An Alternating Semi-Proximal Method for Nonconvex Regularized Structured Total Least Squares Problems

Amir Beck* Shoham Sabach† Marc Teboulle‡

Abstract

We consider a broad class of regularized structured total-least squares problems (RSTLS) encompassing many scenarios in image processing. This class of problems results in a nonconvex and often nonsmooth model in large dimension. To tackle this difficult class of problems we introduce a novel algorithm which blends proximal and alternating minimization methods by beneficially exploiting data information and structures inherently present in RSTLS. The proposed algorithm, which can also be applied to more general problems, is proven to globally converge to critical points, and is amenable to efficient and simple computational steps. We illustrate our theoretical findings by presenting numerical experiments on deblurring large scale images which demonstrate the viability and effectiveness of the proposed method.

Key words: Alternating minimization, Kurdyka-Łojasiewisz property, global convergence, nonconvex-nonsmooth minimization, proximal gradient methods, regularized structured total least squares, semi-algebraic functions.

1 Introduction

In this paper we consider a broad class of optimization problems for Regularized Structure Total Least Squares (RSTLS) which captures many models arising in applications, and consists of solving the following nonconvex and nonsmooth minimization problem:

$$\text{(RSTLS)} \quad \min_{\mathbf{x},\mathbf{y}} \left\{ F(\mathbf{x}) + \frac{1}{\sigma_w^2} \left\| \left(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right\|^2 + \frac{1}{\sigma_e^2} \|\mathbf{y}\|^2 : \ \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^p \right\}, \tag{1.1}$$

*Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 3200003, Israel. E-mail: becka@ie.technion.ac.il. The research of Amir Beck was partially supported by the Israel Science Foundation under grant ISF No.253/12.

†Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 3200003, Israel. E-mail: ssabach@ie.technion.ac.il.

‡School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel. E-mail: teboulle@post.tau.ac.il. The research of Marc Teboulle partially supported by the Israel Science Foundation, ISF Grant 998/12.

where the model matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and the measurements vector $\mathbf{b} \in \mathbb{R}^m$ are contaminated by noise, σ_e and σ_w are the standard deviations of the corresponding noise components (see Section 5 for more details), and $F : \mathbb{R}^d \to (-\infty, +\infty]$ is a given convex (possibly nonsmooth) function, which allows to capture some desired regularized features of the solution. We are focusing in this paper on the situation in which the model matrix $\mathbf{A}$ admits a linear structure, and therefore it is natural to assume that the noise matrix contaminating the "true" model matrix shares the same structure described by the structure matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p \in \mathbb{R}^{m\times n}$ with $y_1, y_2, \ldots, y_p \in \mathbb{R}$ being the unknown structure components. The proposed model is motivated by our desire to tackle and extend various well known total least squares based models which have been studied in the literature and will be briefly recalled and discussed in the forthcoming section.

A key difficulty in the RSTLS problem is the nonconvexity in the variables $(\mathbf{x}, \mathbf{y})$ due to the coupling term in the squared norm of the objective function in (1.1). Another difficulty is the large scale nature of the problem which naturally arises in many applications of interest, as well as, the possible nonsmoothness of the regularizer $F(\cdot)$. Thus, we face a problem sharing the three most difficult properties an optimization problem can have – nonconvexity, nonsmoothness, and large size – precluding the direct use of any standard optimization schemes to its solution. However, the RSTLS problem also shares some particular structures and data information, such as convexity in separate arguments $(\mathbf{x}, \mathbf{y})$, and smoothness of the least-squares term that can be beneficially exploited.

In the present work we will strongly exploit the alluded structures and properties, and our main objective is to devise a simple and efficient algorithm proven to globally converge to a critical point of the nonconvex objective function of (1.1) and capable of handling large scale instances. To achieve this goal we blend alternating minimization and proximal methods. These two very well known paradigms have recently attracted intensive research activities in many disparate applications due to their simplicity and remarkable theoretical and practical performance, mainly in the convex setting, see *e.*g., [3, 6, 13, 37] and references therein for a small representative sample of these activities. However here, the RSTLS problem under consideration is *nonconvex* and nonsmooth. Motivated by the recent algorithmic and convergence analysis framework developed in Bolte *et al.* [12] which builds on the powerful Kurdyka-Łojasiewicz property [19, 21] to handle genuine nonconvex and nonsmooth minimization problems, we address the inherent nonconvex difficulty present in the RSTLS problem by further exploiting the problem's data information. This leads us to introduce a novel algorithm which relies on alternating minimization and on *semi-proximal* regularization, which for ease of reference is called SPA. While the focus of this paper is on the RSTLS problems, our analysis is developed for broader class of problems which captures the class of RSTLS problems and the corresponding algorithm as a particular case. Thus, our results can also be applied to other applications and contexts sharing this proposed broader formulation. For the RSTLS problem, the resulting algorithm involves two simple computational steps. One, asking for the solution of a small scale $(p \times p)$ linear system, while the other step, depending on the choice of the regularizer $F(\cdot)$, either admits a closed

form solution, or can be efficiently computed via a fast dual proximal method [7]. We prove that the proposed algorithm SPA globally converges to a critical point of the problem at hand. Finally, we illustrate our theoretical findings by presenting some numerical examples on image deblurring which demonstrate the effectiveness of the proposed method.

Outline of the paper. The paper is organized as follows. In the next section we describe various total least squares models which have motivated the proposed RSTLS, and make our setting more precise. While the focus of this paper is on the RSTLS problem, our analysis is developed for broader class of problems which captures the class of RSTLS problems and the corresponding algorithm as a particular case; this is developed in Section 3. In Section 4, we present a general analysis framework and we state and prove the main convergence results. Finally, in Section 5, numerical results on image deblurring problems illustrate our theoretical findings. To make the paper self-contained, a short appendix includes some relevant technical additional material. Our notations are quite standards, and will be defined throughout the text if and when necessary.

2 Total Least Squares: Approaches and Models

2.1 Motivation

Many data fitting and linear inverse problems arising in a wide variety of applications lead to study overdetermined linear systems of the form

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}, \tag{2.1}$$

where both the model matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and the measurement vector $\mathbf{b} \in \mathbb{R}^m$ are contaminated by noise, leading to the model

$$(\mathbf{A} + \mathbf{E})\,\mathbf{x} = \mathbf{b} + \mathbf{w}, \tag{2.2}$$

where $\mathbf{E} \in \mathbb{R}^{m\times n}$ and $\mathbf{w} \in \mathbb{R}^m$ are unknown perturbation matrix and vector to the model matrix and to the right-hand side vector, respectively. A well known and classical approach to this problem is the total least squares (TLS) which seeks to find a triple $(\mathbf{x}, \mathbf{E}, \mathbf{w})$ that minimizes $\|\mathbf{E}\|_F^2 + \|\mathbf{w}\|^2$ subject to the consistency equation (2.2); that is, to solve the following nonconvex minimization problem

$$\text{(TLS)} \qquad \min_{\mathbf{x},\mathbf{w},\mathbf{E}} \left\{ \|\mathbf{E}\|_F^2 + \|\mathbf{w}\|^2 : \, (\mathbf{A} + \mathbf{E})\,\mathbf{x} = \mathbf{b} + \mathbf{w} \right\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The popularity of TLS mainly stems from the fact that although nonconvex, this problem admits an explicit solution expressed by the singular value decompostion of the augmented matrix $(\mathbf{A}, \mathbf{b})$, see [16, 38].

However, it also well known that in the cases where the model matrix $\mathbf{A}$ has a special structure, TLS methods may not always be appropriate, since they do not take into

consideration the given structure. It is thus desirable to exploit the special structure in order to reduce the number of unknown parameters and to improve the performance. In the context of image processing, which is one of the main application of interest in this work, it is natural and adequate to consider the same structure on the pertubation matrix $\mathbf{E}$. Thus, for a model matrix $\mathbf{A}$ admitting a linear structure, which as just mentioned is a common situation in ill-posed linear inverse problems, in this paper we assume that the perturbation matrix $\mathbf{E}$ shares the same structure, and is defined by

$$\mathbf{E} = \sum_{i=1}^{p} y_i \mathbf{A}_i,$$

where $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p \in \mathbb{R}^{m\times n}$ are the structure matrices and $y_1, y_2, \ldots, y_p \in \mathbb{R}$ are the unknown structure components. Then, the Structured Total Least Squares method (STLS) aims at solving the following minimization problem [1, 5, 20, 23, 25, 26, 34]:

$$\text{(STLS)} \qquad \min_{\mathbf{x},\mathbf{y}} \left\{ \frac{1}{\sigma_w^2} \left\| \left(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right\|^2 + \frac{1}{\sigma_e^2} \|\mathbf{y}\|^2 \right\}, \tag{2.3}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_p)^T$.

The STLS problem is nonconvex, and in contrast to the unstructured TLS problem, (which can easily be seen as a special case of (2.3)), finding its global solution is difficult. Note also that by minimizing with respect to $\mathbf{y}$, we can reformulate the (2.3) as a minimization problem with respect to the $\mathbf{x}$ variables only via,

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ (\mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{C}(\mathbf{x}) (\mathbf{A}\mathbf{x} - \mathbf{b}) \right\}, \text{ with } \mathbf{C}(\mathbf{x}) := \left(\sigma_w^2 \mathbf{I} + \sigma_e^2 \sum_{i=1}^{p} \mathbf{A}_i \mathbf{x}\mathbf{x}^T \mathbf{A}_i^T \right)^{-1},$$

which however remains highly nonconvex. Several algorithms have been proposed to find a stationary point of this formulation of the STLS problem (see, for instance, [22, 24, 34].

In [34], the authors developed the Structured Total Least Norm (STLN) method to solve such class of problems, which turns out to be useful for many structured linear problems. However, for many problems, such as, image deblurring, the matrix is often ill-conditoned and applying STLN results in poor images recovery. Thus regularization is needed. Such an approach was considered for instance in [29], who implemented Tihkhonov regularization [35] to derive the regularized structured total least norm (RSTLN), with an algorithm preserving linear structure of the matrix and which minimizes the l_p-norm error, with $p = 1, 2$ or ∞. For another work which focused on quadratic regularizers and improves the computational efficiency of the algorithm [29], see for instance [23]. RSTLN based methods rely on a variation of the Gauss-Newton method[1] and one of its main drawback is that it

[1]When the regularizer is quadratic, *e.g.*, $p = 2$, Gauss-Newton is directly applicable, while for $p = 1, \infty$, the objective function is not anymore differentiable, but can similarly be applied, see [29] for details.

requires solving a least-squares problem at each iteration (see Section 5.1 for more details on the method), which in many scenarios becomes too difficult or impossible for large scale models. In addition, the objective function to be minimized is nonconvex, so there is no guarantee that the algorithm converges to global minimum.

Linear inverse problems where the model matrix $\mathbf{A}$ is ill-conditioned, is a common phenomena which directly implies that either solutions obtained by the LS, TLS or STLS methods have a huge norm, resulting with meaningless solutions. This was discussed above within the RSTLN approach. Alternatively, this difficulty is often handle for example by adding a penalty term which plays the role of regularization and is somehow able to control the size of the solution. A popular one is the Tikhonov regularization method [36], but other choices are also possible depending on the application at hand. In this case, the resulting model is either called the regularized TLS (RTLS) or regularized structured TLS (RSTLS), depending on whether the model is structured or not, see [4, 15, 17] and the references therein for several regularization ideas.

Motivated by the above, in this paper we focus on a broad class of regularized structured total least squares (RSTLS) problems which captures the ideas and models just described, and allows for unification and extension, by considering the following nonconvex and nonsmooth regularized structured total least squares:

$$\text{(RSTLS)} \quad \min_{\mathbf{x},\mathbf{y}} \left\{ F(\mathbf{x}) + \frac{1}{\sigma_w^2} \left\| \left(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right\|^2 + \frac{1}{\sigma_e^2} \|\mathbf{y}\|^2 : \ (\mathbf{x},\mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^p \right\}, \tag{2.4}$$

where $F : \mathbb{R}^d \to (-\infty, +\infty]$ is a proper, lsc and convex function (possibly nonsmooth), and which as illustrated next, allows to encompass various scenarios.

2.2 Some Models Examples

Now we describe some interesting special cases of the RSTLS problem that can be obtained by specific choices of the regularizing function $F(\cdot)$.

Example 2.1. (The Constrained RSTLS problem). Let the nonsmooth function F to be the indicator function of the convex set

$$C_q := \left\{ \mathbf{x} \in \mathbb{R}^n : \ \|\mathbf{L}\mathbf{x}\|_q \leq \delta \right\},$$

where $\mathbf{L} \in \mathbb{R}^{d \times n}$, $\delta > 0$ and $q = 1, 2, \infty$. In this case, we obtain from (1.1) the constrained RSTLS model given by

$$\text{(RSTLS-C)} \quad \min_{\mathbf{x},\mathbf{y}} \left\{ \frac{1}{\sigma_w^2} \left\| \left(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right\|^2 + \frac{1}{\sigma_e^2} \|\mathbf{y}\|^2 : \ \|\mathbf{L}\mathbf{x}\|_q \leq \delta \right\}. \tag{2.5}$$

Example 2.2. (Penalized RSTLS). As an alternative to the constrained RSTLS, we can penalize the constraint with $F(\mathbf{x}) := \lambda f(\mathbf{x})$. The penalized RSTLS model is then given by

$$\text{(RSTLS-P)} \quad \min_{\mathbf{x},\mathbf{y}} \left\{ \lambda f(\mathbf{L}\mathbf{x}) + \frac{1}{\sigma_w^2} \left\| \left(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right\|^2 + \frac{1}{\sigma_e^2} \|\mathbf{y}\|^2 \right\}, \tag{2.6}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function (usually a norm), and $\lambda > 0$ is a penalty (regularization) parameter which measures the trade-off between error measurements and constraint satisfaction. One of the popular choices for the penalty function is known as the Tikhonov regularization [36] in which $f(\cdot) = \|\cdot\|^2$. Another penalty function that has attracted a revived interest and considerable amount of attention in the signal processing literature is the use of the ℓ_1 norm which is discussed in more details in the next example.

Example 2.3. (Total Variation Deblurring). This example will serve as a basis of our numerical experiments given in Section 5. We concentrate on RSTLS with a total variation penalized function $f(\mathbf{x}) := TV_1(\mathbf{x})$. More precisely, TV_1 stands for the ℓ_1-based anisotropic TV, given by

$$\begin{aligned} TV_1(\mathbf{x}) = &\sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \left(|x_{i,j} - x_{i+1,j}| + |x_{i,j} - x_{i,j+1}| \right) + \sum_{i=1}^{m-1} |x_{i,n} - x_{i+1,n}| \\ &+ \sum_{j=1}^{n-1} |x_{m,j} - x_{m,j+1}| . \end{aligned} \tag{2.7}$$

It is well-known that $TV_1(\mathbf{x}) = \left\| \mathcal{L}^T \mathbf{x} \right\|_1$ where $\mathcal{L} : \mathbb{R}^{(m-1)\times n} \times \mathbb{R}^{m\times(n-1)} \to \mathbb{R}^{m\times n}$ is defined by the formula

$$\mathcal{L}(\mathbf{C}, \mathbf{D})_{i,j} = c_{i,j} + d_{i,j} - c_{i-1,j} - d_{i,j-1}, i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n,$$

where we assume that $c_{0,j} = c_{m,j} = d_{i,0} = d_{i,n} = 0$ for every $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. This means that in this case $\mathbf{L} = \mathcal{L}^T$ and the operator $\mathcal{L}^T : \mathbb{R}^{m\times n} \to \mathbb{R}^{(m-1)\times n} \times \mathbb{R}^{m\times(n-1)}$ which is the adjoint to $\mathcal{L}$ is given by

$$\mathcal{L}^T(\mathbf{x}) = (\mathbf{C}, \mathbf{D}),$$

where $\mathbf{C} \in \mathbb{R}^{(m-1)\times n}$ and $\mathbf{D} \in \mathbb{R}^{m\times(n-1)}$ are the matrices defined by

$$\begin{aligned} c_{i,j} &= x_{i,j} - x_{i+1,j}, i = 1, 2, \ldots, m-1 \text{ and } j = 1, 2, \ldots, n, \\ d_{i,j} &= x_{i,j} - x_{i,j+1}, i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n-1. \end{aligned}$$

3 Problem Formulation and The Main Algorithm

Although the emphasis of this paper is on solving the class of RSTLS problems, for the purpose of the forthcoming development and analysis we will consider a more general class of nonconvex and nonsmooth problems which captures the RSTLS as a special case, and which we hope could also be beneficially applied in other contexts.

3.1 The Problem Model

Consider the following nonconvex and nonsmooth optimization model

$$(M) \qquad \min_{\mathbf{x},\mathbf{y}} \{\Psi(\mathbf{x},\mathbf{y}) := H(\mathbf{x},\mathbf{y}) + F(\mathbf{x}) + G(\mathbf{y})\}, \tag{3.1}$$

with the following assumption.

Assumption 1. (i) $F : \mathbb{R}^n \to (-\infty, +\infty]$ is a proper, closed and convex function.

(ii) $G : \mathbb{R}^p \to (-\infty, +\infty]$ is a proper, closed and strongly convex function with strong convexity constant σ.

(iii) $H : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is a C^1 function and for any fixed $\mathbf{x}$, the function $\mathbf{y} \to H(\mathbf{x},\mathbf{y})$ is convex.

The general optimization model (M) is naturally motivated by the RSTLS problem which is clearly a special case of problem (M) and satisfies Assumption 1 with:

$$H(\mathbf{x},\mathbf{y}) = \frac{1}{\sigma_w^2}\left\|\left(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{A}_i\right)\mathbf{x} - \mathbf{b}\right\|^2 \quad \text{and} \quad G(\mathbf{y}) = \frac{1}{\sigma_e^2}\|\mathbf{y}\|^2, \tag{3.2}$$

and the choice of any proper, lsc and convex function F. Indeed, G is strongly convex with $\sigma = 2/\sigma_e^2$; the above H is clearly C^1 on $\mathbb{R}^n \times \mathbb{R}^p$, and for any fixed $\mathbf{x} \in \mathbb{R}^n$, the function $\mathbf{y} \to H(\mathbf{x},\mathbf{y})$ is convex on $\mathbb{R}^p$.

Note that the first item in Assumption 1 allows to consider various interesting scenarios (see previous examples), while the second item naturally generalizes the quadratic function $\|\cdot\|^2$, through strong convexity (note that smoothness is not needed). The third item which asks for the smoothness of $H(\cdot,\cdot)$ and partial convexity with respect to second variable will be beneficially exploited to build the algorithm proposed below, where each step involves the solution of a convex minimization problem in each block $(\mathbf{x},\mathbf{y})$.

3.2 The Algorithm SPA

The block structure of problem (M) naturally suggested to apply the Alternating Minimization (AM) (also known as Gauss-Seidel [8]) method. That is, starting with any given

initial point $(\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^n \times \mathbb{R}^p$, we generate a sequence $\left\{\left(\mathbf{x}^k, \mathbf{y}^k\right)\right\}_{k\in\mathbb{N}}$ via the following two steps: for $k \in \mathbb{N}$

$$\mathbf{x}^{k+1} \in \operatorname{argmin}_{\mathbf{x}} \Psi\left(\mathbf{x}, \mathbf{y}^{k+1}\right), \tag{3.3}$$

$$\mathbf{y}^{k+1} = \operatorname{argmin}_{\mathbf{y}} \Psi\left(\mathbf{x}^{k+1}, \mathbf{y}\right), \tag{3.4}$$

where the second step produces a *unique* minimizer, thanks to the strong convexity of $\mathbf{y} \to \Psi\left(\mathbf{x}^{k+1}, \mathbf{y}\right)$, see Assumption 1(ii).

The algorithm (AM) is quite attractive, as it permits a natural decoupling of the function H and amounts to solving two simpler problems at each iteration. The AM algorithm and many of its variants has been studied in several works, both in the convex and nonconvex setting, and we refer the reader to [2, 12] and references therein for details on the origin, advantages, difficulties, and convergence results of several (AM)-based schemes.

Both subproblems in the AM algorithm above are convex, but usually cannot be solved exactly. Very recently, [12] suggests to overcome this difficulty by solving *approximately* both iterations of AM. More precisely, since each iteration of AM consists of minimizing the sum of a smooth function with a nonsmooth one, they apply *one step* of the so-called proximal gradient method (see, *e*.g., [6]), resulting in the Proximal Alternating Linearization Minimization (PALM) algorithm [12], which actually can solve a broader class of nonsmooth and nonconvex problems (*i*.e., where F and G also nonconvex), and was proven to globally converge to a critical point of Ψ under suitable assumptions; see more in the next section. At this juncture, it should be noted that very recently AM and PALM based methods have been applied in various important applications. For instance, in [9], the authors use the regularized version of the AM algorithm for solving blind image recovery problems. A variable metric version of PALM can be found in [14]. Other variants of PALM have also been very recently proposed and used in [31] to solve Ptychographic diffraction imaging problems, and in [30] for another variant which was used in the context of sparse blind deconvoluion.

Here, we follow the approach of [12], by further exploiting the data information in problem (M). More specifically, since for the model (M) the $\mathbf{y}$-step consists of solving a *strongly convex* minimization problem, we keep it intact. As we shall see below, for the RSTLS problem this step can be written explicitly as the solution of a linear system which can be efficiently approximated. This motivates the proposed algorithm which can be seen as a variant of PALM, whereby only the $\mathbf{x}$-step (3.3) is solved approximately through one shot of a proximal gradient step, *i*.e.,

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}} \left\{ F(\mathbf{x}) + \left\langle \mathbf{x} - \mathbf{x}^k, \nabla_{\mathbf{x}} H\left(\mathbf{x}, \mathbf{y}^k\right)\right\rangle + \frac{t_k}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|^2 \right\}, \ (t_k > 0),$$

while the $\mathbf{y}$-step is solved exactly through the global optimality condition for (3.4). For ease of reference, we call the resulting algorithm: Semi-Proximal-Alternating (SPA).

Before presenting the algorithm, it will be convenient to recall the definition of the Moreau proximal map [27] to re-write compactly the above step. Given a proper, closed and convex function $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$, the *proximal map* associated with φ is uniquely defined by

$$\text{prox}_{\varphi}(\mathbf{x}) := \text{argmin}\left\{\varphi(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2 : \ \mathbf{u} \in \mathbb{R}^m\right\}. \tag{3.5}$$

For a recent and comprehensive review on proximal methods, we refer the reader to [28].

To solve (M) we thus propose the following scheme (the definition of $L_1(\mathbf{y})$ is given in Assumption 2 that follows the description of the method).

Algorithm SPA

(1) Initialization: start with any $(\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^n \times \mathbb{R}^p$.

(2) For each $k = 1, 2, \ldots$ generate a sequence $\left\{\left(\mathbf{x}^k, \mathbf{y}^k\right)\right\}_{k\in\mathbb{N}}$ as follows:

(2.1) Set $L_k := L_1\left(\mathbf{y}^k\right)$ and compute

$$\mathbf{x}^{k+1} = \text{prox}_{\frac{1}{L_k}F}\left(\mathbf{x}^k - \frac{1}{L_k}\nabla_{\mathbf{x}} H\left(\mathbf{x}^k, \mathbf{y}^k\right)\right). \tag{3.6}$$

(2.2) Solve the strongly convex minimization problem:

$$\mathbf{y}^{k+1} = \text{argmin}_{\mathbf{y}\in\mathbb{R}^p}\left\{G(\mathbf{y}) + H\left(\mathbf{x}^{k+1}, \mathbf{y}\right)\right\}. \tag{3.7}$$

To analyze and derive the convergence of SPA, we need the following minimal assumptions.

Assumption 2. (i) For any fixed $\mathbf{y}$, the function $\mathbf{x} \to H(\mathbf{x}, \mathbf{y})$ is $C^{1,1}_{L_1(\mathbf{y})}$, namely, the partial gradient $\nabla_{\mathbf{x}} H(\cdot, \mathbf{y})$ is globally Lipschitz with constant $L_1(\mathbf{y})$, that is,

$$\|\nabla_{\mathbf{x}} H(\mathbf{x}_1, \mathbf{y}) - \nabla_{\mathbf{x}} H(\mathbf{x}_2, \mathbf{y})\| \le L_1(\mathbf{y})\|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n.$$

In addition, there exist two positive numbers $\underline{L}$ and $\overline{L}$ such that $\inf_{\mathbf{y}\in B} L_1(\mathbf{y}) = \overline{L} > 0$ and $\sup_{\mathbf{y}\in B} L_1(\mathbf{y}) = \overline{L} < \infty$ for any compact set $B \subseteq \mathbb{R}^m$.

(ii) The gradient ∇H is Lipschitz continuous on bounded subsets of $\mathbb{R}^n \times \mathbb{R}^p$.

(iii) $\inf_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^n\times\mathbb{R}^p} \Psi(\mathbf{x}, \mathbf{y}) = \underline{\Psi} > -\infty$.

Remark 3.1. As we shall see below, $L_1(\mathbf{y})$ is explicitly available for the RSTLS problem, and we can set the parameter $L_k = L_1\left(\mathbf{y}^k\right)$. However, note that if this is unknown, or still too difficult to compute, then a backtracking scheme [6] can be incorporated and the convergence results developed below remain true, for simplicity of exposition we omit the details.

3.3 Applying SPA: An Algorithm for RSTLS

As we shall see now, the particular realization of SPA when applied to RSTLS yields an attractive scheme for solving a broad class of RSTLS problems. Before doing so, let us first verify that Assumption 2 also holds for RSTLS, which corresponds to the choice of H and G as given in (3.2). A simple computation then shows that

$$\nabla_{\mathbf{x}} H(\mathbf{x},\mathbf{y}) = \frac{2}{\sigma_w^2}\left(\mathbf{A}+\sum_{i=1}^{p} y_i\mathbf{A}_i\right)^T\left(\left(\mathbf{A}+\sum_{i=1}^{p} y_i\mathbf{A}_i\right)\mathbf{x}-\mathbf{b}\right), \tag{3.8}$$

$$\nabla_{\mathbf{y}} H(\mathbf{x},\mathbf{y}) = 2\left(\frac{1}{\sigma_e^2}\mathbf{I}_{p\times p}+\frac{1}{\sigma_w^2}\mathbf{B}(\mathbf{x})^T\mathbf{B}(\mathbf{x})\right)\mathbf{y}+\frac{1}{\sigma_w^2}\mathbf{B}(\mathbf{x})^T\left(\mathbf{A}\mathbf{x}-\mathbf{b}\right), \tag{3.9}$$

where I_p stands for the usual $p\times p$ identity matrix, and we define

$$\mathbf{B}(\mathbf{x}) := \left(\mathbf{A}_1\mathbf{x},\mathbf{A}_2\mathbf{x},\ldots,\mathbf{A}_p\mathbf{x}\right). \tag{3.10}$$

Now, it is easy to see that Assumption 2(ii) and (iii) hold for the RSTLS problem. For the former, this is immediate with H defined in (3.2), while the latter follows since Ψ is nonnegative in most applications of interest, *i*.e., with $F(\cdot)$ being chosen as a norm. Now, for any fixed $\mathbf{y}\in\mathbb{R}^p$, it is easy to see from (3.8) that the function $H(\cdot,\mathbf{y})$ admits a Lipschitz continuous partial gradient $\nabla_{\mathbf{x}} H(\cdot,\mathbf{y})$ with constant $L_1(\mathbf{y}) = (2/\sigma_w^2)\left\|\mathbf{A}+\sum_{i=1}^{p} y_i\mathbf{A}_i\right\|^2$. Moreover, it follows by the continuity of $L_1(\cdot)$ that the moduli $L_1(\mathbf{y})$ is bounded from below and from above over compact sets, and hence Assumption 2(i) holds.

Remark 3.2. It should be noted that the computation of $L_1(\mathbf{y})$ requires to find the largest singular value of the corresponding matrix. For large scale problems, such as those arising in image applications, this could be quite involved, but in several scenarios it can handled efficiently. For example, the operator represents a blurring operation with a PSF under reflexive or periodic boundary conditions, the operator norm can be efficiently computed by directly computing the eigenvalues of the matrix by the techniques described, for example, in [18]. In other cases, as mentioned in Remark 3.1, a version of the algorithm with a backtracking scheme can be used in order to refrain from the computation of the exact Lipschitz constant.

When applied to the RSTLS problem, namely with H and G defined in (3.2), the SPA reduces to the following:

Algorithm for RSTLS. Start with any $(\mathbf{x}^0,\mathbf{y}^0)\in\mathbb{R}^n\times\mathbb{R}^p$, and for each $k\geq 1$ generate the sequence $\left\{\left(\mathbf{x}^k,\mathbf{y}^k\right)\right\}_{k\in\mathbb{N}}$ via

(a) Compute $\mathbf{x}^{k+1}$ via (3.6), with H defined in (3.2) and $L_k = (2/\sigma_w^2)\left\|\mathbf{A}+\sum_{i=1}^{p} y_i^k\mathbf{A}_i\right\|^2$.

(b) Compute $\mathbf{y}^{k+1}$ via (3.11) below.

The first step of the algorithm reduces to compute the proximal map of the given convex function F. Depending on the choice of F the proximal map can be computed explicitly or via an efficient algorithm. For the case of interest in this paper, namely with F being the total variation function, we will use the recent efficient scheme FDPG of [7], see more details in the numerical section below.

Now, let us derive the second step (b). Writing the optimality condition for the second step of (3.7) consists of finding the unique $\mathbf{y}^{k+1}$ which solves the equation $\nabla_{\mathbf{y}} H\left(\mathbf{x}^{k+1}, y^{k+1}\right) + (1/\sigma_e^2)\,\mathbf{y}^{k+1} = 0$. Using (3.9), and denoting $\mathbf{B}_k := \mathbf{B}(\mathbf{x}^k)$, this reduces to solving a linear system of equations of dimension $p \times p$ and given explicitly by

$$\mathbf{y}^{k+1} = -\sigma_e^2 \left(\sigma_w^2 \mathbf{I}_{p\times p} + \sigma_e^2 \mathbf{B}_{k+1}^T \mathbf{B}_{k+1}\right)^{-1} \mathbf{B}_{k+1}^T \left(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\right). \tag{3.11}$$

Since p is in many scenarios small, and clearly much smaller than mn, the solution of this linear system can be computed very efficiently, see Section 5.

In the next section we develop the theoretical framework and prove the promised convergence results.

4 The Analysis Framework and Convergence Results

The main goal of this section is to derive the convergence properties of SPA and hence of its special case, the algorithm for the RSTLS problem.

4.1 Methodology: An Abstract Convergence Result

To establish the main convergence result of SPA, we follow the scope of a general convergence mechanism first described in [2] and more recently in [12]. Here we use the methodology of [12], whereby a systematic and simple procedure was developed and which essentially can be applied to any given algorithm. We summarize here the key elements and main results of the approach [12] under an abstract convergence result which will then be applied to prove the convergence of SPA.

Let $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ be a sequence in $\mathbb{R}^d$ with starting point $\mathbf{z}^0$. The set of all limit points is denoted by $\omega\left(\mathbf{z}^0\right)$, and defined by

$$\left\{\overline{\mathbf{z}} \in \mathbb{R}^d : \ \exists \text{ an increasing sequence of integers } \left\{k_l\right\}_{l\in\mathbb{N}} \text{ such that } \ \mathbf{z}^{k_l} \to \overline{\mathbf{z}} \text{ as } l \to \infty\right\}.$$

For $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ proper and lsc, $\operatorname{crit}\varphi := \left\{\mathbf{x} \in \mathbb{R}^d : \ \mathbf{0} \in \partial\varphi\left(\mathbf{x}\right)\right\}$ denotes the set of critical points of φ, where $\partial\varphi$ stands for the subdifferential of φ (see Appendix A).

The following definition will be useful to capture two key ingredients of the forthcoming methodology.

Definition 4.1. Let $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lsc function. A sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ is called *a gradient-like descent sequence* for φ if for $k \in \mathbb{N}$ the following two conditions hold:

(C1) *Sufficient decrease property:* There exists a positive scalar ρ_1 such that

$$\rho_1 \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2 \leq \varphi\left(\mathbf{z}^k\right) - \varphi\left(\mathbf{z}^{k+1}\right).$$

(C2) *A subgradient*[2] *lower bound for the iterates gap:*

- $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ is bounded.
- There exists a positive scalar ρ_2 such that

$$\left\|\mathbf{w}^{k+1}\right\| \leq \rho_2 \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|, \ \mathbf{w}^{k+1} \in \partial\varphi\left(\mathbf{z}^{k+1}\right)$$

Remark 4.1. We note that the assumption that $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ is bounded can also be relaxed, and condition (C2) can be replaced by: For any compact subset $Q \subset \mathbb{R}^d$, there exists $\rho_2 > 0$, (possibly depending on Q) such that:

$$\operatorname{dist}\left(\mathbf{0}, \partial\varphi\left(\mathbf{z}^{k+1}\right)\right) \leq \rho_2 \left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|, \ \forall\, \mathbf{z}^k, \mathbf{z}^{k+1} \in Q.$$

The two conditions (C1) and (C2) defining a gradient-like descent sequence for a given φ are typical for any descent type algorithms (see, *e*.g., [2]), and provide the basic tools to prove that the limit of any convergent subsequence of $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ is a critical point of φ. More precisely, from [12, Lemma 5 and Remark 5] we have

Lemma 4.1. *If $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ is a gradient-like descent sequence for a given function φ, which is lsc and proper on $\mathbb{R}^d$, then $\omega\left(\mathbf{z}^0\right)$ is a nonempty, compact and connected set, and we have*

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

This result can thus be applied to any algorithm that produces a gradient-like descent to establish convergence in accumulation points. The main goal is to establish global convergence, *i*.e., that the *whole* sequence converges to a critical point of φ. This can be achieved by imposing an additional assumption on the class of functions φ: it must satisfy the so-called Kurdyka-Łojasiewicz (KL) property [19, 21]. We refer the reader to [11] for an in depth study of the class KL, as well as references therein.

As proven in [12], relying on a key uniformization of the KL property (see [12, Lemma 6], it is possible to establish global convergence of any gradient-like descent sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$, independently of the algorithm used. Verifying the KL property of a given function might often be a difficult task. However, thanks to a result established in [10], any proper and lsc function φ which is *semi-algebraic* satisfies the KL property at any point in $\operatorname{dom}\varphi$, see also the appendix. We can now conveniently summarize the general methodology and convergence results of [12] in the following abstract convergence result.

[2]Here, $\partial\varphi$ stands for the limiting subdifferential of φ [33], which in the convex case reduces to the usual subgradient map [32]. For the reader's convenience, more details are given in the appendix.

Theorem 4.1. *Let $\varphi : \mathbb{R}^d \to (-\infty, \infty]$ be a proper, lsc and semi-algebraic function with $\inf \varphi > -\infty$, and assume that $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ is a gradient-like descent sequence for φ. If $\omega\left(\mathbf{z}^0\right) \subset \operatorname{crit} \varphi$, then the sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ converges to a critical point $\mathbf{z}^*$ of φ.*

Remark 4.2. Under the premises of this theorem, it is also possible to derive a rate of convergence result for the sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ of the form $\left\|\mathbf{z}^k - \mathbf{z}^*\right\| \leq C\ k^{-\gamma}$, for some positive constant C and where $\gamma > 0$ is the so-called KL exponent, see [2] for details.

Semi-algebraic functions abound in applications and include a broad list of functions arising in many optimization models. Moreover, sum and composition of semi-algebraic functions are also semi-algebraic. We refer the reader to [12] and the references therein for more details, and illustrating examples. For the reader's convenience some properties/examples of semi-algebraic sets and functions are also given in the appendix.

For our RSTLS model, it is immediate to see that both $\left(1/\sigma_e^2\right)\|\mathbf{y}\|^2$ and the quadratic function $H(\mathbf{x}, \mathbf{y})$ (*cf.* (3.2)) are semi-algebraic, and hence with any choice of $F(\cdot)$ semi-algebraic, the RSTLS problem thus admits of a semi-algebraic objective function. Therefore, the forthcoming global convergence result established below for SPA, clearly applies to the RSTLS algorithm described in Section 3.3. It should be noted that the in the RSTLS case, the generated sequence is bounded (see condition (C2) in Definition 4.1) as long as the function $F(\cdot)$ is coercive. In this case the objective function $\Psi(\cdot, \cdot)$ has bounded level sets which is enough to guarantee boundedness of the sequence (see [12, Remark 6(i), Page 482]).

4.2 Convergence of SPA

Equipped with the abstract Theorem 4.1, our main objective is now to prove the global convergence of the sequence generated by SPA to a critical point of Ψ. Before doing so, we need to recall some well-known basic results. We begin by recalling the sufficient decrease property of the objective function at a proximal gradient step (*cf.* [6, Lemma 2.3, Page 190]).

Lemma 4.2 (Sufficient decrease property)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with gradient ∇h assumed L_h-Lipschitz continuous and let $g : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper, lsc and convex function. Then, for any $\mathbf{u} \in \operatorname{dom} g$ and*

$$\mathbf{u}^+ = \operatorname{prox}_{\frac{1}{t}g}\left(\mathbf{u} - \frac{1}{t}\nabla h(\mathbf{u})\right),$$

we have

$$h\left(\mathbf{u}^+\right) + g\left(\mathbf{u}^+\right) \leq h(\mathbf{u}) + g(\mathbf{u}) - \left(t - \frac{L}{2}\right)\left\|\mathbf{u}^+ - \mathbf{u}\right\|^2, \tag{4.1}$$

where $L \geq L_h$.

The next result recalls useful basic properties of subdifferential maps (see [33]).

Proposition 4.1 (Subdifferentiability property)**.** *Assume that the coupling function H in problem (M) is continuously differentiable on $\mathbb{R}^n \times \mathbb{R}^p$. Then, for all $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^m$, we have*

$$\partial \Psi(\mathbf{x}, \mathbf{y}) = \left(\nabla_{\mathbf{x}} H(\mathbf{x}, \mathbf{y}) + \partial F(\mathbf{x}), \nabla_{\mathbf{y}} H(\mathbf{x}, \mathbf{y}) + \partial G(\mathbf{y})\right) = \left(\partial_{\mathbf{x}} \Psi(\mathbf{x}, \mathbf{y}), \partial_{\mathbf{y}} \Psi(\mathbf{x}, \mathbf{y})\right). \tag{4.2}$$

Moreover, $\partial \Psi(\cdot, \cdot)$ is a closed map.

We are now ready to state and prove the main convergence result for SPA. In the sequel, for convenience we often use the notation

$$\mathbf{z}^k := \left(\mathbf{x}^k, \mathbf{y}^k\right), \quad \text{for all} \quad k \geq 0,$$

for the sequence generated by SPA.

Theorem 4.2 (Global convergence)**.** *Suppose that Assumptions 1 and 2 hold, and assume that Ψ is semi-algebraic. Let $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ be a sequence generated by SPA, which is assumed to be bounded. Then, the sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ converges to a critical point $\mathbf{z}^*$ of Ψ.*

Proof. We invoke Theorem 4.1. For that purpose the proof is divided in two main parts. We first need to show that the sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ generated by ASPOX is a gradient-like descent with respect to Ψ, that is, according to Definition 4.1, we need to show that conditions (C1) and (C2) hold. In the second part we then have to prove that $\omega(\mathbf{z}^0) \subset \operatorname{crit} \Psi$.

(i) Proving that conditions (C1)-(C2) hold. Thanks to Assumption 1(ii) and Assumption 2(i), we can apply Lemma 4.2 with $h(\cdot) := H\left(\cdot, y^k\right)$, $L_h := L_1\left(y^k\right) = L_k \geq \underline{L} > 0$ (since here $\mathbf{y}^k$ is assumed bounded) and the convex function $g(\cdot) := F(\cdot)$ to get from (3.6) that for any $k \geq 0$,

$$H\left(\mathbf{x}^{k+1}, \mathbf{y}^k\right) + F\left(\mathbf{x}^{k+1}\right) \leq H\left(\mathbf{x}^k, \mathbf{y}^k\right) + F\left(\mathbf{x}^k\right) - \frac{\underline{L}}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2. \tag{4.3}$$

Now, from Assumption 1(ii) and (iii), it follows that $G(\cdot) + H\left(\mathbf{x}^{k+1}, \cdot\right)$ is σ-strongly convex, and hence by the well-known subgradient inequality for a σ-strongly convex function [33, Chapter 12] we obtain for any $k \geq 0$:

$$\begin{aligned} G\left(\mathbf{y}^k\right) + H\left(\mathbf{x}^{k+1}, \mathbf{y}^k\right) &\geq G\left(\mathbf{y}^{k+1}\right) + H\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}\right) \\ &\quad + \left\langle \eta^{k+1} + \nabla_{\mathbf{y}} H\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}\right), \mathbf{y}^k - \mathbf{y}^{k+1}\right\rangle + \frac{\sigma}{2}\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2 \\ &= G\left(\mathbf{y}^{k+1}\right) + H\left(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}\right) + \frac{\sigma}{2}\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2, \end{aligned} \tag{4.4}$$

where $\eta^{k+1} \in \partial G\left(\mathbf{y}^{k+1}\right)$ and the last equality follows immediately from the optimality condition for (3.7). Combining (4.3) and (4.4) and using the fact that $\Psi(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}, \mathbf{y}) + F(\mathbf{x}) + G(\mathbf{y})$ yields, for all $k \geq 0$, that

$$\Psi\left(\mathbf{z}^k\right) - \Psi\left(\mathbf{z}^{k+1}\right) \geq \frac{\underline{L}}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 + \frac{\sigma}{2}\left\|\mathbf{y}^{k+1} - \mathbf{y}^k\right\|^2 \geq \frac{\rho_1}{2}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|^2, \tag{4.5}$$

where $\rho_1 = \min\{\underline{L}, \sigma\}$, thus proving condition (C1) for the sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ with respect to Ψ. Moreover, as a byproduct, it also follows that

$$\lim_{k\to\infty}\left(\mathbf{z}^{k+1}-\mathbf{z}^k\right)=\mathbf{0}. \tag{4.6}$$

Indeed, from (4.5), the sequence $\left\{\Psi\left(\mathbf{z}^k\right)\right\}_{k\geq 0}$ is nonincreasing, and since Ψ is assumed to be bounded from below (see Assumption 2(iii)), it converges to some real number $\underline{\Psi}$. Summing (4.5) from $k=0$ to $N-1$ we thus get

$$\sum_{k=0}^{N-1}\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|^2 \leq \frac{2}{\rho_1}\left(\Psi\left(\mathbf{z}^0\right)-\Psi\left(\mathbf{z}^N\right)\right) \leq \frac{2}{\rho_1}\left(\Psi\left(\mathbf{z}^0\right)-\underline{\Psi}\right), \tag{4.7}$$

and taking the limit as $N\to\infty$ the claim (4.6) follows.

Now, we prove that condition (C2) holds for the sequence $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$. Writing the optimality conditions for the convex iterations of SPA defined in (3.6) and (3.7) we have for any $k\geq 0$:

$$\nabla_{\mathbf{x}}H\left(\mathbf{x}^k,\mathbf{y}^k\right)+L_k\left(\mathbf{x}^{k+1}-\mathbf{x}^k\right)+\mathbf{u}^{k+1}=\mathbf{0},\ \mathbf{u}^{k+1}\in\partial F\left(\mathbf{x}^{k+1}\right), \tag{4.8}$$

$$\nabla_{\mathbf{y}}H\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1}\right)+\mathbf{v}^{k+1}=\mathbf{0},\ \mathbf{v}^{k+1}\in\partial G\left(\mathbf{y}^{k+1}\right). \tag{4.9}$$

On the other hand, using Proposition 4.1 we have

$$\partial\Psi\left(\mathbf{z}^k\right)=\left(\nabla_{\mathbf{x}}H\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1}\right)+\partial F\left(\mathbf{x}^{k+1}\right),\nabla_{\mathbf{y}}H\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1}\right)+\partial G\left(\mathbf{y}^{k+1}\right)\right). \tag{4.10}$$

Therefore, defining the quantity

$$\xi_{\mathbf{x}}^{k+1}:=L_k\left(\mathbf{x}^k-\mathbf{x}^{k+1}\right)+\nabla_{\mathbf{x}}H\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1}\right)-\nabla_{\mathbf{x}}H\left(\mathbf{x}^k,\mathbf{y}^k\right),$$

and using (4.10) together with (4.8) and (4.9), we obtain

$$\mathbf{w}^{k+1}:=\left(\xi_{\mathbf{x}}^{k+1},\mathbf{0}\right)\in\partial\Psi\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1}\right).$$

To complete the proof of condition (C2) it remains to estimate the norm of $\mathbf{w}^{k+1}$. For that, since we assumed that $\left\{\mathbf{z}^k\right\}_{k\in\mathbb{N}}$ is bounded, thanks to Assumption 2(i) it follows that $L_k\leq\overline{L}$, and since by Assumption 2(ii) ∇H is Lipschitz continuous over bounded subsets of $\mathbb{R}^n\times\mathbb{R}^m$, there exists $M>0$ such that

$$\begin{aligned}\left\|\mathbf{w}^{k+1}\right\|=\left\|(\xi_{\mathbf{x}}^{k+1},\mathbf{0})\right\| &\leq L_k\left\|\mathbf{x}^k-\mathbf{x}^{k+1}\right\|+\left\|\nabla_{\mathbf{x}}H\left(\mathbf{x}^{k+1},\mathbf{y}^{k+1}\right)-\nabla_{\mathbf{x}}H\left(\mathbf{x}^k,\mathbf{y}^k\right)\right\|\\ &\leq\overline{L}\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|+M\left(\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|+\left\|\mathbf{y}^{k+1}-\mathbf{y}^k\right\|\right)\\ &\leq\overline{L}\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|+\sqrt{2}M\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|\\ &\leq\left(\sqrt{2}M+\overline{L}\right)\left\|\mathbf{z}^{k+1}-\mathbf{z}^k\right\|,\end{aligned} \tag{4.11}$$

thus proving that condition (C2) holds with $\rho_2:=\sqrt{2}M+\overline{L}$.

(ii) Proving that $\omega(\mathbf{z}^0) \subset \operatorname{crit}\Psi$. Let $\mathbf{z}^* = (\mathbf{x}^*, \mathbf{y}^*)$ be a limit point of $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$. Thus, there exists a subsequence $\{(\mathbf{x}^{k_q}, \mathbf{y}^{k_q})\}_{q\in\mathbb{N}}$ such that $(\mathbf{x}^{k_q}, \mathbf{y}^{k_q}) \to (\mathbf{x}^*, \mathbf{y}^*)$ as $q \to \infty$. We need to show that

$$\lim_{q\to\infty} \Psi(\mathbf{x}^{k_q}, \mathbf{y}^{k_q}) = \Psi(\mathbf{x}^*, \mathbf{y}^*) \text{ and } (\mathbf{0}, \mathbf{0}) \in \partial\Psi(\mathbf{x}^*, \mathbf{y}^*).$$

Following the same proof as [12, Lemma 5(i), Page 476] for the iterative Step (3.6), since both the sequences $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ and L_k are bounded, ∇H is continuous and from (4.6) the distance between two successive iterates tends to zero, it follows that $F(\mathbf{x}^{k_q})$ tends to $F(\mathbf{x}^*)$ as $q \to \infty$. On the other hand, from the iterative Step (3.7), we have for all $k \geq 0$

$$H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + G(\mathbf{y}^{k+1}) \leq H(\mathbf{x}^{k+1}, \mathbf{y}^*) + G(\mathbf{y}^*).$$

Since $H(\cdot,\cdot)$ is continuous (see Assumption 1(iii)), choosing $k = k_q - 1$ in the above inequality and letting q goes to ∞, yields

$$\limsup_{q\to\infty} G(\mathbf{y}^{k_q}) \leq G(\mathbf{y}^*),$$

and since G is lower semicontinuous (see Assumption 1(ii)), this shows that $G(\mathbf{y}^{k_q})$ tends to $G(\mathbf{y}^*)$ as $q \to \infty$. Therefore,

$$\begin{aligned}\lim_{q\to\infty} \Psi(\mathbf{x}^{k_q}, \mathbf{y}^{k_q}) &= \lim_{q\to\infty} \{H(\mathbf{x}^{k_q}, \mathbf{y}^{k_q}) + F(\mathbf{x}^{k_q}) + G(\mathbf{y}^{k_q})\} \\ &= H(\mathbf{x}^*, \mathbf{y}^*) + F(\mathbf{x}^*) + G(\mathbf{y}^*) = \Psi(\mathbf{x}^*, \mathbf{y}^*).\end{aligned}$$

Finally, since $\mathbf{w}^{k+1} = (\xi_{\mathbf{x}}^{k+1}, \mathbf{0}) \in \partial\Psi(\mathbf{x}^{k+1}, \mathbf{y}^k)$ and since from (4.6) and (4.11) $\|\mathbf{w}^{k+1}\| \to 0$ as $k \to \infty$, then recalling the closedness property of the map $\partial\Psi$ (*cf.* Proposition 4.1), we obtain that $(\mathbf{0}, \mathbf{0}) \in \partial\Psi(\mathbf{x}^*, \mathbf{y}^*)$, which completes the proof that $(\mathbf{x}^*, \mathbf{y}^*)$ is a critical point of Ψ. □

5 Numerical Results

In this section we report on several numerical experiments of the SPA algorithm applied to the problem of deblurring images. We split this section into two subsections. In the first subsection we present experiments on small scale images (21×21), while in the second subsection we present results for large images (256×256). The motivation for this partition is the fact that on small images we can compare our algorithm to the RSTLN method [29]. Since the RSTLN method is based on the Gauss-Newton method, it can be used efficiently to solve small scale problems.

To be more precise, in this section we are interested in solving the problem

$$\min_{\mathbf{x},\mathbf{y}} \left\{ \lambda f(L\mathbf{x}) + \frac{1}{\sigma_w^2} \left\| \left(\mathbf{A} + \sum_{i=1}^{p} y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right\|^2 + \frac{1}{\sigma_e^2} \|\mathbf{y}\|^2 \right\},$$

where

- $\mathbf{b}$ is the vectorized observed image (that is, a vector obtained by stacking the columns of the observed image). This image was obtained from the original "vectorized image $\mathbf{b}_t$ whose pixels were scaled to be between 0 and 1. The image goes through a Gaussian blur point spread function (PSF) of size $q \times q$ and standard deviation γ (given by the MATLAB command psfGauss([q,q],γ)), by using the MATLAB command imfilter. After that we add a zero-mean white Gaussian noise with standard deviation σ_w.

- We assume that the blurring operator is not exactly known and that the observed PSF is constructed by taking the original PSF and adding to it a $q \times q$ matrix with the same structure as the original PSF such that the components of each structure matrix $\mathbf{A}_i$ are with noise which independently generated from a zero-mean normal distribution with standard deviation σ_e.

- We consider here the periodic boundary condition. This means that the blurring matrix has a structure of BCCB. For more details, we refer the reader to the book [18].

- $f : \mathbb{R}^n \to (-\infty, +\infty]$ is a regularizing function and $\lambda > 0$ is a regularizing parameter.

5.1 Small Scale Images - Comparison with RSTLN

We begin this section by recalling the RSTLN algorithm (see [34, 29]). This method was designed to solve the following class of RSTLS problems:

$$\min_{\mathbf{x}\in\mathbb{R}^n, \mathbf{y}\in\mathbb{R}^p} \left\| \begin{pmatrix} (1/\sigma_w)\left((\mathbf{A} + \sum_{i=1}^p y_i \mathbf{A}_i)\,\mathbf{x} - \mathbf{b}\right) \\ (1/\sigma_e)\,\mathbf{y} \\ \lambda L\mathbf{x} \end{pmatrix} \right\|_\mu , \tag{5.1}$$

where $\mu = 1, 2, \infty$. It should be noted that when $\mu = 2$, this formulation is equivalent to (1.1) where the regularizer $f(\cdot)$ is chosen as $\lambda \|\mathbf{x}\|_2^2$. This model is more flexible in the sense that it allows different types of norms to measure the fidelity to the data (in (1.1) only ℓ_2 is possible for the objective function). A major drawback of this model is the fact that the type of the chosen norm should be the same for the data fidelity and for the regularizer. On the other hand, in (1.1) any type of regularizer can be considered, which gives us the possibility of tackling the problem in the common situation that the data fidelity is measured with ℓ_2 norm and the regularizer is given by ℓ_1 norm.

As we mentioned above, the RSTLN method is based on the Gauss-Newton method, which means that at each iteration one should solve an optimization problem of the following form:

$$\min_{\mathbf{x},\mathbf{y}} \left\{ \left\| \begin{pmatrix} \mathbf{A} + \sum_{i=1}^p y_i^k \mathbf{A}_i & \mathbf{B}_k \\ \mathbf{0}_{p\times n} & \frac{\sigma_w}{\sigma_e}\mathbf{I}_{p\times p} \end{pmatrix} \begin{pmatrix} \mathbf{x} - \mathbf{x}^k \\ \mathbf{y} - \mathbf{y}^k \end{pmatrix} \right\|^2 + \lambda f(L\mathbf{x}) \right\}, \tag{5.2}$$

where $\left(\mathbf{x}^k, \mathbf{y}^k\right)$ is the given iterate and $\mathbf{B}_k$ is the $m \times p$ matrix defined in (3.10), that is, $\mathbf{B}_k := \left[\mathbf{A}_1\mathbf{x}^k, \mathbf{A}_2\mathbf{x}^k, \ldots, \mathbf{A}_p\mathbf{x}^k\right]$.

It should be noted that this is not exactly the problem that was solved in [29], because here we apply the Gauss-Newton method on the model (1.1) and not on the model (5.1), meaning we look at general regularizers. Here we use CVX in order to solve these subproblems at each iteration.

Example 5.1. Consider the 21×21 'plus' image given in Figure 1(1). The "true" image goes through a Gaussian blur point spread function (PSF) of size 5×5 and standard deviation 2, and then the observed image is constructed by adding to each of the components of the blurred image a zero-mean normally distributed random variable with standard deviation $\sigma_w = 10^{-4}$. See Figure 1 for the blurred and noised image (displayed on the right-hand side). We also add noise to the structure components vector of the PSF which is normally distributed with zero-mean and variance $\sigma_e = 10^{-4}$. In this example we have 6 structure matrices, *i*.e., $p = 6$, and each matrix is of size 5×5. Specifically , the structure matrices are

$$A_1 = \begin{bmatrix} 0.0352 & 0 & 0 & 0 & 0.0352 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0352 & 0 & 0 & 0 & 0.0352 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0.0387 & 0 & 0.0387 & 0 \\ 0.0387 & 0 & 0 & 0 & 0.0387 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0387 & 0 & 0 & 0 & 0.0387 \\ 0 & 0.0387 & 0 & 0.0387 & 0 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 0 & 0 & 0.0399 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0399 & 0 & 0 & 0 & 0.0399 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0399 & 0 & 0 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0425 & 0 & 0.0425 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0425 & 0 & 0.0425 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_5 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0438 & 0 & 0 \\ 0 & 0.0438 & 0 & 0.0438 & 0 \\ 0 & 0 & 0.0438 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_6 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0452 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Here we regularize the problem with the ℓ_1 norm, that is, $f(\mathbf{x}) := \lambda \|\mathbf{x}\|_1$ with a regularization parameter $\lambda = 10^{-3}$. In this case the proximal mapping is the so-called soft-thresholding operator which is given by

$$\operatorname{prox}_{\lambda f}(\mathbf{x})_i := (|x_i| - \lambda)_+ \operatorname{sign}(x_i),$$

where x_i is the i-coordinate of $\mathbf{x}$ and $(x)_+ := \max\{x, 0\}$.

In this experiment we used periodic boundary condition and initialize SPA with $\mathbf{x}^0 = \mathbf{b}$, the observed image, and $\mathbf{y} = \mathbf{0}$. We first solved the problem with the RSTLN method which after 8 outer iterations (of the Gauss-Newton method) produced the solution given

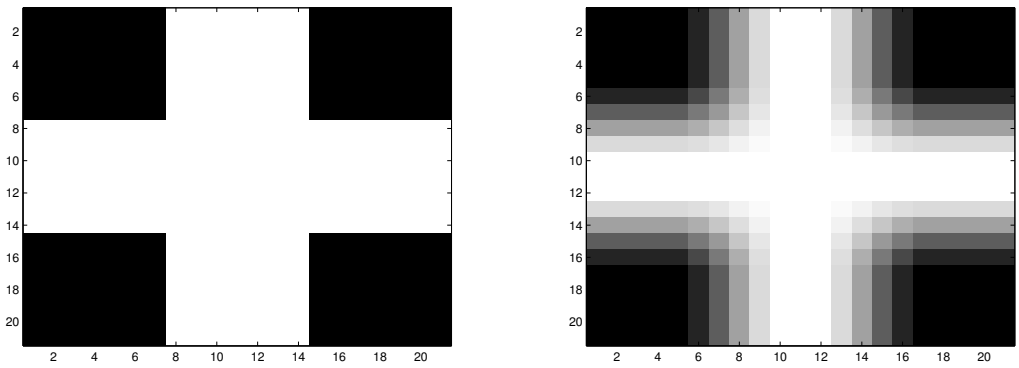

Figure 1: The plus test image: original image (left), blurred and noised image (right)

in Figure 2 (in the top-left corner). The relative error of this solution is 0.0405 (*i*.e., the value of $\|\mathbf{x}_{RSTLN} - \mathbf{x}_{orig}\| / \|\mathbf{x}_{orig}\|$) when the relative error of the blurred and noisy image is 0.2725. The SPA algorithm achieves such an error (*i*.e., 0.0405) after 972 iterations (see the solution obtained by the SPA in Figure 2, the top-right corner). The SPA obtained this solution after 2.15 seconds while the RSTLN method did it after 9.88 seconds. The relative error between the two solutions (SPA and RSTLN) is $\|\mathbf{x}_{RSTLN} - \mathbf{x}_{SPA}\| / \|\mathbf{x}_{SPA}\| = 0.047$. The two figures in the bottom obtained by the SPA algorithm after 2500 (left corner) and 4500 (right corner) iterations.

It should be noted that the 4500 iterations of the SPA algorithm took 10 seconds which is almost the time that the RSTLN ran. So, for the same running time, the SPA algorithm generated a much better solution than the RSTLN method (the relative error of the SPA after 4500 iterations is 0.0252 and after 2500 iterations is 0.0288). It should be noted that the poor performance of the RSTLN method here is also due to the fact that we have used CVX in order to solve (5.2) at each iteration of the Gauss-Newton method.

We made more extensive set of tests in which we solved 10 problems – each corresponding to a realization of σ_w and σ_e for different values 10^{-2}, 10^{-3} and 10^{-4}. In Table 1, for each value of standard deviation, and for each choice of an algorithm (SPA and RSTLN), we indicate the average CPU time of each algorithm (out of the 100) with tolerance parameter $\varepsilon > 0$ for $\varepsilon = 10^{-3}, 10^{-4}, 10^{-5}$ after 100 iterations; The parameter ε is used in the stopping criterion, which is $\left\|\mathbf{x}^{N+1} - \mathbf{x}^{N}\right\| / \left\|\mathbf{x}^{N}\right\| \leq \varepsilon$ for some $N \in \mathbb{N}$).

For the moderate accuracy $\varepsilon = 10^{-3}$, the SPA method is significantly better (more than 10 times faster) than the RSTLN method. For the greater accuracies $\varepsilon = 10^{-4}, 10^{-5}$, it is clear that the SPA method is better than RSTLN (about twice faster). It also should be noted that the maximum CPU time for RSTLN could be very high for larger accuracies, for example, for the parameter $\varepsilon = 10^{-5}$, RSTLN run for more than 230 seconds (for $\sigma_w = \sigma_w = 10^{-2}$), which for the SPA, only required about 11 seconds.

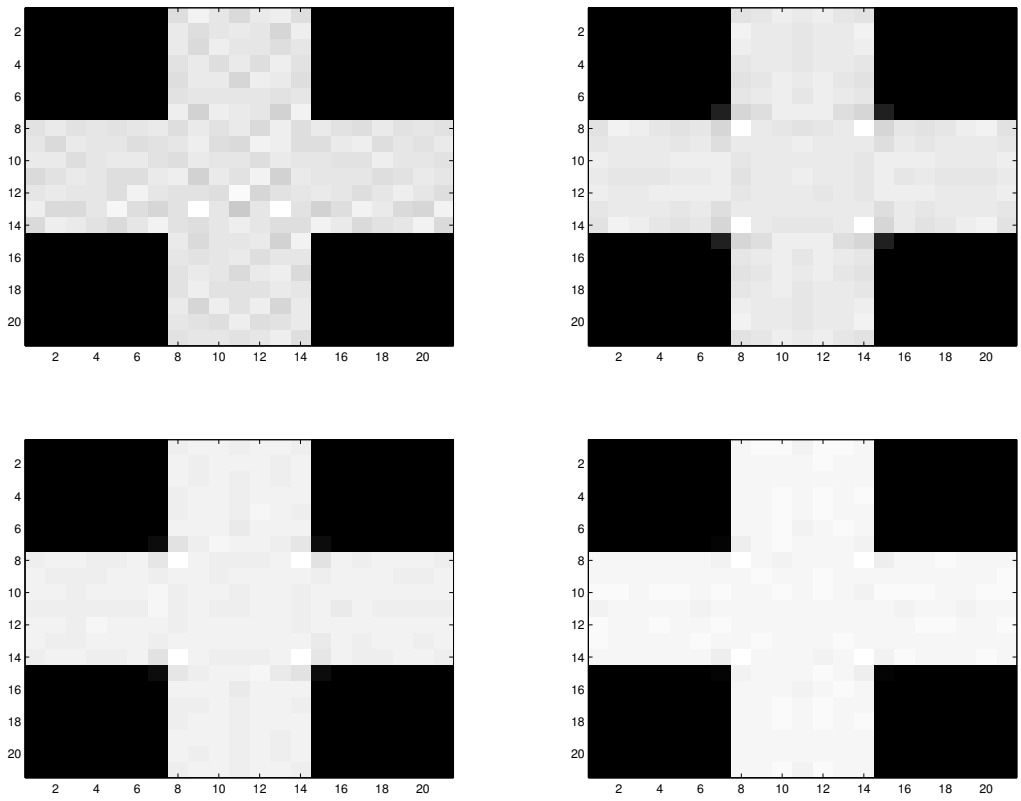

Figure 2: The plus test image: RSTLN solution (top-left corner), SPA solution - 972 iterations (top-right corner), SPA solution - 2500 iterations (bottom- right corner), SPA solution - 4500 iterations (bottom-left corner)

5.2 Large Scale Image

In this section we present numerical examples on large scale images. We consider the 256×256 camerman test image whose pixels were scaled into the range between 0 and 1. We will demonstrate the viability and effectiveness of the proposed SPA method in deblurring large scale images.

Example 5.2. The camerman test image went through a Gaussian blur point spread function (PSF) of size 5×5 and standard deviation 4, and then the observed image is constructed by adding to each of the components of the blurred image a zero-mean normally distributed random variable with standard deviation $\sigma_w = 10^{-3}$. See Figure 3 for the original image (displayed on the left-hand side) and the blurred and noised image (displayed on the right-hand side). We also add noise to the structure components vector of the PSF, which is normally distributed with zero-mean and variance $\sigma_e = 10^{-3}$. The number and size of the structure matrices are the same as in the previous example.

σ_w	σ_e	$\varepsilon = 10^{-3}$		$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-5}$	
		SPA	RSTLN	SPA	RSTLN	SPA	RSTLN
10^{-4}	10^{-4}	0.0769	1.7733	0.3423	2.2879	1.256	2.3395
		0.09 0.07	1.84 1.69	0.36 0.33	2.68 2.12	1.35 1.23	2.91 2.22
10^{-3}	10^{-4}	0.0789	1.9428	0.35	2.1357	1.5889	3.0093
		0.1 0.07	2.42 1.66	0.36 0.34	2.5 2.19	1.69 1.57	3.91 2.3
10^{-3}	10^{-3}	0.0774	1.8299	0.3484	2.4399	1.5506	3.2463
		0.12 0.07	2.26 1.55	0.38 0.33	3.09 2.29	1.69 1.52	4 2.38
10^{-2}	10^{-2}	0.0802	7.4978	0.5838	16.2789	10.5208	19.243
		0.11 0.07	39.35 3.37	0.68 0.57	185.61 4.84	10.73 10.4	230.59 5.8

Table 1: average CPU time(out of 100) in the first line, maximum and minimum CPU time (out of 100) in the second line for which an ε-optimal solution is reached.

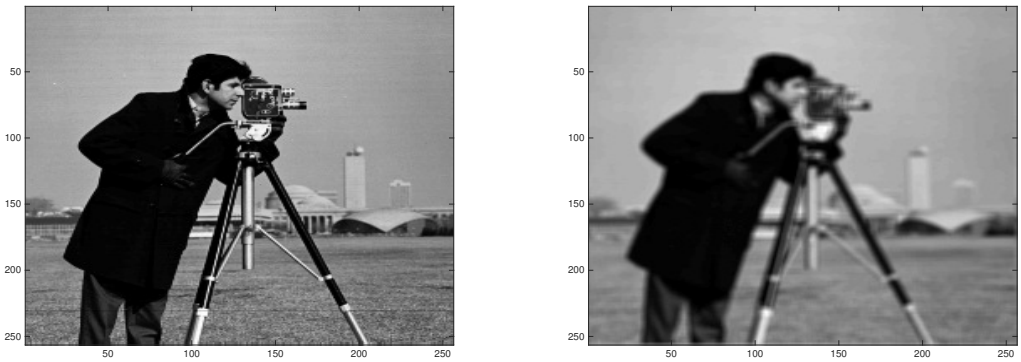

Figure 3: The camerman test image: original image (left), blurred and noised image (right)

In this experiment we used periodic boundary conditions and considered the total variation (TV)-based regularization, that is, $F(\mathbf{x}) = \lambda TV_1(\mathbf{x})$ where $\lambda = 10^{-3}$ (see Example 2.3 for the precise definition of TV_1). We initialized the algorithm in the same way as in the previous example, namely the initial $\mathbf{x}$ is the observed image and $\mathbf{y} = \mathbf{0}$. The main effort in implementing SPA for deblurring the image involves the computation of the proximal mapping of TV_1. It is well-known that there is no an explicit expression for the proximal mapping of TV_1, and therefore it requires an additional iterative method. We will use the FDPG method of [7] in order to compute the prox of TV_1. When implementing the FDPG at each iteration of SPA, we bound the number of the iterations of FDPG by 12.

The figure below (see Figure 4) obtained by the SPA after 50 and 200 iterations.

The relative error of the solution obtained by the SPA after 50 iterations is 0.0868 (*i*.e., the value of $\|\mathbf{x}_{SPA} - \mathbf{x}_{orig}\| / \|\mathbf{x}_{orig}\|$) while after 200 iterations the relative error is 0.0657. The relative error of the blurred and noisy image is 0.1327.

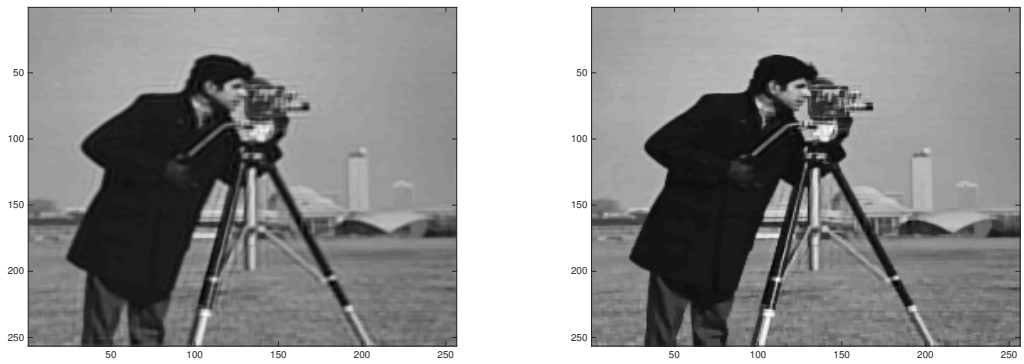

Figure 4: The camerman test image: SPA solution after 50 (left) and 200 (right) iterations

A Appendix

A.1 Nonsmooth Calculus

Let us recall a few definitions concerning subdifferential calculus (see, for instance, [33]). Recall that for a proper and lower semicontinuous function $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$, the domain of φ is defined through $\operatorname{dom}\varphi := \left\{\mathbf{x} \in \mathbb{R}^d : \ \varphi(\mathbf{x}) < +\infty\right\}$.

Definition A.1 (Subdifferentials)**.** Let $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function.

(i) For a given $\mathbf{x} \in \operatorname{dom}\varphi$, the *Fréchet subdifferential* of φ at $\mathbf{x}$, written $\widehat{\partial}\varphi(\mathbf{x})$, is the set of all vectors $\mathbf{u} \in \mathbb{R}^d$ which satisfy

$$\liminf_{\mathbf{y}\neq\mathbf{x}\ \mathbf{y}\to\mathbf{x}} \frac{\varphi(\mathbf{y}) - \varphi(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y}-\mathbf{x}\rangle}{\|\mathbf{y}-\mathbf{x}\|} \geq 0.$$

When $\mathbf{x} \notin \operatorname{dom}\varphi$, we set $\widehat{\partial}\varphi(\mathbf{x}) = \emptyset$.

(ii) The *limiting-subdifferential*, or simply the subdifferential, of φ at $\mathbf{x} \in \mathbb{R}^n$, written $\partial\varphi(\mathbf{x})$, is defined through the following closure process

$$\partial\varphi(\mathbf{x}) := \left\{\mathbf{u} \in \mathbb{R}^d : \ \exists\, \mathbf{x}^k \to \mathbf{x}, \varphi\left(\mathbf{x}^k\right) \to \varphi(\mathbf{x}) \text{ and } \mathbf{u}^k \in \widehat{\partial}\varphi\left(\mathbf{x}^k\right) \to \mathbf{u} \text{ as } k \to \infty\right\}.$$

Note that in this nonsmooth and nonconvex context, the well-known Fermat's rule remains barely unchanged. It formulates as: "if $\mathbf{x} \in \mathbb{R}^d$ is a local minimizer of φ then $\mathbf{0} \in \partial\varphi(\mathbf{x})$". Points whose subdifferential contains $\mathbf{0}$ are called *critical points*, and the set of critical points of φ is denoted by $\operatorname{crit}\varphi$.

A.2 Semi-algebraic sets and functions

We recall here basic definitions and properties, see *e.*g., [12] and references therein for more details.

Definition A.2 (Semi-algebraic sets and functions)**.** (i) A subset S of $\mathbb{R}^d$ is a real semi-algebraic set if there exists a finite number of real polynomial functions $g_{ij}, h_{ij} : \mathbb{R}^d \to \mathbb{R}$ such that

$$S = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \left\{ u \in \mathbb{R}^d : \; g_{ij}(u) = 0 \,\text{and}\; h_{ij}(u) < 0 \right\}.$$

(ii) A function $h : \mathbb{R}^d \to (-\infty, +\infty]$ is called semi-algebraic if its graph

$$\left\{ (u,t) \in \mathbb{R}^{d+1} : \; h(u) = t \right\},$$

is a semi-algebraic subset of $\mathbb{R}^{d+1}$.

The following useful result can be found in [10].

Theorem A.1. *Let $\varphi : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. If φ is semi-algebraic then it satisfies the KL property at any point of* $\operatorname{dom} \varphi$.

The class of semi-algebraic sets is stable under the following operations: finite unions, finite intersections, complementation and Cartesian products. There is broad class of semi-algebraic functions arising in optimization: Real polynomial functions; indicator functions of semi-algebraic sets; finite sums product and composition of semi-algebraic functions; Sup/Inf type functions; the sparsity measure (or the counting norm) $\|\mathbf{x}\|_0$ of a vector $\mathbf{x}$, and much more, see *e.*g., [12] and references therein.

References

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

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

[3] A. Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM J. Optim.*, 25(1):185–209, 2015.

[4] A. Beck and A. Ben-Tal. On the solution of the Tikhonov regularization of the total least squares problem. *SIAM J. Optim.*, 17(1):98–118, 2006.

[5] A. Beck, A. Ben-Tal, and C. Kanzow. A fast method for finding the global solution of the regularized structured total least squares problem for image deblurring. *SIAM J. Matrix Anal. Appl.*, 30(1):419–443, 2008.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[7] A. Beck and M. Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Oper. Res. Lett.*, 42(1):1–6, 2014.

[8] D. P. Bertsekas. *Nonlinear Programming*. Belmont MA: Athena Scientific, 2nd edition, 1999.

[9] J. Bolte, P. L. Combettes, and J.-C. Pesquet. Alternating proximal algorithm for blind image recovery. In *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP)*, pages 1673–1676, 2010.

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

[11] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: Subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362(6):3319–3363, 2010.

[12] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program. Ser. A*, 146:459–494, 2014.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

[14] A.Repetti E. Chouzenoux, J.-C. Pesquet. A block coordinate variable metric forward-backward algorithm. 2013.

[15] G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.*, 21(1):185–194, 1999.

[16] G. H. Golub and C. F. van Loan. An analysis of the total least squares problem. *SIAM J. Numer. Anal.*, 17(6):883–893, 1980.

[17] P. C. Hansen. Regularization tools: a Matlab package for analysis and solution of discrete ill-posed problems. *Numer. Algorithms*, 6(1-2):1–35, 1994.

[18] P. C. Hansen, J. G. Nagy, and D. P. O'Leary. *Deblurring Images: Matrices, Spectra, and Filtering*, volume 3 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

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

[20] P. Lemmerling and S. van Huffel. Analysis of the structured total least squares problem for Hankel/Toeplitz matrices. *Numer. Algorithms*, 27(1):89–114, 2001.

[21] 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.

[22] I. Markovsky, S. van Huffel, and R. Pintelon. Block-Toeplitz/Hankel structured total least squares. *SIAM J. Matrix Anal. Appl.*, 26(4):1083–1099, 2005.

[23] N. Mastronardi, P. Lemmerling, and S. van Huffel. Fast structured total least squares algorithm for solving the basic deconvolution problem. *SIAM J. Matrix Anal. Appl.*, 22(2):533–553, 2000.

[24] B. De Moor. Structured total least squares and {L2} approximation problems. *Linear Algebra Appl*, 188-189:163–205, 1993.

[25] B. De Moor. Total least squares for affinely structured matrices and the noisy realization problem. *IEEE Trans. Signal Processing*, 42(11):3104–3113, 1994.

[26] J. J. Moré. Generalizations of the trust region subproblem. *Optim. Methods Softw.*, 2(3):189–209, 1993.

[27] J. J. Moreau. Proximitéet dualité dans un espace Hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[28] N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, 2014.

[29] A. Pruessner and D. P. O'Leary. Blind deconvolution using a regularized structured total least norm algorithm. *SIAM J. Matrix Anal. Appl.*, 24(4):1018–1037, 2003.

[30] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet. Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1/ℓ_2 regularization. *IEEE Signal Process. Lett*, 22(5):539–543, 2015.

[31] 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 J. Imaging Sci.*, 8(1):426–457, 2015.

[32] R. T. Rockafellar. *Convex Analysis*. Princeton Math. Ser. 28. Princeton University Press, Princeton, NJ, 1970.

[33] R. T Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundl. Math. Wiss.* Springer-Verlag, Berlin, 1998.

[34] J. B. Rosen, H. Park, and J. Glick. Total least norm formulation and solution for structured problems. *SIAM J. Matrix Anal. Appl.*, 17(1):110–126, 1996.

[35] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet. Math. Dokl.*, 5:1035–1038, 1963.

[36] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems.* Scripta Ser. Math., V. H. Winston & Sons, Washington, DC; John Wiley & Sons, New York, Toronto, London, 1977. translated from the Russian; preface by F. John, ed.

[37] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

[38] S. van Huffel and J. Vandewalle. *The Total Least Squares Problem: Computational Aspects and Analysis*, volume 9 of *Frontiers in Appl. Math.* SIAM, 1991.