Journal of Optimization Theory and Applications (2023) 199:1130–1157
https://doi.org/10.1007/s10957-023-02310-4

Nested Alternating Minimization with FISTA for Non-convex and Non-smooth Optimization Problems

Eyal Gur[1] · **Shoham Sabach**[1] · **Shimrit Shtern**[1]

Received: 21 December 2022 / Accepted: 12 September 2023 / Published online: 3 October 2023

Abstract

Motivated by a recent framework for proving global convergence to critical points of nested alternating minimization algorithms, which was proposed for the case of smooth subproblems, we first show here that non-smooth subproblems can also be handled within this framework. Specifically, we present a novel analysis of an optimization scheme that utilizes the FISTA method as a nested algorithm. We establish the global convergence of this nested scheme to critical points of non-convex and non-smooth optimization problems. In addition, we propose a hybrid framework that allows to implement FISTA when applicable, while still maintaining the global convergence result. The power of nested algorithms using FISTA in the non-convex and non-smooth setting is illustrated with some numerical experiments that show their superiority over existing methods.

Keywords Non-convex and non-smooth optimization · Alternating minimization · Global convergence · Nested algorithms · FISTA

Mathematics Subject Classification 90C06 · 90C26 · 90C30 · 90C90

Communicated by Russel Luke.

✉ Eyal Gur
eyal.gur@campus.technion.ac.il

Shoham Sabach
ssabach@technion.ac.il

Shimrit Shtern
shimrits@technion.ac.il

[1] Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, 3200003 Haifa, Israel

1 Introduction

Non-convex and non-smooth optimization has garnered significant attention in recent years across various applied fields [8–10, 12, 14, 17, 29]. However, the available tools for proving convergence to globally optimal solutions for non-convex and non-smooth (namely, not continuously differentiable) optimization problems are extremely limited. Consequently, in the past decade, a new trend of research in this domain has concentrated on developing methodologies to establish global convergence of algorithms toward critical points of non-convex and non-smooth minimization problems (as the criticality of a point serves as a necessary condition for its optimality). To clarify, by global convergence, we mean that the entire sequence of iterations generated by the algorithm in question converges to a unique point, which is also a critical point of the minimized objective function.

In this paper, our focus is on problems that possess a general and widely used block structure for the variables (please refer to the exact formulation below). A popular strategy for addressing high-dimensional non-convex and non-smooth problems with such a block structure is to employ the well-known alternating minimization (AM) scheme. In this scheme, the original high-dimensional problem is divided into several lower-dimensional subproblems, and the vector of variables is partitioned into distinct blocks. The AM scheme then iterates between these blocks, concentrating on the minimization of one block at a time while keeping all other blocks fixed.

However, even solving these subproblems can be challenging at times. To overcome this challenge, an alternative approach is to approximate the solution of each subproblem instead of solving them exactly. Several works in the last few years (such as [5, 22] and references therein) have demonstrated that by approximating the solution of each subproblem using one iteration of a sufficient decrease algorithm, such as the classical proximal gradient method, an overall globally convergent algorithm that reaches a critical point of the objective function can be obtained. This approach is known as the one iteration approximation (OIA) approach. For theoretical convergence results regarding OIA, along with relevant references, we refer the reader to [11], which incorporates OIA within the framework of AM.

An alternative approach to approximating solutions of subproblems within the AM scheme is to apply multiple inner iterations of a nested algorithm. This approach is known as the nested approximation (NA) approach. Empirical evidence from several works (such as [6, 11, 25]) has demonstrated that the NA approach can converge to superior solutions in terms of objective function values while maintaining low computational complexity compared to the OIA approach.

A highly desirable property of the NA approach, which is notoriously challenging to achieve, is the global convergence of the overall algorithm. Recall that by global convergence, we mean that the entire algorithm (i.e., the AM scheme integrated with the nested sub-algorithms) converges to a critical point of the objective function.

Several studies have tackled the issue of global convergence of AM with NA algorithms in the context of non-convex and non-smooth optimization problems. One such technique involves utilizing the essentially cyclic rule (ECR) of [26]. In ECR, a predefined total number $T \in \mathbb{N}$ of inner iterations is set. Subsequently, the subproblems are minimized using nested algorithms in a cyclic order, ensuring that the total number of

inner iterations across all subproblems is exactly T. For instance, in [25] (also refer to [6]), the authors demonstrate that ECR with the sufficient decrease Bregman proximal gradient method as a nested algorithm globally converges to a critical point of the objective function.

Another technique was recently proposed in our paper [11]. This novel technique allowed for the first time to incorporate non-descent algorithms in the AM scheme. Specifically, we introduced a class of nested alternating minimization (NAM) algorithms and demonstrated that global convergence can be achieved using any nested algorithm for each subproblem, provided that the chosen nested algorithms satisfy specific block-wise conditions and that the subproblems are continuously differentiable. The introduction of NAM in [11] represents an important contribution, as it allows for the nesting of non-descent methods, including the accelerated gradient method by Nesterov [19], the heavy ball method by Polyak [23] and other accelerated variants of gradient descent.

1.1 Our Contribution

In this work, we make several contributions to the field of non-convex optimization within the context of the nested alternating minimization (NAM) framework. These contributions expand the applicability of NAM to address non-smooth subproblems and provide enhanced convergence results. Our main contributions can be summarized as follows:

- Extension to Non-Smooth Subproblems: We extend previous results that focused on strongly convex subproblems to the non-smooth setting, where the function is not continuously differentiable. Specifically, we explore the integration of the FISTA method [4], which is a non-descent and accelerated variant of the proximal gradient method, as a nested algorithm within the NAM framework to handle non-differentiable subproblems.
- Hybrid NAM Framework with FISTA: We introduce a hybrid NAM framework that allows for the utilization of FISTA even in the non-convex setting. At each outer iteration, the framework selects FISTA as a nested accelerated algorithm if the subproblem can be verified to be strongly convex. If not, any sufficient decrease method can be applied.

These new theoretical results are made possible by leveraging our recent block-wise variant, first introduced in [11], of the well-known global convergence methodology presented in [5]. Importantly, our work demonstrates that the block-wise conditions can be satisfied for non-descent methods even in the absence of differentiability, expanding the scope of applicability beyond the traditional smooth and strongly convex subproblems. Notably, to the best of our knowledge, this is the first work that establishes global convergence to critical points for non-descent methods in the presence of a non-differentiable objective function.

While we specifically consider the popular non-descent and accelerated FISTA method in our analysis, we believe that the insights and techniques presented in our paper can be extended to other accelerated non-descent methods that are based on

the proximal gradient step. This opens up new possibilities for applying optimization techniques in the non-differentiable setting.

This paper is organized as follows: In Sect. 2 we formulate the problem and review the NAM optimization scheme together with the FISTA method, while in Sect. 3 we provide an analysis of FISTA when nested in the NAM scheme and prove global convergence to critical points results. In Sect. 4, we provide a numerical analysis, and in Sect. 5, we give concluding remarks.

2 The NAM Scheme with FISTA

We consider the following composite model for minimizing a bounded from below function $F\colon \mathbb{R}^d \times \mathbb{R}^{d_0} \to (-\infty, \infty]$, which consists of two main block variables $\mathbf{z} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathbb{R}^{d_0}$

$$\min_{(\mathbf{z},\mathbf{u})\in\mathbb{R}^d\times\mathbb{R}^{d_0}} F(\mathbf{z},\mathbf{u}) \equiv G(\mathbf{z},\mathbf{u}) + \sum_{i=1}^{p} g_i(\mathbf{z}_i) + g_0(\mathbf{u}), \tag{P}$$

where the block $\mathbf{z} \equiv \left(\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_p^T\right)^T \in \mathbb{R}^d$ is decomposed into $p \in \mathbb{N}$ sub-blocks $\mathbf{z}_i \in \mathbb{R}^{d_i}$, $i = 1, 2, \ldots, p$. The function $G\colon \mathbb{R}^d \times \mathbb{R}^{d_0} \to \mathbb{R}$ is continuously differentiable and (possibly) non-convex, and each function $g_i\colon \mathbb{R}^{d_i} \to (-\infty, \infty]$, $i = 0, 1, \ldots, p$, is an extended real-valued, proper, lower semi-continuous, (possibly) non-convex and non-differentiable function. Notice that using the functions g_i, $i = 0, 1, \ldots, p$, one can introduce constraints to Problem (P). In order to obtain convergence results, we further assume that F satisfies the Kurdyka-Łojasiewicz (KL) property [15, 16], which is a common structural assumption in convergence analysis as many functions, including semi-algebraic and real sub-analytic functions, posses this property [1].

The set of critical points of the function F, denoted by $\operatorname{crit}(F)$, is defined as the set of points where the sub-differential set contains the zero vector. Mathematically,

$$\operatorname{crit}(F) = \left\{(\mathbf{z},\mathbf{u}) \in \mathbb{R}^d \times \mathbb{R}^{d_0} : \mathbf{0}_{d+d_0} \in \partial F(\mathbf{z},\mathbf{u})\right\}.$$

In this paper, since we are dealing with non-convex and non-smooth functions, when we write ∂F we mean the limiting sub-differential of F [18].

As previously mentioned, the nested alternating minimization (NAM) scheme of [11] is employed to tackle large-scale optimization problems with a block structure. In the case of Problem (P), applying the NAM scheme involves first finding the exact minimizer of the partial function $\mathbf{u} \mapsto F(\mathbf{z},\mathbf{u})$, followed by minimizing the p partial functions $\mathbf{z}_i \mapsto F(\mathbf{z},\mathbf{u})$, $i = 1, 2, \ldots, p$, in a sequential order using some chosen nested algorithm. This means that in any (outer) iteration $k \geq 0$, we apply inner iterations of some nested algorithm to minimize these partial functions.

Mathematically, for a specific sub-block $i = 1, 2, \ldots, p$ and an outer iteration $k \geq 0$, the minimization problem associated with the partial function

$$\mathbf{z}_i \mapsto F\left(\mathbf{z}_1^{k+1}, \ldots, \mathbf{z}_{i-1}^{k+1}, \mathbf{z}_i, \mathbf{z}_{i+1}^{k}, \ldots, \mathbf{z}_p^{k}, \mathbf{u}^{k+1}\right), \tag{1}$$

can be represented as

$$\min_{\mathbf{x}\in\mathbb{R}^{d_i}} \quad \Psi_i^k(\mathbf{x}) \equiv \varphi_i^k(\mathbf{x}) + g_i(\mathbf{x}). \tag{P_i^k}$$

Here, the function $\varphi_i^k : \mathbb{R}^{d_i} \to \mathbb{R}$ is continuously differentiable with L_i^k-Lipschitz continuous gradient, for some $L_i^k > 0$, while the function $g_i : \mathbb{R}^{d_i} \to (-\infty, \infty]$ is non-smooth. Specifically, by setting $\mathbf{x} = \mathbf{z}_i$ in

$$\varphi_i^k(\mathbf{x}) = G\left(\mathbf{z}_1^{k+1}, \ldots, \mathbf{z}_{i-1}^{k+1}, \mathbf{x}, \mathbf{z}_{i+1}^{k}, \ldots, \mathbf{z}_p^{k}, \mathbf{u}^{k+1}\right), \tag{2}$$

we recover the partial function in (1). It should be emphasized that for any $i = 1, 2, \ldots, p$ and $k \geq 0$, the functions Ψ_i^k and φ_i^k in Problem (P_i^k) depend on the p vectors

$$\begin{aligned}\left(\mathbf{z}_1^{k+1}, \ldots, \mathbf{z}_{i-1}^{k+1}, \mathbf{z}_{i+1}^{k}, \ldots, \mathbf{z}_p^{k}, \mathbf{u}^{k+1}\right) \subset \mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_{i-1}} \\ \times \mathbb{R}^{d_{i+1}} \times \ldots \times \mathbb{R}^{d_p} \times \mathbb{R}^{d_0} \equiv \mathbb{E}_i.\end{aligned} \tag{3}$$

Given the structure of Problem (P_i^k) as a sum of a continuously differentiable function φ_i^k with a Lipschitz continuous gradient, and non-smooth function g_i, a natural candidate for solving this subproblem is the popular FISTA method of [4], which is an accelerated and non-descent variant of the well-known proximal gradient method. Hence, in this paper, our focus is on utilizing the FISTA method as a nested algorithm to approximate solutions of the corresponding subproblems. The resulting optimization scheme for solving Problem (P) is outlined in Algorithm 1.

Remark 2.1 (i) Algorithm 1 is a specific instance of the broader NAM framework of [11]. In Algorithm 1, the subproblems are minimized using FISTA, which is the main focus of this work. However, NAM scheme allows the utilization of any nested algorithm, and it also allows to mix nested algorithms for different blocks $\mathbf{z}_i$.

(ii) For simplicity, Algorithm 1 specifies the order of variable updates as $\mathbf{u} \to \mathbf{z}_1 \to \ldots \to \mathbf{z}_p$. However, it is important to note that this update order is arbitrary, and any other order of updating the $p + 1$ variables $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p$, and $\mathbf{u}$ can be chosen. Additionally, in some outer iteration $k \geq 0$, it is possible to choose not to update one of the subproblems, as long as all subproblems are updated infinitely many times across all outer iterations.

To provide a clearer understanding of the NAM with FISTA method presented in Algorithm 1, it is essential to delve into the concept of outer and inner iterations.

Algorithm 1 Nested Alternating Minimization (NAM) Scheme With FISTA

1: **Initialization:** $\left(\mathbf{z}^0, \mathbf{u}^0\right) \in \mathbb{R}^d \times \mathbb{R}^{d_0}$ and a sequence of integers $\left\{j_i^k\right\}_{k\geq 0}$ (see (4)).
2: **for** $k \geq 0$ **do**
3: Update $\mathbf{u}^{k+1} \in \operatorname{argmin}\left\{F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \ldots, \mathbf{z}_p^k, \mathbf{u}\right) : \mathbf{u} \in \mathbb{R}^{d_0}\right\}$.
4: **for** $i = 1, 2, \ldots, p$ **do**
5: Set $\mathbf{z}_i^{k,0} = \mathbf{z}_i^k \in \mathbb{R}^{d_i}$, $\mathbf{y}^{k,0} = \mathbf{z}_i^{k,0}$ and $t_0 = 1$.
6: **for** $j = 0, 1, \ldots, j_i^k - 1$ **do**
7: Update $\mathbf{z}_i^{k,j+1} = \operatorname{argmin}\left\{g_i(\mathbf{y}) + \frac{L_i^k}{2}\left\|\mathbf{y} - \mathbf{y}^{k,j} + \frac{1}{L_i^k}\nabla\varphi_i^k\left(\mathbf{y}^{k,j}\right)\right\|^2 : \mathbf{y} \in \mathbb{R}^{d_i}\right\}$.
8: Set $t_{j+1} = \frac{1+\sqrt{1+4t_j^2}}{2}$.
9: Set $\mathbf{y}^{k,j+1} = \mathbf{z}_i^{k,j+1} + \frac{t_j - 1}{t_{j+1}}\left(\mathbf{z}_i^{k,j+1} - \mathbf{z}_i^{k,j}\right)$.
10: **end for**
11: Set $\mathbf{z}_i^{k+1} = \mathbf{z}_i^{k,j_i^k}$.
12: **end for**
13: Define $\mathbf{z}^{k+1} = \left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^{k+1}, \ldots, \mathbf{z}_p^{k+1}\right)$.
14: **end for**

Specifically, when focusing on the sub-block $\mathbf{z}_i$, where $i = 1, 2, \ldots, p$, within each outer iteration of NAM (indexed by $k \geq 0$), FISTA is employed to minimize the corresponding partial functions mentioned in (1). This implies that in each outer iteration, a finite sequence of inner FISTA iterations is executed for each sub-block. As a result, the sequence $\mathbf{z}_i^{k,j}$, where $j = 1, 2, \ldots, j_i^k$, is generated by applying FISTA to solve the subproblem defined in Problem $\left(\mathrm{P}_i^k\right)$.

It is important to note that in step 5 of Algorithm 1, for any outer iteration $k \geq 0$, the initial point of FISTA, which is $\mathbf{z}_i^{k,0}$, is set to be $\mathbf{z}_i^k$, and the last inner iterate in step 11 corresponds to the next outer iterate, i.e., $\mathbf{z}_i^{k,j_i^k} = \mathbf{z}_i^{k+1}$. This ensures the seamless continuation of the optimization process from one outer iteration to the next.

Before we proceed, a few comments about the number of inner iterations j_i^k (see step 6 in Algorithm 1). For any outer iteration $k \geq 0$, the number j_i^k is the number of FISTA iterations applied to minimize Ψ_i^k in Problem $\left(\mathrm{P}_i^k\right)$. Following [11, Section 4.2.1], for any $k \geq 0$, we use the following rule for determining the number of inner iterations

$$j_i^k = s + 2^{\lfloor k/r \rfloor} - 1, \tag{4}$$

where s and r are some fixed and predefined integers. One may use different integers for different blocks $\mathbf{z}_i$, but here for simplicity we take them to be equal. This rule is quite general due to the flexibility in choosing the parameters s and r.

In this paper, our objective is to prove that the sequence generated by NAM with FISTA (Algorithm 1) globally converges to a critical point of the original non-convex and non-differentiable function F. To achieve this, we utilize the global convergence theory presented in [11]. This theory states that if the generated sequence by the

algorithm satisfies certain block-wise conditions (see conditions (B1)–(B4) in Sect. 3 for details), then the sequence globally converges to a critical point of F.

In [11], it was shown that the block-wise conditions hold for a wide class of nested algorithms (and in particular, by the non-descent accelerated gradient method of Nesterov [19]), when the subproblems are strongly convex and continuously differentiable (i.e., $\Psi_i^k(\mathbf{x})$ is strongly convex and $g_i \equiv 0$). In this paper, we extend these results by proving that FISTA also fulfills the block-wise conditions in the setting where the subproblems are strongly convex and *non-differentiable*. Consequently, one can opt to minimize a subproblem using a nested accelerated algorithm, irrespective of its differentiability.

3 Convergence Analysis

In this section, we present novel results that establish the global convergence of NAM with FISTA (Algorithm 1) to critical points of the original non-convex and non-differentiable objective function F in Problem (P). Notably, unlike existing frameworks such as ECR, the nested algorithm employed here is *not* a sufficient decrease method.

Before we proceed, we provide a brief overview of the block-wise conditions for global convergence to critical points, as outlined in [11]. As mentioned above, these conditions serve as necessary requirements for a nested algorithm to ensure convergence of the NAM scheme toward critical points of the objective function F. The summarized block-wise conditions are as follows:

(B1) Overall approximate sufficient decrease in function value: The subproblem should exhibit a decrease in function value between the input and output vectors proportional to the squared norm of the difference between these vectors, up to a squared root of an error term.
(B2) Approximate sub-gradient bound: a sub-gradient of the subproblem should be approximately bounded by the norm of the gap between the input and output vectors, up to an error term.
(B3) Upper semi-continuity property of F with respect to the generated iterations.
(B4) Summable errors.

This section is organized into two subsections. In the first subsection, we utilize the block-wise conditions and prove the global convergence of NAM with FISTA to critical points in the setting of strong convexity and non-differentiability of the involved subproblems. Subsequently, in the second subsection, we relax the strong convexity assumption by introducing a hybrid NAM scheme, denoted as H-NAM. The H-NAM scheme incorporates a verification step for strong convexity of each Problem (P_i^k) at every outer iteration. If the involved subproblem is found to be strongly convex, FISTA is employed. Otherwise, a sufficient decrease method is utilized while preserving the global convergence property.

In the analysis to follow, we will make the following assumption regarding the continuously differentiable functions φ_i^k, for all $k \geq 0$ and for all $i = 1, 2, \ldots, p$ (see

(2)). Recall that each such function has L_i^k-Lipschitz continuous gradient for some $L_i^k > 0$, and this constant depends on the set $\mathbb{E}_i$ (see (3)).

Assumption 1 On any compact subset of $\mathbb{E}_i$, there exists $\bar{L}_i > 0$ such that $L_i^k \leq \bar{L}_i$, for all $k \geq 0$.

In some practical cases, the values of L_i^k and $\bar{L}_i$ may be unknown. In such cases, a backtracking procedure could be used for the update step within FISTA [2].

3.1 The Strongly Convex Setting

Throughout this subsection, in addition to Assumption 1, we make the following structural assumption.

Assumption 2 (a) The functions $\varphi_i^k : \mathbb{R}^{d_i} \to \mathbb{R}$ and $g_i : \mathbb{R}^{d_i} \to (-\infty, \infty]$ are convex over their domain.
(b) The function $\Psi_i^k : \mathbb{R}^{d_i} \to \mathbb{R}$ is σ_i^k-strongly convex for some $\sigma_i^k > 0$.
(c) On any compact subset of $\mathbb{E}_i$, there exist $\underline{\sigma}_i > 0$ such that $\underline{\sigma}_i \leq \sigma_i^k$ for all $k \geq 0$.

It should be noted that the strong convexity of the function Ψ_i^k, for $k \geq 0$, can arise from either the continuously differentiable function φ_i^k or the non-differentiable function g_i.

Now, we are ready to state the main result of this section, which states that NAM with FISTA globally converges to critical points.

Theorem 3.1 *Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ be a bounded sequence generated by NAM with FISTA (Algorithm 1) under Assumptions 1 and 2, where $j_i^k \in \mathbb{N}$ is set according to (4). Then, there exists a unique $\mathbf{z}^* \in \mathbb{R}^d$ such that $\left\{\mathbf{z}^k\right\}_{k\geq 0}$ converges to $\mathbf{z}^*$. In addition, if $\mathbf{u}^* \in \mathbb{R}^{d_0}$ is a limit point of the sequence $\left\{\mathbf{u}^k\right\}_{k\geq 0}$, then $(\mathbf{z}^*, \mathbf{u}^*)$ is a critical point of the function F of Problem (P).*

To prove the global convergence of the sequence $\left\{\mathbf{z}^k\right\}_{k\geq 0}$ to a critical point of the original non-convex and non-differentiable function F (as formulated in Theorem 3.1), we utilize the global convergence theory presented in [11]. To this end, and as mentioned above, we will prove that the generated sequence $\left\{\mathbf{z}^k\right\}_{k\geq 0}$ satisfies the block-wise conditions. Since these conditions are block-wise, it is sufficient to establish them for each block separately.

Therefore, to simplify the analysis, we focus on a single block and omit the index i in the discussion that follows. Thus, we consider one subproblem with the following structure (replacing Problem $\left(\mathrm{P}_i^k\right)$)

$$\min_{\mathbf{x}\in\mathbb{R}^{d_i}} \quad \Psi^k(\mathbf{x}) \equiv \varphi^k(\mathbf{x}) + g(\mathbf{x}). \tag{P^k}$$

With the notations introduced earlier, we denote the sequence of inner iterations $\mathbf{z}_i^{k,j}$, where $j = 1, 2, \ldots, j_i^k$, simply as $\mathbf{x}^{k,j}$, where $j = 1, 2, \ldots, j^k$, for a specific

outer iteration $k \geq 0$. Similarly, the sequence of outer iterations $\left\{\mathbf{z}_i^k\right\}_{k\geq 0}$ is denoted as $\left\{\mathbf{x}^k\right\}_{k\geq 0}$. Notice that $\mathbf{x}^{k,j^k} = \mathbf{x}^{k+1} = \mathbf{z}_i^{k+1}$ represents the output obtained by FISTA after j^k iterations for minimizing Problem $\left(\mathrm{P}^k\right)$, with an initial point $\mathbf{x}^{k,0} = \mathbf{x}^k = \mathbf{z}_i^k$. Also, we omit the index i and replace L_i^k, $\bar{L}_i$, σ_i^k and $\underline{\sigma}_i$ with L^k, $\bar{L}$, σ^k and $\underline{\sigma}$, respectively.

As for the number of inner FISTA iterations j^k, $k \geq 0$, notice that if the sequence generated by NAM is bounded, then from Assumptions 1 and 2 (c) it follows that $\underline{\sigma} \leq \sigma^k \leq L^k \leq \bar{L}$. We denote $\kappa \equiv \bar{L}/\underline{\sigma} \geq 1$. The value of κ, or an upper bound of it, can be unknown or difficult to compute. However, since this value is constant, then by the update rule of (4) we know that there exists $K \in \mathbb{N}$ such that for any $k \geq K$ it holds that

$$j^k \geq 2\sqrt{\kappa}. \tag{5}$$

To simplify notations in the proof of Theorem 3.1, we assume that inequality (5) is satisfied for $K = 0$.

An important property of FISTA that we use throughout the analysis below is the rate of convergence in terms of function values. Fixing some $k \geq 0$, following [4], we have for any $j \geq 0$ that

$$\Psi^k\left(\mathbf{x}^{k,j}\right) - \Psi^k\left(\mathbf{x}_*^k\right) \leq \beta_{\mathrm{F}}^{k,j}\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\|^2, \tag{6}$$

where $\mathbf{x}_*^k$ is the unique minimizer of the strongly convex function Ψ^k, and

$$\beta_{\mathrm{F}}^{k,j} \equiv \frac{2L^k}{(j+1)^2} \leq \frac{2\bar{L}}{(j+1)^2}, \tag{7}$$

where the inequality follows from Assumption 1.

The FISTA method satisfies other fundamental properties that we use throughout the analysis below. However, in order to make the reading of this subsection simple, we state and prove these results in appendix (see Lemmas A.1 and A.2 in Appendix A).

In the following result, we state three inequalities regarding the boundedness of the sequences generated by FISTA when nested in NAM. Due to the technical nature of its proof, we also postpone it to Appendix B.

Proposition 3.1 *For all $k \geq 0$, let $\left\{\mathbf{x}^{k,j}\right\}_{j\geq 0}$ and $\left\{\mathbf{y}^{k,j}\right\}_{j\geq 0}$ be sequences generated by FISTA (steps 7–9 in Algorithm 1) for minimizing Problem $\left(\mathrm{P}^k\right)$. Assume that the sequence generated by Algorithm 1 is bounded. Then, there exists $M > 0$ such that for any $k \geq 0$ it holds that*

(i) $\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| \leq M$.
(ii) $\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\| \leq M$.
(iii) $\left\|\nabla\varphi^k\left(\mathbf{y}^{k,j}\right) - L^k\mathbf{y}^{k,j}\right\| \leq M$ *for any* $j \geq 0$.

Now we can proceed to prove the main result stated in Theorem 3.1. Following the approach in [11, Theorem 2], our task is to prove that the block-wise conditions (B1)–(B4), as mathematically defined in [11], hold true. By establishing the fulfillment of these conditions, we establish the global convergence of the NAM with FISTA scheme.

Proof of Theorem 3.1 First, recall that $\mathbf{x}^k = \mathbf{x}^{k,0} = \mathbf{z}_i^{k,0}$ and that $\mathbf{x}^{k+1} = \mathbf{x}^{k,j^k} = \mathbf{z}_i^{k,j^k}$. The point $\mathbf{x}^k$ is the initial point of FISTA when applied to minimize the strongly convex function Ψ^k of Problem $\left(\mathrm{P}^k\right)$, and the point $\mathbf{x}^{k+1}$ is defined as the output obtained after j^k iterations. In addition, recall that $\mathbf{x}_*^k$ is the minimizer of Ψ^k. We now prove conditions (B1)–(B4) as they are stated in [11].

Now we prove condition (B1). To this end, we first need to show that

$$\Psi^k\left(\mathbf{x}^k\right) - \Psi^k\left(\mathbf{x}^{k+1}\right) \geq 0. \tag{8}$$

Since the function Ψ^k is strongly convex, from Lemma A.1(i) and Assumption 2(c) we get

$$\Psi^k\left(\mathbf{x}^k\right) - \Psi^k\left(\mathbf{x}_*^k\right) \geq \frac{\sigma^k}{2}\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\|^2 \geq \frac{\underline{\sigma}}{2}\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\|^2. \tag{9}$$

In addition, from (6) and (7) we have

$$\Psi^k\left(\mathbf{x}^{k+1}\right) - \Psi^k\left(\mathbf{x}_*^k\right) \leq \frac{2\bar{L}\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\|^2}{\left(j^k+1\right)^2} \leq \frac{\bar{L}\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\|^2}{2\kappa}, \tag{10}$$

where the last inequality follows from (5). Combining (9) and (10), we get (recall that $\kappa = \bar{L}/\underline{\sigma}$)

$$\Psi^k\left(\mathbf{x}^{k+1}\right) - \Psi^k\left(\mathbf{x}_*^k\right) \leq \Psi^k\left(\mathbf{x}^k\right) - \Psi^k\left(\mathbf{x}_*^k\right),$$

and the required inequality (8) is established. Now, from Lemma A.2(ii) and Proposition 3.1(i), (ii) together with (7) and Assumption 2(c) we get

$$\begin{aligned}
\Psi^k\left(\mathbf{x}^k\right) - \Psi^k\left(\mathbf{x}^{k+1}\right) &\geq \frac{\sigma^k}{2}\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|^2 - \sqrt{2\sigma^k\beta_{\mathrm{F}}^{k,j^k}}M^2 - \beta_{\mathrm{F}}^{k,j^k}M^2 \\
&\geq \frac{\underline{\sigma}}{2}\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|^2 - \frac{2M^2\bar{L}}{j^k+1} - \frac{2M^2\bar{L}}{\left(j^k+1\right)^2} \\
&\geq \frac{\underline{\sigma}}{2}\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|^2 - \frac{4M^2\bar{L}}{j^k+1}.
\end{aligned} \tag{11}$$

For any $k \geq 0$, we denote

$$e_1^{k,j} \equiv \frac{4M^2\bar{L}}{j+1} \geq 0,$$

and $e_1^{k+1} \equiv e_1^{k,j^k}$. Combining (8) and (11), we get for all $k \geq 0$ that

$$\Psi^k\left(\mathbf{x}^k\right) - \Psi^k\left(\mathbf{x}^{k+1}\right) \geq \max\left\{0, \frac{\sigma}{2}\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\|^2 - e_1^{k+1}\right\}, \tag{12}$$

which is the required condition (B1) with $\rho_1 = \sigma/2$ (where we follow the notations of [11]).

Now we prove condition (B2). From Proposition 3.1(ii) and Lemma A.2(iii) together with (7), we get (recall that $j^k \geq 2$ for all $k \geq 0$)

$$\left\|\mathbf{w}^{k+1}\right\| \leq 8L^k\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\|\sqrt{\frac{2\beta_{\mathrm{F}}^{k,j^k-2}}{\sigma^k}} \leq \frac{16M\bar{L}\sqrt{\kappa}}{j^k - 1}, \tag{13}$$

for some $\mathbf{w}^{k+1} \equiv \mathbf{w}^{k,j^k} \in \partial\Psi^k\left(\mathbf{x}^{k+1}\right)$. For any $k \geq 0$, we denote

$$e_2^{k,j} \equiv \frac{16M\bar{L}\sqrt{\kappa}}{j-1} \geq 0,$$

and $e_2^{k+1} \equiv e_2^{k,j^k}$. The required condition (B2) now follows from (13) with $\rho_2 = 0$.

Now we prove condition (B3). To this end, let $\left\{\mathbf{x}^{k_l}\right\}_{l\in\mathcal{K}_i\subseteq\mathbb{N}}$ be a subsequence of $\left\{\mathbf{x}^k\right\}_{k\geq 0}$, which converges to some limit point $\bar{\mathbf{x}}$. It follows from step 7 in Algorithm 1 that

$$\begin{aligned} &g\left(\mathbf{x}^{k+1}\right) + \frac{L^k}{2}\left\|\mathbf{x}^{k+1} - \mathbf{y}^{k,j^k-1} + \frac{1}{L^k}\nabla\varphi^k\left(\mathbf{y}^{k,j^k-1}\right)\right\|^2 \\ &\leq g\left(\bar{\mathbf{x}}\right) + \frac{L^k}{2}\left\|\bar{\mathbf{x}} - \mathbf{y}^{k,j^k-1} + \frac{1}{L^k}\nabla\varphi^k\left(\mathbf{y}^{k,j^k-1}\right)\right\|^2, \end{aligned}$$

and after simple algebraic manipulations, we get

$$\begin{aligned} g\left(\bar{\mathbf{x}}\right) &\geq g\left(\mathbf{x}^{k+1}\right) + \frac{L^k}{2}\left\|\mathbf{x}^{k+1}\right\|^2 - \frac{L^k}{2}\|\bar{\mathbf{x}}\|^2 + \left(\mathbf{x}^{k+1} - \bar{\mathbf{x}}\right)^T\left(\nabla\varphi^k\left(\mathbf{y}^{k,j^k-1}\right) - L^k\mathbf{y}^{k,j^k-1}\right) \\ &\geq g\left(\mathbf{x}^{k+1}\right) + \frac{L^k}{2}\left(\left\|\mathbf{x}^{k+1}\right\|^2 - \|\bar{\mathbf{x}}\|^2\right) - \left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}\right\| \cdot \left\|\nabla\varphi^k\left(\mathbf{y}^{k,j^k-1}\right) - L^k\mathbf{y}^{k,j^k-1}\right\| \\ &\geq g\left(\mathbf{x}^{k+1}\right) + \frac{L^k}{2}\left(\left\|\mathbf{x}^{k+1}\right\|^2 - \|\bar{\mathbf{x}}\|^2\right) - M\left\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}\right\|, \end{aligned} \tag{14}$$

where the second inequality follows from the Cauchy–Schwartz inequality, and the last inequality follows from Proposition 3.1(iii). Now, from Assumption 1 it holds that $0 \leq L^k \leq \bar{L}$ for any $k \geq 0$; thus, by replacing k with $k_l - 1$ in (14) and taking the limit $l \to \infty$ we get

$$\limsup_{l\to\infty} g\left(\mathbf{x}^{k_l}\right) \leq g\left(\bar{\mathbf{x}}\right),$$

and condition (B3) is established.

Finally, we prove condition (B4). To this end, notice that

$$\sum_{k=1}^{\infty}\sqrt{e_1^k}=\sum_{k=1}^{\infty}\sqrt{\frac{4M^2\bar{L}}{j^k+1}}=\sum_{k=1}^{\infty}\sqrt{\frac{4M^2\bar{L}}{s+2^{\lfloor k/r\rfloor}}}<\infty,$$

where we used the rule (4). Also, we have

$$\sum_{k=1}^{\infty}e_2^k=\sum_{k=1}^{\infty}\frac{16M\bar{L}\sqrt{\kappa}}{j^k-1}=\sum_{k=1}^{\infty}\frac{16M\bar{L}\sqrt{\kappa}}{s+2^{\lfloor k/r\rfloor}-2}<\infty,$$

and the proof is completed. □

Remark 3.1 (i) Upon closer examination of the proof for condition (B1), it becomes evident that it solely relies on the inequality (6), which characterizes the convergence rate of the nested algorithm in terms of function values. Consequently, any algorithm, not exclusively FISTA, that satisfies an inequality of this form (as described in Assumption 1 in Appendix A) will also satisfy condition (B1), provided that the number of inner iterations j^k satisfies (5) starting from some $K \geq 0$.

(ii) To establish the validity of conditions (B2) and (B3), we relied on the specific structure of the nested algorithm, namely FISTA. Therefore, when considering other algorithms apart from FISTA, these conditions must be independently verified for each individual nested algorithm. In the proof of Theorem 3.1, we made use of the fact that the iteration $\mathbf{x}^{k+1}$, for $k \geq 0$, is defined as a proximal gradient step (refer to step 7 in Algorithm 1). Hence, the results presented in Theorem 3.1 can be extended to other accelerated or, in general, non-descent variants of the proximal gradient method, such as the inertial proximal gradient method introduced in [20].

(iii) Since the function Ψ^k is assumed to be σ^k-strongly convex, we can leverage a variant of FISTA known as restarted FISTA, which exhibits a linear rate of convergence in terms of function values (refer to [2, Section 10.7.6]). In this scenario, we can substitute the rate factor in (6) with

$$\beta_{\mathrm{F}}^{k,j}=\frac{L^k}{2}\left(\frac{1}{2}\right)^j.$$

It is worth noting that all the arguments previously discussed concerning FISTA can be adapted to this variant as well, thereby ensuring the satisfaction of the block-wise conditions. In this case, the sequence $\left\{j^k\right\}_{k\geq 0}$, representing the total number of inner iterations, can be chosen to be any increasing sequence of integers which provide a decrease in function values.

3.2 Hybrid NAM Scheme

In the previous subsection, we showed that by nesting FISTA within NAM, we achieved a globally convergent algorithm to critical points of the function F in Problem (P) under Assumption 2. However, in many practical applications, the subproblems may not always exhibit strong convexity, or it may be challenging to ascertain whether the strong convexity parameters are bounded from below.

To address this issue, we propose the hybrid NAM algorithmic framework in this subsection, which is applicable in scenarios where strong convexity throughout the algorithm cannot be guaranteed or verified easily. By incorporating this hybrid approach, we can still ensure global convergence to critical points while relaxing the strong convexity requirement.

During each (outer) iteration $k \geq 0$ of the H-NAM scheme, we check whether the function Ψ_i^k from Problem $\left(\mathrm{P}_i^k\right)$ exhibits strong convexity. If it is confirmed to be strongly convex, then the FISTA method is employed for the minimization process. However, if the function is not strongly convex, one may utilize any sufficient decrease method instead. This combined approach is outlined in Algorithm 2.

Algorithm 2 Hybrid NAM Scheme

1: **Input:** Nested sufficient decrease algorithms $\mathcal{B}_i$ and $\underline{\sigma}_i > 0$, $i = 1, 2, \ldots, p$.
2: **Initialization:** $\left(\mathbf{z}^0, \mathbf{u}^0\right) \in \mathbb{R}^d \times \mathbb{R}^{d_0}$.
3: **for** $k \geq 0$ **do**
4: Update $\mathbf{u}^{k+1} \in \operatorname{argmin}\left\{ F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \ldots, \mathbf{z}_p^k, \mathbf{u}\right) : \mathbf{u} \in \mathbb{R}^{d_0} \right\}$.
5: **for** $i = 1, 2, \ldots, p$ **do**
6: **if** Ψ_i^k is σ_i^k-strongly convex with $\sigma_i^k \geq \underline{\sigma}_i$, **then** update $\mathbf{z}_i^{k+1}$, starting from $\mathbf{z}_i^k$, by performing $j_i^k \in \mathbb{N}$ iterations of FISTA (see steps 7–9 in Algorithm 1).
7: **else** update $\mathbf{z}_i^{k+1}$ by performing $j_i^k \in \mathbb{N}$ iterations of $\mathcal{B}_i$, starting from $\mathbf{z}_i^k$.
8: **end for**
9: Define $\mathbf{z}^{k+1} = \left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^{k+1}, \ldots, \mathbf{z}_p^{k+1}\right)$.
10: **end for**

As a result of employing the hybrid scheme, we are able to relax not only the requirement that the functions Ψ_i^k must be strongly convex for all $k \geq 0$, but also the boundedness assumption of the strong convexity parameters σ_i^k for $i = 1, 2, \ldots, p$ and $k \geq 0$ (as stated in Assumption 2(c)). Instead, we adopt the following approach. For each $i = 1, 2, \ldots, p$, we define a predetermined scalar $\underline{\sigma}_i > 0$. Then, during each outer iteration $k \geq 0$, we check whether Ψ_i^k is σ_i^k-strongly convex, where σ_i^k satisfies $\sigma_i^k \geq \underline{\sigma}_i$.

This algorithmic framework is particularly advantageous because it allows for the use of the FISTA method in certain outer iterations, even if the user is unable to verify in advance that all functions Ψ_i^k, for $k \geq 0$, are strongly convex, or to verify that the strong convexity parameters are bounded from below by $\underline{\sigma}_i > 0$. In these cases, H-NAM suggests applying inner iterations of a sufficient decrease algorithm $\mathcal{B}_i$ (as described in step 7 of Algorithm 2). For instance, when dealing with Problem (P) and its corresponding subproblems $\left(\mathrm{P}_i^k\right)$ for $k \geq 0$, one can choose to utilize inner

iterations of the proximal gradient method while maintaining the global convergence result of H-NAM. Additionally, if the selected algorithms $\mathcal{B}_i$ satisfy the block-wise conditions individually, their combination in steps 6 and 7 of H-NAM also satisfies the block-wise conditions. Therefore, the global convergence result remains valid. While we omit the proof here, it follows a similar line of reasoning to the arguments presented in the proof of Theorem 3.1.

As an illustrative example of the applicability of the H-NAM framework, we consider the nonnegative matrix factorization (NMF) problem [7, 28]. Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, the NMF problem aims to find two matrices $\mathbf{U} \in \mathbb{R}^{n\times q}$ and $\mathbf{V} \in \mathbb{R}^{q\times m}$ with nonnegative entries, where $q < \min\{n, m\}$, such that $\mathbf{A} \approx \mathbf{UV}$. The NMF problem can be formulated as the following non-convex and non-differentiable optimization problem [21]:

$$\min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times q}\\ \mathbf{V}\in\mathbb{R}^{q\times m}}} \quad H(\mathbf{U}, \mathbf{V}) \equiv \frac{1}{2}\|\mathbf{UV} - \mathbf{A}\|_F^2 + \delta_+(\mathbf{U}) + \delta_+(\mathbf{V}), \tag{NMF}$$

where the indicator function δ_+ is defined such that $\delta_+(\mathbf{U}) = 0$ if all elements of the matrix $\mathbf{U}$ are nonnegative, and ∞ otherwise. Notice that Problem (NMF) has a block structure with respect to the variables $\mathbf{U}$ and $\mathbf{V}$, making it suitable for the NAM framework.

Consider a sequence $\left\{\left(\mathbf{U}^k, \mathbf{V}^k\right)\right\}_{k\geq 0}$ generated by NAM to minimize Problem (NMF). When focusing on the minimization of the partial function $\mathbf{V} \mapsto H\left(\mathbf{U}^k, \mathbf{V}\right)$ for some $k \geq 0$, it can be verified that if the matrix $\mathbf{U}^k$ has full column rank, then this function is $\sigma_{\mathbf{U}}^k$-strongly convex, where

$$\sigma_{\mathbf{U}}^k = \lambda_{\min}\left(\left(\mathbf{U}^k\right)^T \mathbf{U}^k \otimes \mathbf{I}_n\right) = \lambda_{\min}\left(\left(\mathbf{U}^k\right)^T \mathbf{U}^k\right),$$

for $\otimes$ the Kronecker matrix product and $\mathbf{I}_n$ the $n \times n$ identity matrix. Similarly, it can be verified that if the matrix $\mathbf{V}^k$ is of full row rank, then the partial function $\mathbf{U} \mapsto H\left(\mathbf{U}, \mathbf{V}^k\right)$, $k \geq 0$, is $\sigma_{\mathbf{V}}^k$-strongly convex where

$$\sigma_{\mathbf{V}}^k = \lambda_{\min}\left(\mathbf{I}_m \otimes \mathbf{V}^k\left(\mathbf{V}^k\right)^T\right) = \lambda_{\min}\left(\mathbf{V}^k\left(\mathbf{V}^k\right)^T\right).$$

Calculating $\sigma_{\mathbf{U}}^k$ and $\sigma_{\mathbf{V}}^k$ is a relatively simple computational task when the integer q is small, which is often desirable in the context of Problem (NMF). However, there is no guarantee that the strong convexity parameters $\sigma_{\mathbf{U}}^k$ and $\sigma_{\mathbf{V}}^k$ are bounded from below by a positive value independent of $k \geq 0$.

To address this, H-NAM can be applied by introducing threshold parameters $\varrho_{\mathbf{U}} > 0$ and $\varrho_{\mathbf{V}} > 0$. The following procedure is then followed for any $k \geq 0$: If $\sigma_{\mathbf{U}}^k \geq \varrho_{\mathbf{U}}$ or $\sigma_{\mathbf{V}}^k \geq \varrho_{\mathbf{V}}$, then FISTA is applied to minimize the partial function $\mathbf{V} \mapsto H\left(\mathbf{U}^k, \mathbf{V}\right)$ or $\mathbf{U} \mapsto H\left(\mathbf{U}, \mathbf{V}^k\right)$, respectively. Otherwise, a sufficient decrease method is employed.

4 Numerical Experiments

In this section, our objective is to demonstrate the advantages of utilizing NAM with the accelerated non-descent method FISTA for addressing non-convex and non-smooth optimization problems. To illustrate this, we focus on the application of regularized structured total least squares (RSTLS) in the context of image deblurring. RSTLS is formulated as a non-convex and non-smooth minimization problem, making it an ideal candidate for our analysis.

4.1 RSTLS for Image Deblurring

The objective of image deblurring is to recover a given vectorized image $\mathbf{b} \in \mathbb{R}^d$ that has been blurred and corrupted by noise. We consider the scenario where the blurred image is generated by convolving the true image, denoted as $\mathbf{z}^{\text{real}} \in \mathbb{R}^d$, with a Gaussian blur point spread function (PSF) of size $q \times q$ and standard deviation $\gamma > 0$. The PSF is assumed to have periodic boundary conditions, which is a common assumption in image processing (for more details, refer to [13, Chapter 3]). Additionally, the blurred image is further corrupted by additive noise. Hence, we assume that the PSF has the form

$$\text{PSF}(\mathbf{u}) = \sum_{i=1}^{d_0} \mathbf{u}_i \mathbf{A}_i,$$

for some unknown weight vector $\mathbf{u} \in \mathbb{R}^{d_0}$ and given matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{d_0} \in \mathbb{R}^{d \times d}$ called structure matrices. In the experiments below, the blurred and noisy image was constructed by setting

$$\mathbf{b} = \text{PSF}^{\text{real}} \mathbf{z}^{\text{real}} + \boldsymbol{\mu}, \quad \boldsymbol{\mu} \sim \text{Gauss}(\mathbf{0}_d, \delta_w), \tag{15}$$

where PSF^{real} denotes the real PSF used for deblurring and $\delta_w > 0$ is some noise factor. In practice, the blurring operator PSF^{real} is unknown, and we assume that we only have the observed PSF denoted by PSF^{obs}, which was constructed by

$$\text{PSF}^{\text{obs}} = \text{PSF}^{\text{real}} + \sum_{i=1}^{d_0} \boldsymbol{\eta}_i \mathbf{u}_i^{\text{real}} \mathbf{A}_i, \quad \boldsymbol{\eta} \sim \text{Uniform}\left[\mathbf{0}_{d_0}, \delta_e\right], \tag{16}$$

where $\text{PSF}\left(\mathbf{u}^{\text{real}}\right) = \text{PSF}^{\text{real}}$ and for some noise factor $\delta_e > 0$.

Thus, the problem of reconstructing the blurred image $\mathbf{b} \in \mathbb{R}^d$ can be formulated as the following RSTLS problem (see, for example, [24] and [3])

$$\min_{\mathbf{z} \in \mathbb{R}^d, \mathbf{u} \in \mathbb{R}^{d_0}} F(\mathbf{z}, \mathbf{u}) \equiv \delta_w^2 \lambda f(\mathbf{z}) + \left\| \sum_{i=1}^{d_0} \mathbf{u}_i \mathbf{A}_i \mathbf{z} - \mathbf{b} \right\|^2 + \frac{\delta_w^2}{\delta_e^2} \|\mathbf{u}\|^2, \tag{RSTLS}$$

where $f\colon \mathbb{R}^d \to (-\infty, \infty]$ is a regularizing function and $\lambda > 0$ is a regularizing parameter. In the experiments below, we use the non-smooth elastic net regularization, which is considered as a successful regularizer in matrix recovery (see, for example, [27]). Mathematically, for some $\alpha \in (0, 1)$ we take

$$f(\mathbf{z}) = \alpha \|\mathbf{z}\|_1 + (1-\alpha)\|\mathbf{z}\|_2^2 .$$

4.2 Methods for the RSTLS Problem

We consider two blocks of variables, which are $\mathbf{z} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathbb{R}^{d_0}$. All methods below approximate the solution of the subproblems with respect to the $\mathbf{z}$ block using some nested algorithm, whereas the subproblems with respect to the $\mathbf{u}$ block are exactly solved. The reason is that the dimension of the $\mathbf{z}$ block can be large, while the dimension of the $\mathbf{u}$ block can be relatively small. (For example, in the experiments below we consider a $256 \times 256 \times 3$ test image, and therefore, $d = 3 \cdot 256^2$, while $d_0 = 6$.) We consider the following three methods:

1. NAM-F: our NAM scheme with FISTA as a nested algorithm.
2. ECR-PG: proximal gradient as a nested algorithm (see, for example, [25] and [6]).
3. SPA: the semi-proximal alternating (SPA) algorithm of [3], which performs one iteration approximation of the proximal gradient method. (We note that if the number of inner iterations of NAM-F and ECR-PG is fixed and set to 1 in any outer iteration, then the two coincide with SPA.)

It is easy to verify using the first-order optimality condition that the minimizer of the strongly convex partial function $\mathbf{u} \mapsto F\left(\mathbf{z}^{k+1}, \mathbf{u}\right)$ is given by

$$\mathbf{u} = \left(\mathbf{B}\left(\mathbf{z}^{k+1}\right)^T \mathbf{B}\left(\mathbf{z}^{k+1}\right) + \frac{\delta_w^2}{\delta_e^2}\mathbf{I}_{d_0 \times d_0}\right)^{-1} \mathbf{B}\left(\mathbf{z}^{k+1}\right)^T \mathbf{b}, \tag{17}$$

where $\mathbf{B}(\mathbf{z}) \equiv \left[\mathbf{A}_1\mathbf{z}\ \mathbf{A}_2\mathbf{z}\ \cdots\ \mathbf{A}_{d_0}\mathbf{z}\right]$.

Notice that with respect to the block $\mathbf{z}$, the partial functions $\mathbf{z} \mapsto F\left(\mathbf{u}^k, \mathbf{z}\right)$, $k \geq 0$, are σ^k-strongly convex, and can be written as a sum of a smooth part with L^k-Lipschitz continuous gradient and a non-smooth part. It is easy to verify that

$$\sigma^k = 2\lambda_{\min}\left(\mathrm{PSF}\left(\mathbf{u}^k\right)^T \mathrm{PSF}\left(\mathbf{u}^k\right)\right) + 2\delta_w^2\lambda(1-\alpha) \geq 2\delta_w^2\lambda(1-\alpha) > 0,$$

and that

$$L^k = L\left(\mathbf{u}^k\right) \equiv 2\lambda_{\max}\left(\mathrm{PSF}\left(\mathbf{u}^k\right)^T \mathrm{PSF}\left(\mathbf{u}^k\right)\right) + 2\delta_w^2\lambda(1-\alpha). \tag{18}$$

We should mention that since the PSF used in this paper has periodic boundary conditions, then the matrix $\mathrm{PSF}\left(\mathbf{u}^k\right)^T \mathrm{PSF}\left(\mathbf{u}^k\right)$ is a block circulant with circulant blocks (BCCB) matrix (see [13, Chapter 4]), and therefore, its eigenvalues can be computed

efficiently (see [13, section 4.2.1]). Therefore, in the implementation of the algorithms below we use the step-size $1/L^k$, where L^k is the tight Lipschitz constant given in (18).

From the continuity of $L\left(\mathbf{u}^k\right)$, it follows that the sequence $\left\{L^k\right\}_{k\geq 0}$ is bounded from above over compact subsets of $\mathbb{R}^{d_0}$. Moreover, the function $F(\mathbf{z}, \mathbf{u})$ is coercive and as such has bounded level sets. Therefore, since the sequence $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ generated by the algorithms is non-increasing, it follows that the sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is contained in the level set of F at level $F\left(\mathbf{z}^0, \mathbf{u}^0\right)$ and hence bounded.

Now, since the function F is a sum of norms, it surely satisfies the KL property. Hence, it follows that NAM-F and ECR-PG globally converge to a critical point of the function F. In addition, the SPA algorithm was proved to be a globally convergent algorithm (see [3]).

Notice that in all three methods we calculate a certain prox-grad step with respect to the ℓ_1-norm. One can verify that for any $t > 0$ it holds that (see, for example, [2, Example 6.8])

$$\begin{aligned}\operatorname{prox}_{t\|\cdot\|_1}(\mathbf{x}) &\equiv \operatorname{argmin}\left\{t\|\mathbf{y}\|_1 + \tfrac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2 : \mathbf{y}\in\mathbb{R}^d\right\}\\ &= \max\left\{|\mathbf{x}| - t\mathbf{e}, \mathbf{0}_d\right\}\odot \operatorname{sign}(\mathbf{x}),\end{aligned}$$

where $|\cdot|$ is the element-wise absolute value, $\mathbf{e}\in\mathbb{R}^d$ is the vector of all ones, the maximum and sign functions are taken element-wise, and $\odot$ is element-wise vector multiplication.

The explicit NAM-F, ECR-PG and SPA algorithms for solving the RSTLS problem are explicitly given in Algorithms 3, 4 and 5, respectively. To make the steps in the algorithms easy to read, we use the following notation for any $\mathbf{x}\in\mathbb{R}^d$ and $k\geq 0$

$$\mathcal{T}^k(\mathbf{x}) \equiv \operatorname{prox}_{\frac{\lambda\alpha\delta_w^2}{L^k}\|\cdot\|_1}\left(\mathbf{x} - \frac{2}{L^k}\left(\lambda(1-\alpha)\delta_w^2\mathbf{x} - \operatorname{PSF}\left(\mathbf{u}^k\right)^T\operatorname{PSF}\left(\mathbf{u}^k\right)\mathbf{x}\right)\right).$$

Algorithm 3 NAM-F

1: **Input:** $\left(\mathbf{z}^0, \mathbf{u}^0\right)\in\mathbb{R}^d\times\mathbb{R}^{d_0}$ and $s, r\in\mathbb{N}$.
2: **for** $k\geq 0$ **do**
3: Set $\mathbf{y}^{k,0} = \mathbf{z}^k$, $t_0 = 1$ and update L^k according to (18).
4: **for** $j = 0, 1, \ldots, s + 2^{\lfloor k/r\rfloor} - 2$ **do**
5: Update $\mathbf{z}^{k,j+1} = \mathcal{T}^k\left(\mathbf{y}^{k,j}\right)$.
6: Set $t_{j+1} = \frac{1+\sqrt{1+4t_j^2}}{2}$ and then update $\mathbf{y}^{k,j+1} = \mathbf{z}^{k,j+1} + \frac{t_j-1}{t_{j+1}}\left(\mathbf{z}^{k,j+1} - \mathbf{z}^{k,j}\right)$.
7: **end for**
8: Set $\mathbf{z}^{k+1}$ as the output of step 4 and then update $\mathbf{u}^{k+1}$ according to (17).
9: **end for**

Algorithm 4 ECR-PG

1: **Input:** $\left(\mathbf{z}^0, \mathbf{u}^0\right) \in \mathbb{R}^d \times \mathbb{R}^{d_0}$ and $s \in \mathbb{N}$.
2: **for** $k \geq 0$ **do**
3: Update L^k according to (18).
4: **for** $j = 0, 1, \ldots, s-1$ **do**
5: Update $\mathbf{z}^{k,j+1} = \mathcal{T}^k\left(\mathbf{z}^{k,j}\right)$.
6: **end for**
7: Set $\mathbf{z}^{k+1}$ as the output of step 4 and then update $\mathbf{u}^{k+1}$ according to (17).
8: **end for**

Algorithm 5 SPA

1: **Input:** $\left(\mathbf{z}^0, \mathbf{u}^0\right) \in \mathbb{R}^d \times \mathbb{R}^{d_0}$.
2: **for** $k \geq 0$ **do**
3: Update L^k according to (18) and then update $\mathbf{z}^{k+1} = \mathcal{T}^k\left(\mathbf{z}^k\right)$.
4: Update $\mathbf{u}^{k+1}$ according to (17).
5: **end for**

4.3 Results

We ran the three algorithms on a blurred $256 \times 256 \times 3$ gantry crane test image taken from the MATLAB Image Processing Toolbox. Following [3, 11], in all experiments we set $\lambda = 1$, $\alpha = 0.5$, $\delta_w = \delta_e = 10^{-4}$, $q = 5$ and $\gamma = 2$.

For NAM-F, we considered four cases that differ by the number of inner iterations in every outer iteration. Specifically, we set $s \in \{1, 10, 100, 1000\}$, and in all four cases, we set $r = 10$. For example, in the case where $s = 10$ and $r = 10$, following the formula in (4), the number of inner iterations is $j^0 = j^1 = \ldots = j^9 = 10$, $j^{10} = j^{11} = \ldots = j^{19} = 11$, $j^{20} = j^{21} = \ldots = j^{29} = 13$, etc. For ECR-PG, we considered three cases where $s \in \{10, 100, 1000\}$ (notice that ECR-PG with $s = 1$ coincides with the SPA algorithm, which is not the case for NAM-F since in NAM-F the number of inner iterations increases along the outer iterations).

We performed $R = 50$ Monte Carlo trials, where in each we drew a different blurred and noisy image $\mathbf{b} \in \mathbb{R}^d$ and a different PSF^{obs} according to (15) and (16), respectively. See panel (a) of Fig. 1 for the real image, and panel (b) for one of the blurred and noisy images. In all experiments, we set the starting points $\mathbf{z}^0 = \mathbf{b}$ and $\mathbf{u}^0 = \mathbf{0}_{d_0}$.

We let all methods run for $N = 15,000$ total iterations, where the total iterations counter counts inner iterations. Recall that in the inner iterations we only update the $\mathbf{z} \in \mathbb{R}^d$ block. Therefore, for $N = 15,000$ the SPA algorithm updates each of the blocks $\mathbf{z} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathbb{R}^{d_0}$ exactly 15,000 times. This is not the case for the nested methods NAM-F and ECR-PG, where for the same number of total iterations, the block $\mathbf{u} \in \mathbb{R}^{d_0}$ is updated less times. For example, if $j_0 = 100$ and $j_1 = 200$, then the current total number of iterations is 300, which means that the $\mathbf{z}$ block was updated 300 times while the $\mathbf{u}$ block was updated only twice.

We used the following two measures in order to compare the three methods:

(a) Original image (b) One of the 50 blurred and noisy images

Fig. 1 The gantry crane image used in the experiments

1. Function value of Problem (RSTLS) averaged over all Monte Carlo trials at each iteration $1 \leq n \leq N$

$$F_n = \frac{1}{R} \sum_{t=1}^{R} F\left(\mathbf{z}_t^n, \mathbf{u}_t^n\right),$$

where $(\mathbf{z}^n, \mathbf{u}^n)$ is the nth iterate (out of N) in the tth trial (out of R). Since the $\mathbf{u}$ block can be updated less times then the $\mathbf{z}$ block, then by $\mathbf{u}_t^n$ we mean the last obtained iterate of $\mathbf{u}$. Notice that since $F \geq 0$, then a closer value of F_n to 0 is an indication of the performance of the method in minimizing Problem (RSTLS).

2. The deviation of the function value at the current iteration, $1 \leq n \leq N$, from the function value at the point of convergence, averaged over all Monte Carlo trails

$$\Delta F_n = \frac{1}{R} \sum_{t=1}^{R} \left(F\left(\mathbf{z}_t^n, \mathbf{u}_t^n\right) - F\left(\mathbf{z}_t^{\text{con}}, \mathbf{u}_t^{\text{con}}\right)\right),$$

where $\left(\mathbf{z}_t^{\text{con}}, \mathbf{u}_t^{\text{con}}\right)$ denotes the point of convergence of the corresponding method in some trial t. For each method and for each trial, the point of convergence was obtained by running the corresponding method for 30, 000 total iterations and then setting the output as $\mathbf{z}^{\text{con}}$. Since all involved methods are globally convergent, the convergence point is uniquely defined, and therefore $\mathbf{z}^{\text{con}}$ is an estimator of this unique point. Notice that since the sequence of function values is non-increasing over the outer iterations (for each of the methods), then a value of ΔF_n closer to 0 indicates how close the method is to the value at its point of convergence.

All experiments were ran on an Intel Core i7-8565 CPU at 1.80 GHz and 1.99 GHz with 40.0 GB RAM, using MATLAB 2021a on the Windows 10 Pro 64-bit operating

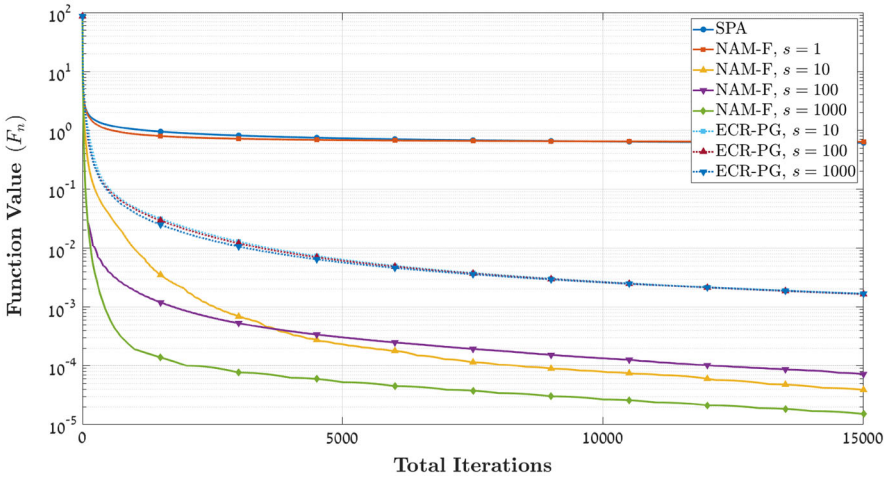

Fig. 2 Average function value at each iteration, for $N = 15,000$ over $R = 50$ Monte Carlo trials, in a logarithmic scale. For all NAM-F variants, we set $r = 10$

system. The datasets generated during this study are available from the corresponding author on a reasonable request.

In Fig. 2, we compare the average function values obtained by the methods (F_n). We see that the three nested methods NAM-F with $s \in \{10, 100, 1000\}$ performed better than other methods. Hence, there is a benefit in using nested methods for solving Problem (RSTLS) relatively to non-nested methods, which are SPA and NAM-F with $s = 1$. (We point out that NAM-F with $s = 1$ is not a "truly" nested method, since it performs only one inner iteration at the first $r = 10$ outer iterations.) In addition, we see that there is a benefit in using accelerated methods compared to the non-accelerated methods, which are ECR-PG with $s \in \{10, 100, 1000\}$, as nested algorithms.

In Fig. 3, we compare the gap in function values obtained by the methods between the current iteration and the convergence point (ΔF_n). We see that the three nested methods with accelerated inner algorithms, which are NAM-F with $s \in \{10, 100, 1000\}$, get closer to the value of their output in much fewer iterations. We conclude that these nested methods require fewer iterations to reach a better solution, relatively to the other methods.

Last, in Fig. 4 we provide the images obtained by SPA, NAM-F with $s = \{1, 1000\}$ and ECR-PG with $s = 1000$ after 1, 500, and 1500 total iterations for one of the Monte Carlo trials. We see that the nested methods all yield better images compared to SPA and NAM-F with $s = 1$. In addition, we see that the accelerated method NAM-F with $s = 1000$ yields a better image in comparison to ECR-PG with $s = 1000$.

5 Conclusion

In this paper, we showed that one can utilize the NAM algorithmic framework in order to design a nested, non-descent and accelerated algorithm using FISTA to solve

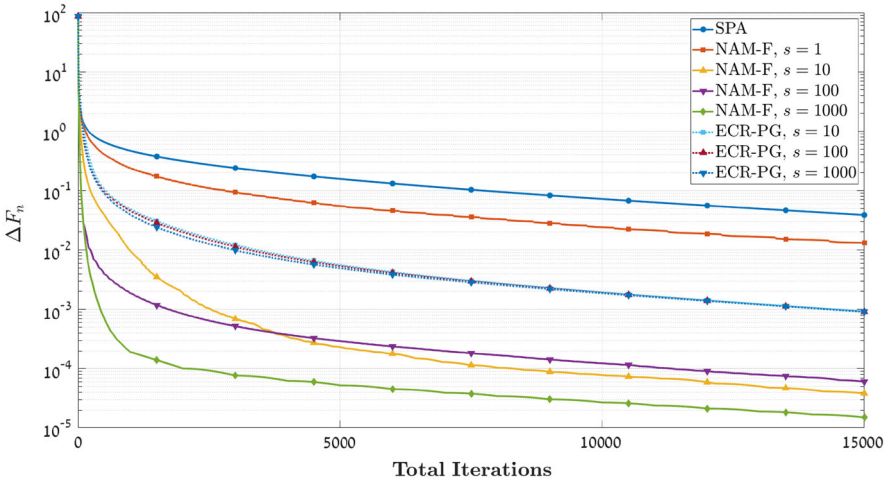

Fig. 3 Average gap of the function value from the value at the convergence point, for $N = 15,000$ over $R = 50$ Monte Carlo trials, in a logarithmic scale. For all NAM-F variants, we set $r = 10$

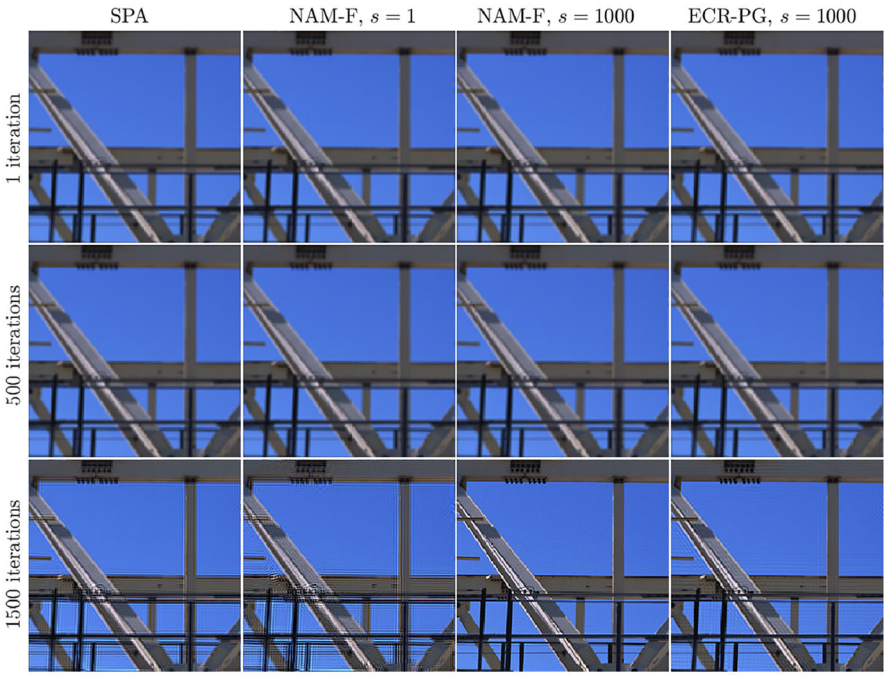

Fig. 4 Images obtained by the methods at selected iterations

non-convex and non-smooth optimization problems, while guaranteeing global convergence to a unique critical point of the objective function. In addition, we provided the H-NAM algorithmic framework, which allows to use FISTA whenever possible, while still keeping the global convergence result.

Acknowledgements We express our gratitude to the anonymous reviewers whose valuable feedback has greatly contributed to enhancing the paper and making it more concise.

Funding The work of Shoham Sabach and Eyal Gur was supported by the Israel Science Foundation, grant no. ISF 2480/21.

Data Availability Data will be made available on request.

Appendix

A FISTA Convergence Results

Here we prove several results about the FISTA method, which are used in the proof of Theorem 3.1 (see Sect. 3.1).

A.0 Let $\mathcal{A}$ be some algorithm and let $\left\{\mathbf{v}^j\right\}_{j\geq 0}$ be a sequence generated by $\mathcal{A}$ for minimizing a σ-strongly convex function $f:\mathbb{R}^n \to (-\infty,\infty]$. Assume that there exists a sequence of scalars $\left\{\beta^j\right\}_{j\geq 0}$ such that

(a) $\beta^j \to 0$ as $j \to \infty$.
(b) $f\left(\mathbf{v}^j\right) - f\left(\mathbf{v}^*\right) \leq \beta^j \left\|\mathbf{v}^0 - \mathbf{v}^*\right\|^2$, where $\mathbf{v}^* \in \mathbb{R}^n$ is the unique minimizer of f.

Notice that any convergent algorithm with a known convergence rate in terms of function values satisfies Assumption 1. In particular, following inequality (6), we see that FISTA satisfies this assumption.

Lemma A.1 *Let $f:\mathbb{R}^n \to (-\infty,\infty]$ be a σ-strongly convex function, and let $\mathbf{v}^* \in \mathbb{R}^n$ be its minimizer. Then,*

(i) $f(\mathbf{v}) \geq f(\mathbf{v}^*) + (\sigma/2)\cdot\|\mathbf{v}-\mathbf{v}^*\|^2$ *for any* $\mathbf{v}\in\mathbb{R}^n$.

Let $\left\{\mathbf{v}^j\right\}_{j\geq 0}$ be a sequence generated by algorithm $\mathcal{A}$ that satisfies Assumption 1. Then,

(ii) $\left\|\mathbf{v}^j - \mathbf{v}^*\right\| \leq \sqrt{2\beta^j/\sigma}\left\|\mathbf{v}^0-\mathbf{v}^*\right\|$ *for any* $j \geq 0$. *In particular,* $\mathbf{v}^j \to \mathbf{v}^*$ *as* $j\to\infty$.
(iii) *If, in addition,* $\beta^j \geq \beta^{j+1}$ *for any* $j\geq 0$, *then* $\left\|\mathbf{v}^{j+1}-\mathbf{v}^j\right\| \leq 2\sqrt{2\beta^j/\sigma}\left\|\mathbf{v}^0-\mathbf{v}^*\right\|$.

Proof Since f is σ-strongly convex, for any $\mathbf{v}\in\mathbb{R}^n$ and for any $\boldsymbol{\xi}\in\partial f(\mathbf{v}^*)$, we have

$$f(\mathbf{v}) \geq f\left(\mathbf{v}^*\right) + \boldsymbol{\xi}^T\left(\mathbf{v}-\mathbf{v}^*\right) + \frac{\sigma}{2}\left\|\mathbf{v}-\mathbf{v}^*\right\|^2.$$

Since $\mathbf{v}^*$ is a minimizer of the function f, then from the first-order optimality condition we have $\mathbf{0}_n \in \partial f(\mathbf{v}^*)$, and item (i) follows. Now, item (ii) immediately follows from item (i) by plugging $\mathbf{v}=\mathbf{v}^j$ and using Assumption 1(b). Moreover, from Assumption 1(a) we have that $\mathbf{v}^j \to \mathbf{v}^*$ as $j\to\infty$, as required.

To prove item (iii), notice that if $\beta^j \geq \beta^{j+1}$ for any $j \geq 0$, then from the triangle inequality and item (ii) we get

$$\left\|\mathbf{v}^{j+1} - \mathbf{v}^j\right\| \leq \sqrt{\frac{2\beta^{j+1}}{\sigma}}\left\|\mathbf{v}^0 - \mathbf{v}^*\right\| + \sqrt{\frac{2\beta^j}{\sigma}}\left\|\mathbf{v}^0 - \mathbf{v}^*\right\| \leq 2\sqrt{\frac{2\beta^j}{\sigma}}\left\|\mathbf{v}^0 - \mathbf{v}^*\right\|,$$

and the proof is completed. □

Using the inequalities obtained in Lemma A.1, in the following lemma we prove convergence results for the FISTA method in the strongly convex setting.

Lemma A.2 *For $k \geq 0$, let $\left\{\mathbf{x}^{k,j}\right\}_{j\geq 0}$ and $\left\{\mathbf{y}^{k,j}\right\}_{j\geq 0}$ be sequences generated by FISTA for Problem $\left(\mathrm{P}^k\right)$ (steps 7–9 in Algorithm 1). Then,*

(i) $\left\|\mathbf{x}^{k,j+1} - \mathbf{y}^{k,j}\right\| \leq 4\left\|\mathbf{x}^{k,j} - \mathbf{x}^k_*\right\|\sqrt{2\beta_{\mathrm{F}}^{k,j-1}/\sigma^k}$ *for all* $j \geq 1$.
(ii) *For all $j \geq 0$ we have*

$$\Psi^k\left(\mathbf{x}^k\right) - \Psi^k\left(\mathbf{x}^{k,j}\right) \geq \frac{\sigma^k}{2}\left\|\mathbf{x}^k - \mathbf{x}^{k,j}\right\|^2 - \sqrt{2\sigma^k\beta_{\mathrm{F}}^{k,j}}\left\|\mathbf{x}^k - \mathbf{x}^k_*\right\|\left\|\mathbf{x}^k - \mathbf{x}^{k,j}\right\| - \beta_{\mathrm{F}}^{k,j}\left\|\mathbf{x}^k - \mathbf{x}^{k,j}\right\|^2.$$

(iii) $\left\|\mathbf{w}^{k,j}\right\| \leq 8L^k\left\|\mathbf{x}^k - \mathbf{x}^k_*\right\|\sqrt{2\beta_{\mathrm{F}}^{k,j-2}/\sigma^k}$ *for all* $j \geq 2$ *and some* $\mathbf{w}^{k,j} \in \partial\Psi^k\left(\mathbf{x}^{k,j}\right)$.

Proof First, recall that $\mathbf{x}^k = \mathbf{x}^{k,0}$ and that $\mathbf{x}^{k,j^k} = \mathbf{x}^{k+1}$. In addition, recall that $\mathbf{x}^k_*$ is the minimizer of the strongly convex function Ψ^k of Problem $\left(\mathrm{P}^k\right)$.

Since FISTA satisfies Assumption 1 with $\beta_{\mathrm{F}}^{k,j}$ (see (6)), we can use Lemma A.1. Now we prove item (i). From Lemma A.1(iii), we have for all $j \geq 0$ that

$$\left\|\mathbf{x}^{k,j+1} - \mathbf{x}^{k,j}\right\| \leq 2\left\|\mathbf{x}^k - \mathbf{x}^k_*\right\|\sqrt{\frac{2\beta_{\mathrm{F}}^{k,j}}{\sigma^k}}. \tag{19}$$

Hence, for any $j \geq 1$ it follows that

$$\begin{aligned}\left\|\mathbf{x}^{k,j+1}-\mathbf{y}^{k,j}\right\| &= \left\|\mathbf{x}^{k,j+1}-\mathbf{x}^{k,j}-\frac{t_{j-1}-1}{t_j}\left(\mathbf{x}^{k,j}-\mathbf{x}^{k,j-1}\right)\right\| \\ &\leq \left\|\mathbf{x}^{k,j+1}-\mathbf{x}^{k,j}\right\|+\frac{t_{j-1}-1}{t_j}\left\|\mathbf{x}^{k,j}-\mathbf{x}^{k,j-1}\right\| \\ &\leq \left\|\mathbf{x}^{k,j+1}-\mathbf{x}^{k,j}\right\|+\left\|\mathbf{x}^{k,j}-\mathbf{x}^{k,j-1}\right\| \\ &\leq 2\left\|\mathbf{x}^{k}-\mathbf{x}_*^{k}\right\|\sqrt{\frac{2\beta_{\mathrm{F}}^{k,j}}{\sigma^k}}+2\left\|\mathbf{x}^{k}-\mathbf{x}_*^{k}\right\|\sqrt{\frac{2\beta_{\mathrm{F}}^{k,j-1}}{\sigma^k}} \\ &\leq 4\left\|\mathbf{x}^{k}-\mathbf{x}_*^{k}\right\|\sqrt{\frac{2\beta_{\mathrm{F}}^{k,j-1}}{\sigma^k}},\end{aligned}$$

where first equality follows from step 9 in Algorithm 1, the second inequality follows from the fact that $t_0 = 1$ and $t_{j-1} \leq t_j$ for any $j \geq 1$ (see step 8 in Algorithm 1), the third inequality follows from (19), and the last inequality follows from the fact that $\beta_{\mathrm{F}}^{k,j} \leq \beta_{\mathrm{F}}^{k,j-1}$ for any $j \geq 1$.

Now we prove item (ii). From Lemma A.1(i), we have, for any $j \geq 0$, that

$$\Psi^k\left(\mathbf{x}^k\right)-\Psi^k\left(\mathbf{x}_*^k\right)+\Psi^k\left(\mathbf{x}^{k,j}\right)-\Psi^k\left(\mathbf{x}^{k,j}\right) \geq \frac{\sigma^k}{2}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\|^2.$$

Rearranging of the terms yields

$$\begin{aligned}\Psi^k\left(\mathbf{x}^k\right)-\Psi^k\left(\mathbf{x}^{k,j}\right) &\geq \frac{\sigma^k}{2}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\|^2-\left(\Psi^k\left(\mathbf{x}^{k,j}\right)-\Psi^k\left(\mathbf{x}_*^k\right)\right) \\ &= \frac{\sigma^k}{2}\left\|\mathbf{x}^{k,j}-\mathbf{x}_*^k\right\|^2+\frac{\sigma^k}{2}\left\|\mathbf{x}^k-\mathbf{x}^{k,j}\right\|^2 \\ &\quad+\sigma\left(\mathbf{x}^{k,j}-\mathbf{x}_*^k\right)^T\left(\mathbf{x}^k-\mathbf{x}^{k,j}\right)-\left(\Psi^k\left(\mathbf{x}^{k,j}\right)-\Psi^k\left(\mathbf{x}_*^k\right)\right).\end{aligned} \tag{20}$$

Since $\left(\sigma^k/2\right)\cdot\left\|\mathbf{x}^{k,j}-\mathbf{x}_*^k\right\| \geq 0$, we get from (20) using the Cauchy–Schwartz inequality

$$\begin{aligned}\Psi^k\left(\mathbf{x}^k\right)-\Psi^k\left(\mathbf{x}^{k,j}\right) &\geq \frac{\sigma^k}{2}\left\|\mathbf{x}^k-\mathbf{x}^{k,j}\right\|^2-\sigma^k\left\|\mathbf{x}^{k,j}-\mathbf{x}_*^k\right\|\cdot\left\|\mathbf{x}^k-\mathbf{x}^{k,j}\right\| \\ &\quad-\left(\Psi^k\left(\mathbf{x}^{j}\right)-\Psi^k\left(\mathbf{x}_*^k\right)\right) \\ &\geq \frac{\sigma^k}{2}\left\|\mathbf{x}^k-\mathbf{x}^{k,j}\right\|^2-\sqrt{2\sigma^k\beta_{\mathrm{F}}^j}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\|\cdot\left\|\mathbf{x}^k-\mathbf{x}^{k,j}\right\| \\ &\quad-\beta_{\mathrm{F}}^j\left\|\mathbf{x}^k-\mathbf{x}^{k,j}\right\|^2,\end{aligned}$$

where the second inequality follows from (6) and Assumption 1(b) with $\beta^j = \beta_{\mathrm{F}}^{k,j}$, and Lemma A.1(ii).

Now we prove item (iii). For any $j \geq 1$, denote

$$\mathbf{w}^{k,j} \equiv L^k\left(\mathbf{y}^{k,j-1} - \mathbf{x}^{k,j}\right) + \nabla\varphi^k\left(\mathbf{x}^{k,j}\right) - \nabla\varphi^k\left(\mathbf{y}^{k,j-1}\right) \in \partial\Psi^k\left(\mathbf{x}^{k,j}\right), \quad (21)$$

where the inclusion follows from the first-order optimality condition of step 7 in Algorithm 1. Now, for any $j \geq 2$ we get

$$\begin{aligned}\left\|\mathbf{w}^{k,j}\right\| &\leq L^k\left\|\mathbf{x}^{k,j} - \mathbf{y}^{k,j-1}\right\| + \left\|\nabla\varphi^k\left(\mathbf{x}^{k,j}\right) - \nabla\varphi^k\left(\mathbf{y}^{k,j-1}\right)\right\| \\ &\leq L^k\left\|\mathbf{x}^{k,j} - \mathbf{y}^{k,j-1}\right\| + L^k\left\|\mathbf{x}^{k,j} - \mathbf{x}^{k,j-1}\right\| \leq 8L^k\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\|\sqrt{\frac{2\beta_{\mathrm{F}}^{k,j-2}}{\sigma^k}},\end{aligned}$$

where the last inequality follows from item (i).

B Proof of Proposition 3.1

Proposition B.1 *For all $k \geq 0$, let $\left\{\mathbf{x}^{k,j}\right\}_{j\geq 0}$ and $\left\{\mathbf{y}^{k,j}\right\}_{j\geq 0}$ be sequences generated by FISTA (steps 7–9 in Algorithm 1) for minimizing Problem $\left(\mathrm{P}^k\right)$. Assume that the sequence generated by Algorithm 1 is bounded. Then, there exists $M > 0$ such that for any $k \geq 0$ it holds that*

(i) $\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| \leq M$.
(ii) $\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\| \leq M$.
(iii) $\left\|\nabla\varphi^k\left(\mathbf{y}^{k,j}\right) - L^k\mathbf{y}^{k,j}\right\| \leq M$ *for any* $j \geq 0$.

Proof Since the sequence generated by NAM is bounded, there exists $M_1 > 0$ such that

$$\left\|\mathbf{z}^k\right\| \leq M_1, \quad (22)$$

and that

$$\left\|\mathbf{z}_i^k - \mathbf{z}_i^{k+1}\right\| = \left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| \leq M_1, \quad (23)$$

for all $k \geq 0$ and item (i) is established. To prove item (ii), notice that from item (i) we have for any $k \geq 0$ that

$$\left\|\mathbf{x}^k - \mathbf{x}_*^k\right\| \leq \left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| + \left\|\mathbf{x}^{k+1} - \mathbf{x}_*^k\right\| \leq M_1 + \left\|\mathbf{x}^{k+1} - \mathbf{x}_*^k\right\|. \quad (24)$$

From Lemma A.1(ii) and (7), we get

$$\left\|\mathbf{x}^{k+1}-\mathbf{x}_*^k\right\| \leq \sqrt{\frac{2\beta_{\mathrm{F}}^{k,j^k}}{\sigma^k}}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\| \leq \sqrt{\frac{4\tilde{L}}{\varrho\left(j^k+1\right)^2}}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\| = \frac{2\sqrt{\kappa}}{j^k+1}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\| \leq \frac{2\sqrt{\kappa}}{2\sqrt{\kappa}+1}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\|, \tag{25}$$

where the second inequality follows from Assumptions 1, 2(c) and 1 (c), and the last inequality follows from (5). Combining (24) and (25) we get

$$0 \leq \left(1-\frac{2\sqrt{\kappa}}{2\sqrt{\kappa}+1}\right)\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\| \leq M_1.$$

Therefore, there exists $M_2 > 0$, such that

$$\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\| \leq M_2, \tag{26}$$

for all $k \geq 0$, and item (ii) is established.

Now we prove item (iii). To this end, we first prove that the sequences $\left\{\mathbf{x}^{k,j}\right\}_{j\geq 0}$ and $\left\{\mathbf{y}^{k,j}\right\}_{j\geq 0}$ are bounded. For any $j \geq 0$, we have

$$\left\|\mathbf{x}^{k,j}\right\| \leq \left\|\mathbf{x}^{k,j}-\mathbf{x}_*^k\right\|+\left\|\mathbf{x}_*^k\right\| \leq \frac{2\sqrt{\kappa}}{2\sqrt{\kappa}+1}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\|+\left\|\mathbf{x}_*^k\right\| \leq \frac{2M_2\sqrt{\kappa}}{2\sqrt{\kappa}+1}+\left\|\mathbf{x}_*^k\right\|, \tag{27}$$

where the second inequality follows by similar arguments as in (25), and the last inequality follows from (26). In addition,

$$\left\|\mathbf{x}_*^k\right\| \leq \left\|\mathbf{x}^k-\mathbf{x}_*^k\right\|+\left\|\mathbf{x}^k\right\| \leq M_2+\left\|\mathbf{x}^k\right\|, \tag{28}$$

and since the sequence $\left\{\mathbf{x}^k\right\}_{k\geq 0}$ is assumed to be bounded, it follows from (27) and (28) that there exists $M_3 > 0$ such that $\left\|\mathbf{x}^{k,j}\right\| \leq M_3$ for any $k \geq 0$ and for any $j \geq 0$. In addition,

$$\left\|\mathbf{y}^{k,j}\right\| \leq \left\|\mathbf{x}^{k,j+1}-\mathbf{y}^{k,j}\right\|+\left\|\mathbf{x}^{k,j+1}\right\| \leq 8\sqrt{\kappa}\left\|\mathbf{x}^k-\mathbf{x}_*^k\right\|+M_3, \tag{29}$$

where the second inequality follows from Lemma A.2(i) and the fact that $\left\|\mathbf{x}^{k,j+1}\right\| \leq M_3$. Therefore, it follows from (26) and (29) that there exists $M_4 > 0$ such that $\left\|\mathbf{y}^{k,j}\right\| \leq M_4$ for any $k \geq 0$ and for any $j \geq 0$.

Last, notice that since the function G is continuously differentiable, then over the compact set containing the involved bounded iterates (which is a subset of the domain $\mathbb{R}^d \times \mathbb{R}^{d_0}$), the gradient of the function $G\colon \mathbb{R}^d \times \mathbb{R}^{d_0} \to \mathbb{R}$ in Problem (P)

is $\mathcal{L}$-Lipschitz continuous, for some $\mathcal{L} > 0$. Hence, we have (recall that $\varphi^k(\mathbf{x}) = G\left(\mathbf{z}_1^{k+1}, \ldots, \mathbf{z}_{i-1}^{k+1}, \mathbf{x}, \mathbf{z}_{i+1}^{k}, \ldots, \mathbf{z}_p^{k}, \mathbf{u}^{k+1}\right)$)

$$\begin{aligned}\left\|\nabla\varphi^k\left(\mathbf{y}^{k,j}\right)\right\| - \left\|\nabla_{\mathbf{z}_i} G\left(\mathbf{0}_{d+d_0}\right)\right\| &\leq \left\|\nabla\varphi^k\left(\mathbf{y}^{k,j}\right) - \nabla_{\mathbf{z}_i} G\left(\mathbf{0}_{d+d_0}\right)\right\| \\ &\leq \mathcal{L}\left(\sum_{j=1}^{i-1}\left\|\mathbf{z}_j^{k+1}\right\| + \left\|\mathbf{y}^{k,j}\right\| + \sum_{j=i}^{p}\left\|\mathbf{z}_j^{k}\right\|\right) \\ &\leq \mathcal{L}\left(M_4 + pM_1\right),\end{aligned}$$

where we used (22). Since $\left\|\nabla_{\mathbf{z}_i} G\left(\mathbf{0}_{d+d_0}\right)\right\|$ is a constant independent of $k \geq 0$ and $j \geq 0$, it follows that there exists $M_5 > 0$ such that $\left\|\nabla\varphi^k\left(\mathbf{y}^{k,j}\right)\right\| \leq M_5$. Hence,

$$\left\|\nabla\varphi^k\left(\mathbf{y}^{k,j}\right) - L^k\mathbf{y}^{k,j}\right\| \leq M_5 + \bar{L}M_4, \tag{30}$$

for any $k \geq 0$ and for any $j \geq 0$.

Finally, by setting $M \equiv \max\left\{M_1, M_2, M_5 + \bar{L}M_4\right\}$, the required results follow from (23), (26) and (30).

References

1. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
2. Beck, A.: First-Order Methods in Optimization, vol. 25. SIAM (2017)
3. Beck, A., Sabach, S., Teboulle, M.: An alternating semiproximal method for nonconvex regularized structured total least squares problems. SIAM J. Matrix Anal. Appl. **37**(3), 1129–1150 (2016)
4. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
5. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**(1–2), 459–494 (2014)
6. Bonettini, S., Prato, M., Rebegoldi, S.: A block coordinate variable metric linesearch based proximal gradient method. Comput. Optim. Appl. **71**(1), 5–52 (2018)
7. Gan, J., Liu, T., Li, L., Zhang, J.: Non-negative matrix factorization: a survey. Comput. J. **64**(7), 1080–1092 (2021)
8. Gorissen, B.L., Yanıkoğlu, İ, den Hertog, D.: A practical guide to robust optimization. Omega **53**, 124–137 (2015)
9. Groenen, P.J.F., van de Velden, M.: Multidimensional scaling by majorization: a review. J. Stat. Softw. **73**, 1–26 (2016)
10. Gur, E., Sabach, S., Shtern, S.: Alternating minimization based first-order method for the wireless sensor network localization problem. IEEE Trans. Signal Process. **68**, 6418–6431 (2020)
11. Gur, E., Sabach, S., Shtern, S.: Convergent nested alternating minimization algorithms for nonconvex optimization problems. Math. Oper. Res., (2022)
12. Gutjahr, W.J., Pichler, A.: Stochastic multi-objective optimization: a survey on non-scalarizing methods. Ann. Oper. Res. **236**(2), 475–499 (2016)
13. Hansen, P.C., Nagy, J.G., O'leary, D.P.: Deblurring Images: Matrices, Spectra, and Filtering. SIAM, (2006)
14. Jain, P., Kar, P.: Non-convex optimization for machine learning. Found. Trends® Mach. Learn. **10**(3–4), 142–336 (2017)

15. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. Inst. Fourier (Grenoble) **48**(3), 769–783 (1998)
16. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels. In Les Équations aux Dérivées Partielles (Paris, 1962), pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, (1963)
17. Mohammadi, F.G., Amini, M.H., Arabnia, H.R.: Evolutionary computation, optimization, and learning algorithms for data science. In: Optimization, Learning, and Control for Interdependent Complex Networks, pages 37–65. Springer, (2020)
18. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation I: Basic Theory, volume 330. Springer Science & Business Media, (2006)
19. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk SSSR **269**(3), 543–547 (1983)
20. Ochs, P., Chen, Y., Brox, T., Pock, T.: iPiano: Inertial proximal algorithm for nonconvex optimization. SIAM J. Imag. Sci. **7**(2), 1388–1419 (2014)
21. Paatero, P., Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994)
22. Pock, T., Sabach, S.: Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM J. Imag. Sci. **9**(4), 1756–1787 (2016)
23. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
24. Pruessner, A., O'Leary, D.P.: Blind deconvolution using a regularized structured total least norm algorithm. SIAM J. Matrix Anal. Appl. **24**(4), 1018–1037 (2003)
25. Teboulle, M., Vaisbourd, Y.: Novel proximal gradient methods for nonnegative matrix factorization with sparsity constraints. SIAM J. Imag. Sci. **13**(1), 381–421 (2020)
26. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**(3), 475–494 (2001)
27. Wang, H., Pan, J., Zhixun, S., Liang, S.: Blind image deblurring using elastic-net based rank prior. Comput. Vis. Image Underst. **168**, 157–171 (2018)
28. Wang, Y.-X., Zhang, Y.-J.: Nonnegative matrix factorization: a comprehensive review. IEEE Trans. Knowl. Data Eng. **25**(6), 1336–1353 (2012)
29. Wen, F., Chu, L., Liu, P., Qiu, R.C.: A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. IEEE Access **6**, 69883–69906 (2018)