A NEW CONVERGENCE PROOF FOR THE HIGHER-ORDER POWER METHOD AND GENERALIZATIONS

ANDRÉ USCHMAJEW*

ABSTRACT. A proof for the point-wise convergence of the factors in the higher-order power method for tensors towards a critical point is given. It is obtained by applying established results from the theory of Łojasiewicz inequalities to the equivalent, unconstrained alternating least squares algorithm for best rank-one tensor approximation.

1. INTRODUCTION

Finding the best rank-one approximation to a given higher-order tensor is equivalent to finding its largest tensor singular value (also known as its spectral norm), which is defined as the maximum of the associated multilinear form on the product of unit spheres. This simplest of all low-rank tensor approximation tasks is of large interest in its own, but also constitutes the main building-block when constructing approximations of higher rank by means of rank-one updates, see for example the references given in [7, Sec. 3.3].

The *higher-order power method* (HOPM) [4, 5] is a simple, effective, and widely used optimization algorithm to approximately solve the task. The name comes from the fact that it is the straight-forward generalization of an alternating power method for finding a pair of dominant left and right singular vectors of a matrix. Depending on the scaling strategy used for the iterates during the process, the higher-order power method can be seen as an *alternating least squares* (ALS) algorithm, see [10] and references therein.

Despite its importance, a satisfactory convergence theory for the HOPM was missing until recently. Clearly, the convergence of the generated sequence of approximated singular values follows easily from the monotonicity of the method [17]. More interesting and important, however, is the question of single-point (and not just sub-sequential) convergence of the sequences of generated rank-one tensors or even their factors to a critical point of the problem. The local convergence for starting guesses close enough to a critical point was established in [24] and [21], but the made assumptions remained somewhat restrictive. Concerning global convergence, the investigations of Mohlenkamp [15] showed that the sequence of rank-one tensors generated by ALS is bounded, and that their consecutive differences are absolutely square summable and hence converge to zero. This would imply convergence of the method, if the set of cluster points, each of which must be a critical point, contains at least one isolated point which then is the limit. In a recent work by Wang and Chu [22] this last issue was addressed by arguing that for almost every tensor the second-order derivative at zeros of the projected gradient of the cost function is regular, and

*MATHICSE-ANCHP, Section de Mathématiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. Current address: Hausdorff Center for Mathematics & Institute for Numerical Simulation, University of Bonn, 53115 Bonn, Germany (uschmajew@ins.uni-bonn.de).

2010 *Mathematics Subject Classification.* 15A69, 49M20, 65K05, 68W25, 90C26.

Key words and phrases. Tensors, rank-one approximation, higher-order power method, alternating least squares, global convergence, Łojasiewicz inequality.

hence critical points isolated. In this way, global convergence of the higher-order power method has been established, at least for almost every tensor.

The outlined argumentation appears, however, somewhat intricate. In this paper, we propose an alternative convergence proof based on an elegant method from the theory of analytic gradient flows, whose foundation is the *Łojasiewicz gradient inequality* – a powerful feature of real-analytic functions. Simply speaking, the validity of this inequality at a cluster point of a gradient-related descent iteration enforces absolute summability of increments, which implies convergence [1]. The continuous counterpart of this methodology is mentioned in [22], but the possibility to directly apply the available results on discrete gradient flows to ALS was not explored. This is what we shall do in the present paper.

In [23], Xu and Yin used a further generalization, the Kurdyka-Łojasiewicz inequality, to obtain convergence results for a variety of cyclic block coordinate descent methods when applied to a large class of strongly block multiconvex functions. This includes a wide range of alternating block techniques for regularized low-rank tensor optimization tasks. In principle, our considerations will show that even *without* regularization, the ALS algorithm for rank-one approximation is a member of this problem class. The key observation is an insight gained in [12], that the norms of the factors generated by ALS remain bounded from above and below, even when no normalization is used. In particular, norm constraints can be avoided in the analysis for this reason.

The focus on one specific method allows us to present the logic of the convergence proof in a simplified form compared to the very general reasoning in [23]. As a result, we obtain the global convergence of the higher-order power method as the last link in a transparent chain of simple arguments. Admittedly, the abstract results based on the Łojasiewicz gradient inequality, that are invoked at one point in the presentation, constitute a nontrivial ingredient in our proof, but they can be regarded as well-established by now.

The paper is organized as follows. In Sec. 2 we introduce the notation used, define the higher-order power method, and the equivalent alternating least squares algorithm. In Sec. 3 we state the abstract convergence results from the literature on which we rely, and then prove that they can be applied to rank-one ALS. The main result is Theorem 3.10. Finally, Sec. 4 is devoted to generalizations of the used arguments to strongly convex optimization tasks in other multilinear tensor formats by means of ALS-type algorithms [3, 9, 10, 18, 16, 20]. We explain why for formats other than rank-one, regularization is typically unavoidable to achieve similar strong results.

2. Best rank-one approximation

In this section, we recall the higher-order power method and the alternating least squares algorithm, and explain their connection in more detail.

2.1. Preliminaries.

Let $d \geq 3$ and $n_1, n_2, \dots, n_d \in \mathbb{N}$ be given. The elements of the Cartesian product $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_d}$ will be either explicitly denoted by tuples $(x^1, x^2, \dots, x^d)$, or abbreviated by

$$\mathbf{x} = (x^1, x^2, \dots, x^d).$$

The elements in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ will be called *tensors* and are treated as multi-dimensional arrays with entries labeled by multi-indices $i_1, i, \dots, i_d$. For tensors we use $\langle \cdot, \cdot \rangle_{\mathsf{F}}$ and $\| \cdot \|_{\mathsf{F}}$ to denote the Frobenius inner product and norm, respectively. For vectors, we omit the subscript F when denoting the Euclidean inner product and norm. Similarly, the norm of a tuple $\mathbf{x}$ will be denoted by $\|\mathbf{x}\| = (\|x^1\|^2 + \|x^2\|^2 + \cdots + \|x^d\|^2)^{1/2}$.

Consider the multilinear map $\tau_1 : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_d} \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ defined by

$$\tau_1(\mathbf{x}) = x^1 \circ x^2 \circ \cdots \circ x^d, \tag{1}$$

where $\circ$ is the outer product. It means that $[\tau_1(\mathbf{x})]_{i_1,i_2,\ldots,i_d} = x^1_{i_1} x^2_{i_2} \cdots x^d_{i_d}$. The non-zero tensors in the range of τ_1 are called *rank-one tensors*. The vectors x^μ will be called *factors* of $\tau_1(\mathbf{x})$. A crucial property of τ_1 is

$$\langle \tau_1(\mathbf{x}), \tau_1(\mathbf{y}) \rangle_{\mathsf{F}} = \langle x^1, y^1 \rangle \langle x^2, y^2 \rangle \cdots \langle x^d, y^d \rangle, \tag{2}$$

and therefore

$$\|\tau_1(\mathbf{x})\|_{\mathsf{F}} = \|x^1\| \|x^2\| \cdots \|x^d\|. \tag{3}$$

To a tensor $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ we associate a multilinear form F defined as

$$F(\mathbf{x}) = \langle \mathcal{F}, \tau_1(\mathbf{x}) \rangle_{\mathsf{F}} = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} \mathcal{F}_{i_1,i_2,\ldots,i_d} x^1_{i_1} x^2_{i_2} \cdots x^d_{i_d}.$$

For $\mu = 1, 2, \ldots, d$, we also define partial contractions $F^\mu(\mathbf{x})$ which are the vectors in $\mathbb{R}^{n_\mu}$, whose i_μth entry is

$$\sum_{i_1=1}^{n_1} \cdots \sum_{i_{\mu-1}=1}^{n_{\mu-1}} \sum_{i_{\mu+1}=1}^{n_{\mu+1}} \cdots \sum_{i_d=1}^{n_d} \mathcal{F}_{i_1,\ldots,i_{\mu-1},i_\mu,i_{\mu+1},\ldots,i_d} x^1_{i_1} \cdots x^{\mu-1}_{i_{\mu-1}} x^{\mu+1}_{i_{\mu+1}} \cdots x^d_{i_d},$$

that is, the contraction with x^μ is omitted. Equivalently, $F^\mu(\mathbf{x})$ may be defined as the unique vector in $\mathbb{R}^{n_\mu}$ satisfying

$$\langle F^\mu(\mathbf{x}), x^\mu \rangle = F(\mathbf{x}) \tag{4}$$

for all x^μ.

The algorithms we consider produce sequences of iterates $(x^\mu_k)_k$ for every component $\mu = 1, 2, \ldots, d$. We hence introduce the notation

$$\mathbf{x}^\mu_k = (x^1_k, \ldots, x^\mu_k, x^{\mu+1}_{k-1}, \ldots, x^d_{k-1}).$$

For convenience, let further $\mathbf{x}^0_{k+1} = \mathbf{x}_k$.

2.2. **Higher-order power method.** The critical values of $\mathbf{x} \mapsto F(\mathbf{x})/(\|x^1\| \|x^2\| \cdots \|x^d\|)$ are called the *singular values* of the tensor $\mathcal{F}$ [13]. The *maximum singular value* is

$$\lambda^* = \max_{\|x^1\|=\|x^2\|=\cdots=\|x^d\|=1} F(\mathbf{x}). \tag{5}$$

This expression defines a norm (the usual norm of a multilinear form), and so λ^* is sometimes referred to as the *spectral norm* of the underlying tensor $\mathcal{F}$ in the literature. The higher-order power method (HOPM) is a cyclic block coordinate method to approximate λ^*. By (4), the optimal choice for x^μ when fixing the other factors is

$$x^\mu_{\text{HOPM}} = \frac{F^\mu(\mathbf{x})}{\|F^\mu(\mathbf{x})\|}. \tag{6}$$

This already constitutes the HOPM summarized as Algorithm 1. For clarity we use the notation $\mathbf{y}^\mu_k$ for the iterates of HOPM, and reserve $\mathbf{x}^\mu_k$ for the iterates produced by the ALS algorithm (Algorithm 2) introduced next.

Note that since $F(\mathbf{y}^\mu_k)$ is not decreasing and $F(\mathbf{y}^1_1) > 0$, a division by zero will never occur.

Algorithm 1: Higher-order power method (HOPM)

Input: Tensor $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, starting guess $\mathbf{y}_0$ with $F^1(\mathbf{y}_0) \neq 0$.
$k \leftarrow 0$, $\lambda_0 = F(\mathbf{y}_0)$
while *not converged* **do**
  for $\mu = 1, 2, \dots, d$ **do**
    $y^\mu_{k+1} = \dfrac{F^\mu(\mathbf{y}^{\mu-1}_{k+1})}{\|F^\mu(\mathbf{y}^{\mu-1}_{k+1})\|}$
  end
  $\lambda_{k+1} = F(\mathbf{y}_{k+1})$
  $k \leftarrow k+1$
end

2.3. **Alternating least squares.** Given $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, let

$$f(\mathbf{x}) = \frac{1}{2}\|\mathcal{F} - \tau_1(\mathbf{x})\|^2_{\mathsf{F}}. \tag{7}$$

The best rank-one approximation problem consists in finding a minimizer for f. The corresponding block coordinate descent method is called alternating least squares (ALS). The name comes from the fact that the problem for a single factor x^μ with the others held fixed is a least squares problem with normal equation

$$\begin{aligned} 0 &= \langle \mathcal{F} - \tau_1(\mathbf{x}), \tau_1(x^1, \dots, x^{\mu-1}, y^\mu, x^{\mu+1}, \dots, x^d)\rangle_{\mathsf{F}} \\ &= \left\langle F^\mu(\mathbf{x}) - \left(\prod_{\nu \neq \mu} \|x^\nu\|^2\right) x^\mu, y^\mu \right\rangle \quad \text{for all } y^\mu \in \mathbb{R}^{n_\mu}, \end{aligned} \tag{8}$$

where we have used (2) and (4). Assuming $x^\nu \neq 0$ for all $\nu \neq \mu$, the unique solution is

$$x^\mu_{\text{ALS}} = \frac{F^\mu(\mathbf{x})}{\prod_{\nu \neq \mu} \|x^\nu\|^2}. \tag{9}$$

The resulting ALS algorithm is noted as Algorithm 2.

Algorithm 2: Alternating Least Squares (ALS)

Input: Tensor $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, starting guess $\mathbf{x}_0$ with $F^1(\mathbf{x}_0) \neq 0$.
$k \leftarrow 0$
while *not converged* **do**
  for $\mu = 1, 2, \dots, d$ **do**
    $x^\mu_{k+1} = \dfrac{F^\mu(\mathbf{x}^{\mu-1}_{k+1})}{\|x^1_{k+1}\|^2 \cdots \|x^{\mu-1}_{k+1}\|^2 \cdot \|x^{\mu+1}_k\|^2 \cdots \|x^d_k\|^2}$
  end
  $k \leftarrow k+1$
end

Note that $F^1(\mathbf{x}_0) \neq 0$ implies $x^\mu_0 \neq 0$ for $\mu = 2, \dots, d$, so the very first step of the algorithm is feasible and $x^1_1 \neq 0$. As we show in the next section, the sequences $\|x^\mu_k\|$ are monotonically increasing for every μ, so the subsequent update steps also never fail.

In contrast to what is recommended in practice (see e.g. [10]), our version of ALS omits any normalization of the factors x_k^μ during the process. This is by purpose, as it simplifies the analysis. From a theoretical viewpoint it also makes no difference, as any rescaling strategy does not affect the generated sequence of subspaces $\operatorname{span}(x_k^\mu)$. In particular, comparing (6) and (9) it is plain to see the equivalence of HOPM and ALS, but the detailed proof requires some notational effort.

Proposition 2.1. *Let $(\lambda_k, \mathbf{y}_k)$ and $(\mathbf{x}_k)$ denote the iterates generated by Algorithms 1 and 2, respectively, when applied to the same starting guess $\mathbf{y}_0 = \mathbf{x}_0$. Then it holds*

$$y_k^\mu = \frac{x_k^\mu}{\|x_k^\mu\|}, \quad \textit{and} \quad \lambda_k = \|\tau_1(\mathbf{x}_k)\|_{\mathsf{F}}$$

for all $k \geq 1$ and all μ. Also, if $\mathbf{x}_*$ is a critical point of the function (7) *with $\tau_1(\mathbf{x}_*) \neq 0$, then $\mathbf{y}_*$ with $y_*^\mu = x_*^\mu / \|x_*^\mu\|$ is a critical point (w.r.t. the spherical constraints) of* (5)*.*

Proof. We show by induction that for every $k \geq 0$ and μ there exists $\alpha_k^\mu > 0$ such that $x_k^\mu = \alpha_k^\mu y_\mu^k$. For $k \geq 1$ this obviously implies $\alpha_k^\mu = \|x_k^\mu\|$, as $\|y_k^\mu\| = 1$ by construction. We introduce an ordering of the pairs (k, μ) according to their appearance in the algorithms, i.e., $(k, \mu) > (\ell, \nu)$ if $k > \ell$ or if $k = \ell$ and $\mu > \nu$. Setting $\alpha_0^\mu = 1$ the assertion $y_0^\mu = \alpha_0^\mu x_0^\mu$ obviously holds for all pairs (k, μ) with $k = 0$. Now fix $(k+1, \mu)$ and assume the relation has been proved for all previous pairs. Exploiting the multilinearity of $F^\mu(\mathbf{x})$ w.r.t. $x^1, \dots, x^{\mu-1}, x^{\mu+1}, \dots, x^d$, and using $\alpha_{k+1}^\nu = \|x_{k+1}^\nu\|$ for $\nu < \mu$, it holds

$$x_{k+1}^\mu = \frac{\alpha_k^{\mu+1} \cdots \alpha_k^d F^\mu(\mathbf{y}_{k+1}^{\mu-1})}{\alpha_{k+1}^1 \cdots \alpha_{k+1}^{\mu-1} \|x_k^{\mu+1}\|^2 \cdots \|x_k^d\|^2} = \frac{\alpha_k^{\mu+1} \cdots \alpha_k^d \|F(\mathbf{y}_{k+1}^{\mu-1})\|}{\alpha_{k+1}^1 \cdots \alpha_{k+1}^{\mu-1} \|x_k^{\mu+1}\|^2 \cdots \|x_k^d\|^2} \cdot y_{k+1}^\mu \tag{10}$$

(note that $\alpha_k^{\mu+1} \cdots \alpha_k^d$ and $\|x_k^{\mu+1}\| \cdots \|x_k^d\|$ also cancel once $k \geq 1$). Hence α_{k+1}^μ equals the fraction on the right side of (10), which is positive.

Now that we have proved $x_k^\mu = \alpha_k^\mu y_k^\mu$ with $\alpha_k^\mu = \|x_k^\mu\|$ for all $k \geq 1$, (10) and (3) imply

$$\|\tau_1(\mathbf{x}_{k+1}^\mu)\|_{\mathsf{F}} = \|F^\mu(\mathbf{y}_{k+1}^{\mu-1})\|.$$

By definition of y_{k+1}^μ and (4),

$$\|F^\mu(\mathbf{y}_{k+1}^{\mu-1})\| = \langle F^\mu(\mathbf{y}_{k+1}^{\mu-1}), y_{k+1}^\mu \rangle = \langle F^\mu(\mathbf{y}_{k+1}^\mu), y_{k+1}^\mu \rangle = F(\mathbf{y}_{k+1}^\mu),$$

where the second equality holds because $F^\mu(\mathbf{x})$ never depends on x^μ. In summary,

$$\|\tau_1(\mathbf{x}_{k+1}^\mu)\|_{\mathsf{F}} = F(\mathbf{y}_{k+1}^\mu), \tag{11}$$

and in particular $\|\tau_1(\mathbf{x}_{k+1})\|_{\mathsf{F}} = F(\mathbf{y}_{k+1}) = \lambda_{k+1}$.

Finally, let $\mathbf{x}_*$ be a critical point of the function (7) with $\tau_1(\mathbf{x}_*) \neq 0$. Then $\mathbf{x}^*$ is a stationary point of (9), so that for all μ it holds $x_*^\mu = \alpha_*^\mu F^\mu(\mathbf{x}_*)$ with $\alpha_*^\mu = \prod_{\nu \neq \mu} \|x_*^\nu\|^2 \neq 0$. Let $y_*^\mu = x_*^\mu / \|x_*^\mu\|$, then using multilinearity it also follows that

$$\beta_*^\mu y_*^\mu = F^\mu(\mathbf{y}_*) \tag{12}$$

for some $\beta_*^\mu \neq 0$. The tangent space to the unit sphere at y_*^μ consists of all vectors δy_*^μ orthogonal to y_*^μ. Hence $\mathbf{y}_*$ is a critical point of F on the Cartesian product of spheres, if for every μ it holds $\langle \nabla_\mu F(\mathbf{y}_*), \delta y_*^\mu \rangle = 0$ for all such δy_*^μ. But since F is linear with respect to every block variable, this is the case, as

$$\langle \nabla_\mu F(\mathbf{y}_*), \delta y_*^\mu \rangle = \langle F^\mu(\mathbf{y}_*), \delta y_*^\mu \rangle = \beta_*^\mu \langle y_*^\mu, \delta y_*^\mu \rangle = 0$$

by (4) and (12). □

As a result, we can prove convergence of HOPM by proving convergence of ALS.

3. Convergence of alternating least squares

The global, point-wise convergence of the iterates generated by Algorithm 2 will be deduced from known results based on the Łojasiewicz gradient inequality. We first recall the required abstract properties, and then show that they hold for Algorithm 2.

3.1. Point-wise convergence via Łojasiewicz inequality.

Our aim is to apply the following result [1, Theorem 3.2].

Theorem 3.1. *Let $f : V \to \mathbb{R}$ be a real-analytic function on a finite-dimensional real vector space V, and let $(\mathbf{x}_k) \subset \mathbb{R}^n$ be a sequence satisfying*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \sigma \|\nabla f(\mathbf{x}_k)\| \, \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \tag{13}$$

for all large enough k and some $\sigma > 0$. Assume further that the implication

$$[f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k)] \quad \Rightarrow \quad \mathbf{x}_{k+1} = \mathbf{x}_k \tag{14}$$

holds. Then a cluster point $\mathbf{x}_*$ of the sequence $(\mathbf{x}_k)$ must be its limit. In particular, if the sequence is bounded, it is convergent.

The key ingredient in the proof of this theorem is the Łojasiewicz gradient inequality,

$$|f(\mathbf{x}) - f(\mathbf{x}_*)|^{1-\theta} \leq \Lambda \|\nabla f(\mathbf{x})\|, \tag{15}$$

which can be shown to hold in some (unknown) neighborhood of $\mathbf{x}_*$ when f is real-analytic [14, p. 92]. The constants $\Lambda > 0$ and $\theta \in (0, 1/2]$ are typically not explicitly known as well. Yet, in combination with (13) and (14), the Łojasiewicz gradient inequality allows to prove that the norms $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ of increments are summable, which implies convergence of the sequence $(\mathbf{x}_k)$.

Under stronger conditions, one can conclude that the limit is a critical point of f. The following theorem will be applicable to Algorithm 2, although the convergence rate estimates remain of minor use as long as no a-priori results on the expected value of the Łojasiewicz exponent θ are available.

Theorem 3.2. *Under the conditions of Theorem 3.1, assume further that there exists $\kappa > 0$ such that*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \geq \kappa \|\nabla f(\mathbf{x}_k)\| \tag{16}$$

for all large enough k. Then $\nabla f(\mathbf{x}_*) = 0$, and the convergence rate can be estimated as follows:

$$\|\mathbf{x}_* - \mathbf{x}_k\| \lesssim \begin{cases} q^k & \text{if } \theta = \frac{1}{2} \text{ (for some } 0 < q < 1\text{)}, \\ k^{-\frac{\theta}{1-2\theta}} & \text{if } 0 < \theta < \frac{1}{2}. \end{cases} \tag{17}$$

We were not able to identify the original reference for this theorem which seems rather scattered through the literature, see e.g. [2, 11]. For concreteness we point to [19], where Theorems 3.1 and 3.2 are proved in the stated form.

3.2. Application to Algorithm 2.

Let now f be the function (7) again, and $(\mathbf{x}_k)$ the sequence generated by Algorithm 2. By a chain of simple arguments, we will show that the required properties (13), (14), and (16) are satisfied. As $F^1(\mathbf{x}_0)$ only depends on $x_0^2, \dots, x_0^d$, we assume now without loss in generality that $x_0^1 = x_1^1$ to avoid special treatment of the very first update in the following proofs.

The first two results are well-known, and express the monotonicity of the algorithms.

Proposition 3.3. *For all $k \geq 1$ and $\mu = 1, 2, \ldots, d$ it holds*

$$\|\mathcal{F}\|_{\mathsf{F}}^2 = \|\tau_1(\mathbf{x}_k^\mu)\|_{\mathsf{F}}^2 + \|\mathcal{F} - \tau_1(\mathbf{x}_k^\mu)\|_{\mathsf{F}}^2.$$

Proof. This is a necessary optimality condition for the least squares problem that was solved to obtain x_k^μ, since, by homogeneity, $\tau_1(\mathbf{x}_k^\mu)$ is in particular the Euclidean best approximation of $\mathcal{F}$ in $\operatorname{span}(\tau_1(\mathbf{x}_k^\mu))$. (More concretely, it follows from choosing $x^\mu = y^\mu = x_{\text{ALS}}^\mu$ in (8) that $F(\mathbf{x}_n^\mu) = \|\tau_1(\mathbf{x}_n^\mu)\|_{\mathsf{F}}$.) □

Proposition 3.4. *For $\nu \geq \mu$ and $\ell \geq k$ it holds*

$$\|\tau_1(\mathbf{x}_\ell^\nu)\|_{\mathsf{F}} \geq \|\tau_1(\mathbf{x}_k^\mu)\|_{\mathsf{F}}.$$

Proof. This is an immediate consequence of Proposition 3.3, as, by the decreasing property of ALS, $\|\mathcal{F} - \tau_1(\mathbf{x}_\ell^\nu)\|_{\mathsf{F}}^2 \leq \|\mathcal{F} - \tau_1(\mathbf{x}_k^\mu)\|_{\mathsf{F}}^2$. Alternatively, the statement follows from (11) and the monotonicity of HOPM. □

The next two key conclusions were drawn in [12, Lemma 4.1]. The first is as crucial as it is trivial.

Proposition 3.5. *For every $\mu = 1, 2, \ldots, d$ the sequence $(\|x_k^\mu\|)$ of norms is monotonically increasing.*

Proof. As in every inner step of Algorithm 2 only one block is updated, this follows from Proposition 3.4 and (3). □

As a result, the norms of the factors x_k^μ remain bounded from below *and* from above.

Proposition 3.6. *For all $k \geq 1$ and $\mu = 1, 2, \ldots, d$ it holds*

$$0 < \|x_0^\mu\| \leq \|x_k^\mu\| \leq \|\mathcal{F}\|_{\mathsf{F}} \left(\prod_{\nu \neq \mu} \|x_0^\nu\|^{-1} \right).$$

Proof. Since $F^1(\mathbf{x}_0) \neq 0$, we have $\|x_0^\mu\| > 0$ for all $\mu \geq 2$. Then also $x_0^1 = x_1^1 \neq 0$ (the equality was assumed at the beginning of the section). The inequality $\|x_0^\mu\| \leq \|x_k^\mu\|$ holds by Proposition 3.5. Combining with (3) and Proposition 3.3 gives

$$\left(\prod_{\nu \neq \mu} \|x_0^\nu\| \right) \|x_k^\mu\| \leq \|\tau_1(\mathbf{x}_k)\|_{\mathsf{F}} \leq \|\mathcal{F}\|_{\mathsf{F}},$$

that is, the third inequality in the assertion. □

We now turn to the assumptions in Theorems 3.1 and 3.2.

Proposition 3.7. *In loop k, the decrease in function value of block update μ satisfies*

$$f(\mathbf{x}_{k+1}^{\mu-1}) - f(\mathbf{x}_{k+1}^\mu) = \frac{\sigma_{k+1}^\mu}{2} \|x_{k+1}^\mu - x_k^\mu\|^2,$$

where

$$\sigma_{k+1}^\mu = \|x_{k+1}^1\|^2 \cdots \|x_{k+1}^{\mu-1}\|^2 \cdot \|x_k^{\mu+1}\|^2 \cdots \|x_k^d\|^2.$$

Proof. This is standard least squares theory: the update x_{k+1}^μ is chosen such that the gradient of the quadratic form $x^\mu \mapsto f(x_{k+1}^1, \ldots, x_{k+1}^{\mu-1}, x^\mu, x_k^{\mu+1}, \ldots, x_k^d)$ is zero. Its quadratic term is, using (3),

$$x_\mu \mapsto \frac{1}{2} \|\tau_1(x_{k+1}^1, \ldots, x_{k+1}^{\mu-1}, x^\mu, x_k^{\mu+1}, \ldots, x_k^d\|_{\mathsf{F}}^2 = \frac{\sigma_{k+1}^\mu}{2} \|x^\mu\|^2,$$

hence the Hessian in every point is $\sigma_{k+1}^\mu I_{n_\mu}$. A Taylor expansion around x_{k+1}^μ proves the claim. □

Proposition 3.8. *The decrease in function value per outer loop satisfies*

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{\sigma_0}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,$$

where

$$\sigma_0 = \min_{\mu=1,2,\dots,d} \sigma_1^\mu > 0.$$

Proof. By Proposition 3.6, we have $\sigma_k^\mu \geq \sigma_1^\mu \geq \sigma_0 > 0$ for all $\mu = 1, 2, \dots, d$. Building a telescopic sum, Proposition 3.7 yields

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) = \sum_{\mu=1}^d f(\mathbf{x}_{k+1}^{\mu-1}) - f(\mathbf{x}_{k+1}^\mu) \geq \frac{\sigma_0}{2}\sum_{\mu=1}^d \|x_{k+1}^\mu - x_k^\mu\|^2 = \frac{\sigma_0}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,$$

as asserted. □

Proposition 3.9. *There exists $\kappa > 0$ such that* (16) *holds.*

Proof. By Proposition 3.6, the iterates $\mathbf{x}_k^\mu$ (so in particular the $\mathbf{x}_k$) remain in some compact set B for all k. Let $\nabla_\mu f(\mathbf{x})$ denote the partial block gradient at $\mathbf{x}$ with respect to x^μ. As f is twice continuously differentiable on B, there exists $L > 0$ such that

$$\|\nabla_\mu f(\mathbf{x}) - \nabla_\mu f(\mathbf{y})\| \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in B$. Since by construction of the iterates it holds $\nabla_\mu f(\mathbf{x}_{k+1}^\mu) = 0$, we deduce

$$\begin{aligned}\|\nabla f(\mathbf{x}_k)\|^2 = \sum_{\mu=1}^d \|\nabla_\mu f(\mathbf{x}_k)\|^2 &= \sum_{\mu=1}^d \|\nabla_\mu f(\mathbf{x}_{k+1}^\mu) - \nabla_\mu f(\mathbf{x}_k)\|^2 \\ &\leq L^2 \sum_{\mu=1}^d \|\mathbf{x}_{k+1}^\mu - \mathbf{x}_k\|^2 \\ &\leq L^2 \sum_{\mu=1}^d \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 = L^2 d\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.\end{aligned}$$

The second inequality follows from the fact that $\mathbf{x}_{k+1}^\mu$ shares the first μ blocks with $\mathbf{x}_{k+1}$, and the last $d - \mu$ blocks with $\mathbf{x}_k$. Hence (16) holds with $\kappa = 1/(L\sqrt{d})$. □

In summary, we obtain our main result.

Theorem 3.10. *The iterates $(\mathbf{x}_k)$ generated by Algorithm 2 converge to a point $\mathbf{x}_*$ with $\nabla f(\mathbf{x}_*) = 0$, where f is given by* (7)*. The convergence rate estimates* (17) *in terms of the (a-priori unknown) exponent in the Łojasiewicz gradient inequality* (15) *at $\mathbf{x}_*$ apply.*

Proof. As stated in Proposition 3.9, relation (16) holds for all k and some $\kappa > 0$. Proposition 3.8 then implies that both (13) and (14) also hold. The result is therefore an instance of Theorems 3.1 and 3.2. □

Without going into detail, we shall not conceal that the appearance of the tensor order d in the constant κ obtained in the proof of Proposition 3.9 may ultimately deteriorate the convergence rate stated in Theorem 3.2 for growing d. This rate, however, is not explicitly available anyway. Generally speaking, a dependence on the dimensionality has to be taken into account when relying on black-box tools like Theorems 3.1 and 3.2.

Due to the equivalence of ALS and HOPM in the sense of Proposition 2.1, Theorem 3.10 in particular states that the limit $\lambda_* = \|\tau(\mathbf{x}_*)\|_{\mathsf{F}}$ is a singular value of the tensor $\mathcal{F}$. There is no guarantee that it is the maximum singular value λ^*. Of course, by the monotonicity of HOPM, we would have $\lambda_* = \lambda^*$, if the starting guess $\lambda_0 = F(\mathbf{x}_0)/(\|x_0^1\|\|x_0^2\|\cdots\|x_0^d\|)$ happened to be larger than the second largest critical value (singular value), but ensuring this seems comparably hard as finding λ^* itself.

4. On generalizations to compositions of strongly convex functions with multilinear maps

In this section we take a second look at the main arguments used in Sec. 3.2 from an abstract perspective, much in the spirit of [23]. We explain why these arguments do not easily apply to general low-rank tensor optimization tasks by means of cyclic block coordinate descent (BCD), unless regularization is used.

A generic low-rank optimization problem is the following. One is given a function

$$J : \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} \to \mathbb{R},$$

and a multilinear map

$$\tau : V^1 \times V^2 \times \cdots \times V^D \to \mathbb{R}^{n_1\times n_2\times\cdots\times n_d},$$

where $V^1, V^2, \ldots, V^D$, $D \geq d$, are finite-dimensional vector spaces. Now denoting the elements of $V^1 \times V^2 \times \cdots \times V^D$ by $\mathbf{x} = (x^1, x^2, \ldots, x^D)$, the task is to minimize the function

$$f(\mathbf{x}) = J(\tau(\mathbf{x})) + \frac{\sigma_*}{2}\sum_{\mu=1}^{D}\|x^\mu\|^2, \tag{18}$$

where $\sigma_* \geq 0$ is a regularization parameter.

The most common examples for J are the squared Euclidean distance $\|\mathcal{F} - \mathcal{X}\|_{\mathsf{F}}^2$ to a given tensor $\mathcal{F}$, and the energy functional $\frac{1}{2}\langle \mathsf{A}\mathcal{X}, \mathcal{X}\rangle_{\mathsf{F}} - \langle \mathcal{B}, \mathcal{X}\rangle_{\mathsf{F}}$ of a "high-dimensional" linear system of equations $\mathsf{A}\mathcal{X} = \mathcal{B}$ with A being a symmetric positive definite linear operator on $\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$. The map τ, on the other hand, represents a low-rank tensor format of some fixed rank. A notable example is the rank-r CP format, which for $d = D = 3$ reads

$$\tau_r(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \sum_{i=1}^{r} a_i \circ b_i \circ c_i \tag{19}$$

with a_i, b_i, and c_i being the columns of $\mathbf{A} \in \mathbb{R}^{n_1\times r}$, $\mathbf{B} \in \mathbb{R}^{n_2\times r}$, and $\mathbf{C} \in \mathbb{R}^{n_3\times r}$, respectively. For $r = 1$ we recover (1). Other important examples of tensor formats are the Tucker format, the hierarchical Tucker format, and the tensor train format. We refer to [7, 8, 10] and references therein.

The generalization of Algorithm 2 to f given by (18) is the cyclic BCD method noted in Algorithm 3. It is feasible whenever J has bounded sub-level sets. We shall investigate to what extent one can prove convergence using the same ideas as in Sec. 3.2. To this end, we assume that J is real-analytic, convex, and coercive, that is, $\|\mathcal{X}\|_{\mathsf{F}} \to \infty$ implies $J(\mathcal{X}) \to \infty$. The two above-mentioned examples have this property. Then f in (18) is real-analytic, and at least the restriction to every block-variable is convex. For fixed $\mathcal{X}_0 = \tau(\mathbf{x}_0)$, let $\mathcal{L}_0 = \{\mathcal{X} : J(\mathcal{X}) \leq J(\mathcal{X}_0)\}$. Letting $\gamma_0 \geq 0$ be a lower spectral bound for the Hessians $\nabla^2 J(\mathcal{X})$ on the compact convex set $\mathcal{L}_0$, it follows from Taylor's theorem that

$$J(\mathcal{Y}) \geq J(\mathcal{X}) + \langle \nabla J(\mathcal{X}), \mathcal{Y} - \mathcal{X}\rangle_{\mathsf{F}} + \frac{\gamma_0}{2}\|\mathcal{Y} - \mathcal{X}\|_{\mathsf{F}}^2 \tag{20}$$

Algorithm 3: Cyclic BCD for low-rank optimization

Input: Starting guess $\mathbf{x}_0$.
$k \leftarrow 0$
while *not converged* **do**
  for $\mu = 1, 2, \dots, D$ **do**
    $x^\mu_{k+1} \in \operatorname{argmin}_{x^\mu \in V^\mu} f(x^1_{k+1}, \dots, x^{\mu-1}_{k+1}, x^\mu, x^{\mu+1}_k, \dots, x^D_k)$
  end
  $k \leftarrow k+1$
end

for all $\mathcal{X}, \mathcal{Y}$ in $\mathcal{L}_0$. Let us further introduce the quantities

$$\sigma^\mu_{k+1} = \min_{x^\mu \neq 0} \frac{\|\tau(x^1_{k+1}, \dots, x^{\mu-1}_{k+1}, x^\mu, x^{\mu+1}_k, \dots, x^D_k)\|_{\mathsf{F}}^2}{\|x^\mu\|^2}, \tag{21}$$

which are easily identified as the squared minimal singular values of the restricted linear maps $x^\mu \mapsto \tau(x^1_{k+1}, \dots, x^{\mu-1}_{k+1}, x^\mu, x^{\mu+1}_k, \dots, x^D_k)$ that arise during the iteration. They will play a similar same role as the σ^μ_{k+1} in Proposition 3.7. Note that if $\max(\gamma_0 \sigma^\mu_{k+1}, \sigma_*) > 0$, then the update x^μ_{k+1} is a unique choice, since in this case the corresponding restricted minimization problem is strongly convex (see (22)).

The new entry x^μ_{k+1} satisfies $\nabla_\mu f(\mathbf{x}^\mu_{k+1}) = 0$. Specifically, by the chain rule and the linearity of τ with respect to every block variable, this implies

$$\begin{aligned} 0 &= \langle \nabla_\mu f(\mathbf{x}^\mu_{k+1}), x^\mu_k - x^\mu_{k+1} \rangle \\ &= \langle \nabla J(\tau(\mathbf{x}^\mu_{k+1})), \tau(\mathbf{x}^{\mu-1}_{k+1}) - \tau(\mathbf{x}^\mu_{k+1}) \rangle_{\mathsf{F}} + \sigma_* \langle x^\mu_{k+1}, x^\mu_k - x^\mu_{k+1} \rangle. \end{aligned}$$

Since all generated tensors $\tau(\mathbf{x}^\mu_k)$ remain in $\mathcal{L}_0$, it then follows from (20) that

$$f(\mathbf{x}^{\mu-1}_{k+1}) - f(\mathbf{x}^\mu_{k+1}) \geq \frac{\gamma_0 \sigma^\mu_{k+1} + \sigma_*}{2} \|x^\mu_{k+1} - x^\mu_k\|^2, \tag{22}$$

which is the analog to Proposition 3.7.

4.1. **The regularized case.** Suppose we have chosen $\sigma_* > 0$. Then we can easily deduce an analog of Proposition 3.8 from (22). Further, the sub-level sets of f are bounded when $\sigma_* > 0$ (as J is bounded below). Hence, since f is decreasing, the sequences (x^μ_k) themselves remain bounded for every μ, which in turn allows to prove an analog of Proposition 3.9. These two propositions were sufficient to prove Theorem 3.10, which therefore can be generalized as follows.

Theorem 4.1 (cf. Xu and Yin [23, Theorems 2.8 and 2.9])**.** *Let J be real-analytic, convex, and coercive, and τ be multilinear as considered above. A sequence $(\mathbf{x}_k)$ of iterates generated by Algorithm 3 for the function f given by* (18) *with $\sigma_* > 0$ is uniquely determined by $\mathbf{x}_0$, and converges to a point $\mathbf{x}_*$ satisfying $\nabla f(\mathbf{x}_*) = 0$. The convergence rate estimates* (17) *apply correspondingly.*

4.2. **The non-regularized case.** When $\sigma_* = 0$, we need $\gamma_0 > 0$ (which is always possible if f is strictly convex), but also have to assume that

$$\liminf_{k,\mu} \sigma^\mu_k = \sigma_0 > 0, \tag{23}$$

in order to deduce an analog of Proposition 3.8 from (22). As the $\tau(\mathbf{x}_k^\mu)$ remain bounded, property (23) then implies, in light of (21), that the sequences (x_k^μ) also remain bounded for every μ, so that an analog of Proposition 3.9 again can be established.

Theorem 4.2. *Let J be real-analytic, strictly convex, and coercive, and τ be multilinear as considered above. A sequence $(\mathbf{x}_k)$ of iterates generated by Algorithm 3 for the function f given by* (18) *with $\sigma_* = 0$ satisfying* (23) *converges to a point $\mathbf{x}_*$ satisfying $\nabla f(\mathbf{x}_*) = 0$. The convergence rate estimates* (17) *apply correspondingly.*

Condition (23) can be interpreted as a stability requirement on the used tensor format during the iteration. Unless one finds a good argument to guarantee it in advance (for instance some condition on the starting guess), Theorem 4.2 remains an a-posteriori statement of minor practical value. For Algorithm 2, the product formula (3) and Proposition 3.6 (which itself is proved using (3)) imply (23). Unfortunately, a property like (3) is a unique feature of rank-one tensors. For none of the aforementioned tensor formats involving notions of higher rank an argument ensuring the stability condition (23) is currently available. In the case of optimization using rank-r CP tensors (19) with $r > 1$, this may be explained by the fact that the problem itself can be ill-posed [6]. Another reason for (23) to fail can be that the used rank in the multilinear tensor format overestimates the actual rank of the sought solution.

REFERENCES

[1] P.-A. ABSIL, R. MAHONY, AND B. ANDREWS, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM J. Optim., 16 (2005), pp. 531–547.

[2] H. ATTOUCH AND J. BOLTE, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Program., 116 (2009), pp. 5–16.

[3] G. BEYLKIN AND M. J. MOHLENKAMP, *Algorithms for numerical analysis in high dimensions*, SIAM J. Sci. Comput., 26 (2005), pp. 2133–2159 (electronic).

[4] L. DE LATHAUWER, P. COMON, B. DE MOOR, AND J. VANDEWALLE, *High-order power method – Application in Independent Component Analysis*, in Proceedings of the 1995 International Symposium on Nonlinear Theory and its Applications (NOLTA'95), 1995, pp. 91–96.

[5] L. DE LATHAUWER, B. DE MOOR, AND J. VANDEWALLE, *On the best rank-1 and rank-$(R_1, R_2, \cdots, R_N)$ approximation of higher-order tensors*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1324–1342 (electronic).

[6] V. DE SILVA AND L.-H. LIM, *Tensor rank and the ill-posedness of the best low-rank approximation problem*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 1084–1127.

[7] L. GRASEDYCK, D. KRESSNER, AND C. TOBLER, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitt., 36 (2013), pp. 53–78.

[8] W. HACKBUSCH, *Tensor Spaces and Numerical Tensor Calculus*, Springer-Verlag, Heidelberg, 2012.

[9] S. HOLTZ, T. ROHWEDDER, AND R. SCHNEIDER, *The alternating linear scheme for tensor optimization in the tensor train format*, SIAM J. Sci. Comput., 34 (2012), pp. A683–A713.

[10] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[11] A. LEVITT, *Convergence of gradient-based algorithms for the Hartree-Fock equations*, ESAIM Math. Model. Numer. Anal., 46 (2012), pp. 1321–1336.

[12] Z. LI, A. USCHMAJEW, AND S. ZHANG, *On Convergence of the Maximum Block Improvement Method*, SIAM J. Optim., 25 (2015), pp. 210–233.

[13] L.-H. LIM, *Singular values and eigenvalues of tensors: a variational approach*, in Proceedings of the 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2005), Dec. 2005, pp. 129–132.

[14] S. ŁOJASIEWICZ, *Ensemble semi-analytique*. Note des cours, Institut des Hautes Etudes Scientifique, 1965.

[15] M. J. MOHLENKAMP, *Musings on multilinear fitting*, Linear Algebra Appl., 438 (2013), pp. 834–852.

[16] I. V. OSELEDETS AND S. V. DOLGOV, *Solution of linear systems and matrix inversion in the TT-format*, SIAM J. Sci. Comput., 34 (2012), pp. A2718–A2739.

[17] P. REGALIA AND E. KOFIDIS, *The higher-order power method revisited: convergence proofs and effective initialization*, in Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00), vol. 5, 2000, pp. 2709–2712.

[18] 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 (2013), pp. 1134–1162.

[19] 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 (2015), pp. 622–646.

[20] C. TOBLER, *Low-rank tensor methods for linear systems and eigenvalue problems*, PhD thesis, ETH Zürich, 2012.

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

[22] L. WANG AND M. T. CHU, *On the global convergence of the alternating least squares method for rank-one approximation to generic tensors*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 1058–1072.

[23] Y. XU AND W. YIN, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci., 6 (2013), pp. 1758–1789.

[24] T. ZHANG AND G. H. GOLUB, *Rank-one approximation to high order tensors*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 534–550 (electronic).