Convergent Nested Alternating Minimization Algorithms for Non-Convex Optimization Problems

Eyal Gur, Shoham Sabach[*], Shimrit Shtern[†]

July 27, 2021

Abstract

We introduce a new algorithmic framework for solving non-convex optimization problems, that is called Nested Alternating Minimization, which aims at combining the classical Alternating Minimization technique with inner iterations of any optimization method. We provide a global convergence analysis of the new algorithmic framework to critical points of the problem at hand, which to the best of our knowledge, is the first of this kind for nested methods in the non-convex setting. Central to our global convergence analysis is a new extension of classical proof techniques in the non-convex setting that allows for errors in the conditions. The power of our framework is illustrated with some numerical experiments that show the superiority of this algorithmic framework over existing methods.

1 Introduction

In recent years non-convex and non-smooth optimization problems have gained a lot of interest in numerous applied fields such as machine learning [23], signal processing [36, 32],

[*] This research was partially supported by the D.F.G German Research Foundation 800240.

[†] Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology, Technion city, Haifa 3200003, Israel (e-mail: eyal.gur@campus.technion.ac.il, ssabach@technion.ac.il, shimrits@technion.ac.il).

data science [19, 26], operations research [20, 18] and more. While the theoretical results in the convex setting (such as convergence analysis and rates of convergence) are extensive (see, for instance, the recent book [6] and references therein), the tools and results available regarding the non-convex setting are very limited, making this setting much more challenging. An important task in the non-convex problem domain, which has already gained success in several settings, is to establish global convergence of algorithms to critical points (see more detail in Section 2). In this paper, we are motivated by the analysis and application of a large and important class of nested algorithms, which we call Nested Alternating Minimization (NAM) and will be precisely defined in Section 3. Our contribution consists of providing a new general proof procedure that can be applied to obtain global convergence results for this class of algorithms. As far as we know, this is the first theoretical guarantee for this class of algorithms in the non-convex setting.

1.1 Global Convergence

Proving global convergence of an algorithm to a critical point involves showing that for any starting point the entire sequence generated by the algorithm converges to a single critical point of the optimization problem at hand. In recent years, this task received a lot of attention starting with the works [3, 4, 5] and later with the work [11], which formulates a simple proof technique to obtain global convergence results of algorithms that satisfy three conditions in addition to the KL property of the objective function (see [25, 24, 10]). These works pave the way for many modifications and extensions that have been, and are still being developed (e.g., [35, 30] and references therein). The first condition consists of a sufficient decrease requirement of the algorithm with respect to all variables (see, for instance, [11]) or part of them (see, for instance, [35]), where the sufficient decrease can be measured in terms of the original objective function as in [11] or a related Lyapunov function as in [31, 33]. The second condition consists of bounding the iterates gap from below by the function's sub-gradient norm.

In this work, we propose another proof technique to obtain global convergence results in the non-convex setting. This extension allows more flexibility in accommodating the conditions of previous proposed recipes, in the sense that we propose a relaxed variant of the first two conditions by adding some non-negative error terms (our recipe is a further relaxation of the recipe suggested in [30], which only relaxes the second condition).

In Section 2, following the approach introduced in [11] and recently summarized in [12], we develop the extended proof methodology in detail. Even though the techniques used to derive the new methodology are not new and are based on [11, 12], the newly developed technique allows us to easily analyze non necessarily sufficient decrease methods (due to the relaxation of the first condition), which in turn opens the gate for obtaining global convergence results of the very important and challenging class of Nested Alternating Minimization algorithms.

1.2 Nested Alternating Minimization Algorithms

The celebrated Alternating Minimization (AM) methodology is a technique for handling complicated optimization problems that has regained a huge popularity in the last decade in the context of non-convex optimization. The AM methodology splits an optimization problem into smaller sub-problems, which would hopefully be easier to solve in a closed form. Unfortunately, in many cases, especially in the non-convex setting, even the resulting sub-problems remain too complicated or too large to be solved explicitly. This led to a stream of papers in recent years that suggest to approximate the solution of the sub-problems by computing one (or several) iteration of a certain descent optimization algorithm (see, for instance, [22, 33, 12, 35, 13]). This approach of One Iteration Approximation (OIA) led, in many cases, to algorithms that enjoy global convergence guarantees using the techniques mentioned in Section 1.1. However, the idea of OIA is limited to algorithms with a certain sufficient decrease property (see our discussion above). Motivated by considering to approximate the solution of the sub-problems using non sufficient decrease methods (like Accelerated Gradient Descent Method [29] and its variants for strongly convex or non-smooth functions [9, 6]), we study the class of Nested Alternating Minimization (NAM) algorithms, which allow to approximate the solution of the sub-problems using several iterations instead of one, in order to achieve the relaxed sufficient decrease property discussed above. It should be noted that our NAM framework also covers sub-problems which are solved using OIA, and therefore allows for combining different approximations (one iterations or several) for different sub-problems.

In Section 3, we propose the general algorithmic framework of NAM for solving non-convex and non-smooth problems. Our main contribution is to show that this algorithmic framework, which captures a wide spectrum of algorithms, generates a globally convergent sequence to critical points of the problem at hand. To this end, we define the notion of Nested Friendly Algorithms (NFA), which captures the minimal necessary requirements of the nested algorithm to be used within NAM in order to achieve the global convergence of NAM. In Section 4, we illustrate the concept of NFA by presenting several examples that satisfy the requirements and therefore can be used in NAM. Another central aspect in applying NAM is to determine the number of inner iterations which is sufficient to guarantee the global convergence. This is also discussed in Section 4. Finally, in Section 5 we illustrate the advantage of the NAM framework in tackling the Regularized Structured Total Least Squares (RSTLS) problem via its application on an image deblurring task.

2 A Procedure for Global Convergence with Errors

Consider an extended real-valued function $F\colon \mathbb{R}^d \times \mathbb{R}^{d_0} \to (-\infty, \infty]$ that is bounded from below and assumed to be proper and lower semi-continuous. We consider the following two-block minimization problem

$$\min_{(\mathbf{z},\mathbf{u})\in\mathbb{R}^d\times\mathbb{R}^{d_0}} F\left(\mathbf{z},\mathbf{u}\right). \tag{1}$$

Starting with any pair $(\mathbf{z}^0, \mathbf{u}^0)$, let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ be a sequence generated by a generic algorithm, which we denote by $\mathcal{A}$, that aims at tackling Problem (1) in the sense of global convergence to critical points of the function F. As mentioned above, our goal in this section is to extend recent proof techniques with a set of relaxed conditions that still guarantee global convergence (see [11] and also [12] for a recent and concise version).

Before proceeding we recall the following definition from non-smooth analysis (see, for instance, [27]), which will be useful in our developments below.

Definition 1. [Limiting Sub-differential]. Let $f\colon \mathbb{R}^n \to (\infty, \infty]$ be a proper and lower semicontinuous function. The *limiting sub-differential* (or simply the sub-differential) of f at $\mathbf{x} \in \mathbb{R}^n$ is denoted by $\partial f(\mathbf{x})$ and is defined as the set

$$\partial f(\mathbf{x}) \equiv \left\{ \mathbf{v} \in \mathbb{R}^n \colon \exists \mathbf{x}^k \to \mathbf{x}, f\left(\mathbf{x}^k\right) \to f(\mathbf{x}), \mathbf{v}^k \in \hat{\partial} f\left(\mathbf{x}^k\right) \to \mathbf{v} \text{ as } k \to \infty \right\},$$

where $\hat{\partial} f(\mathbf{x})$ is the Fréchet sub-differential of f at $\mathbf{x} \in \operatorname{dom}(f)$, which is defined as

$$\hat{\partial} f(\mathbf{x}) \equiv \left\{ \mathbf{v} \in \mathbb{R}^n \colon \liminf_{\mathbf{y} \neq \mathbf{x}, \mathbf{y} \to \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \mathbf{v}^T (\mathbf{y} - \mathbf{v})}{\|\mathbf{y} - \mathbf{x}\|} \geq 0 \right\},$$

and when $\mathbf{x} \notin \operatorname{dom}(f)$ we set $\hat{\partial} f(\mathbf{x}) \equiv \varnothing$.

Now, following [11], we extend the definition of gradient-like descent sequences to include non-negative errors in the following way.

Definition 2. [Approximate gradient-like descent sequence]. A sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$ is called an *approximate gradient-like descent sequence* for minimizing the function F of Problem (1), if the following conditions hold:

(C1) *Approximate sufficient decrease property.* The sequence $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$ is non-increasing and there exist a positive scalar $\rho_1 > 0$ and a non-negative error term $e_1^k \geq 0$ such that

$$\rho_1 \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 - e_1^k \leq F\left(\mathbf{z}^k, \mathbf{u}^k\right) - F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right), \quad \forall\, k \geq 0.$$

(C2) *Approximate sub-gradient lower bound on the iterates gap.* There exist a vector $\mathbf{w}^{k+1} \in \partial F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right)$, a positive scalar $\rho_2 > 0$ and a non-negative error term $e_2^k \geq 0$ such that

$$\left\|\mathbf{w}^{k+1}\right\| \leq \rho_2 \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| + e_2^k, \quad \forall\, k \geq 0.$$

(C3) *Continuity.* If $(\bar{\mathbf{z}}, \bar{\mathbf{u}})$ is a limit point of some sub-sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k \in \mathcal{K} \subseteq \mathbb{N}}$, then

$$\limsup_{k \in \mathcal{K} \subseteq \mathbb{N}} F\left(\mathbf{z}^k, \mathbf{u}^k\right) \leq F(\bar{\mathbf{z}}, \bar{\mathbf{u}}).$$

(C4) *Summability of the errors.* The sequences $\left\{\sqrt{e_1^k}\right\}_{k \geq 0}$ and $\left\{e_2^k\right\}_{k \geq 0}$ are summable, that is, $\sum_{k=1}^{\infty} \sqrt{e_1^k} < \infty$ and $\sum_{k=1}^{\infty} e_2^k < \infty$.

Few words about the new conditions (C1) and (C2). In the case that $e_1^k = e_2^k = 0$, for all $k \geq 0$, we recover the conditions introduced in [35] (note that in this case condition (C1) reduces to the "regular" sufficient decrease property and condition (C4) trivially holds) and the notion defined in Definition 2 coincides with the notion of partial gradient-like descent sequence introduced in [35]. In addition, we obviously recover the original recipe of [11] when the second block of the function F vanishes.

Condition (C1) includes two main requirements, which are: (i) approximate sufficient decrease property of the sequence $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ in the sense of the additional non-negative error term, and (ii) a non-increasing property of the sequence $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$. It should be noted that these two requirements together are obviously weaker than the "regular" sufficient decrease property, which appears in all previous recipes as mentioned above. We illustrate below that condition (C1) can be guaranteed in several settings.

Condition (C2) requires finding a sub-gradient with a norm, which is bounded from above by the gap between two successive iterations, plus some non-negative error term. This additional error term allows for some more flexibility in bounding the sub-gradient, but due to condition (C4), these errors should be small enough such that the infinite sum of them is finite. See [30] for a set of conditions which include errors only in condition (C2).

We would also like to stress the fact that now, in Definition 2, the two-block structure comes into a play since the first two conditions are only measured in terms of the variable $\mathbf{z}$ (see also [35]). This adds another level of flexibility in proving the approximate gradient-like descent sequence property.

Now, we are in a position to prove that these relaxed conditions are enough to guarantee global convergence to critical points of F. To this end, we denote by $\omega\left(\mathbf{x}^0\right)$ the set of limit points of a sequence $\left\{\mathbf{x}^k\right\}_{k\geq 0}$ and for any function f we denote by $\operatorname{crit}(f)$ the set of its critical points.

Lemma 1. *Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ be a bounded approximate gradient-like descent sequence for minimizing F of Problem* (1)*. Then, $\omega\left(\mathbf{z}^0, \mathbf{u}^0\right)$ is a non-empty and compact subset of* $\operatorname{crit}(F)$ *and*

$$\lim_{k\to\infty} \operatorname{dist}\left(\left(\mathbf{z}^k, \mathbf{u}^k\right), \omega\left(\mathbf{z}^0, \mathbf{u}^0\right)\right) = 0.$$

In addition, the function F is finite and constant on $\omega\left(\mathbf{z}^0, \mathbf{u}^0\right)$.

Proof. Since $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is bounded, $\omega\left(\mathbf{z}^0, \mathbf{u}^0\right)$ is a non-empty and compact set. Therefore, it follows that $\operatorname{dist}\left(\left(\mathbf{z}^k, \mathbf{u}^k\right), \omega\left(\mathbf{z}^0, \mathbf{u}^0\right)\right)$ converges to 0 as k goes to infinity. The non-increasing property required in condition (C1) and the boundedness from below of F, yields the convergence of the sequence $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$. From condition (C1) we have

$$\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 \leq \frac{1}{\rho_1}\left(F\left(\mathbf{z}^k, \mathbf{u}^k\right) - F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) + e_1^k\right), \tag{2}$$

which together with condition (C4) implies from (2) that

$$\lim_{k\to\infty}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| = 0. \tag{3}$$

Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\in\mathcal{K}}$ be a convergent sub-sequence to some pair $\left(\mathbf{z}^*, \mathbf{u}^*\right)$. Denote $F^* \equiv F\left(\mathbf{z}^*, \mathbf{u}^*\right)$. From condition (C3) and the lower semi-continuity of F we derive that

$$\lim_{k\to\infty} F\left(\mathbf{z}^k, \mathbf{u}^k\right) = \lim_{k\in\mathcal{K}} F\left(\mathbf{z}^k, \mathbf{u}^k\right) = F^*. \tag{4}$$

From condition (C2) there exists a sequence $\left\{\mathbf{w}^k\right\}_{k\geq 0}$, for which $\mathbf{w}^k \in \partial F\left(\mathbf{z}^k, \mathbf{u}^k\right)$ for all $k \geq 0$, that satisfies

$$\lim_{k\to\infty}\left\|\mathbf{w}^k\right\| \leq \lim_{k\to\infty}\left(\rho_2\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| + e_2^k\right) = 0,$$

where the equality follows from (3) and condition (C4). This means that the sequence $\left\{\mathbf{w}^k\right\}_{k\geq 0}$ converges to $\mathbf{0}$ as $k \to \infty$ and therefore $\left(\mathbf{z}^*, \mathbf{u}^*\right) \in \operatorname{crit}(F)$, as follows from the closedness property of the sub-differential (see [34]). Finally, since $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ converges we can assume that $\lim\limits_{k\to\infty} F\left(\mathbf{z}^k, \mathbf{u}^k\right) = c \in \mathbb{R}$. Then, $\lim\limits_{k\in\mathcal{K}} F\left(\mathbf{z}^k, \mathbf{u}^k\right) = c$ and from (4) we derive that F is finite and constant over the set $\omega\left(\mathbf{z}^0, \mathbf{u}^0\right)$. □

In order to obtain global convergence, based on Lemma 1, all we need to add is the KL property [10, 24, 25], which was shown in many papers to be a central tool in achieving global convergence results in the non-convex setting.

Definition 3. [Kurdyka–Łojasiewicz Property]. Let $f\colon \mathbb{R}^n \to (\infty, \infty]$ be a proper and lower semi-continuous function. The function f satisfies the KL property [24, 25] at $\bar{\mathbf{x}} \in$

$\mathrm{dom}\left(\partial f\right) \equiv \left\{\mathbf{x} \in \mathbb{R}^n \colon \partial f\left(\mathbf{x}\right) \neq \varnothing\right\}$ if there exist $\eta \in (0, \infty]$, a neighborhood U of $\bar{\mathbf{x}}$ and a function $\varphi \in \Phi_\eta$, such that for all $\mathbf{x} \in U \cap \left[f\left(\bar{\mathbf{x}}\right) < f\left(\mathbf{x}\right) < f\left(\bar{\mathbf{x}}\right) + \eta\right]$ it holds that

$$\varphi'\left(f\left(\mathbf{x}\right) - f\left(\bar{\mathbf{x}}\right)\right) \mathrm{dist}\left(\mathbf{0}_n, \partial f\left(\mathbf{x}\right)\right) \geq 1,$$

where Φ_η is the set of all desingularizing functions, which are concave and continuous functions $\varphi \colon [0, \eta) \to \mathbb{R}_+$ such that $\varphi\left(0\right) = 0$, φ is C^1 on $(0, \eta)$ and continuous at 0, and for all $s \in (0, \eta)$ it holds that $\varphi'\left(s\right) > 0$.

We now prove the main result of this section. It states that if a generic optimization method $\mathcal{A}$ generates a bounded approximate gradient-like descent sequence for minimizing F of Problem (1), then it globally converges to a critical point of F. It should be noted that this unique limit point is dependent on the initialization point of $\mathcal{A}$.

Theorem 1. [Global convergence]. *Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ be a bounded approximate gradient-like descent sequence for minimizing F of Problem* (1)*. If F satisfies the KL property, then the sequence $\left\{\mathbf{z}^k\right\}_{k\geq 0}$ has finite length and it globally converges to a point $\mathbf{z}^*$. Moreover, let $\mathbf{u}^*$ be any limit point of the sequence $\left\{\mathbf{u}^k\right\}_{k\geq 0}$, then $\left(\mathbf{z}^*, \mathbf{u}^*\right) \in \mathrm{crit}\left(F\right)$.*

Proof. Let $\left(\mathbf{z}^*, \mathbf{u}^*\right)$ be a limit point of the sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$, which exists due to the boundedness assumption. From Lemma 1 (see (4)) we derive

$$\lim_{k\to\infty} F\left(\mathbf{z}^k, \mathbf{u}^k\right) = F\left(\mathbf{z}^*, \mathbf{u}^*\right).$$

Assume that there exists $\tilde{k} \in \mathbb{N}$ such that $F\left(\mathbf{z}^k, \mathbf{u}^k\right) = F\left(\mathbf{z}^*, \mathbf{u}^*\right)$ and $e_1^k = 0$ for all $k \geq \tilde{k}$. Then, from condition (C1) it follows that $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ has a "regular" sufficient decrease property and therefore $\mathbf{z}^{k+1} = \mathbf{z}^k$ for all $k \geq \tilde{k}$. This proves a finite convergence of the sequence $\left\{\mathbf{z}^k\right\}_{k\geq 0}$ and obviously the desired result. If $F\left(\mathbf{z}^k, \mathbf{u}^k\right) = F\left(\mathbf{z}^*, \mathbf{u}^*\right)$ and $e_1^k > 0$ for all $k \geq \tilde{k}$, then summing condition (C1) and using condition (C4) for all $k \geq \tilde{k}$ yields

$$\sum_{k=\tilde{k}}^{\infty} \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| \leq \frac{1}{\sqrt{\rho_1}} \sum_{k=\tilde{k}}^{\infty} \sqrt{e_1^k} < \infty.$$

Hence, the sequence $\left\{\mathbf{z}^k\right\}_{k\geq 0}$ is of a finite length, and it globally converges to some $\mathbf{z}^*$.

Now, on the other hand, assume that $F\left(\mathbf{z}^k, \mathbf{u}^k\right) > F\left(\mathbf{z}^*, \mathbf{u}^*\right)$ for all $k \geq 0$. We know from condition (C1) that the sequence $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is non-increasing and converges to

$F\left(\mathbf{z}^*, \mathbf{u}^*\right)$. Thus, for any $\gamma > 0$ there exists an integer k_0 such that $F\left(\mathbf{z}^k, \mathbf{u}^k\right) < F\left(\mathbf{z}^*, \mathbf{u}^*\right) + \gamma$ for all $k > k_0$. From Lemma 1

$$\lim_{k \to \infty} \operatorname{dist}\left(\left(\mathbf{z}^k, \mathbf{u}^k\right), \omega\left(\mathbf{z}^0, \mathbf{u}^0\right)\right) = 0.$$

So, for any $\varepsilon > 0$ there exists an integer k_1 such that $\operatorname{dist}\left(\left(\mathbf{z}^k, \mathbf{u}^k\right), \omega\left(\mathbf{z}^0, \mathbf{u}^0\right)\right) < \varepsilon$ for any $k > k_1$. Again from Lemma 1 the function F is finite and constant on the non-empty and compact set $\omega\left(\mathbf{z}^0, \mathbf{u}^0\right)$. Hence, from the Uniformized KL property [11, Lemma 6] there exists a desingularizing function $\varphi \in \Phi_\gamma$ (see Definition 3) such that for any $k > s \equiv \max\left\{k_0, k_1\right\} + 1$ we have

$$\varphi'\left(F\left(\mathbf{z}^k, \mathbf{u}^k\right) - F\left(\mathbf{z}^*, \mathbf{u}^*\right)\right) \cdot \operatorname{dist}\left(\mathbf{0}, \partial F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right) \geq 1. \tag{5}$$

Since $\operatorname{dist}\left(\mathbf{0}, \partial F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right) \leq \left\|\mathbf{w}^k\right\|$ for any $\mathbf{w}^k \in \partial F\left(\mathbf{z}^k, \mathbf{u}^k\right)$ it follows from condition (C2) and (5) that

$$\varphi'\left(F\left(\mathbf{z}^k, \mathbf{u}^k\right) - F\left(\mathbf{z}^*, \mathbf{u}^*\right)\right) \geq \frac{1}{\rho_2\left\|\mathbf{z}^k - \mathbf{z}^{k-1}\right\| + e_2^{k-1}}. \tag{6}$$

For any $n_1, n_2 \in \mathbb{N}$ we denote

$$\Delta^{n_1,n_2} \equiv \varphi\left(F\left(\mathbf{z}^{n_1}, \mathbf{u}^{n_1}\right) - F\left(\mathbf{z}^*, \mathbf{u}^*\right)\right) - \varphi\left(F\left(\mathbf{z}^{n_2}, \mathbf{u}^{n_2}\right) - F\left(\mathbf{z}^*, \mathbf{u}^*\right)\right).$$

Since φ is concave it follows from the gradient inequality that

$$\Delta^{k,k+1} \geq \varphi'\left(F\left(\mathbf{z}^k, \mathbf{u}^k\right) - F\left(\mathbf{z}^*\right)\right) \cdot \left(F\left(\mathbf{z}^k, \mathbf{u}^k\right) - F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right)\right). \tag{7}$$

By combining (6) and (7) with condition (C1), we establish

$$\Delta^{k,k+1} \geq \frac{\rho_1\left(\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 - e_1^k/\rho_1\right)}{\rho_2\left(\left\|\mathbf{z}^k - \mathbf{z}^{k-1}\right\| + e_2^{k-1}/\rho_2\right)}. \tag{8}$$

Denote $c = \rho_2/\rho_1$, $\tilde{e}_1^k := e_1^k/\rho_1$ and $\tilde{e}_2^k := e_2^k/\rho_2$. We obtain from (8) that

$$\sqrt{c\Delta^{k,k+1}\left(\left\|\mathbf{z}^k - \mathbf{z}^{k-1}\right\| + \tilde{e}_2^{k-1}\right) + \tilde{e}_1^k} \geq \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|. \tag{9}$$

Now, since $c\Delta^{k,k+1}\left(\left\|\mathbf{z}^k - \mathbf{z}^{k-1}\right\| + \tilde{e}_2^{k-1}\right) \geq 0$ and $\tilde{e}_1^k \geq 0$ we obtain from (9) and the geometric-arithmetic mean inequality that

$$\begin{aligned}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| &\leq \sqrt{c\Delta^{k,k+1}\left(\left\|\mathbf{z}^k - \mathbf{z}^{k-1}\right\| + \tilde{e}_2^{k-1}\right)} + \sqrt{\tilde{e}_1^k} \\ &\leq \frac{c}{2}\Delta^{k,k+1} + \frac{1}{2}\left\|\mathbf{z}^k - \mathbf{z}^{k-1}\right\| + \frac{1}{2}\tilde{e}_2^{k-1} + \sqrt{\tilde{e}_1^k}.\end{aligned} \tag{10}$$

Summing up inequalities (10) for $r = s+1, \ldots, k$ we derive that

$$\begin{aligned}
2\sum_{r=s+1}^{k} \left\|\mathbf{z}^{r+1} - \mathbf{z}^{r}\right\| &\leq \sum_{r=s+1}^{k} \left\|\mathbf{z}^{r} - \mathbf{z}^{r-1}\right\| + c\sum_{r=s+1}^{k} \Delta^{r,r+1} + \sum_{r=s+1}^{k} \left(2\sqrt{\tilde{e}_1^r} + \tilde{e}_2^{r-1}\right) \\
&\leq \sum_{r=s+1}^{k} \left\|\mathbf{z}^{r+1} - \mathbf{z}^{r}\right\| + \left\|\mathbf{z}^{s+1} - \mathbf{z}^{s}\right\| + c\sum_{r=s+1}^{k} \Delta^{r,r+1} \\
&+ \sum_{r=s+1}^{k} \left(2\sqrt{\tilde{e}_1^r} + \tilde{e}_2^{r-1}\right) \\
&= \sum_{r=s+1}^{k} \left\|\mathbf{z}^{r+1} - \mathbf{z}^{r}\right\| + \left\|\mathbf{z}^{s+1} - \mathbf{z}^{s}\right\| + c\Delta^{s+1,k+1} \\
&+ \sum_{r=s+1}^{k} \left(2\sqrt{\tilde{e}_1^r} + \tilde{e}_2^{r-1}\right), \qquad (11)
\end{aligned}$$

where the last equality follows from the definition of $\Delta^{r,r+1}$. We now see that

$$\begin{aligned}
\sum_{r=s+1}^{k} \left\|\mathbf{z}^{r+1} - \mathbf{z}^{r}\right\| &\leq \left\|\mathbf{z}^{s+1} - \mathbf{z}^{s}\right\| + \sum_{r=s+1}^{k} \left(2\sqrt{\tilde{e}_1^r} + \tilde{e}_2^{r-1}\right) + c\Delta^{s+1,k+1} \\
&\leq \left\|\mathbf{z}^{s+1} - \mathbf{z}^{s}\right\| + \sum_{r=s+1}^{k} \left(2\sqrt{\tilde{e}_1^r} + \tilde{e}_2^{r-1}\right) + c\varphi\left(F\left(\mathbf{z}^{s+1}, \mathbf{u}^{s+1}\right) - F\left(\mathbf{z}^{*}, \mathbf{u}^{*}\right)\right),
\end{aligned}$$

where the last inequality follows from the definition of $\Delta^{s+1,k+1}$ and the fact that $\varphi \geq 0$. Due to condition (C4), the infinite sum of errors above is also finite and therefore we finally derive that $\sum_{r=1}^{\infty} \|\mathbf{z}^{r+1} - \mathbf{z}^{r}\| < \infty$. Hence, the sequence $\left\{\mathbf{z}^{k}\right\}_{k\geq 0}$ is of a finite length, and it globally converges to some $\mathbf{z}^{*}$. Last, from Lemma 1 it follows that $(\mathbf{z}^{*}, \mathbf{u}^{*})$ is a critical point of the function F. □

Remark 1. It should be noted that in Theorem 1 we prove convergence only of the sequence $\left\{\mathbf{z}^{k}\right\}_{k\geq 0}$. Therefore, if the problem and the algorithm at hand satisfy all the required conditions of Definition 2 with respect to all blocks, then the $\mathbf{u}$ block can be removed.

3 Nested Alternating Minimization Algorithms

The global convergence of Section 2 was obtained on the function F of Problem (1), which consists of two blocks: $\mathbf{z}$ and $\mathbf{u}$. The main reason for this split is flexibility in accommodating the requirements given in Definition 2. Indeed, the first two requirements (conditions (C1) and (C2)) are measured only with respect to the block $\mathbf{z}$. Therefore, when one designs an algorithm to tackle a certain optimization problem, the split into two blocks $\mathbf{z}$ and $\mathbf{u}$ should be done according to the structure of the involved functions and the analyzed algorithm in order to guarantee these conditions. In this section, we would like to describe another level of flexibility in splitting the variables into sub-blocks. We focus on the block $\mathbf{z} \in \mathbb{R}^d$ and split it into p sub-blocks, i.e., $\mathbf{z} = \left(\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_p^T\right)^T$ where $\mathbf{z}_i \in \mathbb{R}^{d_i}$ for any $i = 1, 2, \ldots, p$. Throughout this section, we will assume that the function $F\left(\mathbf{z}, \mathbf{u}\right)$ of Model (1) is of the following additive composite block structure

$$\min_{(\mathbf{z},\mathbf{u}) \in \mathbb{R}^d \times \mathbb{R}^{d_0}} \left\{ F\left(\mathbf{z}, \mathbf{u}\right) \equiv G\left(\mathbf{z}, \mathbf{u}\right) + \sum_{i=1}^{p} g_i\left(\mathbf{z}_i\right) + g_0\left(\mathbf{u}\right) \right\}, \tag{12}$$

where the function $G \colon \mathbb{R}^d \times \mathbb{R}^{d_0} \to \mathbb{R}$ is non-convex but smooth, and the functions $g_i \colon \mathbb{R}^{d_i} \to (-\infty, \infty]$, $i = 0, 1, 2, \ldots, p$, are non-convex and non-smooth. We will also make the following assumption regarding the smooth function G.

Assumption 1. The gradient $\nabla G\left(\mathbf{z}, \mathbf{u}\right)$ is L-Lipschitz continuous over bounded subsets of $\mathbb{R}^d \times \mathbb{R}^{d_0}$.

The most classical approach to tackle Model (12) is the algorithmic framework of Alternating Minimization (AM), which suggests to alternate in each iteration between the $p + 1$ blocks of Problem (12) in a cyclic fashion, while all other blocks are kept fixed.

Applying AM on Problem (12) results with $p + 1$ sub-problems that should be exactly solved at each iteration. However, as we already discussed in the Introduction, in this work we are interested in their approximations. Therefore, our main reason to consider this block model, is due to the fact that the sub-problems with respect to the sub-blocks of $\mathbf{z}$ will be approximated, while the $\mathbf{u}$ block will still be solved exactly. The approximation of each of the sub-blocks of $\mathbf{z}$ is performed using some nested algorithm (each sub-block can be updated using a different algorithm). Therefore, we refer to the algorithmic scheme as Nested Alternating Minimization (NAM), and it is recorded below in Algorithm 1.

In order to make the presentation and developments simple, we introduce the following notation. For any iteration $k \geq 0$ and any sub-block $i = 1, 2, \ldots, p$ of $\mathbf{z} \in \mathbb{R}^d$, we denote

$$\mathbf{z}^{k,i} \equiv \left(\mathbf{z}_1^k, \ldots, \mathbf{z}_i^k, \mathbf{z}_{i+1}^{k-1}, \ldots, \mathbf{z}_p^{k-1}\right) \in \mathbb{R}^d,$$

where $\mathbf{z}_i^k$, $i = 1, 2, \ldots, p$, is the k-th update of the sub-block $\mathbf{z}_i$. Notice that we have $\mathbf{z}^{k,p} = \left(\mathbf{z}_1^k, \mathbf{z}_2^k, \ldots, \mathbf{z}_p^k\right) = \mathbf{z}^k$ and for simplicity we define $\mathbf{z}^{k,0} = \mathbf{z}^{k-1}$.

Algorithm 1 Nested Alternating Minimization (NAM) Scheme

```
Input: Nested algorithms $\mathcal{A}_i$, $i = 1, 2, \ldots, p$.
Initialization: $\left(\mathbf{z}^{-1}, \mathbf{u}^0\right) \in \mathbb{R}^d \times \mathbb{R}^{d_0}$.
Iterative step:
for $k \geq 0$ do
    for $i = 1, 2, \ldots, p$ do
        Update $\mathbf{z}_i^k$ by applying iterations of algorithm $\mathcal{A}_i$ for minimizing the partial function
            $\mathbf{z}_i \mapsto F\left(\mathbf{z}_1^k, \ldots, \mathbf{z}_{i-1}^k, \mathbf{z}_i, \mathbf{z}_{i+1}^{k-1}, \ldots, \mathbf{z}_p^{k-1}, \mathbf{u}^k\right)$,
        with a starting point $\mathbf{z}_i^{k-1}$.
    end for
    Define $\mathbf{z}^k = \mathbf{z}^{k,p}$.
    Update $\mathbf{u}^{k+1} \in \text{argmin}\left\{F\left(\mathbf{z}^k, \mathbf{u}\right) : \mathbf{u} \in \mathbb{R}^{d_0}\right\}$.
end for
```

3.1 Global Convergence of NAM

In this sub-section, we would like to prove global convergence of NAM using Theorem 1. To establish this result, we will show that it is enough to prove that each of the sub-blocks satisfies the following block-wise properties.

Definition 4. [Block-wise approximate gradient-like descent sequence]. A sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$ is called a *block-wise approximate gradient-like descent sequence* for minimizing the function F of Problem (12), if the following conditions hold:

(B0) If $\bar{\mathbf{u}}$ is a limit point of some sub-sequence $\left\{\mathbf{u}^{k_j}\right\}_{j \in \mathcal{J}_0 \subseteq \mathbb{N}}$, then

$$\limsup_{j \in \mathcal{J}_0 \subseteq \mathbb{N}} g_0\left(\mathbf{u}_j^k\right) \leq g_0\left(\bar{\mathbf{u}}\right).$$

For any $i = 1, 2, \ldots, p$:

(B1) There exist $\rho_{1,i} > 0$ and a sequence of non-negative scalars $\left\{e_{1,i}^k\right\}_{k \geq 0}$ such that

$$F\left(\mathbf{z}^{k+1,i-1}, \mathbf{u}^{k+1}\right) \geq F\left(\mathbf{z}^{k+1,i}, \mathbf{u}^{k+1}\right) + \max\left\{0, \rho_{1,i}\left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\|^2 - e_{1,i}^{k+1}\right\}, \quad \forall k \geq 0.$$

(B2) There exist $\rho_{2,i} \geq 0$, a sequence of non-negative scalars $\left\{e_{2,i}^{k}\right\}_{k\geq 0}$, and a sequence of sub-gradients $\left\{\mathbf{w}_i^k\right\}_{k\geq 0}$ where $\mathbf{w}_i^k \in \partial_{\mathbf{z}_i} F\left(\mathbf{z}^{k,i}, \mathbf{u}^k\right)$ such that

$$\left\|\mathbf{w}_i^{k+1}\right\| \leq \rho_{2,i}\left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\| + e_{2,i}^{k+1}, \quad \forall k \geq 0.$$

(B3) If $\bar{\mathbf{z}}_i$ is a limit point of some sub-sequence $\left\{\mathbf{z}_i^{k_j}\right\}_{j\in\mathcal{J}_i\subseteq\mathbb{N}}$, then

$$\limsup_{j\in\mathcal{J}_i\subseteq\mathbb{N}} g_i\left(\mathbf{z}_i^{k_j}\right) \leq g_i\left(\bar{\mathbf{z}}_i\right).$$

(B4) $\sum_{k=1}^{\infty}\sqrt{e_{1,i}^{k}} < \infty$ and $\sum_{k=1}^{\infty} e_{2,i}^{k} < \infty$.

Next, we show that under Assumption 1, any block-wise approximate gradient-like descent sequence for minimizing the function F of Problem (12) generated by NAM, is also an approximate gradient-like descent sequence according to Definition 2, and therefore is a globally convergent sequence using Theorem 1. Hence, the task of proving global convergence of an algorithm that updates several blocks (could be any finite number) becomes much easier. Indeed, our results in this section show that we can focus on each sub-block separately and verify that the conditions (B0) to (B4) hold true. Therefore, as we will illustrate in Section 4, each sub-block can be updated in its own fashion and is not influenced from the updates of the other sub-blocks.

Lemma 2. *Let* $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ *be a block-wise approximate gradient-like descent sequence generated by NAM for minimizing* F *of Problem* (12)*. Then, condition (C1) of Definition 2 is satisfied, i.e., there exist* $\rho_1 > 0$ *and a sequence of non-negative scalars* $\left\{e_1^k\right\}_{k\geq 0}$ *such that*

$$F\left(\mathbf{z}^k, \mathbf{u}^{k+1}\right) \geq F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) + \max\left\{0, \rho_1\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 - e_1^{k+1}\right\}, \quad \forall k \geq 0.$$

Proof. Summing the inequalities in condition (B1) of Definition 4 for all $i = 1, 2, \ldots, p$, and using the short notations $\mathbf{z}^{k+1,p} = \mathbf{z}^{k+1}$ and $\mathbf{z}^{k+1,0} = \mathbf{z}^k$, we have

$$\begin{aligned} F\left(\mathbf{z}^k, \mathbf{u}^{k+1}\right) &\geq F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) + \sum_{i=1}^{p}\max\left\{0, \rho_{1,i}\left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\|^2 - e_{1,i}^{k+1}\right\} \\ &\geq F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) + \max\left\{0, \sum_{i=1}^{p}\left(\rho_{1,i}\left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\|^2 - e_{1,i}^{k+1}\right)\right\} \\ &\geq F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) + \max\left\{0, \rho_1\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 - e_1^{k+1}\right\}, \end{aligned} \tag{13}$$

where we set $\rho_1 = \min\left\{\rho_{1,1}, \rho_{1,2}, \ldots, \rho_{1,p}\right\} > 0$ and $e_1^k = \sum_{i=1}^{p} e_{1,i}^k \geq 0$. Finally, by combining inequality (13) with the fact that $F\left(\mathbf{z}^k, \mathbf{u}^k\right) \geq F\left(\mathbf{z}^k, \mathbf{u}^{k+1}\right)$, which follows directly from the definition of $\mathbf{u}^{k+1}$ in NAM (see step 9), the required result follows. □

Lemma 3. *Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ be a bounded block-wise approximate gradient-like descent sequence generated by NAM for minimizing F of Problem* (12) *and suppose that Assumption 1 holds. Then, condition (C2) of Definition 2 is satisfied, i.e., there exist $\rho_2 > 0$, a sequence of non-negative scalars $\left\{e_2^k\right\}_{k\geq 0}$, and a sequence of sub-gradients $\left\{\mathbf{w}^k\right\}_{k\geq 0}$ where $\mathbf{w}^k \in \partial F\left(\mathbf{z}^k, \mathbf{u}^k\right)$ such that*

$$\left\|\mathbf{w}^{k+1}\right\| \leq \rho_2 \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| + e_2^{k+1}, \quad \forall k \geq 0.$$

Proof. From condition (B2) of Definition 4 it follows that there exists $\bar{\mathbf{w}}_i^{k+1}$, $i = 1, 2, \ldots, p$, for which

$$\bar{\mathbf{w}}_i^{k+1} \in \partial_{\mathbf{z}_i} F\left(\mathbf{z}^{k+1,i}, \mathbf{u}^{k+1}\right) = \nabla_{\mathbf{z}_i} G\left(\mathbf{z}^{k+1,i}, \mathbf{u}^{k+1}\right) + \partial g_i\left(\mathbf{z}_i^{k+1}\right),$$

and

$$\left\|\bar{\mathbf{w}}_i^{k+1}\right\| \leq \left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\| + e_{2,i}^{k+1}.$$

By defining

$$\mathbf{w}_i^{k+1} \equiv \bar{\mathbf{w}}_i^{k+1} - \nabla_{\mathbf{z}_i} G\left(\mathbf{z}^{k+1,i}, \mathbf{u}^{k+1}\right) + \nabla_{\mathbf{z}_i} G\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right), \tag{14}$$

it is easy to check that $\mathbf{w}_i^{k+1} \in \partial_{\mathbf{z}_i} F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right)$. In addition, we wish to bound the norm of each $\mathbf{w}_i^{k+1}$, $i = 1, 2, \ldots, p$. Using the triangle inequality, condition (B2) of Definition 4 and Assumption 1 (which can be used since the sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is bounded) we have

$$\begin{aligned}
\left\|\mathbf{w}_i^{k+1}\right\| &= \left\|\bar{\mathbf{w}}_i^{k+1} - \nabla_{\mathbf{z}_i} G\left(\mathbf{z}^{k+1,i}, \mathbf{u}^{k+1}\right) + \nabla_{\mathbf{z}_i} G\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right)\right\| \\
&\leq \left\|\bar{\mathbf{w}}_i^{k+1}\right\| + \left\|\nabla_{\mathbf{z}_i} G\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) - \nabla_{\mathbf{z}_i} G\left(\mathbf{z}^{k+1,i}, \mathbf{u}^{k+1}\right)\right\| \\
&\leq \rho_{2,i} \left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\| + e_{2,i}^{k+1} + L \sum_{j=i+1}^{p} \left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\|, \qquad (15)
\end{aligned}$$

where we used the definition of the short notation of $\mathbf{z}^{k+1,i}$. Additionally, since $\mathbf{u}^{k+1}$ is the exact minimizer of the partial function $\mathbf{u} \mapsto F\left(\mathbf{z}^k, \mathbf{u}\right)$, from the first-order optimality condition we derive that

$$\mathbf{0}_{d_0} \in \nabla_{\mathbf{u}} G\left(\mathbf{z}^k, \mathbf{u}^{k+1}\right) + \partial g_0\left(\mathbf{u}^{k+1}\right).$$

Therefore, $-\nabla_{\mathbf{u}} G\left(\mathbf{z}^k, \mathbf{u}^{k+1}\right) \in \partial g_0\left(\mathbf{u}^{k+1}\right)$ and by defining

$$\mathbf{w}_{\mathbf{u}}^{k+1} \equiv \nabla_{\mathbf{u}} G\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) - \nabla_{\mathbf{u}} G\left(\mathbf{z}^k, \mathbf{u}^{k+1}\right), \tag{16}$$

we easily obtain from Assumption 1 that

$$\left\|\mathbf{w}_{\mathbf{u}}^{k+1}\right\| \leq L\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|. \tag{17}$$

Overall, by constructing the vector $\mathbf{w}^{k+1} \equiv \left(\mathbf{w}_1^{k+1}, \mathbf{w}_2^{k+1}, \ldots, \mathbf{w}_p^{k+1}, \mathbf{w}_{\mathbf{u}}^{k+1}\right)$ it follows from (14) and (16) that $\mathbf{w}^{k+1} \in \partial F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right)$. Finally, using the triangle inequality it follows from (15) and (17) that

$$\begin{aligned}
\left\|\mathbf{w}^{k+1}\right\| &\leq \sum_{i=1}^{p}\left(\rho_{2,i}\left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\| + e_{2,i}^{k+1} + L\sum_{j=i+1}^{p}\left\|\mathbf{z}_i^{k+1} - \mathbf{z}_i^k\right\|\right) + L\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| \\
&\leq \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|\sum_{i=1}^{p}\rho_{2,i} + \sum_{i=1}^{p} e_{2,i}^{k+1} + L\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|\sum_{i=1}^{p}(p-i) + L\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| \\
&= \rho_2\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\| + e_2^{k+1},
\end{aligned}$$

where we set

$$\rho_2 \equiv L + \frac{L \cdot p\,(p-1)}{2} + \sum_{i=1}^{p}\rho_{2,i} > 0 \quad \text{and} \quad e_2^k \equiv \sum_{i=1}^{p} e_{2,i}^k \geq 0,$$

and the proof is completed. □

Lemma 4. *Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ be a block-wise approximate gradient-like descent sequence generated by NAM for minimizing F of Problem (12). Then, condition (C3) of Definition 2 is satisfied, i.e., if $(\bar{\mathbf{z}}, \bar{\mathbf{u}})$ is a limit point of some sub-sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\in\mathcal{K}\subseteq\mathbb{N}}$, then*

$$\limsup_{k\in\mathcal{K}\subseteq\mathbb{N}} F\left(\mathbf{z}^k, \mathbf{u}^k\right) \leq F\left(\bar{\mathbf{z}}, \bar{\mathbf{u}}\right).$$

Proof. Recall that $F(\mathbf{z}, \mathbf{u}) \equiv G(\mathbf{z}, \mathbf{u}) + \sum_{i=1}^{p} g_i(\mathbf{z}_i) + g_0(\mathbf{u})$ and G is continuous. Therefore, the result immediately follows from conditions (B0) and (B3) for all $i = 1, 2, \ldots, p$ of Definition 4 . □

Equipped with the three above lemmas, we can now state a global convergence result of NAM for minimizing the function F of Problem (12).

Theorem 2. *Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$ be a bounded block-wise approximate gradient-like descent sequence generated by NAM for minimizing F of Problem* (12) *under Assumption 1. If F satisfies the KL property, then the sequence $\left\{\mathbf{z}^k\right\}_{k \geq 0}$ has finite length and it globally converges to some $\mathbf{z}^* \in \mathbb{R}^d$. Moreover, let $\mathbf{u}^* \in \mathbb{R}^{d_0}$ be any limit point of $\left\{\mathbf{u}^k\right\}_{k \geq 0}$, then $\left(\mathbf{z}^*, \mathbf{u}^*\right) \in \operatorname{crit}(F)$.*

Proof. According to Theorem 1, all we need to show is that the sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$ is an approximate gradient-like descent sequence for minimizing F according to Definition 2. Conditions (C1), (C2) and (C3) of Definition 2 follow from Lemmas 2, 3 and 4, respectively. Finally, condition (C4) follows from condition (B4) of Definition 4 using the definitions of e_1^k and e_2^k introduced in the proofs of Lemma 2 and 3, respectively. □

4 Examples of Nested Algorithms

In the previous section, we have developed a theory that guarantees global convergence of NAM for any finite number of sub-blocks p, where we only need to verify the block-wise properties of Definition 4. Satisfying these block-wise properties is done by choosing suitable nested algorithms $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_p$. Therefore, a natural question is which nested algorithms can be chosen in order to satisfy these block-wise properties.

In this section, our main goal is to study nested algorithms that satisfy the block-wise conditions of Definition 4. To this end, we will introduce a new class of nested algorithms, called Nested Friendly Algorithms (NFA), and we will show that this class indeed accommodate the required block-wise properties. We will show below that some very well-known algorithms (such as Accelerated Gradient Descent Method [29]) are NFAs. It should be noted that NFA can be used for sub-blocks which satisfy the following structural assumption:

($\mathcal{T}1$) The partial function $\mathbf{z}_i \mapsto F(\mathbf{z}, \mathbf{u})$ is smooth and strongly convex (i.e., in this case $g_i(\mathbf{z}_i) \equiv 0$ in Model (12)).

Even though this assumption is not mild, there are several challenging and important non-convex applications that have sub-blocks with this structure, and we will discuss one such an application in detail below (see Section 5).

Before discussing the new concept of NFA, we would like to mention that our NAM framework also covers the following scenario. Suppose that some sub-blocks of $\mathbf{z}$ do not satisfy the structural assumption ($\mathcal{T}1$). In this case, we can still approximate the solution of the corresponding sub-problem using a One Iteration Approximation (OIA) as we discussed in the Introduction. In this case, the corresponding partial function could be any non-smooth and non-convex function. Thus, we define the following additional structural assumption:

($\mathcal{T}2$) The partial function $\mathbf{z}_i \mapsto F(\mathbf{z}, \mathbf{u})$ is non-smooth and non-convex.

We will show below, under a mild and standard assumption on the corresponding partial function, that OIA of descent algorithms indeed satisfy the block-wise properties of Definition 4 and thus can be easily integrated into NAM. Therefore, the NAM framework generalizes existing globally convergent algorithms which are only based on OIA, such as PALM [11] and many more (see, for instance, [30, 14, 15, 37]). Thus, the NAM framework can easily combine sub-blocks which are updated using NFA with sub-blocks which are updated using an OIA.

Remark 2. Notice that if the partial function $\mathbf{z}_i \mapsto F(\mathbf{z}, \mathbf{u})$ is strongly convex, then another immediate approach to solve the corresponding sub-problem that satisfies all the block-wise properties is exact minimization.

As we emphasized in Section 3, in order to guarantee the global convergence of any NAM, all we need is to verify block-wise conditions and the integration is already proved above in Theorem 2. This means that in the rest of this section we can focus on proving that both NFA and OIA indeed satisfy these block-wise conditions for sub-blocks of type ($\mathcal{T}1$) and ($\mathcal{T}2$), respectively. Therefore, the number of sub-blocks and the order of updating the sub-blocks (could be different at each iteration), do not play any role and thus, for the sake of simplicity of the presentation, we skip all sub-block notations and focus on two sub-blocks of $\mathbf{z}$: $\mathbf{z}_1$ of type ($\mathcal{T}1$) and $\mathbf{z}_2$ of type ($\mathcal{T}2$). In other words, from now on, our model is as follows:

$$\min_{(\mathbf{z},\mathbf{u}) \in \mathbb{R}^d \times \mathbb{R}^{d_0}} \left\{ F(\mathbf{z}_1, \mathbf{z}_2, \mathbf{u}) \equiv G(\mathbf{z}_1, \mathbf{z}_2, \mathbf{u}) + g_2(\mathbf{z}_2) + g_0(\mathbf{u}) \right\}, \tag{18}$$

where F is bounded from below by some $\underline{F} \in \mathbb{R}$, G is a non-convex but smooth function, which satisfies Assumption 1, and the functions $g_i \colon \mathbb{R}^{d_i} \to (-\infty, \infty]$, $i = 0, 2$, are non-convex and non-smooth (notice that in this model we set $g_1 \equiv 0$). For the sake of convenience of the reader, we therefore recall the NAM algorithm in its two sub-blocks form, where the sub-block $\mathbf{z}_1$ of type $(\mathcal{T}1)$ is approximated using NFA and the sub-block $\mathbf{z}_2$ of type $(\mathcal{T}2)$ is approximated using one iteration of a descent algorithm.

Algorithm 2 Nested Alternating Minimization (NAM) Scheme for Two Sub-blocks

```
**Input:** A nested NFA algorithm $\mathcal{A}_1$ and a descent algorithm $\mathcal{A}_2$.
**Initialization:** $\left(\mathbf{z}_1^{-1}, \mathbf{z}_2^{-1}, \mathbf{u}^0\right) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_0}$.
**Iterative step:**
**for** $k \geq 0$ **do**
    Update $\mathbf{z}_1^k$ by $j_k \in \mathbb{N}$ iterations of $\mathcal{A}_1$ for minimizing $\mathbf{z}_1 \mapsto F\left(\mathbf{z}_1, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right)$ starting from $\mathbf{z}_1^{k-1}$.
    Update $\mathbf{z}_2^k$ by one iteration of $\mathcal{A}_2$ for minimizing $\mathbf{z}_2 \mapsto F\left(\mathbf{z}_1^k, \mathbf{z}_2, \mathbf{u}^k\right)$ starting from $\mathbf{z}_2^{k-1}$.
    Update $\mathbf{u}^{k+1} \in \operatorname{argmin}\left\{F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}\right) : \mathbf{u} \in \mathbb{R}^{d_0}\right\}$.
**end for**
```

We now analyze the above algorithm, and according to the discussion above our results can be used to any NAM with any finite number of sub-blocks in any combination of NFA/OIA updates. The rest of this section is organized as follows: we will first show that sub-blocks of type $(\mathcal{T}2)$, which are solved using an OIA, satisfy the block-wise properties of Definition 4. It should be noted that the results obtained in [5] and [11] actually show these conditions, but for the sake of completeness and for a unified presentation we provide below the proofs. Then, we will show that sub-blocks of type $(\mathcal{T}1)$, which are solved using NFA, satisfy the block-wise properties and we will prove that some well-known algorithms are NFA.

4.1 Solving Sub-blocks of Type $(\mathcal{T}2)$ via an OIA

In this sub-section, we will focus on the minimization with respect to the sub-block $\mathbf{z}_2$, which has the form of a classical additive composite model, i.e., sum of a smooth function and a non-smooth function. Therefore, we will focus on a specific OIA, which is the celebrated Proximal Grdaient method. More precisely, in this case, the update step of the sub-block $\mathbf{z}_2$

is given by

$$\begin{aligned}\mathbf{z}_2^k &= \underset{\mathbf{z}_2 \in \mathbb{R}^{d_1}}{\operatorname{argmin}} \left\{ \nabla_{\mathbf{z}_2} G\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right)^T \left(\mathbf{z}_2 - \mathbf{z}_2^{k-1}\right) + \frac{t_k}{2} \left\|\mathbf{z}_2 - \mathbf{z}_2^{k-1}\right\|^2 + g_2\left(\mathbf{z}_2\right) \right\} \\ &\equiv \operatorname{prox}_{t_k}^{g_2}\left(\mathbf{z}_2^{k-1} - \frac{1}{t_k} \nabla_{\mathbf{z}_2} G\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right)\right), \end{aligned} \tag{19}$$

where by $\operatorname{prox}_{t_k}^{g_2}(\cdot)$ we denote the Moreau proximal mapping [28] of g_2, and $t_k = 2L_2\left(\mathbf{z}_1^k, \mathbf{u}^k\right)$ when $L_2\left(\mathbf{z}_1, \mathbf{u}\right)$ is the Lipschitz constant of the gradient of the partial function $\mathbf{z}_2 \mapsto G\left(\mathbf{z}_1, \mathbf{z}_2, \mathbf{u}\right)$ for fixed $\mathbf{z}_1 \in \mathbb{R}^{d_1}$ and $\mathbf{u} \in \mathbb{R}^{d_0}$. It should be noted that the step-size t_k, which is used in (19), can be chosen in a more delicate way, and actually any $t_k > L_2\left(\mathbf{z}_1^k, \mathbf{u}^k\right)$ could work, but for the sake of simplicity we take $t_k = 2L_2\left(\mathbf{z}_1^k, \mathbf{u}^k\right)$.

Now, we will show that OIA using Proximal Gradient indeed satisfies the block-wise conditions. To this end, we will make the following assumption regarding the partial function of the sub-block $\mathbf{z}_2$.

Assumption 2. For any pair $\left(\mathbf{z}_1, \mathbf{u}\right)$ the partial function $\mathbf{z}_2 \mapsto G\left(\mathbf{z}_1, \mathbf{z}_2, \mathbf{u}\right)$ has an $L_2\left(\mathbf{z}_1, \mathbf{u}\right)$-Lipschitz continuous gradient. In addition, for any compact subset $B \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_0}$ there exist constants $\bar{L}_2, \underline{L}_2 > 0$ such that

$$\inf\left\{L_2\left(\mathbf{z}_1, \mathbf{u}\right) : \left(\mathbf{z}_1, \mathbf{u}\right) \in B\right\} = \underline{L}_2,$$

and

$$\sup\left\{L_2\left(\mathbf{z}_1, \mathbf{u}\right) : \left(\mathbf{z}_1, \mathbf{u}\right) \in B\right\} = \bar{L}_2.$$

Proposition 1. *Let $\left\{\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$ be a bounded sequence generated by NAM under Assumption 2. Then, conditions (B1) to (B4) of Definition 4 hold true for the sub-block $\mathbf{z}_2$.*

Proof. Recall that each sub-problem with respect to the sub-block $\mathbf{z}_2$ is solved using one iteration of the Proximal Gradient method, as in (19). To prove condition (B1), we obtain from [11, Lemma 2] that

$$\begin{aligned} F\left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right) - F\left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^{k+1}, \mathbf{u}^{k+1}\right) &\geq \frac{L_2\left(\mathbf{z}_1^{k+1}, \mathbf{u}^{k+1}\right)}{2} \left\|\mathbf{z}_2^{k+1} - \mathbf{z}_2^k\right\|^2 \\ &\geq \frac{\underline{L}_2}{2} \left\|\mathbf{z}_2^{k+1} - \mathbf{z}_2^k\right\|^2, \end{aligned} \tag{20}$$

where the last inequality follows from Assumption 2. Now, condition (B1) easily follows from (20) by setting $\rho_{1,2} = \underline{L}_1/2$ and $e_{1,2}^k = 0$ for all $k \geq 0$.

Now we prove condition (B2). Using the update rule (19) it follows from the first-order optimality condition that there exists $\mathbf{g}^k \in \partial g_2\left(\mathbf{z}_2^k\right)$ such that

$$\nabla_{\mathbf{z}_2} G\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) + t_k\left(\mathbf{z}_2^k - \mathbf{z}_2^{k-1}\right) + \mathbf{g}^k = \mathbf{0}_{d_2},$$

which results in the inclusion

$$-\nabla_{\mathbf{z}_2} G\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) + t_k\left(\mathbf{z}_2^{k-1} - \mathbf{z}_2^k\right) \in \partial g_2\left(\mathbf{z}_2^k\right).$$

This means that

$$\mathbf{w}_2^k \equiv \nabla_{\mathbf{z}_2} G\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right) - \nabla_{\mathbf{z}_2} G\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) + t_k\left(\mathbf{z}_2^{k-1} - \mathbf{z}_2^k\right) \in \partial_{\mathbf{z}_2} F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right).$$

Now, from the triangle inequality and Assumption 2 it follows that

$$\begin{aligned}\left\|\mathbf{w}_2^{k+1}\right\| &= L_2\left(\mathbf{z}_1^{k+1}, \mathbf{u}^{k+1}\right)\left\|\mathbf{z}_2^{k+1} - \mathbf{z}_2^k\right\| + t_{k+1} \cdot \left\|\mathbf{z}_2^{k+1} - \mathbf{z}_2^k\right\| \\ &= 3L_2\left(\mathbf{z}_1^{k+1}, \mathbf{u}^{k+1}\right)\left\|\mathbf{z}_2^{k+1} - \mathbf{z}_2^k\right\| \\ &\le 3\bar{L}_2\left\|\mathbf{z}_2^{k+1} - \mathbf{z}_2^k\right\|,\end{aligned}$$

and the result follows by setting $\rho_{2,2} = 3\bar{L}_2$ and $e_{2,2}^k = 0$ for all $k \ge 0$.

To prove condition (B3), we take a sub-sequence $\left\{\mathbf{z}_2^{k_j}\right\}_{j\ge1}$, which converges to some $\bar{\mathbf{z}}_2$. From the definition of $\mathbf{z}_2^k$ as a minimizer of the proximal gradient operator (see (19)), we have for all $k \ge 0$ that

$$\nabla_{\mathbf{z}_2} G\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right)^T\left(\mathbf{z}_2^k - \bar{\mathbf{z}}_2\right) + \frac{t_k}{2}\left\|\mathbf{z}_2^k - \mathbf{z}_2^{k-1}\right\|^2 - \frac{t_k}{2}\left\|\bar{\mathbf{z}}_2 - \mathbf{z}_2^{k-1}\right\|^2 + g_2\left(\mathbf{z}_2^k\right) \le g_2\left(\mathbf{z}_2^*\right).$$

Therefore, substituting k with k_j and taking $j \to \infty$, it follows from Assumption 2 that

$$\limsup_{j\to\infty} g_2\left(\mathbf{z}_2^{k_j}\right) \le g_2\left(\bar{\mathbf{z}}_2\right),$$

since from (20) it it easily deduced that $\left\|\mathbf{z}_2^k - \mathbf{z}_2^{k-1}\right\| \to \infty$ as $k \to \infty$. Last, condition (B4) immediately follows from the fact that $e_{1,2}^k = e_{2,2}^k = 0$ for all $k \ge 0$, and the proof is completed. □

4.2 Solving Sub-blocks of Type ($\mathcal{T}1$) via NFA

In this sub-section, we will show that the class of Nested Friendly Algorithms for solving sub-problems in NAM with respect to sub-blocks of type ($\mathcal{T}1$) satisfies the properties of a block-wise approximate gradient like-descent sequence according to Definition 4. To this end, we will make the following assumption regarding the sub-block $\mathbf{z}_1$ of type ($\mathcal{T}1$).

Assumption 3. For any pair $(\mathbf{z}_2, \mathbf{u})$ the partial function $\mathbf{z}_1 \mapsto F(\mathbf{z}_1, \mathbf{z}_2, \mathbf{u})$ is $\sigma_1(\mathbf{z}_2, \mathbf{u})$-strongly convex with an $L_1(\mathbf{z}_2, \mathbf{u})$-Lipschitz continuous gradient. In addition, for any compact subset $B \subset \mathbb{R}^{d_2} \times \mathbb{R}^{d_0}$ there exist constants $\underline{\sigma}_1, \bar{L}_1 > 0$ such that

$$\inf \{\sigma_1(\mathbf{z}_2, \mathbf{u}) : \ (\mathbf{z}_2, \mathbf{u}) \in B\} = \underline{\sigma}_1,$$

and

$$\sup \{L_1(\mathbf{z}_2, \mathbf{u}) : \ (\mathbf{z}_2, \mathbf{u}) \in B\} = \bar{L}_1.$$

Now we are ready to define the class of NFA.

Definition 5. [Nested friendly algorithm]. Let $\varphi_k \colon \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbb{N}$, be a family of convex functions, each with a minimizer $\mathbf{v}_k^*$ and an optimal value φ_k^*. Let $\mathcal{A}$ be an optimization algorithm that generates, for any $k \in \mathbb{N}$, a sequence $\left\{\mathbf{v}^{k,i}\right\}_{i \geq 0}$ starting from $\mathbf{v}^{k,0}$. We say that $\mathcal{A}$ is a Nested Friendly Algorithm (NFA) with respect to $\{\varphi_k\}_{k \geq 0}$, if for any $k \in \mathbb{N}$ there exist an index $j_k \in \mathbb{N}$ and a non-increasing sequence of non-negative scalars $\{c_{\mathcal{A}}(k,j)\}_{j \geq 0}$ such that

(N1) $\lim\limits_{j \to \infty} c_{\mathcal{A}}(k,j) = 0$.

(N2) $\varphi_k\left(\mathbf{v}^{k,j}\right) - \varphi_k^* \leq \frac{c_{\mathcal{A}}^2(k,j)}{2} \cdot \left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|^2$ for all $j \geq j_k$.

(N3) $\sum\limits_{k=1}^{\infty} \sqrt{c_{\mathcal{A}}(k, j_k)} < \infty$.

Remark 3. It should be noted that if algorithm $\mathcal{A}$ is NFA with some $\{j_k\}_{k \geq 0}$, then it is obviously also NFA with any $\{l_k\}_{k \geq 0}$, where $l_k \geq j_k$ for all $k \geq 0$.

In this sub-section, we will prove that under Assumption 3, when choosing an algorithm $\mathcal{A}_1$ for minimizing the smooth partial function (see step 5 in Algorithm 2)

$$F_1^k(\mathbf{z}_1) \equiv F\left(\mathbf{z}_1, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) = G\left(\mathbf{z}_1, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right),$$

to be NFA, then conditions (B1) to (B4) of Definition 4 hold true. Before getting into the proof, we would like to show that NFAs satisfy some properties, which will be useful for us in proving this required result.

Lemma 5. *Let $\varphi_k \colon \mathbb{R}^n \to \mathbb{R}$, $k \in \mathbb{N}$, be a family of continuously differentiable and σ_k-strongly convex functions with an L_k-Lipschitz continuous gradient such that $\sigma_k \geq \underline{\sigma} > 0$ for all $k \geq 0$. Let $\mathcal{A}$ be an NFA with respect to $\{\varphi_k\}_{k\geq 0}$ that generates sequences $\left\{\mathbf{v}^{k,j}\right\}_{j\geq 0}$ with $\mathbf{v}^{k,0} = \mathbf{v}^{k-1,j_{k-1}}$ for some starting point $\mathbf{v}^{0,0} = \mathbf{v}^0$. Then, there exists $l_k \geq j_k$, for all $k \geq 0$, such that the following properties hold for all $k \geq 0$*

(i) Overall non-increasing function value. $c_{\mathcal{A}}^2\left(k, l_k\right) \leq \underline{\sigma}$ *and*

$$\varphi_k\left(\mathbf{v}^{k,l_k}\right) \leq \varphi_k\left(\mathbf{v}^{k,0}\right) = \varphi_k\left(\mathbf{v}^{k-1,l_{k-1}}\right).$$

(ii) Convergence of the gradients.

$$\left\|\nabla\varphi_k\left(\mathbf{v}^{k,l_k}\right)\right\| \leq \delta^{k,l_k} \equiv \frac{L_k \cdot c_{\mathcal{A}}\left(k, l_k\right)\left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|}{\sqrt{\sigma_k}}.$$

Proof. Since $\mathcal{A}$ is NFA, we know from Definition 5 that there exists a sequence $\{j_k\}_{k\geq 0}$ such that (N2) and (N3) of Definition 5 hold. To prove that item (i) holds, we use the strong convexity of φ_k to establish that

$$\varphi_k\left(\mathbf{v}^{k,0}\right) \geq \varphi_k^* + \nabla\varphi_k\left(\mathbf{v}_k^*\right)^T\left(\mathbf{v}^{k,0} - \mathbf{v}_k^*\right) + \frac{\sigma_k}{2}\left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|^2 = \varphi_k^* + \frac{\sigma_k}{2}\left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|^2, \quad (21)$$

where the last equality follows from the fact that $\mathbf{v}_k^*$ is a minimizer of φ_k and therefore $\nabla\varphi_k\left(\mathbf{v}_k^*\right) = \mathbf{0}$. This fact with (N2) of Definition 5 and Remark 3 yields

$$\varphi_k\left(\mathbf{v}^{k,j_k}\right) - \varphi_k^* \leq \frac{c_{\mathcal{A}}^2\left(k, j_k\right)}{\sigma_k}\left(\varphi_k\left(\mathbf{v}^{k,0}\right) - \varphi_k^*\right) \leq \frac{c_{\mathcal{A}}^2\left(k, j_k\right)}{\underline{\sigma}}\left(\varphi_k\left(\mathbf{v}^{k,0}\right) - \varphi_k^*\right),$$

where the last inequality follows from the assumption that $\sigma_k \geq \underline{\sigma}$ for all $k \geq 0$. Therefore, item (i) holds for any $l_k \geq j_k$ that satisfies $c_{\mathcal{A}}^2\left(k, l_k\right) \leq \underline{\sigma}$. Since (N1) of Definition 5 states that $c_{\mathcal{A}}\left(k, j\right) \to 0$ as $j \to \infty$, there must exist such an index $l_k \geq j_k$.

To show that item (ii) holds, we use the fact that the functions φ_k, for all $k \geq 0$, are strongly convex with an L_k-Lipschitz continuous gradient to obtain that

$$\varphi_k\left(\mathbf{v}^{k,l_k}\right) - \varphi_k^* \geq \frac{\sigma_k}{2}\left\|\mathbf{v}^{k,l_k} - \mathbf{v}_k^*\right\|^2 \geq \frac{\sigma_k}{2L_k^2}\left\|\nabla\varphi_k\left(\mathbf{v}^{k,l_k}\right) - \nabla\varphi\left(\mathbf{v}_k^*\right)\right\|^2 = \frac{\sigma_k}{2L_k^2}\left\|\nabla\varphi_k\left(\mathbf{v}^{k,l_k}\right)\right\|^2,$$

where the first inequality follows again from the fact that $\nabla\varphi_k\left(\mathbf{v}_k^*\right) = \mathbf{0}$. This fact with (N2) of Definition 5 yields

$$\left\|\nabla\varphi_k\left(\mathbf{v}^{k,l_k}\right)\right\| \leq \frac{c_{\mathcal{A}}\left(k, l_k\right)}{\sqrt{\sigma_k}} \cdot L_k\left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|, \quad (22)$$

and item (ii) holds true with $\delta^{k,l_k} \equiv L_k \cdot c_{\mathcal{A}}(k, l_k) \left\| \mathbf{v}^{k,0} - \mathbf{v}_k^* \right\| / \sqrt{\sigma_k}$, which completes the proof. □

Remark 4. Equipped with the properties of NFAs obtained in Lemma 5, we would like to describe the concept of NFA in the context of NAM for the sub-block $\mathbf{z}_1$ of type ($\mathcal{T}1$) under Assumption 3. According to Lemma 5, saying that algorithm $\mathcal{A}_1$ in NAM (see step 5 in Algorithm 1) is NFA means that with $\varphi_k := F_1^k(\mathbf{z}_1)$, at each outer iteration $k \geq 0$, the algorithm $\mathcal{A}_1$ iteratively generates a sequence $\left\{\mathbf{z}_1^{k,j}\right\}_{j \geq 0}$, which starts at $\mathbf{z}_1^{k,0} = \mathbf{z}_1^{k-1}$ and stops after j_k inner iterations at $\mathbf{z}_1^{k,j_k} = \mathbf{z}_1^k$ (for simplicity and in order to guarantee that also Lemma 5 holds true, we assume without the loss of generality that $j_k = l_k$), such that

(a) Overall non-increasing sequence:

$$F\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) \leq F\left(\mathbf{z}_1^{k-1}, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right), \quad \forall k \geq 0.$$

We emphasize that the sequences $\left\{F_1^k\left(\mathbf{z}_1^{k,j}\right)\right\}_{j \geq 0}$ for all $k \geq 0$ need not be non-increasing.

(b) Bounded gradient:

$$\left\| \nabla_{\mathbf{z}_1} F\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) \right\| \leq \delta_1^k, \quad \forall k \geq 0,$$

where

$$\delta_1^k \equiv \delta_1^{k,j_k} = \frac{L_1\left(\mathbf{z}_2^{k-1}, \mathbf{u}^k\right)}{\sqrt{\sigma_1\left(\mathbf{z}_2^{k-1}, \mathbf{u}^k\right)}} \cdot c_{\mathcal{A}_1}(k, j_k) \left\| \mathbf{z}_1^k - (\mathbf{z}_1)_k^* \right\|,$$

and $(\mathbf{z}_1)_k^* \in \mathbb{R}^{d_1}$ is the minimizer of the strongly convex partial function $F_1^k(\mathbf{z}_1)$.

Now, equipped with the above theory, we are ready to prove that NFAs satisfy the conditions of a block-wise gradient-like descent sequence.

Proposition 2. *Let $\left\{\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$ be a bounded sequence generated by NAM under Assumption 3. Then, conditions (B1) to (B4) of Definition 4 hold true for the sub-block $\mathbf{z}_1$.*

Proof. To prove condition (B1), we notice from Remark 4(a) that after j_k (inner) iterations of $\mathcal{A}_1$ (which is NFA) for any (outer) iteration $k \geq 0$ it holds that

$$F\left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right) \leq F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right). \tag{23}$$

From the update of the sub-block $\mathbf{z}_2$ by a descent algorithm (see (20)), and from the update of the $\mathbf{u}$ block, it follows that the sequence $\left\{F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is non-increasing. In addition, since the sequence $\left\{\mathbf{z}_1^k\right\}_{k\geq 0}$ is assumed to be bounded, there exists $M_1 > 0$ such that $\left\|\mathbf{z}_1^{k+1} - \mathbf{z}_1^k\right\| \leq M_1$ for all $k \geq 0$. Now, the strong convexity of the partial function $F_1^k\left(\mathbf{z}_1\right)$ yields

$$\begin{aligned} F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right) - F\left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right) &\geq \nabla_{\mathbf{z}_1} F\left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right)^T \left(\mathbf{z}_1^k - \mathbf{z}_1^{k+1}\right) \\ &+ \frac{\sigma_1\left(\mathbf{z}_2^k, \mathbf{u}^{k+1}\right)}{2} \left\|\mathbf{z}_1^k - \mathbf{z}_1^{k+1}\right\|^2 \\ &\geq -\left\|\nabla_{\mathbf{z}_1} F\left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right)\right\| \cdot \left\|\mathbf{z}_1^k - \mathbf{z}_1^{k+1}\right\| \\ &+ \frac{\sigma_1\left(\mathbf{z}_2^k, \mathbf{u}^{k+1}\right)}{2} \left\|\mathbf{z}_1^k - \mathbf{z}_1^{k+1}\right\|^2 \\ &\geq \frac{\underline{\sigma}_1}{2} \left\|\mathbf{z}_1^k - \mathbf{z}_1^{k+1}\right\|^2 - M_1 \delta_1^{k+1}, \end{aligned} \tag{24}$$

where the second inequality follows from the Cauchy-Schwartz inequality, and the last inequality follows from Assumption 3 and Remark 4(b). Therefore, condition (B1) is established by combining inequalities (23) with (24) and by setting $\rho_{1,1} = \underline{\sigma}_1/2 > 0$ and $e_{1,1}^k = M_1 \delta_1^k \geq 0$.

Condition (B2) follows from Remark 4(b), since $\mathbf{w}_1^{k+1} = \nabla_{\mathbf{z}_1} F\left(\mathbf{z}_1^{k+1}, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right)$, by setting $\rho_{2,1} = 0$ and $e_{2,1}^k = \delta_1^k$. Condition (B3) easily follows from the smoothness of this partial function.

Now we will prove condition (B4). By the definition of $e_{1,1}^k$ and $e_{2,1}^k$ it is enough to show that the sequence $\left\{\sqrt{\delta_1^k}\right\}_{k\geq 0}$ is summable. From the $\underline{\sigma}_1$-strong convexity of the partial function $F_1^k\left(\mathbf{z}_1\right)$, for all $k \geq 0$, we have

$$\begin{aligned} F\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) - F\left(\mathbf{z}_1^*, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) &\geq \nabla_{\mathbf{z}_1} F\left(\mathbf{z}_1^*, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right)^T \left(\mathbf{z}_1^k - \left(\mathbf{z}_1\right)_k^*\right) \\ &+ \frac{\underline{\sigma}_1}{2} \left\|\mathbf{z}_1^k - \left(\mathbf{z}_1\right)_k^*\right\|^2 \\ &= \frac{\underline{\sigma}_1}{2} \left\|\mathbf{z}_1^k - \left(\mathbf{z}_1\right)_k^*\right\|^2, \end{aligned} \tag{25}$$

where the last equality follows from the fact that $\nabla_{\mathbf{z}_1} F\left(\mathbf{z}_1^*, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) = \mathbf{0}_{d_1}$. Therefore, since we assume that F is bounded from below by some $\underline{F} \in \mathbb{R}$, it follows from (25) that

$$\left\|\mathbf{z}_1^k - \left(\mathbf{z}_1\right)_k^*\right\|^2 \leq \frac{2}{\underline{\sigma}_1}\left(F\left(\mathbf{z}_1^k, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right) - F\left(\mathbf{z}_1^*, \mathbf{z}_2^{k-1}, \mathbf{u}^k\right)\right) \leq \frac{2}{\underline{\sigma}_1}\left(F\left(\mathbf{z}^{-1}, \mathbf{u}^0\right) - \underline{F}\right), \tag{26}$$

where the last inequality follows from the fact that the sequence $\left\{F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is non-increasing as we proved above. Now, using Remark 4(b) we derive from (26) that

$$\begin{aligned}
\delta_1^k &= \frac{L_1\left(\mathbf{z}_2^{k-1}, \mathbf{u}^k\right)}{\sqrt{\sigma_1\left(\mathbf{z}_2^{k-1}, \mathbf{u}^k\right)}} \cdot c_{\mathcal{A}_1}\left(k, j_k\right)\left\|\mathbf{z}_1^k - \left(\mathbf{z}_1\right)_k^*\right\| \\
&\leq \frac{\bar{L}_1}{\sqrt{\underline{\sigma}_1}} \cdot c_{\mathcal{A}_1}\left(k, j_k\right)\left\|\mathbf{z}_1^k - \left(\mathbf{z}_1\right)_k^*\right\| \\
&\leq \kappa_1 \cdot c_{\mathcal{A}_1}\left(k, j_k\right)\sqrt{2\left(F\left(\mathbf{z}_1^{-1}, \mathbf{z}_2^{-1}, \mathbf{u}^0\right) - \underline{F}\right)},
\end{aligned}$$

where $\kappa_1 \equiv \bar{L}_1/\underline{\sigma}_1$. Hence, from Definition 5(iii) we obtain that $\sum_{k=0}^{\infty}\sqrt{\delta_1^k} < \infty$ and therefore $\sum_{k=1}^{\infty}\sqrt{e_{1,1}^k} < \infty$ (from the triangle inequality) and $\sum_{k=1}^{\infty}\sqrt{e_{2,1}^k} < \infty$. Hence, condition (B4) is established and the proof is completed. □

Remark 5. In Propositions 1 and 2 we showed that solving the sub-problems using an OIA or NFA, respectively, indeed generates sequences $\left\{\mathbf{z}_1^k\right\}_{k\geq 0}$ and $\left\{\mathbf{z}_2^k\right\}_{k\geq 0}$ that satisfy conditions (B1) to (B4) of Definition 4. Therefore, all we left to show is that the updates of the **u** block satisfy condition (B0), and therefore NAM generates a block-wise approximate gradient-like descent sequence for minimizing F of Problem (12), which is globally convergent from Theorem 2. To see that the updates of the block **u** indeed satisfy condition (B0), we take a sub-sequence $\left\{\mathbf{u}^{k_j}\right\}_{j\geq 1}$ which converges to some $\bar{\mathbf{u}}$. Recall that $\mathbf{u}^{k+1}$ is an exact minimizer by its definition (see step 9 in Algorithm 1). Therefore

$$G\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right) + g_2\left(\mathbf{z}_2^k\right) + g_0\left(\mathbf{u}^{k+1}\right) \leq G\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \bar{\mathbf{u}}\right) + g_2\left(\mathbf{z}_2^k\right) + g_0\left(\bar{\mathbf{u}}\right),$$

and thus from Assumption 1 it follows that

$$g_0\left(\mathbf{u}^{k+1}\right) \leq g_0\left(\bar{\mathbf{u}}\right) + \nabla_{\mathbf{u}}G\left(\mathbf{z}\right)\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^{k+1}\right)^T\left(\bar{\mathbf{u}} - \mathbf{u}^{k+1}\right) + \frac{L}{2}\left\|\bar{\mathbf{u}} - \mathbf{u}^{k+1}\right\|^2,$$

where we used the Descent Lemma applied on the function G. Substituting k with k_j and taking $j \to \infty$ gives

$$\limsup_{j\to\infty} g_0\left(\mathbf{u}^{k_j}\right) \leq g_0\left(\bar{\mathbf{u}}\right).$$

Hence condition (B0) is also satisfied.

Remark 6. In the proof of Proposition 2 we used the fact that the sequence $\left\{F\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is non-increasing. It should be noted that this property does not require the boundedness of the sequence $\left\{\left(\mathbf{z}_1^k, \mathbf{z}_2^k, \mathbf{u}^k\right)\right\}_{k \geq 0}$. Therefore, the boundedness requirement can be obtained if, for example, the function F has a bounded level sets.

4.2.1 Examples of Nested Friendly Algorithms

In the previous sub-section, we showed that using NFA to solve sub-problems of sub-blocks of type ($\mathcal{T}1$) generates a sequence that satisfies conditions (B1) to (B4) of Definition 4. In this sub-section, we would like to show the reader that many well-known algorithms are indeed NFAs. To this end, we recall that Definition 5 includes three conditions on the algorithm to be NFA:

(N1) $\lim\limits_{j \to \infty} c_{\mathcal{A}}\left(k, j\right) = 0$.

(N2) $\varphi_k\left(\mathbf{v}^{k,j}\right) - \varphi_k^* \leq \frac{c_{\mathcal{A}}^2(k,j)}{2} \cdot \left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|^2$, for all $j \geq j_k$.

(N3) $\sum\limits_{k=1}^{\infty} \sqrt{c_{\mathcal{A}}\left(k, j_k\right)} < \infty$.

A closer inspection reveals that only the first two conditions are related to the performance of the algorithm itself (mainly in terms of rate of convergence), while the third condition concerns with controlling the number of inner iterations j_k, $k \geq 0$, to be performed at each outer iteration such that the errors over all outer iterations are summable.

Therefore, we split now our discussion into two parts. First, we show that several classical optimization methods satisfy the first two conditions. Afterwards, we will discuss how to set the number of inner iterations in general.

The first two conditions of Definition 5 revolve around a rate of convergence property of the nested algorithm in terms of function values. Hence, any algorithm with a known rate of convergence of the function values can be NFA. For example, if we take the Accelerated Gradient (AG) method of Nesterov [29] as our nested algorithm, then we have the classical result of a fast rate of convergence in terms of function values. Mathematically speaking, using the setting of Definition 5 with a family of convex functions $\left\{\varphi_k\right\}_{k \geq 0}$, each with an L_k-Lipschitz continuous gradient, then for any $k \geq 0$ we have that AG satisfies the first two conditions since in this case (see, for instance, [6, Theorem 10.34]) it holds that

$$\varphi_k\left(\mathbf{v}^{k,j}\right) - \varphi_k^* \leq \frac{2L_k \cdot \left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|^2}{\left(j+1\right)^2},$$

and therefore

$$c_{\mathcal{A}}(k,j) \equiv \frac{2\sqrt{L_k}}{j+1} \to 0 \quad \text{as} \quad j \to \infty.$$

So, as can be seen, any optimization method with a provable rate of convergence in terms of function values satisfies the first two conditions of Definition 5.

As another example, if φ_k, $k \geq 0$, is additionally σ_k-strongly convex (which is a requirement on sub-blocks of type $(\mathcal{T}1)$ that are solved using NFA), then we can use a variant of AG called V-AG [6] which exploits the strong convexity and enjoys a better rate (see [6, Theorem 10.42])

$$\varphi_k\left(\mathbf{v}^{k,j}\right) - \varphi_k^* \leq \left(1 - \frac{1}{\sqrt{\kappa_k}}\right)^j \frac{\sigma_k + L_k}{2} \left\|\mathbf{v}^{k,0} - \mathbf{v}_k^*\right\|^2,$$

and therefore

$$c_{\mathcal{A}}(k,j) \equiv \sqrt{\sigma_k + L_k}\left(1 - \frac{1}{\sqrt{\kappa_k}}\right)^{\frac{j}{2}} \to 0 \quad \text{as} \quad j \to \infty,$$

where $\kappa_k = L_k/\sigma_k \geq 1$.

Next, without focusing on a specific algorithm, we would like to focus on the issue of determining the required number of inner iterations j_k for solving sub-blocks of type $(\mathcal{T}1)$ using algorithm $\mathcal{A}$ as NFA. We recall that the number of inner iterations is dictated by two factors. The first is satisfying condition (iii) of Definition 5 (summability of the errors). The second is satisfying the inequality $c_{\mathcal{A}}^2(k, j_k) \leq \underline{\sigma}$ of condition (i) in Lemma 5, which guarantees overall non-increasing function values. Hence, we require that the number of inner iterations satisfies

(a) $c_{\mathcal{A}}^2(k, j_k) \leq \underline{\sigma}$ for all $k \geq 0$.

(b) $\sum_{k=1}^{\infty} \sqrt{c_{\mathcal{A}}(k, j_k)} < \infty$.

We should mention that it is enough to satisfy the above two requirements for all $k \geq K$ for some $K \in \mathbb{N}$ (and we use this fact later). Notice that since $c_{\mathcal{A}}(k,j) \to 0$ as $j \to \infty$ (condition (i) of Definition 5), then each requirement can be satisfied for some sequence of indices $\{j_k\}_{k \geq 0}$. Hence, in order to satisfy both requirements simultaneously, we will take a sequence which bounds both sequences from above (see below).

Requirement (a) can be easily verified since $c_{\mathcal{A}}\left(k, j_k\right)$ has an explicit dependency on j_k. For instance, in the case of AG we have

$$c_{\mathcal{A}}\left(k, j_k\right) = \frac{2\sqrt{L_k}}{j_k + 1}.$$

Hence, we can take $\left\lceil 2\sqrt{L_k/\underline{\sigma}} - 1 \right\rceil \leq j_k$. Notice that since we solve with NFA sub-blocks of type ($\mathcal{T}1$) then from Assumption 3 there exits $\bar{L}$ such that $L_k \leq \bar{L}$ for all $k \geq 0$. Therefore, we can also take

$$\left\lceil 2\sqrt{\kappa} - 1 \right\rceil \leq j_k,$$

where $\kappa = \bar{L}/\underline{\sigma} \geq 1$. Similarly, we can show that requirement (a) for V-AG is satisfied if we take

$$\left\lceil \frac{\log\left(4\right) + 2\text{log}\left(\bar{L}\right) - 4\log\left(\underline{\sigma}\right)}{\log\left(\underline{\sigma}\right) - 2\log\left(\sqrt{\bar{L}} - \sqrt{\underline{\sigma}}\right)} \right\rceil \leq j_k,$$

where we used the fact that $L_k \geq \underline{\sigma}$.

As for requirement (b), it can also be easily satisfied since all we need is to guarantee that the sequence $\left\{\sqrt{c_{\mathcal{A}}\left(k, j_k\right)}\right\}_{k\geq 0}$ converges to 0 fast enough such that its sum is finite. Since $c_{\mathcal{A}}\left(k, j\right)$ converges to 0 as $j \to \infty$ independently on the specific structure of $c_{\mathcal{A}}\left(k, j\right)$, it is enough to take a sequence $\left\{j_k\right\}_{k\geq 0}$ that grows fast enough to obtain the summability. For example, if we set $j_k = s + 2^{\lfloor k/r \rfloor} - 1$ for some fixed integers s and r, we easily obtain the desired requirement.

Last, recall that we need requirements (a) and (b) to hold together, so as mentioned above, for each $k \geq 0$, we can take the number of inner iterations to be an upper bound on the number of iterations required for (a) and (b). To achieve this, since $\bar{L}$ and $\underline{\sigma}$ are fixed for any compact set, then taking $j_k = s + 2^{\lfloor k/r \rfloor} - 1$ (as described above) guarantees that both (a) and (b) are satisfied *simultaneously* for all $k \geq K$ for some $K \in \mathbb{N}$, and there is *no need* to calculate the value of the constants $\bar{L}$ and $\underline{\sigma}$. Meaning, by taking, for example, $j_k = s + 2^{\lfloor k/r \rfloor} - 1$ inner iterations of *any* algorithm that satisfies conditions (i) and (ii) of Definition 5 guarantees that conditions (B1) to (B4) of Definition 4 are satisfied for this sub-block.

5 Numerical Experiments

In this section, we provide a specific application to show the advantage of using NAM, and more importantly NAM which is incorporated with NFA, compared to OIA methods. Unlike OIA descent methods, NAM with NFA allows several inner iterations, at each outer iteration, in order to approximate the solution of each sub-problem. We will show that these additional inner iterations at each outer iteration lead to a superior algorithm performance compared to OIA based descent methods in terms of the obtained function value and in terms of a relative error of the obtained solution. In our experiments, we will tackle the challenging non-convex problem of Regularized Structured Total Least Squares (RSTLS).

5.1 RSTLS Problem Description

The RSTLS problem arises in the world of data estimation, when trying to estimate a vector $\mathbf{x} \in \mathbb{R}^d$ such that $\mathbf{A}\mathbf{x} \approx \mathbf{b}$, where the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ are contaminated by noise. Known formulations to tackle this problem are the celebrated Least Squares (LS) approach, Total Least Squares (TLS) and Structured Total Least Squares (STLS). See, for example [17] and [2], respectively. These approaches are solved using some linear inverse methods, which often result in solutions that have large norms (see [8]). Therefore, a regularizing function is added to the problem formulation, that ends up with the RSTLS problem formulation (see, for example, [16, 7]).

In this paper, we consider the class of RSLTS problems which are formulated as the following non-convex and possibly non-smooth optimization problems

$$\min_{\mathbf{z}\in\mathbb{R}^d, \mathbf{u}\in\mathbb{R}^{d_0}} \left\{ F\left(\mathbf{z},\mathbf{u}\right) \equiv \sigma_w^2 f\left(\mathbf{z}\right) + \left\| \left(\sum_{i=1}^{d_0} \mathbf{u}_i \mathbf{A}_i \right) \mathbf{z} - \mathbf{b} \right\|^2 + \frac{\sigma_w^2}{\sigma_e^2} \left\|\mathbf{u}\right\|^2 \right\}, \tag{27}$$

where $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{d_0} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$ are given and contain some noise, σ_w and σ_e are the noise factors (see below), respectively, and $f \colon \mathbb{R}^d \to (-\infty, \infty]$ is a regularizing function. For more details about this problem formulation and its applications in image processing see [8] and references therein.

In this section, we use NAM in order to solve Problem (27). We will consider two blocks, $\mathbf{z}$ and $\mathbf{u}$, where the sub-problem with respect to $\mathbf{z}$ will be approximated, while with respect to

$\mathbf{u}$ it will be exactly solved. We chose to approximate the solution with respect to the $\mathbf{z}$ block since its dimension can be large (for example, in the experiments to follow we have $d = 256^2$). In addition, we chose to find exact solutions with respect to the $\mathbf{u}$ block since its dimension in the classical settings of this application is small (for example, in our experiments $d_0 = 6$). In this section, we do not decompose the block $\mathbf{z}$ into sub-blocks, even though our NAM framework allows us to do so.

Our goal in this section is to show that there is an advantage in taking several inner iterations for solving the sub-problem with respect to the block $\mathbf{z}$, instead of one iteration of a descent method. To this end, in each outer iteration we will take several AG inner iterations (which was proved to be NFA in Section 4), and we will compare the obtained results with the results of the Semi-Proximal Alternating (SPA) algorithm of [8], which is an OIA based descent algorithm (see details below).

In order to be compatible with the concept of NFA in the $\mathbf{z}$ block, we choose in Problem (27) the regularizing function f to be $\lambda \left\|\mathbf{z}\right\|^2$ for some regularization parameter $\lambda > 0$. This yields a non-convex formulation which is proper, lower semi-continuous, bounded from below by 0 and satisfies the KL property (all these properties can be easily verified in this case, since the function F is a quartic polynomial function). Under the notations introduced in Section 4, we can write Problem (27) as $F\left(\mathbf{z}, \mathbf{u}\right) = G\left(\mathbf{z}, \mathbf{u}\right) + g_0\left(\mathbf{u}\right)$ where

$$G\left(\mathbf{z}, \mathbf{u}\right) = \sigma_w^2 f\left(\mathbf{z}\right) + \tilde{G}\left(\mathbf{z}, \mathbf{u}\right) \equiv \lambda\sigma_w^2 \left\|\mathbf{z}\right\|^2 + \left\|\left(\sum_{i=1}^{d_0} \mathbf{u}_i \mathbf{A}_i\right)\mathbf{z} - \mathbf{b}\right\|^2,$$

and

$$g_0\left(\mathbf{u}\right) = \frac{\sigma_w^2}{\sigma_e^2}\left\|\mathbf{u}\right\|^2.$$

In the following lemma, which its proof can be found in the Appendix, we see that applying NAM (see exact steps below) on the RSTLS problem formulation generates a globally convergent sequence $\left\{\mathbf{z}^k\right\}_{k\geq 0}$, as follows from Theorem 2 and Proposition 2.

Lemma 6. *Let $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ be a sequence generated by NAM (see Algorithm 4). Then, the sequence is bounded, the function F satisfies Assumption 1 and the partial function $\mathbf{z} \mapsto F\left(\mathbf{z}, \mathbf{u}\right)$ satisfies Assumption 3.*

We point out that in the general NAM framework, the function g_0 is allowed to be nonsmooth, which is not the case in this specific application. In addition, it is easy to verify that in this application the function $\mathbf{u} \mapsto F(\mathbf{z}, \mathbf{u})$, for a fixed $\mathbf{z} \in \mathbb{R}^d$, is strongly convex with a linear gradient function (and hence satisfies Assumption 3). Therefore, if d_0 (the dimension of the block $\mathbf{u}$) is small enough, then the sub-problem of minimizing the function F with respect to $\mathbf{u}$ can be solved exactly, as also done in [8] (see more details below).

5.2 Description of the Algorithms for Solving the RSTLS Problem

As mentioned above, we would like to compare the SPA algorithm of [8], which is an OIA based algorithm, with our NAM method incorporated with inner iterations of AG (in short NAM-AG). The two algorithms, for the RSTLS problem, are described in Algorithms 3 and 4, respectively.

Algorithm 3 SPA Algorithm for RSTLS

```
1: Initialization: $(\mathbf{z}^0, \mathbf{u}^0) \in \mathbb{R}^d \times \mathbb{R}^{d_0}$.
2: Iterative step:
3: for $k \geq 0$ do
4:     Pick $L_k$ (a Lipschitz constant of $\nabla_{\mathbf{z}} \tilde{G}(\mathbf{z}, \mathbf{u}^k)$) and update
           $$\mathbf{z}^{k+1} = \operatorname{prox}_{\frac{\sigma_w^2}{L_k} f}\left(\mathbf{z}^k - \frac{1}{L_k} \nabla_{\mathbf{z}} \tilde{G}\left(\mathbf{z}^k, \mathbf{u}^k\right)\right).$$
5:     Update $\mathbf{u}^{k+1} = \operatorname{argmin}\left\{F\left(\mathbf{z}^{k+1}, \mathbf{u}\right) : \mathbf{u} \in \mathbb{R}^{d_0}\right\}$.
6: end for
```

We see that in step 4 of the SPA algorithm we need to calculate the proximal mapping of F. Since the function f is set to be $\lambda \|\mathbf{x}\|^2$, simple calculations show that an explicit update rule is given by the formula

$$\mathbf{z}^{k+1} = \frac{L_k}{L_k + 2\lambda\sigma_w^2}\left(\mathbf{z}^k - \frac{1}{L_k} \nabla_{\mathbf{z}} \tilde{G}\left(\mathbf{z}^k, \mathbf{u}^k\right)\right). \tag{28}$$

We notice that the update (28) and also the initialization step in NAM-AG (see step 4 of Algorithm 4) depend on L_k. We will describe how to calculate L_k in the next sub-section.

Now, in step 5 of SPA and in step 11 of NAM-AG we need to find an exact minimizer of an optimization problem with respect to the block $\mathbf{u}$. Since the function $\mathbf{u} \mapsto F\left(\mathbf{z}^{k+1}, \mathbf{u}\right)$

Algorithm 4 NAM with AG (NAM-AG) for RSTLS

```
1: **Initialization:** $\left(\mathbf{z}^{-1}, \mathbf{u}^{0}\right) \in \mathbb{R}^{d} \times \mathbb{R}^{d_0}$.
2: **Iterative step:**
3: **for** $k \geq 0$ **do**
4:     Set $\mathbf{w}^{k_0} = \mathbf{z}^{k_0} = \mathbf{z}^{k-1}$, $t_0 = 1$ and pick $L_k$ (a Lipschitz constant of $\nabla_{\mathbf{z}} \tilde{G}\left(\mathbf{z}, \mathbf{u}^{k}\right)$).
5:     **for** $j = 0, 1, 2, \ldots, j_k - 1$ **do**
6:         Update $\mathbf{z}^{k_{j+1}} = \mathbf{w}^{k_j} - \frac{1}{2\lambda\sigma_w^2 + L_k} \nabla_{\mathbf{z}} F\left(\mathbf{w}^{k_j}, \mathbf{u}^{k}\right)$.
7:         Set $t_{j+1} = \frac{1+\sqrt{1+4t_j^2}}{2}$.
8:         Update $\mathbf{w}^{k_{j+1}} = \mathbf{z}^{k_{j+1}} + \frac{t_j - 1}{t_{j+1}}\left(\mathbf{z}^{k_{j+1}} - \mathbf{z}^{k_j}\right)$.
9:     **end for**
10:    Set $\mathbf{z}^{k} = \mathbf{z}^{k_{j_k}}$.
11:    Update $\mathbf{u}^{k+1} = \operatorname{argmin}\left\{F\left(\mathbf{z}^{k}, \mathbf{u}\right) : \mathbf{u} \in \mathbb{R}^{d_0}\right\}$.
12: **end for**
```

is strongly convex, a unique minimizer for the problem can be obtained by the first-order optimality condition, that is $\nabla_{\mathbf{u}} F\left(\mathbf{z}^{k+1}, \mathbf{u}^{k+1}\right) = \mathbf{0}_{d_0}$, and is given by the formula

$$\mathbf{u}^{k+1} = \left(\mathbf{B}\left(\mathbf{z}^{k+1}\right)^{T} \mathbf{B}\left(\mathbf{z}^{k+1}\right) + \frac{\sigma_w^2}{\sigma_e^2} \mathbf{I}_{d_0 \times d_0}\right)^{-1} \mathbf{B}\left(\mathbf{z}^{k+1}\right)^{T} \mathbf{b}, \tag{29}$$

where $\mathbf{B}(\mathbf{z}) = \begin{bmatrix} \mathbf{A}_1\mathbf{z} & \mathbf{A}_2\mathbf{z} & \ldots & \mathbf{A}_{d_0}\mathbf{z} \end{bmatrix}$. A similar update formula, where we substitute $\mathbf{z}^{k+1}$ with $\mathbf{z}^{k}$, can also be derived for step 11 of NAM-AG. Since in practice d_0 is small, the system of linear equations in (29) can be solved efficiently.

5.3 Generating the Data

We follow the paper [8], and utilize the RSTLS formulation of (27) in order to reconstruct a given blurred and noisy image. To this end, we consider the vector $\mathbf{b}$ as a vectorized observed blurred image. We denote by $\mathbf{z}^{\text{real}}$ the real image.

In the experiments below, the vector $\mathbf{b}$ is obtained from the real image $\mathbf{z}^{\text{real}}$ by first scaling the pixels of $\mathbf{z}^{\text{real}}$ to be between 0 and 1. Then, it is transferred through a blurring operator, which is Gaussian blur Point Spread Function (PSF) of size $q \times q$ and standard deviation γ. The PSF in this paper uses periodic boundary conditions (see [21, Chapter 3] for more information regarding PSFs in image processing). After blurring, a different noise drawn from a Gaussian distribution with parameters $(0, \sigma_w)$ is added to each coordinate of the image. We denote by $\mathrm{PSF}^{\mathrm{real}}$ the real blurring operator used to generate the observed blurred and noisy image $\mathbf{b}$. Hence,

$$\mathbf{b} = \mathrm{PSF}^{\mathrm{real}}\mathbf{z}^{\mathrm{real}} + \boldsymbol{\mu}, \quad \boldsymbol{\mu} \sim \mathrm{Gauss}\left(\mathbf{0}_d, \sigma_w\right). \tag{30}$$

In the RSTLS problem we assume that the real blurring operator $\mathrm{PSF}^{\mathrm{real}}$ is not exactly known. Instead, we assume that $\mathrm{PSF}^{\mathrm{real}}$ is of the form

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

for some unknown weight vector $\mathbf{u} \in \mathbb{R}^{d_0}$ and known matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{d_0}$ called structure matrices. In practice, the blurring operator $\mathrm{PSF}^{\mathrm{real}}$ is unknown and only the observed PSF, denoted by $\mathrm{PSF}^{\mathrm{obs}}$, is known. In the experiments below the observed blurring operator $\mathrm{PSF}^{\mathrm{obs}}$ is constructed by summing $\mathrm{PSF}^{\mathrm{real}}$ with the known structure matrices after each of them was subjected to a different uniformly distributed noise from the interval $[0, \sigma_e]$. Thus,

$$\mathrm{PSF}^{\mathrm{obs}} = \mathrm{PSF}^{\mathrm{real}} + \sum_{i=1}^{d_0} \boldsymbol{\eta}_i \mathbf{u}_i^{\mathrm{real}} \mathbf{A}_i, \quad \boldsymbol{\eta} \sim \mathrm{Unifrom}\left[\mathbf{0}_{d_0}, \sigma_e\right], \tag{31}$$

where $\mathrm{PSF}\left(\mathbf{u}^{\mathrm{real}}\right) = \mathrm{PSF}^{\mathrm{real}}$. Using the notations above, the objective of the RSTLS problem is to find the following: (i) a reconstructed image $\mathbf{z}$ from the blurred image $\mathbf{b}$ such that $\mathbf{z} \approx \mathbf{z}^{\mathrm{real}}$, where $\mathbf{z}$ is the output of Algorithm 3 or 4, and (ii) a reconstructed blurring operator $\mathrm{PSF}\left(\mathbf{u}\right)$ of the observed $\mathrm{PSF}^{\mathrm{obs}}$ such that $\mathrm{PSF}\left(\mathbf{u}\right) \approx \mathrm{PSF}^{\mathrm{real}}$, where $\mathbf{u}$ is the output of Algorithm 3 or 4.

In the previous sub-section, we pointed out that both algorithms depend on L_k, the Lipschitz constant of the partial function $\mathbf{z} \mapsto \nabla_{\mathbf{z}} \tilde{G}\left(\mathbf{z}, \mathbf{u}^k\right)$, at each outer iteration $k \geq 0$. We recall that the PSF used in this paper has periodic boundary conditions, which results with a PSF operator that is a Block Circulant with Circulant Blocks (BCCB) matrix (see [21, Chapter 4] for exact definition, examples and relation to PSFs of BCCB matrices in the context of image processing). One property of BCCB matrices is that its eigenvalues can be computed efficiently. Therefore, in this paper we use the tight Lipschitz constant

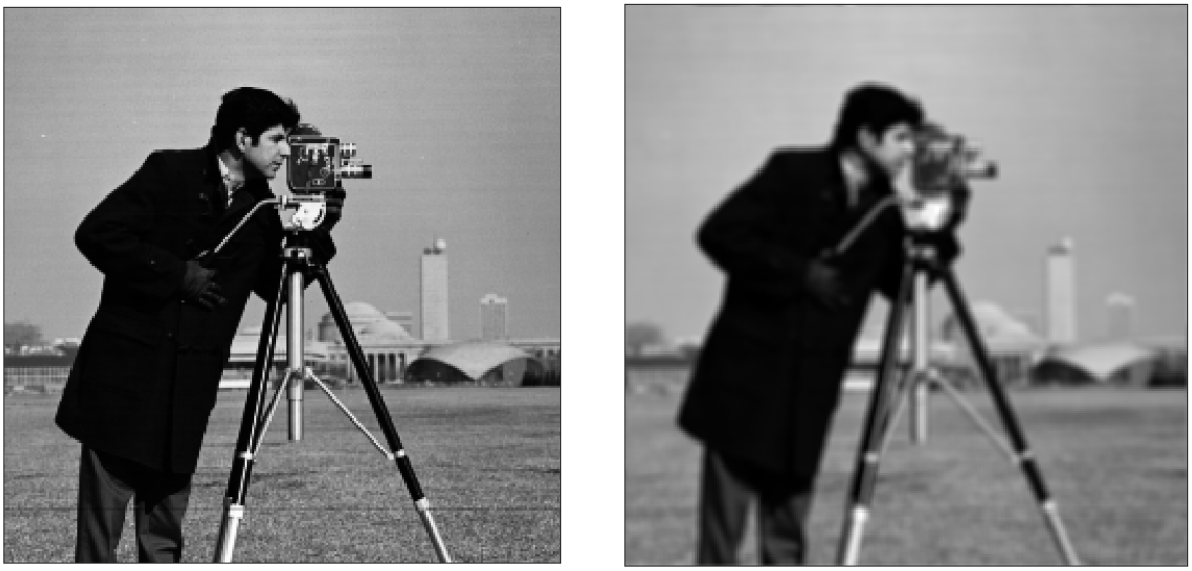
$$L_k = 2\lambda_{\max}\left(\left(\sum_{i=1}^{d_0} \mathbf{u}_i^k \mathbf{A}_i\right)^T \left(\sum_{i=1}^{d_0} \mathbf{u}_i^k \mathbf{A}_i\right)\right). \tag{32}$$

We refer the reader to [21, Section 4.2.1] for information on how to efficiently calculate the above constant for BCCB matrices.

5.4 Description of the Experiments

We ran the SPA and NAM-AG algorithms on a blurred version of the 256 × 256 pixels cameraman test image taken from the MATLAB Image Processing Toolbox available in [1]. Following Sub-section 4.2.1, the number of inner AG iterations is given by the formula $j_k = s + 2^{\lfloor k/r \rfloor} - 1$. In the experiments to follow, we set $r = 10$ and $s \in \{1, 10, 50, 100, 200\}$, which results in five different variants of NAM-AG, each differs by the number of inner AG iterations. In order to blur the original image, we used a Gaussian $\mathrm{PSF}^{\mathrm{real}}$ of size 5×5 with a standard deviation of $\gamma = 2$.

We performed $R = 100$ Monte Carlo trails, where in each trial we drew a different observed image $\mathbf{b}$ and a different $\mathrm{PSF}^{\mathrm{obs}}$, i.e., we drew different $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ according to (30) and (31), respectively. See Figure 1(a) and 1(b) for both the original image and one of the 100 blurred images.



(a) Original image

(b) One of the 100 blurred and noisy images

Figure 1: The cameraman image used in the experiments.

In all experiments we set $\lambda = 0.02$, $\sigma_w = 10^{-4}$ and $\sigma_e = 10^{-3}$. In each Monte Carlo trail, the starting point of all methods is set to be $\mathbf{z}^{-1} = \mathbf{b}$ (the observed blurred and noisy image) and $\mathbf{u}^0 = \mathbf{0}_{d_0}$.

We let SPA and the five different variants of NAM-AG run for $N = 25,000$ total iterations. For all the methods, the total iteration counter also counts inner iterations. Recall that in the inner iterations we only update the $\mathbf{z}$ variable. Therefore, for $N = 25,000$ total iterations of SPA, each variable $\mathbf{z}$ and $\mathbf{u}$ is updated $25,000$ times. This is not the case for the NAM-AG variants, which relatively to OIA methods update less the $\mathbf{u}$ variable for the same number of inner iterations. For example, if $j_0 = 100$ and $j_1 = 200$, then the total iteration counter after the first two outer iterations displays 300 out of N, which means that the $\mathbf{z}$ variable was updated 300 times, while the $\mathbf{u}$ variable was updated only twice.

We compare all of the above methods using three different measures for a total number of iterations of $N = 25,000$:

1. The function value F_j of Problem (27) averaged over all Monte Carlo trials at each iteration $1 \leq j \leq N$, calculated by

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

   where $\mathbf{z}_t^j$ and $\mathbf{u}_t^j$ are the j-th iterate out of N, in the t-th trail out of R. Since $\mathbf{u}$ can be updated less than $\mathbf{z}$, then by $\mathbf{u}_t^j$ we mean the last update of $\mathbf{u}$. Since the optimal function value is bounded from below by 0, then a value of F_j closer to 0 is an indication for each method's performance on solving the Model (27).

2. The deviation of the function value $F\left(\mathbf{z}_t^j, \mathbf{u}_t^j\right)$ from the function value at $(\mathbf{z}_t^{\text{con}}, \mathbf{u}_t^{\text{con}})$ averaged over all Monte Carlo trials at each iteration $1 \leq j \leq N$, calculated by

   $$\text{Dev}_j^{\text{con}} = \frac{1}{R} \sum_{t=1}^{R} \left(F\left(\mathbf{z}_t^j, \mathbf{u}_t^j\right) - F\left(\mathbf{z}_t^{\text{con}}, \mathbf{u}_t^{\text{con}}\right)\right).$$

   Notice that the RSTLS problem is non-convex and there is no guaranty that the real image $\mathbf{z}^{\text{real}}$ is a global solution of the problem. For each Monte Carlo trial and for each of the above methods, the point $\mathbf{z}_t^{\text{con}}$ was obtained by running the corresponding method for $N = 50,000$ total iterations, and then setting the output image as $\mathbf{z}_t^{\text{con}}$

(i.e., $\mathbf{z}_t^{\mathrm{con}} = \mathbf{z}_t^N$). We point out that since all involved methods are globally convergent, then the convergence point is uniquely defined for each method. Since the generated sequence of function values is non-increasing, then $\mathrm{Dev}_j^{\mathrm{con}}$ closer to 0 is an indication on how close the method is to its value at the point $(\mathbf{z}_t^{\mathrm{con}}, \mathbf{u}_t^{\mathrm{con}})$.

3. The average convergence gap over the Monte Carlo trials at each iteration $1 \leq j \leq N$ with respect to the point of convergence $\mathbf{z}_t^{\mathrm{con}}$ of each method in trail t, calculated by

$$\mathrm{ConGap}_j = \frac{1}{R} \sum_{t=1}^{R} \frac{\left\| \mathbf{z}_t^j - \mathbf{z}_t^{\mathrm{con}} \right\|}{\mathbf{z}_t^{\mathrm{con}}}.$$

   We use this measure as an indication on how close the method is to its convergence point $\mathbf{z}_t^{\mathrm{con}}$.

5.5 Results

All experiments were ran on an Intel(R) Core(TM) i7-8565 CPU @ 1.80GHz 1.99GHz with 40.0GB RAM, Windows 10 Pro 64-bit using MATLAB 2020a.

In Figure 2, we compare the average function values at each iteration of SPA and all NAM-AG variants. We see that the OIA method SPA and NAM-AG with $s = 1$ (which is OIA at the first $r = 10$ outer iterations) converge to a point with a higher function value relatively to the nested NAM-AG variants with $s \in \{10, 50, 100, 200\}$. Moreover, while there is almost no decrease in function values of SPA and NAM-AG with $s = 1$ after roughly $5,000$ total iterations, the other nested NAM-AG variants continue to decrease. We therefore see that there is a clear benefit in using nested algorithms compared to OIA algorithms.

In Figure 3, we compare the average deviation of the function value at each iteration from the output's value (obtained after $50,000$ total iterations). It is clear that the nested methods NAM-AG with $s \in \{10, 50, 100, 200\}$ get closer to the value of their output in much less iterations. Combining this result with Figure 2, we conclude that these four methods require less iterations to reach a better solution, relatively to SPA and NAM-AG with $s = 1$.

In Figure 4, we compare the average relative error of the iterates $\mathbf{z}_t^j$ from $\mathbf{z}_t^{\mathrm{con}}$ obtained by running $50,000$ total iterations. We see that SPA and the four nested methods NAM-AG

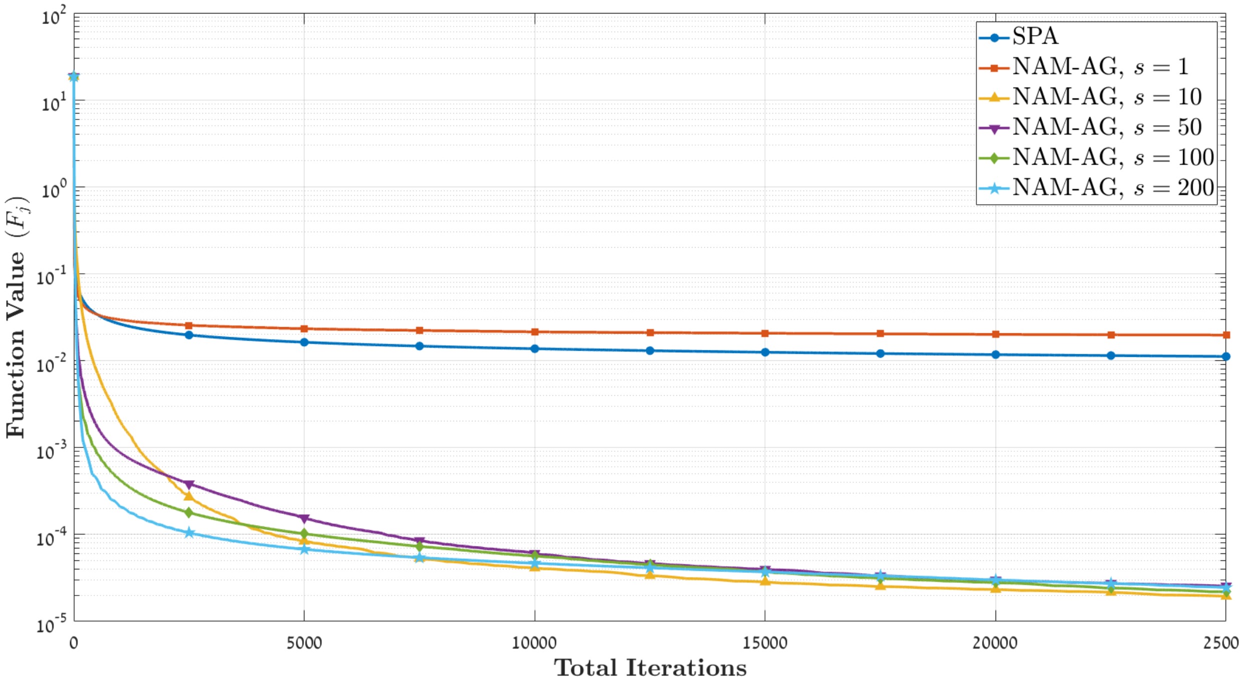


Figure 2: Average function value at each iteration $1 \leq j \leq N$ for $N = 25,000$ over $R = 100$ Monte Carlo trials, in a logarithmic scale. For all NAM-AG variants we set $r = 10$.

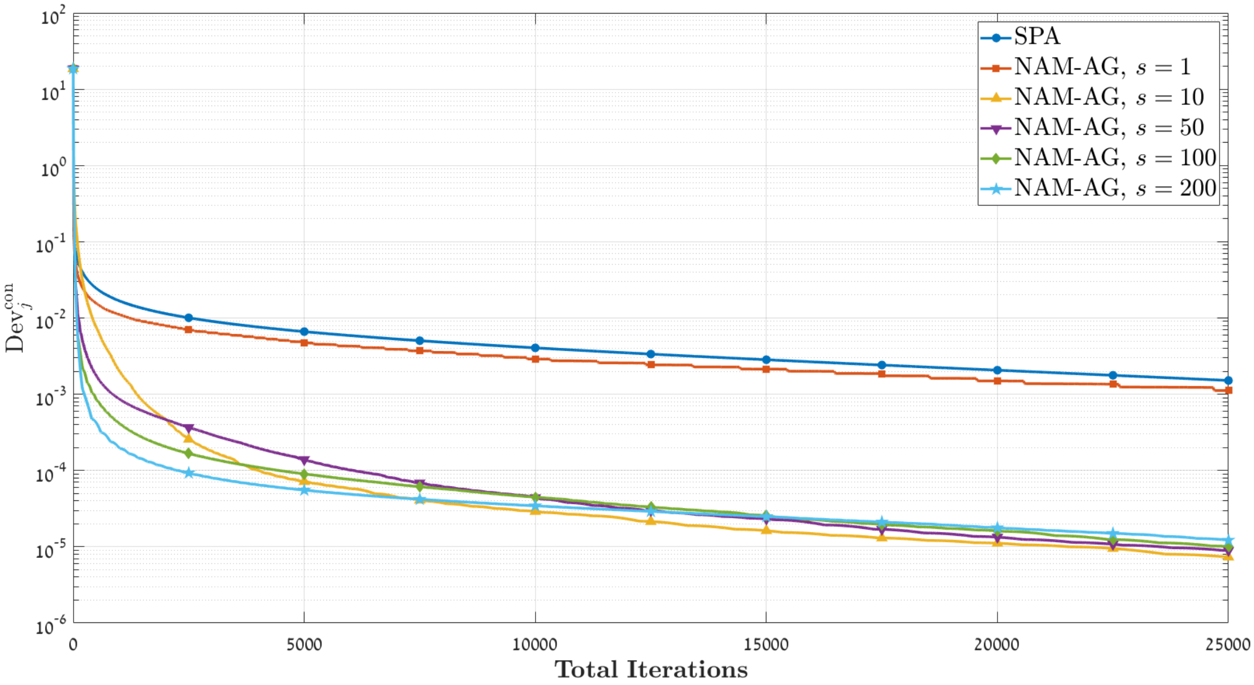


Figure 3: Average deviation of the function value from the value obtained after $50,000$ total iterations, at each iteration $1 \leq j \leq N$ for $N = 25,000$ over $R = 100$ Monte Carlo trials, in a logarithmic scale. For all NAM-AG variants we set $r = 10$.

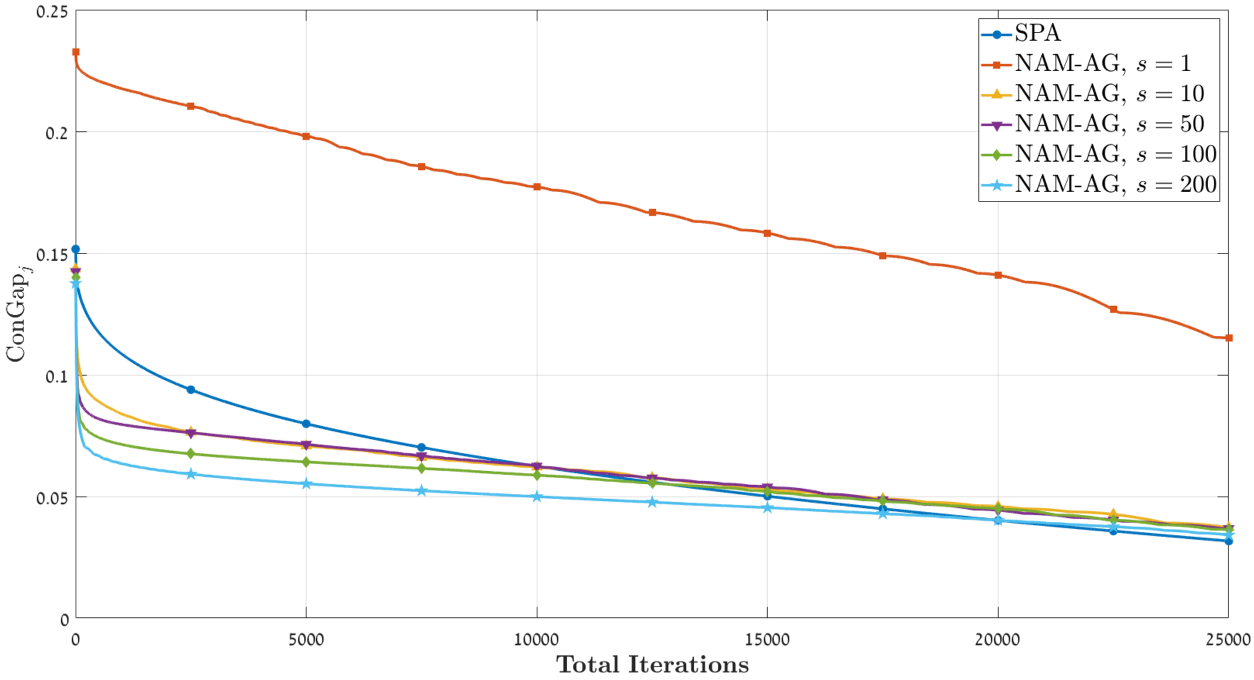


Figure 4: Average relative error from $\mathbf{z}_t^{\mathrm{con}}$ at each iteration $1 \leq j \leq N$ for $N = 25,000$ over $R = 100$ Monte Carlo trials, in a linear scale. For all NAM-AG variants we set $r = 10$.

with $s \in \{10, 50, 100, 200\}$ approach the point of convergence more rapidly. Combining this result with previous results, we derive that these four nested methods reach a good solution at the beginning of the methods' run, while SPA reaches its inaccurate solution with the same amount of iterations. We also see that NAM-AG with $s = 1$ requires more iterations to reach a worse relative error.

Last, in Figure 5, we provide the images obtained by each of the methods after 2, 10, 100, 1000 and 3000 total iterations for one of the Monte Carlo trials. We see that the four nested methods NAM-AG with $s \in \{10, 50, 100, 200\}$ reach a sharper image within this range of inner iterations, while SPA and NAM-AG with $s = 1$ does not.

6 Conclusion

In this paper, we provided a general recipe which serves as a base for proving global convergence of algorithms that tackle non-convex optimization problems. We have generalized the methodology of [11] to allow errors in the conditions. We utilized the new methodology

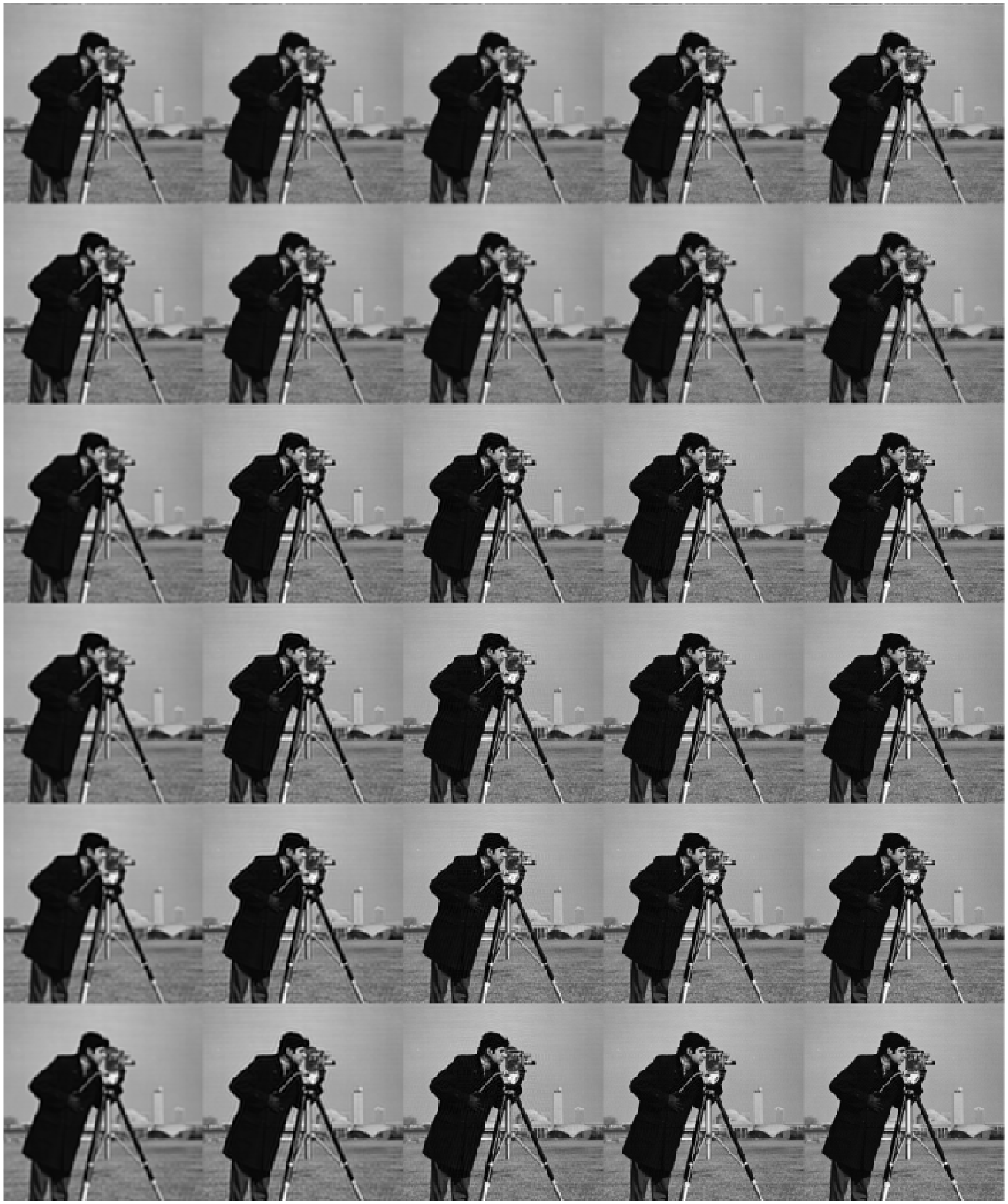


Figure 5: Images obtained by the methods at selected iterations. From top to bottom: SPA and NAM-AG with $s \in \{1, 10, 50, 100, 200\}$. From left to right: iteration 2, 10, 100, 1000 and 3000.

in order to establish global convergence of Nested Alternating Minimization methods that are incorporated with a Nested Friendly Algorithm. In addition, we also showed that it is enough to verify block-wise properties in order to establish global convergence of the whole sequence. To the best of our knowledge, this is the first result that proves global convergence of general nested algorithms, which use non-descent inner algorithms, in the non-convex setting. Moreover, to the best of our knowledge, this is the first methodology which can be used to derive global convergence of methods which are not necessarily sufficient decrease methods.

Appendix

We give here the proof of Lemma 6.

Proof. We established in the proof of Proposition 2 that the sequence $\left\{F\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is non-increasing. It is easy to verify that the objective function $F\left(\mathbf{z}, \mathbf{u}\right)$ in (27) is coercive for $f\left(\mathbf{z}\right) = \lambda \left\|\mathbf{z}\right\|^2$. Therefore, F has bounded level sets, and the sequence $\left\{\left(\mathbf{z}^k, \mathbf{u}^k\right)\right\}_{k\geq 0}$ is bounded since it is contained in the level set of F at level $F\left(\mathbf{z}^{-1}, \mathbf{u}^0\right) \in \mathbb{R}_+$.

Now, as we already mentioned, it is easy to check that the partial function $\mathbf{z} \mapsto F\left(\mathbf{z}, \mathbf{u}\right)$, for F of Problem (27), is strongly convex for any $\mathbf{u} \in \mathbb{R}^{d_0}$. In addition, the Hessian matrix of the partial function $\mathbf{z} \mapsto F\left(\mathbf{z}, \mathbf{u}\right)$ is given by

$$\nabla^2_{\mathbf{z}} F\left(\mathbf{z}, \mathbf{u}\right) = 2\lambda\sigma_w^2 \mathbf{I}_{d\times d} + 2\left(\sum_{i=1}^{d_0} \mathbf{u}_i \mathbf{A}_i\right)^T \left(\sum_{i=1}^{d_0} \mathbf{u}_i \mathbf{A}_i\right).$$

Therefore, we derive that $\bar{\sigma} = 2\lambda\sigma_w^2$ is a positive lower bound on the strong convexity parameter of the partial function $\mathbf{z} \mapsto F\left(\mathbf{z}, \mathbf{u}\right)$ for any $\mathbf{u} \in \mathbb{R}^{d_0}$. Additionally, one can verify that the partial gradient $\nabla_{\mathbf{z}} G\left(\mathbf{z}, \mathbf{u}\right)$ is Lipschitz continuous with a constant

$$L\left(\mathbf{u}\right) = 2\lambda\sigma_w^2 + 2\left\|\sum_{i=1}^{d_0} \mathbf{u}_i \mathbf{A}_i\right\|^2. \tag{33}$$

From the continuity of $L\left(\mathbf{u}\right)$ it follows that there exists some $\bar{L} \geq L\left(\mathbf{u}\right)$ over any compact subset of $\mathbb{R}^{d_0}$. This implies that Assumption 3 holds. Last, since $G\left(\mathbf{z}, \mathbf{u}\right)$ is a quadratic polynomial function we easily derive that Assumption 1 is also satisfied for some $L > 0$. □

References

[1] *Image Processing Toolbox*. The MathWorks Inc. Natick, Massachusetts, United States. Online: mathworks.com/products/image, 2020.

[2] T. J. Abatzoglou, J. M. Mendel, and G. A. Harada. The constrained total least squares technique and its applications to harmonic superresolution. *IEEE Transactions on Signal Processing*, 39(5):1070–1087, 1991.

[3] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

[4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[5] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[6] A. Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

[7] A. Beck and A. Ben-Tal. On the solution of the tikhonov regularization of the total least squares problem. *SIAM Journal on Optimization*, 17(1):98–118, 2006.

[8] A. Beck, S. Sabach, and M. Teboulle. An alternating semiproximal method for non-convex regularized structured total least squares problems. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1129–1150, 2016.

[9] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE transactions on image processing*, 18(11):2419–2434, 2009.

[10] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[11] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[12] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[13] S. Bonettini, M. Prato, and S. Rebegoldi. A block coordinate variable metric linesearch based proximal gradient method. *Computational Optimization and Applications*, 71(1):5–52, 2018.

[14] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward–backward algorithm. *Journal of Global Optimization*, 66(3):457–485, 2016.

[15] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.

[16] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

[17] G. H. Golub and C. F. Van Loan. An analysis of the total least squares problem. *SIAM journal on numerical analysis*, 17(6):883–893, 1980.

[18] B. L. Gorissen, İ. Yanıkoğlu, and D. den Hertog. A practical guide to robust optimization. *Omega*, 53:124–137, 2015.

[19] P. J. Groenen, M. van de Velden, et al. Multidimensional scaling by majorization: A review. *Journal of Statistical Software*, 73(8):1–26, 2016.

[20] W. J. Gutjahr and A. Pichler. Stochastic multi-objective optimization: a survey on non-scalarizing methods. *Annals of Operations Research*, 236(2):475–499, 2016.

[21] P. C. Hansen, J. G. Nagy, and D. P. O'leary. *Deblurring images: matrices, spectra, and filtering*. SIAM, 2006.

[22] R. Hesse, D. R. Luke, S. Sabach, and M. K. Tam. Proximal heterogeneous block implicit-explicit method and application to blind ptychographic diffraction imaging. *SIAM Journal on Imaging Sciences*, 8(1):426–457, 2015.

[23] P. Jain, P. Kar, et al. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[24] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Ann. Inst. Fourier (Grenoble)*, 48(3):769–783, 1998.

[25] S. Łojasiewicz. 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.

[26] F. G. Mohammadi, M. H. Amini, and H. R. Arabnia. Evolutionary computation, optimization, and learning algorithms for data science. In *Optimization, Learning, and Control for Interdependent Complex Networks*, pages 37–65. Springer, 2020.

[27] B. S. Mordukhovich. *Variational analysis and generalized differentiation I: Basic theory*, volume 330. Springer Science & Business Media, 2006.

[28] J.-J. Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[29] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983.

[30] P. Ochs. Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano. *SIAM Journal on Optimization*, 29(1):541–570, 2019.

[31] P. Ochs, Y. Chen, T. Brox, and T. Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[32] N. Piovesan and T. Erseghe. Cooperative localization in wsns: A hybrid convex/nonconvex solution. *IEEE Transactions on Signal and Information Processing over Networks*, 4(1):162–172, 2018.

[33] T. Pock and S. Sabach. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787, 2016.

[34] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[35] S. Sabach, M. Teboulle, and S. Voldman. A smoothing alternating minimization-based algorithm for clustering with sum-min of Euclidean norms. *Pure and Applied Functional Analysis*, 3:653–679, 2018.

[36] F. Wen, L. Chu, P. Liu, and R. C. Qiu. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, 6:69883–69906, 2018.

[37] Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, 2017.