A QUADRATICALLY CONVERGENT PROXIMAL ALGORITHM FOR NONNEGATIVE TENSOR DECOMPOSITION

Nico Vervliet[1] Andreas Themelis[1] Panagiotis Patrinos[1] Lieven De Lathauwer[1,2]

[1] KU Leuven, Department of Electrical Engineering ESAT–STADIUS, Kasteelpark Arenberg 10, bus 2446, B-3001 Leuven, Belgium
[2] KU Leuven–Kulak, Group Science, Engineering and Technology, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium
{Nico.Vervliet, Andreas.Themelis, Panos.Patrinos, Lieven.DeLathauwer}@kuleuven.be

***Abstract*—The decomposition of tensors into simple rank-1 terms is key in a variety of applications in signal processing, data analysis and machine learning. While this canonical polyadic decomposition (CPD) is unique under mild conditions, including prior knowledge such as nonnegativity can facilitate interpretation of the components. Inspired by the effectiveness and efficiency of Gauss–Newton (GN) for unconstrained CPD, we derive a proximal, semismooth GN type algorithm for nonnegative tensor factorization. Global convergence to local minima is achieved via backtracking on the forward-backward envelope function. If the algorithm converges to a global optimum, we show that *Q*-quadratic rates are obtained in the exact case. Such fast rates are verified experimentally, and we illustrate that using the GN step significantly reduces number of (expensive) gradient computations compared to proximal gradient descent.**

***Index Terms*—nonnegative tensor factorization, canonical polyadic decomposition, proximal methods, Gauss–Newton**

I. Introduction

The canonical polyadic decomposition (CPD) expresses an Nth-order tensor $\mathcal{T}$ as a minimal number of R rank-1 terms, each of which is the outer product, denoted by $\otimes$, of N nonzero vectors, with R the tensor rank. Mathematically, we have

$$\mathcal{T} = \sum_{r=1}^{R} \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(N)} =: \left[\!\left[\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)} \right]\!\right], \qquad (1)$$

in which factor matrix $\mathbf{A}^{(n)}$ has $\boldsymbol{a}_r^{(n)}$ as its columns. (Elementwise, we have $t_{i_1,i_2,\ldots,i_N} = \sum_{r=1}^{R} a_{i_1 r}^{(1)} a_{i_2 r}^{(2)} \cdots a_{i_N r}^{(N)}$.) The CPD is essentially unique under mild conditions, which is an attractive property often exploited in, e.g., data analysis, signal processing and machine learning [1,2]. To improve interpretability of the components and to mitigate ill-posedness, nonnegativity constraints can be imposed on the factor vectors [1,2], i.e., $\boldsymbol{a}_r^{(n)} \geq 0$, $\forall n, r$, where the inequality is meant elementwise.

The CPD can be cast as the following nonlinear least squares problem (NLS):

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \quad \text{with} \quad f(\boldsymbol{x}) \coloneqq \tfrac{1}{2}\|F(\boldsymbol{x})\|^2,$$

where $F(\boldsymbol{x})$ is a vector-valued polynomial (multilinear) function. The Gauss–Newton method (GN) is a powerful tool to address this kind of problems, as despite requiring only first-order information of $F(\boldsymbol{x})$ it can exhibit up to quadratic rates of convergence. The idea behind GN is using the Gramian $JF(\boldsymbol{x})^\top JF(\boldsymbol{x})$ as a surrogate for $\nabla^2 f(\boldsymbol{x})$, which well approximates the true Hessian around solutions $\boldsymbol{x}^\star$ whenever $F(\boldsymbol{x}^\star) = \mathbf{0}$. One iteration $\boldsymbol{x} \mapsto \boldsymbol{x}^+$ of GN amounts to solving the linear system

$$(JF(\boldsymbol{x})^\top JF(\boldsymbol{x}))(\boldsymbol{x}^+ - \boldsymbol{x}) = -JF(\boldsymbol{x})^\top F(\boldsymbol{x}),$$

in which $JF(\boldsymbol{x})^\top F(\boldsymbol{x})$ is the gradient of $f(\boldsymbol{x})$. Thanks to the multilinear structure of the problem, the linear system can be solved efficiently using iterative methods [3,4].

As is typical for higher-order methods, GN converges only if the starting point is already close enough to a local solution, whence the need of a globalization strategy ensuring that the iterates eventually enter a basin of (fast) local convergence. Thanks to the smoothness of the cost function $f(\boldsymbol{x})$, many line-search or trust region approaches can efficiently be employed for the purpose; see, e.g. [3,4]. However, the nonsmoothness arising from the constraints makes the approach not applicable to *nonnegative* CPD problems, namely

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathbb{R}^d} \tfrac{1}{2}\|F(\boldsymbol{x})\|^2 \quad \text{subject to} \;\; \boldsymbol{x} \geq 0. \qquad (2)$$

In this paper we leverage the globalization technique of [5,6] to obtain a globally and quadratically convergent algorithm for NCPD directly addressing the constrained formulation (2). Convergence is global in the sense that convergence holds regardless of the initialization, albeit possibly to a local solution. Nevertheless, to further reduce the number of singularities and nonoptimal stationary points, we impose an additional nonconvex constraint that singles out ambiguities in the tensor decomposition arising because of its equivalence up to scaling factors. We defer the details to Section II.

A. Related work

A number of alternating least squares or block coordinate descent (BCD) type methods have been proposed to solve (2). In these algorithms, one factor matrix or one row or column is fixed at every iteration, after which a linear least squares subproblem with nonnegativity constraints is solved [7,8] by, e.g., using multiplicative updates [9], active set methods [10], or the alternating direction method of multipliers (ADMM) [11]. To compute a nonnegative CPD, other cost functions based on divergences can be used as well; see, e.g., [7,9,12].

This work was supported by the *Research Foundation Flanders (FWO)* via projects G086518N, G086318N, and via postdoc grant 12ZM220N; *KU Leuven Internal Funds* via projects C14/18/068, C16/15/059, and IDN/19/014; *Fonds de la Recherche Scientifique—FNRS and the Fonds Wetenschappelijk Onderzoek—Vlaanderen* under EOS project No. 30468160 (SeLMA). This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

While these BCD methods are often easy to implement, their convergence is slow. Therefore, a few algorithms based on GN or Levenberg–Marquardt (LM) have been proposed. Nonnegativity constraints can then be enforced using logarithmic penalty functions [13] or active set methods [3,4,14,15]. By change of variable, e.g., by replacing x_i with x_i^2, (2) can be converted to an unconstrained problem [16,17], which may lead to a prohibitive increase of nonoptimal stationary points. For nonnegative matrix factorization, a proximal LM type algorithm which solves an optimization problem using ADMM in every iteration has been proposed [18].

B. Notation

Scalars, vectors, matrices are denoted by lower case, e.g., a, bold lower case, e.g., $\boldsymbol{a}$ and bold upper case, e.g., $\mathbf{A}$, respectively. Calligraphic letters are used for a tensor $\mathcal{T}$, a constraint set $\mathcal{C}$, or the uniform distribution $\mathcal{U}(a, b)$. Sets are indexed by superscripts within parentheses, e.g., $\mathbf{A}^{(n)}$, $n = 1, \ldots, N$. The Kronecker and Khatri–Rao (column-wise Kronecker) products are denoted by $\otimes$ and $\odot$, respectively. The notation $\operatorname{blkdiag}(\{\boldsymbol{x}_r^{(n)}\}_{r,n=1}^{R,N})$ is used for a block-diagonal matrix with blocks $\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_R^{(1)}, \boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_R^{(N)}$. The identity matrix is denoted by $\mathbf{I}$, the column-wise concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$ by $[\boldsymbol{a}; \boldsymbol{b}]$, and the closed ε-ball around $\boldsymbol{x}$ by $\mathrm{B}(\boldsymbol{x}, \varepsilon)$.

II. A semismooth Gauss–Newton method

We derive a GN method to compute the nonnegative CPD of an $I_1 \times I_2 \times \cdots \times I_N$ tensor $\mathcal{T}$. By requiring each vector to have unit norm, $(N-1)R$ degrees of freedom associated with the scaling ambiguity are removed. The magnitudes λ_r of each term in the sum are scalar quantities that we collect in a vector $\boldsymbol{\lambda} \in \mathbb{R}^R$, resulting in the normalized decomposition

$$\mathcal{T} = \textstyle\sum_{r=1}^R \lambda_r \cdot \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(N)} \quad \text{with} \quad \|\boldsymbol{a}_r^{(n)}\| = 1, \forall n, r.$$

The normalized version of problem (2) thus becomes

$$\underset{\boldsymbol{a} \in \mathbb{R}^d, \boldsymbol{\lambda} \in \mathbb{R}^R}{\text{minimize}} \ \tfrac{1}{2}\|F(\boldsymbol{a}, \boldsymbol{\lambda})\|^2 \ \text{ subject to } \begin{cases} \boldsymbol{a}, \boldsymbol{\lambda} \geq 0, \\ \|\boldsymbol{a}_r^{(n)}\| = 1, \end{cases} \tag{3}$$

in which $F(\boldsymbol{a}, \boldsymbol{\lambda}) = \sum_{r=1}^R \lambda_r \cdot \boldsymbol{a}_r^{(1)} \otimes \cdots \otimes \boldsymbol{a}_r^{(N)} - \mathcal{T}$, $\boldsymbol{a} = [\boldsymbol{a}_1^{(1)}; \boldsymbol{a}_2^{(1)}; \ldots; \boldsymbol{a}_R^{(1)}; \boldsymbol{a}_1^{(2)}; \ldots; \boldsymbol{a}_R^{(N)}]$, and $d = R \sum_{n=1}^N I_n$. The feasible set $\mathcal{C} \coloneqq \{(\boldsymbol{a}, \boldsymbol{\lambda}) \in \mathbb{R}^d \times \mathbb{R}^R \mid \boldsymbol{a}, \boldsymbol{\lambda} \geq 0, \|\boldsymbol{a}_r^{(n)}\| = 1\}$ is nonconvex, but projecting onto it is a simple block-separable operation: one has $\Pi_{\mathcal{C}}(\boldsymbol{a}, \boldsymbol{\lambda}) = (\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{\lambda}})$ where

$$\tilde{\boldsymbol{a}}_r^{(n)} = \Pi_{\mathbb{S}}\left(\Pi_{\mathbb{O}_+}(\boldsymbol{a}_r^{(n)})\right) = \frac{[\boldsymbol{a}_r^{(n)}]_+}{\|[\boldsymbol{a}_r^{(n)}]_+\|} \quad \text{and} \quad \tilde{\boldsymbol{\lambda}} = [\boldsymbol{\lambda}]_+. \tag{4}$$

Here, $\mathbb{S}$ and $\mathbb{O}_+$ denote the unit sphere and the positive orthant of suitable size, respectively, and $[\,\cdot\,]_+ = \max\{0, \cdot\,\}$ elementwise; in case $\|[\boldsymbol{a}_r^{(n)}]_+\| = 0$, the (set-valued) projection is easily seen to equal $\tilde{\boldsymbol{a}}_r^{(n)} = \{e_i \mid i \in \arg\max_j \boldsymbol{a}_{r,j}^{(n)}\}$ where e_i is the vector whose ith entry is 1 and is 0 elsewhere. First-order necessary condition for optimality in this constrained minimization setting can be cast as the nonlinear equation $\mathcal{R}_\gamma(\boldsymbol{x}) = \mathbf{0}$, with $\boldsymbol{x} \coloneqq (\boldsymbol{a}, \boldsymbol{\lambda})$ as optimization variable and

$$\mathcal{R}_\gamma(\boldsymbol{x}) \coloneqq \boldsymbol{x} - \Pi_{\mathcal{C}}(\boldsymbol{x} - \gamma JF(\boldsymbol{x})^\top F(\boldsymbol{x})) \tag{5}$$

is the projected-gradient residual mapping. This map is everywhere piecewise smooth (up to a negligible set of points that we may disregard, as shown in the proof of Theorem 1). As such, its Clarke Jacobian $J\mathcal{R}_\gamma$ [19, §7.1] furnishes a first-order approximation. The chain rule [19, Prop. 7.1.11(a)] gives

$$J\mathcal{R}_\gamma(\boldsymbol{x}) = \mathbf{I} - J\Pi_{\mathcal{C}}(\boldsymbol{w}) \cdot \left[\mathbf{I} - \gamma JF(\boldsymbol{x})^\top JF(\boldsymbol{x}) - \underbrace{\gamma \textstyle\sum_i F_i(\boldsymbol{x}) \nabla^2 F_i(\boldsymbol{x})}\right],$$

where $F_i(\boldsymbol{x})$ is the ith element of vector $F(\boldsymbol{x})$,

$$\boldsymbol{w} = \boldsymbol{x} - \gamma JF(\boldsymbol{x})^\top F(\boldsymbol{x}) \tag{6}$$

is a gradient descent step at $\boldsymbol{x}$, and $J\Pi_{\mathcal{C}}(\boldsymbol{w})$ is a (set of) $(d+R) \times (d+R)$ block-diagonal matrices. In order to avoid Hessian evaluations, in the same spirit of (unconstrained) GN we replace $J\mathcal{R}_\gamma$ with

$$\hat{J}\mathcal{R}_\gamma(\boldsymbol{x}) \coloneqq \mathbf{I} - J\Pi_{\mathcal{C}}(\boldsymbol{w}) \cdot \left[\mathbf{I} - \gamma JF(\boldsymbol{x})^\top JF(\boldsymbol{x})\right], \tag{7}$$

which is $O(\|\boldsymbol{x} - \boldsymbol{x}^\star\|)$-close to $J\mathcal{R}_\gamma(\boldsymbol{x})$ around a solution $\boldsymbol{x}^\star$ of (3) provided that $F(\boldsymbol{x}^\star) = \mathbf{0}$, as is apparent from the bracketed term in the expression of $J\mathcal{R}_\gamma(\boldsymbol{x})$. Since the feasible set $\mathcal{C}$ is the product of small dimensional sets $\mathbb{S}_+ \coloneqq \mathbb{S} \cap \mathbb{O}_+$ and $\mathbb{O}_+$, $J\Pi_{\mathcal{C}}$ is a structured set of block-diagonal matrices whose computation can be easily carried out using the chain rule $J\Pi_{\mathbb{S}_+}(\boldsymbol{w}) = J\Pi_{\mathbb{S}}([\boldsymbol{w}]_+) J\Pi_{\mathbb{O}_+}(\boldsymbol{w})$ and the formulas

$$J\Pi_{\mathbb{S}}([\boldsymbol{w}]_+) = \frac{\mathbf{I} - \boldsymbol{z}\boldsymbol{z}^\top}{\|[\boldsymbol{w}]_+\|}, \text{with } \boldsymbol{z} \coloneqq \Pi_{\mathbb{S}}([\boldsymbol{w}]_+) = \frac{[\boldsymbol{w}]_+}{\|[\boldsymbol{w}]_+\|}, \tag{8}$$

and (see [20, §15.6.2d])

$$J\Pi_{\mathbb{O}_+}(\boldsymbol{w})_{i,j} \begin{cases} = 1 & \text{if } i = j \wedge w_i > 0 \\ \in [0, 1] & \text{if } i = j \wedge w_i = 0 \\ = 0 & \text{otherwise.} \end{cases} \tag{9}$$

Theorem 1 (Local quadratic convergence)**.** *Let $\boldsymbol{x}^\star$ be such that $F(\boldsymbol{x}^\star) = \mathbf{0}$, and suppose that all matrices in $\hat{J}\mathcal{R}_\gamma(\boldsymbol{x}^\star)$ are nonsingular. Then there exists $\varepsilon > 0$ such that the iterations*

$$\begin{cases} \boldsymbol{x}^0 \in \mathrm{B}(\boldsymbol{x}^\star, \varepsilon) \\ \boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \boldsymbol{d}^k, \end{cases} \quad \textit{where} \quad \hat{\mathbf{H}}_k \boldsymbol{d}^k = -\mathcal{R}_\gamma(\boldsymbol{x}^k) \tag{10}$$

with $\hat{\mathbf{H}}_k$ being any element of $\hat{J}\mathcal{R}_\gamma(\boldsymbol{x}^k)$, are Q-quadratically convergent to $\boldsymbol{x}^\star$.

Proof. We start by remarking that the projection onto the (product of) sphere(s) is C^∞ wherever it is well defined. Since $\boldsymbol{x}^\star$ is optimal, it follows from [5, Thm. 3.4(iii)] that $\mathcal{R}_\gamma(\boldsymbol{x}^\star) = \{\mathbf{0}\}$, and that consequently the projection onto the spheres it entails, cf. (4), is well defined and is thus C^∞ in a neighborhood. Combined with the strong semismoothness of the projection onto the positive orthant, see [19, Prop. 7.4.7], by invoking [19, Prop. 7.4.4] we conclude that $\mathcal{R}_\gamma$ is strongly semismooth around $\boldsymbol{x}^\star$.

Next, observe that $\mathbf{H}_k \coloneqq \hat{\mathbf{H}}_k + \Delta(\boldsymbol{x}^k) \in J\mathcal{R}_\gamma(\boldsymbol{x}^k)$, for some $\Delta(\boldsymbol{x}) \in J\Pi_{\mathcal{C}}(\boldsymbol{w}) \sum_i F_i(\boldsymbol{x}) \nabla^2 F_i(\boldsymbol{x})$ (with $\boldsymbol{w}$ as in (6)) is a locally bounded quantity such that $\Delta(\boldsymbol{x}) \to \mathbf{0}$ as $\boldsymbol{x} \to \boldsymbol{x}^\star$. Therefore, denoting $\boldsymbol{d}^k \coloneqq \boldsymbol{x}^{k+1} - \boldsymbol{x}^k$,

$$\|(\mathcal{R}_\gamma(\boldsymbol{x}^k) + \mathbf{H}_k)\boldsymbol{d}^k\| = \|\Delta(\boldsymbol{x}^k)\boldsymbol{d}^k\| \leq \|\Delta(\boldsymbol{x}^k)\| \|\hat{\mathbf{H}}_k^{-1}\| \|\mathcal{R}_\gamma(\boldsymbol{x}^k)\|.$$

We have that $\sup_{\hat{\mathbf{H}} \in \hat{J}\mathcal{R}_\gamma(\boldsymbol{x})} \|\hat{\mathbf{H}}^{-1}\|$ is bounded by a same quantity c_ε for all $\boldsymbol{x} \in \mathrm{B}(\boldsymbol{x}^\star, \varepsilon)$ when ε is small enough, as it follows from [19, Lem. 7.5.2]. Consequently, for ε small enough [19, Thm. 7.5.5] guarantees that $\boldsymbol{x}^k \to \boldsymbol{x}^\star$ Q-linearly. In turn, this implies that $\Delta(\boldsymbol{x}^k) \to \mathbf{0}$, hence invoking again the same result, the claimed Q-quadratic convergence is obtained. □

Theorem 1 requires that all matrices in $\hat{J}\mathcal{R}_\gamma(\boldsymbol{x}^\star)$ in (7) are nonsingular. In Theorem 3, we show that this is the case for the exact decomposition problem if the Gramian $JF(\boldsymbol{x})^\top JF(\boldsymbol{x})$ has an NR-dimensional null space (which is usually true for a unique CPD). This null space is derived in the next lemma.

Lemma 2 (Kernel of Gramian)**.** *The Gramian of the unconstrained problem $JF(\boldsymbol{x})^\top JF(\boldsymbol{x})$ has at least NR zero eigenvalues, and a basis $\mathbf{K}$ for the subspace corresponding to these NR zero eigenvalues is given by*

$$\mathbf{K} = \operatorname{blkdiag}(\{\operatorname{diag}(\boldsymbol{k}^{(n)}) \odot \mathbf{A}^{(n)}\}_n, \operatorname{diag}(\boldsymbol{k}^{(N+1)}) \odot \boldsymbol{\lambda}^\top), \quad (11)$$

for $\boldsymbol{k}^{(n)} \in \mathbb{R}^R$, $n = 1, \ldots, N+1$, and $\sum_{n=1}^{N+1} \boldsymbol{k}^{(n)} = \mathbf{0}$.

Proof. It suffices to check that $JF(\boldsymbol{x})\mathbf{K} = \mathbf{0}$ and that the dimension of $\mathbf{K}$ is NR. Using the expressions for $JF(\boldsymbol{x})$ (see, e.g., [4]) and multilinear identities, we have

$$JF(\boldsymbol{x})\mathbf{K} = \left(\bigodot_{n=1}^{N} \mathbf{A}^{(n)} \odot \boldsymbol{\lambda}^\top\right) \sum_{n=1}^{N+1} \boldsymbol{k}^{(n)} = \mathbf{0}. \quad (12)$$

As $\left(\bigodot_n \mathbf{A}^{(n)} \odot \boldsymbol{\lambda}^\top\right)$ usually has full column rank for an essentially unique decomposition defined by $\boldsymbol{x}$, we need $\sum_{n=1}^{N+1} \boldsymbol{k}^{(n)} = \mathbf{0}$. Since $\boldsymbol{k} = \left[\boldsymbol{k}^{(1)}; \ldots; \boldsymbol{k}^{(N+1)}\right] \in \mathbb{R}^{(N+1)R}$ and the summation imposes R linearly independent constraints, the columns of $\mathbf{K}$ span an NR-dimensional subspace. □

Theorem 3. *Let $\hat{\mathbf{H}} \in \hat{J}\mathcal{R}_\gamma(\boldsymbol{x}^\star)$ and $F(\boldsymbol{x}^\star) = \mathbf{0}$. If the Gramian $JF(\boldsymbol{x}^\star)^\top JF(\boldsymbol{x}^\star)$ has NR zero eigenvalues, $\hat{\mathbf{H}}$ is nonsingular.*

Proof. In the global optimum $\boldsymbol{x}^\star$, $JF(\boldsymbol{x}^\star)^\top F(\boldsymbol{x}^\star) = \mathbf{0}$ and $\boldsymbol{x}^\star \in \mathcal{C}$, hence $J\Pi_{\mathcal{C}}(\boldsymbol{w}) = J\Pi_{\mathcal{C}}(\boldsymbol{x}^\star)$. Let $\mathbf{P} \in J\Pi_{\mathcal{C}}(\boldsymbol{x}^\star)$, and $\mathbf{G} = JF(\boldsymbol{x}^\star)^\top JF(\boldsymbol{x}^\star)$. Before proving that $\hat{\mathbf{H}} \in \hat{J}\mathcal{R}_\gamma(\boldsymbol{x}^\star)$ has full rank, we show that $\operatorname{range}(\mathbf{PG}) = \operatorname{range}(\mathbf{P})$ which is the case if $\operatorname{range}(\mathbf{P}) \cap \operatorname{null}(\mathbf{G}) = \emptyset$. Let $\boldsymbol{a}_r^{(n)}$ and λ_r be the factor vectors and scaling factors corresponding to $\boldsymbol{x}^\star$. By assumption, $\mathbf{G}$ has NR zero eigenvalues and $\operatorname{null}(\mathbf{G}) = \operatorname{range}(\mathbf{K})$; see Lemma 2. Any $\boldsymbol{b} \in \operatorname{range}(\mathbf{P})$ can be written as $\boldsymbol{b} = [\boldsymbol{b}_1^{(1)}; \ldots; \boldsymbol{b}_R^{(1)}; \boldsymbol{b}_1^{(2)}; \ldots; \boldsymbol{b}_R^{(N)}; b_1^{(N+1)}; \ldots; b_R^{(N+1)}]$ in which either $\boldsymbol{b}_r^{(n)} \perp \boldsymbol{a}_r^{(n)}$ or $\boldsymbol{b}_r^{(n)} = \mathbf{0}$; see (8). If $\boldsymbol{b} \in \operatorname{range}(\mathbf{K})$, then the following should hold with $\sum_{n=1}^{N+1} \boldsymbol{k}^{(n)} = \mathbf{0}$:

$$\begin{aligned} \boldsymbol{b}_r^{(n)} &= k_r^{(n)} \boldsymbol{a}_r^{(n)} &\Leftrightarrow\quad k_r^{(n)} &= 0, &\forall n, r, \\ b_r^{(N+1)} &= k_r^{(N+1)} \lambda_r &\Leftrightarrow\quad k_r^{(N+1)} &= b_r^{(N+1)} / \lambda_r, &\forall r, \end{aligned}$$

which is false, hence $\boldsymbol{b} \notin \operatorname{range}(\mathbf{K})$ and $\operatorname{range}(\mathbf{PG}) = \operatorname{range}(\mathbf{P})$.

Let $\hat{\mathbf{H}} = (\mathbf{I} - \mathbf{P}) + \gamma \mathbf{PG}$, and l_0 the number of active constraints for which $J\Pi_{\mathbb{O}_+}(\boldsymbol{x}^\star)_{i,i} = 0$. As $\operatorname{range}(\mathbf{PG}) = \operatorname{range}(\mathbf{P})$, we can show that there exists an $NR + l_0$ dimensional subspace U_1 of $\mathbf{I} - \mathbf{P}$, and a $d + R - NR - l_0$ dimensional subspace U_2 of $\mathbf{PH}$, such that $\boldsymbol{u}_1^\top \boldsymbol{u}_2 = 0$, $\forall \boldsymbol{u}_1 \in U_1$ and $\forall \boldsymbol{u}_2 \in U_2$. Therefore, $\operatorname{range}((\mathbf{I} - \mathbf{P}) + \gamma \mathbf{PG}) = \mathbb{R}^{d+R}$ and $\hat{\mathbf{H}}$ has full rank. □

III. The forward-backward envelope

Theorem 1 highlights an appealing property that the constrained GN directions (10) enjoy close to the solutions of (3). Unfortunately, however, there is no practical way of initializing the iterations in such a way that the quadratic convergence is triggered. In fact, not only is fast convergence not guaranteed without a proper initialization, but iterates may not converge at all and even diverge otherwise. Because of the constraints, classical linesearch strategies cannot be adopted for *nonnegative* CPDs.

Here, we overcome this limitation by integrating the fast GN directions (10) in a *nonsmooth* globalization strategy proposed in [6], based on the *forward-backward envelope* [5,21]

$$\varphi_\gamma^{\mathrm{FB}}(\boldsymbol{x}) = \tfrac{1}{2}\|F(\boldsymbol{x})\|^2 - \langle JF(\boldsymbol{x})^\top F(\boldsymbol{x}), \boldsymbol{r}\rangle + \tfrac{1}{2\gamma}\|\boldsymbol{r}\|^2, \quad (13)$$

where $\gamma > 0$ is a stepsize parameter,

$$\boldsymbol{z} \coloneqq \Pi_{\mathcal{C}}(\boldsymbol{x} - \gamma JF(\boldsymbol{x})^\top F(\boldsymbol{x})), \quad \text{and} \quad \boldsymbol{r} \coloneqq \boldsymbol{x} - \boldsymbol{z}. \quad (14)$$

The key properties of the FBE are summarized next. Although an easy adaptation of that of [5, Prop. 4.3 and Rem. 5.2], the proof is included for the sake of self containedness.

Lemma 4 (Basic properties of the FBE)**.** *For every $\gamma > 0$, $\varphi_\gamma^{\mathrm{FB}}(\boldsymbol{x})$ is locally Lipschitz continuous and real-valued. Moreover, denoting $f(\boldsymbol{x}) \coloneqq \frac{1}{2}\|F(\boldsymbol{x})\|^2$, the following hold:*

(i) $\varphi_\gamma^{\mathrm{FB}}(\boldsymbol{x}) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{C}$.

(ii) For all $\boldsymbol{x} \in \mathbb{R}^d$, if $\boldsymbol{z} \coloneqq \Pi_{\mathcal{C}}[\boldsymbol{x} - \gamma JF(\boldsymbol{x})^\top F(\boldsymbol{x})]$ satisfies

$$f(\boldsymbol{z}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x}\rangle + \tfrac{L}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 \quad (15)$$

for some $L > 0$, then $F(\boldsymbol{z}) \leq \varphi_\gamma^{\mathrm{FB}}(\boldsymbol{x}) - \frac{1-\gamma L}{2\gamma}\|\boldsymbol{x} - \boldsymbol{z}\|^2$. In particular, $\varphi_\gamma^{\mathrm{FB}}(\boldsymbol{x}) \geq f(\boldsymbol{z}) \geq 0$ whenever $\gamma \leq 1/L$.

(iii) On every bounded set $\Omega \subseteq \mathbb{R}^n$, there exists $L_\Omega > 0$ such that inequality (15) holds for every $\boldsymbol{x} \in \Omega$ and $L \geq L_\Omega$.

Proof. Real valuedness is apparent from (13). Moreover,

$$\begin{aligned} \varphi_\gamma^{\mathrm{FB}}(\boldsymbol{x}) &= \min_{\boldsymbol{v} \in \mathcal{C}} \{f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} - \boldsymbol{x}\rangle + \tfrac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{v}\|^2\} \\ &= f(\boldsymbol{x}) - \tfrac{\gamma}{2}\|\nabla f(\boldsymbol{x})\|^2 + \tfrac{1}{2\gamma}\operatorname{dist}(\boldsymbol{x} - \gamma \nabla f(\boldsymbol{x}), \mathcal{C})^2 \end{aligned}$$

hence $\varphi_\gamma^{\mathrm{FB}}(\boldsymbol{x}) \leq f(\boldsymbol{x})$ whenever $\boldsymbol{x} \in \mathcal{C}$ (by simply replacing $\boldsymbol{v} = \boldsymbol{x}$ in the minimization). Local Lipschitz continuity owes to that of f, ∇f and $\operatorname{dist}(\,\cdot\,, \mathcal{C})$, see [22, Ex. 9.6]. Moreover, since the minimum above is obtained at $\boldsymbol{v} = \boldsymbol{z}$, the second claim follows. Finally, since $\Pi_{\mathcal{C}}$ is locally bounded (cf. [22, Ex. 5.23(a)]) and so is ∇f, $\Pi_{\mathcal{C}}[\mathrm{id} - \gamma \nabla f]$ maps the bounded set Ω into a bounded set. Therefore, there exists a convex set $\bar{\Omega}$ that contains any $\boldsymbol{x}$ and $\boldsymbol{z} = \Pi_{\mathcal{C}}[\boldsymbol{x} - \gamma \nabla f(\boldsymbol{x})]$ with $\boldsymbol{x} \in \Omega$. The claimed L_Ω satisfying the last condition can thus be taken as the Lipschitz modulus of ∇f over $\bar{\Omega}$, see [23, Prop. A.24]. □

IV. A globally convergent algorithm

Lemma 4 contains all the key properties that lead to Algorithm I, which amounts to PANOC algorithm [6] specialized to this setting. Having shown the efficacy of the fast GN directions (10), the following result is a direct consequence of the more general ones in [5,6]. We remark that the differentiability

Algorithm I PANOC for NCPD

REQUIRE Starting point $\boldsymbol{x} \in \mathbb{R}^{d+R}$; $\alpha, \beta \in (0,1)$; tolerance $\varepsilon > 0$; estimate of Lipschitz modulus $L > 0$

INITIALIZE $\gamma = \alpha/L$

I.0: $\boldsymbol{z} = \Pi_{\mathcal{C}}[\boldsymbol{x} - \gamma JF(\boldsymbol{x})^\top F(\boldsymbol{x})]$ and $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{z}$
if $\frac{1}{2}\|F(\boldsymbol{z})\|^2 > \varphi_\gamma^{\text{FB}}(\boldsymbol{x}) - \frac{1-\alpha}{2\gamma}\|\boldsymbol{r}\|^2$ **then**
$\gamma \leftarrow \gamma/2$ and go back to step I.0

I.1: **if** $\frac{1}{\gamma}\|\boldsymbol{r}\|^2 \leq \varepsilon$ **then**
return $\boldsymbol{z}$

I.2: Pick $\hat{\mathbf{H}} \in \hat{J}\mathcal{R}_\gamma(\boldsymbol{x})$ and let $\boldsymbol{d}$ be such that $\hat{\mathbf{H}}^\top \hat{\mathbf{H}} \boldsymbol{d} = -\hat{\mathbf{H}}^\top \mathcal{R}_\gamma(\boldsymbol{x})$
Set stepsize $\tau = 1$

I.3: $\boldsymbol{x}^+ = (1-\tau)\boldsymbol{z} + \tau(\boldsymbol{x} + \boldsymbol{d})$
$\boldsymbol{z}^+ = \Pi_{\mathcal{C}}[\boldsymbol{x}^+ - \gamma JF(\boldsymbol{x}^+)^\top F(\boldsymbol{x}^+)]$ and $\boldsymbol{r}^+ = \boldsymbol{x}^+ - \boldsymbol{z}^+$

I.4: **if** $\frac{1}{2}\|F(\boldsymbol{z}^+)\|^2 > \varphi_\gamma^{\text{FB}}(\boldsymbol{x}^+) - \frac{1-\alpha}{2\gamma}\|\boldsymbol{r}^+\|^2$ **then**
$\gamma \leftarrow \gamma/2$ and go back to step I.0
else if $\varphi_\gamma^{\text{FB}}(\boldsymbol{x}^+) > \varphi_\gamma^{\text{FB}}(\boldsymbol{x}) - \frac{1-\alpha}{2\gamma}\beta\|\boldsymbol{r}\|^2$
$\tau \leftarrow \tau/2$ and go back to step I.3

I.5: $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{r}) \leftarrow (\boldsymbol{x}^+, \boldsymbol{z}^+, \boldsymbol{r}^+)$ and proceed to step I.1

assumptions of $\mathcal{R}_\gamma$ therein are only needed for showing the efficacy of quasi-Newton directions, whereas acceptance of unit stepsize only requires strong local minimality as shown in [24, Thm. 5.23].

Theorem 5 (Convergence of Algorithm I)**.** *Suppose that the sequence of points $\boldsymbol{z}$ remains bounded (as can be enforced by intersecting $\mathcal{C}$ with any large box, cf. (16)), then Algorithm I terminates in finitely many iterations. Moreover, with tolerance $\varepsilon = 0$ the following hold:*

(i) γ is reduced only finitely many times and the sequence of points $\boldsymbol{z}$ converges to a stationary point for (3).
(ii) If the conditions of Theorem 1 are satisfied at the limit point, then eventually stepsize $\tau = 1$ is always accepted and the sequence of points $\boldsymbol{z}$ converges Q-quadratically.

Although the proof is already subsumed by previous work, in conclusion of this section we briefly outline the main details of the globalization strategy. The algorithm revolves around the upper bound (15); since the modulus L (initialized as $L = \alpha/\gamma$) is not known a priori, it is adjusted adaptively throughout the iterations at steps I.0 and I.4 by halving γ (hence doubling $L = \alpha/\gamma$ as a byproduct) until (15) is satisfied. Since $\boldsymbol{z}$ is the result of a projection on $\mathcal{C}$, its $\boldsymbol{\lambda}$-component is nonnegative and its $\boldsymbol{a}$-component is bounded on unit spheres. Consequently, boundedness of all the iterates may be artificially imposed by changing the feasible set $\mathcal{C}$ into

$$\tilde{\mathcal{C}} \coloneqq \{(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^R \mid \boldsymbol{a} \geq 0, \|\boldsymbol{a}_r^{(n)}\| = 1, 0 \leq \boldsymbol{\lambda} \leq M\}, \tag{16}$$

where M is a large constant. Up to possibly resorting to this modification, as ensured by Lemma 4*(iii)* the stepsize γ is halved only a finite number of times and eventually remains constant. Since $\gamma = \alpha/L < 1/L$, Lemma 4*(ii)* guarantees that

$$\tfrac{1}{2}\|F(\boldsymbol{z})\|^2 \leq \varphi_\gamma^{\text{FB}}(\boldsymbol{x}) - \tfrac{1-\alpha}{2\gamma}\|\boldsymbol{r}\|^2 < \varphi_\gamma^{\text{FB}}(\boldsymbol{x}) - \tfrac{1-\alpha}{2\gamma}\beta\|\boldsymbol{r}\|^2$$

holds at every iteration; strict inequality holds because $\beta < 1$ and $\boldsymbol{r} \neq \boldsymbol{0}$ (for otherwise the algorithm would have stopped at step I.1). It then follows from the continuity of the FBE, Lemma 4*(i)*, and the fact that $\boldsymbol{x}^+ \to \boldsymbol{z}$ as $\tau \searrow 0$ (cf. step I.3), that at every iteration τ is halvened only a finite number of times at step I.4. In particular, the algorithm is well defined and, when γ becomes constant, produces a sequence satisfying

$$0 \leq \varphi_\gamma^{\text{FB}}(\boldsymbol{x}^+) \leq \varphi_\gamma^{\text{FB}}(\boldsymbol{x}) - \tfrac{1-\alpha}{2\gamma}\beta\|\boldsymbol{r}\|^2.$$

By telescoping the inequality, the vanishing of the *residual* $\boldsymbol{r}$ follows, hence the finite termination of the entire algorithm.

V. Complexity

The computational complexity of Algorithm I is dominated by the same operations as in the unconstrained case: computing the function value $f(\boldsymbol{x})$ (steps I.0 and I.4) and the gradient $JF(\boldsymbol{x})^\top F(\boldsymbol{x})$ (steps I.0 and I.3), and solving the linear system (10) (step I.2). If an iterative solver is used to solve (10)—which is common practice for medium to large scale problems—the computational complexity is dominated by the computation of the function evaluation and the gradient, which require $O(R \prod_n I_n)$ and $O(NR \prod_n I_n)$ operations, respectively, for an Nth-order $I_1 \times I_2 \times \cdots \times I_N$ tensor [3]. Given a good initial guess for the Lipschitz modulus L, step I.0 is computed only a few times. Hence, the total complexity mainly depends on the number of backtracking steps on τ at step I.4.

VI. Experiments

We validate the theoretical properties of Algorithm I by two experiments. Algorithm I is implemented in MATLAB 2019b with Tensorlab 3.0 [14]. Default values for the parameters $\alpha = 0.95$ and $\beta = 0.5$ are used. If more than five backtracking steps on τ at step I.4 are needed, a proximal gradient step is taken. The stopping tolerance is set to $\varepsilon = 10^{-20}$ and the maximum number of iterations to 2000. The Lipschitz modulus L is estimated using finite differences in a random direction. To prevent slower convergence due to small γ, we heuristically set $\gamma \leftarrow \max\{\gamma, \eta\}$ in step I.5 with $\eta = \boldsymbol{g}^\top \mathbf{G} \boldsymbol{g} / \|\boldsymbol{g}\|^2$ in which $\boldsymbol{g}$ and $\mathbf{G}$ are the gradient and Gramian for the unconstrained problem, respectively. (The scaling factor η is often used in trust region methods to compute the Cauchy point.)

We show that Q-quadratic convergence can be achieved for exact nonnegative CPD. In this experiment, 250 random $10 \times 10 \times 10$ tensors of rank-5 are constructed using random factor matrices with entries drawn from $\mathcal{U}(0,1)$. In each factor matrix, ten entries are set to zero at random to ensure some constraints are active. For each random tensor, Algorithm I is initialized by perturbing the exact solution such that approximately one digit is correct. During the algorithm the distance to the exact solution $\|\boldsymbol{x}^k - \boldsymbol{x}^\star\|$ is tracked. This error should decrease as $\|\boldsymbol{x}^+ - \boldsymbol{x}^\star\| = O(\|\boldsymbol{x} - \boldsymbol{x}^\star\|^q)$ with $q = 2$ to achieve Q-quadratic convergence, which is indeed the case, cf. Fig. 1.

In the second experiment, we show that even for nonnegative tensor approximation problems, the GN step can lead to a faster convergence. Similar to the previous experiment, random $10 \times 10 \times 10$ tensors of rank-5 are constructed using random factor matrices with entries drawn from $\mathcal{U}(0,1)$. In each factor matrix, ten entries are replaced by small negative entries drawn from $\mathcal{U}(-0.01, 0)$. Hence, no exact nonnegative CPD exists. Starting from a random initialization, the proposed

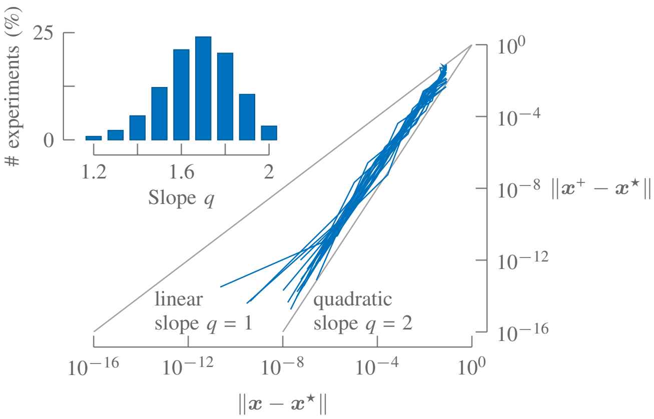

Figure 1. Up to Q-quadratic convergence can be achieved near the global optimum if an exact, unique solutions exists ($F(\boldsymbol{x}^\star) = \mathbf{0}$). The histogram shows the slope for the penultimate iteration for 500 experiments. The convergence curves for ten randomly chosen experiments have a slope close to 2. Results shown for a rank-5, $10 \times 10 \times 10$ tensor with a unique and nonnegative CPD, starting close to the global optimum.

method and standard proximal gradient descent are run until convergence ($\|\boldsymbol{r}\|^2 < \gamma\varepsilon$). To eliminate excess iterations due to nonoptimal stopping criteria, the number of gradient iterations is counted until the algorithm converges to $1.01 f(\boldsymbol{z}_{\text{final}})$, in which $\boldsymbol{z}_{\text{final}}$ is the value returned by the algorithm. (This mainly benefits proximal gradient descent.) As can be seen in Fig. 2, using the GN step clearly reduces the number of gradient evaluations, which are the dominant cost. Note that both algorithms may converge to local optima in which one or more of the rank-1 terms become zero.

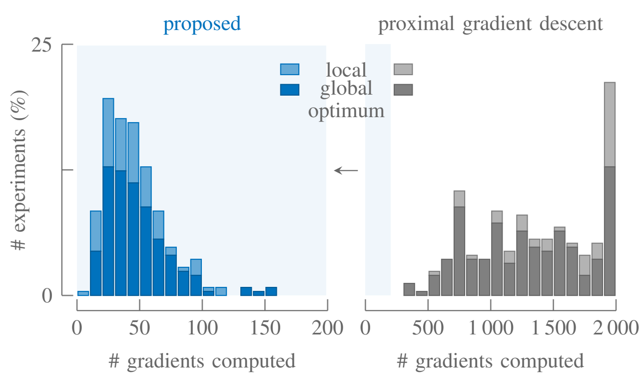

Figure 2. By using (approximate) second-order information as in the proposed algorithm, fewer gradients are computed compared to proximal gradient descent. The global optimum is attained in 67% (proposed) and 77% (PGD) of the cases. Histograms created using 250 experiments.

VII. Conclusion and future work

By combining nonnegativity and unit-norm constraints, a proximal Gauss–Newton (GN) type algorithm is derived. Global convergence is achieved by backtracking the GN step to the proximal gradient descent (PGD) step based on the forward-backward envelope function. While Q-quadratic convergence is only shown for the global optima in the case of an exact, essentially unique decomposition, the GN directions effectively reduce the computational cost compared to PGD.

In the current work we focused on theoretical properties; large-scale implementations are part of future work.

References

[1] A. Cichocki, D. Mandic, A.-H. Phan, C. Caiafa, G. Zhou, Q. Zhao, and L. De Lathauwer, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 145–163, Mar 2015.

[2] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3551–3582, July 2017.

[3] L. Sorber, M. Van Barel, and L. De Lathauwer, "Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-$(L_r, L_r, 1)$ terms, and a new generalization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 695–720, Apr 2013.

[4] N. Vervliet and L. De Lathauwer, "Numerical optimization based algorithms for data fusion," in *Data Fusion Methodology and Applications*, 1st ed., ser. Data Handling in Science and Technology, M. Cocchi, Ed. Elsevier, 2019, vol. 31, ch. 4, pp. 81–128.

[5] A. Themelis, L. Stella, and P. Patrinos, "Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms," *SIAM J. Optim.*, vol. 28, no. 3, pp. 2274–2303, Aug 2018.

[6] L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos, "A simple and efficient algorithm for nonlinear model predictive control," in *IEEE 56th Annu. Conf. Decision and Control (CDC)*, Dec 2017, pp. 1939–1944.

[7] A. Cichocki, R. Zdunek, A.-H. Phan, and S.-I. Amari, *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. UK: John Wiley, 2009.

[8] A. Cichocki and A.-H. Phan, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE Trans. Fundamentals*, vol. E92-A, no. 3, pp. 708–721, Mar 2009.

[9] E. C. Chi and T. G. Kolda, "On tensors, sparsity, and nonnegative factorizations," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 4, pp. 1272–1299, Dec 2012.

[10] R. Bro, "Multi-way analysis in the food industry: Models, algorithms, and applications," Ph.D. dissertation, University of Amsterdam, 1998.

[11] K. Huang, N. D. Sidiropoulos, and A. P. Liavas, "A flexible and efficient algorithmic framework for constrained matrix and tensor factorization," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5052–5065, June 2016.

[12] S. Hansen, T. Plantenga, and T. G. Kolda, "Newton-based optimization for Kullback–Leibler nonnegative tensor factorizations," *Optimization Methods and Software*, vol. 30, no. 5, pp. 1002–1029, Apr 2015.

[13] P. Paatero, "A weighted non-negative least squares algorithm for three-way "PARAFAC" factor analysis," *Chemometr. Intell. Lab.*, vol. 38, no. 2, pp. 223–242, Oct 1997.

[14] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer, "Tensorlab 3.0," Mar 2016, available online at https://www.tensorlab.net.

[15] C. Kelley, *Iterative Methods for Optimization*. SIAM, 1999.

[16] J.-P. Royer, N. Thirion-Moreau, and P. Comon, "Computing the polyadic decomposition of nonnegative third order tensors," *Signal Processing*, vol. 91, no. 9, pp. 2159–2171, Sep 2011.

[17] L. Sorber, M. Van Barel, and L. De Lathauwer, "Structured data fusion," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 586–600, June 2015.

[18] K. Huang and X. Fu, "Low-complexity proximal Gauss–Newton algorithm for nonnegative matrix factorization," in *IEEE Glob. Conf. Signal and Information Processing (GlobalSIP)*. IEEE, Nov 2019.

[19] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003, vol. II.

[20] A. Themelis, M. Ahookhosh, and P. Patrinos, "On the acceleration of forward-backward splitting via an inexact Newton method," in *Splitting Algorithms, Modern Operator Theory, and Applications*, H. H. Bauschke, R. S. Burachik, and D. R. Luke, Eds. Cham: Springer International Publishing, Nov 2019, pp. 363–412.

[21] P. Patrinos and A. Bemporad, "Proximal Newton methods for convex composite optimization," in *52nd IEEE Conf. Decision and Control*, Dec 2013, pp. 2358–2363.

[22] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2011, vol. 317.

[23] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2016.

[24] A. Themelis, "Proximal algorithms for structured nonconvex optimization," Ph.D. dissertation, KU Leuven, Dec 2018.