On convergence of the maximum block improvement method

Z. Li, A. Uschmajew, S. Zhang

ON CONVERGENCE OF THE MAXIMUM BLOCK IMPROVEMENT METHOD

ZHENING LI*, ANDRÉ USCHMAJEW†, AND SHUZHONG ZHANG‡

Abstract. The so-called MBI (maximum block improvement) method is a greedy type approach to solve optimization problems where the decision variables can be grouped into a finite number of blocks. Assuming that optimizing over one block of variables while fixing all others is relatively easy, the MBI method then advocates to update the block of variables corresponding to the maximally improving block at each iteration, which is arguably a most natural and simple process to tackle block-structured problems with great potentials for engineering applications. It is therefore imperative to fully understand its convergence properties. In this paper we establish global and local linear convergence results for this method. The global convergence is established under the Łojasiewicz inequality assumption, while the local analysis invokes second-order assumptions. We study in particular the tensor optimization model with spherical constraints. Conditions for linear convergence of the famous power method for computing the maximum eigenvalue of a matrix follow in this framework as a special case. The condition is interpreted in various other forms for the rank-one tensor optimization model under spherical constraints. Numerical experiments are shown to support the convergence property of the MBI method for the randomly generated instances.

Key words. maximum block improvement, block coordinate descent, nonconvex optimization, rank-one tensor approximation, convergence, Łojasiewicz inequality

AMS subject classifications. 65N12, 90C26, 15A69, 41A25, 49M27

1. Introduction. Updating the values of certain variables sequentially and iteratively is a common practice in scientific computation. For instance, the Gauss-Seidel method or the Jacobi method are such schemes for solving linear equations. In the domain of optimization, the example extends to the so-called coordinate descent method [26]. Further, a substantial extension of coordinate descent is the *block* coordinate descent to solve problems where one variable is replaced by a *block* of variables.

If multiple blocks of variables are present, then the issue of selecting which order of blocks to update becomes imperative. A natural choice is to predetermine an ordering and then adhere to it. In the context of rank-one approximation of tensors, which is an application we also consider in this paper, such a cyclic coordinate descent is known as the *alternating least squares* (ALS) method. The difficult choice of orderings notwithstanding, the ALS method may fail to converge at all. However, if the ALS method converges under certain non-degeneracy assumption, then it actually has a local linear convergence rate [39]. This is also true for higher rank approximations, see Uschmajew [35].

For the general block optimization model, Chen et al. [7] proposed an ordering-free approach to select an appropriate block for updates. The main idea is to implement a block update which is the best possible among all the block updates. Hence the method was termed *maximum block improvement* (MBI). A global convergence result

---

*Department of Mathematics, University of Portsmouth, Portsmouth PO1 3HF, United Kingdom (zheningli@gmail.com).

†MATHICSE-ANCHP, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland (andre.uschmajew@epfl.ch). Parts of the paper were written while the author was employed at Technische Universität Berlin, Germany, supported by DFG Priority Program SPP 1324.

‡Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455 (zhangs@umn.edu). The research of this author was supported in part by U.S. National Science Foundation (Grant CMMI-1161242).

was proven in [7] under fairly loose conditions. However, the rate of convergence remained unknown. In this paper, we prove that the local rate of convergence of the MBI method for solving the general block optimization model is actually linear under a non-degeneracy condition, even if the overall objective is nonconvex.

The organization of the paper is as follows. In Section 2 we recall the MBI method, and provide a global convergence result (here and in the following understood as critical point convergence) for block strongly multi-convex functions assuming a Łojasiewicz inequality at a cluster point (Section 2.3). Global convergence results for this kind of functions have been recently obtained for the cyclic (Gauss-Seidel) BCD approach even in a non-smooth setting [38] based on the more involved Kurdyka-Łojasiewicz inequality [19], which itself recently attracted considerable attention in nonlinear optimization, see e.g. [4, 2]. Thanks to the greedy nature of the MBI method, the analysis is much simpler than that for Gauss-Seidel.

In Section 3, we present the local linear convergence analysis for the MBI method using second order information. The key assumption for linear convergence is a sort of non-degeneracy of the limiting KKT point. For the unconstrained minimization model, this boils down to the strict positivity of the Hessian at the KKT solution. By a local diffeomorphism, we study the non-degeneracy condition in the case of spherical constraints.

In the literature, global convergence results of block coordinate descent (BCD) methods are often obtained for functions which separate into sums of functions of single block variables. Instead, a particular situation we have in mind is the optimization of smooth (e.g. quadratic) functions in low-rank tensor formats [11], which usually indeed leads to block multi-convex and real-analytic problems (for which the Łojasiewicz inequality holds), but separate into (sums of) products of lower-variate functions. This typically makes it difficult to guarantee lower bounds of the moduli of strong block convexity. A modified abstract global convergence statement for this situation is formulated in Section 2.4.

More concretely, as promoted in [7], a possible application for the MBI method is the calculation of the best rank-one approximation to a (possibly super-symmetric) d-th order tensor $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ in the Frobenius norm. This problem can be both formulated as a block multi-convex unconstrained optimization problem, and as spherically constrained multi-linear maximization problem which is also known as the tensor maximum eigenvalue problem. In Section 4, we apply our global and local convergence results on this problem, and the corresponding numerical results will be presented in Section 5. To the best of our knowledge, Theorem 4.2 is the first global convergence result for best rank-one approximation by a block coordinate descent method. The key is that it is possible to bound from below the mentioned convexity moduli of the block functions when using a BCD method on this problem (Lemma 4.1), an observation which might be also used to prove global convergence of the ALS algorithm for rank-one approximation based on the results in [38].

Admittedly, rank-one tensors are the simplest objects in tensor approximation, but serve here as a perfect illustration of our abstract results, without involving too heavy tensor tools. Generalizations to optimization using, e.g. tensors in the TT tensor format [29, 28], in particular the global convergence result of Theorem 4.2, will be an interesting future research topic to be reported elsewhere.

Notation. We denote by $\langle \mathbf{x}, \mathbf{y} \rangle$ the Euclidian inner product of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and by $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ the Euclidian norm. For tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ we will add an additional subscript F to emphasize that we deal with the Frobenius norm

and inner product.

2. The MBI method and its global convergence property.

2.1. Definition of MBI. Given finite-dimensional spaces $X^1, X^2, \ldots, X^d$ and subsets $S^i \subseteq X^i$ a generic problem one wants to solve is

$$
\begin{aligned}
\min \quad & f(x^1, x^2, \ldots, x^d) \\
\text{s.t.} \quad & x^i \in S^i, \ i = 1, 2, \ldots, d,
\end{aligned}
\tag{2.1}
$$

where $f : S^1 \times S^2 \times \cdots \times S^d \to \mathbb{R}$ is a function of d block variables $x^i \in X^i$. In [7] the MBI method was proposed to solve (2.1) for certain problem classes. We first recall its definition. For every $\mathbf{x} \in S^1 \times S^2 \times \cdots \times S^d$ define the *block search sets*

$$
E_{\mathbf{x}}^i = \{(x^1, \ldots, x^{i-1})\} \times S^i \times \{(x^{i+1}, \ldots, x^d)\}, \ i = 1, 2, \ldots, d,
$$

and the shifted *block coordinate cross*

$$
E_{\mathbf{x}} = \bigcup_i E_{\mathbf{x}}^i.
$$

Let further

$$
\mathcal{S}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y} \in E_{\mathbf{x}}} f(\mathbf{y}),
$$

which is a set-valued function. Given a current iterate $\mathbf{x}_n \in S^1 \times S^2 \times \cdots \times S^d$, the MBI method chooses

$$
\mathbf{x}_{n+1} \in \mathcal{S}(\mathbf{x}_n)
\tag{2.2}
$$

by some rule. In practice, this amounts in calculating all restricted block solutions

$$
\boldsymbol{\xi}_{n+1}^i = \operatorname*{argmin}_{\mathbf{y} \in E_{\mathbf{x}_n}^i} f(\mathbf{y}),
\tag{2.3}
$$

and then choosing $\mathbf{x}_{n+1}$ to be the $\boldsymbol{\xi}_{n+1}^i$ which gives the best improvement. For reference the procedure is reproduced in Algorithm 1.

Algorithm 1 MBI method

Require: $\mathbf{x}_0 \in S^1 \times S^2 \times \cdots \times S^d$
 1: **for** $n = 0, 1, 2, \ldots$ **do**
 2:     **for** $i = 1, 2, \ldots, d$ **do**
 3:         Find some $\boldsymbol{\xi}_{n+1}^i \in \operatorname{argmin}_{\mathbf{y} \in E_{\mathbf{x}_n}^i} f(\mathbf{y})$.
 4:     **end for**
 5:     Choose $\mathbf{x}_{n+1} \in \operatorname{argmin}_{i \in \{1, 2, \ldots, d\}} f(\boldsymbol{\xi}_{n+1}^i)$.
 6: **end for**

2.2. MBI in the context of coordinate descent methods. The MBI method is a greedy-type *block coordinate descent* or *relaxation* method. As such it is to be distinguished from the following well-known relaxation methods.

1. The *cyclic relaxation* (or *nonlinear block Gauss-Seidel method*), where the optimization is looped over the spaces $E_{\mathbf{x}}^i$ in some specified order without a greedy comparison. Variants include the forward, the backward and the symmetric Gauss-Seidel iteration [32]. An extreme case is the almost cyclic iteration [24] where one only ensures one update per block coordinate within one fixed iteration length.

2. The *Gauss-Southwell rule* [33], where the block coordinate with the directional derivative of largest modulus is chosen as the next update direction. Defined in this way the Gauss-Southwell method is meaningful for optimization in open sets of linear spaces, although the required modifications to constrained optimization are obvious. Since the gradient provides only local information, the Gauss-Southwell procedure differs from MBI in general. Even for quadratic functions, more precisely, for minimizing

$$\frac{1}{2}\langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle$$

with positive definite matrix A they can be different unless, e.g. $A = aI$. In the context of quadratic functions, Griebel and Oswald call the Gauss-Southwell method the *greedy multiplicative Schwarz method* [12].

3. The *random relaxation*, where the block coordinate to be optimized is chosen randomly. A detailed investigation of this method started rather recently [25, 31], in particular for linear equations; see [20, 12].

In comparison with the mentioned methods, MBI can be considered as the most expensive one, since all block improvements have to be calculated per iteration, and much of information is discarded. It is therefore understood that its applicability is restricted to problems, where the one coordinate optimization can be done efficiently.

However, the additional cost lead to improved theoretical convergence properties compared to the other methods. Powell [30] gave examples of even smooth functions with a bounded sublevel set in which the cyclic relaxation can produce non-trivial loops, i.e., with the gradient remaining bounded from below. Typically some sort of convexity assumptions are required to prove global sub-sequential convergence to critical points [13, 34]. This is also the case for the nonlinear Gauss-Southwell method [24]. In contrast, for the MBI method a compact sublevel set $\{\mathbf{x} : f(\mathbf{x}) \geq f(\mathbf{x}_0)\}$ for the starting value is sufficient. This is a major theoretical advantage. In [7, Theorem 3.1] the following theorem has been proved.

THEOREM 2.1. *Assume that the S^i $(i = 1, 2, \ldots, d)$ are closed, and the MBI sequence $(\mathbf{x}_n)$ is bounded. Then every cluster point $\mathbf{x}_*$ of the MBI method will be a stationary point for (2.2), i.e.,*

$$\mathbf{x}_* \in \mathcal{S}(\mathbf{x}_*).$$

2.3. Global convergence based on multi-convexity and Łojasiewicz inequality. We consider the case that $S^1 \times S^2 \times \cdots \times S^d$ is open and f is differentiable. Then the statement of Theorem 2.1 means that $\nabla f(\mathbf{x}_*) = 0$; i.e., $\mathbf{x}_*$ is a critical point of (2.1). Hence in the unconstrained case it suffices to prove that a sequence generated by the MBI method converges at all to conclude that the limit point is a critical point.

To establish such a global convergence result, we would like to make use of a powerful principle from analytical dynamical systems, the Łojasiewicz gradient

inequality, which states that

$$\|\nabla f(\mathbf{x})\| \geq c|f(\mathbf{x}) - f(\mathbf{x}_*)|^{1-\theta} \tag{2.4}$$

for all $\mathbf{x}$ in some neighborhood of $\mathbf{x}_*$. Here $c > 0$ and $\theta \in (0, 1)$. We will not go into detail, but as Łojasiewicz showed, this inequality holds for instance at any point $\mathbf{x}_*$ (with the c and θ, as well as the valid neighborhood depending on $\mathbf{x}_*$) in whose neighborhood f is real-analytic [23, p. 92]. One can prove the following result [1].

THEOREM 2.2. *Let* $S^1 \times S^2 \times \cdots \times S^d$ *be open. Assume a sequence* $(\mathbf{x}_n) \subset S^1 \times S^2 \times \cdots \times S^d$ *satisfies the* strong descent condition

$$f(\mathbf{x}_n) - f(\mathbf{x}_{n+1}) \geq \sigma \|\nabla f(\mathbf{x}_n)\| \|\mathbf{x}_{n+1} - \mathbf{x}_n\|,$$

and the stationarity condition

$$[f(\mathbf{x}_{n+1}) = f(\mathbf{x}_n)] \Rightarrow [\mathbf{x}_{n+1} = \mathbf{x}_n]$$

for some $\sigma > 0$ *and all sufficiently large* n*. Then, if a cluster point* $\mathbf{x}_*$ *of the sequence satisfies the Łojasiewicz gradient inequality* (2.4)*, it is its limit point. This is for instance the case, if* f *is real-analytic and* $(\mathbf{x}_n)$ *is bounded.*

We next provide a problem class of interest – *block strongly multi-convex functions* – for which MBI turns out to meet the strong descent and stationarity requirements of Theorem 2.2. For simplicity, we restrict to the case that $S^i = X^i$ for $i = 1, 2, \ldots, d$. We say that f is *block downwards strongly multi-convex at* $\mathbf{x}$, if its restrictions to S^i,

$$f_{\mathbf{x}}^i : X^i \to \mathbb{R}, \ x^i \mapsto f(x^1, \ldots, x^i, \ldots, x^d),$$

are strongly convex on the current level-set, i.e.,

$$f_{\mathbf{x}}^i(z) \geq f_{\mathbf{x}}^i(y) + \langle \nabla f_{\mathbf{x}}^i(y), z - y \rangle + \frac{\ell_{\mathbf{x}}^i}{2}\|z - y\|^2 \quad \text{for all } y, z \text{ with } f^i(y), f^i(z) \leq f^i(x^i), \tag{2.5}$$

where for every $i = 1, 2, \ldots, d$ the constant $\ell_{\mathbf{x}}^i > 0$ can depend on all block-variables except x^i. Similarily, we say that ∇f is *block downwards multi-Lipschitz at* $\mathbf{x}$, if

$$\|\nabla f^i(z) - \nabla f^i(y)\| \leq L_{\mathbf{x}}^i \|z - y\| \quad \text{for all } y, z \text{ with } f^i(y), f^i(z) \leq f^i(x^i) \tag{2.6}$$

with a constants $L_{\mathbf{x}}^i$ which can depend on all blocks except x^i.

We now analyze one step of the MBI method from some $\mathbf{x}_n$ for such functions. For readability, we will use the subscript n instead of $\mathbf{x}_n$ in the following. Clearly, every f_n^i has a unique minimizer x_*^i with $\nabla f^i(x_*^i) = 0$, and trivially $f^i(x_*^i) \leq f(x_n^i)$. Let i^* be an index with maximum improvement, such that $\mathbf{x}_{n+1} \in E_{\mathbf{x}_n}^{i^*}$ is the response of the MBI method. Further, let

$$\ell_n = \min_{i=1,2,\ldots,d} \ell_n^i, \quad L_n = \max_{i=1,2,\ldots,d} L_n^i,$$

then we deduce from (2.5) and (2.6) that for any index $j \in \{1, 2, \ldots, d\}$ it holds

$$f(\mathbf{x}_n) - f(\mathbf{x}_{n+1}) = \max_{i=1,2,\ldots,d}(f(x_n^i) - f(x_*^i))$$

$$\geq \max_{i=1,2,\ldots,d} \frac{\ell_n^i}{2}\|x_n^i - x_*^i\|^2$$

$$\geq \frac{\ell_n}{2}\|x_n^j - x_*^j\| \|x_n^{i^*} - x_*^{i^*}\|$$

$$\geq \frac{1}{2}\frac{\ell_n}{L_n}\|\nabla f^j(x_n^j)\| \|\mathbf{x}_n - \mathbf{x}_{n+1}\|.$$

If we pick j as the index for which $\|\nabla f^j(x_n^j)\|$ is maximal (i.e., the index of the Gauss-Southwell rule), then it follows from $\|\nabla f(\mathbf{x}_n)\|^2 = \sum_{i=1}^d \|\nabla f^i(x_n^i)\|^2$ that

$$f(\mathbf{x}_n) - f(\mathbf{x}_{n+1}) \geq \frac{1}{2\sqrt{d}} \frac{\ell_n}{L_n} \|\nabla f(\mathbf{x}_n)\| \|\mathbf{x}_{n+1} - \mathbf{x}_n\|.$$

Applying Theorem 2.2, we arrive at the following result.

THEOREM 2.3. *Let* $(\mathbf{x}_n)$ *be a sequence generated by the MBI method (Algorithm 1) for problem* (2.1) *with* $S^i = X^i$ *for* $i = 1, 2, \ldots, d$ *(the unconstrained case). Assume that*

(i) *f is differentiable and block downwards strongly multi-convex at all $\mathbf{x}_n$, and ∇f is block downwards multi-Lipschitz at all $\mathbf{x}_n$ with constants ℓ_n^i and L_n^i, respectively,*

(ii) *there is a constant $\bar{\sigma}$ such that*

$$\frac{\min_i \ell_n^i}{\max_i L_n^i} \geq \bar{\sigma} > 0$$

for all n, and

(iii) *the sequence $(\mathbf{x}_n)$ has a cluster point $\mathbf{x}_*$ which satisfies the Lojasiewicz gradient inequality* (2.4).

Then $\mathbf{x}_*$ is the limit point of the sequence and $\nabla f(\mathbf{x}_*) = 0$.

Note that the assumption implies that f is strictly convex on every $E_{\mathbf{x}_n}^i \cap \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_n)\}$ so that the stationarity condition in Theorem 2.2 is indeed satisfied. We also remark that the Gauss-Southwell rule allows for almost the same analysis.

2.4. Example: multi-convex functions in tensor optimization. Of course, every strongly convex function on $X^1 \times X^2 \times \cdots \times X^d$ is trivially block strongly multi-convex. Less trivial examples of such functions naturally arise in tensor optimization by composition with multi-linear functions as follows. Let $J : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}$ be a function of tensors (multi-dimensional arrays) of size $n_1 \times n_2 \times \cdots \times n_d$. Let further $\tau : X^1 \times X^2 \times \cdots \times X^d \to \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ be *multi-linear* with respect to the block variables $x^i \in X^i$. The mapping τ parameterizes certain tensors of low-rank. One then wishes to minimize $f = J \circ \tau$ which is multi-convex. Examples are the problem of the best canonical low-rank tensor approximation (4.1) below, where $J(\mathcal{X}) = \frac{1}{2}\|\mathcal{F} - \mathcal{X}\|^2$ and $\tau(x^1, x^2, \ldots, x^d) = \sum_{k=1}^r x_k^1 \otimes x_k^2 \otimes \cdots \otimes x_k^d$ with $x^i \in \mathbb{R}^{n_i \times r}$. Another example is the approximate solution of linear equations in the TT format [17].

We assume that J is differentiable and uniformly strongly convex with modulus λ, i.e.,

$$J(\mathcal{Y}) \geq J(\mathcal{X}) + \langle \nabla J(\mathcal{X}), \mathcal{Y} - \mathcal{X} \rangle + \frac{\lambda}{2}\|\mathcal{Y} - \mathcal{X}\|^2,$$

and its gradient Lipschitz continuous with Lipschitz constant Λ:

$$\|\nabla J(\mathcal{Y}) - \nabla J(\mathcal{X})\| \leq \Lambda \|\mathcal{Y} - \mathcal{X}\|.$$

Further, we need bounds

$$m_{\mathbf{x}}^i \|y^i\| \leq \|\tau_{\mathbf{x}}^i(y^i)\| \leq M_{\mathbf{x}}^i \|y_i\|, \tag{2.7}$$

where we have denoted by $\tau_{\mathbf{x}}^i$ the restrictions $x^i \mapsto \tau(x_n^1, \ldots, x^i, \ldots, x_n^d)$ of τ on X^i. Since these functions are linear, it holds $\langle \nabla \tau_{\mathbf{x}}^i(x^i), y^i - x^i \rangle = \tau_{\mathbf{x}}^i(y^i) - \tau_{\mathbf{x}}^i(x^i)$. Now we

can estimate

$$f_{\mathbf{x}}^i(y^i) = J(\tau_{\mathbf{x}}^i(y^i))$$

$$\geq J(\tau_{\mathbf{x}}^i(x^i)) + \langle \nabla J(\tau_{\mathbf{x}}^i(x^i)), \tau_{\mathbf{x}}^i(y^i) - \tau_{\mathbf{x}}^i(x^i) \rangle + \frac{\lambda}{2} \|\tau_{\mathbf{x}}^i(y^i) - \tau_{\mathbf{x}}^i(x^i)\|^2$$

$$\geq f^i(x^i) + \langle \nabla f^i(x^i), y^i - x^i \rangle + \frac{\lambda \cdot (m_{\mathbf{x}}^i)^2}{2} \|y^i - x^i\|^2,$$

and similarly,

$$\|\nabla f^i(y^i) - \nabla f^i(x^i)\| \leq \Lambda M_{\mathbf{x}}^i \|y^i - x^i\|.$$

Hence, in this special case Theorem 2.3 reads as follows.

COROLLARY 2.4. *Let $(\mathbf{x}_n)$ be a sequence generated by the MBI method (Algorithm 1) for problem (2.1) with $f = J \circ \tau$ and $S^i = X^i$ for $i = 1, 2, \ldots, d$ (unconstrained case). Assume that*

(i) *there is a constant $\tilde{\sigma}$ such that*

$$\frac{\min_i (m_{\mathbf{x}_n}^i)^2}{\max_i M_n^i} \geq \tilde{\sigma} > 0$$

for all n, and

(ii) *the sequence $(\mathbf{x}_n)$ has a cluster point $\mathbf{x}_*$ which satisfies the Łojasiewicz inequality (2.4).*

Then $\mathbf{x}_*$ is the limit point of the sequence and $\nabla f(\mathbf{x}_*) = 0$.

Since τ is multi-linear, f will be real-analytic, if J is real-analytic. Hence condition (ii) can be replaced by requiring the sequence to be bounded and J to be real-analytic. We should however note that both conditions in the theorem are usual not trivial to validate for most tensor formats. An exception are rank-one tensors for which we present the result in Section 4.2.

3. Local convergence analysis. We return from multi-convex functions to general ones. The local convergence analysis below is classical in the sense that it will invoke second-order assumptions. Therefore we now assume f to be twice differentiable.

3.1. The unconstrained case. The proof of Theorem 2.1 relies on compactness and is nonconstructive. Nothing can be said about the rate of convergence or about how many cluster points there can be. Ideally, a numerical method is also backed by a local convergence theorem. Local convergence analysis is again the simplest in the unconstrained case, i.e., $S^i = X^i$ for $i = 1, 2, \ldots, d$. Being local in nature, the results for the unconstrained case are obviously applicable to stationary points which lie in the interior of $S^1 \times S^2 \times \cdots \times S^d$.

As a BCD method, we expect local linear convergence from the MBI method. As a rule of thumb [26], one proves local linear convergence of a method by investigating its linearization. For instance, it is well known [27] that the linearization of the nonlinear Gauss-Seidel method around a stationary point $\mathbf{x}_*$ (under some smoothness assumptions) is just the linear Gauss-Seidel iteration for solving $Ax = 0$, where $A = \nabla^2 f(\mathbf{x}_*)$ is the Hessian at $\mathbf{x}_*$. Therefore, it is locally linearly convergent if $\nabla^2 f(\mathbf{x}_*)$ is positive definite [37] and this condition cannot be much improved in general.

However, for the MBI method it is not clear whether its linearization is the MBI method applied to solve $Ax = 0$ for the Hessian (which would coincide with the

Gauss-Southwell/greedy multiplicative Schwarz method). The simple reason is that even if $\mathbf{x}_{n+1} = \mathcal{S}(\mathbf{x}_n)$ defines a map, we cannot prove that it is differentiable. Still it should hold (as a general principle, e.g, also for nonlinear Gauss-Southwell) that the local contractivity of a nonlinear relaxation method is governed by its application to the second order model of the function f. This is the statement of Lemma 3.2 below. It then remains to prove the local convergence of MBI for quadratic functions under definiteness assumptions on the Hessian. But such results are known.

Still, we have to enforce (or assume) that the MBI iteration stays local, since otherwise the question of local convergence would not make much sense. Let $\mathbf{x}_*$ be a local minimum of (2.1) with $S^i = X^i$ for $i = 1, 2, \ldots, d$. Let further $A = \nabla^2 f(\mathbf{x}_*)$ denote the Hessian. We assume that A is *positive definite*. Then

$$\|\mathbf{x}\|_A = \langle \mathbf{x}, A\mathbf{x} \rangle^{1/2}$$

defines the *energy norm* of A. By the implicit function theorem, there exists a neighborhood V of $\mathbf{x}_*$ and a possibly smaller neighborhood U such that for every $\mathbf{x}_n \in U$ the problems

$$\nabla_i f(\mathbf{x}) = 0, \quad \mathbf{x} \in V \cap E_{\mathbf{x}_n}^i, \tag{3.1}$$

have unique solutions. For V small enough, they are local minima of the coordinate-restricted problems, since the second directional derivatives are, as the diagonal blocks of A, positive definite. The least trouble we encounter if f is strictly convex and has bounded level sets. Then we can choose V as the whole space.

The solutions of

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in V \cap E_{\mathbf{x}_n} \end{aligned} \tag{3.2}$$

are among the d possible solutions of (3.1). Let $\mathbf{x}_{n+1}^{\mathrm{loc}}$ be one solution of (3.2), the particular choice does not matter. We call $\mathbf{x}_{n+1}^{\mathrm{loc}}$ the *local response* of the MBI method on $\mathbf{x}_n$.

THEOREM 3.1. *Let $\mathbf{x}_*$ be a local minimum of (2.1) with $S^i = X^i$ for $i = 1, 2, \ldots, d$. Assume $A = \nabla^2 f(\mathbf{x}_*)$ is positive definite. For $\mathbf{x}_0$ close enough to $\mathbf{x}_*$, a sequence $(\mathbf{x}_n^{\mathrm{loc}})$ of local MBI responses exists. Any such sequence converges at least Q-linearly to $\mathbf{x}_*$ in the energy norm, i.e., there exists $0 \leq q < 1$ such that*

$$\|\mathbf{x}_{n+1}^{\mathrm{loc}} - \mathbf{x}_*\|_A \leq q\|\mathbf{x}_n^{\mathrm{loc}} - \mathbf{x}_*\|_A.$$

To prove the theorem, we continue now, without losing generality, by assuming $\mathbf{x}_* = 0$ and $f(\mathbf{x}_*) = 0$. By Taylor expansion, we then have

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_A^2 + \frac{1}{2}\eta(\mathbf{x})\|\mathbf{x}\|_A^2,$$

where η is some function with $\eta(\mathbf{x}) \to 0$ for $\mathbf{x} \to 0$. The comparison of the local MBI method with any other coordinate descent method for A is based on the following lemma.

LEMMA 3.2. *Let $A = \nabla^2 f(\mathbf{x}_*)$ be positive definite. For every $\epsilon > 0$ there exists a $\delta > 0$, such that $\|\mathbf{x}\|_A, \|\mathbf{y}\|_A \leq \delta$ and $f(\mathbf{x}) \leq f(\mathbf{y})$ implies $\|\mathbf{x}\|_A \leq (1 + \epsilon)\|\mathbf{y}\|_A$.*

Proof. For any $\epsilon > 0$, let $\epsilon_0 := \min\{1/2, \epsilon/2\} > 0$, and there exists $\delta > 0$, such that $\|\mathbf{x}\|_A \leq \delta$ implies $|\eta(\mathbf{x})| \leq \epsilon_0$. When $\|\mathbf{x}\|_A, \|\mathbf{y}\|_A \leq \delta$ and $f(\mathbf{x}) \geq f(\mathbf{y})$, as $|\eta(\mathbf{x})|, |\eta(\mathbf{y})| \leq \epsilon_0$, we have that

$$\frac{1}{2}(1 - \epsilon_0)\|\mathbf{x}\|_A^2 \leq f(\mathbf{x}) \leq f(\mathbf{y}) \leq \frac{1}{2}(1 + \epsilon_0)\|\mathbf{y}\|_A^2,$$

implying

$$\frac{\|\mathbf{x}\|_A^2}{\|\mathbf{y}\|_A^2} \leq \frac{1 + \epsilon_0}{1 - \epsilon_0} = 1 + \frac{2\epsilon_0}{1 - \epsilon_0} \leq 1 + \frac{2 \cdot \epsilon/2}{1 - 1/2} = 1 + 2\epsilon \leq (1 + \epsilon)^2. \qquad \square$$

Proof of Theorem 3.1. Recall that $\mathbf{x}_* = 0$ and $f(\mathbf{x}_*) = 0$. Given $\mathbf{x}_n$, let $\mathbf{y}_{n+1}$ be the result of some coordinate descent method applied to $\mathbf{x}_n$ for minimizing the convex function $\mathbf{x} \mapsto \|\mathbf{x}\|_A^2$, i.e., solving $A\mathbf{x} = 0$, which is known to converge Q-linearly in the energy norm, e.g. the Gauss-Southwell method [24, 12].

Let V and $U \subseteq V$ be balls in the energy norm such that for $\mathbf{x}_n \in U$ both $\mathbf{x}_{n+1}^{\text{loc}}$ and $\mathbf{y}_{n+1}$ are defined, belong to $V \cap (\bigcup_i E_{\mathbf{x}_n}^i)$, and satisfy

$$\|\mathbf{y}_{n+1}\|_A \leq \tilde{q}\|\mathbf{x}_n\|_A$$

for some $0 \leq \tilde{q} < 1$ related to the chosen method. By definition of $\mathbf{x}_{n+1}^{\text{loc}}$ it holds $f(\mathbf{x}_{n+1}^{\text{loc}}) \leq f(\mathbf{y}_{n+1})$. Let $\epsilon > 0$ be such small that $q = (1 + \epsilon)\tilde{q} < 1$. Then we can make V smaller, so that Lemma 3.2 becomes applicable, i.e.,

$$\|\mathbf{x}_{n+1}^{\text{loc}}\|_A \leq (1 + \epsilon)\|\mathbf{y}_{n+1}\|_A \leq q\|\mathbf{x}_n\|_A.$$

The rest follows by induction. $\square$

3.2. Spherical constraints. The local convergence result obtained above is easily applicable to constrained problems when the S^i are smooth manifolds. The idea is to consider a local diffeomorphism to the tangent space to get into the locally unconstrained setting. We present the idea for the important class of spherically constrained problems,

$$\begin{aligned} \min \quad & f(x^1, x^2, \ldots, x^d) \\ \text{s.t.} \quad & x^i \in \mathbb{R}^{n_i}, \; \|x^i\| = 1, \; i = 1, 2, \ldots, d. \end{aligned} \tag{3.3}$$

Let now $\mathbf{x}_*$ denote a stationary point of the MBI method for (3.3), i.e., a stationary point of (2.2) with $S^i = \{x^i \in X^i : \|x^i\| = 1\}$. If we set $Y^i = \{x_*^i\}^\perp$ and

$$h(y^1, y^2, \ldots, y^d) = \left(\frac{x_*^1 + y^1}{\|x_*^1 + y^1\|}, \frac{x_*^2 + y^2}{\|x_*^2 + y^2\|}, \ldots, \frac{x_*^d + y^d}{\|x_*^d + y^d\|} \right),$$

then all directional derivatives of h at 0 will be zero, i.e., 0 is a critical point of the function

$$g : Y^1 \times Y^2 \times \cdots \times Y^d \to \mathbb{R}, \; \mathbf{y} \mapsto f(h(\mathbf{y})),$$

and moreover, since $\mathbf{y} \to h(\mathbf{y})$ is a local diffeomorphism to a neighborhood of $\mathbf{x}_*$ on the spheres, a *locally* responding MBI method applied to (3.3) in the neighborhood of $\mathbf{x}_*$ is in one-to-one correspondence to a local MBI method applied to g in some neighborhood of 0, in the sense that

$$\mathbf{x}_n^{\text{loc}} = h(\mathbf{y}_n^{\text{loc}})$$

would be a feasible choice when $\mathbf{x}_0 = h(\mathbf{y}_0)$.[1] If we show local linear convergence of $\mathbf{y}_n^{\mathrm{loc}} \to 0$, then also $\mathbf{x}_n^{\mathrm{loc}} \to \mathbf{x}_*$ at least R-linearly. Hence Theorem 3.1 proves

THEOREM 3.3. *Assume*

$$\nabla^2 g(0) \text{ is positive definite on } Y^1 \times Y^2 \times \cdots \times Y^d. \tag{3.4}$$

Then a local MBI iteration $(\mathbf{x}_n^{\mathrm{loc}})$ *for* (3.3) *exists and is R-linearly convergent to* $\mathbf{x}_*$.

Condition (3.4) is reasonable as a general assumption when 0 is a local minimum of g. But it is typically difficult to verify for particular applications. Let us derive the explicit formula for $\nabla^2 g(0)$. By the chain rule,

$$\langle \mathbf{y}, \nabla^2 g(0)\mathbf{y} \rangle = \langle \nabla h(0)\mathbf{y}, \nabla^2 f(\mathbf{x}_*) \cdot \nabla h(0)\mathbf{y} \rangle + \nabla f(\mathbf{x}_*) \cdot \nabla^2 h(0)[\mathbf{y}, \mathbf{y}].$$

Using the Taylor expansion $(1 + t)^{-1/2} = 1 - t/2 + O(|t|^2)$ we obtain for $\mathbf{y} \in Y^1 \times Y^2 \times \cdots \times Y^d$, by using orthogonality,

$$\begin{aligned}
(x_*^i + y^i)/\|x_*^i + y^i\| &= (x_*^i + y^i)(1 + \|y^i\|^2)^{-1/2} \\
&= (x_*^i + y^i)(1 - \|y^i\|^2/2 + O(\|y^i\|^4)) \\
&= x_*^i + y^i - x_*^i \|y^i\|^2/2 + O(\|y^i\|^3).
\end{aligned}$$

Hence,

$$\nabla h(0)\mathbf{y} = \mathbf{y}$$

and

$$\begin{aligned}
\nabla^2 h(0)[\mathbf{y}, \mathbf{y}] &= -(x_*^1 \|y^1\|^2, x_*^2 \|y^2\|^2, \dots, x_*^d \|y^d\|^2) \\
&= -\mathrm{diag}(\|y^1\|^2, \|y^2\|^2, \dots, \|y^d\|^2)\mathbf{x}_*.
\end{aligned}$$

Therefore,

$$\langle \mathbf{y}, \nabla^2 g(0)\mathbf{y} \rangle = \langle \mathbf{y}, \nabla^2 f(\mathbf{x}_*)\mathbf{y} \rangle - \nabla f(\mathbf{x}_*) \cdot \mathrm{diag}(\|y^1\|^2, \|y^2\|^2, \dots, \|y^d\|^2)\mathbf{x}_*. \tag{3.5}$$

It remains to mention that using different local diffeomorphisms h from the sphere to its tangent plane, such as stereographic or orthogonal projection leads to slightly different convergence criteria. However, we found (3.5) reasonably simple.

3.3. Example: The interpretation of the power method as MBI. As a model problem for the reasoning so far, consider the task

$$\begin{aligned}
\max \quad & \langle x^1, Mx^2 \rangle \\
\text{s.t.} \quad & \|x^1\| = \|x^2\| = 1
\end{aligned}$$

for calculating the largest eigenvalue $\lambda_{\max}$ of a positive definite matrix M. The global maximum is attained for $x_*^1 = x_*^2$ being a normed eigenvector for $\lambda_{\max}$. Denote $f(x^1, x^2) = -\langle x^1, Mx^2 \rangle$, then we have

$$\nabla f(\mathbf{x}_*) \cdot \mathbf{z} = -\langle z^1, Mx_*^2 \rangle - \langle x_*^1, Mz^2 \rangle,$$

so in particular

$$\nabla f(\mathbf{x}_*) \cdot \mathrm{diag}(\|y^1\|^2, \dots, \|y^d\|^2)\mathbf{x}_* = -\lambda_{\max}(\|y^1\|^2 + \|y^2\|^2) = -\lambda_{\max}\|\mathbf{y}\|^2. \tag{3.6}$$

---

[1] As above, our assumption (3.4) on the Hessian will imply that this choice is locally unique when using the same coordinate for each iteration of $(\mathbf{x}_n^{\mathrm{loc}})$ and $(\mathbf{y}_n^{\mathrm{loc}})$.

Further,

$$\langle \mathbf{y}, \nabla^2 f(\mathbf{x}_*)\mathbf{y}\rangle = -2\langle y^1, M y^2\rangle.$$

Let $\lambda' \leq \lambda_{\max}$ denote the second largest eigenvalue (counting multiplicity) of M. Then, for $y^1, y^2 \in \{x_*^1\}^\perp$ it holds

$$-\langle \mathbf{y}, \nabla^2 f(\mathbf{x}_*)\mathbf{y}\rangle = 2\langle y^1, M y^2\rangle \leq 2\lambda'\|y^1\| \cdot \|y^2\| \leq \lambda'(\|y^1\|^2 + \|y^2\|^2) = \lambda'\|\mathbf{y}\|^2. \quad (3.7)$$

This estimate is sharp (consider $y^1 = y^2$ being the eigenvector for λ'). Inserting (3.6) and (3.7) into (3.5) we obtain

$$\langle \mathbf{y}, \nabla^2 g(0)\mathbf{y}\rangle \geq (\lambda_{\max} - \lambda')\|\mathbf{y}\|^2.$$

In the case $\lambda' < \lambda_{\max}$, i.e., the largest eigenvalue has multiplicity one, the Hessian $\nabla^2 g(0)$ is positive definite at the eigenvector for the largest eigenvalue. Otherwise we could have some trouble which might be avoided by considering as Y^i only the eigenspaces of strictly smaller eigenvalues.

Let us see how the MBI method looks like for this problem. Starting from $\mathbf{x}_0 = (\xi, \eta)$ and assuming that $\xi^T M^2 \xi > \eta^T M^2 \eta$ it is easy to see that the MBI method produces the normalized version of the sequence

$$\mathbf{x}^1 = (\xi, M\xi), \quad \mathbf{x}^2 = (M^2\xi, M\xi), \quad \mathbf{x}^3 = (M^2\xi, M^3\xi), \quad \dots$$

This is the classical power method in each component! Under the assumption that $\lambda_{\max}$ has multiplicity one, both components tend to the corresponding eigenvector $x_*^1 = x_*^2$, as long as $\mathbf{x}_0 \notin Y^1 \times \cdots \times Y^d$, which is almost surely the case. If it is not the case, we obtain convergence to some other eigenvector under the assumption that *all* eigenvalues have multiplicity one. The analysis can then be restricted to the smallest invariant subspace of M the starting point lies in. Hence, in this model application, the subsequential global convergence of MBI is in fact global (if all eigenvalues are distinct), and the convergence is asymptotically linear (which, of course, is known for the power method). We also see that a limit point might only be a saddle point of the initial problem, although here this scenario has probability zero.

In summary, for the matrix eigenvalue model, MBI is no new method, but it is interesting to see that the power method is a particular instance of it. In any case, it is a very illustrative, convincing example for our general reasoning and gives hope for "similar" applications such as the calculation of the maximum eigenvalue of tensors, which is the topic of the section 4.

4. Convergence results for best rank-one approximation of tensors. The rank-r decomposition/approximation

$$\begin{array}{ll} \min & \frac{1}{2}\left\|\mathcal{F} - \sum_{k=1}^r x_k^1 \otimes x_k^2 \otimes \cdots \otimes x_k^d\right\|_F^2 \\ \text{s.t.} & x_k^i \in \mathbb{R}^{n_i}, \ k = 1, 2, \dots, r, \ i = 1, 2, \dots, d \end{array} \quad (4.1)$$

of a tensor $\mathcal{F}$ (in the Frobenius norm) is an important task in some branches of statistics such as Psychometrics or Chemometrics, and also in signal processing, where it serves as a model known as PARAFAC or CANDECOMP to analyze multidimensional data; see [18] and the references therein. The task also naturally occurs in high-dimensional tensor calculus [14].

A popular method to solve (4.1) is the cyclic coordinate descent method with respect to the block variables $x^i = (x_1^i, x_2^i, \dots, x_r^i) \in \mathbb{R}^{n_i \times r}$, which in this context

is called ALS algorithm. It was originally proposed by Carroll and Chang [6] and Harshman [15]. However, as already mentioned above, the ALS method is not even guaranteed to globally converge to a stationary point, only to a solution where the objective function ceases to decrease; see e.g. numerical observed cases in [10]. Local convergence of ALS has been proved in [39] for rank $r = 1$ and in [35] for the general case under natural, yet controversial, assumptions on the Hessian of (4.2), which never is positive definite due to scaling indeterminacy of the tensor product.

4.1. Best rank-one approximation. What makes (4.1) a very difficult task besides its high nonlinearity is its possible ill-posedness for ranks r larger than one [8]. Global minima might not exist, and minimizing sequences might not stay bounded. These problems do not occur for the well behaved rank-one approximation problem

$$\min_{\mathbf{x} \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_d}} \quad \tfrac{1}{2} \left\| \mathcal{F} - x^1 \otimes x^2 \otimes \cdots \otimes x^d \right\|_F^2 . \tag{4.2}$$

We note that if $\mathbf{x}_*$ is a critical point of (4.2) and $x_*^1 \otimes x_*^2 \otimes \cdots \otimes x_*^d \neq 0$, then $x_*^1 \otimes x_*^2 \otimes \cdots \otimes x_*^d$ indeed is a critical point of $\mathcal{X} \mapsto \tfrac{1}{2} \| \mathcal{F} - \mathcal{X} \|_F^2$ on the smooth submanifold of rank-one tensors. See, e.g. [36, p. 155] for further explanation.

One point we want to focus on is the equivalence of (4.2) to the maximization of homogenous polynomials. Specifically, given $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, its corresponding multi-linear form F is defined as

$$F(x^1, x^2, \ldots, x^d) := \sum_{i_1=1}^{n_1} \sum_{i_1=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} \mathcal{F}_{i_1 i_2 \ldots i_d} x_{i_1}^1 x_{i_2}^2 \ldots x_{i_d}^d .$$

Note that $F(x^1, x^2, \ldots, x^d)$ is the Frobenius inner product of $\mathcal{F}$ and the rank-one tensor $x^1 \otimes x^2 \otimes \cdots \otimes x^d$. It is therefore clear that, up to scaling, (4.2) is equivalent to

$$\begin{aligned} \max \quad & F(x^1, x^2, \ldots, x^d) \\ \text{s.t.} \quad & x^i \in \mathbb{R}^{n_i}, \ \|x^i\| = 1, \ i = 1, 2, \ldots, d, \end{aligned} \tag{4.3}$$

which is an instance of (3.3). The possibility to reformulate (4.2) into (4.3) distinguishes the rank-one approximation from the higher rank case in many respects.

The equivalence (up to scaling) between (4.2) and (4.3) also holds on the level of a single block update. Given $\mathbf{x}_n$ such that $\hat{\mathbf{x}}_n = (x_n^1 / \|x_n^1\|, x_n^2 / \|x_n^2\|, \ldots, x_n^d / \|x_n^d\|)$ is defined, $\hat{x}_{n+1}^i$ is a possible MBI response for (4.3) starting from $\hat{\mathbf{x}}_n$ if and only if $x_{n+1}^i = F(x_n^1, \ldots, \hat{x}_{n+1}^i, \ldots, x_n^d) \hat{x}_{n+1}^i$ is a valid MBI response for (4.2) from $\mathbf{x}_n$. Conversely, if $(\mathbf{x}_n)$ is an MBI sequence for (4.2) with $\tau_1(\mathbf{x}_n) \neq 0$ for all n, then $(x_n^1 / \|x_n^1\|, x_n^2 / \|x_n^2\|, \ldots, x_n^d / \|x_n^d\|)$ is an MBI sequence for (4.3) when using the starting point $(x_0^1 / \|x_0^1\|, x_0^2 / \|x_0^2\|, \ldots, x_0^d / \|x_0^d\|)$.

Multi-linear form optimization over spheres (4.3) is a fundamental model in polynomial optimization. Recently, this model is frequently used as a relaxation for homogeneous polynomial optimizations; see e.g. [16]. Problem (4.3) is also known as the largest singular value of a high order tensor [21, 22]. Variants of problem (4.2) are for instance also important because it is the key step in some greedy tensor approximation strategies; see [5, 9] and the references therein. It might be worth mentioning that also in the case of a super-symmetric tensor the best rank-one approximation is symmetric [7, 40] (i.e., $x^1 = x^2 = \cdots = x^d$ in (4.3)), so there is no need to impose this; unfortunately the symmetry does not hold for local minima.

4.2. Global convergence. We prove the convergence of the MBI method for (4.2). As previously remarked this implies an analogous result for the MBI method applied to (4.3). The advantage of (4.2) is that it is of the form described in section 2.4 with $J(\mathcal{X}) = \frac{1}{2}\|\mathcal{F} - \mathcal{X}\|_F^2$ and

$$\tau = \tau_1 : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_d} \to \mathbb{R}, \ \tau_1(x^1, x^2, \ldots, x^d) = x^1 \otimes x^2 \otimes \cdots \otimes x^d.$$

We hence aim to apply Corollary 2.4.

The following lemma basically shows that the representation of the rank-one tensors in an MBI method does not become arbitrarily degenerate, and essentially allows to bound the constants in (2.7). This useful fact applies to any BCD method for solving (4.2), e.g. Gauss-Seidel (ALS).

LEMMA 4.1. *Let* $(\mathbf{x}_n)$ *be an MBI sequence for problem* (4.2) *with initial guess* $\mathbf{x}_0$ *and* $\tau_1(\mathbf{x}_1) \neq 0$.[2] *Then it holds for all* $n \geq 1$ *that*

$$\prod_{i=1}^{n} \|x_n^i\| = \|\tau_1(\mathbf{x}_n)\|_F \leq \|\mathcal{F}\|_F, \tag{4.4}$$

and

$$\|x_{n+1}^i\| \geq \|x_n^i\|, \quad i = 1, 2, \ldots, d. \tag{4.5}$$

As a consequence,

$$0 < \|x_1^i\| \leq \|x_n^i\| \leq \|\mathcal{F}\|_F \prod_{j \neq i} \|x_1^j\|^{-1}, \quad i = 1, 2, \ldots, d.$$

Proof. By homogeneity of the tensor product, it is clear that, for $n \geq 1$, $\tau_1(\mathbf{x}_n)$ is the Euclidian (Frobenius) best approximation to $\mathcal{F}$ on $\operatorname{span} \tau_1(\mathbf{x}_n)$, and therefore satisfies

$$\|\tau_1(\mathbf{x}_n)\|_F^2 + \|\mathcal{F} - \tau_1(\mathbf{x}_n)\|_F^2 = \|\mathcal{F}\|_F^2.$$

This immediately proves (4.4), but moreover, since $\|\mathcal{F} - \tau_1(\mathbf{x}_n)\|_F$ does not increase with n, it inductively also shows

$$0 < \prod_{i=1}^{n} \|x_n^i\| = \|\tau_1(\mathbf{x}_n)\|_F \leq \|\tau_1(\mathbf{x}_{n+1})\|_F = \prod_{i=1}^{n} \|x_{n+1}^i\|$$

for all $n \geq 1$. Since except for possibly one index all the blocks x_n^i and x_{n+1}^i are equal, (4.5) follows. ☐

Lemma 4.1 not only tells us that $(\mathbf{x}_n)$ is bounded in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_d}$, but also that there exist constants $0 < m < M$ independent of i and n such that

$$m\|x_n^i\| \leq \|\tau_1(\mathbf{x}_n)\|_F = \prod_{j=1}^{n} \|x_n^j\| \leq M\|x_n^i\|$$

for all i and n. The convergence of the sequence to a critical point now follows from Corollary 2.4.

THEOREM 4.2. *If* $\tau(\mathbf{x}_1) \neq 0$, *the sequence of iterates of an MBI method applied to the rank-one approximation problem (either* (4.2) *or* (4.3)*) converges to a critical point of the problem.*

---

[2]Note that $\tau_1(\mathbf{x}_0)$ can be orthogonal to $\mathcal{F}$ in the Frobenius inner product as long as there exists one block i for which $\operatorname{span}(x_0^1) \otimes \cdots \otimes \mathbb{R}^{n_i} \otimes \cdots \otimes \operatorname{span}(x_0^d)$ is not an orthogonal subspace to $\mathcal{F}$.

4.3. Local linear convergence for perturbed rank-one tensors. We continue to use the notation τ_1 for the rank-one map, but focus the anlysis on the equivalent maximization problem (4.3) in order to apply Theorem 3.3.[3] It is easy to see (and known) that the MBI method will find the optimal solution after d steps if

$$\mathcal{F}_* = \lambda_* \tau_1(\mathbf{x}_*) \neq 0, \quad \|x_*^i\| = 1, \quad \lambda_* > 0,$$

itself is a rank-one tensor, and if the starting point $\mathbf{x}_0$ satisfies $F(\mathbf{x}_0) = \langle \mathcal{F}, \tau_1(\mathbf{x}_0)\rangle_F \neq 0$, which is almost surely the case when choosing it randomly. In fact, in that case

$$F_*(x^1, x^2, \ldots, x^d) = \lambda_* \langle \tau_1(\mathbf{x}_*), \tau_1(\mathbf{x})\rangle_F = \lambda_* \prod_{i=1}^{d} \langle x_*^i, x^i\rangle, \tag{4.6}$$

and the unique solution of a subproblem

$$\min_{x^j \in \mathbb{R}^{n_j},\ \|x^j\|=1} \lambda_* \prod_{i=1}^{d} \langle x_*^i, x^i\rangle$$

is $x^j = \pm x_*^j$ as long as $\prod_{i \neq j} \langle x_*^i, x^i\rangle \neq 0$.

In other words, given $F(\mathbf{x}_0)$, $\mathbf{x}_* = (x_*^1, x_*^2, \ldots, x_*^d)$ and some of its sign-flipped variants are the only cluster points of the MBI method satisfying $F_*(\mathbf{x}_*) > 0$. Furthermore, using multi-linearity and (4.6),

$$\langle \mathbf{y}, \nabla^2 F_*(\mathbf{x}_*)\mathbf{y}\rangle = \lambda_* \sum_{i \neq j} \langle x_*^i, y^i\rangle \langle x_*^j, y^j\rangle = 0 \tag{4.7}$$

for $\mathbf{y} \in \{x_*^1\}^\perp \times \{x_*^2\}^\perp \times \cdots \times \{x_*^d\}^\perp$, and

$$\nabla F_*(\mathbf{x}_*) \cdot \mathrm{diag}(\|y^1\|^2, \|y^2\|^2, \ldots, \|y^d\|^2)\mathbf{x}_* = F(\mathbf{x}_*)\|\mathbf{y}\|^2 = \lambda_*\|\mathbf{y}\|^2. \tag{4.8}$$

Hence we see from (3.5) that the assumption (3.4) of Theorem 3.3 is formally fulfilled. This of course holds for all sign flipped solutions.

By continuity, we expect the same behaviour if $\mathcal{F}$ does not lie too far away from the set of rank-one decomposable tensors. However, this statement will be rather weak. We first note that also in the case of general $\mathcal{F}$ the corresponding restricted subsolutions (2.3) are unique, if $F(\mathbf{x}_n) \neq 0$, since they are minimization problems of then non-vanishing linear functionals on spheres.

THEOREM 4.3. *For every $\delta > 0$ there exists an $\epsilon > 0$ such that if the distance of $\mathcal{F}/\|\mathcal{F}\|_F$ to the set of rank-one tensors is smaller than ϵ, and if the initial guess satisfies $|\langle \mathcal{F}, x_0^1 \otimes x_0^2 \otimes \cdots \otimes x_0^d\rangle_F| \geq \delta\|\tau_1(\mathbf{x}_0)\|_F$, the sequence of iterates of an MBI method applied to the rank-one approximation problem (either (4.2) or (4.3)) converges R-linearly to a critical point of the problem.*

In fact, it would be possible to prove Q-linear convergence using the results from [35] instead of Theorem 3.3.

Proof. We sketch the proof leaving out some details. Assume the claim is wrong. Then there exist $\delta > 0$ and a sequence $(\mathcal{F}^n)$ of tensors with norm $\|\mathcal{F}^n\| = 1$, converging to some $\mathcal{F}^* = \tau_1(\mathbf{x}^*)$ with $\|(x^*)^i\| = 1$ (the set of rank-one tensors is closed), such that for all n there exists a starting point $\mathbf{x}_0^n$ (consisting of normalized blocks) for which the

---

[3]Due to the scaling indeterminacy of the rank-one representation, Theorem 3.1 is not directly applicable to the unconstrained version (4.2), cf. the discussion in [35].

MBI method applied to (4.3) with F_n converges to some $\mathbf{x}_*^n$, but not R-linearly (the convergence follows from Theorem 4.2). We can assume that the sequence of starting points converges to some $\mathbf{x}_0^*$. Using that $F(\mathbf{x})$ depends continuously on the generating tensor $\mathcal{F}$, it is fairly easy to see that $\mathbf{x}_0^*$ has to be stationary points of the MBI method for F_*. Furthermore, $F_*(\mathbf{x}_0^*) \geq \delta$ by the monotonicity of MBI. As just discussed, for exact decompositions of $\mathcal{F}^*$ the only cluster points of MBI having this property are $\mathbf{x}^*$ itself and its sign-flipped variants, and they satisfy the main assumption (3.4) of Theorem 3.3 that the corresponding Hessian $\nabla^2 g_{\mathcal{F}^*,\mathbf{x}^*}(0)$ is positive definite on $\{(x^*)^1\}^\perp \times \{(x^*)^2\}^\perp \times \cdots \times \{(x^*)^d\}^\perp$. One now has to argue by continuity that $\nabla^2 g_{\mathcal{F}^n,\mathbf{x}^n}(0)$ will be positive definite on $\{(x^n)^1\}^\perp \times \{(x^n)^2\}^\perp \times \cdots \times \{(x^n)^d\}^\perp$ for large enough n, which contradicts Theorem 3.3. $\square$

In a way, Theorem 4.3 is an odd statement. Although it is not vacuous (unless δ is not implausibly large), it gives no concrete information. Namely, it is not very meaningful to guarantee the quality of a starting point (by choosing δ) before knowing the actual function (whose "choice" is restricted by ϵ). The theorem merely gives an idea that small perturbations of rank-one tensors lead to linear convergence rate if the starting point is not too bad. Indeed, it states that there is at least one pair (δ, ϵ) with the desired property. Of course, ideally, one would like to see δ depend on ϵ, i.e., interchange the roles of both, and state a theorem like "for all $\epsilon > 0$ (perhaps smaller than some constant) there exists $\delta > 0$ such that $\ldots$". This can be achieved when we only consider global minima, as we show next.

4.4. A precise bound for optimal solutions. Let $\mathcal{F}_* = \lambda_* \tau_1(\mathbf{x}_*)$ with $\|x_*^i\| = 1$ for $i = 1, 2, \ldots, d$, be a *best* rank-one approximation to $\mathcal{F}$, such that

$$\mathcal{F} = \mathcal{F}_* + \mathcal{E},$$

where $\|\mathcal{E}\|$ is the (Frobenius) distance of $\mathcal{F}$ to the set of rank-one tensors. At the possible solution $\mathbf{x}_*$ of (4.3), (3.5) reads

$$\langle \mathbf{y}, \nabla^2 g(0)\mathbf{y} \rangle = 2 \sum_{i<j} F(x_*^1, \ldots, y^i, \ldots, y^j, \ldots, x_*^d) - F(\mathbf{x}_*)\|\mathbf{y}\|^2.$$

We again have to show that this expression is negative for nonzero $\mathbf{y} \in \{x_*^1\}^\perp \times \{x_*^2\}^\perp \times \cdots \times \{x_*^d\}^\perp$. First note that (4.7) is again valid. Next, by optimality, $\mathcal{E}$ and $\tau_1(\mathbf{x}_*) = \lambda_*^{-1}\mathcal{F}_*$ have to be orthogonal in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, i.e., $E(\mathbf{x}_*) = 0$, where E denotes the multi-linear form associated to $\mathcal{E}$. Therefore, with (4.8),

$$\langle \mathbf{y}, \nabla^2 g(0)\mathbf{y} \rangle = 2 \sum_{i<j} E(x_*^1, \ldots, y^i, \ldots, y^j, \ldots, x_*^d) - \lambda_*\|\mathbf{y}\|^2. \tag{4.9}$$

By (4.6), the rank-one tensors $x_*^i \otimes \cdots \otimes y^i \otimes \cdots \otimes y^j \otimes \cdots \otimes x_*^d$ are pairwise orthogonal for all $i < j$ in the Frobenius inner product. Using that $E(x_*^1, \ldots, y^i, \ldots, y^j, \ldots, x_*^d)$

is the inner product with $\mathcal{E}$, Pythagoras and Cauchy-Schwarz yield

$$\left(\sum_{i<j} E(x_*^1, \ldots, y^i, \ldots, y^j, \ldots, x_*^d)\right)^2 = \left\langle \mathcal{E}, \sum_{i<j} x_*^1 \otimes \cdots \otimes y^i \otimes \cdots \otimes y^j \otimes \cdots \otimes x_*^d \right\rangle_F^2$$

$$\leq \|\mathcal{E}\|_F^2 \left\|\sum_{i<j} x_*^1 \otimes \cdots \otimes y^i \otimes \cdots \otimes y^j \otimes \cdots \otimes x_*^d\right\|_F^2$$

$$= \|\mathcal{E}\|_F^2 \sum_{i<j} \left\|x_*^1 \otimes \cdots \otimes y^i \otimes \cdots \otimes y^j \otimes \cdots \otimes x_*^d\right\|_F^2$$

$$= \|\mathcal{E}\|_F^2 \sum_{i<j} \|y^i\|^2 \|y^j\|^2$$

where the third line is due to (4.6). For a fixed value of $\|\mathbf{y}\|^2 = \sum_{i=1}^d \|y^i\|^2$ the last sum is minimal if every y^i has the same norm $\|\mathbf{y}\|/\sqrt{d}$. Hence

$$\sum_{i<j} E(x_*^1, \ldots, y^i, \ldots, y^j, \ldots, x_*^d) \leq \sqrt{\frac{d-1}{2d}} \|\mathcal{E}\|_F \|\mathbf{y}\|^2.$$

Inserting this in (4.9) gives the estimate

$$\langle \mathbf{y}, \nabla^2 g(0)\mathbf{y} \rangle \leq \left(\sqrt{\frac{2d-2}{d}} \|\mathcal{E}\|_F - \lambda_*\right) \|\mathbf{y}\|^2.$$

Thus, we end up with the condition

$$\sqrt{\frac{2d-2}{d}} \|\mathcal{E}\|_F < \lambda_*$$

for (3.5) to hold, which, using the optimality condition $\lambda_*^2 = \|\mathcal{F}\|_F^2 - \|\mathcal{E}\|_F^2$, is equivalent to

$$\|\mathcal{E}\|_F < \sqrt{\frac{d}{3d-2}} \|\mathcal{F}\|_F. \tag{4.10}$$

The same bound has been obtained, in a slightly different way, in [35] to ensure local convergence of the alternating least squares algorithm (ALS).

Theorem 3.3 can now be reformulated into the following result.

THEOREM 4.4 (cf. [35]). *Let $\mathcal{F}_* = \lambda_* x_*^1 \otimes x_*^2 \otimes \cdots \otimes x_*^d$ be a best rank-one approximation of $\mathcal{F} = \mathcal{F}_* + \mathcal{E}$. Then, if (4.10) holds, also (3.4) holds at $\mathbf{x}_*$. Consequently, the MBI method is linearly convergent in a neighborhood of $\mathbf{x}_*$.*

We emphasize that this is actually a condition on the distance of a tensor $\mathcal{F}$ to the set of rank-one tensors. Although it looks surprisingly soft, and essentially independent of the order d, one should be aware that its restrictiveness lies in the sizes $n_1, n_2, \ldots, n_d$ of the tensors under consideration. The larger they are, the more structured tensors satisfying (4.10) have to be.[4]

---

[4]Consider hyper-diagonal tensors of size n^d with entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ on the diagonal. The distance to the set of rank-one tensors in the Frobenius norm is $\|\mathcal{E}\| = (|\lambda_2|^2 + \cdots + |\lambda_n|^2)^{1/2}$, while the norm is $\|\mathcal{F}\| = (|\lambda_1|^2 + \|\mathcal{E}\|^2)^{1/2}$.
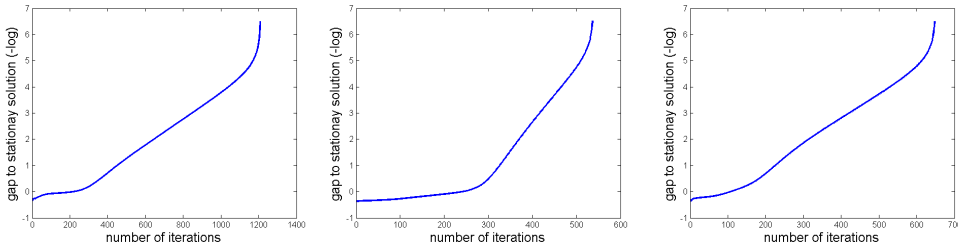
FIG. 5.1. *Three instances of* (4.3) *for* $d = 3$ *and* $n_i = 50$ ($i = 1, 2, 3$) *by MBI*
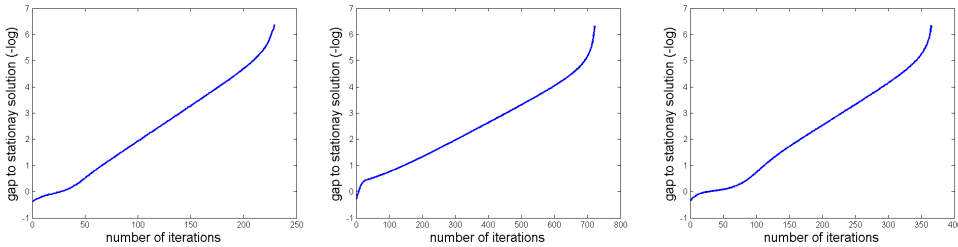

FIG. 5.2. *Three instances of* (4.3) *for* $d = 4$ *and* $n_i = 10$ ($i = 1, 2, 3, 4$) *by MBI*

5. Numerical experiments. In the final section, let us present some numerical experiments of the MBI method, to make the linear convergence being visible. We focus on the problem of finding the largest singular value of a high-order tensor, that is, given a tensor $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ with its associated multi-linear form F, on the optimization model (4.3).

All the computations are conducted in an Intel Core2 Quad CPU 2.70GHz computer with 4GB RAM. The supporting software is MATLAB 7.12.0 (R2011a) as a platform. We use MATLAB Tensor Toolbox Version 2.5 [3] whenever tensor operations are called. We test instances of (4.3) with $d = 3$ or $d = 4$, and the data are randomly generated. In particular, the components of $\mathcal{F}$ follow i.i.d. standard Gaussian. The initial solution $\mathbf{x}_0 = (x_0^1, x_0^2, \ldots, x_0^d)$ are also randomly generated, with each block being drawn from uniform distribution on the unit sphere.

We follow the standard MBI procedure (Algorithm 1) to solve (4.3) and output a sequence $(\mathbf{x}_n)$ of approximate solutions, which itself always (by observation) converges to a stationary point $\mathbf{x}_*$. Here the stopping criterion is both $\|\mathbf{x}_n - \mathbf{x}_{n-1}\| \leq 10^{-8}$ and $F(\mathbf{x}_n) - F(\mathbf{x}_{n-1}) \leq 10^{-12}$. In the following, we regard $\mathbf{x}_*$ as a solution to (4.3). This approach is common since it lacks alternatives: only for trivially structured tensors a best rank-one approximation can be exactly determined.

The sequence $(-\log_{10} \|\mathbf{x}_n - \mathbf{x}_*\|)$ is plotted in Figures 5.1 and 5.2. Three randomly generated instances (random $\mathcal{F}$ and $\mathbf{x}_0$) are presented in each figure, which clearly show the linear convergence of the MBI method. In fact, these curves appear to be convex when the iterations are near the stationary points, suggesting that the local rate of convergence may even be super-linear, at least for very small perturbations. Clearly, since F is locally Lipschitz, the plots for the logarithm of $F(\mathbf{x}_*) - F(\mathbf{x}_n)$ would look almost the same and are not of importance here.

5.1. Rank-one approximation of perturbed rank-one tensors. As discussed in Section 4.3, if the tensor $\mathcal{F}$ does not lie too far away from the set of rank-one

$\|\mathcal{E}\|/\|\mathcal{F}\|$	≤ 0.1	0.2	0.3	0.4	0.5	0.6	$\sqrt{\frac{d}{3d-2}} \approx 0.632$	0.7	0.8	0.9
Recovery (%)	100	99	99	99	94	90	84	72	40	16

TABLE 5.1
Recovery of perturbed rank-one tensors using random starting points by MBI

decomposable tensors (within $\sqrt{\frac{d}{3d-2}}$ multiple of the Frobenius norm of the perturbed tensor), then the MBI method is linearly convergent in a neighborhood of a best rank-one approximation tensor of $\mathcal{F}$. In this set of tests, we set out to check the performance of the MBI method in recovering perturbed rank-one tensors, which gives an idea of how large that neighborhood can expected to be.

Let $d = 4$ and $n_i = 10$ for $i = 1, 2, 3, 4$ in (4.3). We randomly generate a normalized rank-one tensor $\mathcal{F}_0$ (from randomly generated vectors), then put $\mathcal{F} = \mathcal{F}_0 + \mathcal{E}_0$ where $\mathcal{E}_0$ is a random orthogonal perturbation satisfying $\|\mathcal{E}_0\|_F/\|\mathcal{F}\|_F = \epsilon$. For $\mathcal{F}$ we first obtain a "best" normalized rank-one approximation $\mathcal{F}_*$ using MBI. Note that it holds for $\mathcal{E}_* = \mathcal{F} - \mathcal{F}_*$,

$$\sqrt{1 + \|\mathcal{E}_0\|_F^2} - 1 \leq \|\mathcal{E}_*\|_F \leq \|\mathcal{E}_0\|_F$$

(the first term is the distance of $\mathcal{F}$ to the unit sphere), that is, $\|\mathcal{E}_*\|_F/\|\mathcal{F}\|_F$ is also of order ϵ.

We then apply the MBI method to solve (4.3) using random initial solution 100 more times, and check if they *all* return to the same stationary point, indicating the exact recovery or not. For different level of perturbation ($\|\mathcal{E}\|_F/\|\mathcal{F}\|_F$), the chance of exact recover by the MBI method is listed in Table 5.1 (according to 100 random test instances). It clearly shows that for a large allowed distance to the set of rank-one tensors (e.g. 50%), the MBI method is able to recover the original rank-one tensor. This somewhat confirms Theorem 4.3 that for small perturbations the convergence result is almost global.

In the same context we expect that the local convergence region for the stationary point $\mathcal{F}^* = x_*^1 \otimes x_*^2 \otimes x_*^3 \otimes x_*^4$ should be larger, the smaller the perturbation is, that is, the smaller the distance of $\mathcal{F}$ to the set of rank-one tensors is. Next we try to confirm this. We let the algorithm start from initial points

$$\mathbf{x}_0 = (x_0^1, x_0^2, x_0^3, x_0^4) = \left(\frac{x_*^1 + y^1}{\|x_*^1 + y^1\|}, \frac{x_*^2 + y^2}{\|x_*^2 + y^2\|}, \frac{x_*^3 + y^3}{\|x_*^3 + y^3\|}, \frac{x_*^4 + y^4}{\|x_*^4 + y^4\|} \right), \quad (5.1)$$

where $y^i \perp x_*^i$ and $\frac{\|y^i\|}{\|x_*^i + y^i\|} = \delta$ for $i = 1, 2, 3, 4$ are randomly generated for fixed δ. We generate the initial points 100 times and check if they *all* return to $\mathcal{F}^*$, indicating the exact recovery or not. The percentage of exact recovery for different perturbation $\|\mathcal{E}\|_F/\|\mathcal{F}\|_F$ of $\mathcal{F}$ and radius δ for starting points is listed in Table 5.2. It clearly confirms the expectation. Indeed, the local convergence region is quite large even for largely perturbed tensors.

5.2. Comparison with ALS and random block improvement. In this set of tests, we compare different block search procedures to solve (4.3). Apart from the greedy-type search method (MBI), we apply the workhorse algorithm ALS for low-rank approximations of tensors, which is a cyclic search method. Besides, we also

$\dfrac{\|y^i\|}{\|x_*^i+y^i\|}$ $\diagdown$ $\|\mathcal{E}\|/\|\mathcal{F}\|$	≤ 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
≤ 0.5	100	100	100	100	100	100	100	100	100
0.6	100	100	100	100	100	100	100	100	99
0.7	100	100	100	100	100	100	100	99	94
0.8	100	100	100	100	100	100	100	98	86
0.9	100	100	100	100	100	100	98	89	56
1.0	100	99	96	93	84	64	40	18	1

TABLE 5.2
Convergence regions of perturbed rank-one tensors by MBI

try random block search procedure, to be named by RAN, where the next block to be updated is uniformly (with probability $\frac{1}{d-1}$) chosen among all the blocks except the block updated previously. As a side remark, when implementing MBI, the next block to be updated is chosen the best one among all the blocks except the block updated previously, for the sake of saving computational cost.

In order to observe the local convergence properties of these three methods (MBI, ALS and RAN), we first apply the MBI method to get a stationary solution $\mathbf{x}_*$. Then we add $\mathbf{x}_*$ with a small perturbation over its null space as the initial solution like (5.1). Specifically, $x_0^i = \frac{x_*^i+y^i}{\|x_*^i+y^i\|}$ with random $y^i \perp x_*^i$ and $\frac{\|y^i\|}{\|x_*^i+y^i\|} = 0.05$ for $i = 1,2,3,4$ (hence $\|\mathbf{x}_0 - \mathbf{x}_*\| = 0.1$). Apply the three methods independently with the same initial solution $\mathbf{x}_0$, and report the results *only* when all the three methods converge to the original stationary solution $\mathbf{x}_*$. Five randomly generated instances are plotted in Figure 5.3. Each row of figures are for the same instance, representing curves of the three methods according to number of iterations, the gap to the stationary solution $(-\log_{10}\|\mathbf{x}_n - \mathbf{x}_*\|)$, the loss of function value $(-\log_{10}(v_* - v_n))$, as well as the CPU time.

The following observations are clearly visible by comparing these three methods (Figure 5.3).

1. The linear convergence of all the methods are validated. In most cases, the convergence rate of MBI is slightly better than that of ALS, and both of them are clearly better than RAN.
2. Under random data tensors, the plots for gap to the stationary solution are almost same as that for loss of function value.
3. In terms of the CPU time, MBI is the most costly, mainly because it checks $d-1$ blocks in each iteration. In fact, one can use a parallel implementation for the block search in each MBI iteration, which ALS and RAN cannot apply. We estimated the time for parallelized MBI by dividing by $d-1$, which should be adequate for random data. Then the CPU time plots (the last column in Figure 5.3) are almost same as the iteration plots (the first two columns).
4. MBI is a more stable and reliable method to use, as almost all the plots look the same. In comparison, the prediction of ALS and RAN convergence is quite uncertain, especially for RAN.

All the test instances are quite uniform and random. In our next and final test, we incorporate some artificial structure. We adopt exactly the same test procedures for comparing the three methods as above, the only slight difference is the choice of the starting point. After having applied the MBI method to get the stationary point
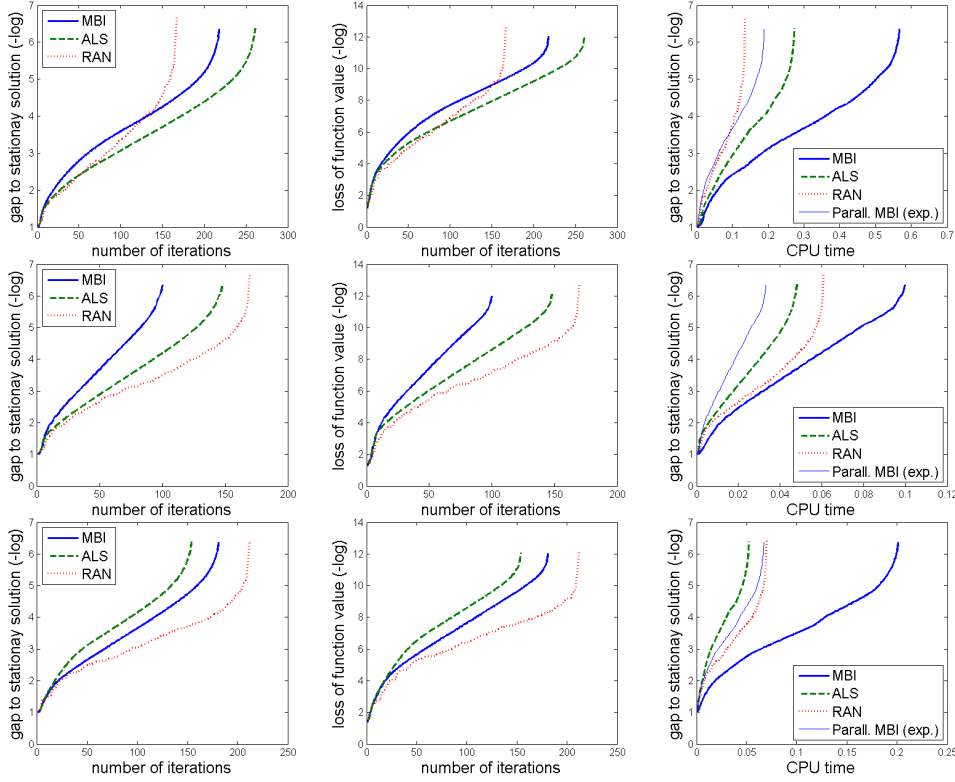
Fig. 5.3. *Three instances of* (4.3) *by MBI, ALS, and RAN*

$\mathbf{x}_*$, the initial solution is chosen as

$$\mathbf{x}_0 = (x_0^1, x_0^2, x_0^3, x_0^4) = \left(\frac{x_*^1 + y^1}{\|x_*^1 + y^1\|}, \frac{x_*^2 + y^2}{\|x_*^2 + y^2\|}, \frac{x_*^1 + y^3}{\|x_*^1 + y^3\|}, \frac{x_*^2 + y^4}{\|x_*^2 + y^4\|} \right)$$

where the y_i are random. Both the first two blocks and the last two blocks of the initial solution, are hence small perturbations of the first two blocks of the stationary point.

Again, we apply the three methods using the same initial solution $\mathbf{x}_0$, and report the results only when all the three methods output the original stationary solution $\mathbf{x}_*$. Three randomly generated instances are plotted in Figure 5.4. In contrast to the previous observations in Figure 5.3, a clearer advantage of the MBI method for the structured initial data can be seen in this set of tests.

1. The convergence rate of MBI outperforms that of ALS and RAN significantly.
2. In terms of the CPU time, MBI is the fastest to get a low accuracy (say at least 10^{-2}) even under this costly subroutine (not considering the parallelization). This suggests us to propose MBI at least as a starting procedure.
3. MBI is much more robust than ALS and RAN. In same cases, the convergence rates of ALS and RAN can be very slow.

To conclude this section, in terms of the number of iterations, the search for a better block in each iteration is important. The MBI method usually converges faster than other block coordinate search type methods. Though this greedy type search method may increase the overall computational cost, parallel implementation
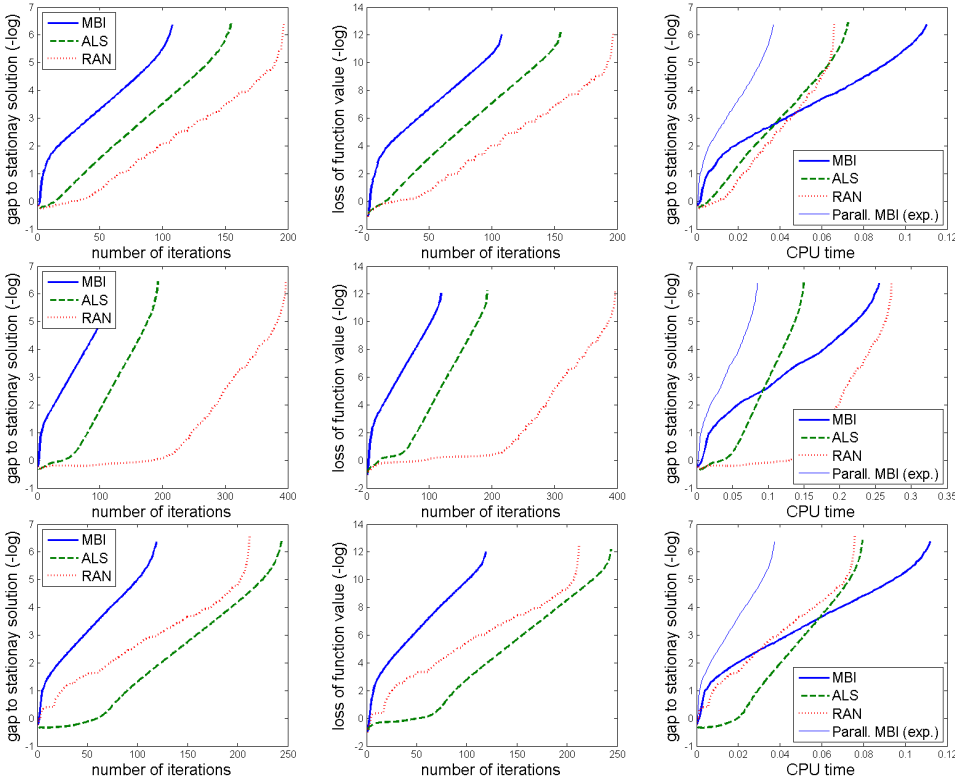
Fig. 5.4. *Three instances of* (4.3) *using structured starting points by MBI, ASL and RAN*

for the block search in each MBI iteration is strongly recommended. Besides, applying the MBI method in the beginning to a point close to the stationary point, and then switching to other searching methods would be suggested. No matter what type of search method is applied, the convergence is only linear. The MBI method is the only one (amongst the considered) to guarantee global convergence to a stationary point, as shown in [7].

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, J. BOLTE, P. REDONT, AND A. SOUBEYRAN, *Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.

[3] B. W. BADER, T. G. KOLDA, ET AL., *Matlab tensor toolbox version 2.5.* Available online, January 2012.

[4] J. BOLTE, A. DANIILIDIS, AND A. LEWIS, *The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim., 17 (2006), pp. 1205–1223 (electronic).

[5] E. CANCÈS, V. EHRLACHER, AND T. LELIÈVRE, *Convergence of a greedy algorithm for high-dimensional convex nonlinear problems*, Math. Models Methods Appl. Sci., 21 (2011), pp. 2433–2467.

[6] J. D. CARROLL AND J.-J. CHANG, *Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition*, Psychometrika, 35 (1970), pp. 283–319.

[7] B. Chen, S. He, Z. Li, and S. Zhang, *Maximum block improvement and polynomial optimization*, SIAM J. Optim., 22 (2012), pp. 87–107.

[8] 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.

[9] A. Falcó and A. Nouy, *Proper generalized decomposition for nonlinear convex problems in tensor Banach spaces*, Numer. Math., 121 (2012), pp. 503–530.

[10] E. F. Gonzalez and Y. Zhang, *Accelerating the Lee-Seung algorithm for nonnegative matrix factorization*, tech. rep., Rice University, Houston, TX, 2005.

[11] L. Grasedyck, D. Kressner, and C. Tobler, *A literature survey of low-rank tensor approximation techniques*, GAMM-Mitt., 36 (2013), pp. 53–78.

[12] M. Griebel and P. Oswald, *Greedy and randomized versions of the multiplicative Schwarz method*, Linear Algebra Appl., 437 (2012), pp. 1596–1610.

[13] L. Grippo and M. Sciandrone, *On the convergence of the block nonlinear Gauss-Seidel method under convex constraints*, Oper. Res. Lett., 26 (2000), pp. 127–136.

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

[15] R. A. Harshman, *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA Working Papers in Phonetics, 16 (1970), pp. 1–84.

[16] S. He, Z. Li, and S. Zhang, *Approximation algorithms for homogeneous polynomial optimization with quadratic constraints*, Math. Program., 125 (2010), pp. 353–383.

[17] 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.

[18] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[19] K. Kurdyka, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier (Grenoble), 48 (1998), pp. 769–783.

[20] D. Leventhal and A. S. Lewis, *Randomized methods for linear constraints: convergence rates and conditioning*, Math. Oper. Res., 35 (2010), pp. 641–654.

[21] G. Li, L. Qi, and G. Yu, *Semismoothness of the maximum eigenvalue function of a symmetric tensor and its application*, Linear Algebra Appl., 438 (2013), pp. 813–833.

[22] L.-H. Lim, *Singular values and eigenvalues of tensors: a variational approach*, in Computational Advances in Multi-Sensor Adaptive Processing, 2005 1st IEEE International Workshop on, dec. 2005, pp. 129 –132.

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

[24] Z. Q. Luo and P. Tseng, *On the convergence of the coordinate descent method for convex differentiable minimization*, J. Optim. Theory Appl., 72 (1992), pp. 7–35.

[25] Y. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM J. Optim., 22 (2012), pp. 341–362.

[26] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.

[27] J. M. Ortega and M. L. Rockoff, *Nonlinear difference equations and Gauss-Seidel type iterative methods*, SIAM J. Numer. Anal., 3 (1966), pp. 497–513.

[28] I. V. Oseledets, *Tensor-train decomposition*, SIAM J. Sci. Comput., 33 (2011), pp. 2295–2317.

[29] I. V. Oseledets and E. E. Tyrtyshnikov, *Breaking the curse of dimensionality, or how to use SVD in many dimensions*, SIAM J. Sci. Comput., 31 (2009), pp. 3744–3759.

[30] M. J. D. Powell, *On search directions for minimization algorithms*, Math. Programming, 4 (1973), pp. 193–201.

[31] P. Richtárek and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, tech. rep., School of Mathematics, University of Edinburgh, 2011.

[32] S. Schechter, *Iteration methods for nonlinear problems*, Trans. Amer. Math. Soc., 104 (1962), pp. 179–189.

[33] R. V. Southwell, *Relaxation Methods in Engineering Science. A treatise on approximate computation*, Oxford Engineering Science Series, Oxford University Press, New York, 1940.

[34] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl., 109 (2001), pp. 475–494.

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

[36] A. Uschmajew and B. Vandereycken, *The geometry of algorithms using hierarchical tensors*, Linear Algebra Appl., 439 (2013), pp. 133–166.

[37] J. WEISSINGER, *Verallgemeinerungen des Seidelschen Iterationsverfahrens*, Z. Angew. Math. Mech., 33 (1953), pp. 155–163.

[38] Y. XU AND W. YIN, *A block coordinate descent method for regularized multi-convex optimization with applications to nonnegative tensor factorization and completion*, tech. rep., Rice University, 2012.

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

[40] X. ZHANG, C. LING, AND L. QI, *The best rank-1 approximation of a symmetric tensor and related spherical optimization problems*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 806–821.

Recent publications :

MATHEMATICS INSTITUTE OF COMPUTATIONAL SCIENCE AND ENGINEERING
Section of Mathematics
Ecole Polytechnique Fédérale
CH-1015 Lausanne

16.2013   D. KRESSNER, M. MILOLOZA PANDUR, M. SHAO:
An indefinite variant of LOBPCG for definite matrix pencils

17.2013   A. ABDULLE, M. J. GROTE, C. STOHRER:
FE heterogeneous multiscale method for long time wave propagation

18.2013   A. ABDULLE, Y. BAI, G. VILMART:
An online-offline homogenization strategy to solve quasilinear two-scale problems at the cost of one-scale problems

19.2013   C.M. COLCIAGO, S. DEPARIS, A. QUARTERONI:
Comparison between reduced order models and full 3D models for fluid-structure interaction problems in haemodynamics

20.2013   D. KRESSNER, M. STEINLECHNER, B. VANDEREYCKEN:
Low-rank tensor completion by Riemannian optimization

21.2013   M. KAROW, D. KRESSNER, E. MENGI:
Nonlinear eigenvalue problems with specified eigenvalues

22.2013   T. LASSILA, A. MANZONI, A. QUARTERONI, G. ROZZA:
Model order reduction in fluid dynamics: challenges and perspectives

23.2013   M. DISCACCIATI, P. GERVASIO, A. QUARTERONI:
The interface control domain decomposition (ICDD) method for the Stokes problem

24.2013   V. LEVER, G. PORTA, L. TAMELLINI, M. RIVA:
Characterization of basin-scale systems under mechanical and geochemical compaction

25.2013   D. DEVAUD, A. MANZONI, G. ROZZA:
A combination between the reduced basis method and the ANOVA expansion: on the computation of sensitivity indices

26.2013   M. SHAO:
On the finite section method for computing exponentials of doubly-infinite skew-Hermitian matrices

27.2013   A. ABDULLE, G. VILMART, K. C. ZYGALAKIS:
High order numerical approximation of the invariant measure of ergodic SDEs

28.2013   S. ROSSI, T. LASSILA, R. RUIZ-BAIER, A. SEQUEIRA, A. QUARTERONI:
Thermodynamically consistent orthotropic activation model capturing ventricular systolic wall thickening in cardiac electromechanics

29.2013   F. BONIZZONI, F. NOBILE:
Perturbation analysis for the Darcy problem with log-normal permeability

30.2013   Z. LI, A. USCHMAJEW, S. ZHANG:
On convergence of the maximum block improvement method