A QUADRATICALLY CONVERGENT PROXIMAL ALGORITHM FOR NONNEGATIVE TENSOR DECOMPOSITION

Nico Vervliet1 Andreas Themelis1 Panagiotis Patrinos1 Lieven De Lathauwer1,2

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

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 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 \( \mathbf{T} \) 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

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

in which factor matrix \( \mathbf{A}^{(n)} \) has \( a_r^{(n)} \) as its columns. (Element-wise, we have \( t_{i_1,\ldots,i_N} = \sum_{r=1}^{R} a_{i_{r_1}}^{(1)} \cdots a_{i_{r_N}}^{(N)} \).) 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., \( a_r^{(n)} \geq 0, \forall n, r \), where the inequality is meant elementwise.

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

\[
\text{minimize } f(x) \text{ with } f(x) := \frac{1}{2} \| F(x) \|^2, \]

where \( F(x) \) 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 \( F(x) \) it can exhibit up to quadratic rates of convergence. The idea behind GN is using the Gramian \( JF(x)^\top JF(x) \) as a surrogate for \( \nabla^2 f(x) \), which well approximates the true Hessian around solutions \( x^* \) whenever \( F(x^*) = 0 \). One iteration \( x \mapsto x^* \) of GN amounts to solving the linear system

\[
(JF(x)^\top JF(x))(x^* - x) = -JF(x)^\top F(x),
\]

in which \( JF(x)^\top F(x) \) is the gradient of \( f(x) \). 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 \( f(x) \), 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

\[
\text{minimize } \frac{1}{2} \| F(x) \|^2 \text{ subject to } x \geq 0.
\]

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].
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 [12] 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., $a$ and bold upper case, e.g., $A$, 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., $\mathbb{R}^n$, $n = 1, \ldots, N$. The Kronecker and Khatri–Rao (column-wise Kronecker) products are denoted by $\otimes$ and $\odot$, respectively. The notation $\text{blkdiag}(x_r^{(m)})$ is used for a block-diagonal matrix with blocks $x_1^{(r)}, x_2^{(r)}, \ldots, x_r^{(m)}$. The identity matrix is denoted by $I$, the column-wise concatenation of $a$ and $b$ by $[a; b]$, and the closed $\varepsilon$-ball around $x$ by $B(x, \varepsilon)$.

II. A semismooth Gaussian–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 $l_r$ of each term in the sum are scalar quantities that we collect in a vector $\lambda \in \mathbb{R}^R$, resulting in the normalized decomposition

$$\mathcal{T} = \sum_{r=1}^R l_r \cdot a_r^{(1)} \otimes \cdots \otimes a_r^{(N)} \quad \text{with} \quad \|a_r^{(n)}\| = 1, \forall n, r.$$  

The normalized version of problem (2) thus becomes

$$\text{minimize} \quad \frac{1}{2} \|F(\lambda, a)\|^2 \quad \text{subject to} \quad \begin{cases} \lambda, a \geq 0, \\ \|a_r^{(n)}\| = 1, \end{cases} \quad (3)$$

in which $F(\lambda, a) = \sum_{r=1}^R l_r \cdot a_r^{(1)} \otimes \cdots \otimes a_r^{(N)} - \mathcal{T}$.

$\lambda = [a_1^{(1)}; a_2^{(1)}; \ldots; a_r^{(1)}; \ldots; a_r^{(n)}; \ldots; a_r^{(N)}]$, and $d = R \sum_{r=1}^N L_r$. The feasible set $\mathcal{C} = \{(a, \lambda) \in \mathbb{R}^r \times \mathbb{R}^\lambda | a, \lambda \geq 0, \|a_r^{(n)}\| = 1\}$ is nonconvex, but projecting onto it is a simple block-separable operation: one has $\Pi_C(a, \lambda) = (\tilde{a}, \tilde{\lambda})$ where

$$\tilde{a}_r^{(n)} = \Pi_C(\lambda_r^{(n)}) = \begin{cases} a_r^{(n)}, & n \neq r, \\ a_r^{(n)}, & n = r. \end{cases}$$

Here, $\mathcal{S}$ and $\mathcal{O}_\circ$ denote the unit sphere and the positive orthant of suitable size, respectively, and $\{ \cdot \}$, $\text{max}(\{ \cdot \})$ elementwise; in case $\|a_r^{(n)}\| = 0$, the (set-valued) projection is easily seen to equal $\tilde{a}_r^{(n)} = \{ e_r \} | e_r = \text{arg max}_i a_r^{(i)} \}$. $e_r$ is the vector whose $r$th 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 $\mathcal{R}_\gamma(x) = 0$, with $x := (\lambda, a)$ as optimization variable and

$$\mathcal{R}_\gamma(x) := x - \Pi_C(x - \gamma JF(x)^\top F(x))$$

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. 7.1.1(iii)] gives

$$J_{\mathcal{R}_\gamma}(x) = I - \Pi_C(w) \cdot \left[ 1 - \gamma JF(x)^\top F(x) - \gamma \sum F_i(x) \nabla^2 F_i(x) \right],$$

where $F_i(x)$ is the $i$th element of vector $F(x)$,

$$w = x - \gamma JF(x)^\top F(x)$$

is a gradient descent step at $x$, and $J\Pi_C(w)$ 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 $\mathcal{R}_\gamma$ with

$$\mathcal{J}_{\mathcal{R}_\gamma}(x) := I - J\Pi_C(w) \cdot \left[ 1 - \gamma JF(x)^\top F(x) \right],$$

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

$$J_{\Pi_C([w]), i} = \begin{cases} 1 & \text{if } i \in {j \mid \omega_i > 0} \\ 0 & \text{otherwise.} \end{cases}$$

and (see [20, §15.6.2d])

$$J_{\Pi_C([w]), i} = \begin{cases} 1 & \text{if } i \in {j \mid \omega_i > 0} \\ 0 & \text{otherwise.} \end{cases}$$

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

$$x^0 \in B(x^*, \varepsilon), \quad x^{k+1} = x^k + d^k,$$

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

**Proof.** We start by remarking that the projection onto the (product of) sphere(s) is $C^\infty$ wherever it is well defined. Since $x^*$ is optimal, it follows from [5, Thm. 3.4(iii)] that $\mathcal{R}_\gamma(x^*) = \{0\}$, and that consequently the projection onto the spheres it entails, cf. (4), is well defined and is thus $C^\infty$ 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 $x^*$.

Next, observe that $\hat{H}_k := \hat{H}_k + \Delta(x^k) \in J\mathcal{R}_\gamma(x^k)$, for some $\Delta(x) \in J\Pi_C([w]) \sum F_i(x) \nabla^2 F_i(x)$ (with $w$ as in (6)) is a locally bounded quantity such that $\Delta(x) \to 0$ as $x \to x^*$. Therefore, denoting $d^k := x^{k+1} - x^k$,

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

1021
We have that \( \sup_{x \in J_{\gamma}(x)} \| \hat{H}^{-1} \| \) is bounded by a same quantity \( c_\varepsilon \) for all \( x \in B(x^*, \varepsilon) \) when \( \varepsilon \) is small enough, as it follows from [19, Lem. 7.5.2]. Consequently, for \( \varepsilon \) small enough [19, Thm. 7.5.5] guarantees that \( x^k \to x^* \) \( Q \)-linearly. In turn, this implies that \( \Delta(x^k) \to 0 \), hence invoking again the same result, the claimed \( Q \)-quadratic convergence is obtained.

\[ \Box \]

Theorem 1 requires that all matrices in \( \hat{J} \gamma (x^*) \) in (7) are nonsingular. In Theorem 3, we show that this is the case for the exact decomposition problem if the Gramian \( JF(x)^\top JF(x) \) 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 \( JF(x)^\top JF(x) \) has at least \( NR \) zero eigenvalues, and a basis \( K \) for the subspace corresponding to these zero eigenvalues is given by

\[ K = \text{blkdiag}(\{ \text{diag}(k^{(n)}(n) \circ A^{(n)}) \}, \text{diag}(k^{(N+1)}(n) \circ \Lambda^\top)) \doteq \sum_{k=1}^{N} k^{(n)} = 0. \]  

\[ \text{for } k^{(n)}(n) \in R^{n}, n = 1, \ldots, N + 1, \text{ and } \sum_{n=1}^{N+1} k^{(n)}(n) = 0. \]

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

\[ JF(x)K = \sum_{n=1}^{N} k^{(n)}(n) = 0. \]

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

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

**Proof.** In the global optimum \( x^* \), \( JF(x)^* \doteq 0 \) and \( x^* \in C \), hence \( f \in C \). Let \( P \) be a full-rank projection matrix. Let \( G = JF(x^*) \) and \( H = \hat{H} \). Before proving that \( \hat{H} \in \hat{J} \gamma (x^*) \) has full rank, we show that range(\( P \)) \( = \text{range}(P) \) is the case if range(\( G \)) \( \cap \text{null}(G) = 0. \) Let \( a_r \) be the factor vectors and scaling factors corresponding to \( a_r \). By assumption, \( G \) has \( NR \) zero eigenvalues and null(\( G \)) \( = \text{range}(K); \) see Lemma 2. Any \( b \in \text{range}(P) \) can be written as \( b = [b_1; \ldots; b_R] \) and \( b = \sum_{r=1}^{N} b^{(n)}(n) a_n^{(n)} \). If \( b^{(n)}(n) \) \( \perp a_n^{(n)} \) or \( b^{(n)}(n) = 0 \), see (8). If \( b \in \text{range}(K) \), then the following should hold with \( \sum_{n=1}^{N+1} k^{(n)} = 0: \)

\[ b^{(n)}(n) a_n^{(n)} \iff k^{(n)} = 0, \forall n, r, \]

\[ b^{(n)}(n) a_n^{(n)} = 0, \forall r, \]

\[ b^{(n)}(n + 1) a^{(n)}_r \iff k^{(n)}(n + 1) = 0, \forall r, \]

which is false, hence \( b \in \text{range}(K) \) and \( \text{range}(P) = \text{range}(P) \).

Let \( \hat{H} = (I - P) + \gamma P \), and \( \delta \) the number of active constraints for which \( J_{\gamma} \delta_{\gamma}(x^*) \). As range(\( PG \)) \( \neq \text{range}(P) \), we can show that there exists an \( NR + \delta \)-dimensional subspace \( U_1 \) of \( I - P \), and a \( d + R - N\delta - \delta \)-dimensional subspace \( U_2 \) of \( \text{null}(P) \), such that \( U_1 \cap U_2 = 0 \), \( U_1 \subseteq U_1 \), and \( U_2 \subseteq U_2 \). Therefore, range(\( (I - P) + \gamma P \)) \( = R^{d + R} \) and \( \hat{H} \) has full rank. \[ \Box \]

**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]

\[ \varphi^{\gamma}_{\psi}(x) = \frac{1}{2} \| F(x) \|^2 - \langle JF(x)^\top F(x), r \rangle + \frac{1}{2} \| r \|^2, \]

(13)

where \( y > 0 \) is a stepsize parameter.

\[ z = \Pi_{L} (x - \gamma JF(x)^\top F(x)), \quad \text{and} \quad r := x - z. \]

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 \( \gamma > 0 \), \( \varphi^{\gamma}_{\psi}(x) \) is locally Lipschitz continuous and real-valued. Moreover, denoting \( \varphi(z) := \frac{1}{2} \| F(x) \|^2 \), the following hold:

(i) \( \varphi^{\gamma}_{\psi}(x) \leq f(x) \) for all \( x \in C \).

(ii) For all \( x \in \mathbb{R}^d \), if \( z := \Pi_{L} (x - \gamma JF(x)^\top F(x)) \) satisfies

\[ f(z) \leq f(x) + \langle \nabla f(x), z - x \rangle + \frac{1}{2} \| x - z \|^2 \]

for some \( L > 0 \), then \( F(z) \leq \varphi^{\gamma}_{\psi}(x) - \frac{1}{2} \| x - z \|^2. \) In particular, \( \varphi^{\gamma}_{\psi}(x) \geq f(z) \geq 0 \) whenever \( \gamma \leq 1/L \).

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

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

\[ \varphi^{\gamma}_{\psi}(x) = \min_{v \in C} \{ f(x) + \langle \nabla f(x), v - x \rangle + \frac{1}{2} \| v - x \|^2 \} \]

\[ = f(x) - \frac{1}{2} \| \nabla f(x) \|^2 + \frac{1}{2} \| \nabla f(x) \| R^2 \]

hence \( \varphi^{\gamma}_{\psi}(x) \leq f(x) \) whenever \( x \in C \) (by simply replacing \( v = x \) in the minimization). Local Lipschitz continuity owes to that of \( f, \nabla f \) and \( \text{dist}(\cdot, C) \), see [22, Ex. 9.6]. Moreover, since the minimum above is obtained at \( v = z \), the second claim follows. Finally, since \( \Pi_{L} \) is locally bounded (cf. [22, Ex. 5.23(a)]) and so is \( \nabla f, \Pi_{L} \); \( \Pi_{L} \) maps the bounded set \( \Omega \) into a bounded set. Therefore, there exists a convex set \( \hat{\Omega} \) that contains any \( x \) and \( z = \Pi_{L} (x - \gamma JF(x)^\top F(x)) \) with \( x \in \Omega \). The claimed \( L_{\Omega} \) satisfying the last condition can thus be taken as the Lipschitz modulus of \( \nabla f \) over \( \hat{\Omega} \), see [23, Prop. A.24]. \[ \Box \]

**IV. A globally convergent algorithm**

**Lemma 4** contains all the key properties that lead to Algorithm 1, 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 \( x \in \mathbb{R}^{d+R} \); \( \alpha, \beta \in (0, 1) \); tolerance \( \varepsilon > 0 \); estimate of Lipschitz modulus \( L > 0 \)

Initialize \( \gamma = \frac{\alpha}{2} \)

1.0: \( z = \Pi_{C}[x - \gamma JF(x)^{\top} F(x)] \) and \( r = x - z \)
   if \( \frac{1}{2} \left\| F(z) \right\|^{2} > \varphi_{\mu}^{n}(x) - \frac{1-\alpha}{2\gamma} \left\| r \right\|^{2} \) then
   \( \gamma \leftarrow \gamma/2 \) and go back to step 1.0
   else \( \varphi_{\mu}^{n}(x) > \varphi_{\mu}^{n}(x) - \frac{1-\alpha}{2\gamma} \left\| r \right\|^{2} \) then
   \( \gamma \leftarrow \gamma/2 \) and go back to step 1.0
   \( \tau \leftarrow \gamma/2 \) and go back to step 1.3

1.1: if \( \frac{1}{2} \left\| r \right\|^{2} \leq \varepsilon \) then
   return \( z \)

1.2: Pick \( \bar{d} \in J_{R}(x) \) and let \( d \) be such that \( \bar{d} \Pi_{R} \bar{d} = -\bar{d} \Pi_{R}(x) \)
   Set stepsize \( \tau = 1 \)

1.3: \( x^{+} = (1 - \tau) z + \tau(x + d) \)
   \( z^{+} = \Pi_{C}[x - \gamma JF(x^{+})^{	op} F(x^{+})] \) and \( r^{+} = x^{+} - z^{+} \)

1.4: if \( \frac{1}{2} \left\| F(z^{+}) \right\|^{2} > \varphi_{\mu}^{n}(x^{+}) - \frac{1-\alpha}{2\gamma} \left\| r \right\|^{2} \) then
   \( \gamma \leftarrow \gamma/2 \) and go back to step 1.0
   else \( \varphi_{\mu}^{n}(x^{+}) > \varphi_{\mu}^{n}(x^{+}) - \frac{1-\alpha}{2\gamma} \left\| r \right\|^{2} \) then
   \( \tau \leftarrow \gamma/2 \) and go back to step 1.3

1.5: \( (x, z, r) \leftarrow (x^{+}, z^{+}, r^{+}) \) and proceed to step 1.1

assumptions of \( R_{\gamma} \) 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 \( z \) remains bounded (as can be enforced by intersecting \( C \) with any large box, cf. (16)), then Algorithm I terminates in finitely many iterations. Moreover, with tolerance \( \varepsilon = 0 \) the following hold:

(i) \( \gamma \) is reduced only finitely many times and the sequence of points \( z \) 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 \( z \) 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 of the globalization strategy. The algorithm revolves around the upper bound (15); since the modulus of the globalization strategy. The algorithm revolves around the upper bound (15); since the modulus of the globalization strategy. The algorithm revolves around the upper bound (15); since the modulus of the globalization strategy.
method and standard proximal gradient descent are run until convergence ($||r||^2 < \epsilon$). To eliminate excess iterations due to nonoptimal stopping criteria, the number of gradient iterations is counted until the algorithm converges to $1.01f(z_{\text{final}})$, in which $z_{\text{final}}$ 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.

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