mathbf{N}}_i = 0$ considered here arises in two notable cases.

[1] To see it observe (we omit the subscript i) that $(\bar{\mathbf{P}}^{\bar{\mathbf{A}}}x)^T\bar{\mathbf{A}}(\mathbf{I}-\bar{\mathbf{P}}^{\bar{\mathbf{A}}})x = x^T(\bar{\mathbf{A}}\bar{\mathbf{B}}\bar{\mathbf{P}}\bar{\mathbf{A}} - \bar{\mathbf{A}}\bar{\mathbf{B}}\bar{\mathbf{P}}\bar{\mathbf{A}}\bar{\mathbf{P}}\bar{\mathbf{P}}^{\bar{\mathbf{A}}})x = 0$ for all $x\in\mathbf{V}$ since $\bar{\mathbf{P}}\bar{\mathbf{A}}\bar{\mathbf{P}}\bar{\mathbf{P}}^{\bar{\mathbf{A}}} = \bar{\mathbf{B}}^{-1}\bar{\mathbf{B}}\bar{\mathbf{P}}\bar{\mathbf{A}} = \bar{\mathbf{P}}\bar{\mathbf{A}}$. Hence $\bar{\mathbf{P}}^{\bar{\mathbf{A}}}x$ is $\bar{\mathbf{A}}$-orthogonal to $(\mathbf{I}-\bar{\mathbf{P}}^{\bar{\mathbf{A}}})x$.

2.2.1 Locally constant subspaces

If the subspaces $T_i(x)$ are the same for all x in a neighborhood of $\bar{x}$, then $\mathbf{P}_i'(\bar{x}) = 0$. This case occurs in the classical BCD method discussed in the introduction, which is also known as *nonlinear block Gauss-Seidel* method, and based on a fixed, non-overlapping decomposition of the full space $T(\bar{x}) = \mathbf{V}$. Hence in this case we recover the well known fact, that the local convergence rate of the nonlinear Gauss-Seidel method equals the rate of the linear block Gauss-Seidel method with the Hessian as the system matrix; cf., e.g., [10].

2.2.2 Zero gradient

The operators $\bar{\mathbf{N}}_i$ are also zero in fixed points satisfying $\nabla f(\bar{x}) = 0$. Again, this is true in the BCD method.

Another interesting scenario for this situation is low-rank optimization where a globally critical point lies on a considered manifold of low-rank matrices or tensors. This scenario is presented in sec. 3, see in particular Lemma 3.1.

2.3 A nontrivial example including curvature ($\bar{\mathbf{N}}_i \neq 0$)

A case with $\bar{\mathbf{N}}_i \neq 0$, but allowing for considerable simplification, is obtained for $d = 2$ when $T(\bar{x}) = T_1(\bar{x}) + T_2(\bar{x})$ can be decomposed into its intersection and two other $\bar{\mathbf{A}}$-orthogonal parts. This case occurs for problems of low-rank best approximation. In these cases, f is a quadratic function with Hessian equal to identity matrix: $\bar{\mathbf{A}} = \mathbf{I}$; see sec. 3 below.

Theorem 2.5. *Additionally to the assumption of Theorem 2.3, suppose the following two conditions hold:*

(i) $\bar{\mathbf{P}}_1^{\bar{\mathbf{A}}}$ *and* $\bar{\mathbf{P}}_2^{\bar{\mathbf{A}}}$ *commute,*[2]

(ii) $\bar{\mathbf{N}}_i = 0$ *on* $T_i(\bar{x})$ *for* $i = 1, \dots, d$.

Then

$$\rho(\mathbf{S}'(\bar{x})\bar{\mathbf{P}}) = \rho(\bar{\mathbf{B}}_2\bar{\mathbf{N}}_2\bar{\mathbf{B}}_1\bar{\mathbf{N}}_1\bar{\mathbf{P}}).$$

Proof. When $\bar{\mathbf{P}}_1^{\bar{\mathbf{A}}}$ and $\bar{\mathbf{P}}_2^{\bar{\mathbf{A}}}$ commute, it is easily verified that the operator $(\mathbf{I} - \bar{\mathbf{P}}_1^{\bar{\mathbf{A}}})\bar{\mathbf{P}}$ maps to $T_2(\bar{x})$. By (2.7) and assumption (ii), it hence holds

$$\mathbf{S}'(\bar{x})\bar{\mathbf{P}} = (\mathbf{I} - \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}} + \bar{\mathbf{B}}_2\bar{\mathbf{N}}_2)\bar{\mathbf{B}}_1\bar{\mathbf{N}}_1\bar{\mathbf{P}} = (\mathbf{I} - \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}} + \bar{\mathbf{B}}_2\bar{\mathbf{N}}_2)\bar{\mathbf{P}}\bar{\mathbf{B}}_1\bar{\mathbf{N}}_1\bar{\mathbf{P}}.$$

It is a well known fact that the spectral radius of the product of two operators is invariant under the order of factors. Thus, by the above formula, the spectral radius of $\mathbf{S}'(\bar{x})\bar{\mathbf{P}}$ is the same as the spectral radius of $\bar{\mathbf{B}}_1\bar{\mathbf{N}}_1\bar{\mathbf{P}}(\mathbf{I} - \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}} + \bar{\mathbf{B}}_2\bar{\mathbf{N}}_2)\bar{\mathbf{P}} = \bar{\mathbf{B}}_1\bar{\mathbf{N}}_1\bar{\mathbf{P}}\bar{\mathbf{B}}_2\bar{\mathbf{N}}_2\bar{\mathbf{P}}$. Here we have used that $(\mathbf{I} - \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}})\bar{\mathbf{P}}$ maps to $T_1(\bar{x})$ by (i). Changing the order of factors again, we obtain the result. □

Remark 2.6. An even stronger result is obviously obtained when again $\bar{\mathbf{N}}_i = 0$ on the whole space $T(\bar{x})$ as in Sec. 2.2. Then $\mathbf{S}'(\bar{x})\bar{\mathbf{P}} = 0$ and we expect at least quadratic convergence (given sufficient smoothness of $\mathbf{S}$). This happens for instance when $\nabla f(\bar{x}) = 0$. If additionally f is quadratic, then the sequential solution of $\mathbf{P}_i(\bar{x})\nabla f(y) = 0$ on $T_i(\bar{x})$ provides a critical point on the whole space $T(\bar{x})$ after only one sweep through $i = 1, \dots, N$ (due to orthogonal residuals). Of course, the condition (i) in Theorem 2.5 is very strong when $\bar{\mathbf{A}}$ is not the identity operator, as it implies that we are given a possibly overlapping, but otherwise $\bar{\mathbf{A}}$-orthogonal splitting of the space $T(\bar{x})$.

[2] This condition is equivalent to the fact that the $\bar{\mathbf{A}}$-orthogonal projector on $T(\bar{x})$ allows the two decompositions $\bar{\mathbf{P}}^{\bar{\mathbf{A}}} = \bar{\mathbf{P}}_1^{\bar{\mathbf{A}}} + \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}} - \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}}\bar{\mathbf{P}}_1^{\bar{\mathbf{A}}}$ and $\bar{\mathbf{P}}^{\bar{\mathbf{A}}} = \bar{\mathbf{P}}_1^{\bar{\mathbf{A}}} + \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}} - \bar{\mathbf{P}}_1^{\bar{\mathbf{A}}}\bar{\mathbf{P}}_2^{\bar{\mathbf{A}}}$.

3 Alternating optimization for low-rank matrices

We return to the alternating optimization method (1.1) for solving the problem

$$\min_{\operatorname{rank}(X)\le k} f(X) \tag{3.1}$$

for a function $f\colon \mathbb{R}^{m\times n}\to\mathbb{R}$, as outlined in the introduction. We first give an overview how the abstract setup developed above looks like in this case. We then deal with the alternating least squares method for quadratic functions f, and its relation to power iterations in the case that the Hessian of the quadratic is the identity operator.

Starting from an initial guess $X_0 = U_0V_0^T$ of rank k, the method produces a sequence $X_\ell = U_\ell V_\ell^T$ of matrices of rank at most k by alternatingly minimizing the function $f(UV^T)$ with respect to U and V only. As long as the matrices U_ℓ and V_ℓ remain of rank k, this method is equivalently described as alternating optimization on the varying subspaces defined in (1.3) and (1.4).[3] Using the orthogonal projectors

$$\mathbf{P}_1(X)[Z] = ZX^+X, \qquad \mathbf{P}_2(X)[Z] = XX^+Z, \tag{3.2}$$

these subspaces can also be written as

$$\begin{aligned} T_1(X) &= \{Y\in\mathbb{R}^{m\times n}\colon Y = \mathbf{P}_1(X)[Y]\},\\ T_2(X) &= \{Y\in\mathbb{R}^{m\times n}\colon Y = \mathbf{P}_2(X)[Y]\}. \end{aligned}$$

We recall that the Moore-Penrose inverse of $X\in\mathbb{R}^{m\times n}$ is defined as $X^+ = V\Sigma^{-1}U^T\in\mathbb{R}^{n\times m}$ where $X = U\Sigma V^T$ is a singular value decomposition of X. Observe that in difference to previous notation in $\mathbb{R}^n$ we now use square brackets in $\mathbf{P}(X)[Z]$ to describe the linear action of $\mathbf{P}(X)$ on Z in order to avoid confusion with matrix multiplication.

It is obvious that $\mathbf{P}_i(X)$ are projections whose ranges are the subspaces $T_i(X)$ as defined in (1.3). We also see that XX^+ and X^+X are themselves orthogonal projections in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Since the Frobenius inner product of two matrices can be computed column- or row-wise, it easily follows that the projectors $\mathbf{P}_i(X)$ are orthogonal with respect to this inner product, that will be denoted by $\langle\cdot,\cdot\rangle_F$.

It is well known that for every k the set

$$\mathcal{M}_k = \{X\in\mathbb{R}^{m\times n}\colon \operatorname{rank}(X) = k\}$$

is a smooth embedded submanifold of $\mathbb{R}^{m\times n}$, and the space

$$T(X) = T_1(X) + T_2(X)$$

is the tangent space to that manifold at $X\in\mathcal{M}_k$. Therefore, if $\bar X$ has rank k and is a fixed point of a (locally) smooth map $\mathbf{S}\colon\mathbb{R}^{m\times n}\to\mathbb{R}^{m\times n}$ satisfying

$$\operatorname{rank}(\mathbf{S}(X))\le\operatorname{rank}(X) \tag{3.3}$$

for all X, then the condition $\rho(\mathbf{S}'(\bar X)\bar{\mathbf{P}}) < 1$ is sufficient for R-linear convergence

$$\limsup_{\ell\to\infty}\|X_\ell - \bar X\|^{1/\ell}\le\rho(\mathbf{S}'(\bar X)\bar{\mathbf{P}}) \tag{3.4}$$

[3] When the rank drops, some formal subtleties appear. In the alternating subspace method the rank can only decrease, but never increase again, whereas in the BCD method for U and V the size of the blocks is not changed, and even if, say, U with rank less than k is fixed, the minimizer for V is then not unique and a full rank matrix V could be selected.

(in any norm, since we are now in a finite-dimensional setting) of an iteration

$$X_{\ell+1} = \mathbf{S}(X_\ell),$$

with starting guess X_0 of rank k close enough to $\bar{X}$.[4]

From (3.2) it is obvious that $\mathbf{P}_1(X)$ and $\mathbf{P}_2(X)$ commute. Correspondingly,

$$\mathbf{P}(X) = \mathbf{P}_1(X) + \mathbf{P}_2(X) - \mathbf{P}_1(X)\mathbf{P}_2(X) = \mathbf{P}_1(X) + \mathbf{P}_2(X) - \mathbf{P}_2(X)\mathbf{P}_1(X) \tag{3.5}$$

is the orthogonal projection on $T(X)$.

3.1 Derivatives of projections

The reader will have noticed that the mappings $X \mapsto \mathbf{P}_i(X)$ as defined in (3.2) are not differentiable on $\mathbb{R}^{m\times n}$ unless $\operatorname{rank}(X) = \min(m, n)$, since the map $X \mapsto X^+$ is not. Since fixed points $\bar{X}$ of alternating optimization method have rank at most k we shall comment on how to resolve this conflict to the theory developed above. There are two possible solutions.

The first is to observe that all derivations of Sec. 2 remain valid if directional derivatives $\mathbf{P}_i'(\bar{X}; H)$ and $\mathbf{P}'(\bar{X}; H)$ exist for all $H \in T(\bar{X})$, where $\bar{X}$ is a fixed-point of the iteration. For the projections (3.2), and hence for P given by (3.5), this is the case. In fact, the Moore-Penrose pseudoinverse is a smooth map on manifolds of constant rank, and its Riemannian derivative at $X \in \mathcal{M}_k$ is given by

$$\mathrm{D}X^+[H] = -X^+HX^+ + X^+(X^+)^TH^T(I - XX^+) + (I - X^+X)H^T(X^+)^TX^+ \tag{3.6}$$

with $H \in T(X)$; see [6]. Hence, for such H, we compute from (3.2) that

$$\begin{aligned}\mathbf{P}_1'(X; H)[Z] &= ZX^+H + Z \cdot \mathrm{D}X^+[H] \cdot X \\ &= ZX^+H - ZX^+HX^+X + Z(I - X^+X)H^T(X^+)^T.\end{aligned} \tag{3.7}$$

Here we have used $(I - XX^+)X = 0$ and $(X^+)^TX^+X = (X^+)^T$. Correspondingly,

$$\begin{aligned}\mathbf{P}_2'(X; H)[Z] &= HX^+Z + X \cdot \mathrm{D}X^+[H] \cdot Z \\ &= HX^+Z - XX^+HX^+Z + (X^+)^TH^T(I - XX^+)Z,\end{aligned} \tag{3.8}$$

since $X(I - X^+X) = 0$ and $XX^+(X^+)^T = (X^+)^T$.

A second solution to ensure differentiability is to formally define $\mathbf{P}_1(X)$ resp. $\mathbf{P}_2(X)$ as the projections on the subspaces of matrices whose column resp. row spaces are contained in the subspaces spanned by the *dominant* k left resp. right singular vectors of X. These maps are smooth in the neighborhood of any rank k matrix $\bar{X}$. In fact, their derivatives along the tangent space $T(\bar{X})$ are the same as (3.7) and (3.8), respectively, so the result is the same.[5]

The formulas (3.7) and (3.8) can be considerably simplified if Z is orthogonal to the tangent space $T(X)$, since in this case $\mathbf{P}_1(X)[Z] = 0$, implying $ZX^+ = 0$, and $\mathbf{P}_2(X)[Z] = 0$, implying $X^+Z = 0$. For such Z, (3.7) and (3.8) become

$$\mathbf{P}_1'(X; H)[Z] = ZH^T(X^+)^T$$

[4] Let us prove this. If $\bar{X}$ is a fixed point, then, by continuity, $\mathbf{S}(X)$ is close to $\bar{X}$ when X is close to $\bar{X}$. Hence under the given assumptions, $\operatorname{rank}(\mathbf{S}(X)) = k$ for all X with $\operatorname{rank}(X) = k$ that are close enough to $\bar{X}$ (by semi-continuity of rank). Therefore, $\mathbf{S}$ can be locally regarded as a map between smooth submanifolds of $\mathcal{M}_k$, $\mathbf{S}'(\bar{X})$ maps the tangent space $T(\bar{X})$ into itself, and the sufficiency of the condition $\rho(\mathbf{S}'(\bar{X})) < 1$ on $T(\bar{X})$ for local contractivity follows in the same way as in linear space using differential calculus on manifolds.

[5] The directional derivatives orthogonal to the tangent space are zero, since small perturbations in this direction have orthogonal column and row spaces, and hence increase the rank, but do not change the dominant singular vectors.

and

$$\mathbf{P}_2'(X;H)[Z] = (X^+)^T H^T Z.$$

Note that since we need to derive the operators $\bar{\mathbf{N}}_i[H] = \mathbf{P}_i'(\bar{X};H)[\nabla f(\bar{X})]$ defined in (2.6) at critical points $\bar{X}$ of f on $\mathcal{M}_k$, where $\nabla f(\bar{X})$ is orthogonal to $T(\bar{X})$, this is indeed the case of interest. For reference we state this as a lemma.

Lemma 3.1. *Let $X \in \mathcal{M}_k$ be a critical point of f on $\mathcal{M}_k$, that is, $\mathbf{P}(\bar{X})\nabla f(\bar{X}) = 0$. Then for the projections* (3.2) *it holds*

$$\bar{\mathbf{N}}_1[H] \coloneqq \mathbf{P}_1'(\bar{X};H)\nabla f(\bar{X}) = \nabla f(\bar{X}) H^T (\bar{X}^+)^T$$

and

$$\bar{\mathbf{N}}_2[H] \coloneqq \mathbf{P}_2'(\bar{X};H)[\nabla f(\bar{X})] = (\bar{X}^+)^T H^T \nabla f(\bar{X})$$

for all $H \in T(\bar{X})$. In particular, $\bar{\mathbf{N}}_i = 0$ on $T_i(\bar{X})$.

Remark. Regarding the initial problem (3.1) on $\mathcal{M}_{\leq k}$, we remark that the "smoothness" assumption $\operatorname{rank}(\bar{X}) = k$, which has been crucial in the above derivations, is plausible in most applications, except for very special or artificial cases. It has been shown in [12, Corollary 3.4] that critical points $\bar{X}$ of (3.1), for example local minima, do either satisfy $\operatorname{rank}(\bar{X}) = k$ or $\nabla f(\bar{X}) = 0$.

3.2 Alternating least squares algorithm

When f is a strictly convex quadratic function, the outlined method is known as *alternating least squares* (ALS) method. Let us give formulas for this important special case in more detail.

For simplicity, we assume that $f(0) = 0$. Then f takes the form

$$f(X) = \frac{1}{2}\langle X, \mathbf{A}[X]\rangle_F - \langle X, B\rangle_F, \tag{3.9}$$

where $\mathbf{A}$ is a symmetric positive definite linear operator on $\mathbb{R}^{m\times n}$, and $B \in \mathbb{R}^{m\times n}$. We have $\nabla f(X) = \mathbf{A}[X] - B$, and the Hessian at every point is the operator $\mathbf{A}$. Minimizing the function f without constraint is equivalent to solving the linear matrix equation $\mathbf{A}[X] = B$. The ALS algorithm is used to find an approximate low-rank solution to this equation, as it tries to minimize function f subject to $\operatorname{rank}(X) \leq k$.

At a given iterate X_ℓ, the first step of ALS computes

$$\mathbf{S}_1(X_\ell) = \operatorname*{argmin}_{X \in T_1(X_\ell)} f(X).$$

Since $\mathbf{A}$ is positive definite, there is indeed a unique solution, and it is given as

$$\mathbf{S}_1(X) = (\mathbf{P}_1(X)\mathbf{A}\mathbf{P}_1(X))^{-1}[\mathbf{P}_1(X)[B]]. \tag{3.10}$$

Here, as usual, $(\mathbf{P}_1(X)\mathbf{A}\mathbf{P}_1(X))^{-1}$ is understood as the inverse of the operator $\mathbf{P}_1(X)\mathbf{A}\mathbf{P}_1(X)$ on its invariant subspace $T_1(X)$. The map $X \mapsto \mathbf{S}_1(X)$ is differentiable on the manifold $\mathcal{M}_k$ of rank-k matrices, since $\mathbf{P}_1$ is.

If $\mathbf{S}_1(X_\ell)$ has rank k,[6] then the next step of ALS computes

$$X_{\ell+1} = \mathbf{S}(X_\ell) \coloneqq \mathbf{S}_2(\mathbf{S}_1(X_\ell)) = \operatorname*{argmin}_{X \in T_2(\mathbf{S}_1(X_\ell))} f(X).$$

[6] If not, there are several options, but we ignore this case.

The solution map $\mathbf{S}_2$ is given as

$$\mathbf{S}_2(X) = (\mathbf{P}_2(X)\mathbf{A}\mathbf{P}_2(X))^{-1}[\mathbf{P}_2(X)[B]]. \tag{3.11}$$

We repeat once more that $X \in T_1(X)$ and $X \in T_2(X)$ for every X, so we are in the abstract framework developed in Sec. 2.

The original idea of AO for low-rank optimization is to operate on a (non-unique) factorization $X_\ell = U_\ell V_\ell^T$. In terms of these factors, more precisely their vectorizations, the ALS method becomes the algorithm displayed as Alg. 3, where $\mathbf{A}$ is to be understood as an $mn \times mn$ matrix. As explained in the introduction, the QR decompositions are not mandatory in theory, but highly recommended in practice for numerical stability.

Algorithm 3: Alternating least squares algorithm for (3.9)

Input: $B \in \mathbb{R}^{m\times n}$, $U_0 \in \mathbb{R}^{m\times k}$, $V_0 \in \mathbb{R}^{n\times k}$.
for $\ell = 0, 1, 2, \dots$ **do**
  $\operatorname{vec}(U) = \big((V_\ell^T \otimes I)\,\mathbf{A}\,(V_\ell \otimes I)\big)^{-1} \operatorname{vec}(BV_\ell), \quad QR = \operatorname{qr}(U)$
  $U \coloneqq Q$
  $\operatorname{vec}(V^T) = \big((I \otimes U^T)\,\mathbf{A}\,(I \otimes U)\big)^{-1} \operatorname{vec}(U^T B), \quad QR = \operatorname{qr}(V)$
  $V_{\ell+1} \coloneqq Q, \quad U_{\ell+1} \coloneqq UR^{-T}$
end

3.3 SVD block power method

As a special case, we now consider the quadratic function (3.9) with $\mathbf{A} = \mathbf{I}$. It corresponds to the task

$$\min_{\operatorname{rank}(X)\le k} \frac{1}{2}\|X - B\|_F^2 \tag{3.12}$$

of computing a best rank-k approximation to matrix B in Frobenius norm. Since $\mathbf{A} = \mathbf{I}$, we have $(\mathbf{P}_i(X)\mathbf{A}\mathbf{P}_i(X))^{-1}\mathbf{P}_i(X) = \mathbf{P}_i(X)$ for $i = 1, 2$, and hence the update formulas (3.10) and (3.11) for alternating optimization simplify to

$$\mathbf{S}_1(X) = \mathbf{P}_1(X)[B] = BX^+X, \qquad \mathbf{S}_2(X) = \mathbf{P}_2(X)[B] = XX^+B. \tag{3.13}$$

The resulting ALS iteration $X_{\ell+1} = \mathbf{S}(X_\ell)$ becomes

$$X_{\ell+1/2} = BX_\ell^+X_\ell, \qquad X_{\ell+1} = X_{\ell+1/2}X_{\ell+1/2}^+B. \tag{3.14}$$

Writing $X_\ell = U_\ell\Sigma_\ell V_\ell^T$, it is easily seen (and shown below) that the sequence generated by (3.14) is the same as in the *simultaneous orthogonal iteration*, which is a two-sided block power method for computing the dominant k left and right singular subspaces of B, displayed as Alg. 4 (provided X_0 has the row space spanned by V_0).

Algorithm 4: Simultaneous orthogonal iteration

Input: $B \in \mathbb{R}^{m\times n}$, $k \le \min(m, n)$, $V_0 \in \mathbb{R}^{n\times k}$ such that $V_0^TV_0 = I$
for $\ell = 0, 1, 2, \dots$ **do**
  $QR = \operatorname{qr}(BV_\ell)$ `// tall QR decomposition`
  $U_{\ell+1} \coloneqq Q$
  $QR = \operatorname{qr}(B^TU_{\ell+1})$ `// tall QR decomposition`
  $V_{\ell+1} \coloneqq Q, \quad S_{\ell+1} = R^T$
end

Let

$$B = \sum_{i=1}^{\min(m,n)} \sigma_i \bar{u}_i \bar{v}_i^T \tag{3.15}$$

be the singular value decomposition of B, that is, $\bar{u}_1, \bar{u}_2, \dots$ and $\bar{v}_1, \bar{v}_2, \dots$ are orthonormal systems in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. If $\sigma_k > \sigma_{k+1}$, then it can be shown that the sequence $X_\ell = U_\ell S_\ell V_\ell^T$ generated in Algorithm 4 converges to the unique best rank-k approximation

$$\bar{X} = \sum_{i=1}^{k} \sigma_i \bar{u}_i \bar{v}_i^T \tag{3.16}$$

for almost every starting guess $X_0 = U_0 V_0^T$. In fact, the method produces the same subspaces as the corresponding orthogonal iterations for the symmetric matrices $B^T B$ and BB^T, respectively, whose eigenvalues are the σ_i^2 (zero may be a further eigenvalue). Hence, by well known results, $U_\ell \to [\bar{u}_1, \dots, \bar{u}_k]$ and $V_\ell \to [\bar{v}_1, \dots, \bar{v}_k]$ in terms of subspaces for almost every initialization, with a convergence rate $O\left(\sigma_{k+1}^2/\sigma_k^2\right)$; see [3, 7]. As an application of our abstract framework we are going to derive this rate of convergence by inspecting Algorithm 3 instead.

Theorem 3.2. *Let B have singular values $\sigma_1 \geq \cdots \geq \sigma_k > \sigma_{k+1} \geq \dots$, and the unique best rank-k approximation $\bar{X}$. Then the sequence $X_{\ell+1} = \mathbf{S}(X_\ell) = \mathbf{S}_2(\mathbf{S}_1(X_\ell))$ defined via* (3.13) *resp.* (3.14) *(AO for problem* (3.12)*) is, in exact arithmetic, identical to the sequence $X_\ell = U_\ell S_\ell V_\ell^T$ generated by the simultaneous orthogonal iteration (Alg. 4). With $\bar{\mathbf{P}} = \mathbf{P}(\bar{X})$ as before, it holds*

$$\rho(\mathbf{S}'(\bar{X})\bar{\mathbf{P}}) = \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^2. \tag{3.17}$$

Consequently, by (3.4)*, the sequence (X_ℓ) converges (for close enough starting guesses) R-linearly to $\bar{X}$ at a rate*

$$\limsup_{\ell\to\infty} \|X_\ell - \bar{X}\|^{1/\ell} \leq \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^2 \tag{3.18}$$

(in any norm). The convergence of the column and row spaces can be estimated correspondingly in the sense of operator norm of projectors as

$$\limsup_{\ell\to\infty} \|\mathbf{P}_i(X_\ell) - \mathbf{P}_i(\bar{X})\|^{1/\ell} \leq \left(\frac{\sigma_{k+1}}{\sigma_k}\right)^2, \quad i = 1, 2. \tag{3.19}$$

Proof. We first show by induction that the methods are the same. If X_ℓ has the row space spanned by V_ℓ, then $X_{\ell+1/2}$ in (3.14) can be written $BV_\ell V_\ell^T$, which has the same column space as BV_ℓ. Therefore, using $U_{\ell+1}$ from Alg. 4, we get that $X_{\ell+1}$ from (3.14) equals $U_{\ell+1}U_{\ell+1}^T B = U_{\ell+1}\Sigma_{\ell+1}V_{\ell+1}^T$.

One may attempt to compute the spectral radius of $\mathbf{S}'(\bar{X})\bar{\mathbf{P}}$ from the explicit formulas (3.13) and (3.6), but it will be more elegant to invoke Theorem 2.5. Since $\bar{\mathbf{A}} = \mathbf{A} = \mathbf{I}$, the condition in item (i) of that theorem is obviously satisfied ($\mathbf{P}_1(X)$ and $\mathbf{P}_2(X)$ commute, see (3.2)). The condition (ii), that $\bar{\mathbf{N}}_i = 0$ on $T_i(\bar{X})$, is stated in Lemma 3.1. Taking further into account that $\bar{\mathbf{B}}_i$ are identities, Theorem 2.5 yields the formula

$$\rho(\mathbf{S}'(\bar{X})\bar{\mathbf{P}}) = \rho(\bar{\mathbf{N}}_2\bar{\mathbf{N}}_1\bar{\mathbf{P}}) \tag{3.20}$$

for the iteration (3.14). By Lemma 3.1,

$$\bar{\mathbf{N}}_2[\bar{\mathbf{N}}_1[H]] = (\bar{X}^+)^T\bar{X}^+H(\nabla f(\bar{X}))^T\nabla f(\bar{X}) \quad \text{for } H \in T(\bar{X}).$$

Taking further into account that $\nabla f(\bar{X}) = \bar{X} - B$, this shows that the restriction of $\bar{\mathbf{N}}_2\bar{\mathbf{N}}_1$ to the space $T(\bar{X})$ is a tensor product operator:

$$\bar{\mathbf{N}}_2\bar{\mathbf{N}}_1 = (\bar{X}^+)^T\bar{X}^+ \otimes (\bar{X} - B)^T(\bar{X} - B) \quad \text{on } T(\bar{X}). \tag{3.21}$$

By (3.15) and (3.16), the rank-one matrices $E_{ij} = \bar{u}_i\bar{v}_j^T$, $i = 1, \dots, m$, $j = 1, \dots, n$, form an orthonormal system of eigenvectors of the operator $(\bar{X}^+)^T\bar{X}^+ \otimes (\bar{X} - B)^T(\bar{X} - B)$, corresponding to eigenvalues $\lambda_{ij} = \sigma_j^2/\sigma_i^2$ for $i \le k$ and $j > k$, and $\lambda_{ij} = 0$ otherwise. Since $E_{k,k+1} = \bar{u}_k\bar{v}_{k+1}^T \in T_2(\bar{X}) \subseteq T(\bar{X})$, it follows from (3.21) that $\bar{\mathbf{N}}_2\bar{\mathbf{N}}_1\bar{\mathbf{P}}$ has the largest eigenvalue $\lambda_{k,k+1}$, which due to (3.20) proves the assertion (3.17).

Since $\operatorname{rank}(\bar{X}) = k$, it follows that $\mathbf{S}$ is a local contraction on the manifold $\mathcal{M}_k$ in the neighborhood of $\bar{X}$, and the R-linear convergence rate of $\|X_\ell - \bar{X}\|_F$ is as asserted (see the explanations for (3.4) in Sec. 3).

Let us show that (3.18) implies (3.19) fo $\mathbf{P}_1$. For all Z with $\|Z\|_F = 1$, we can estimate

$$\begin{aligned}\|(\mathbf{P}_1(X_\ell) - \mathbf{P}_1(\bar{X}))[Z]\|_F &= \|Z(X_\ell^+X_\ell - \bar{X}^+\bar{X})\|_F \\ &\le \|X_\ell^+X_\ell - \bar{X}^+\bar{X}\|_2 \\ &= \|X_\ell^+(X_\ell - \bar{X}) + (X_\ell^+ - \bar{X}^+)\bar{X}\|_2 = O(\|X_\ell - \bar{X}\|_2),\end{aligned}$$

since $X_\ell \to \bar{X}$ on $\mathcal{M}_k$ (implying that $\|X_\ell^+\|_2$ is bounded). □

4 Numerical experiments

The goal of this section is to investigate the agreement between the theoretical estimates and the numerical behaviour. Namely, we consider optimization on a set of matrices with rank bounded from above by k and a quadratic functional f of the form (3.9) using the ALS Algorithm 3. In this setting we test two cases: when the Hessian is the identity operator, that is $\mathbf{A} = \bar{\mathbf{A}} = \mathbf{I}$, and the case when the Hessian is a general symmetric positive definite (SPD) operator.

In all experiments, the initial guesses X_0 in Algorithm 3 have been chosen randomly. In the figures we depict lines corresponding to the theoretical rate of convergence $\rho(\mathbf{S}'(\bar{X})\bar{\mathbf{P}})$ by black colour, which has been computed numerically at the observed limit point $\bar{X}$ by forming (2.7) and solving a full eigenvalue problem to find the spectral radius. Note that in order to assemble $\mathbf{S}'(\bar{x}) = [(\mathbf{I} - \bar{\mathbf{P}}_2^{\bar{\mathbf{A}}}) - \bar{\mathbf{B}}_2\bar{\mathbf{N}}_2][(\mathbf{I} - \bar{\mathbf{P}}_1^{\bar{\mathbf{A}}}) - \bar{\mathbf{B}}_1\bar{\mathbf{N}}_1]$ we avoid finding its matrix representation explicitly. Instead, we utilize $\mathbf{S}'(\bar{x})$ matrix-vector multiplication (using Lemma 3.1) and successively apply it to the columns of the identity matrix.

4.1 Case $\bar{\mathbf{A}} = \mathbf{I}$

Consider the ALS method for problem (3.12), that is, minimizing the function

$$f(X) = \frac{1}{2}\|X - B\|_F^2$$

subject to $\operatorname{rank}(X) \le k$, where $B \in \mathbb{R}^{n\times n}$ is a given matrix with the predefined distribution of singular values. The goal is to find the best rank-k approximation $\bar{X}$ of B.

Specifically, we consider $n = 50$, $k = 2$, set $\sigma_2(B) = 10^{-3}$ and test with different $\sigma_3(B)$. By Theorem 3.2, the ALS method in this case is locally linearly convergent at least with the rate $(\sigma_{k+1}/\sigma_k)^2$, and in fact, this bound is sharp.[7] As illustrated in Figure 1, we observe close experimental agreement with this bound.

[7] Using the classical linear algebra approach related to spectral decomposition and power method one should see that this rate is in fact attained for almost every starting guess.

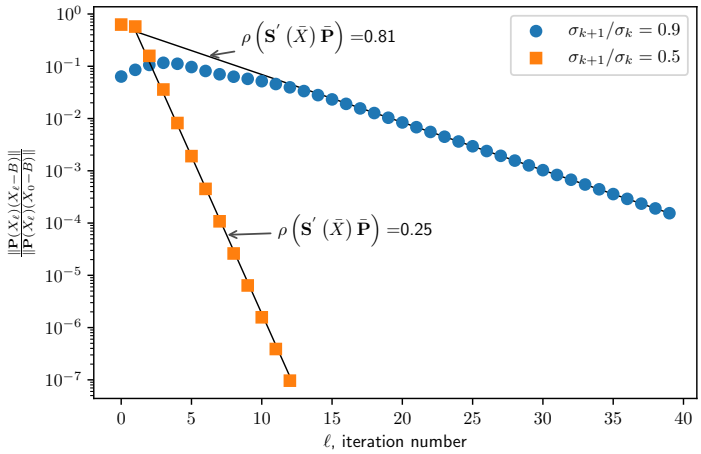


Figure 1: Relative projected residual w.r.t. the iteration number for $\bar{\mathbf{A}} = \mathbf{I}$, $k = 2$, $\sigma_k = 10^{-3}$. Black lines have slopes $\rho(\mathbf{S}'(\bar{X})\bar{\mathbf{P}}) = (\sigma_{k+1}/\sigma_k)^2$ for different σ_{k+1}, while colored dots represent the observed convergence.

Note that if $\sigma_{k+1} = 0$, then the method (generically) converges in one iteration since row and column spaces of B are found immediately. From another point of view, it holds $\nabla f(\bar{X}) = 0$ in this case, and hence $\bar{N}_i = 0$. Since $\bar{\mathbf{P}}_1$ and $\bar{P}_2$ commute, Remark 2.6 applies.

In the another extreme case, when $\sigma_{k+1} = \sigma_k$, there is an absence of convergence due to the nonuniqueness of the best rank-k approximation.

4.2 General symmetric positive definite $\bar{\mathbf{A}}$

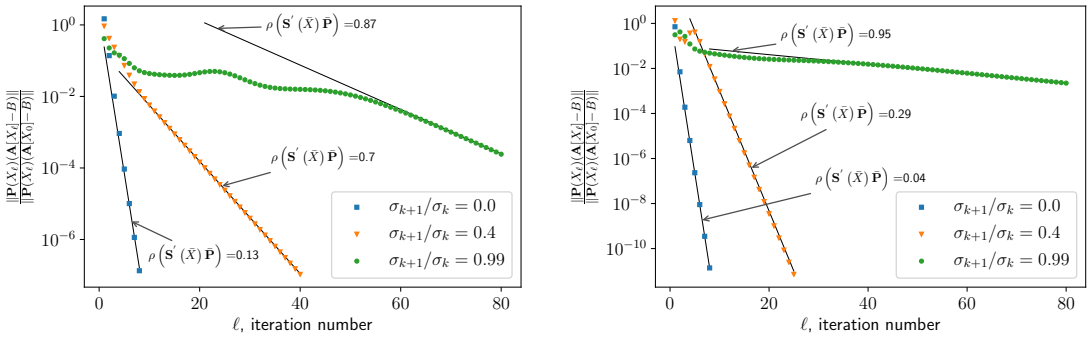
The goal is to minimize the quadratic cost function (3.9) subject to $\operatorname{rank}(X) \leq k$. We consider two examples for the operator $\mathbf{A} = \bar{\mathbf{A}}$. The first is a random SPD matrix

$$\mathbf{A} = \mathbf{R}^\top \mathbf{R} \in \mathbb{R}^{n^2 \times n^2},$$

where $\mathbf{R}$ is a matrix with each element produced by the standard normal distribution. As a second example we take the well-known matrix arising in the discretization of a 2D Laplacian on uniform tensor product grid with zero Dirichlet boundary conditions:

$$\mathbf{A} = I_n \otimes D_n + D_n \otimes I_n \in \mathbb{R}^{n^2 \times n^2}, \quad D_n = (n+1)^2 \operatorname{tridiag}(-1, 2, -1)_{n \times n}.$$

Figure 2 displays experimental results for the ALS algorithm with $n = 50$ and $k = 2$. The matrix B has been chosen such that the solution of the matrix equation $\mathbf{A}[Y] = B$ (that is, the global minimizer of (3.9) without low-rank constraint) has a predefined distribution of singular values. Similarly to the experiments for $\bar{\mathbf{A}} = \mathbf{I}$ we set $\sigma_2(Y) = 10^{-3}$, while $\sigma_3(Y)$ varies and the goal is to find the best rank-2 approximation $\bar{X}$. As we observe, numerical behaviour is in close agreement with theoretical estimates. While the convergence rate $\rho(\bar{\mathbf{S}}'(\bar{X})\bar{\mathbf{P}})$ does not equal σ_{k+1}/σ_k as in the case $\bar{\mathbf{A}} = \mathbf{I}$, it still seems related to this ratio for both choices of $\bar{\mathbf{A}}$. Remarkably, $\rho(\bar{\mathbf{S}}'(\bar{X})\bar{\mathbf{P}})$ is smaller than one even when σ_{k+1}/σ_k is close to one. A decisive question for future work would be for which combinations of $\bar{\mathbf{A}}$ and B this can be rigorously shown.



(a) (b)

Figure 2: Relative projected residual w.r.t. the iteration number for $\mathbf{A}$: (a) random SPD and (b) Laplace matrix, $k = 2$, $\sigma_k = 10^{-3}$ – predefined singular value of the solution of $\mathbf{A}[X] = B$. Black lines have slopes corresponding to the theoretical convergence rate for different σ_{k+1}.

Also note that, in contrast to the case $\bar{\mathbf{A}} = \mathbf{I}$, there is no superlinear convergence in this experiment when $\sigma_{k+1} = 0$. In this situation the minimizer of (3.9) on the rank-k variety is the same as the global one, so the curvature free case considered in sec. 2.2.2 (zero gradient) applies. Local linear convergence of the ALS method to this minimizer is then guaranteed by Theorem 2.4.

5 Conclusion

The goal of this paper was to derive transparent conditions for the local linear convergence of alternating optimization (AO) algorithms for multilinear and low-rank optimization, specifically the ALS algorithm, which reflect the underlying geometry and do not depend on the representation of low-rank tensors as in previous works. Due to multilinearity of the cost function, single optimization steps take place on linear subspaces, leading (in particular for multiquadratic cost functions) to an interpretation of AO as a nonlinear subspace correction method (with changing subspaces). Using a sufficiently general framework, a formula for the derivative of the nonlinear iteration function can be obtained (Theorem 2.3), which displays the interplay of terms from the classic linear subspace correction method with the curvature of the underlying low-rank manifold and the gradient of the cost function in a clear way. The main task remains to show that the spectral radius of this derivative is less than one in applications of interest. This is true in low-rank optimization tasks where the global minimizer lies on the considered low-rank manifold. The case where this is not true is more subtle. For AO for low-rank matrices, the curvature terms can be considerably simplified, which allows for an alternative, analytic proof for the well known convergence rate of the simultaneous orthogonal iteration for computing the dominant left and right singular subspaces of a matrix. While the main trick (Theorem 2.5) that was used to obtain this result may not apply in more general situations, we hope that our framework can be a useful starting point in future works for finding rigorous statements for the observed linear convergence of AO and ALS in other applications, like low-rank solutions of Lyapunov equations (cost function (3.9)) and low-rank tensor approximation.

Related work

In [14] and [11] the local convergence of the alternating least square algorithm has been analyzed for low-rank tensor approximation in the CP and tensor train format, respectively, using the nonlinear Gauss-Seidel approach for a cost function of the form (1.5), e.g., using an explicit representation of low-rank tensors. To address the problem that the Hessian of this cost function cannot be positive definite due to non-uniqueness of tensor representations, equivalence classes of representations (level sets of the function τ in (1.5)) are introduced. Linear convergence is then established for the case that the null space of the Hessian equals the tangent space of the orbit of equivalent representations. The idea is certainly analogous to restricting the operator $\mathbf{S}'(\bar{x})$ to the subspace $T(\bar{x})$ as in the present paper, but we believe that our approach provides a much clearer picture by avoiding the unintuitive concept of equivalent representations. A formula

$$\langle h, \nabla F^2(\xi)h\rangle = \langle \tau'(\xi)h, \nabla^2 f(x)\tau'(\xi)h\rangle + \langle \nabla f(x), \tau''(\xi)[h,h]\rangle \tag{5.1}$$

for the Hessian at $x = \tau(\xi)$ is given in [11], which features the Hessian $\nabla^2 f(x)$ on the tangent space of the image of τ, and the interaction of curvature ($\tau''(x)$) and gradient ∇f as in our work (cf. the definition (2.6) of $\bar{\mathbf{N}}_i$). In particular, it is concluded that local convergence is guaranteed if $\nabla f(\bar{x}) = 0$ under an injectivity assumption on $\tau'(\bar{\xi})$.

For optimization problems on manifolds, the interplay of global Hessian, gradient and curvature as displayed in (5.1) is gathered in the important concept of the *Riemannian Hessian*. This is thoroughly discussed in [2], see in particular sec. 6 therein. Similar to its role in smooth optimization in linear spaces, the positivity of the Riemannian Hessian ensures local (Riemannian) convexity and hence contractivity of many Riemannian optimization methods; see the book [1]. For manifolds of low-rank matrices, the curvature terms in this Hessian has been obtained in other works. Specifically, [15, Proposition 2.2] features a formula that makes the Kronecker type interplay between $\bar{X}^+$ and $\nabla f(\bar{X})$ in the curvature (see Lemma 3.1) clearly visible, albeit for a special case of the cost function (3.9) related to matrix completion (with $\mathbf{A} = \mathbf{P}_\Omega$ being a projection on given entries Ω). In [8], the curvature term in the Riemannian Hessian is explicitly neglected to derive Riemannian Gauss-Newton type methods on low-rank tensor manifolds.

In the works [9] and [4], convergence of the ALS method for low-rank tensor approximation has been investigated using alternative techniques. In particular, questions on cluster points and global convergence are addressed. Also, examples for sublinear, linear and superlinear convergence are presented in [4]. Also, we mention the work [13], in which global convergence of a related method (called scaled alternating steepest descent) for matrix completion is investigated.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[2] P.-A. Absil, J. Trumpf, R. Mahony, and B. Andrews. All roads lead to Newton: Feasible second-order methods for equality-constrained optimization. *Tech. Report UCL-INMA-2009.024*, 2009.

[3] F. L. Bauer. Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. *Z. Angew. Math. Phys.*, 8:214–235, 1957.

[4] M. Espig, W. Hackbusch, and A. Khachatryan. On the convergence of alternating least squares optimisation in tensor format representations. *arXiv:1506.00062*, May, 2015.

[5] A. Frommer, R. Nabben, and D. B. Szyld. Convergence of stationary iterative methods for Hermitian semidefinite linear systems and applications to Schwarz methods. *SIAM J. Matrix Anal. Appl.*, 30(2):925–938, 2008.

[6] G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM J. Numer. Anal.*, 10:413–432, 1973.

[7] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[8] D. Kressner, M. Steinlechner, and B. Vandereycken. Preconditioned low-rank riemannian optimization for linear systems with tensor product structure. *SIAM J. Sci. Comput.*, 38(4):A2018–A2044, 2016.

[9] M. J. Mohlenkamp. Musings on multilinear fitting. *Linear Algebra Appl.*, 438(2):834–852, 2013.

[10] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.

[11] T. Rohwedder and A. Uschmajew. On local convergence of alternating schemes for optimization of convex problems in the tensor train format. *SIAM J. Numer. Anal.*, 51(2):1134–1162, 2013.

[12] R. Schneider and A. Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM J. Optim.*, 25(1):622–646, 2015.

[13] J. Tanner and K. Wei. Low rank matrix completion by alternating steepest descent methods. *Appl. Comput. Harmon. Anal.*, 40(2):417–429, 2016.

[14] A. Uschmajew. Local convergence of the alternating least squares algorithm for canonical tensor approximation. *SIAM J. Matrix Anal. Appl.*, 33(2):639–652, 2012.

[15] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013.