

## AN ALTERNATING SEMIPROXIMAL METHOD FOR NONCONVEX REGULARIZED STRUCTURED TOTAL LEAST SQUARES PROBLEMS\*

AMIR BECK<sup>†</sup>, SHOHAM SABACH<sup>†</sup>, AND MARC TEBoulLE<sup>‡</sup>

**Abstract.** We consider a broad class of regularized structured total least squares (RSTLS) problems 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–Lojasiewicz property, global convergence, nonconvex-nonsmooth minimization, proximal gradient methods, regularized structured total least squares, semialgebraic functions

**AMS subject classifications.** 90C26, 90C30, 49N45, 65K05

**DOI.** 10.1137/15M1017557

**1. Introduction.** In this paper we consider a broad class of optimization problems for regularized structured total least squares (RSTLS), which captures many models arising in applications and consists of solving the following nonconvex and nonsmooth minimization problem:

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

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,  $\sigma_e$  and  $\sigma_w$  are the standard deviations of the corresponding noise components (see section 5 for more details), and  $F : \mathbb{R}^d \rightarrow (-\infty, +\infty]$  is a given convex (possibly nonsmooth) function that allows us 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  $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_p \in \mathbb{R}^{m \times n}$ , with  $y_1, y_2, \dots, y_p \in \mathbb{R}$  being the unknown structure components. The proposed model is motivated by our desire to tackle and extend various well-known total least squares (TLS)-based models which have been studied in the literature and will be briefly recalled and discussed in the next section.

\*Received by the editors April 20, 2015; accepted for publication (in revised form) by J. G. Nagy February 29, 2016; published electronically August 24, 2016.

<http://www.siam.org/journals/simax/37-3/M101755.html>

<sup>†</sup>Department of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 3200003, Israel (becka@ie.technion.ac.il, ssabach@ie.technion.ac.il). The research of the first author was partially supported by the Israel Science Foundation under grant ISF 253/12.

<sup>‡</sup>School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel (teboulle@post.tau.ac.il). The research of this author was partially supported by the Israel Science Foundation under grant ISF 998/12.

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). 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 in 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 aforementioned structures and properties; 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) 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 activity 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, 39] and references therein for a small representative sample of this activity. However, here the RSTLS problem under consideration is *nonconvex* and nonsmooth. Motivated by the recent algorithmic and convergence analysis framework developed in [12], which builds on the powerful Kurdyka–Lojasiewicz property [22, 24] 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 *semiproximal* regularization, which for ease of reference is called SPA. While the focus of this paper is on 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 asks 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 TLS models which have motivated the proposed RSTLS approach and make our setting more precise. While the focus of this paper is on the RSTLS problem, our analysis is developed for a 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 additional technical material. Our notation is quite standard 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 one to study overdetermined linear systems of the form

$$(2) \quad \mathbf{Ax} \approx \mathbf{b},$$

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

$$(3) \quad (\mathbf{A} + \mathbf{E}) \mathbf{x} = \mathbf{b} + \mathbf{w},$$

where  $\mathbf{E} \in \mathbb{R}^{m \times n}$  and  $\mathbf{w} \in \mathbb{R}^m$  are the 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 TLS method, 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 (3); that is, to solve the following nonconvex minimization problem:

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

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 decomposition of the augmented matrix  $(\mathbf{A}, \mathbf{b})$ ; see [16, 21].

However, it is 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 performance. In the context of image processing, which is one of the main applications of interest in this work, it is natural and adequate to consider the same structure on the perturbation 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  $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_p \in \mathbb{R}^{m \times n}$  are the structure matrices and  $y_1, y_2, \dots, y_p \in \mathbb{R}$  are the unknown structure components. Then, the structured total least squares (STLS) method aims at solving the following minimization problem [1, 5, 23, 26, 28, 29, 36]:

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

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

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

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

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, [25, 27, 36]).

In [36], the authors developed the structured total least norm (STLN) method to solve this 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-conditioned, and applying STLN results in poor image recovery. Thus regularization is needed. Such an approach was considered, for instance, in [32], which implemented Tikhonov regularization [37] 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  $\infty$ . For another work which focused on quadratic regularizers and improved the computational efficiency of the algorithm from [32], see, for instance, [26]. RSTLN-based methods rely on a variation of the Gauss–Newton method,<sup>1</sup> one of the main drawbacks of which is that it 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 a global minimum.

Linear inverse problems where the model matrix  $\mathbf{A}$  is ill-conditioned are common phenomena, which directly implies that solutions obtained by the LS (least squares), TLS, or STLS methods have a huge norm, resulting in meaningless solutions. This was discussed above within the RSTLN approach. Alternatively, this difficulty is often handled by, for example, 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 [38], 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; see [4, 15, 18] and the references therein for several regularization ideas.

Motivated by the above, in this paper we focus on a broad class of RSTLS problems which capture the ideas and models just described and allow for unification and extension, by considering the following nonconvex and nonsmooth RSTLS problem:

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

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

**2.2. Some model 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 1* (the constrained RSTLS problem). Let the nonsmooth function  $F$  be

<sup>1</sup>When the regularizer is quadratic, e.g.,  $p = 2$ , the Gauss–Newton method is directly applicable, while for  $p = 1, \infty$ , the objective function is no longer differentiable but can be applied similarly; see [32] for details.

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) the constrained RSTLS model given by

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

*Example 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

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

where  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  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 [38], in which  $f(\cdot) = \|\cdot\|^2$ . Another penalty function that has attracted revived interest and considerable attention in the signal processing literature is the use of the  $\ell_1$ -norm, which is discussed in more detail in the next example.

*Example 3* (total variation (TV) 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  $\ell_1$ -based anisotropic TV, given by

$$(4) \quad \begin{aligned} TV_1(\mathbf{x}) = & \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (|x_{i,j} - x_{i+1,j}| + |x_{i,j} - x_{i,j+1}|) + \sum_{i=1}^{m-1} |x_{i,n} - x_{i+1,n}| \\ & + \sum_{j=1}^{n-1} |x_{m,j} - x_{m,j+1}|. \end{aligned}$$

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

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

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

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

**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 nonconvex and nonsmooth optimization model

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

with the following assumption.

*Assumption 1.*

- (i)  $F : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is a proper, closed, and convex function.
- (ii)  $G : \mathbb{R}^p \rightarrow (-\infty, +\infty]$  is a proper, closed, and strongly convex function with strong convexity constant  $\sigma$ .
- (iii)  $H : \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}$  is a  $C^1$  function, and for any fixed  $\mathbf{x}$  the function  $\mathbf{y} \rightarrow H(\mathbf{x}, \mathbf{y})$  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

$$(5) \quad H(\mathbf{x}, \mathbf{y}) = \frac{1}{\sigma_w^2} \left\| \left( \mathbf{A} + \sum_{i=1}^p y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right\|^2 \quad \text{and} \quad G(\mathbf{y}) = \frac{1}{\sigma_e^2} \|\mathbf{y}\|^2$$

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  $\mathbf{y} \rightarrow H(\mathbf{x}, \mathbf{y})$  is convex on  $\mathbb{R}^p$ .

Note that the first item in Assumption 1 allows us to consider various interesting scenarios (see the 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 in building 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 suggests applying 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  $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k \in \mathbb{N}}$  via the following two steps: for  $k \in \mathbb{N}$

$$(6) \quad \mathbf{x}^{k+1} \in \operatorname{argmin}_{\mathbf{x}} \Psi(\mathbf{x}, \mathbf{y}^{k+1}),$$

$$(7) \quad \mathbf{y}^{k+1} = \operatorname{argmin}_{\mathbf{y}} \Psi(\mathbf{x}^{k+1}, \mathbf{y}),$$

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

The AM algorithm 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 have 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 they usually cannot be solved exactly. Very recently, [12] suggested overcoming this difficulty by solving both iterations of AM *approximately*. More precisely, since each iteration of AM consists of minimizing the sum of a smooth function and 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  $\Psi$  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 [20] to solve ptychographic diffraction imaging problems, and in [33] for another variant which was used in the context of sparse blind convolution.

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 (6) is solved approximately through one shot of a proximal gradient step, i.e.,

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

while the  $\mathbf{y}$ -step is solved exactly through the global optimality condition for (7). For ease of reference, we call the resulting algorithm semiproximal-alternating (SPA).

Before presenting the algorithm, it will be convenient to recall the definition of the Moreau proximal map [30], to rewrite the above step compactly. Given a proper, closed, and convex function  $\varphi : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ , the *proximal map* associated with  $\varphi$  is uniquely defined by

$$(8) \quad \operatorname{prox}_{\varphi}(\mathbf{x}) := \operatorname{argmin} \left\{ \varphi(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 : \mathbf{u} \in \mathbb{R}^m \right\}.$$

For a recent and comprehensive review on proximal methods, we refer the reader to [31].

To solve (M) we thus propose the following scheme; see Assumption 2 for a definition of  $L_1(\mathbf{y})$ .

**Algorithm SPA**

- (1) Initialization: Start with any  $(\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^n \times \mathbb{R}^p$ .
- (2) For each  $k = 1, 2, \dots$  generate a sequence  $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k \in \mathbb{N}}$  as follows:
  - (2.1) Set  $L_k := L_1(\mathbf{y}^k)$  and compute

$$(9) \quad \mathbf{x}^{k+1} = \operatorname{prox}_{\frac{1}{L_k} F} \left( \mathbf{x}^k - \frac{1}{L_k} \nabla_{\mathbf{x}} H(\mathbf{x}^k, \mathbf{y}^k) \right).$$

- (2.2) Solve the strongly convex minimization problem:

$$(10) \quad \mathbf{y}^{k+1} = \operatorname{argmin}_{\mathbf{y} \in \mathbb{R}^p} \{ G(\mathbf{y}) + H(\mathbf{x}^{k+1}, \mathbf{y}) \}.$$

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

*Assumption 2.*

- (i) For any fixed  $\mathbf{y}$ , the function  $\mathbf{x} \rightarrow H(\mathbf{x}, \mathbf{y})$  is  $C_{L_1(\mathbf{y})}^{1,1}$ ; namely, the partial gradient  $\nabla_{\mathbf{x}}H(\cdot, \mathbf{y})$  is globally Lipschitz with constant  $L_1(\mathbf{y})$ , that is,

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

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

- (ii) The gradient  $\nabla H$  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 4.* As we shall see below,  $L_1(\mathbf{y})$  is explicitly available for the RSTLS problem, and we can set the parameter  $L_k = L_1(\mathbf{y}^k)$ . 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 presenting the algorithm for RSTLS, let us first verify that Assumption 2 also holds for RSTLS, which corresponds to the choice of  $H$  and  $G$  as given in (5). A simple computation then shows that

$$(11) \quad \nabla_{\mathbf{x}}H(\mathbf{x}, \mathbf{y}) = \frac{2}{\sigma_w^2} \left( \mathbf{A} + \sum_{i=1}^p y_i \mathbf{A}_i \right)^T \left( \left( \mathbf{A} + \sum_{i=1}^p y_i \mathbf{A}_i \right) \mathbf{x} - \mathbf{b} \right),$$

$$(12) \quad \nabla_{\mathbf{y}}H(\mathbf{x}, \mathbf{y}) = 2 \left( \frac{1}{\sigma_e^2} \mathbf{I}_{p \times p} + \frac{1}{\sigma_w^2} \mathbf{B}(\mathbf{x})^T \mathbf{B}(\mathbf{x}) \right) \mathbf{y} + \frac{1}{\sigma_w^2} \mathbf{B}(\mathbf{x})^T (\mathbf{A}\mathbf{x} - \mathbf{b}),$$

where  $\mathbf{I}_p$  stands for the usual  $p \times p$  identity matrix and we define

$$(13) \quad \mathbf{B}(\mathbf{x}) := (\mathbf{A}_1\mathbf{x}, \mathbf{A}_2\mathbf{x}, \dots, \mathbf{A}_p\mathbf{x}).$$

Now, it is easy to see that Assumptions 2(ii) and (iii) hold for the RSTLS problem. For the former, this is immediate with  $H$  defined in (5), while the latter follows since  $\Psi$  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 (11) that the function  $H(\cdot, \mathbf{y})$  admits a Lipschitz continuous partial gradient  $\nabla_{\mathbf{x}}H(\cdot, \mathbf{y})$  with constant  $L_1(\mathbf{y}) = (2/\sigma_w^2) \|\mathbf{A} + \sum_{i=1}^p y_i \mathbf{A}_i\|^2$ . Moreover, it follows by the continuity of  $L_1(\cdot)$  that the modulus  $L_1(\mathbf{y})$  is bounded from below and from above over compact sets, and hence Assumption 2(i) holds.

*Remark 5.* It should be noted that the computation of  $L_1(\mathbf{y})$  requires finding 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 be handled efficiently. For example, the operator represents a blurring operation with a point spread function (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 [19]. In other cases, as mentioned in Remark 4, 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 (5), 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  $\{(\mathbf{x}^k, \mathbf{y}^k)\}_{k \in \mathbb{N}}$  as follows:

- (a) Compute  $\mathbf{x}^{k+1}$  via (9), with  $H$  defined in (5) and with  $L_k = (2/\sigma_w^2) \|\mathbf{A} + \sum_{i=1}^p y_i^k \mathbf{A}_i\|^2$ .
- (b) Compute  $\mathbf{y}^{k+1}$  via (14) below.

The first step of the algorithm reduces to computing 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 TV function, we will use the recent efficient scheme FDPG of [7]; see more details in the numerical section (section 5) below.

Now, let us derive the second step, (b). Writing the optimality condition for the second step of (10) consists of finding the unique  $\mathbf{y}^{k+1}$  which solves the equation  $\nabla_{\mathbf{y}} H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + (1/\sigma_e^2) \mathbf{y}^{k+1} = 0$ . Using (12) and defining  $\mathbf{B}_k := \mathbf{B}(\mathbf{x}^k)$ , this reduces to solving a linear system of equations of dimension  $p \times p$  given explicitly by

$$(14) \quad \mathbf{y}^{k+1} = -\sigma_e^2 (\sigma_w^2 \mathbf{I}_{p \times p} + \sigma_e^2 \mathbf{B}_{k+1}^T \mathbf{B}_{k+1})^{-1} \mathbf{B}_{k+1}^T (\mathbf{A} \mathbf{x}^{k+1} - \mathbf{b}).$$

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 from [12] under an abstract convergence result which will then be applied to prove the convergence of SPA.

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

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

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

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

**DEFINITION 6.** Let  $\varphi : \mathbb{R}^d \rightarrow (-\infty, +\infty]$  be a proper and lsc function. A sequence  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  is called a gradient-like descent sequence for  $\varphi$  if for  $k \in \mathbb{N}$  the following two conditions hold:

(C1) *Sufficient decrease property: There exists a positive scalar  $\rho_1$  such that*

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

(C2) *A subgradient<sup>2</sup> lower bound for the iterates gap:*

- $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  is bounded.
- There exists a positive scalar  $\rho_2$  such that

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

*Remark 7.* We note that the assumption that  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  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

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

The two conditions (C1) and (C2) defining a gradient-like descent sequence for a given  $\varphi$  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  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  is a critical point of  $\varphi$ . More precisely, from [12, Lemma 5 and Remark 5] we have the following result.

**LEMMA 8.** *If  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  is a gradient-like descent sequence for a given function  $\varphi$ , which is lsc and proper on  $\mathbb{R}^d$ , then  $\omega(\mathbf{z}^0)$  is a nonempty, compact, and connected set, and we have*

$$\lim_{k \rightarrow \infty} \text{dist}(\mathbf{z}^k, \omega(\mathbf{z}^0)) = 0.$$

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  $\varphi$ . This can be achieved by imposing an additional assumption on the class of functions  $\varphi$ : it must satisfy the so-called Kurdyka–Łojasiewicz (KL) property [22, 24]. 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  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$ , 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  $\varphi$  which is *semialgebraic* satisfies the KL property at any point in  $\text{dom } \varphi$ ; see also the appendix. We can now conveniently summarize the general methodology and convergence results of [12] in the following abstract convergence result.

**THEOREM 9.** *Let  $\varphi : \mathbb{R}^d \rightarrow (-\infty, \infty]$  be a proper, lsc, and semialgebraic function with  $\inf \varphi > -\infty$ , and assume that  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  is a gradient-like descent sequence for  $\varphi$ . If  $\omega(\mathbf{z}^0) \subset \text{crit } \varphi$ , then the sequence  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  converges to a critical point  $\mathbf{z}^*$  of  $\varphi$ .*

*Remark 10.* Under the premises of this theorem, it is also possible to derive a rate of convergence result for the sequence  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  of the form  $\|\mathbf{z}^k - \mathbf{z}^*\| \leq C k^{-\gamma}$ ,

<sup>2</sup>Here,  $\partial\varphi$  stands for the limiting subdifferential of  $\varphi$  [35], which in the convex case reduces to the usual subgradient map [34]. For the reader's convenience, more details are given in the appendix.

for some positive constant  $C$  and where  $\gamma > 0$  is the so-called KL exponent; see [2] for details.

Semialgebraic functions abound in applications and include a broad list of functions arising in many optimization models. Moreover, sum and composition of semialgebraic functions are also semialgebraic. 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 semialgebraic sets and functions are also given in the appendix.

For our RSTLS model, it is immediate to see that both  $(1/\sigma_e^2) \|\mathbf{y}\|^2$  and the quadratic function  $H(\mathbf{x}, \mathbf{y})$  (cf. (5)) are semialgebraic, and hence with any choice of  $F(\cdot)$  semialgebraic the RSTLS problem thus admits of a semialgebraic objective function. Therefore, the global convergence result established below for SPA clearly applies to the RSTLS algorithm described in section 3.3. It should be noted that in the RSTLS case, the generated sequence is bounded (see condition (C2) in Definition 6) 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), p. 482]).

**4.2. Convergence of SPA.** Equipped with the abstract Theorem 9, our main objective is now to prove the global convergence of the sequence generated by SPA to a critical point of  $\Psi$ . 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, p. 190]).

LEMMA 11 (sufficient decrease property). *Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  be a continuously differentiable function with gradient  $\nabla h$  assumed  $L_h$ -Lipschitz continuous, and let  $g : \mathbb{R}^d \rightarrow (-\infty, +\infty]$  be a proper, lsc, and convex function. Then, for any  $\mathbf{u} \in \text{dom } g$  and*

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

we have

$$(15) \quad h(\mathbf{u}^+) + g(\mathbf{u}^+) \leq h(\mathbf{u}) + g(\mathbf{u}) - \left( t - \frac{L}{2} \right) \|\mathbf{u}^+ - \mathbf{u}\|^2,$$

where  $L \geq L_h$ .

The next result recalls useful basic properties of subdifferential maps (see [35]).

PROPOSITION 12 (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*

$$(16) \quad \partial\Psi(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}}H(\mathbf{x}, \mathbf{y}) + \partial F(\mathbf{x}), \nabla_{\mathbf{y}}H(\mathbf{x}, \mathbf{y}) + \partial G(\mathbf{y})) = (\partial_{\mathbf{x}}\Psi(\mathbf{x}, \mathbf{y}), \partial_{\mathbf{y}}\Psi(\mathbf{x}, \mathbf{y})).$$

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

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

$$\mathbf{z}^k := (\mathbf{x}^k, \mathbf{y}^k) \quad \forall k \geq 0$$

for the sequence generated by SPA.

**THEOREM 13** (global convergence). *Suppose that Assumptions 1 and 2 hold, and assume that  $\Psi$  is semialgebraic. Let  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  be a sequence generated by SPA which is assumed to be bounded. Then, the sequence  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  converges to a critical point  $\mathbf{z}^*$  of  $\Psi$ .*

*Proof.* We invoke Theorem 9. For that purpose the proof is divided into two main parts. We first need to show that the sequence  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$  generated by SPA is a gradient-like descent with respect to  $\Psi$ ; that is, according to Definition 6, we need to show that conditions (C1) and (C2) hold. In the second part we then have to prove that  $\omega(\mathbf{z}^0) \subset \text{crit } \Psi$ .

*Step 1. Proving that conditions (C1)–(C2) hold.* Thanks to Assumptions 1(ii) and 2(i), we can apply Lemma 11 with  $h(\cdot) := H(\cdot, y^k)$ ,  $L_h := L_1(y^k) = L_k \geq \underline{L} > 0$  (since here  $\mathbf{y}^k$  is assumed bounded), and the convex function  $g(\cdot) := F(\cdot)$  to get from (9) that, for any  $k \geq 0$ ,

$$(17) \quad H(\mathbf{x}^{k+1}, \mathbf{y}^k) + F(\mathbf{x}^{k+1}) \leq H(\mathbf{x}^k, \mathbf{y}^k) + F(\mathbf{x}^k) - \frac{L}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2.$$

Now, from Assumptions 1(ii) and (iii), it follows that  $G(\cdot) + H(\mathbf{x}^{k+1}, \cdot)$  is  $\sigma$ -strongly convex, and hence by the well-known subgradient inequality for a  $\sigma$ -strongly convex function [35, Chapter 12], we obtain for any  $k \geq 0$

$$(18) \quad \begin{aligned} G(\mathbf{y}^k) + H(\mathbf{x}^{k+1}, \mathbf{y}^k) &\geq G(\mathbf{y}^{k+1}) + H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) \\ &\quad + \langle \eta^{k+1} + \nabla_{\mathbf{y}} H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}), \mathbf{y}^k - \mathbf{y}^{k+1} \rangle + \frac{\sigma}{2} \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 \\ &= G(\mathbf{y}^{k+1}) + H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + \frac{\sigma}{2} \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2, \end{aligned}$$

where  $\eta^{k+1} \in \partial G(\mathbf{y}^{k+1})$  and the last equality follows immediately from the optimality condition for (10). Combining (17) and (18) 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

$$(19) \quad \Psi(\mathbf{z}^k) - \Psi(\mathbf{z}^{k+1}) \geq \frac{L}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 + \frac{\sigma}{2} \|\mathbf{y}^{k+1} - \mathbf{y}^k\|^2 \geq \frac{\rho_1}{2} \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2,$$

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

$$(20) \quad \lim_{k \rightarrow \infty} (\mathbf{z}^{k+1} - \mathbf{z}^k) = \mathbf{0}.$$

Indeed, from (19), the sequence  $\{\Psi(\mathbf{z}^k)\}_{k \geq 0}$  is nonincreasing, and since  $\Psi$  is assumed to be bounded from below (see Assumption 2(iii)), it converges to some real number  $\underline{\Psi}$ . Summing (19) from  $k = 0$  to  $N - 1$ , we thus get

$$(21) \quad \sum_{k=0}^{N-1} \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \leq \frac{2}{\rho_1} (\Psi(\mathbf{z}^0) - \Psi(\mathbf{z}^N)) \leq \frac{2}{\rho_1} (\Psi(\mathbf{z}^0) - \underline{\Psi}),$$

and taking the limit as  $N \rightarrow \infty$ , the claim (20) follows.

Now, we prove that condition (C2) holds for the sequence  $\{\mathbf{z}^k\}_{k \in \mathbb{N}}$ . Writing the optimality conditions for the convex iterations of SPA defined in (9) and (10), we have for any  $k \geq 0$

$$(22) \quad \nabla_{\mathbf{x}} H(\mathbf{x}^k, \mathbf{y}^k) + L_k(\mathbf{x}^{k+1} - \mathbf{x}^k) + \mathbf{u}^{k+1} = \mathbf{0}, \quad \mathbf{u}^{k+1} \in \partial F(\mathbf{x}^{k+1}),$$

$$(23) \quad \nabla_{\mathbf{y}} H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + \mathbf{v}^{k+1} = \mathbf{0}, \quad \mathbf{v}^{k+1} \in \partial G(\mathbf{y}^{k+1}).$$

On the other hand, using Proposition 12, we have

$$(24) \quad \partial\Psi(\mathbf{z}^k) = (\nabla_{\mathbf{x}}H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + \partial F(\mathbf{x}^{k+1}), \nabla_{\mathbf{y}}H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) + \partial G(\mathbf{y}^{k+1})).$$

Therefore, defining the quantity

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

and using (24) together with (22) and (23), we obtain

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

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

$$(25) \quad \begin{aligned} \|\mathbf{w}^{k+1}\| &= \|(\xi_{\mathbf{x}}^{k+1}, \mathbf{0})\| \leq L_k \|\mathbf{x}^k - \mathbf{x}^{k+1}\| + \|\nabla_{\mathbf{x}}H(\mathbf{x}^{k+1}, \mathbf{y}^{k+1}) - \nabla_{\mathbf{x}}H(\mathbf{x}^k, \mathbf{y}^k)\| \\ &\leq \bar{L} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| + M(\|\mathbf{x}^{k+1} - \mathbf{x}^k\| + \|\mathbf{y}^{k+1} - \mathbf{y}^k\|) \\ &\leq \bar{L} \|\mathbf{x}^{k+1} - \mathbf{x}^k\| + \sqrt{2}M \|\mathbf{z}^{k+1} - \mathbf{z}^k\| \\ &\leq (\sqrt{2}M + \bar{L}) \|\mathbf{z}^{k+1} - \mathbf{z}^k\|, \end{aligned}$$

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

*Step 2. Proving that  $\omega(\mathbf{z}^0) \subset \text{crit } \Psi$ .* 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  $(\mathbf{x}^{k_q}, \mathbf{y}^{k_q}) \rightarrow (\mathbf{x}^*, \mathbf{y}^*)$  as  $q \rightarrow \infty$ . We need to show that

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

Following the same proof as [12, Lemma 5(i), p. 476] for the iterative step (9), since both sequences  $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$  and  $L_k$  are bounded,  $\nabla H$  is continuous, and from (20) the distance between two successive iterates tends to zero, it follows that  $F(\mathbf{x}^{k_q})$  tends to  $F(\mathbf{x}^*)$  as  $q \rightarrow \infty$ . On the other hand, from the iterative step (10), we have for all  $k \geq 0$  that

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

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

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

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

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

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 (20) and (25),  $\|\mathbf{w}^{k+1}\| \rightarrow 0$  as  $k \rightarrow \infty$ , then recalling the closedness property of the map  $\partial\Psi$  (cf. Proposition 12), 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  $\Psi$ .  $\square$

**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 ( $21 \times 21$ ), while in the second subsection we present results for large images ( $256 \times 256$ ). The motivation for this partition is the fact that on small images we can compare our algorithm to the RSTLN method [32]. 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 the following hold:

- $\mathbf{b}$  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 PSF of size  $q \times q$  and standard deviation  $\gamma$  (given by the MATLAB command `psfGauss([q,q],\gamma)`), by using the MATLAB command `imfilter`. After that we add a zero-mean white Gaussian noise with standard deviation  $\sigma_w$ .
- 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  $\mathbf{A}_i$  have been subjected to noise which has been independently generated from a zero-mean normal distribution with standard deviation  $\sigma_e$ .
- 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 [19].
- $f : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  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 [36, 32]). This method was designed to solve the following class of RSTLS problems:

$$(26) \quad \min_{\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^p} \left\| \begin{pmatrix} (1/\sigma_w) \left( (\mathbf{A} + \sum_{i=1}^p y_i \mathbf{A}_i) \mathbf{x} - \mathbf{b} \right) \\ (1/\sigma_e) \mathbf{y} \\ \lambda L\mathbf{x} \end{pmatrix} \right\|_{\mu},$$

where  $\mu = 1, 2, \infty$ . It should be noted that when  $\mu = 2$ , this formulation is equivalent to (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), only  $\ell_2$  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

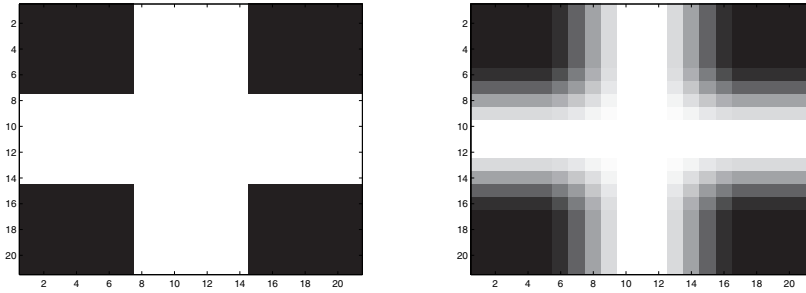


FIG. 1. The plus test image: Original image (left) and blurred and noised image (right).

fidelity and for the regularizer. On the other hand, in (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 an  $\ell_2$ -norm and the regularizer is given by an  $\ell_1$ -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:

$$(27) \quad \min_{\mathbf{x}, \mathbf{y}} \left\{ \left\| \begin{pmatrix} \mathbf{A} + \sum_{i=1}^p y_i^k \mathbf{A}_i & \mathbf{B}_k \\ \mathbf{0}_{p \times n} & \frac{\sigma_w}{\sigma_e} \mathbf{I}_{p \times p} \end{pmatrix} \begin{pmatrix} \mathbf{x} - \mathbf{x}^k \\ \mathbf{y} - \mathbf{y}^k \end{pmatrix} \right\|^2 + \lambda f(L\mathbf{x}) \right\},$$

where  $(\mathbf{x}^k, \mathbf{y}^k)$  is the given iterate and  $\mathbf{B}_k$  is the  $m \times p$  matrix defined in (13), that is,  $\mathbf{B}_k := [\mathbf{A}_1 \mathbf{x}^k, \mathbf{A}_2 \mathbf{x}^k, \dots, \mathbf{A}_p \mathbf{x}^k]$ .

It should be noted that this is not exactly the problem that was solved in [32], because here we apply the Gauss–Newton method on the model (1) and not on the model (26), meaning we look at general regularizers. Here we use CVX [17] in order to solve these subproblems at each iteration.

Here we regularize the problem with the  $\ell_1$ -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

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

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

*Example 14.* Consider the  $21 \times 21$  “plus” image given in Figure 1(left). The “true” image goes through a Gaussian blur PSF of size  $5 \times 5$  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(right) for the blurred and noised image. 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 six structure matrices, i.e.,  $p = 6$ , and each matrix is of size  $5 \times 5$ . Specifically, the

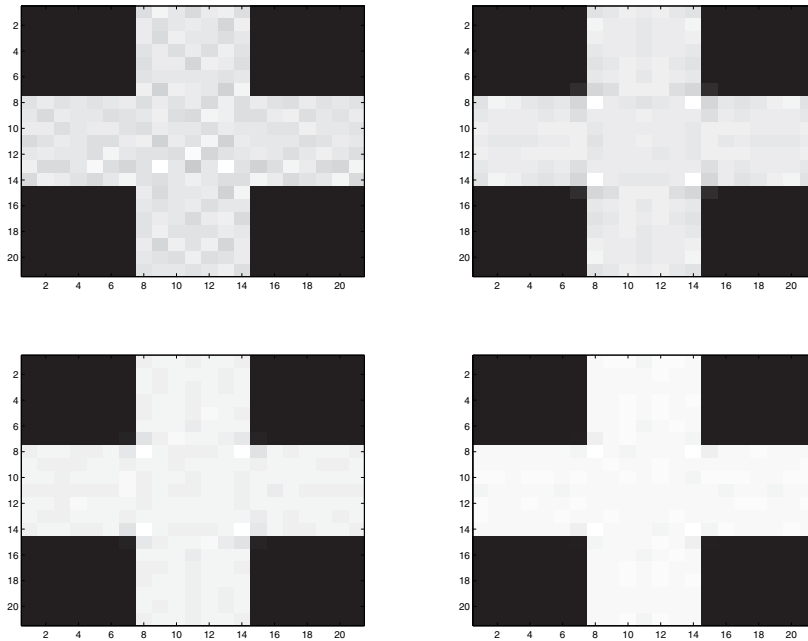


FIG. 2. The plus test image: RSTLN solution (top-left); SPA solution, 972 iterations (top-right); SPA solution, 2500 iterations (bottom-right); SPA solution, 4500 iterations (bottom-left).

structure matrices are

$$\begin{aligned}
 A_1 &= \begin{bmatrix} 0.0352 & 0 & 0 & 0 & 0.0352 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0352 & 0 & 0 & 0 & 0.0352 \end{bmatrix}, & A_2 &= \begin{bmatrix} 0 & 0.0387 & 0 & 0.0387 & 0 \\ 0.0387 & 0 & 0 & 0 & 0.0387 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0387 & 0 & 0 & 0 & 0.0387 \\ 0 & 0.0387 & 0 & 0.0387 & 0 \end{bmatrix}, \\
 A_3 &= \begin{bmatrix} 0 & 0 & 0.0399 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.0399 & 0 & 0 & 0 & 0.0399 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0399 & 0 & 0 \end{bmatrix}, & A_4 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0425 & 0 & 0.0425 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0425 & 0 & 0.0425 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \\
 A_5 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0438 & 0 & 0 \\ 0 & 0.0438 & 0 & 0.0438 & 0 \\ 0 & 0 & 0.0438 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, & A_6 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0452 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.
 \end{aligned}$$

In this experiment we used periodic boundary conditions and initialized 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 eight outer iterations (of the Gauss–Newton method) produced the solution given in Figure 2(top-left). 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(top-right)). 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 row were obtained by the SPA algorithm after 2500 (left) and 4500 (right) 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 (27) at each iteration of the Gauss–Newton method.

We made a more extensive set of tests in which we solved 10 problems—each corresponding to a realization of  $\sigma_w$  and  $\sigma_e$  for different values  $10^{-2}$ ,  $10^{-3}$ , and  $10^{-4}$ . In Table 1, for each value of the standard deviation and for each choice of an algorithm (SPA and RSTLN), we indicate the average CPU time of each algorithm (out of 100) with tolerance parameter  $\varepsilon > 0$  for  $\varepsilon = 10^{-3}, 10^{-4}, 10^{-5}$  after 100 iterations. The parameter  $\varepsilon$  is used in the stopping criterion, which is  $\|\mathbf{x}^{N+1} - \mathbf{x}^N\| / \|\mathbf{x}^N\| \leq \varepsilon$  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 as fast). 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 ran for more than 230 seconds (for  $\sigma_w = \sigma_e = 10^{-2}$ ), while the SPA required only about 11 seconds.

**5.2. Large scale image.** In this section we present numerical examples on large scale images. We consider the  $256 \times 256$  cameraman 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.

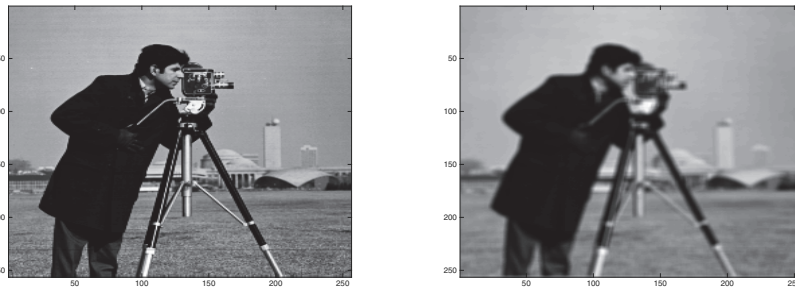


FIG. 3. The cameraman test image: Original image (left) and blurred and noised image (right).

*Example 15.* The cameraman test image went through a Gaussian blur PSF of size  $5 \times 5$  and standard deviation 4, and then the observed image was 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(left)

TABLE 1  
 Average CPU time (out of 100) in the first line, and maximum and minimum CPU times (out of 100) in the second line, for which an  $\epsilon$ -optimal solution was reached.

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

for the original image and Figure 3(right) for the blurred and noised image. We also added 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.

In this experiment we used periodic boundary conditions and considered the TV-based regularization, that is,  $F(\mathbf{x}) = \lambda TV_1(\mathbf{x})$ , where  $\lambda = 10^{-3}$  (see Example 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.

Figure 4 shows data for 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.

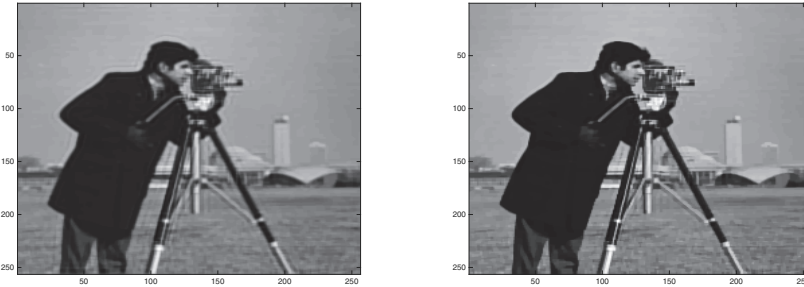


FIG. 4. The cameraman test image: SPA solution after 50 (left) and 200 (right) iterations.

## Appendix A. Appendix.

**A.1. Nonsmooth calculus.** Let us recall a few definitions concerning subdifferential calculus (see, for instance, [35]). Recall that for a proper and lsc function  $\varphi : \mathbb{R}^d \rightarrow (-\infty, +\infty]$  the domain of  $\varphi$  is defined via  $\text{dom } \varphi := \{\mathbf{x} \in \mathbb{R}^d : \varphi(\mathbf{x}) < +\infty\}$ .

DEFINITION 16 (subdifferentials). Let  $\varphi : \mathbb{R}^d \rightarrow (-\infty, +\infty]$  be a proper and lsc function.

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

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

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

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

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

Note that in this nonsmooth and nonconvex context, the well-known Fermat's rule remains barely changed. It formulates as "if  $\mathbf{x} \in \mathbb{R}^d$  is a local minimizer of  $\varphi$ , 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  $\varphi$  is denoted by  $\text{crit } \varphi$ .

**A.2. Semialgebraic sets and functions.** We recall here basic definitions and properties; see, e.g., [12] and references therein for more details.

DEFINITION 17 (semialgebraic sets and functions).

- (i) A subset  $S$  of  $\mathbb{R}^d$  is a real semialgebraic set if there exists a finite number of real polynomial functions  $g_{ij}, h_{ij} : \mathbb{R}^d \rightarrow \mathbb{R}$  such that

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

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

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

is a semialgebraic subset of  $\mathbb{R}^{d+1}$ .

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

THEOREM 18. Let  $\varphi : \mathbb{R}^d \rightarrow (-\infty, +\infty]$  be a proper and lsc function. If  $\varphi$  is semialgebraic, then it satisfies the KL property at any point of  $\text{dom } \varphi$ .

The class of semialgebraic sets is stable under the following operations: finite unions, finite intersections, complementation, and Cartesian products. There is a broad class of semialgebraic functions arising in optimization: real polynomial functions, indicator functions of semialgebraic sets, finite sums product and composition of semialgebraic 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 the 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 (1991), pp. 1070–1087.
- [2] H. ATTOUCH AND J. BOLTE, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Program., 116 (2009), pp. 5–16.
- [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 (2015), pp. 185–209.
- [4] A. BECK AND A. BEN-TAL, *On the solution of the Tikhonov regularization of the total least squares problem*, SIAM J. Optim., 17 (2006), pp. 98–118.
- [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 (2008), pp. 419–443.
- [6] A. BECK AND M. TEBoulLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.
- [7] A. BECK AND M. TEBoulLE, *A fast dual proximal gradient algorithm for convex minimization and applications*, Oper. Res. Lett., 42 (2014), pp. 1–6.
- [8] D. P. BERTSEKAS, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 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), 2010, pp. 1673–1676.
- [10] J. BOLTE, A. DANILIDIS, AND A. LEWIS, *The Lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. Optim., 17 (2007), pp. 1205–1223.
- [11] J. BOLTE, A. DANILIDIS, O. LEY, AND L. MAZET, *Characterizations of Lojasiewicz inequalities: Subgradient flows, talweg, convexity*, Trans. Amer. Math. Soc., 362 (2010), pp. 3319–3363.
- [12] J. BOLTE, S. SABACH, AND M. TEBoulLE, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program. Ser. A, 146 (2014), pp. 459–494.
- [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 (2011), pp. 1–122.
- [14] E. CHOUZENOUX, J.-C. PESQUET, AND A. REPETTI, *A block coordinate variable metric forward-backward algorithm*, J. Global Optim., (2016), doi:10.1007/s10898-016-0405-9.
- [15] G. H. GOLUB, P. C. HANSEN, AND D. P. O’LEARY, *Tikhonov regularization and total least squares*, SIAM J. Matrix Anal. Appl., 21 (1999), pp. 185–194.
- [16] G. H. GOLUB AND C. F. VAN LOAN, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.
- [17] M. GRANT AND S. BOYD, *CVX: Matlab software for disciplined convex programming*, v. 2.0 beta, 2013, <http://cvxr.com/cvx>.
- [18] P. C. HANSEN, *Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems*, Numer. Algorithms, 6 (1994), pp. 1–35.
- [19] P. C. HANSEN, J. G. NAGY, AND D. P. O’LEARY, *Deblurring Images: Matrices, Spectra, and Filtering*, vol. 3 of Fundamentals of Algorithms, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2006.
- [20] 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 (2015), pp. 426–457.
- [21] S. VAN HUFFEL AND J. VANDEWALLE, *The Total Least Squares Problem: Computational Aspects and Analysis*, Frontiers in Appl. Math. 9, SIAM, Philadelphia, 1991.
- [22] K. KURDYKA, *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier (Grenoble), 48 (1998), pp. 769–783.
- [23] P. LEMMERLING AND S. V. HUFFEL, *Analysis of the structured total least squares problem for Hankel/Toeplitz matrices*, Numer. Algorithms, 27 (2001), pp. 89–114.
- [24] S. LOJASIEWICZ, *Une propriété topologique des sous-ensembles analytiques réels*, in Les Équations aux Dérivées Partielles (Paris, 1962), Éditions du Centre National de la Recherche Scientifique, Paris, 1963, pp. 87–89.
- [25] I. MARKOVSKY, S. V. HUFFEL, AND R. PINTELON, *Block-Toeplitz/Hankel structured total least squares*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1083–1099.
- [26] 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 (2000), pp. 533–553.
- [27] B. D. MOOR, *Structured total least squares and  $\{L_2\}$  approximation problems*, Linear Algebra Appl., 188-189 (1993), pp. 163–205.
- [28] B. D. MOOR, *Total least squares for affinely structured matrices and the noisy realization problem*, IEEE Trans. Signal Processing, 42 (1994), pp. 3104–3113.
- [29] J. J. MORE, *Generalizations of the trust region subproblem*, Optim. Methods Softw., 2 (1993), pp. 189–209.
- [30] J. J. MOREAU, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
- [31] N. PARIKH AND S. BOYD, *Proximal algorithms*, Found. Trends Optim., 1 (2014), pp. 127–239.
- [32] A. PRUESSNER AND D. P. O’LEARY, *Blind deconvolution using a regularized structured total least norm algorithm*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 1018–1037.
- [33] A. REPETTI, M. Q. PHAM, L. DUVAL, E. CHOUZENOUX, AND J.-C. PESQUET, *Euclid in a taxicab: Sparse blind deconvolution with smoothed  $\ell_1/\ell_2$  regularization*, IEEE Signal Process. Lett., 22 (2015), pp. 539–543.
- [34] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Math. Ser. 28, Princeton University Press, Princeton, NJ, 1970.
- [35] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1998.

- [36] J. B. ROSEN, H. PARK, AND J. GLICK, *Total least norm formulation and solution for structured problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 110–126.
- [37] A. N. TIKHONOV, *Solution of incorrectly formulated problems and the regularization method*, Soviet. Math. Dokl., 5 (1963), pp. 1035–1038.
- [38] A. N. TIKHONOV AND V. Y. ARSEININ, *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.).
- [39] P. TSENG, *Convergence of a block coordinate descent method for nondifferentiable minimization*, J. Optim. Theory Appl., 109 (2001), pp. 475–494.