On some steplength approaches for proximal algorithms

Federica Porta[a,*], Ignace Loris[b]

[a] *Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia*
[b] *Département de Mathématique, Université Libre de Bruxelles.*

Abstract

We discuss a number of novel steplength selection schemes for proximal-based convex optimization algorithms. In particular, we consider the problem where the Lipschitz constant of the gradient of the smooth part of the objective function is unknown. We generalize two optimization algorithms of Khobotov type and prove convergence. We also take into account possible inaccurate computation of the proximal operator of the non-smooth part of the objective function. Secondly, we show convergence of an iterative algorithm with Armijo-type steplength rule, and discuss its use with an approximate computation of the proximal operator. Numerical experiments show the efficiency of the methods in comparison to some existing schemes.

Keywords: proximal algorithms, steplength selection, non-smooth optimization, signal recovering
2010 MSC: 65J22, 65K10, 68U10

1. Introduction

In this paper we consider the problem of minimizing the sum of two given functions

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \mathcal{F}(\boldsymbol{x}) \equiv f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{1}$$

where $f : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a convex, continuously differentiable function and $g : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ is an extended-value convex function, possibly including constraints on the unknown.

The minimization problem (1) has been handled by several algorithms especially tailored to deal with a non-differentiable function g. In particular, numerical schemes known in the literature as *proximal gradient methods* have earned a great popularity in the last years. They find a very general applicability in problems concerning with large or high-dimensional datasets from several

*Corresponding author
Email addresses: federica.porta@unimore.it (Federica Porta), igloris@ulb.ac.be (Ignace Loris)

scientific areas, like compressed sensing, machine learning and signal processing (see for example [1, 2, 3, 4]).

In this paper we discuss new techniques to select the steplength in the proximal gradient methods, without the assumption of knowing the Lipschitz constant of the gradient of the smooth part of the objective function. We start our analysis from two approaches developed in the constrained differentiable optimization context: the Khobotov extra-gradient method [5, 6] and the gradient projection method along the feasible directions [7]. Due to the non-smooth term g in the function (1), a possible generalization of these algorithms has to account for the presence of the proximal operator of this term instead of a projection onto a suitable set.

We study several extensions of constrained optimization algorithms of Khobotov type. In particular, we propose an extension of Khobotov's original scheme [5] to the more general proximal case. We prove convergence, even when the proximal operators cannot be computed exactly. We also consider the saddle-point formulation for the minimization problem (1). The so-called Alternating Extragradient Method (AEM) [8] is a variant of Khobotov's method for constrained smooth saddle-point problems. We propose a generalization of the AEM algorithm for a general (not necessarily smooth) saddle-point problem. This extension is again achieved through the use of the proximal operator of the non-smooth part of the objective function. Again, none of these algorithms require any knowledge of the Lipschitz constant of the gradient of the smooth part of the objective function. Such a problem is also recently discussed in [9].

Secondly, and following the basic idea behind the gradient projection methods, we suggest an iterative proximal algorithm that exploits an Armijo-type steplength selection rule similar to [10]. A proof of convergence of the algorithm is provided. We also explore its use in case only an approximation for the required proximal operator is available.

Finally, in order to evaluate the effectiveness of the presented methods, we conduct a numerical study on some signal recovering test problems that can be modeled by equation (1): the performance of the discussed schemes is assessed through a comparison with some algorithms already known in the literature and designed to solve this type of problems.

Several problems arising from real-world applications [11, 12, 13] can be formalized through the mathematical model introduced in equation (1): the applications of this work will be focused on one-dimensional and two-dimensional signal restoration problems with data perturbed by Poisson noise [14, 15, 16]. Signal and image restoration consist in recovering an approximation of an object detected by an acquisition system, starting from the data provided by the instrument and a model representing the distortion occurring during the acquisition process itself. More precisely, the signal formation process is an inverse problem that can be formalized through a linear system $\boldsymbol{g} = H\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{\eta}$ where $\boldsymbol{g} \in \mathbb{R}^M$ is the observed data, $\boldsymbol{x} \in \mathbb{R}^N$ represents an ideal, undistorted object to be recovered, $H \in \mathbb{R}^{M\times N}$ is a typically ill-conditioned matrix describing the acquisition instrument effect, $\boldsymbol{b} \in \mathbb{R}^M$ expresses a non-negative constant background radiation and $\boldsymbol{\eta} \in \mathbb{R}^M$ is the noise corrupting the data. In this paper we will work under the hypothesis of having non-negative signals, therefore we will take into account this type of constraint in the problem formulation. In the Bayesian approach [17, 18], the approximated restored signal is found by solving the following optimization problem

$$\min_{\boldsymbol{x}\geq\boldsymbol{0}} J_0(\boldsymbol{x}) + \mu J_R(\boldsymbol{x}), \tag{2}$$

where $J_0 : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a continuously differentiable function measuring the distance between the model and the data, $J_R : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a regularization term adding a priori information on the solution and μ is a positive parameter balancing the role of the two objective function components J_0 and J_R. When the data are affected by Poisson noise, the so-called Kullback-Leibler divergence is used to describe J_0:

$$J_0(\boldsymbol{x}) = \mathrm{KL}(\boldsymbol{x}) = \sum_{i=1}^{N} \left\{ \boldsymbol{g}_i \ln \frac{\boldsymbol{g}_i}{(H\boldsymbol{x}+\boldsymbol{b})_i} + (H\boldsymbol{x}+\boldsymbol{b})_i - \boldsymbol{g}_i \right\} \tag{3}$$

with $\boldsymbol{g}_i \ln(\boldsymbol{g}_i) = 0$ if $\boldsymbol{g}_i = 0$. As for the regularization term, we will consider properly chosen functionals that enforce a priori information depending on the features of the problem.

2. Mathematical tools

This section recalls some useful definitions and properties on proximal operators and describes a well-known proximal gradient method. For a more complete discussion of proximal operator methods we refer the reader to [4, 3, 19, 20]. In the following we consider convex function that are proper (nowhere equal to $-\infty$ and not identically equal to $+\infty$) and lower semi-continuous.

2.1. Proximal operators

The *proximal operator* $\mathbf{prox}_h : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ of a convex function $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ is defined as:

$$\mathbf{prox}_h(\boldsymbol{u}) = \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^N} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|^2 + h(\boldsymbol{x}).$$

We remark that if h is convex and closed then $\mathbf{prox}_h(\boldsymbol{u})$ exists and is unique for all $\boldsymbol{u} \in \mathbb{R}^N$.

Lemma 1 (Subgradient characterization)**.** *Let $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ be an extended-value function. The following characterization for the proximal operator of h holds true: $\boldsymbol{x} = \mathbf{prox}_h(\boldsymbol{u})$ if and only if $\boldsymbol{u}-\boldsymbol{x} \in \partial h(\boldsymbol{x})$ if and only if $h(\boldsymbol{z}) \geq h(\boldsymbol{x}) + \langle \boldsymbol{u}-\boldsymbol{x}, \boldsymbol{z}-\boldsymbol{x}\rangle$, $\forall \boldsymbol{z} \in \mathbb{R}^N$.*

Proof. See [3]. □

Remark 1. From lemma 1 and by setting $\boldsymbol{w} = \boldsymbol{u}-\boldsymbol{x}$, it follows that $\boldsymbol{w} \in \partial h(\boldsymbol{x})$ *iff* $\boldsymbol{x} = \mathbf{prox}_h(\boldsymbol{x}+\boldsymbol{w})$.

Remark 2. The minimizer $\hat{\boldsymbol{x}}$ of problem (1) is characterized by the inclusion $\mathbf{0} \in \nabla f(\hat{\boldsymbol{x}}) + \partial g(\hat{\boldsymbol{x}})$, or equivalently by the relations $\alpha \nabla f(\hat{\boldsymbol{x}}) + \boldsymbol{w} = 0$ and $\boldsymbol{w} \in \partial \alpha g(\hat{\boldsymbol{x}})$ with $\alpha > 0$. Using Remark 1, these can be rewritten as the single condition

$$\hat{\boldsymbol{x}} = \mathbf{prox}_{\alpha g}(\hat{\boldsymbol{x}} - \alpha \nabla f(\hat{\boldsymbol{x}})), \qquad \alpha > 0. \tag{4}$$

Lemma 2. *Let $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ be an extended-value convex function. The proximal operator of h is a Lipschitz continuous function with constant 1:*

$$\| \mathbf{prox}_h(\boldsymbol{u}) - \mathbf{prox}_h(\tilde{\boldsymbol{u}})\| \leq \|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|, \qquad \forall \boldsymbol{u}, \tilde{\boldsymbol{u}} \in \mathbb{R}^N. \tag{5}$$

Proof. See [3]. □

Remark 3. For fixed $\boldsymbol{u} \in \mathbb{R}^N$ and fixed function g, it can also be shown that $\mathbf{prox}_{\alpha g}(\boldsymbol{u})$ is a continuous functions of α, for $\alpha > 0$. It then follows from Lemma 2 that $\mathbf{prox}_{\alpha g}(\boldsymbol{x})$ is continuous in $(\alpha, \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^N$ and all $\alpha > 0$.

The *conjugate function* $h^* : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ of a convex function $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ is defined as

$$h^*(\boldsymbol{w}) = \sup_{\boldsymbol{x} \in \mathbb{R}^N} \langle \boldsymbol{w}, \boldsymbol{x} \rangle - h(\boldsymbol{x}).$$

Lemma 3 (Moreau decomposition)**.** *Given a convex function $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ and its conjugate $h^* : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$, their proximal operators are related by the identity:* $\mathbf{prox}_{h^*}(\boldsymbol{u}) + \mathbf{prox}_h(\boldsymbol{u}) = \boldsymbol{u}, \qquad \forall \boldsymbol{u} \in \mathbb{R}^N$.

Lemma 4. *Let $\boldsymbol{x}^+, \boldsymbol{x}^-$ and $\Delta \in \mathbb{R}^N$ and $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ be an extended-value convex function. If $\boldsymbol{x}^+ = \mathbf{prox}_h(\boldsymbol{x}^- + \Delta)$ then*

$$\|\boldsymbol{x}^+ - \boldsymbol{x}\| \leq \|\boldsymbol{x}^- - \boldsymbol{x}\|^2 - \|\boldsymbol{x}^+ - \boldsymbol{x}^-\|^2 + 2\langle \boldsymbol{x}^+ - \boldsymbol{x}, \Delta \rangle + 2h(\boldsymbol{x}) - 2h(\boldsymbol{x}^+), \quad \forall \boldsymbol{x} \in \mathbb{R}^N. \tag{6}$$

Proof. A simple application of the lemma 1 is sufficient to prove the required inequality. Indeed, if $\boldsymbol{x}^+ = \mathbf{prox}_h(\boldsymbol{x}^- + \Delta)$ then

$$\begin{aligned} h(\boldsymbol{x}) &\geq h(\boldsymbol{x}^+) + \langle \boldsymbol{x}^- - \boldsymbol{x}^+, \boldsymbol{x} - \boldsymbol{x}^+ \rangle + \langle \Delta, \boldsymbol{x} - \boldsymbol{x}^+ \rangle \\ &= h(\boldsymbol{x}^+) + \frac{1}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}^-\|^2 + \frac{1}{2}\|\boldsymbol{x}^+ - \boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{x}^- - \boldsymbol{x}\|^2 - \langle \boldsymbol{x}^+ - \boldsymbol{x}, \Delta \rangle \end{aligned}$$

which gives (6). □

Let us remark that it is possible to find an explicit expression for the proximal operator of some special functions. If $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ is the indicator function of a closed, non-empty and convex set C: the proximal operator of h reduces to the euclidean projection onto C: $\mathbf{prox}_h(\boldsymbol{x}) = \mathbb{P}_C(\boldsymbol{x}) = \operatorname{argmin}_{\boldsymbol{z} \in C} \|\boldsymbol{z} - \boldsymbol{x}\|^2$. When $h : \mathbb{R}^N \longrightarrow \overline{\mathbb{R}}$ is defined as $h(\boldsymbol{x}) = \lambda \|\boldsymbol{x}\|_1$ (with $\lambda \geq 0$), the proximal operator of h is the so-called soft-thresholding operator $\mathbf{prox}_h(\boldsymbol{x}) = S_\lambda(\boldsymbol{x})$, with

$$S_\lambda(x_i) = \begin{cases} x_i - \dfrac{x_i}{|x_i|}\lambda & \text{if} \quad |x_i| > \lambda \\ 0 & \text{if} \quad |x_i| \leq \lambda, \end{cases}$$

applied component-wise.

2.2. Proximal gradient method

A famous approach [21, 22, 3, 19, 23, 24] for solving the minimization problem (1) is based on the *proximal gradient method* which consists of a proximal step at a gradient point (see also equation (4)):

$$\boldsymbol{x}_{n+1} = \mathbf{prox}_{\alpha g}(\boldsymbol{x}_n - \alpha \nabla f(\boldsymbol{x}_n)) \tag{7}$$

where α is a suitable positive steplength. If the function f has a Lipschitz continuous gradient and the Lipschitz constant L is known, a classical choice for the steplength is

$$\alpha = \frac{1}{L} \tag{8}$$

(convergence can be shown for $0 < \alpha < 2/L$). However, the knowledge or the computation of the Lipschitz constant is not always evident. To overcome these difficulties, a backtracking steplength rule is proposed in [22].

Another possible formulation for the proximal gradient method is suggested in [19] and it can be described by the scheme:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \lambda_n \left(\mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)) - \boldsymbol{x}_n \right), \tag{9}$$

where $\{\alpha_n\}_{n\in\mathbb{N}}$ is a sequence in $]0, +\infty[$ verifying:

$$0 < \inf_{n\in\mathbb{N}} \alpha_n \le \sup_{n\in\mathbb{N}} \alpha_n < \frac{2}{L} \tag{10}$$

and $\{\lambda_n\}_{n\in\mathbb{N}}$ is a sequence in $]0, 1]$ such that $\inf_{n\in\mathbb{N}} \lambda_n > 0$.

The algorithm (7) equipped by a steplength rule as in (8) and the algorithm (9) with the steplength strategy depicted in (10) are strongly connected to the existence of the Lipschitz constant for the gradient of the function f. Clearly, some problems occur in applying this type of method if the function f does not admit a Lipschitz continuous gradient: this is, for example, the case of the Kullback-Leibler divergence when the background is null. If the function f has Lipschitz continuous gradient, the corresponding Lipschitz constant may be hard (impossible) to compute.

In order to consider a function class broader than the one required to use the criteria (8) and (10), we investigate other possible approaches for choosing the steplength in the proximal gradient methods for both the versions (7) and (9).

3. Khobotov-type algorithms

Khobotov's method [5, 25] is a modified version of the extra-gradient method [6] for solving minimization problems of type:

$$\min_{\boldsymbol{x}\in X} f(\boldsymbol{x}), \tag{11}$$

where X is a closed convex subset of $\mathbb{R}^N$ and f is a convex C^1 function from X to $\mathbb{R}$. In particular, Khobotov's idea for solving problem (11) consists of a two step projection algorithm defined by the following relations:

$$\begin{cases} \bar{\boldsymbol{x}}_n = \mathbb{P}_X(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)) \\ \boldsymbol{x}_{n+1} = \mathbb{P}_X(\boldsymbol{x}_n - \alpha_n \nabla f(\bar{\boldsymbol{x}}_n)), \end{cases} \tag{12}$$

where $\mathbb{P}_X$ denotes the projection on to the feasible set X. If the steplength parameters α_n are chosen such that:

$$\alpha_{\min} \le \alpha_n \le \alpha_{\max} \quad \text{and} \quad \alpha_n^2 \frac{\|\nabla f(\boldsymbol{x}_n) - \nabla f(\bar{\boldsymbol{x}}_n)\|^2}{\|\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n\|^2} \le \rho^2, \tag{13}$$

for some $0 < \alpha_{\min} < \alpha_{\max}$ and $0 < \rho < 1$, the algorithm converges to a solution of problem (11). If ∇f is Lipschitz continuous, the conditions (13) can be satisfied [25]. A practical steplength selection procedure for this iteration is also given in [25].

The choice of the steplength parameters however is not dependent on the value of Lipschitz constant of the gradient of the objective function. This convenient aspect encourages us to apply the Khobotov technique also for the proximal methods introduced in (7).

3.1. A generalization of Khobotov's algorithm

The goal of this section is to provide an extension of algorithm (12) to the more general problem (1), and to prove convergence. Let us consider the algorithm

$$\begin{cases} \bar{\boldsymbol{x}}_n = \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)) \\ \boldsymbol{x}_{n+1} = \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\bar{\boldsymbol{x}}_n)) \end{cases} \tag{14}$$

($\boldsymbol{x}_0$ arbitrary) for solving the minimization problem (1). It is a generalization of the method (12) as the latter belongs to the class of the schemes described by (14) when the function g defines the indicator function of the set X.

Before introducing the main theorem of this section, we recall a simple property of the differentiable convex functions.

Lemma 5 (Monotonicity)**.** *If $f : \mathbb{R}^N \longrightarrow \mathbb{R}$ is a differentiable convex function, then $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \ge 0, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$.*

Proof. This follows by expressing the subdifferential inequality in $\boldsymbol{x}$ and $\boldsymbol{y}$. □

Theorem 1. *Let $\hat{\boldsymbol{x}}$ be a minimizer of minimization problem* (1) *and $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ be the sequence defined by* (14)*. Then we have, for any positive sequence $\{\alpha_n\}_{n\in\mathbb{N}}$:*

$$\|\boldsymbol{x}_{n+1} - \hat{\boldsymbol{x}}\|^2 \le \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n\|^2 \left[1 - \alpha_n^2 \frac{\|\nabla f(\boldsymbol{x}_n) - \nabla f(\bar{\boldsymbol{x}}_n)\|^2}{\|\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n\|^2}\right]. \tag{15}$$

Furthermore, if the steplength parameters α_n satisfy the relations (13), the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ converges to a minimizer of problem (1).

Proof. We apply Lemma 4 three times.

(i) Let us consider $\boldsymbol{x}_{n+1} = \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\bar{\boldsymbol{x}}_n))$, i.e. $\boldsymbol{x}^+ = \boldsymbol{x}_{n+1},\ \boldsymbol{x}^- = \boldsymbol{x}_n,\ \Delta = -\alpha_n \nabla f(\bar{\boldsymbol{x}}_n)$ and $\boldsymbol{x} = \hat{\boldsymbol{x}}$:

$$\begin{aligned}\|\boldsymbol{x}_{n+1} - \hat{\boldsymbol{x}}\|^2 &\leq \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{x}_{n+1} - \boldsymbol{x}_n\|^2 + 2\langle \boldsymbol{x}_{n+1} - \hat{\boldsymbol{x}}, -\alpha_n \nabla f(\bar{\boldsymbol{x}}_n)\rangle \\ &\quad +2\alpha_n g(\hat{\boldsymbol{x}}) - 2\alpha_n g(\boldsymbol{x}_{n+1}).\end{aligned}$$

(ii) Let us consider $\hat{\boldsymbol{x}} = \mathbf{prox}_{\alpha_n g}(\hat{\boldsymbol{x}} - \alpha_n \nabla f(\hat{\boldsymbol{x}}))$ (see equation (4), i.e. $\boldsymbol{x}^+ = \hat{\boldsymbol{x}},\ \boldsymbol{x}^- = \hat{\boldsymbol{x}},\ \Delta = -\alpha_n \nabla f(\hat{\boldsymbol{x}})$ and $\boldsymbol{x} = \bar{\boldsymbol{x}}_n$:

$$\begin{aligned}\|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}_n\|^2 &\leq \|\hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}_n\|^2 - \|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}\|^2 + 2\langle \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}_n, -\alpha_n \nabla f(\hat{\boldsymbol{x}})\rangle \\ &\quad +2\alpha_n g(\bar{\boldsymbol{x}}_n) - 2\alpha_n g(\hat{\boldsymbol{x}})\end{aligned}$$

and, consequently,

$$0 \leq 2\langle \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}_n, -\alpha_n \nabla f(\hat{\boldsymbol{x}})\rangle + 2\alpha_n g(\bar{\boldsymbol{x}}_n) - 2\alpha_n g(\hat{\boldsymbol{x}}).$$

(iii) Let us consider $\bar{\boldsymbol{x}}_n = \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n))$, i.e. $\boldsymbol{x}^+ = \bar{\boldsymbol{x}}_n,\ \boldsymbol{x}^- = \boldsymbol{x}_n,\ \Delta = -\alpha_n \nabla f(\boldsymbol{x}_n)$ and $\boldsymbol{x} = \boldsymbol{x}_{n+1}$:

$$\begin{aligned}\|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_{n+1}\|^2 &\leq \|\boldsymbol{x}_n - \boldsymbol{x}_{n+1}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2 + 2\langle \bar{\boldsymbol{x}}_n - \boldsymbol{x}_{n+1}, -\alpha_n \nabla f(\boldsymbol{x}_n)\rangle \\ &\quad +2\alpha_n g(\boldsymbol{x}_{n+1}) - 2\alpha_n g(\bar{\boldsymbol{x}}_n).\end{aligned}$$

The sum of the inequalities obtained in (i), (ii) and (iii) provides the following relation:

$$\begin{aligned}\|\boldsymbol{x}_{n+1} - \hat{\boldsymbol{x}}\|^2 + \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_{n+1}\|^2 &\leq \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 + 2\langle \boldsymbol{x}_{n+1} - \hat{\boldsymbol{x}}, -\alpha_n \nabla f(\bar{\boldsymbol{x}}_n)\rangle \\ &\quad +2\langle \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}_n, -\alpha_n \nabla f(\hat{\boldsymbol{x}})\rangle \qquad (16) \\ &\quad -\|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2 + 2\langle \bar{\boldsymbol{x}}_n - \boldsymbol{x}_{n+1}, -\alpha_n \nabla f(\boldsymbol{x}_n)\rangle.\end{aligned}$$

Starting from (16), we obtain:

$$\begin{aligned}\|\boldsymbol{x}_{n+1} - \hat{\boldsymbol{x}}\|^2 &\leq \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_{n+1}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2 \\ &\quad +2\alpha_n\langle \boldsymbol{x}_{n+1} - \bar{\boldsymbol{x}}_n, \nabla f(\boldsymbol{x}_n) - \nabla f(\bar{\boldsymbol{x}}_n)\rangle \\ &\quad +2\alpha_n\langle \hat{\boldsymbol{x}} - \bar{\boldsymbol{x}}_n, \nabla f(\bar{\boldsymbol{x}}_n) - \nabla f(\hat{\boldsymbol{x}})\rangle \\ &\leq \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_{n+1}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2 \\ &\quad +2\alpha_n\langle \boldsymbol{x}_{n+1} - \bar{\boldsymbol{x}}_n, \nabla f(\boldsymbol{x}_n) - \nabla f(\bar{\boldsymbol{x}}_n)\rangle \\ &= \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_{n+1}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2 + \|\boldsymbol{x}_{n+1} - \bar{\boldsymbol{x}}_n\|^2 + \\ &\quad +\alpha_n^2\|\nabla f(\boldsymbol{x}_n) - \nabla f(\bar{\boldsymbol{x}}_n)\|^2 + \\ &\quad -\|\boldsymbol{x}_{n+1} - \bar{\boldsymbol{x}}_n - \alpha_n\left[\nabla f(\boldsymbol{x}_n) - \nabla f(\bar{\boldsymbol{x}}_n)\right]\|^2 \\ &\leq \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 - \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2\left[1 - \alpha_n^2\frac{\|\nabla f(\boldsymbol{x}_n) - \nabla f(\bar{\boldsymbol{x}}_n)\|^2}{\|\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n\|^2}\right],\end{aligned}$$

where the second inequality follows from Lemma 5 and the other relations are consequence of the scalar product properties.

Furthermore, combining steplength conditions (13) and inequality (15) yields:

$$\|\boldsymbol{x}_{n+1} - \hat{\boldsymbol{x}}\|^2 \leq \|\boldsymbol{x}_n - \hat{\boldsymbol{x}}\|^2 - (1-\rho^2)\,\|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2, \tag{17}$$

and:

$$\|\boldsymbol{x}_N - \hat{\boldsymbol{x}}\|^2 \leq \|\boldsymbol{x}_M - \hat{\boldsymbol{x}}\|^2 - (1-\rho^2)\sum_{n=M}^{N-1} \|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2 \qquad N > M. \tag{18}$$

This implies that the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ is bounded and that $\|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|$ tends to zero (as n tends to infinity). Therefore there exists a converging subsequence $\boldsymbol{x}_{n_j} \overset{j\to\infty}{\longrightarrow} \boldsymbol{x}^\dagger$ and $\overline{\boldsymbol{x}}_{n_j} \overset{j\to\infty}{\longrightarrow} \boldsymbol{x}^\dagger$ also. It follows from the first line of iteration (14), and from the continuity of the proximal operator and of ∇f that $\boldsymbol{x}^\dagger$ satisfies $\boldsymbol{x}^\dagger = \mathbf{prox}_{\alpha g}(\boldsymbol{x}^\dagger - \alpha\nabla f(\boldsymbol{x}^\dagger))$ (by taking a further subsequence for which $\alpha_{n_{j_k}} \overset{k\to\infty}{\longrightarrow} \alpha > 0$). This means that $\boldsymbol{x}^\dagger$ is a minimizer of (1).

Replacing $\hat{\boldsymbol{x}}$ by $\boldsymbol{x}^\dagger$ in relation (18) yields:

$$\|\boldsymbol{x}_N - \boldsymbol{x}^\dagger\|^2 \quad \leq \quad \|\boldsymbol{x}_M - \boldsymbol{x}^\dagger\|^2$$

for $N > M$, and this implies the convergence of the whole sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ to $\boldsymbol{x}^\dagger$. □

In case ∇f is Lipschitz continuous, one can show that the conditions (13) can be satisfied. The proof is omitted here, as it is identical to the one for the Khobotov algorithm for the constrained case [25, p. 260]. The practical steplength selection scheme given in [25] for (12) can also be applied to iteration (14). This practical realization does not depend on the value of the Lipschitz constant (see also Algorithm 1 in Section 5.1 in the case of signal recovering under Poisson noise).

In practice, it may often occur that the proximal operator $\mathbf{prox}_{\alpha_n g}$ may not be computed exactly. The following theorem shows that the algorithm (14) is robust with respect to the inexact computation of $\mathbf{prox}_{\alpha_n g}$.

Theorem 2. *Let $\hat{\boldsymbol{x}}$ be a minimizer of minimization problem* (1) *and $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ be the sequence defined by*

$$\begin{cases} \bar{\boldsymbol{x}}_n = \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)) + \boldsymbol{e}_n \\ \boldsymbol{x}_{n+1} = \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\bar{\boldsymbol{x}}_n)) + \boldsymbol{f}_n, \end{cases} \tag{19}$$

with errors $\boldsymbol{e}_n, \boldsymbol{f}_n$ satisfying $\sum_n \|\boldsymbol{e}_n\| < \infty$ and $\sum_n \|\boldsymbol{f}_n\| < \infty$. If ∇f is Lipschitz continuous and the steplength parameters α_n satisfy relations (13), the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ converges to a minimizer of problem (1).

The proof follows the same lines as the proof of Theorem 1; it can be found in Appendix A. The conditions of absolute summability of the error terms $\boldsymbol{e}_n$ and $\boldsymbol{f}_n$ that appear in algorithm (19) are identical to conditions on error terms introduced in [19] for algorithm (9). It is necessary to assume that ∇f is Lipschitz continuous, but the value of the Lipschitz constant does not have to be known.

3.2. A generalization of the AEM algorithm

In certain cases it is worthwhile to cast the minimization problem (1) in the form of a saddle-point problem. E.g., if $g(\boldsymbol{x})$ can be written as $\max_{\boldsymbol{y}} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - g^*(\boldsymbol{y})$, then problem (1) takes the form

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} f(\boldsymbol{x}) + \langle \boldsymbol{x}, \boldsymbol{y} \rangle - g^*(\boldsymbol{y}). \tag{20}$$

E.g. if g^* is the indicator function of the ℓ_∞-ball $\{\boldsymbol{x}, \|\boldsymbol{x}\|_\infty \leq \lambda\}$, then $g(\boldsymbol{x}) = \lambda \|\boldsymbol{x}\|_1$.

A variant of Khobotov's algorithm (12) has also been developed for constrained saddle-point problems, i.e. for problems of type

$$\min_{\boldsymbol{x} \in X} \max_{\boldsymbol{y} \in Y} F(\boldsymbol{x}, \boldsymbol{y}) \tag{21}$$

where F is convex in x, concave in y and continuously differentiable with respect to x and y. In this case, [8] have proposed the so-called alternating extragradient method (AEM)

$$\begin{cases} \bar{\boldsymbol{y}}_n &= \mathbb{P}_Y \left[\boldsymbol{y}_n + \alpha_n \nabla_y F(\boldsymbol{x}_n, \boldsymbol{y}_n) \right] \\ \boldsymbol{x}_{n+1} &= \mathbb{P}_X \left[\boldsymbol{x}_n - \alpha_n \nabla_x F(\boldsymbol{x}_n, \bar{\boldsymbol{y}}_n) \right] \\ \boldsymbol{y}_{n+1} &= \mathbb{P}_Y \left[\boldsymbol{y}_n + \alpha_n \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n) \right] \end{cases} \tag{22}$$

under the conditions

$$\begin{array}{cc} 1 - 2\alpha_n A_n - 2\alpha_n^2 B_n^2 \geq \epsilon > 0, & 1 - 2\alpha_n C_n \geq \epsilon > 0, \\ 0 < \alpha_{\min} \leq \alpha_n \leq \alpha_{\max}, & 0 < \epsilon < 1 \end{array} \tag{23}$$

where A_n, B_n, C_n are given by:

$$\begin{aligned} A_n &= \frac{\|\nabla_x F(\boldsymbol{x}_{n+1}, \bar{\boldsymbol{y}}_n) - \nabla_x F(\boldsymbol{x}_n, \bar{\boldsymbol{y}}_n)\|}{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_n\|} \\ B_n &= \frac{\|\nabla_y F(\boldsymbol{x}_n, \boldsymbol{y}_n) - \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\|}{\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_n\|} \\ C_n &= \frac{\|\nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n) - \nabla_y F(\boldsymbol{x}_{n+1}, \bar{\boldsymbol{y}}_n)\|}{\|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|}. \end{aligned} \tag{24}$$

If a saddle point of problem (21) exists, convergence is shown in [8, Theorem 1]. The paper also shows that the conditions (23) are feasible if some Lipschitz conditions are imposed on $\nabla_x F$ and $\nabla_y F$. A practical implementation is also given (again not depending on the knowledge of the Lipschitz constant). We do not reproduce it here for lack of space.

We propose a generalization of the AEM algorithm (22):

$$\begin{cases} \bar{\boldsymbol{y}}_n &= \mathbf{prox}_{\alpha_n g_2} \left[\boldsymbol{y}_n + \alpha_n \nabla_y F(\boldsymbol{x}_n, \boldsymbol{y}_n) \right] \\ \boldsymbol{x}_{n+1} &= \mathbf{prox}_{\alpha_n g_1} \left[\boldsymbol{x}_n - \alpha_n \nabla_x F(\boldsymbol{x}_n, \bar{\boldsymbol{y}}_n) \right] \\ \boldsymbol{y}_{n+1} &= \mathbf{prox}_{\alpha_n g_2} \left[\boldsymbol{y}_n + \alpha_n \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n) \right] \end{cases} \tag{25}$$

designed to solve the more general saddle-point problem:

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} g_1(\boldsymbol{x}) + F(\boldsymbol{x}, \boldsymbol{y}) - g_2(\boldsymbol{y}), \tag{26}$$

where F and g_1 are convex in $\boldsymbol{x}$ and F and $-g_2$ are concave in $\boldsymbol{y}$. F is also C^1 with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$. If g_1 and g_2 are indicator functions of the sets X and Y, this reduces to the constrained problem (21).

Theorem 3. *Let $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n\in\mathbb{N}}$ be the sequence defined by algorithm (25) with steplength parameters satisfying conditions (23). If a solution to the saddle point problem (26) exists, the sequence $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n\in\mathbb{N}}$ converges to such a saddle-point.*

The proof of this theorem is included in Appendix B. It can also be shown that the steplength conditions (23) are feasible if some Lipschitz conditions are imposed on $\nabla_x F$ and $\nabla_y F$. The proof is identical to the one for a similar property of the AEM iteration (22) in [8, lemma 3.2]. It follows that the same practical step-size rule as for the AEM method can be used here too [8, p 11 and corollary 1]. Such a rule does not depend on the actual value of the Lipschitz constants of $\nabla_x F$ and $\nabla_y F$.

4. Armijo-type algorithm

This section is devoted to the study of a strategy for choosing the parameters α_n and λ_n in the proximal scheme of the form (9); in particular we are interested in a procedure that does not need the Lipschitz continuity assumption of the gradient of the function f in (1). Our analysis is similar to the paper by Tseng and Yun [10], where the authors adapted the Armijo steplength selection rule [7, Chapter 2 - Section 3], often used in gradient-type methods, to the presence of a proximal operator instead of a projection. We explain how it's possible to use this generalization of the Armijo scheme in the algorithm (9), but we provide a slightly different approach to the convergence analysis with respect to the one proposed in [10]. In particular, we underline that the exact knowledge of the proximal operator of the function g in (1) is not needed to obtain a descent direction.

We are concerned with solving the optimization problem (1) by means of an iterative method of type:

$$\begin{cases} \boldsymbol{x}_{n+1} &= \boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n \\ \boldsymbol{d}_n &= \bar{\boldsymbol{x}}_n - \boldsymbol{x}_n = \mathbf{prox}_{\alpha_n g}\left(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)\right) - \boldsymbol{x}_n \end{cases} \tag{27}$$

where α_n are positive steplengths ($0 < \alpha_{\min} \leq \alpha_n \leq \alpha_{\max}$) and $\lambda_n \in (0, 1]$. The parameters λ_n are chosen in order to ensure the decrease of the objective function at each iteration. In particular, for a fixed $\beta \in (0, 1)$ and $\sigma \in (0, 1)$, we set $\lambda_n = \beta^{m_n}$ where m_n is the first non-negative integer for which:

$$f(\boldsymbol{x}_n)+g(\boldsymbol{x}_n)-f(\boldsymbol{x}_n+\beta^{m_n}\boldsymbol{d}_n)-g(\boldsymbol{x}_n+\beta^{m_n}\boldsymbol{d}_n) \geq -\sigma\beta^{m_n}\left[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\right]. \tag{28}$$

There is no further restriction on the steplengths α_n.

We will show that every limit point of the sequence $\{\boldsymbol{x}_n\}$ is a stationary point of the algorithm and hence a minimizer of functional (1). This is a similar result as in the constrained case [7]. The proof requires the use of several lemmas.

Lemma 6. *If $\bar{\boldsymbol{x}}_n = \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n))$ and $\bar{\boldsymbol{x}}_n \neq \boldsymbol{x}_n$ then*

$$\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) \leq -\frac{1}{\alpha_n}\|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2 < 0 \qquad \forall n.$$

Proof. We apply the Lemma 1 by setting $\boldsymbol{u} = \boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)$, $\boldsymbol{x} = \bar{\boldsymbol{x}}_n$ and $\boldsymbol{z} = \boldsymbol{x}_n$. In particular the following inequalities hold true:

$$\langle \boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n) - \bar{\boldsymbol{x}}_n, \boldsymbol{x}_n - \bar{\boldsymbol{x}}_n \rangle \leq \alpha_n g(\boldsymbol{x}_n) - \alpha_n g(\bar{\boldsymbol{x}}_n)$$

or

$$\langle \nabla f(\boldsymbol{x}_n), \bar{\boldsymbol{x}}_n - \boldsymbol{x}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) \leq -\frac{1}{\alpha_n}\|\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n\|^2.$$

From the assumption on the direction $\bar{\boldsymbol{x}}_n - \boldsymbol{x}_n$, the lemma is proved. □

Lemma 7. *For any subsequence $\{\boldsymbol{x}_n\}_{n\in\mathcal{N}}$ (with $\mathcal{N} \subset \mathbb{N}$) that converges to a non-stationary point and for which $\{\alpha_n\}_{n\in\mathcal{N}}$ also converges, the corresponding sequence $\{\boldsymbol{d}_n\}_{n\in\mathcal{N}}$ is bounded and satisfies*

$$\limsup_{n\to+\infty,\ n\in\mathcal{N}} \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) < 0$$

Proof. Suppose the subsequence $\{\boldsymbol{x}_n\}_{n\in\mathcal{N}}$ converges to a non-stationary point $\tilde{\boldsymbol{x}}$. Let $\lim_{n\in\mathcal{N}} \alpha_n = \alpha$. By the continuity of the proximal operator and the gradient of f we have:

$$\lim_{n\to+\infty,\ n\in\mathcal{N}} \mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)) = \mathbf{prox}_{\alpha g}(\tilde{\boldsymbol{x}} - \alpha \nabla f(\tilde{\boldsymbol{x}}))$$

and

$$\limsup_{n\to+\infty,\ n\in\mathcal{N}} \left\|\mathbf{prox}_{\alpha_n g}(\boldsymbol{x}_n - \alpha_n \nabla f(\boldsymbol{x}_n)) - \boldsymbol{x}_n\right\| = \left\|\mathbf{prox}_{\alpha g}(\tilde{\boldsymbol{x}} - \alpha \nabla f(\tilde{\boldsymbol{x}})) - \tilde{\boldsymbol{x}}\right\|.$$

This implies that $\{\boldsymbol{d}_n\}_{n\in\mathcal{N}}$ is bounded.

To prove the second part of the thesis, we recall Lemma 6:

$$\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) \leq -\frac{1}{\alpha_n}\|\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n\|^2.$$

By taking the limit on the above inequality, we obtain

$$\limsup_{n\to+\infty,\ n\in\mathcal{N}} [\langle \nabla f(\boldsymbol{x}_n), \bar{\boldsymbol{x}}_n - \boldsymbol{x}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)] \leq -\frac{1}{\alpha}\|\tilde{\boldsymbol{x}} - \mathbf{prox}_{\alpha g}(\tilde{\boldsymbol{x}} - \alpha \nabla f(\tilde{\boldsymbol{x}}))\|^2$$

Since $\tilde{\boldsymbol{x}}$ is a non-stationary point, it follows that:

$$\limsup_{n\to+\infty,\ n\in\mathcal{N}} [\langle \nabla f(\boldsymbol{x}_n), \bar{\boldsymbol{x}}_n - \boldsymbol{x}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)] < 0.$$

□

Lemma 8. *Let $\sigma \in (0,1)$. If the direction $\boldsymbol{d}_n \neq \boldsymbol{0}$ (i.e. $\boldsymbol{x}_n$ not a fixed-point of (4)), then the line-search (28) determining λ_n is well-defined, i.e. it is always possible to find $\lambda_n \in (0,1]$ such that:*

$$\mathcal{F}(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n) - \mathcal{F}(\boldsymbol{x}_n) \leq \lambda_n \sigma \big[\langle \nabla f(\boldsymbol{x}_n), \bar{\boldsymbol{x}}_n - \boldsymbol{x}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big].$$

Proof. Choosing $\tilde{\sigma}$ such that $0 < \sigma < \tilde{\sigma} < 1$, one finds from Lemma 6 and the fact that $\lambda_n > 0$ that:

$$(1-\tilde{\sigma})\lambda_n \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - (1-\tilde{\sigma})\lambda_n g(\boldsymbol{x}_n) + (1-\tilde{\sigma})\lambda_n g(\bar{\boldsymbol{x}}_n) < 0,$$

and hence:

$$\langle \nabla f(\boldsymbol{x}_n), \lambda_n \boldsymbol{d}_n \rangle < \lambda_n \tilde{\sigma} \big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big] + \lambda_n g(\boldsymbol{x}_n) - \lambda_n g(\bar{\boldsymbol{x}}_n). \tag{29}$$

Since the function g is convex and $\lambda_n \in (0,1]$, we remark that:

$$g(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n) = g\big((1-\lambda_n)\boldsymbol{x}_n + \lambda_n \bar{\boldsymbol{x}}_n\big) \leq (1-\lambda_n) g(\boldsymbol{x}_n) + \lambda_n g(\bar{\boldsymbol{x}}_n),$$

and therefore:

$$\lambda_n g(\boldsymbol{x}_n) - \lambda_n g(\bar{\boldsymbol{x}}_n) \leq g(\boldsymbol{x}_n) - g(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n). \tag{30}$$

By considering equations (29) and (30), it follows that:

$$\langle \nabla f(\boldsymbol{x}_n), \lambda_n \boldsymbol{d}_n \rangle < \lambda_n \tilde{\sigma} \big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big] + g(\boldsymbol{x}_n) - g(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n)$$

or

$$f(\boldsymbol{x}_n) + \langle \nabla f(\boldsymbol{x}_n), \lambda_n \boldsymbol{d}_n \rangle - f(\boldsymbol{x}_n) - g(\boldsymbol{x}_n) + g(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n) < \lambda_n \tilde{\sigma} \big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big].$$

This can be written as:

$$\mathcal{F}(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n) - \mathcal{F}(\boldsymbol{x}_n) - r(\lambda_n) < \lambda_n \tilde{\sigma} \big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big], \tag{31}$$

where $f(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n) = f(\boldsymbol{x}_n) + \langle \nabla f(\boldsymbol{x}_n), \lambda_n \boldsymbol{d}_n \rangle + r(\lambda_n)$ and $\lim_{\lambda_n \to 0} r(\lambda_n)/\lambda_n = 0$ (by Taylor's theorem).

As $\lambda_n(\sigma - \tilde{\sigma})\big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big] > 0$, it follows from the definition of $r(\lambda_n)$ that there exists λ_n small enough such that:

$$r(\lambda_n) < \lambda_n(\sigma - \tilde{\sigma})\big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big]. \tag{32}$$

The inequalities (31) and (32) allow to reach the proof of the proposition. In particular

$$\begin{aligned} \mathcal{F}(\boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n) - \mathcal{F}(\boldsymbol{x}_n) \quad &< \quad \lambda_n \tilde{\sigma} \big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big] + r(\lambda_n) \\ &< \quad \lambda_n \sigma \big[\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)\big]. \end{aligned}$$

□

Theorem 4. *Let $\hat{\boldsymbol{x}}$ be a minimizer for the problem* (1)*, $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ a sequence generated by the algorithm* (27) *and assume that λ_n is chosen by the backtracking rule* (28)*. Then every limit point of the sequence $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$ is a stationary point.*

Proof. We assume that $\tilde{\boldsymbol{x}}$ is a limit point of $\{\boldsymbol{x}_n\}_{n\in\mathbb{N}}$. This means that there exists a subsequence that converges to $\tilde{\boldsymbol{x}}$. As the α_n are bounded, there exists a further subsequence (i.e. $n \in \mathcal{N}$, an infinite subset of $\mathbb{N}$), for which $\lim_{n\to\infty, n\in\mathcal{N}} \boldsymbol{x}_n = \tilde{\boldsymbol{x}}$ and $\lim_{n\to\infty, n\in\mathcal{N}} \alpha_n = \alpha$. To arrive at a contradiction, we assume that $\tilde{\boldsymbol{x}} \neq \mathbf{prox}_{\alpha g}(\tilde{\boldsymbol{x}} - \alpha\nabla f(\tilde{\boldsymbol{x}}))$.

Since the sequence $\{\mathcal{F}(\boldsymbol{x}_n)\}_{n\in\mathbb{N}}$ is monotonically non-increasing and there exists a minimizer $\hat{\boldsymbol{x}}$, it follows that the entire sequence $\{\mathcal{F}(\boldsymbol{x}_n)\}_{n\in\mathbb{N}}$ is convergent. Hence

$$\lim_{n\to+\infty} \mathcal{F}(\boldsymbol{x}_n) - \mathcal{F}(\boldsymbol{x}_{n+1}) = 0.$$

By definition (28) and lemma (6) we have

$$\mathcal{F}(\boldsymbol{x}_n) - \mathcal{F}(\boldsymbol{x}_{n+1}) \geq -\sigma\lambda_n[\langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)] > 0,$$

and therefore

$$\lim_{n\to+\infty} \lambda_n[\langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)] = 0. \tag{33}$$

By applying Lemma 7, we have

$$\limsup_{n\to\infty, n\in\mathcal{N}} \langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) < 0 \tag{34}$$

and, from equation (33), $\lim_{n\to\infty, n\in\mathcal{N}} \lambda_n = 0$ (i.e. m_n in expression (28) tends to infinity).

By the definition of the rule (28), for some index $\bar{n} \geq 0$, we must have:

$$f(\boldsymbol{x}_n) - f\left(\boldsymbol{x}_n + \frac{\lambda_n}{\beta}\boldsymbol{d}_n\right) + g(\boldsymbol{x}_n) - g\left(\boldsymbol{x}_n + \frac{\lambda_n}{\beta}\boldsymbol{d}_n\right) < -\sigma\frac{\lambda_n}{\beta}[\langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)]$$

($\forall n \in \mathcal{N}, n \geq \overline{n}$) or equivalently:

$$\frac{f(\boldsymbol{x}_n) - f(\boldsymbol{x}_n + \lambda_n/\beta\, \boldsymbol{d}_n)}{\lambda_n/\beta} + \frac{g(\boldsymbol{x}_n) - g(\boldsymbol{x}_n + \lambda_n/\beta\, \boldsymbol{d}_n)}{\lambda_n/\beta} < -\sigma[\langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)]$$

($\forall n \in \mathcal{N}, n \geq \overline{n}$). From the mean value theorem and the convexity of the function g, the previous relation can be rewritten as

$$\begin{aligned} - &\quad \langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle + r(\lambda_n) + \sigma[\langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n)] < \\ < &\quad \frac{g(\boldsymbol{x}_n + \lambda_n/\beta\, \boldsymbol{d}_n) - g(\boldsymbol{x}_n)}{\lambda_n/\beta} \leq \frac{(1 - \lambda_n/\beta)\, g(\boldsymbol{x}_n) + \lambda_n/\beta g(\bar{\boldsymbol{x}}_n) - g(\boldsymbol{x}_n)}{\lambda_n/\beta} \\ = &\quad -g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) \end{aligned}$$

($\forall n \in \mathcal{N}, n \geq \overline{n}$ and with $r(\lambda)/\lambda \overset{\lambda\to 0}{\longrightarrow} 0$) and consequently:

$$(\sigma-1)\langle\nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - (\sigma-1)g(\boldsymbol{x}_n) + (\sigma-1)g(\bar{\boldsymbol{x}}_n) + r(\lambda_n) \leq 0 \quad \forall n \in \mathcal{N}, n \geq \overline{n}.$$

Since $\sigma - 1 < 0$, it holds true that:

$$\lim_{n \in \mathcal{N}} \langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) \geq 0$$

which contradicts (34). $\square$

Remark 4. In practice, it may happen that the map $\mathbf{prox}_{\alpha g}$ is difficult to evaluate exactly. In that case, one may use an approximation of the proximal operator. As long as the $\bar{\boldsymbol{x}}_n$ obtained in this way satisfies the inequality:

$$\langle \nabla f(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - g(\boldsymbol{x}_n) + g(\bar{\boldsymbol{x}}_n) < 0, \tag{35}$$

it is always possible to find a step length λ_n that satisfies the inequality (28) (see proof of Lemma 8) and that guarantees a decrease of the objective function.

5. Signal recovering applications

In this section we describe how to apply the algorithms presented in Sections 3 and 4 to the solution of the one-dimensional and two-dimensional signal restoration problems formalized in (2). Henceforth, we restrict our analysis to the problem of minimizing the following functional

$$\mathcal{F}(\boldsymbol{x}) = \mathrm{KL}(\boldsymbol{x}) + \mu J_R(\boldsymbol{x}) + \mathcal{I}_{\{\boldsymbol{x} \geq \mathbf{0}\}}(\boldsymbol{x}) \tag{36}$$

where the regularization term J_R depends on the application (it will be specified in the next subsections) and $\mathcal{I}_{\{\boldsymbol{x} \geq \mathbf{0}\}}$ is the indicator function of the non-negative orthant.

Hereafter we call the previously suggested generalizations of the Khobotov and Armijo-type algorithms (defined in (14) and (27)-(28) respectively) as Proximal Khobotov Method (PKM) and Proximal Armijo Method (PAM).

5.1. Algorithmic details

This subsection is devoted to explain the use of the PKM and PAM algorithms, in the practical case of signal restoration. Algorithm 1 and Algorithm 2 report the implementation details. We first notice that the gradient of the Kullback-Leibler divergence (3) is given by $\nabla \mathrm{KL}(\boldsymbol{x}) = H^T \boldsymbol{e} - H^T Y^{-1} \boldsymbol{g}$, where $\boldsymbol{e} \in \mathbb{R}^M$ is a vector whose components are all equal to 1 and Y is a diagonal matrix with the entries of the vector $H\boldsymbol{x} + \boldsymbol{b}$. The weighted sum $\mu J_R + \mathcal{I}_{\boldsymbol{x} \geq \mathbf{0}}$ is denoted by $\mathcal{Z}$.

5.2. One-dimensional signal restoration: compressed sensing

We consider a compressed sensing problem, i.e the reconstruction of a sparse vector $\boldsymbol{x}^*$ from noisy measurements. In particular, $\boldsymbol{x}^*$ is a sparse vector of non-negative values whose measurements are distorted by Poisson noise [14]. We assume that the observed signal $\boldsymbol{g} \in \mathbb{Z}_+^M$ is a realization of a Poisson random variable with expected value given by $H\boldsymbol{x}^*$, where $\boldsymbol{x}^* \in \mathbb{R}_+^N$ is the signal of

Algorithm 1 Proximal Khobotov Method (PKM) for signal recovering

Choose the starting point $\boldsymbol{x}^0$ and the parameters $\alpha_0 > 0$, $\rho \in (0,1)$.
for $n = 0, 1, 2, ...$ **do**
STEP 1. $\bar{\boldsymbol{x}}_n = \mathbf{prox}_{\alpha_n \mathcal{Z}}(\boldsymbol{x}_n - \alpha_n \nabla \mathrm{KL}(\boldsymbol{x}_n))$

STEP 2.
if $\alpha_n^2 \dfrac{\|\nabla \mathrm{KL}(\boldsymbol{x}_n) - \nabla \mathrm{KL}(\bar{\boldsymbol{x}}_n)\|}{\|\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n\|} > \rho^2$
then $\alpha_n = \min\left\{\dfrac{\alpha_n}{2}, \dfrac{\|\nabla \mathrm{KL}(\boldsymbol{x}_n) - \nabla \mathrm{KL}(\bar{\boldsymbol{x}}_n)\|}{\sqrt{2}\|\boldsymbol{x}_n - \bar{\boldsymbol{x}}_n\|}\right\}$, **goto** STEP 1
else **goto** STEP 3
endif

STEP 3. $\boldsymbol{x}_{n+1} = \mathbf{prox}_{\alpha_n \mathcal{Z}}(\boldsymbol{x}_n - \alpha_n \nabla \mathrm{KL}(\bar{\boldsymbol{x}}_n))$, $\alpha_{n+1} = \alpha_n$
end for

Algorithm 2 Proximal Armijo Method (PAM) for signal recovering

Choose a feasible $\boldsymbol{x}^0$ and the parameters $\alpha_{\max} \geq \alpha_{\min} > 0$ and $\beta, \sigma \in (0,1)$.
for $n = 0, 1, 2, ...$ **do**
STEP 1.
Choose $\alpha_n \in [\alpha_{\min}, \alpha_{\max}]$
$\bar{\boldsymbol{x}}_n = \mathbf{prox}_{\alpha_n \mathcal{Z}}(\boldsymbol{x}_n - \alpha_n \nabla \mathrm{KL}(\boldsymbol{x}_n))$,
$\boldsymbol{d}_n = \bar{\boldsymbol{x}}_n - \boldsymbol{x}_n$,
$\lambda_n = 1$

STEP 2.
if $\mathcal{F}(\boldsymbol{x}_n + \lambda_n \boldsymbol{x}_n) < \mathcal{F}(\boldsymbol{x}_n) + \sigma\lambda_n \left[\langle \nabla \mathrm{KL}(\boldsymbol{x}_n), \boldsymbol{d}_n\rangle - \mathcal{Z}(\boldsymbol{x}_n) + \mathcal{Z}(\bar{\boldsymbol{x}}_n)\right]$
then **goto** STEP 3
else
$\lambda_n = \beta\lambda_n$, **goto** STEP 2
endif

STEP 3. $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \lambda_n \boldsymbol{d}_n$
end for

interest and $H \in \mathbb{R}^{M \times N}$ is the (known) measurement matrix. A way to recover the true signal $\boldsymbol{x}^*$, starting from the observed one $\boldsymbol{g}$ is to find the solution of the following optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^N} \mathrm{KL}(\boldsymbol{x}) + \mu \|\boldsymbol{x}\|_1 + \mathcal{I}_{\{\boldsymbol{x} \geq \mathbf{0}\}} \tag{37}$$

where KL is the Kullback-Leibler divergence, the ℓ_1-norm $\|\boldsymbol{x}\|_1$ promotes sparsity, $\mu > 0$ is the regularization parameter governing the role of the sparsity-inducing operator and $\mathcal{I}_{\{\boldsymbol{x} \geq \mathbf{0}\}}$ is the indicator function of the non-negative orthant.

In [15], the authors suggest an algorithm, called SPIRAL, designed to solve problem (37). The Matlab code of SPIRAL and a collection of test problems are available on-line [26]. The minimization problem (37) can be addressed also by PKM and PAM, noting that the proximal operator of $\mu \|\boldsymbol{x}\|_1 + \mathcal{I}_{\{\boldsymbol{x} \geq \mathbf{0}\}}$ is a modified version of the soft-thresholding operator:

$$\mathbf{prox}_{\mu \|\boldsymbol{x}\|_1 + \mathcal{I}_{\{\boldsymbol{x} \geq \mathbf{0}\}}}(x_i) = \begin{cases} x_i - \mu & \text{if } \; x_i > \mu \\ 0 & \text{otherwise} \end{cases}$$

applied component-wise [3].

Numerical experiments

The numerical experiments are carried out on two test problems. The first one is taken from the collection provided in the SPIRAL package: the true signal is of length 10^5 with 1500 non-zero entries and the observed data is the results of 4000 compressive measurements; the average number of photons per measurement is 15.03 with a maximum of 145.

The second test problem is generated in three steps: 1) a sensing matrix $A \in \mathbb{R}^{1000 \times 5000}$ is created as proposed in [14]; 2) the true signal $\boldsymbol{x}^* \in \mathbb{R}^{5000}$ has all zeros except for 20 non-zero entries drawn uniformly in the interval $[0, 10^5]$ (the positions of the nonzeros in $\boldsymbol{x}^*$ are uniformly distributed in $\{1, 2, \ldots, 5000\}$); 3) the observed signal $\boldsymbol{g} \in \mathbb{R}^{1000}$ has been fixed by multiplying the sensing matrix and the true signal and by adding Poisson noise with the `imnoise` function of the Matlab Image Processing Toolbox. The background is set to 10^{-10}: in practice this means that the background is zero, but it assures the Lipschitz continuity of the gradient of the Kullback-Leibler divergence (this is also done in [15]).

We chose the regularization parameter μ equal to 10^{-6} for the first test problem and 10^{-3} for the second one. We set $\alpha_{\min} = 10^{-10}$ and $\alpha_{\max} = 10^{10}$. Regarding the parameters used by PKM, we fixed $\alpha_0 = 10^{10}$ and $\rho = 0.99$. We tried several values for both α_0 and ρ and we found that the best performance of the algorithm is gained by selecting α_0 large and ρ close to 1, but this dependence is not critical. On the other hand, for PAM, we selected $\sigma = 10^{-4}$, $\beta = 0.5$ (these values for σ and β are standard for Armijo line-search; the performance is not affected by the precise values) and the steplength α variable with the iterations in an interval $[\alpha_{\min}, \alpha_{\max}]$. In this case the choice of α_n influences

Method	Test problem 1			Test problem 2		
	it.	time (s)	RRE	it.	time (s)	RRE
PKM	431	45.76	0.0607	5000*	109.67	0.5769
PAM ($\alpha_n = \alpha_n^{\mathrm{BB1}}$)	38	2.74	0.0599	1073	13.31	0.0746
PAM ($\alpha_n = \alpha_n^{\mathrm{ABB}}$)	58	3.89	0.0599	577	7.35	0.0746
SPIRAL	37	3.02	0.0599	1397	27.10	0.0746

Table 1: Results for the compressed sensing test problems with Poisson noise

the convergence rate significantly. In particular, we show the results obtained by PAM equipped with two different steplength selection schemes based on the Barzilai-Borwein rules [27]:

$$\alpha_n^{BB1} = \frac{{\boldsymbol{s}_{n-1}}^T \boldsymbol{s}_{n-1}}{{\boldsymbol{s}_{n-1}}^T \boldsymbol{w}_{n-1}} \qquad \text{and} \qquad \alpha_n^{BB2} = \frac{{\boldsymbol{s}_{n-1}}^T \boldsymbol{w}_{n-1}}{{\boldsymbol{w}_{n-1}}^T \boldsymbol{w}_{n-1}}$$

where $\boldsymbol{s}_{n-1} = \boldsymbol{x}_n - \boldsymbol{x}_{n-1}$ and $\boldsymbol{w}_{n-1} = \nabla \mathrm{KL}(\boldsymbol{x}_n) - \nabla \mathrm{KL}(\boldsymbol{x}_{n-1})$. The first choice we make sets $\alpha_n = \alpha_n^{\mathrm{BB1}}$ $\forall n$, while the second one defines α_n through a proper alternation between α_n^{BB1} and α_n^{BB2} proposed in [28]:

$$\alpha_n = \alpha_n^{\mathrm{ABB}} = \begin{cases} \alpha_n^{\mathrm{BB2}} & \text{if } \dfrac{\alpha_n^{\mathrm{BB2}}}{\alpha_n^{\mathrm{BB1}}} < \tau \\ \alpha_n^{\mathrm{BB1}} & \text{otherwise} \end{cases}$$

where τ is a prefixed threshold.

SPIRAL, PKM and PAM are stopped when the relative distance between two successive iterations is lower than 10^{-8}: $\|\boldsymbol{x}_{n+1} - \boldsymbol{x}_n\| / \|\boldsymbol{x}_{n+1}\| \leq 10^{-8}$.

The performance of each algorithm is measured through the evaluation of the relative reconstruction error (RRE) that estimates the difference between the n^{th} approximation of the solution and the true object, in Euclidean norm, namely:

$$\mathrm{RRE}(\boldsymbol{x}_n) = \frac{\|\boldsymbol{x}_n - \boldsymbol{x}^*\|}{\|\boldsymbol{x}^*\|},$$

and where $\boldsymbol{x}^*$ is the original non-blurred and non-noisy image (i.e. we do not compare to the unknown true minimizers of the cost function). Table 1 reports the number of iterations and the computational time needed by the different methods to satisfy the stopping criterion and the corresponding relative reconstruction error (RRE) obtained.

The main conclusion that can be drawn from these experiments is that PAM, combined with both the steplength strategies, provides results comparable with those reached by SPIRAL for the first test problem and shows the best performance, with respect to the other methods, in terms of number of iterations and computational time for the second test problem. On the other hand, PKM presents a slow convergence rate: for the second test problem it reaches the prefixed maximum number of iterations (5000) while still very far from the solution. Analogous considerations can be deduced from the plots of the RRE

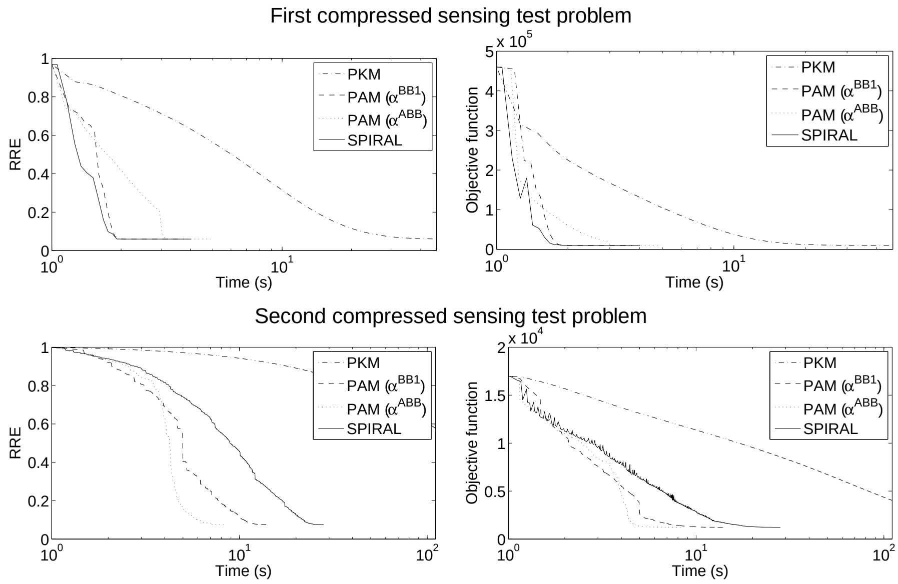


Figure 1: First row: RRE and objective function decrease for the first test problem of Section 5.2. Second row: RRE and objective function decrease for the second test problem.

and the objective function values obtained by the considered algorithms during the iterative process (Figure 1).

Moreover, for both the compressed sensing test problems each algorithm succeeds in recovering the support of the true signal; this is confirmed by similar RRE values provided by all the methods when the stopping criterion is satisfied. The only exception is the PKM algorithm for the second data set: the algorithm stops because the maximum number of iterations is reached and the quality of the corresponding reconstructed support is poor, with many false positives. The support of the PKM reconstructions can be improved (i.e. closer to the true one) by increasing the maximum number of iterations allowed. As a last remark, we observe that PKM and PAM assure a monotone decrease of the objective function with the iterations, while SPIRAL allows a non-monotone decreasing behavior.

5.3. Two-dimensional signal restoration: image deblurring

Image deblurring [16] is the inverse problem of finding an approximation of a true 2D object, given the data $\boldsymbol{g}$, the background $\boldsymbol{b}$ and the blurring matrix H. We recall that a classical assumption for the imaging matrix is $H_{ij} > 0$ $(\forall i, j)$ and $\sum_i H_{ij} = 1$ $(\forall j)$. A possible approach to face up the image deblurring issue is given by the minimization of (36), where the 2D object $\boldsymbol{x} \in \mathbb{R}^{N_1 \times N_2}$ is identified with an element of $\mathbb{R}^N$ (with $N = N_1 N_2$).

In this paper we concentrate on handling the edge-preserving regularization, therefore the regularizer J_R can be expressed by means of the well-known discrete version of the Total Variation (TV) functional [29, 30, 31]:

$$J_R(\boldsymbol{x}) = \mathrm{TV}(\boldsymbol{x}) = \|D\boldsymbol{x}\|_{1,2} \tag{38}$$

where $D\boldsymbol{x}$ denotes a discrete approximation of the gradient (finite differences) of $\boldsymbol{x} \in \mathbb{R}^{N_1 \times N_2}$. In particular, by introducing the local differences $D\boldsymbol{x} = (D_\mathrm{h}\boldsymbol{x}, D_\mathrm{v}\boldsymbol{x})$ as

$$\begin{cases} (D_\mathrm{h}\boldsymbol{x})_{i,j} & = & x_{i+1,j} - x_{i,j} & \quad 1 \le i \le N_1 - 1, \\ (D_\mathrm{v}\boldsymbol{x})_{i,j} & = & x_{i,j+1} - x_{i,j} & \quad 1 \le j \le N_2 - 1, \end{cases} \tag{39}$$

with $(D_\mathrm{h}\boldsymbol{x})_{N_1,j} = (D_\mathrm{v}\boldsymbol{x})_{i,N_2} = 0$, the total variation $\mathrm{TV}(u)$ is given by:

$$\mathrm{TV}(\boldsymbol{x}) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sqrt{((D_\mathrm{h}\boldsymbol{x})_{i,j})^2 + ((D_\mathrm{v}\boldsymbol{x})_{i,j})^2}. \tag{40}$$

As the proximal operator for the total variation does not have an explicit expression, we compute its approximation in an iterative manner in order to compute the first step for both PKM and PAM. The proximal operator of $\alpha\mu\,\mathrm{TV}(\boldsymbol{x})$ is defined as:

$$\mathbf{prox}_{\alpha\mu\,\mathrm{TV}}(\boldsymbol{u}) = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^N} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{u}\|^2 + \alpha\mu\|D\boldsymbol{x}\|_{1,2}. \tag{41}$$

In order to solve problem (41) we used an iterative scheme on the dual formulation similar to the one suggested in [32]. This scheme is stopped with two different stopping criteria depending on the algorithm considered. In the PKM case we decide to set a tolerance value, tol, and break off the iterations when

$$\|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\| \le \mathrm{tol}, \tag{42}$$

while for PAM we require that the inequality introduced in Lemma 6:

$$\langle \nabla\,\mathrm{KL}(\boldsymbol{x}_n), \boldsymbol{d}_n \rangle - \mu\,\mathrm{TV}(\boldsymbol{x}_n) + \mu\,\mathrm{TV}(\bar{\boldsymbol{x}}_n) < 0 \tag{43}$$

is satisfied.

Remark 5. Regarding the non-negativity constraints, we project on the non-negative orthant (at the end of the inner loop) before verifying the stopping criterion (42) or (43). In this way each iterate satisfies the nonnegativity constraints.

Numerical experiments

In order to evaluate the behavior of the proposed algorithms, we perform some numerical experiments on two two-dimensional image restoration test problems, in MATLAB environment. We take into account two different objects: a 128×128 microscopy phantom [33] and the 256×256 cameraman image, well-known in the literature. Blurred and noisy images are generated by convolving each original image with a suitable PSF and by adding Poisson

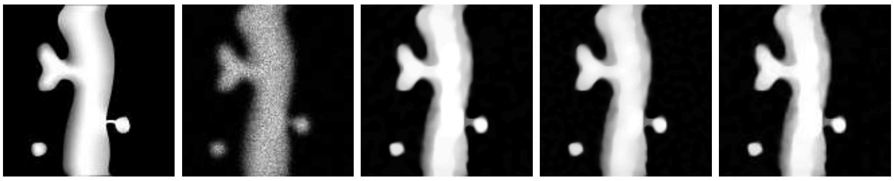


Figure 2: Miroscopy data set: (a) original image, (b) blurred and noisy image, (c) PKM reconstruction, (d) PAM reconstruction, (e) AEM reconstruction.

Figure 3: Cameraman data set: (a) original image, (b) blurred and noisy image, (c) PAM reconstruction, (d) AEM reconstruction.

noise simulated by the `imnoise` function in the Matlab Image Processing Toolbox. The first PSF arises from confocal microscopy framework [33], while the second one is a Gaussian PSF with standard deviation $s = 1.3$. As before the background is set to 10^{-10} for both the test problems. Figures 2 and 3 show the microscopy and cameraman data sets (first two panels).

Both images are reconstructed using model (2) with the TV penalty (38). We fixed the regularization parameter μ equal to 0.09 and 0.0045 for the microscopy and cameraman images, respectively. As for the parameters involved in the description of the PKM method, we defined $\alpha_0 = 2.5$ and $\rho = 0.99$, while for PAM we put $\sigma = 10^{-4}$, $\beta = 0.5$ and α_n equal to 2.5 ($\forall n$) for the microscopy data set or 150 ($\forall n$) for the cameraman. For this application no benefits are gained by using a variable steplength, so we preferred to fix it during the iterations. As before the values chosen for ρ, σ and β do not affect the performance as much as the choice of α. Since the AEM method is designed for the image restoration problem involving the total variation regularization, we use this algorithm as a comparative tool to assess the efficiency of PKM and PAM.

Table 2 reports the values of the RRE reached by the proximal versions of Khobotov and Armijo methods and the AEM algorithm corresponding to different fixed numbers of iterations, for the microscopy test problem. Since the PKM and PAM schemes also have an inner routine to calculate an approximation of the TV proximal operator in a given point, for a better comparison with

nr. of	PKM			PAM			AEM	
iter.	RRE	time (s)	inn. it.	RRE	time (s)	inn. it.	RRE	time (s)
100	0.1126	92.32	17368	0.1040	1.74	200	0.1044	1.99
200	0.1066	157.88	29747	0.0964	3.40	400	0.0972	3.82
500	0.0988	364.45	69269	0.0901	8.26	1000	0.0908	9.29
1000	0.0935	718.85	136602	0.0893	16.34	2000	0.0896	18.40
1500	0.0913	1076.81	205104	0.0898	26.13	3190	0.0903	27.51

Table 2: Microscopy test problem: reconstruction error and computational time reached by the three different methods, corresponding to several fixed numbers of iterations. For PKM and PAM the total number of iterations of the inner solver (inn. it.) are also reported.

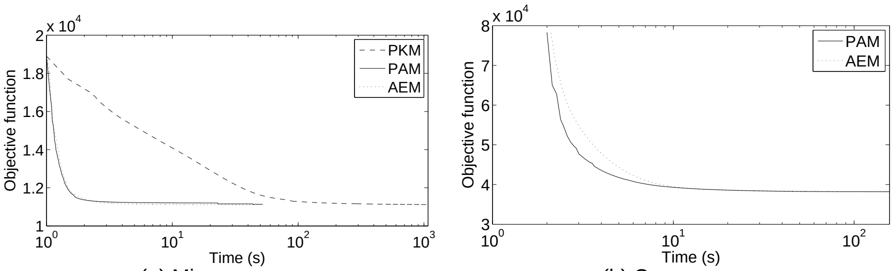
number of	PAM			AEM	
iterations	RRE	time (s)	inn. it.	RRE	time (s)
100	0.0920	10.70	200	0.0934	9.01
200	0.0889	21.09	400	0.0893	18.82
500	0.0867	52.14	1000	0.0868	45.51
1000	0.0864	103.65	2000	0.0865	89.62
1500	0.0865	155.01	3000	0.0866	133.71

Table 3: Cameraman test problem: relative reconstruction error (RRE) and computational time reached by PAM and AEM, corresponding to several fixed numbers of iterations. For PAM the total number of iterations of the inner solver (inn. it.) are also reported.

AEM, we also show the computational time needed by the methods to run until the settled numbers of iterations. Finally, for PKM and PAM, we present the total number of iterations of the inner solver (inn. it.); for PKM, as one can check from Algorithm 1, the number of the inner iterations, corresponding to one outer iteration, is the sum of the number of iterations required in STEP 1 (computed as many times as required by the line-search) and STEP 3.

From the results of Table 2, we can observe that PKM has a very slow convergence rate with respect to the other methods. We found out that the slow PKM behavior doesn't lie as much in the presence of the line-search on the steplength α_n, but rather in a quite high number of iterations required by the inner solver to satisfy the stopping criterion (42). Particularly, for PKM, the average number of iterations of the inner solver is 69. On the other hand the computational costs of PAM and AEM are comparable and, for PAM, the inner routine is not very expensive in terms of computational time (often just 2 inner iterations are sufficient). For these reasons, in solving the cameraman test problem, we decided to compare only the PAM and AEM algorithms: the results can be appreciated in Table 3.

For the sake of completeness, we also present the decrease of the objective function obtained by applying the different methods to the two problems (Figure 4), and the reconstructed images provided by the different methods at the iteration corresponding to the minimum RRE attained (Figures 2 and 3, right



(a) Microscopy (b) Cameraman

Figure 4: Objective functions provided by the different methods for the microscopy test problem (left panel) and the cameraman test problem (right panel)

hand panels). The algorithms provide the same minimum objective function value and similar image approximation. From the plots of the functional decrease it's again possible to conclude that PAM and AEM are faster in finding a solution with respect to PKM.

6. Conclusions

We have presented several proximal operator based iterative algorithms for the minimization of convex cost functions consisting of a smooth data misfit term and a possibly non-smooth penalty term. In case of the proximal Khobotov algorithm (14), we prove convergence to a minimizer of problem (1). We also show convergence in case the proximal operator cannot be computed exactly. In case of the proximal algorithm (27) with Armijo steplength rule (28), we show that any limit point is a minimizer of problem (1). The algorithms do not require knowledge of the Lipschitz constant of the gradient of the smooth term, which is useful in many cases in practice.

We have tested the proposed approaches in facing up to two different real-life applications: the compressed sensing problem and the image deblurring problem, both under Poisson noise. In the first case the proximal operator of the non-differentiable part of the objective function is known, in the second case an approximation strategy is necessary. Numerical experiments confirm the effectiveness of the suggested methods.

7. Acknowledgements

F.P. has been partially supported by the Italian Spinner 2013 PhD Project "High-complexity inverse problems in biomedical applications and social systems". F.P. also acknowledges the Italian GNCS-INdAM (Gruppo Nazionale per il Calcolo Scientifico - Istituto Nazionale di Alta Matematica). I. L. is a research associate of the F.R.S.-FNRS (Belgium).

Appendix A. Proof of Theorem 2

The proof of Theorem 2 relies on the following lemma (see also [34]).

Lemma 9. *Let $a_n, \gamma_n, \delta_n, \epsilon_n \geq 0$, $\sum_n \gamma_n < \infty$, $\sum_n \delta_n < \infty$, $\sum_n \epsilon_n < \infty$ and $a_{n+1}^2 \leq a_n^2 + a_n\gamma_n + \delta_n\epsilon_n$, then the sequence $\{a_n\}_{n\in\mathbb{N}}$ is bounded.*

Proof. One finds:

$$\begin{aligned}
a_{n+1} &\leq \sqrt{a_{n+1}^2 + \gamma_n^2/4} \\
&\leq \sqrt{(a_n + \gamma_n/2)^2 + \delta_n\epsilon_n} \\
&\leq (a_n + \gamma_n/2) + \sqrt{\delta_n\epsilon_n} \\
&\leq (a_n + \gamma_n/2) + (\delta_n + \epsilon_n)/2
\end{aligned}$$

and therefore $a_N \leq a_0 + \sum_{n=0}^{N-1}(\gamma_n + \delta_n + \epsilon_n)/2 \leq a_0 + \sum_{n=0}^{\infty}(\gamma_n + \delta_n + \epsilon_n)/2$ which is independent of N. □

We now prove Theorem 2.

Proof. Set $\bar{\boldsymbol{y}}_n = \bar{\boldsymbol{x}}_n - \boldsymbol{e}_n$ and $\boldsymbol{y}_n = \boldsymbol{x}_n - \boldsymbol{f}_{n-1}$. The iteration (19) becomes:

$$\begin{cases} \bar{\boldsymbol{y}}_n = \mathbf{prox}_{\alpha_n g}\left(\boldsymbol{y}_n - \alpha_n \nabla f(\boldsymbol{y}_n) + \bar{\boldsymbol{e}}_n\right) \\ \boldsymbol{y}_{n+1} = \mathbf{prox}_{\alpha_n g}\left(\boldsymbol{y}_n - \alpha_n \nabla f(\bar{\boldsymbol{y}}_n) + \bar{\boldsymbol{f}}_n\right), \end{cases} \tag{A.1}$$

where $\bar{\boldsymbol{e}}_n = \boldsymbol{f}_{n-1} + \alpha_n\left(\nabla f(\boldsymbol{y}_n) - \nabla f(\boldsymbol{y}_n + \boldsymbol{f}_{n-1})\right)$ and $\bar{\boldsymbol{f}}_n = \boldsymbol{f}_{n-1} + \alpha_n \nabla f(\bar{\boldsymbol{y}}_n) + -\alpha_n \nabla f(\bar{\boldsymbol{y}}_n + \boldsymbol{e}_n)$. One has that:

$$\|\bar{\boldsymbol{e}}_n\| \leq \|\boldsymbol{f}_{n-1}\| + \alpha_n \left\|\nabla f(\boldsymbol{y}_n) - \nabla f(\boldsymbol{y}_n + \boldsymbol{f}_{n-1})\right\| \leq \|\boldsymbol{f}_{n-1}\| + \alpha_{\max} L \|\boldsymbol{f}_{n-1}\|$$

and that:

$$\|\bar{\boldsymbol{f}}_n\| \leq \|\boldsymbol{f}_{n-1}\| + \alpha_n \|\nabla f(\bar{\boldsymbol{y}}_n) - \nabla f(\bar{\boldsymbol{y}}_n + \boldsymbol{e}_n)\| \leq \|\boldsymbol{f}_{n-1}\| + \alpha_{\max} L \|\boldsymbol{e}_n\|$$

as a result of the Lipschitz continuity of ∇f. Hence, it follows that $\sum_n \|\bar{\boldsymbol{e}}_n\| < +\infty$ and $\sum_n \|\bar{\boldsymbol{f}}_n\| < +\infty$. One also finds that:

$$\begin{aligned}
\|\bar{\boldsymbol{y}}_n - \hat{\boldsymbol{x}}\| &= \|\,\mathbf{prox}_{\alpha_n g}\left(\boldsymbol{y}_n - \alpha_n \nabla f(\boldsymbol{y}_n) + \bar{\boldsymbol{e}}_n\right) - \mathbf{prox}_{\alpha_n g}\left(\hat{\boldsymbol{x}} - \alpha_n \nabla f(\hat{\boldsymbol{x}})\right)\| \\
&\leq \|\boldsymbol{y}_n - \alpha_n \nabla f(\boldsymbol{y}_n) + \bar{\boldsymbol{e}}_n - \hat{\boldsymbol{x}} + \alpha_n \nabla f(\hat{\boldsymbol{x}})\| \\
&\leq \|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\| + \alpha_n \|\nabla f(\boldsymbol{y}_n) - \nabla f(\hat{\boldsymbol{x}})\| + \|\bar{\boldsymbol{e}}_n\| \\
&\leq \|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\| + \alpha_n L \|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\| + \|\bar{\boldsymbol{e}}_n\| \\
&= C_1 \|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\| + \|\bar{\boldsymbol{e}}_n\|,
\end{aligned} \tag{A.2}$$

(where C_1 does not depend on n because the sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ is bounded) and that:

$$
\begin{aligned}
\|\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n\| &= \left\|\mathbf{prox}_{\alpha_n g}\left(\boldsymbol{y}_n-\alpha_n\nabla f(\bar{\boldsymbol{y}}_n)+\bar{\boldsymbol{f}}_n\right)-\mathbf{prox}_{\alpha_n g}\left(\boldsymbol{y}_n-\alpha_n\nabla f(\boldsymbol{y}_n)+\bar{\boldsymbol{e}}_n\right)\right\| \\
&\leq \|-\alpha_n\nabla f(\bar{\boldsymbol{y}}_n)+\bar{\boldsymbol{f}}_n+\alpha_n\nabla f(\boldsymbol{y}_n)-\bar{\boldsymbol{e}}_n\| \\
&\leq \alpha_n\|\nabla f(\bar{\boldsymbol{y}}_n)-\nabla f(\boldsymbol{y}_n)\|+\|\bar{\boldsymbol{f}}_n-\bar{\boldsymbol{e}}_n\| \\
&\leq \alpha_n\, L\,\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|+\|\bar{\boldsymbol{f}}_n-\bar{\boldsymbol{e}}_n\| \\
&\leq \alpha_n\, L\,(\|\bar{\boldsymbol{y}}_n-\hat{\boldsymbol{x}}\|+\|\hat{\boldsymbol{x}}-\boldsymbol{y}_n\|)+\|\bar{\boldsymbol{f}}_n-\bar{\boldsymbol{e}}_n\| \\
&\overset{(A.2)}{\leq} \alpha_n\, L\,(C_1\|\boldsymbol{y}_n-\hat{\boldsymbol{x}}\|+\|\bar{\boldsymbol{e}}_n\|+\|\hat{\boldsymbol{x}}-\boldsymbol{y}_n\|)+\|\bar{\boldsymbol{f}}_n-\bar{\boldsymbol{e}}_n\| \\
&\leq C_2\|\boldsymbol{y}_n-\hat{\boldsymbol{x}}\|+C_3\|\bar{\boldsymbol{f}}_n-\bar{\boldsymbol{e}}_n\|+C_4\|\bar{\boldsymbol{e}}_n\|
\end{aligned}
\tag{A.3}
$$

(where C_2, C_3, C_4 do not depend on n because the sequence $\{\alpha_n\}_{n\in\mathbb{N}}$ is bounded).

On the other hand, using the same techniques as in the proof of Theorem 1 (see also equation (16)), one finds from iteration (A.1):

$$
\begin{aligned}
\|\boldsymbol{y}_{n+1}-\hat{\boldsymbol{x}}\|^2 &\leq \|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\
&\qquad +2\langle\boldsymbol{y}_{n+1}-\hat{\boldsymbol{x}},-\alpha_n\nabla f(\bar{\boldsymbol{y}}_n)+\bar{\boldsymbol{f}}_n\rangle+2\langle\hat{\boldsymbol{x}}-\bar{\boldsymbol{y}}_n,-\alpha_n\nabla f(\hat{\boldsymbol{x}})\rangle \\
&\qquad +2\langle\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1},-\alpha_n\nabla f(\boldsymbol{y}_n)+\bar{\boldsymbol{e}}_n\rangle \\
&= \|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\
&\qquad +2\alpha_n\langle\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n,\nabla f(\boldsymbol{y}_n)-\nabla f(\bar{\boldsymbol{y}}_n)\rangle \\
&\qquad +2\langle\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n,-\bar{\boldsymbol{e}}_n+\bar{\boldsymbol{f}}_n\rangle+2\langle\hat{\boldsymbol{x}}-\bar{\boldsymbol{y}}_n,-\bar{\boldsymbol{f}}_n\rangle \\
&\qquad +2\alpha_n\langle\hat{\boldsymbol{x}}-\bar{\boldsymbol{y}}_n,\nabla f(\bar{\boldsymbol{y}}_n)-\nabla f(\hat{x})\rangle.
\end{aligned}
$$

As $\langle\hat{\boldsymbol{x}}-\bar{\boldsymbol{y}}_n,\nabla f(\bar{\boldsymbol{y}}_n)-\nabla f(\hat{x})\rangle\leq 0$ (see Lemma 5), one has:

$$
\begin{aligned}
\|\boldsymbol{y}_{n+1}-\hat{\boldsymbol{x}}\|^2 &\leq \|\boldsymbol{y}_n-\hat{\boldsymbol{x}}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\
&\qquad +2\alpha_n\langle\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n,\nabla f(\boldsymbol{y}_n)-\nabla f(\bar{\boldsymbol{y}}_n)\rangle \\
&\qquad +2\langle\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n,-\bar{\boldsymbol{e}}_n+\bar{\boldsymbol{f}}_n\rangle+2\langle\hat{\boldsymbol{x}}-\bar{\boldsymbol{y}}_n,-\bar{\boldsymbol{f}}_n\rangle \\
&\leq \|\boldsymbol{y}_n-\hat{\boldsymbol{x}}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\
&\qquad +\|\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n\|^2+\alpha_n^2\|\nabla f(\boldsymbol{y}_n)-\nabla f(\bar{\boldsymbol{y}}_n)\|^2 \\
&\qquad +2\|\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n\|\,\|\bar{\boldsymbol{e}}_n-\bar{\boldsymbol{f}}_n\|+2\|\bar{\boldsymbol{y}}_n-\hat{\boldsymbol{x}}\|\,\|\bar{\boldsymbol{f}}_n\| \\
&= \|\boldsymbol{y}_n-\hat{\boldsymbol{x}}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2\left(1-\alpha_n^2\frac{\|\nabla f(\boldsymbol{y}_n)-\nabla f(\bar{\boldsymbol{y}}_n)\|^2}{\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2}\right) \\
&\qquad +2\|\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n\|\,\|\bar{\boldsymbol{e}}_n-\bar{\boldsymbol{f}}_n\|+2\|\bar{\boldsymbol{y}}_n-\hat{\boldsymbol{x}}\|\,\|\bar{\boldsymbol{f}}_n\| \\
&\overset{(13)}{\leq} \|\boldsymbol{y}_n-\hat{\boldsymbol{x}}\|^2-(1-\rho^2)\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2+2\|\boldsymbol{y}_{n+1}-\bar{\boldsymbol{y}}_n\|\,\|\bar{\boldsymbol{e}}_n-\bar{\boldsymbol{f}}_n\| \\
&\qquad +2\|\bar{\boldsymbol{y}}_n-\hat{\boldsymbol{x}}\|\,\|\bar{\boldsymbol{f}}_n\|.
\end{aligned}
\tag{A.4}
$$

Combining expressions (A.4), (A.2) and (A.3), one finally finds the inequality:

$$\|\boldsymbol{y}_{n+1} - \hat{\boldsymbol{x}}\|^2 \leq \|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\|^2 - (1-\rho^2)\|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|^2 + \|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\|\,\gamma_n + \delta_n\epsilon_n, \tag{A.5}$$

where γ_n, ϵ_n and δ_n are three non-negative summable sequences (related to $\bar{\boldsymbol{e}}_n$ and $\bar{\boldsymbol{f}}_n$). It now follows from Lemma 9 that the sequence $\{\boldsymbol{y}_n\}_{n\in\mathbb{N}}$ is bounded. Therefore, a converging subsequence exists: $\boldsymbol{y}_{n_j} \overset{j\to\infty}{\longrightarrow} \boldsymbol{y}^\dagger$. On the other hand, the inequality (A.5) also implies that:

$$(1-\rho^2)\sum_{n=M}^{N-1}\|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|^2 \leq \|\boldsymbol{y}_M - \hat{\boldsymbol{x}}\|^2 - \|\boldsymbol{y}_N - \hat{\boldsymbol{x}}\|^2 + \sum_{n=M}^{N-1}\|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\|\,\gamma_n + \delta_n\epsilon_n \tag{A.6}$$

for $N > M$.

As $\sum_{n=M}^{N-1}\|\boldsymbol{y}_n - \hat{\boldsymbol{x}}\|\,\gamma_n + \delta_n\epsilon_n \leq \sum_{n=M}^{N-1} C_5\gamma_n + \delta_n\epsilon_n \leq C_6$ (independent of N), one finds that $\sum_{n=M}^{N-1}\|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|^2 < \infty$ and hence that $\|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\| \longrightarrow 0$ as $n \to \infty$. Making use of Remark 3, and possibly a further converging subsequence of $\{\alpha_n\}_{n\in\mathbb{N}}$, this in turn implies that $\boldsymbol{y}^\dagger$ is a fixed-point of the iteration (a minimizer of problem (1)).

Replacing $\hat{\boldsymbol{x}}$ by $\boldsymbol{y}^\dagger$ in inequality (A.6) yields:

$$\|\boldsymbol{y}_N - \boldsymbol{y}^\dagger\|^2 \leq \|\boldsymbol{y}_M - \boldsymbol{y}^\dagger\|^2 + \sum_{n=M}^{N-1}\|\boldsymbol{y}_n - \boldsymbol{y}^\dagger\|\,\gamma_n + \delta_n\epsilon_n \tag{A.7}$$

for $N > M$. The right hand side can be made arbitrarily small by choosing M appropriately (the first term because a subsequence converges to zero, the second term because the series $\sum_n \|\boldsymbol{y}_n - \boldsymbol{y}^\dagger\|\,\gamma_n + \delta_n\epsilon_n$ converges). This proves that the whole sequence $\{\boldsymbol{y}_n\}_{n\in\mathbb{N}}$ converges to $\boldsymbol{y}^\dagger$. □

Appendix B. Proof of Theorem 3

Proof. Let $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ be a saddle point of problem (26), i.e.

$$f(\hat{\boldsymbol{x}}) + F(\hat{\boldsymbol{x}}, \boldsymbol{y}) - g(\boldsymbol{y}) \leq f(\hat{\boldsymbol{x}}) + F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) - g(\hat{\boldsymbol{y}}) \leq f(\boldsymbol{x}) + F(\boldsymbol{x}, \hat{\boldsymbol{y}}) - g(\hat{\boldsymbol{y}}) \quad \forall \boldsymbol{x}, \boldsymbol{y}, \tag{B.1}$$

or equivalently

$$\hat{\boldsymbol{x}} = \mathbf{prox}_{\alpha g_1}\left[\hat{\boldsymbol{x}} - \alpha\nabla_x F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})\right], \qquad \hat{\boldsymbol{y}} = \mathbf{prox}_{\alpha g_2}\left[\hat{\boldsymbol{y}} + \alpha\nabla_y F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})\right], \qquad \alpha > 0. \tag{B.2}$$

Let $(\boldsymbol{x}_n, \boldsymbol{y}_n, \bar{\boldsymbol{y}}_n)_{n\in\mathbb{N}}$ satisfy the iteration (25). It follows from the first line of iteration (25) and Lemma 4 that:

$$\begin{aligned}
\|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}\|^2 \quad &\leq \quad \|\boldsymbol{y}_n - \boldsymbol{y}_{n+1}\|^2 - \|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|^2 + 2\left\langle \bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}, \alpha_n \nabla_y F(\boldsymbol{x}_n, \boldsymbol{y}_n)\right\rangle \\
&\qquad +2\alpha_n g(\boldsymbol{y}_{n+1}) - 2\alpha_n g(\bar{\boldsymbol{y}}_n) \\
&= \quad \|\boldsymbol{y}_n - \boldsymbol{y}_{n+1}\|^2 - \|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|^2 + 2\alpha_n \left\langle \bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}, \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\right\rangle \\
&\qquad +2\alpha_n \left\langle \bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}, \nabla_y F(\boldsymbol{x}_n, \boldsymbol{y}_n) - \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\right\rangle \\
&\qquad +2\alpha_n g(\boldsymbol{y}_{n+1}) - 2\alpha_n g(\bar{\boldsymbol{y}}_n)
\end{aligned} \tag{B.3}$$

and it follows from the third line of iteration (25) and Lemma 4 that:

$$\begin{aligned}
\|\boldsymbol{y}_{n+1} - \hat{\boldsymbol{y}}\|^2 \quad &\leq \quad \|\boldsymbol{y}_n - \hat{\boldsymbol{y}}\|^2 - \|\boldsymbol{y}_{n+1} - \boldsymbol{y}_n\|^2 + 2\left\langle \boldsymbol{y}_{n+1} - \hat{\boldsymbol{y}}, \alpha_n \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\right\rangle \\
&\qquad +2\alpha_n g(\hat{\boldsymbol{y}}) - 2\alpha_n g(\boldsymbol{y}_{n+1}).
\end{aligned} \tag{B.4}$$

Summing inequalities (B.3) and (B.4), and rearranging the result yields:

$$\begin{aligned}
\|\boldsymbol{y}_{n+1} - \hat{\boldsymbol{y}}\|^2 \quad &\leq \quad \|\boldsymbol{y}_n - \hat{\boldsymbol{y}}\|^2 - \|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}\|^2 - \|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|^2 \\
&\qquad +2\alpha_n \langle \boldsymbol{y}_n - \hat{\boldsymbol{y}}, \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\rangle \\
&\qquad +2\alpha_n \langle \bar{\boldsymbol{y}}_n - \boldsymbol{y}_n, \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\rangle \\
&\qquad +2\alpha_n \left\langle \bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}, \nabla_y F(\boldsymbol{x}_n, \boldsymbol{y}_n) - \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\right\rangle \\
&\qquad -2\alpha_n g(\bar{\boldsymbol{y}}_n) + 2\alpha_n g(\hat{\boldsymbol{y}}).
\end{aligned} \tag{B.5}$$

By concavity of F in the second variable one has: $\langle \boldsymbol{y}_n - \hat{\boldsymbol{y}}, \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\rangle \leq F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n) - F(\boldsymbol{x}_{n+1}, \hat{\boldsymbol{y}})$ and $F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n) \leq \langle \boldsymbol{y}_n - \bar{\boldsymbol{y}}_n, \nabla_y F(\boldsymbol{x}_{n+1}, \bar{\boldsymbol{y}}_n)\rangle + F(\boldsymbol{x}_{n+1}, \bar{\boldsymbol{y}}_n)$. It thus follows that:

$$\begin{aligned}
\|\boldsymbol{y}_{n+1} - \hat{\boldsymbol{y}}\|^2 \quad &\leq \quad \|\boldsymbol{y}_n - \hat{\boldsymbol{y}}\|^2 - \|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}\|^2 - \|\bar{\boldsymbol{y}}_n - \boldsymbol{y}_n\|^2 \\
&\qquad +2\alpha_n \langle \bar{\boldsymbol{y}}_n - \boldsymbol{y}_n, \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n) - \nabla_y F(\boldsymbol{x}_{n+1}, \bar{\boldsymbol{y}}_n)\rangle \\
&\qquad +2\alpha_n \left\langle \bar{\boldsymbol{y}}_n - \boldsymbol{y}_{n+1}, \nabla_y F(\boldsymbol{x}_n, \boldsymbol{y}_n) - \nabla_y F(\boldsymbol{x}_{n+1}, \boldsymbol{y}_n)\right\rangle \\
&\qquad +2\alpha_n \left[F(\boldsymbol{x}_{n+1}, \bar{\boldsymbol{y}}_n) - F(\boldsymbol{x}_{n+1}, \hat{\boldsymbol{y}})\right] \\
&\qquad -2\alpha_n g(\bar{\boldsymbol{y}}_n) + 2\alpha_n g(\hat{\boldsymbol{y}}).
\end{aligned} \tag{B.6}$$

On the other hand, it follows from the second line of algorithm (25) and

Lemma 4 that:

$$\begin{aligned}
\|\boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}}\|^2 &\leq \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2-\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2+2\left\langle \boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}},-\alpha_n\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\right\rangle \\
&\quad +2\alpha_n f(\hat{\boldsymbol{x}})-2\alpha_n f(\boldsymbol{x}_{n+1}) \\
&= \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2-\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2-2\alpha_n\left\langle \boldsymbol{x}_{n+1}-\boldsymbol{x}_n,\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\right\rangle \\
&\quad +2\alpha_n\left\langle \hat{\boldsymbol{x}}-\boldsymbol{x}_n,\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\right\rangle+2\alpha_n f(\hat{\boldsymbol{x}})-2\alpha_n f(\boldsymbol{x}_{n+1}).
\end{aligned} \tag{B.7}$$

By convexity of F in the first variable one has: $\langle \hat{\boldsymbol{x}}-\boldsymbol{x}_n,\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\rangle \leq F(\hat{\boldsymbol{x}},\bar{\boldsymbol{y}}_n)-F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)$ and $-F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\leq \langle\nabla_x F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n),\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\rangle-F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)$. We thus find that:

$$\begin{aligned}
\|\boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}}\|^2 &\leq \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2-\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2 \\
&\quad +2\alpha_n\left\langle \boldsymbol{x}_{n+1}-\boldsymbol{x}_n,\nabla_x F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)-\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\right\rangle \\
&\quad +2\alpha_n\left[F(\hat{\boldsymbol{x}},\bar{\boldsymbol{y}}_n)-F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)\right]+2\alpha_n f(\hat{\boldsymbol{x}})-2\alpha_n f(\boldsymbol{x}_{n+1}).
\end{aligned} \tag{B.8}$$

Adding inequalities (B.6) and (B.8) one finds:

$$\begin{aligned}
\|\boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_{n+1}-\hat{\boldsymbol{y}}\|^2 &\leq \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2 \\
&\quad -\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\
&\quad +2\alpha_n\left\langle \bar{\boldsymbol{y}}_n-\boldsymbol{y}_n,\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)\right\rangle \\
&\quad +2\alpha_n\left\langle \bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1},\nabla_y F(\boldsymbol{x}_n,\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)\right\rangle \\
&\quad +2\alpha_n\left\langle \boldsymbol{x}_{n+1}-\boldsymbol{x}_n,\nabla_x F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)-\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\right\rangle \\
&\quad +2\alpha_n\left[-F(\boldsymbol{x}_{n+1},\hat{\boldsymbol{y}})-g(\bar{\boldsymbol{y}}_n)+g(\hat{\boldsymbol{y}})\right. \\
&\quad \left.+F(\hat{\boldsymbol{x}},\bar{\boldsymbol{y}}_n)+f(\hat{\boldsymbol{x}})-f(\boldsymbol{x}_{n+1})\right]
\end{aligned} \tag{B.9}$$

As $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ is a saddle point, one has:

$$f(\hat{\boldsymbol{x}})+F(\hat{\boldsymbol{x}},\bar{\boldsymbol{y}}_n)-g(\bar{\boldsymbol{y}}_n)\leq f(\hat{\boldsymbol{x}})+F(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})-g(\hat{\boldsymbol{y}})\leq f(\boldsymbol{x}_{n+1})+F(\boldsymbol{x}_{n+1},\hat{\boldsymbol{y}})-g(\hat{\boldsymbol{y}}), \tag{B.10}$$

such that the expression in square brackets in (B.9) is negative, and one finally finds:

$$\begin{aligned}
\|\boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_{n+1}-\hat{\boldsymbol{y}}\|^2 &\leq \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2 \\
&\quad -\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\
&\quad +2\alpha_n\left\langle \bar{\boldsymbol{y}}_n-\boldsymbol{y}_n,\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)\right\rangle \\
&\quad +2\alpha_n\left\langle \bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1},\nabla_y F(\boldsymbol{x}_n,\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)\right\rangle \\
&\quad +2\alpha_n\left\langle \boldsymbol{x}_{n+1}-\boldsymbol{x}_n,\nabla_x F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)-\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\right\rangle.
\end{aligned} \tag{B.11}$$

The remainder of the proof now follows the same lines as the proof of convergence of algorithm 22 (see [8, Theorem 1]). Using the Cauchy-Schwarz identity three times in inequality (B.11), one has:

$$\begin{aligned}\|\boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_{n+1}-\hat{\boldsymbol{y}}\|^2 \quad \leq \quad & \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2 \\ & -\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\ & +2\alpha_n\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|\ \|\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)\| \\ & +2\alpha_n\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|\ \|\nabla_y F(\boldsymbol{x}_n,\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)\| \\ & +2\alpha_n\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|\ \|\nabla_x F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)-\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\|\,.\end{aligned} \tag{B.12}$$

From the definition (25) of $\bar{\boldsymbol{y}}_n$ and $\boldsymbol{y}_{n+1}$, and from the Lipschitz continuity of the proximal operators (see lemma 2), one also finds:

$$\begin{aligned}\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\| \quad & \leq \quad \left\|\mathbf{prox}_{\alpha_n g}\left[\boldsymbol{y}_n+\alpha_n\nabla_y F(\boldsymbol{x}_n,\boldsymbol{y}_n)\right]-\mathbf{prox}_{\alpha_n g}\left[\boldsymbol{y}_n+\alpha_n\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)\right]\right\| \\ & \leq \quad \alpha_n\left\|\nabla_y F(\boldsymbol{x}_n,\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)\right\|\end{aligned} \tag{B.13}$$

such that (B.12) reduces to:

$$\begin{aligned}\|\boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_{n+1}-\hat{\boldsymbol{y}}\|^2 \quad \leq \quad & \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2 \\ & -\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\ & +2\alpha_n\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|\ \|\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)\| \\ & +2\alpha_n^2\left\|\nabla_y F(\boldsymbol{x}_n,\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)\right\|^2 \\ & +2\alpha_n\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|\ \|\nabla_x F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)-\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\|\,,\end{aligned} \tag{B.14}$$

or:

$$
\begin{aligned}
\|\boldsymbol{x}_{n+1}-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_{n+1}-\hat{\boldsymbol{y}}\|^2 \quad \leq \quad & \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2 \\
& -\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \\
& +2\alpha_n\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \frac{\|\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)\|}{\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|} \\
& +2\alpha_n^2\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2 \frac{\|\nabla_y F(\boldsymbol{x}_n,\boldsymbol{y}_n)-\nabla_y F(\boldsymbol{x}_{n+1},\boldsymbol{y}_n)\|^2}{\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2} \\
& +2\alpha_n\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2 \frac{\|\nabla_x F(\boldsymbol{x}_{n+1},\bar{\boldsymbol{y}}_n)-\nabla_x F(\boldsymbol{x}_n,\bar{\boldsymbol{y}}_n)\|}{\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|} \\
= \quad & \|\boldsymbol{x}_n-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_n-\hat{\boldsymbol{y}}\|^2-\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2 \\
& -\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2\left(1-2\alpha_n A_n-2\alpha_n^2 B_n^2\right) \\
& -\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2\left(1-2\alpha_n C_n\right).
\end{aligned}
\tag{B.15}
$$

It follows from the conditions (23) that the iterates $(\boldsymbol{x}_n,\boldsymbol{y}_n)$ are bounded and that a converging subsequence exists: $(\boldsymbol{x}_{n_j},\boldsymbol{y}_{n_j}) \overset{j\to\infty}{\longrightarrow} (\boldsymbol{x}^\dagger,\boldsymbol{y}^\dagger)$. One also finds that

$$
\begin{aligned}
\|\boldsymbol{x}_N-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_N-\hat{\boldsymbol{y}}\|^2 \quad \leq \quad & \|\boldsymbol{x}_M-\hat{\boldsymbol{x}}\|^2+\|\boldsymbol{y}_M-\hat{\boldsymbol{y}}\|^2-\sum_{n=M}^{N-1}\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|^2 \\
& +\epsilon\,\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|^2+\epsilon\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|^2 \qquad (N>M),
\end{aligned}
\tag{B.16}
$$

which implies that $\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_{n+1}\|$, $\|\boldsymbol{x}_{n+1}-\boldsymbol{x}_n\|$ and $\|\bar{\boldsymbol{y}}_n-\boldsymbol{y}_n\|$ tend to zero as n tends to infinity. It thus follows that $\boldsymbol{x}_{n_j+1}$ (resp. $\bar{\boldsymbol{y}}_{n_j+1},\boldsymbol{y}_{n_j+1}$) also tend to $\boldsymbol{x}^\dagger$ (resp. $\boldsymbol{y}^\dagger$).

By considering a further subsequence for which $\alpha_{n_{j_k}} \overset{k\to\infty}{\longrightarrow} \alpha \neq 0$, it follows then from relations (25) and from the continuity of the proximal operators and of $\nabla_x F$ and of $\nabla_y F$ that $(\boldsymbol{x}^\dagger,\boldsymbol{y}^\dagger)$ satisfies the fixed-point equations (B.2).

Replacing $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ by $(\boldsymbol{x}^\dagger,\boldsymbol{y}^\dagger)$ in expression (B.16) shows that the whole sequence $\{(\boldsymbol{x}_n,\boldsymbol{y}_n)\}_{n\in\mathbb{N}}$ converges to the saddle-point $(\boldsymbol{x}^\dagger,\boldsymbol{y}^\dagger)$. □

References

[1] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Optimization with sparsity-inducing penalties, Foundations and Trends in Machine Learning 4 (2011) 1–106.

[2] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning 3 (2011) 1–222.

[3] P. L. Combettes, J.-C. Pesquet, Proximal splitting methods in signal processing, in: H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, H. Wolkowicz (Eds.), Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Berlin: Springer, 2011, p. 185–212.

[4] N. Parikh, S. Boyd, Proximal algorithms, Foundations and Trends in Optimization 1 (2014) 123–231.

[5] E. N. Khobotov, Modification of the extragradient method for solving variational inequalities and certain optimization problems, USSR Computat. Math. Phys. 27 (1987) 120–127.

[6] G. Korpelevich, The extragradient method for finding saddle points and other problems, Ekonomika i Matematicheskie Metody 12 (1976) 747–756.

[7] D. P. Bertsekas, Nonlinear Programming, Athena Scientific, 2nd edition, 1999.

[8] S. Bonettini, V. Ruggiero, An alternating extragradient method for total variation based image restoration from Poisson data, Inverse Probl. 27 (2011) 095001.

[9] Y. Huang, Y. Dong, New properties of forward-backward splitting and a practical proximal-descent algorithm, Applied Mathematics and Computation 237 (2014) 60–68.

[10] P. Tseng, S. Yun, A coordinate gradient descent method for nonsmooth separable minimization, Math. Program. Ser. B 117 (2009) 387–423.

[11] M. Elad, B. Matalon, M. Zibulevsky, Image denoising with shrinkage and redundant representations, in: Proc. IEEE Conf. Comput. Vision Pattern Recogn., 2006.

[12] M. Figueiredo, R. Nowak, Wavelet-based image estimation: An empirical Bayes' approach using Jeffreys' noninformative prior, IEEE Trans. Image Process. 10 (2001) 1322–1331.

[13] L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D 60 (1992) 259–268.

[14] M. Raginsky, R. Willett, Z. Harmany, R. Marcia, Compressed sensing performance bounds under Poisson noise, IEEE Transactions on Signal Processing 58 (2010) 3990–4002.

[15] Z. Harmany, R. Marcia, R. Willett, This is SPIRAL-TAP: Sparse Poisson Intensity Reconstruction ALgorithms—Theory and practice, IEEE Transactions on Image Processing 21 (2012) 1084–1096.

[16] M. Bertero, P. Boccacci, Introduction to Inverse Problems in Imaging, Bristol: Institute of Physics Publishing, 1998.

[17] L. Shepp, V. Y., Maximum likelihood reconstruction for emission tomography, Trans. Med. Imaging MI–1 (1982) 113–122.

[18] S. Geman, G. D., Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images, IEEE Trans. Pattern Analysis and Machine Intelligence 6 (1984) 721–741.

[19] P. L. Combettes, V. R. Wajs, Signal recovery by proximal forward–backward splitting multiscale model, Multiscale Model. Simul. 4 (2005) 1168–1200.

[20] J. J. Moreau, Proximité et dualité dans un espace hilbertien, Bull. Soc. Math. France 93 (1965) 273–299.

[21] A. Beck, M. Teboulle, A fast iterative shrinkage-threshold algorithm for linear inverse problems, SIAM J. Imaging Sci. 2 (2009) 183–202.

[22] A. Beck, M. Teboulle, Gradient-based algorithms with applications to signal recovery problems, in: D. Palomar, Y. Eldar (Eds.), Convex Optimization in Signal Processing and Communications, Cambribge University Press, 2010, p. 42–88.

[23] A. Iusem, Augmented Lagrangian methods and proximal point methods for convex optimization, Investigación Operativa 8 (1999) 11–49.

[24] B. Lemaire, The proximal algorithm, International Series of Numerical Mathematics (1989) 73–87.

[25] P. Marcotte, Application of Khobotov's algorithm to variational inequalities and network equilibrium problems, INFOR. 29 (1991) 258–270.

[26] Z. Harmany, R. Marcia, R. Willett, The Sparse Poisson Intensity Reconstruction ALgorithms (SPIRAL) toolbox, 2012. URL: `http://drz.ac/code/spiraltap/`, `http://drz.ac/code/spiraltap/`.

[27] J. Barzilai, J. Borwein, Two point step size gradient methods, IMA J. Numer. Anal. 8 (1988) 141–148.

[28] B. Zhou, L. Gao, Y. H. Dai, Gradient methods with adaptive step-sizes, Computional Optimization and Applications 35 (2006) 69–86.

[29] S. Osher, L. Rudin, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D. 60 (1992) 259–268.

[30] S. Alliney, An algorithm for the minimization of mixed l1 and l2 norms with application to bayesian estimation, IEEE Signal Processing 42 (1994) 618–627.

[31] S. Osher, A. Solè, L. Vese, Image decomposition and restoration using total variation minimization and the h1 norm, SIAM Multiscale Modeling and Simulation 1 (2003) 349–370.

[32] A. Chambolle, An algorithm for total variation minimization and applications, Journal of Mathematical Imaging and Vision 20 (2004) 89–97.

[33] R. M. Willett, R. Novak, Platelets: A multiscale approach for recovering edges and surfaces in photon limited medical imaging, IEEE Transactions on Medical Imaging 22 (2003) 332–350.

[34] P. L. Combettes, Quasi-Fejérian analysis of some optimization algorithms, Elsevier, New York., 2001, p. 115–152. Book: Inherently Parallel Algorithms for Feasibility and Optimization.