Journal of Optimization Theory and Applications (2019) 182:110–132
https://doi.org/10.1007/s10957-018-01454-y

The Proximal Alternating Minimization Algorithm for Two-Block Separable Convex Optimization Problems with Linear Constraints

Sandy Bitterlich[1] · **Radu Ioan Boţ**[2] · **Ernö Robert Csetnek**[2,3] · **Gert Wanka**[1]

Received: 1 June 2018 / Accepted: 29 November 2018 / Published online: 24 December 2018

Abstract
The Alternating Minimization Algorithm has been proposed by Paul Tseng to solve convex programming problems with two-block separable linear constraints and objectives, whereby (at least) one of the components of the latter is assumed to be strongly convex. The fact that one of the subproblems to be solved within the iteration process of this method does not usually correspond to the calculation of a proximal operator through a closed formula affects the implementability of the algorithm. In this paper, we allow in each block of the objective a further smooth convex function and propose a proximal version of the algorithm, which is achieved by equipping the algorithm with proximal terms induced by variable metrics. For suitable choices of the latter, the solving of the two subproblems in the iterative scheme can be reduced to the computation of proximal operators. We investigate the convergence of the proposed algorithm in a real Hilbert space setting and illustrate its numerical performances on two applications in image processing and machine learning.

Keywords Proximal AMA · Lagrangian · Saddle points · Subdifferential · Convex optimization · Fenchel duality

✉ Radu Ioan Boţ
radu.bot@univie.ac.at

Sandy Bitterlich
sandy.bitterlich@mathematik.tu-chemnitz.de

Ernö Robert Csetnek
ernoe.robert.csetnek@univie.ac.at

Gert Wanka
gert.wanka@mathematik.tu-chemnitz.de

[1] Faculty of Mathematics, Chemnitz University of Technology, 09126 Chemnitz, Germany

[2] University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

[3] Institute for Numerical and Applied Mathematics, University of Göttingen, 37083 Göttingen, Germany

Mathematics Subject Classification 47H05 · 65K05 · 90C25

1 Introduction

Tseng introduced in [1] the so-called Alternating Minimization Algorithm (AMA) to solve optimization problems with two-block separable linear constraints and two nonsmooth convex objective functions, one of these assumed to be strongly convex. The numerical scheme consists in each iteration of two minimization subproblems, each involving one of the two objective functions, and of an update of the dual sequence which approaches asymptotically a Lagrange multiplier of the dual problem.

The strong convexity of one of the objective functions allows to reduce the corresponding minimization subproblem to the calculation of the proximal operator of a proper, convex and lower semicontinuous function. This is for the second minimization problem in general not the case; thus, with the exception of some very particular cases, one has to use a subroutine in order to compute the corresponding iterate. This may have a negative influence on the convergence behaviour of the algorithm and affect its computational tractability. One possibility to avoid this is to properly modify this subproblem with the aim of transforming it into a proximal step, and, of course, without losing the convergence properties of the algorithm. The papers [2] and [3] provide convincing evidences for the efficiency and versatility of proximal point algorithms for solving nonsmooth convex optimization problems; we also refer to [4] for a block coordinate variable metric forward–backward method.

In this paper, we address in a real Hilbert space setting a more involved two-block separable optimization problem, which is obtained by adding in each block of the objective a further smooth convex function. To solve this problem, we propose a so-called Proximal Alternating Minimization Algorithm (Proximal AMA), which is obtained by inducing in each of the minimization subproblems additional proximal terms defined by means of positively semidefinite operators. The two smooth convex functions in the objective are evaluated via gradient steps. For appropriate choices of these operators, we show that the minimization subproblems turn into proximal steps and the algorithm becomes an iterative scheme formulated in the spirit of the full splitting paradigm. We show that the generated sequence converges weakly to a saddle point of the Lagrangian associated with the optimization problem under investigation. The numerical performances of Proximal AMA are illustrated in particular in comparison with AMA for two applications in image processing and machine learning.

A similarity of AMA to the classical Alternating Direction Method of Multipliers (ADMM) algorithm, introduced by Gabay and Mercier [5], is obvious. In [6–8] (see also [9,10]), proximal versions of the ADMM algorithm have been proposed and proved to provide a unifying framework for primal-dual algorithms for convex optimization. Parts of the convergence analysis for the Proximal AMA are carried out in a similar spirit to the convergence proofs in these papers.

2 Preliminaries

The convex optimization problems addressed in [1] are of the form

$$\inf_{x\in R^n, z\in\mathbb{R}^m} f(x)+g(z) \quad \text{s.t.} \quad Ax+Bz=b, \tag{1}$$

where $f:\mathbb{R}^n\to\overline{\mathbb{R}}:=\mathbb{R}\cup\{\pm\infty\}$ is a proper, γ-strongly convex with $\gamma>0$ (this means that $f-\frac{\gamma}{2}\|\cdot\|^2$ is convex) and lower semicontinuous function, $g:\mathbb{R}^m\to\overline{\mathbb{R}}$ is a proper, convex and lower semicontinuous function, $A\in\mathbb{R}^{r\times n}$, $B\in\mathbb{R}^{r\times m}$ and $b\in\mathbb{R}^r$.

For $c>0$, the augmented Lagrangian associated with problem (1), $L_c:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^r\to\overline{\mathbb{R}}$ reads

$$L_c(x,z,p)=f(x)+g(z)+\langle p,b-Ax-Bz\rangle+\frac{c}{2}\|Ax+Bz-b\|^2.$$

The Lagrangian associated with problem (1) is

$$L:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^r\to\overline{\mathbb{R}},\quad L(x,z,p)=f(x)+g(z)+\langle p,b-Ax-Bz\rangle.$$

Tseng proposed the following so-called Alternating Minimization Algorithm (AMA) for solving (1):

Algorithm 2.1 *(AMA)* Choose $p^0\in\mathbb{R}^r$ and a sequence of strictly positive stepsizes $(c_k)_{k\geq 0}$. For all $k\geq 0$, set:

$$x^k=\operatorname{argmin}_{x\in\mathbb{R}^n}\left\{f(x)-\langle p^k,Ax\rangle\right\} \tag{2}$$

$$z^k\in\operatorname{argmin}_{z\in\mathbb{R}^m}\left\{g(z)-\langle p^k,Bz\rangle+\frac{c_k}{2}\|Ax^k+Bz-b\|^2\right\} \tag{3}$$

$$p^{k+1}=p^k+c_k(b-Ax^k-Bz^k). \tag{4}$$

The main convergence properties of this numerical algorithm are summarized in the theorem below (see [1]).

Theorem 2.1 *Let $A\neq 0$ and $(x,z)\in\operatorname{ri}(\operatorname{dom} f)\times\operatorname{ri}(\operatorname{dom} g)$ be such that the equality $Ax+Bz=b$ holds. Assume that the sequence of stepsizes $(c_k)_{k\geq 0}$ satisfies*

$$\epsilon\leq c_k\leq\frac{2\gamma}{\|A\|^2}-\epsilon\quad\forall k\geq 0,$$

where $0<\epsilon<\frac{\gamma}{\|A\|^2}$. Let $(x^k,z^k,p^k)_{k\geq 0}$ be the sequence generated by Algorithm 2.1. Then there exist $x^*\in\mathbb{R}^n$ and an optimal Lagrange multiplier $p^*\in\mathbb{R}^r$ associated with the constraint $Ax+Bz=b$ such that

$$x^k\to x^*,\quad Bz^k\to b-Ax^*,\quad p^k\to p^*(k\to+\infty).$$

If the function $z \mapsto g(z) + \|Bz\|^2$ has bounded level sets, then $(z^k)_{k\geq 0}$ is bounded and any of its cluster points z^* provides with (x^*, z^*) an optimal solution of (1).

It is the aim of this paper to propose a proximal variant of this algorithm, called Proximal AMA, which overcomes its drawbacks, and to investigate its convergence properties.

In the remainder of this section, we will introduce some notations, definitions and basic properties that will be used in the sequel (see [11]). Let $\mathcal{H}$ and $\mathcal{G}$ be real Hilbert spaces with corresponding inner products $\langle \cdot, \cdot \rangle$ and associated norms $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$. In both spaces, we denote by $\rightharpoonup$ the weak convergence and by $\to$ the strong convergence.

We say that a function $f : \mathcal{H} \to \overline{\mathbb{R}}$ is proper, if its domain satisfies the assumption $\text{dom}\, f := \{x \in \mathcal{H} : f(x) < +\infty\} \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathcal{H}$. Let be $\Gamma(\mathcal{H}) = \{f : \mathcal{H} \to \overline{\mathbb{R}} : f \text{ is proper, convex and lower semicontinuous}\}$.

The (Fenchel) conjugate function $f^* : \mathcal{H} \to \overline{\mathbb{R}}$ of a function $f \in \Gamma(\mathcal{H})$ is defined as

$$f^*(p) = \sup_{x\in\mathcal{H}}\{\langle p, x\rangle - f(x)\} \quad \forall p \in \mathcal{H}$$

and is a proper, convex and lower semicontinuous function. It also holds $f^{**} = f$, where f^{**} is the conjugate function of f^*. The convex subdifferential of f is defined as $\partial f(x) = \{u \in \mathcal{H} : f(y) \geq f(x) + \langle u, y - x\rangle \forall y \in \mathcal{H}\}$, if $f(x) \in \mathbb{R}$, and as $\partial f(x) = \emptyset$, otherwise.

The infimal convolution of two proper functions $f, g : \mathcal{H} \to \overline{\mathbb{R}}$ is the function $f\square g : \mathcal{H} \to \overline{\mathbb{R}}$, defined by $(f\square g)(x) = \inf_{y\in\mathcal{H}}\{f(y) + g(x - y)\}$.

The proximal point operator of parameter γ of f at x, where $\gamma > 0$, is defined as

$$\text{Prox}_{\gamma f} : \mathcal{H} \to \mathcal{H}, \quad \text{Prox}_{\gamma f}(x) = \text{argmin}_{y\in\mathcal{H}}\left\{\gamma f(y) + \frac{1}{2}\|y - x\|^2\right\}.$$

According to Moreau's decomposition formula, we have

$$\text{Prox}_{\gamma f}(x) + \gamma\, \text{Prox}_{(1/\gamma)f^*}(\gamma^{-1}x) = x, \quad \forall x \in \mathcal{H}.$$

Let $C \subseteq \mathcal{H}$ be a convex and closed set. The strong quasi-relative interior of C is

$$\text{sqri}(C) = \{x \in C : \cup_{\lambda>0}\lambda(C - x) \text{ is a closed linear subspace of } \mathcal{H}\}.$$

We always have $\text{int}(C) \subseteq \text{sqri}(C)$, and if $\mathcal{H}$ is finite dimensional, then $\text{sqri}(C) = \text{ri}(C)$, where $\text{ri}(C)$ denotes the interior of C relative to its affine hull.

We denote by $S_+(\mathcal{H})$ the set of operators from $\mathcal{H}$ to $\mathcal{H}$ which are linear, continuous, self-adjoint and positive semidefinite. For $M \in S_+(\mathcal{H})$, we define the seminorm $\|\cdot\|_M : \mathcal{H} \to [0, +\infty)$, $\|x\|_M = \sqrt{\langle x, Mx\rangle}$. We consider the Loewner partial ordering on $S_+(\mathcal{H})$, defined for $M_1, M_2 \in \mathcal{S}_+(\mathcal{H})$ by

$$M_1 \succcurlyeq M_2 \Leftrightarrow \|x\|_{M_1} \geq \|x\|_{M_2}\ \forall x \in \mathcal{H}.$$

Furthermore, we define for $\alpha > 0$ the set $\mathcal{P}_\alpha(\mathcal{H}) := \{M \in \mathcal{S}_+(\mathcal{H}) : M \succcurlyeq \alpha \mathrm{Id}\}$, where $\mathrm{Id} : \mathcal{H} \to \mathcal{H}$, $\mathrm{Id}(x) = x$ for all $x \in \mathcal{H}$, denotes the identity operator on $\mathcal{H}$.

Let $A : \mathcal{H} \to \mathcal{G}$ be a linear continuous operator. The operator $A^* : \mathcal{G} \to \mathcal{H}$, fulfilling $\langle A^* y, x\rangle = \langle y, Ax\rangle$ for all $x \in \mathcal{H}$ and $y \in \mathcal{G}$, denotes the adjoint operator of A, while $\|A\| := \sup\{\|Ax\| : \|x\| \leq 1\}$ denotes the norm of A.

3 The Proximal Alternating Minimization Algorithm

The two-block separable optimization problem we are going to investigate in this paper has the following formulation.

Problem 3.1 Let $\mathcal{H}$, $\mathcal{G}$ and $\mathcal{K}$ be real Hilbert spaces, $f \in \Gamma(H)$ γ-strongly convex with $\gamma > 0$, $g \in \Gamma(G)$, $h_1 : \mathcal{H} \to \mathbb{R}$ a convex and Fréchet differentiable function with L_1-Lipschitz continuous gradient with $L_1 \geq 0$, $h_2 : \mathcal{G} \to \mathbb{R}$ a convex and Fréchet differentiable functions with L_2-Lipschitz continuous gradient with $L_2 \geq 0$, $A : \mathcal{H} \to \mathcal{K}$ and $B : \mathcal{G} \to \mathcal{K}$ linear continuous operators such that $A \neq 0$ and $b \in \mathcal{K}$. Consider the following optimization problem with two-block separable objective function and linear constraints

$$\min_{x\in\mathcal{H}, z\in\mathcal{G}} f(x) + h_1(x) + g(z) + h_2(z) \quad \text{s.t.} \quad Ax + Bz = b. \tag{5}$$

We allow the Lipschitz constant of the gradients of the functions h_1 and h_2 to be zero. In this case, the functions are affine.

The Lagrangian associated with the optimization problem (5) is defined by $L : \mathcal{H} \times \mathcal{G} \times \mathcal{K} \to \overline{\mathbb{R}}$,

$$L(x, z, p) = f(x) + h_1(x) + g(z) + h_2(z) + \langle p, b - Ax - Bz\rangle.$$

We say that $(x^*, z^*, p^*) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ is a saddle point of the Lagrangian L, if

$$(x^*, z^*, p) \leq L(x^*, z^*, p^*) \leq L(x, z, p^*) \quad \forall (x, z, p) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}.$$

It is well known that (x^*, z^*, p^*) is a saddle point of the Lagrangian L if and only if (x^*, z^*) is an optimal solution of (5), p^* is an optimal solution of its Fenchel dual problem

$$\sup_{\lambda\in\mathcal{K}}\{-(f^*\square h_1^*)(A^*\lambda) - (g^*\square h_2^*)(B^*\lambda) + \langle\lambda, b\rangle\}, \tag{6}$$

and the optimal objective values of (5) and (6) coincide. The existence of saddle points for L is guaranteed when (5) has an optimal solution and, for instance, the Attouch–Brézis-type condition

$$b \in \mathrm{sqri}(A(\mathrm{dom}\, f) + B(\mathrm{dom}\, g)) \tag{7}$$

holds (see [12, Theorem 3.4]). In the finite-dimensional setting, this asks for the existence of $x \in \mathrm{ri}(\mathrm{dom}\, f)$ and $z \in \mathrm{ri}(\mathrm{dom}\, g)$ satisfying $Ax + Bz = b$ and coincides with the assumption used by Tseng [1].

The system of optimality conditions for the primal-dual pair of optimization problems (5)–(6) reads:

$$A^*p^* - \nabla h_1(x^*) \in \partial f(x^*),\ B^*p^* - \nabla h_2(z^*) \in \partial g(z^*) \text{ and } Ax^* + Bz^* = b. \quad (8)$$

This means that if (5) has an optimal solution (x^*, z^*) and a qualification condition, like for instance (7), is fulfilled, then there exists an optimal solution p^* of (6) such that (8) holds; consequently, (x^*, z^*, p^*) is a saddle point of the Lagrangian L. Conversely, if (x^*, z^*, p^*) is a saddle point of the Lagrangian L, thus, (x^*, z^*, p^*) satisfies relation (8), then (x^*, z^*) is an optimal solution of (5) and p^* is an optimal solution of (6).

Remark 3.1 If (x_1^*, z_1^*, p_1^*) and (x_2^*, z_2^*, p_2^*) are two saddle points of the Lagrangian L, then $x_1^* = x_2^*$. This follows easily from (8), by using the strong monotonicity of ∂f and the monotonicity of ∂g.

In the following, we formulate the Proximal Alternating Minimization Algorithm to solve (5). To this end, we modify Tseng's AMA by evaluating in each of the two subproblems the functions h_1 and h_2 via gradient steps, respectively, and by introducing proximal terms defined through two sequences of positively semidefinite operators $(M_1^k)_{k\geq 0}$ and $(M_2^k)_{k\geq 0}$.

Algorithm 3.1 *(Proximal AMA)* Let $(M_1^k)_{k\geq 0} \subseteq \mathcal{S}_+(\mathcal{H})$ and $(M_2^k)_{k\geq 0} \subseteq \mathcal{S}_+(\mathcal{G})$. Choose $(x^0, z^0, p^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{K}$ and a sequence of stepsizes $(c_k)_{k\geq 0} \subseteq (0, +\infty)$. For all $k \geq 0$, set:

$$x^{k+1} = \mathrm{argmin}_{x\in\mathcal{H}} \left\{ f(x) - \langle p^k, Ax\rangle + \langle x - x^k, \nabla h_1(x^k)\rangle + \frac{1}{2}\|x - x^k\|^2_{M_1^k} \right\} \quad (9)$$

$$z^{k+1} \in \mathrm{argmin}_{z\in\mathcal{G}} \left\{ g(z) - \langle p^k, Bz\rangle + \frac{c_k}{2}\|Ax^{k+1} + Bz - b\|^2 + \langle z - z^k, \nabla h_2(z^k)\rangle + \frac{1}{2}\|z - z^k\|^2_{M_2^k} \right\} \quad (10)$$

$$p^{k+1} = p^k + c_k(b - Ax^{k+1} - Bz^{k+1}). \quad (11)$$

Remark 3.2 The sequence $(z^k)_{k\geq 0}$ is uniquely determined if there exists $\alpha_k > 0$ such that $c_k B^*B + M_2^k \in \mathcal{P}_{\alpha_k}(\mathcal{G})$ for all $k \geq 0$. This actually ensures that the objective function in subproblem (10) is strongly convex.

Remark 3.3 Let $k \geq 0$ be fixed and $M_2^k := \frac{1}{\sigma_k}\mathrm{Id} - c_k B^*B$, where $\sigma_k > 0$ and $\sigma_k c_k \|B\|^2 \leq 1$. Then M_2^k is positively semidefinite, and the update of z^{k+1} in the Proximal AMA method becomes a proximal step. This idea has been used in the

past with the same purpose for different algorithms involving proximal steps; see, for instance, [7–9,13–16]. Indeed, (10) holds if and only if

$$0 \in \partial g(z^{k+1}) + (c_k B^*B + M_2^k)z^{k+1} + c_k B^*(Ax^{k+1} - b) - M_2^k z^k + \nabla h_2(z^k) - B^*p^k$$

or, equivalently,

$$0 \in \partial g(z^{k+1}) + \frac{1}{\sigma_k} z^{k+1} - \left(\frac{1}{\sigma_k}\,\mathrm{Id} - c_k B^*B\right) z^k + \nabla h_2(z^k) + c_k B^*(Ax^{k+1} - b) - B^*p^k.$$

But this is nothing else than

$$\begin{aligned} z^{k+1} &= \mathrm{argmin}_{z\in\mathcal{G}} \left\{ g(z) + \frac{1}{2\sigma_k} \left\| z - \left(z^k - \sigma_k \nabla h_2(z^k) \right.\right.\right. \\ &\qquad \left.\left.\left. + \sigma_k c_k B^*(b - Ax^{k+1} - Bz^k) + \sigma_k B^* p^k \right) \right\|^2 \right\} \\ &= \mathrm{Prox}_{\sigma_k g}\left(z^k - \sigma_k \nabla h_2(z^k) + \sigma_k c_k B^*(b - Ax^{k+1} - Bz^k) + \sigma_k B^* p^k \right). \end{aligned}$$

The convergence of the Proximal AMA method is addressed in the next theorem.

Theorem 3.1 *In the setting of Problem 3.1, let the set of the saddle points of the Lagrangian L be nonempty. We assume that $M_1^k - \frac{L_1}{2}\,\mathrm{Id} \in \mathcal{S}_+(\mathcal{H})$, $M_1^k \succcurlyeq M_1^{k+1}$, $M_2^k - \frac{L_2}{2}\,\mathrm{Id} \in \mathcal{S}_+(\mathcal{G})$, $M_2^k \succcurlyeq M_2^{k+1}$ for all $k \geq 0$ and that $(c_k)_{k\geq 0}$ is a monotonically decreasing sequence satisfying*

$$\epsilon \leq c_k \leq \frac{2\gamma}{\|A\|^2} - \epsilon \quad \forall k \geq 0, \tag{12}$$

where $0 < \epsilon < \frac{\gamma}{\|A\|^2}$. If one of the following assumptions:

(i) *there exists $\alpha > 0$ such that $M_2^k - \frac{L_2}{2}\,\mathrm{Id} \in \mathcal{P}_\alpha(\mathcal{G})$ for all $k \geq 0$;*
(ii) *there exists $\beta > 0$ such that $B^*B \in \mathcal{P}_\beta(\mathcal{G})$;*

holds true, then the sequence $(x^k, z^k, p^k)_{k\geq 0}$ generated by Algorithm 3.1 converges weakly to a saddle point of the Lagrangian L.

Proof Let (x^*, z^*, p^*) be a fixed saddle point of the Lagrangian L. This means that it fulfils the system of optimality conditions

$$A^*p^* - \nabla h_1(x^*) \in \partial f(x^*) \tag{13}$$
$$B^*p^* - \nabla h_2(z^*) \in \partial g(z^*) \tag{14}$$
$$Ax^* + Bz^* = b \tag{15}$$

We start by proving that

$$\sum_{k\geq 0}\|x^{k+1}-x^*\|^2<+\infty,\quad \sum_{k\geq 0}\|Bz^{k+1}-Bz^*\|^2<+\infty,\quad \sum_{k\geq 0}\|z^{k+1}-z^k\|^2_{M_2^k-\frac{L_2}{2}\mathrm{Id}}<+\infty$$

and that the sequences $(z^k)_{k\geq 0}$ and $(p^k)_{k\geq 0}$ are bounded.

Assume that $L_1>0$ and $L_2>0$. Let $k\geq 0$ be fixed. Writing the optimality conditions for subproblems (9) and (10), we obtain

$$A^*p^k-\nabla h_1(x^k)+M_1^k(x^k-x^{k+1})\in\partial f(x^{k+1}) \tag{16}$$

and

$$B^*p^k-\nabla h_2(z^k)+c_kB^*(-Ax^{k+1}-Bz^{k+1}+b)+M_2^k(z^k-z^{k+1})\in\partial g(z^{k+1}), \tag{17}$$

respectively. Combining (13)–(17) with the strong monotonicity of ∂f and the monotonicity of ∂g, it yields

$$\begin{aligned}&\langle A^*(p^k-p^*)-\nabla h_1(x^k)+\nabla h_1(x^*)\\&\quad+M_1^k(x^k-x^{k+1}),x^{k+1}-x^*\rangle\geq\gamma\|x^{k+1}-x^*\|^2\end{aligned}$$

and

$$\begin{aligned}&\langle B^*(p^k-p^*)-\nabla h_2(z^k)+\nabla h_2(z^*)+c_kB^*(-Ax^{k+1}-Bz^{k+1}+b)\\&\quad+M_2^k(z^k-z^{k+1}),z^{k+1}-z^*\rangle\geq 0,\end{aligned}$$

which after summation lead to

$$\begin{aligned}&\langle p^k-p^*,Ax^{k+1}-Ax^*\rangle+\langle p^k-p^*,Bz^{k+1}-Bz^*\rangle\\&\quad+\langle c_k(-Ax^{k+1}-Bz^{k+1}+b),Bz^{k+1}-Bz^*\rangle\\&\quad-\langle\nabla h_1(x^k)-\nabla h_1(x^*),x^{k+1}-x^*\rangle-\langle\nabla h_2(z^k)-\nabla h_2(z^*),z^{k+1}-z^*\rangle\\&\quad+\langle M_1^k(x^k-x^{k+1}),x^{k+1}-x^*\rangle+\langle M_2^k(z^k-z^{k+1}),z^{k+1}-z^*\rangle\\&\quad\geq\gamma\|x^{k+1}-x^*\|^2.\end{aligned} \tag{18}$$

According to the Baillon–Haddad theorem (see [11, Corollary 18.16]), the gradients of h_1 and h_2 are $\frac{1}{L_1}$ and $\frac{1}{L_2}$-cocoercive, respectively; thus,

$$\begin{aligned}\langle\nabla h_1(x^*)-\nabla h_1(x^k),x^*-x^k\rangle&\geq\frac{1}{L_1}\|\nabla h_1(x^*)-\nabla h_1(x^k)\|^2\\\langle\nabla h_2(z^*)-\nabla h_2(z^k),z^*-z^k\rangle&\geq\frac{1}{L_2}\|\nabla h_2(z^*)-\nabla h_2(z^k)\|^2.\end{aligned}$$

On the other hand, by taking into account (11) and (15), it holds

$$\begin{aligned}&\langle p^k - p^*, Ax^{k+1} - Ax^*\rangle + \langle p^k - p^*, Bz^{k+1} - Bz^*\rangle\\ &\quad = \langle p^k - p^*, Ax^{k+1} + Bz^{k+1} - b\rangle = \frac{1}{c_k}\langle p^k - p^*, p^k - p^{k+1}\rangle\end{aligned}$$

By employing the last three relations in (18), it yields

$$\begin{aligned}&\frac{1}{c_k}\langle p^k - p^*, p^k - p^{k+1}\rangle + c_k\langle -Ax^{k+1} - Bz^{k+1} + b, Bz^{k+1} - Bz^*\rangle\\ &\quad + \langle M_1^k(x^k - x^{k+1}), x^{k+1} - x^*\rangle + \langle M_2^k(z^k - z^{k+1}), z^{k+1} - z^*\rangle\\ &\quad + \langle \nabla h_1(x^*) - \nabla h_1(x^k), x^{k+1} - x^*\rangle + \langle \nabla h_1(x^*) - \nabla h_1(x^k), x^* - x^k\rangle\\ &\quad - \frac{1}{L_1}\|\nabla h_1(x^*) - \nabla h_1(x^k)\|^2 + \langle \nabla h_2(z^*) - \nabla h_2(z^k), z^{k+1} - z^*\rangle\\ &\quad + \langle \nabla h_2(z^*) - \nabla h_2(z^k), z^* - z^k\rangle - \frac{1}{L_2}\|\nabla h_2(z^*) - \nabla h_2(z^k)\|^2\\ &\geq \gamma\|x^{k+1} - x^*\|^2,\end{aligned}$$

which, after expressing the inner products by means of norms, becomes

$$\begin{aligned}&\frac{1}{2c_k}\left(\|p^k - p^*\|^2 + \|p^k - p^{k+1}\|^2 - \|p^{k+1} - p^*\|^2\right)\\ &\quad + \frac{c_k}{2}\left(\|Ax^* - Ax^{k+1}\|^2 - \|b - Ax^{k+1} - Bz^{k+1}\|^2 - \|Ax^* + Bz^{k+1} - b\|^2\right)\\ &\quad + \frac{1}{2}\left(\|x^k - x^*\|^2_{M_1^k} - \|x^k - x^{k+1}\|^2_{M_1^k} - \|x^{k+1} - x^*\|^2_{M_1^k}\right)\\ &\quad + \frac{1}{2}\left(\|z^k - z^*\|^2_{M_2^k} - \|z^k - z^{k+1}\|^2_{M_2^k} - \|z^{k+1} - z^*\|^2_{M_2^k}\right)\\ &\quad + \langle \nabla h_1(x^*) - \nabla h_1(x^k), x^{k+1} - x^k\rangle - \frac{1}{L_1}\|\nabla h_1(x^*) - \nabla h_1(x^k)\|^2\\ &\quad + \langle \nabla h_2(z^*) - \nabla h_2(z^k), z^{k+1} - z^k\rangle - \frac{1}{L_2}\|\nabla h_2(z^*) - \nabla h_2(z^k)\|^2\\ &\geq \gamma\|x^{k+1} - x^*\|^2.\end{aligned}$$

Using again (11), inequality $\|Ax^* - Ax^{k+1}\|^2 \leq \|A\|^2\|x^* - x^{k+1}\|^2$ and the following expressions

$$\begin{aligned}&\langle \nabla h_1(x^*) - \nabla h_1(x^k), x^{k+1} - x^k\rangle - \frac{1}{L_1}\|\nabla h_1(x^*) - \nabla h_1(x^k)\|^2\\ &\quad = -L_1\left\|\frac{1}{L_1}(\nabla h_1(x^*) - \nabla h_1(x^k)) + \frac{1}{2}(x^k - x^{k+1})\right\|^2 + \frac{L_1}{4}\|x^k - x^{k+1}\|^2,\end{aligned}$$

and

$$\langle \nabla h_2(x^*) - \nabla h_2(z^k), z^{k+1} - z^k \rangle - \frac{1}{L_2}\|\nabla h_2(z^*) - \nabla h_2(z^k)\|^2$$
$$= -L_2 \left\| \frac{1}{L_2}(\nabla h_2(z^*) - \nabla h_2(z^k)) + \frac{1}{2}(z^k - z^{k+1}) \right\|^2 + \frac{L_2}{4}\|z^k - z^{k+1}\|^2,$$

it yields

$$\begin{aligned}
&\frac{1}{2}\|x^{k+1} - x^*\|^2_{M_1^k} + \frac{1}{2c_k}\|p^{k+1} - p^*\|^2 + \frac{1}{2}\|z^{k+1} - z^*\|^2_{M_2^k} \\
&\leq \frac{1}{2}\|x^k - x^*\|^2_{M_1^k} + \frac{1}{2c_k}\|p^k - p^*\|^2 + \frac{1}{2}\|z^k - z^*\|^2_{M_2^k} - \frac{c_k}{2}\|Ax^* + Bz^{k+1} - b\|^2 \\
&\quad - \frac{1}{2}\|z^k - z^{k+1}\|^2_{M_2^k} - \left(\gamma - \frac{c_k}{2}\|A\|^2\right)\|x^{k+1} - x^*\|^2 - \frac{1}{2}\|x^k - x^{k+1}\|^2_{M_1^k} \\
&\quad - L_1 \left\| \frac{1}{L_1}(\nabla h_1(x^*) - \nabla h_1(x^k)) + \frac{1}{2}(x^k - x^{k+1}) \right\|^2 + \frac{L_1}{4}\|x^k - x^{k+1}\|^2 \\
&\quad - L_2 \left\| \frac{1}{L_2}(\nabla h_2(z^*) - \nabla h_2(z^k)) + \frac{1}{2}(z^k - z^{k+1}) \right\|^2 + \frac{L_2}{4}\|z^k - z^{k+1}\|^2.
\end{aligned}$$

Finally, by using the monotonicity of $(M_1^k)_{k\geq 0}$, $(M_2^k)_{k\geq 0}$ and $(c_k)_{k\geq 0}$, we obtain

$$\begin{aligned}
&c_{k+1}\|x^{k+1} - x^*\|^2_{M_1^{k+1}} + \|p^{k+1} - p^*\|^2 + c_{k+1}\|z^{k+1} - z^*\|^2_{M_2^{k+1}} \\
&\quad \leq c_k\|x^k - x^*\|^2_{M_1^k} + \|p^k - p^*\|^2 + c_k\|z^k - z^*\|^2_{M_2^k} - R_k,
\end{aligned} \tag{19}$$

where

$$\begin{aligned}
R_k := \; & c_k\left(2\gamma - c_k\|A\|^2\right)\|x^{k+1} - x^*\|^2 + c_k^2\|Bz^{k+1} - Bz^*\|^2 \\
& + c_k\|z^k - z^{k+1}\|^2_{M_2^k - \frac{L_2}{2}\mathrm{Id}} + c_k\|x^k - x^{k+1}\|^2_{M_1^k - \frac{L_1}{2}\mathrm{Id}} \\
& + 2c_kL_1 \left\| \frac{1}{L_1}(\nabla h_1(x^*) - \nabla h_1(x^k)) + \frac{1}{2}(x^k - x^{k+1}) \right\|^2 \\
& + 2c_kL_2 \left\| \frac{1}{L_2}(\nabla h_2(z^*) - \nabla h_2(z^k)) + \frac{1}{2}(z^k - z^{k+1}) \right\|^2.
\end{aligned}$$

If $L_1 = 0$ (and, consequently, ∇h_1 is constant) and $L_2 > 0$, then, by using the same arguments, we obtain again (19), but with

$$\begin{aligned}
R_k := \; & c_k\left(2\gamma - c_k\|A\|^2\right)\|x^{k+1} - x^*\|^2 + c_k^2\|Bz^{k+1} - Bz^*\|^2 \\
& + c_k\|z^k - z^{k+1}\|^2_{M_2^k - \frac{L_2}{2}\mathrm{Id}} + c_k\|x^k - x^{k+1}\|^2_{M_1^k} \\
& + 2c_kL_2 \left\| \frac{1}{L_2}(\nabla h_2(z^*) - \nabla h_2(z^k)) + \frac{1}{2}(z^k - z^{k+1}) \right\|^2.
\end{aligned}$$

If $L_2 = 0$ (and, consequently, ∇h_2 is constant) and $L_2 > 0$, then, by using the same arguments, we obtain again (19), but with

$$\begin{aligned} R_k := \; & c_k \left(2\gamma - c_k\|A\|^2\right)\|x^{k+1} - x^*\|^2 + c_k^2\|Bz^{k+1} - Bz^*\|^2 \\ & + \; c_k\|z^k - z^{k+1}\|^2_{M_2^k} + c_k\|x^k - x^{k+1}\|^2_{M_1^k - \frac{L_1}{2}\mathrm{Id}} \\ & + \; 2c_kL_1\left\|\frac{1}{L_1}(\nabla h_1(x^*) - \nabla h_1(x^k)) + \frac{1}{2}(x^k - x^{k+1})\right\|^2. \end{aligned}$$

Relation (19) follows even if $L_1 = L_2 = 0$, but with

$$\begin{aligned} R_k := \; & c_k \left(2\gamma - c_k\|A\|^2\right)\|x^{k+1} - x^*\|^2 + c_k^2\|Bz^{k+1} - Bz^*\|^2 \\ & + \; c_k\|z^k - z^{k+1}\|^2_{M_2^k} + c_k\|x^k - x^{k+1}\|^2_{M_1^k}. \end{aligned}$$

Notice that, due to $M_1^k - \frac{L_1}{2}\mathrm{Id} \in \mathcal{S}_+(\mathcal{H})$ and $M_2^k - \frac{L_2}{2}\mathrm{Id} \in \mathcal{S}_+(\mathcal{G})$, all summands in R_k are nonnegative.

Let be $N \geq 0$ fixed. By summing the inequality in (19) for $k = 0, \ldots, N$ and using telescoping arguments, we obtain

$$\begin{aligned} & c_{N+1}\|x^{N+1} - x^*\|^2_{M_1^{N+1}} + \|p^{N+1} - p^*\|^2 + c_N\|z^{N+1} - z^*\|^2_{M_2^{N+1}} \\ & \leq c_0\|x^0 - x^*\|^2_{M_1^0} + \|p^0 - p^*\|^2 + c_0\|z^0 - z^*\|_{M_2^0} - \sum_{k=0}^{N} R_k. \end{aligned}$$

On the other hand, from (19) we also obtain that

$$\exists \lim_{k\to\infty}\left(c_k\|x^k - x^*\|^2_{M_1^k} + \|p^k - p^*\|^2 + c_k\|z^k - z^*\|^2_{M_2^k}\right), \tag{20}$$

thus $(p^k)_{k\geq 0}$ is bounded, and $\sum_{k\geq 0} R_k < +\infty$.

Taking (12) into account, we have $c_k(2\gamma - c_k\|A\|^2) \geq \varepsilon^2\|A\|^2$ for all $k \geq 0$. Therefore,

$$\sum_{k\geq 0}\|x^{k+1} - x^*\|^2 < +\infty, \quad \sum_{k\geq 0}\|Bz^{k+1} - Bz^*\|^2 < +\infty \tag{21}$$

and

$$\sum_{k\geq 0}\|z^{k+1} - z^k\|^2_{M_2^k - \frac{L_2}{2}\mathrm{Id}} < +\infty. \tag{22}$$

From here, we obtain

$$x^k \to x^*, \quad Bz^k \to Bz^* \; (k \to +\infty), \tag{23}$$

which, by using (11) and (15), lead to

$$p^k - p^{k+1} \to 0 \ (k \to +\infty). \tag{24}$$

Taking into account the monotonicity properties of $(c_k)_{k\geq 0}$ and $(M_1^k)_{k\geq 0}$, a direct implication of (20) and (23) is

$$\exists \lim_{k\to\infty} \left(\|p^k - p^*\|^2 + c_k \|z^k - z^*\|^2_{M_2^k} \right). \tag{25}$$

Suppose that assumption (i) holds true, namely that there exists $\alpha > 0$ such that $M_2^k - \frac{L_2}{2}\,\mathrm{Id} \in \mathcal{P}_\alpha(\mathcal{G})$ for all $k \geq 0$. From (25), it follows that $(z^k)_{k\geq 0}$ is bounded, while (22) ensures that

$$z^{k+1} - z^k \to 0 \ (k \to +\infty). \tag{26}$$

In the following, let us prove that each weak sequential cluster point of $(x^k, z^k, p^k)_{k\geq 0}$ (notice that the sequence is bounded) is a saddle point of L. Let be $(\bar{z}, \bar{p}) \in \mathcal{G} \times \mathcal{K}$ such that the subsequence $(x^{k_j}, z^{k_j}, p^{k_j})_{j\geq 0}$ converges weakly to $(x^*, \bar{z}, \bar{p})$ as $j \to +\infty$. From (16), we have

$$A^* p^{k_j} - \nabla h_1(x^{k_j}) + M_1^{k_j}(x^{k_j} - x^{k_j+1}) \in \partial f(x^{k_j+1}) \quad \forall j \geq 1.$$

Due to the fact that x^{k_j} converges strongly to x^* and p^{k_j} converges weakly to a $\bar{p}$ as $j \to +\infty$, using the continuity of ∇h_1 and the fact that the graph of the convex subdifferential of f is sequentially closed in the strong-weak topology (see [11, Proposition 20.33]), it follows

$$A^* \bar{p} - \nabla h_1(x^*) \in \partial f(x^*).$$

From (17), we have for all $j \geq 0$

$$\begin{aligned} &B^* p^{k_j} - \nabla h_2(z^{k_j}) + c_{k_j} B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) \\ &+ M_2^{k_j}(z^{k_j} - z^{k_j+1}) \in \partial g(z^{k_j+1}), \end{aligned}$$

which is equivalent to

$$\begin{aligned} &B^* p^{k_j} + \nabla h_2(z^{k_j+1}) - \nabla h_2(z^{k_j}) + c_{k_j} B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) \\ &+ M_2^{k_j}(z^{k_j} - z^{k_j+1}) \in \partial(g + h_2)(z^{k_j+1}) \end{aligned}$$

and further to

$$\begin{aligned} z^{k_j+1} \in \partial(g + h_2)^* \Big(&B^* p^{k_j} + \nabla h_2(z^{k_j+1}) - \nabla h_2(z^{k_j}) \\ &+ c_{k_j} B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) + M_2^{k_j}(z^{k_j} - z^{k_j+1}) \Big). \end{aligned} \tag{27}$$

By denoting for all $j \geq 0$

$$\begin{aligned} v^j &:= z^{k_j+1}, u^j := p^{k_j}, \\ w^j &:= \nabla h_2(z^{k_j+1}) - \nabla h_2(z^{k_j}) \\ &\quad + c_{k_j} B^*(-Ax^{k_j+1} - Bz^{k_j+1} + b) + M_2^{k_j}(z^{k_j} - z^{k_j+1}), \end{aligned}$$

(27) reads

$$v^j \in \partial(g + h_2)^*(B^*u^j + w^j) \quad \forall j \geq 0.$$

According to (26), we have $v^j \rightharpoonup \bar{z}$, $u^j \rightharpoonup \bar{p}$ as $j \to +\infty$; thus, by taking into account (23), $Bv^j \to B\bar{z} = Bz^*$ as $j \to +\infty$. Combining (29) with the Lipschitz continuity of ∇h_2, (24), (26) and (11), one can easily see that $w^j \to 0$ as $j \to +\infty$. Due to the monotonicity of the subdifferential, we have that for all (u, v) in the graph of $\partial(g + h_2)^*$ and for all $j \geq 0$

$$\langle Bv^j - Bv, u^j \rangle + \langle v^j - v, w^j - u \rangle \geq 0.$$

We let j converge to $+\infty$ and receive

$$\langle \bar{z} - v, B^*\bar{p} - u \rangle \geq 0 \ \ \forall (u, v) \text{ in the graph of } \partial(g + h_2)^*.$$

The maximal monotonicity of the convex subdifferential of $(g + h_2)^*$ ensures that $\bar{z} \in \partial(g + h_2)^*(B^*\bar{p})$, which is the same as $B^*\bar{p} \in \partial(g + h_2)(\bar{z})$. In other words, $B^*\bar{p} - \nabla h_2(\bar{z}) \in \partial g(\bar{z})$. Finally, by combining (11) and (24), the equality $Ax^* + B\bar{z} = b$ follows. In conclusion, $(x^*, \bar{z}, \bar{p})$ is a saddle point of the Lagrangian L.

In the following, we show that sequence $(x^k, z^k, p^k)_{k\geq 0}$ converges weakly. To this end, we consider two sequential cluster points (x^*, z_1, p_1) and (x^*, z_2, p_2). Consequently, there exists $(k_s)_{s\geq 0}$, $k_s \to +\infty$ as $s \to +\infty$, such that the subsequence $(x^{k_s}, z^{k_s}, p^{k_s})_{s\geq 0}$ converges weakly to (x^*, z_1, p_1) as $s \to +\infty$. Furthermore, there exists $(k_t)_{t\geq 0}, k_t \to +\infty$ as $t \to +\infty$, such that that a subsequence $(x^{k_t}, z^{k_t}, p^{k_t})_{t\geq 0}$ converges weakly to (x^*, z_2, p_2) as $t \to +\infty$. As seen before, (x^*, z_1, p_1) and (x^*, z_2, p_2) are both saddle points of the Lagrangian L.

From (25), which is fulfilled for every saddle point of the Lagrangian L, we obtain

$$\exists \lim_{k\to+\infty} (\|p^k - p_1\|^2 - \|p^k - p_2\|^2 + c_k\|z^k - z_1\|^2_{M_2^k} - c_k\|z^k - z_2\|^2_{M_2^k}) \coloneqq T. \quad (28)$$

For all $k \geq 0$, we have

$$\begin{aligned} &\|p^k - p_1\|^2 - \|p^k - p_2\|^2 + c_k\|z^k - z_1\|^2_{M_2^k} - c_k\|z^k - z_2\|^2_{M_2^k} \\ &= \|p_2 - p_1\|^2 + 2\langle p_k - p_2, p_2 - p_1 \rangle + c_k\|z_2 - z_1\|^2_{M_2^k} \\ &\quad + 2c_k\langle z_k - z_2, z_2 - z_1 \rangle_{M_2^k}. \end{aligned}$$

Since $M_2^k \geq \left(\alpha + \frac{L_2}{2}\right) \mathrm{Id}$ for all $k \geq 0$ and $(M_2^k)_{k\geq 0}$ is a nonincreasing sequence of symmetric operators in the sense of the Loewner partial ordering, there exists a symmetric operator $M \geq \left(\alpha + \frac{L_2}{2}\right) \mathrm{Id}$ such that $(M_2^k)_{k\geq 0}$ converges pointwise to M in the strong topology as $k \to +\infty$ (see [17, Lemma 2.3]). Furthermore, let $c := \lim_{k\to+\infty} c_k > 0$. Taking the limits in (28) along the subsequences $(k_s)_{s\geq 0}$ and $(k_t)_{t\geq 0}$, it yields

$$T = -\|p_2 - p_1\|^2 - c\|z_2 - z_1\|_M^2 = \|p_2 - p_1\|^2 + c\|z_2 - z_1\|_M^2,$$

thus

$$\|p_2 - p_1\|^2 + c\|z_2 - z_1\|_M^2 = 0.$$

It follows that $p_1 = p_2$ and $z_1 = z_2$; thus, $(x^k, z^k, p^k)_{k\geq 0}$ converges weakly to a saddle point of the Lagrangian L.

Assume now that condition (ii) holds, namely that there exists $\beta > 0$ such that $B^*B \in \mathcal{P}_\beta(\mathcal{H})$. Then $\beta\|z_1 - z_2\|^2 \leq \|Bz_1 - Bz_2\|^2$ for all $z_1, z_2 \in \mathcal{G}$, which means that, if (x_1^*, z_1^*, p_1^*) and (x_2^*, z_2^*, p_2^*) are two saddle points of the Lagrangian L, then $x_1^* = x_2^*$ and $z_1^* = z_2^*$.

For the saddle point (x^*, z^*, p^*) of the Lagrangian L, we fixed at the beginning of the proof and the generated sequence $(x^k, z^k, p^k)_{k\geq 0}$ we receive because of (23) that

$$x^k \to x^*, \quad z^k \to z^*, \quad p^k - p^{k+1} \to 0 \ (k \to +\infty). \tag{29}$$

Moreover,

$$\exists \lim_{k\to\infty} \|p^k - p^*\|^2.$$

The remainder of the proof follows in analogy to the one given under assumption (i). □

If $h_1 = 0$ and $h_2 = 0$, and $M_1^k = 0$ and $M_2^k = 0$ for all $k \geq 0$, then the Proximal AMA method becomes the AMA method as it has been proposed by Tseng [1]. According to Theorem 3.1 (for $L_1 = L_2 = 0$), the generated sequence converges weakly to a saddle point of the Lagrangian, if there exists $\beta > 0$ such that $B^*B \in \mathcal{P}_\beta(\mathcal{G})$. In finite-dimensional spaces, this condition reduces to the assumption that B is injective.

4 Numerical Experiments

In this section, we compare the numerical performances of AMA and Proximal AMA on two applications in image processing and machine learning. The numerical experiments were performed on a computer with an Intel Core i5-3470 CPU and 8 GB DDR3 RAM.

4.1 Image Denoising and Deblurring

We addressed an image denoising and deblurring problem formulated as a nonsmooth convex optimization problem (see [18–20])

$$\inf_{x\in\mathbb{R}^n}\left\{\frac{1}{2}\|Ax-b\|^2+\lambda \mathrm{TV}(x)\right\}, \tag{30}$$

where $A\in\mathbb{R}^{n\times n}$ represents a blur operator, $b\in\mathbb{R}^n$ is a given blurred and noisy image, $\lambda>0$ is a regularization parameter and $\mathrm{TV}:\mathbb{R}^n\to\mathbb{R}$ is a discrete total variation functional. The vector $x\in\mathbb{R}^n$ is the vectorized image $X\in\mathbb{R}^{M\times N}$, where $n=MN$ and $x_{i,j}:=X_{i,j}$ stand for the normalized value of the pixel in the i-th row and the j-th column, for $1\le i\le M$, $1\le j\le N$.

Two choices have been considered for the discrete total variation, namely the isotropic total variation $\mathrm{TV}_{\mathrm{iso}}:\mathbb{R}^n\to\mathbb{R}$,

$$\begin{aligned}\mathrm{TV}_{\mathrm{iso}}(x)=&\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}\sqrt{(x_{i+1,j}-x_{i,j})^2+(x_{i,j+1}-x_{i,j})^2}\\&+\sum_{i=1}^{M-1}|x_{i+1,N}-x_{i,j}|+\sum_{j=1}^{N-1}|x_{M,j+1}-x_{M,j}|,\end{aligned}$$

and the anisotropic total variation $\mathrm{TV}_{\mathrm{aniso}}:\mathbb{R}^n\to\mathbb{R}$,

$$\begin{aligned}\mathrm{TV}_{\mathrm{aniso}}(x)=&\sum_{i=1}^{M-1}\sum_{j=1}^{N-1}|x_{i+1,j}-x_{i,j}|+|x_{i,j+1}-x_{i,j}|\\&+\sum_{i=1}^{M-1}|x_{i+1,N}-x_{i,j}|+\sum_{j=1}^{N-1}|x_{M,j+1}-x_{M,j}|.\end{aligned}$$

Consider the linear operator $L:\mathbb{R}^n\to\mathbb{R}^n\times\mathbb{R}^n$, $x_{i,j}\mapsto\big(L_1x_{i,j},L_2x_{i,j}\big)$, where

$$L_1x_{i,j}=\begin{cases}x_{i+1,j}-x_{i,j}, & \text{if } i<M\\ 0, & \text{if } i=M\end{cases}\quad\text{and}\quad L_2x_{i,j}=\begin{cases}x_{i,j+1}-x_{i,j}, & \text{if } j<N\\ 0, & \text{if } j=N\end{cases}$$

One can easily see that $\|L\|^2\le 8$. The optimization problem (30) can be written as

$$\inf_{x\in\mathbb{R}^n}\{f(Ax)+g(Lx)\}, \tag{31}$$

where $f:\mathbb{R}^n\to\mathbb{R}$, $f(x)=\frac{1}{2}\|x-b\|^2$, and $g:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ is defined by $g(y,z)=\lambda\|(y,z)\|_1$ for the anisotropic total variation, and by $g(y,z)=\lambda\|(y,z)\|_\times:=\lambda\sum_{i=1}^M\sum_{j=1}^N\sqrt{y_{i,j}^2+z_{i,j}^2}$ for the isotropic total variation.

We solved the Fenchel dual problem of (31) by AMA and Proximal AMA and determined in this way an optimal solution of the primal problem, too. The reason for this strategy was that the Fenchel dual problem of (31) is a convex optimization problem with two-block separable linear constraints and objective function.

Indeed, the Fenchel dual problem of (31) reads (see [11,12])

$$\inf_{p\in\mathbb{R}^n, q\in\mathbb{R}^n\times\mathbb{R}^n} \left\{f^*(p)+g^*(q)\right\},\ \text{s.t. } A^*p+L^*q=0. \tag{32}$$

Since f and g have full domains, strong duality for (31)–(32) holds.

As $f^*(p)=\frac{1}{2}\|p\|^2+\langle p,b\rangle$ for all $p\in\mathbb{R}^n$, f^* is 1-strongly convex. We chose $M_1^k=0$ and $M_2^k=\frac{1}{\sigma_k}\mathrm{I}-c_kL^*L$ (see Remark 3.3) and obtained for Proximal AMA the iterative scheme which reads for every $k\geq 0$:

$$\begin{aligned}
p^{k+1} &= Ax^k-b\\
q^{k+1} &= \mathrm{Prox}_{\sigma_k g^*}\left(q^k+\sigma_k c_k L(-A^*p^{k+1}-L^*q^k)+\sigma_k L(x^k)\right)\\
x^{k+1} &= x^k+c_k(-A^*p^{k+1}-L^*q^{k+1}).
\end{aligned}$$

In the case of the anisotropic total variation, the conjugate of g is the indicator function of the set $[-\lambda,\lambda]^n\times[-\lambda,\lambda]^n$; thus, $\mathrm{Prox}_{\sigma_k g^*}$ is the projection operator $\mathcal{P}_{[-\lambda,\lambda]^n\times[-\lambda,\lambda]^n}$ on the set $[-\lambda,\lambda]^n\times[-\lambda,\lambda]^n$. The iterative scheme reads for all $k\geq 0$:

$$\begin{aligned}
p^{k+1} &= Ax^k-b\\
(q_1^{k+1},q_2^{k+1}) &= \mathcal{P}_{[-\lambda,\lambda]^n\times[-\lambda,\lambda]^n}\Big((q_1^k,q_2^k)\\
&\qquad +c_k\sigma_k(-LA^*p^{k+1}-LL^*(q_1^k,q_2^k))+\sigma_k Lx^k\Big)\\
x^{k+1} &= x^k+c_k\left(-A^*p^{k+1}-L^*(q_1^{k+1},q_2^{k+1})\right).
\end{aligned}$$

In the case of the isotropic total variation, the conjugate of g is the indicator function of the set $S:=\left\{(v,w)\in\mathbb{R}^n\times\mathbb{R}^n:\max_{1\leq i\leq n}\sqrt{v_i^2+w_i^2}\leq\lambda\right\}$; thus, $\mathrm{Prox}_{\sigma_k g^*}$ is the projection operator $P_S:\mathbb{R}^n\times\mathbb{R}^n\to S$ on S, defined as

$$(v_i,w_i)\mapsto\lambda\frac{(v_i,w_i)}{\max\left\{\lambda,\sqrt{v_i^2+w_i^2}\right\}},\quad i=1,\ldots,n.$$

The iterative scheme reads for all $k\geq 0$:

$$\begin{aligned}
p^{k+1} &= Ax^k-b\\
(q_1^{k+1},q_2^{k+1}) &= P_S\left((q_1^k,q_2^k)+c_k\sigma_k(-LA^*p^{k+1}-LL^*(q_1^k,q_2^k))+\sigma_k Lx^k\right)
\end{aligned}$$

$$x^{k+1} = x^k + c_k \left(-A^* p^{k+1} - L^*(q_1^{k+1}, q_2^{k+1})\right).$$

We compared the Proximal AMA method with Tseng's AMA method. While in Proximal AMA a closed formula is available for the computation of $(q_1^{k+1}, q_2^{k+1})_{k\geq 0}$, in AMA we solved the resulting optimization subproblem

$$(q_1^{k+1}, q_2^{k+1}) = \operatorname{argmin}_{q_1,q_2} \left\{ g^*(q_1, q_2) - \langle x^{k+1}, L^*(q_1, q_2)\rangle + \frac{1}{2} c_k \|A^* p^{k+1} + L^*(q_1, q_2)\|^2 \right\}$$

in every iteration $k \geq 0$ by making some steps of the FISTA method [2].

We used in our experiments a Gaussian blur of size 9×9 and standard deviation 4, which led to an operator A with $\|A\|^2 = 1$ and $A^* = A$. Furthermore, we added Gaussian white noise with standard deviation 10^{-3}. We used for both algorithms a constant sequence of stepsizes $c_k = 2 - 10^{-7}$ for all $k \geq 0$. One can notice that $(c_k)_{k\geq 0}$ fulfils (12). For Proximal AMA, we considered $\sigma_k = \frac{1}{8.00001 \cdot c_k}$ for all $k \geq 0$, which ensured that every matrix $M_2^k = \frac{1}{\sigma_k} \mathrm{I} - c_k L^* L$ is positively definite for all $k \geq 0$. This is actually the case, if $\sigma_k c_k \|L\|^2 < 1$ for all $k \geq 0$. In other words, assumption (i) in Theorem 3.1 was verified.

In Figs. 1, 2, 3 and 4, we show how Proximal AMA and AMA perform when reconstructing the blurred and noisy coloured MATLAB test image "office_ 4" of 600×903 pixels (see Fig. 5) for different choices for the regularization parameter λ and by considering both the anisotropic and isotropic total variation as regularization functionals. In all considered instances that Proximal AMA outperformed AMA from the point of view of both the convergence behaviour of the sequence of the function values and of the sequence of ISNR (Improvement in signal-to-noise ratio) values. An explanation could be that the number of iterations Proximal AMA makes in a certain amount of time is more than double the number of outer iterations performed by AMA.

4.2 Kernel-Based Machine Learning

In this subsection, we will describe the numerical experiments we carried out in the context of classifying images via support vector machines.

The given data set consisting of 5570 training images and 1850 test images of size 28×28 was taken from http://www.cs.nyu.edu/~roweis/data.html. The problem we considered was to determine a decision function based on a pool of handwritten digits showing either the number five or the number six, labelled by $+1$ and -1, respectively (see Fig. 6). To evaluate the quality of the decision function, we computed the percentage of misclassified images of the test data set.

In order to describe the approach we used, we denote by

$$\mathcal{Z} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \subseteq \mathbb{R}^d \times \{+1, -1\},$$

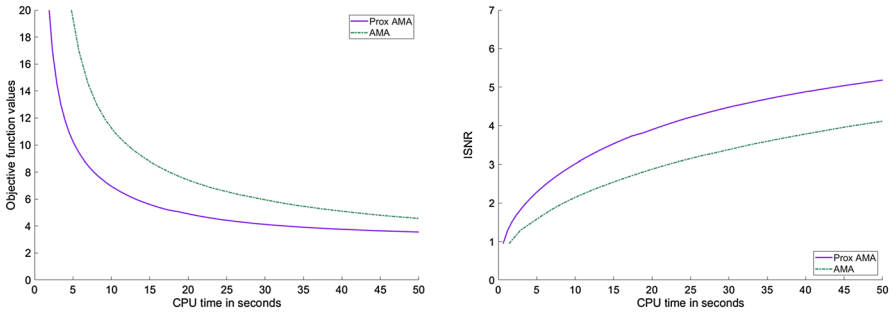

Fig. 1 Objective function values and the ISNR values for the anisotropic TV and $\lambda = 5 \cdot 10^{-5}$

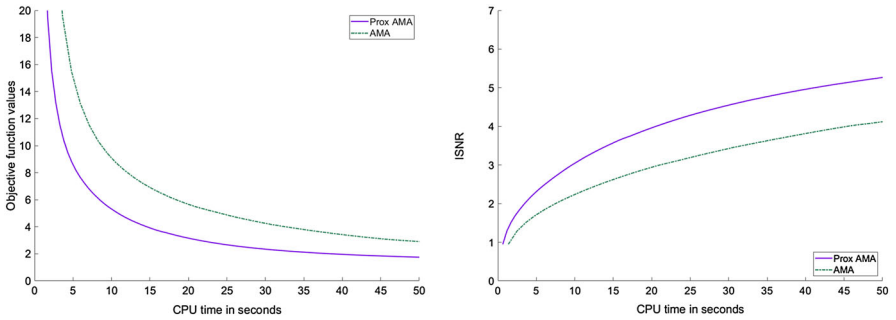

Fig. 2 Objective function values and the ISNR values for the anisotropic TV and $\lambda = 10^{-5}$

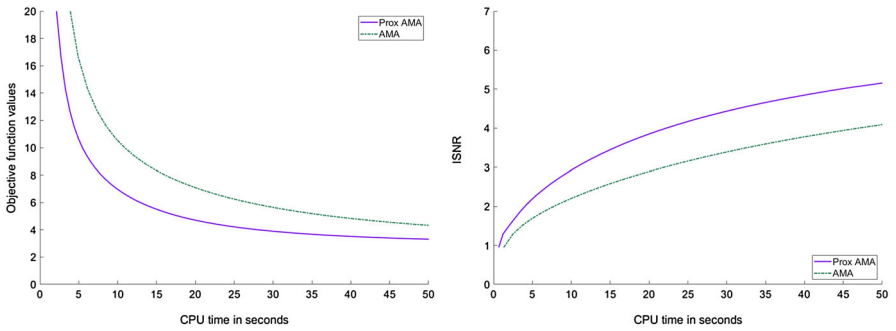

Fig. 3 Objective function values and the ISNR values for the isotropic TV and $\lambda = 5 \cdot 10^{-5}$

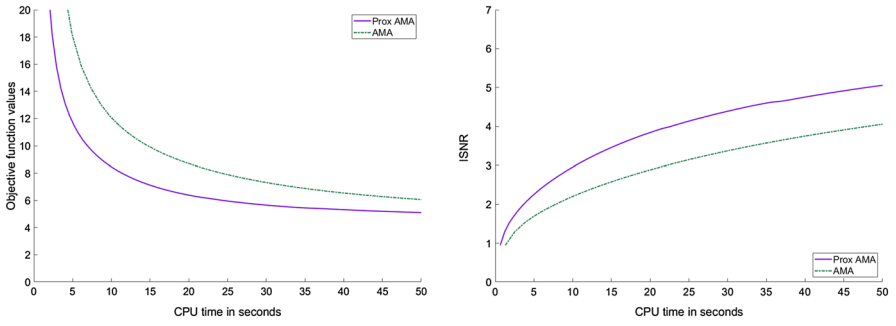

Fig. 4 Objective function values and the ISNR values for the isotropic TV and $\lambda = 10^{-4}$

Fig. 5 Original image, the blurred and noisy image and the reconstructed image after 50 s cpu time

Fig. 6 A sample of images belonging to the classes $+1$ and -1, respectively

the given training data set. The decision functional $\mathtt{f}$ was assumed to be an element of the Reproducing Kernel Hilbert Space (RHKS) $\mathcal{H}_\kappa$, induced by the symmetric and finitely positive definite Gaussian kernel function

$$\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R},\ \kappa(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$

By $K \in \mathbb{R}^{n\times n}$, we denoted the Gram matrix with respect to the training data set $\mathcal{Z}$, namely the symmetric and positive definite matrix with entries $K_{ij} = \kappa(X_i, X_j)$ for $i, j = 1, \ldots, n$. To penalize the deviation between the predicted value $\mathtt{f}(x)$ and the true value $y \in \{+1, -1\}$, we used the hinge loss functional $(x, y) \mapsto \max\{1 - xy, 0\}$.

According to the representer theorem, the decision function $\mathtt{f}$ can be expressed as a kernel expansion in terms of the training data; in other words, $\mathtt{f}(\cdot) = \sum_{i=1}^n x_i \kappa(\cdot, X_i)$, where $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is the optimal solution of the optimization problem

$$\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2} x^T K x + C \sum_{i=1}^n \max\{1 - (Kx)_i Y_i, 0\} \right\}. \tag{33}$$

Here, $C > 0$ denotes the regularization parameter controlling the trade-off between the loss function and the regularization term. Hence, in order to determine the decision function we solved the convex optimization problem (33), which can be written as

$$\min_{x\in\mathbb{R}^n} \{f(x) + g(Kx)\}$$

or, equivalently,

$$\min_{x\in\mathbb{R}^n, z\in\mathbb{R}^n} \{f(x) + g(z)\},\ \text{s.t. } Kx - z = 0$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \frac{1}{2}x^T Kx$, and $g : \mathbb{R}^n \to \mathbb{R}$ is defined by $g(z) = C\sum_{i=1}^n \max\{1 - z_i Y_i, 0\}$.

Since the Gram matrix K is positively definite, the function f is $\lambda_{\min}(K)$-strongly convex, where $\lambda_{\min}(K)$ denotes the minimal eigenvalue of K, and differentiable, and it holds $\nabla f(x) = Kx$ for all $x \in \mathbb{R}^n$. For an element of the form $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$, it holds

$$g^*(p) = \begin{cases} \sum_{i=1}^n p_i Y_i, & \text{if } p_i Y_i \in [-C, 0], \quad i = 1, \ldots, n, \\ +\infty, & \text{otherwise.} \end{cases}$$

Consequently, for every $\mu > 0$ and $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$, it holds

$$\text{Prox}_{\mu g^*}(x) = \left(\mathcal{P}_{Y_1[-C,0]}(p_1 - \sigma Y_1), \ldots, \mathcal{P}_{Y_n[-C,0]}(p_n - \sigma Y_n)\right),$$

where $\mathcal{P}_{Y_i[-C,0]}$ denotes the projection operator on the set $Y_i[-C, 0]$, $i = 1, \ldots, n$.

We implemented Proximal AMA for $M_2^k = 0$ for all $k \geq 0$ and different choices for the sequence $(M_1^k)_{k\geq 0}$. This resulted in an iterative scheme which reads for all $k \geq 0$:

$$x^{k+1} = \text{argmin}_{x\in\mathbb{R}^n}\left\{f(x) - \langle p^k, Kx\rangle + \frac{1}{2}\|x - x^k\|^2_{M_1^k}\right\} = (K + M_1^k)^{-1}(Kp^k + M_1^k x^k) \tag{34}$$

$$z^{k+1} = \text{Prox}_{\frac{1}{c_k}g}\left(Kx^{k+1} - \frac{1}{c^k}p^k\right) = \left(Kx^{k+1} - \frac{1}{c^k}p^k\right) - \frac{1}{c_k}\text{Prox}_{c_k g^*}\left(c_k Kx^{k+1} - p^k\right)$$

$$p^{k+1} = p^k + c_k(-Kx^{k+1} + z^{k+1}). \tag{35}$$

We would like to emphasize that the AMA method updates the sequence $(z^{k+1})_{k\geq 0}$ also via (35), while the sequence $(x^{k+1})_{k\geq 0}$, as $M_1^k = 0$, is updated via $x^{k+1} = p^k$ for all $k \geq 0$. However, it turned out that the Proximal AMA where $M_1^k = \tau_k K$, for $\tau_k > 0$ and all $k \geq 0$, performs better than the version with $M_1^k = 0$ for all $k \geq 0$, which actually corresponds to the AMA method. In this case, (34) becomes $x^{k+1} = \frac{1}{1+\tau_k}(p^k + \tau_k x^k)$ for all $k \geq 0$.

We used for both algorithms a constant sequence of stepsizes given by $c_k = 2 \cdot \frac{\lambda_{\min}(K)}{\|K\|^2} - 10^{-8}$ for all $k \geq 0$. Tables 1 and 2 show for $C = 1$ and different values of the kernel parameter σ that Proximal AMA outperforms AMA in what concerns the time and the number of iterates needed to achieve a certain value for a given fixed misclassification rate (which proved to be the best one among several obtained by varying C and σ) and for the RMSE (root-mean-square deviation) for the sequence of primal iterates.

Table 1 Performance evaluation of Proximal AMA (with $\tau_k = 10$ for all $k \geq 0$) and AMA for the classification problem with $C = 1$ and $\sigma = 0.2$

Algorithm	Misclassification rate at 0.7027%	RMSE $\leq 10^{-3}$
Proximal AMA	8.18 s (145)	23.44 s (416)
AMA	8.65 s (153)	26.64 s (474)

The entries refer to the CPU times in seconds and the number of iterations

Table 2 Performance evaluation of Proximal AMA (with $\tau_k = 102$ for all $k \geq 0$) and AMA for the classification problem with $C = 1$ and $\sigma = 0.25$

Algorithm	Misclassification rate at 0.7027%	RMSE $\leq 10^{-3}$
Proximal AMA	141.78 s (2448)	629.52 s (10,940)
AMA	147.99 s (2574)	652.61 s (11,368)

The entries refer to the CPU times in seconds and the number of iterations

5 Perspectives and Open Problems

In future, it might be interesting to:

(1) carry out investigations related to the convergence rates for both the iterates and objective function values of Proximal AMA; as emphasized in [10] for the Proximal ADMM algorithm, the use of variable metrics can have a determinant role in this context, as they may lead to dynamic stepsizes which are favourable to an improved convergence behaviour of the algorithm (see also [15,21]);

(2) consider a slight modification of Algorithm 3.1, by replacing (11) with

$$p^{k+1} = p^k + \theta c_k(b - Ax^{k+1} - Bz^{k+1}),$$

where $0 < \theta < \frac{\sqrt{5}+1}{2}$ and to investigate the convergence properties of the resulting scheme; it has been noticed in [22] that the numerical performances of the classical ADMM algorithm for convex optimization problems in the presence of a relaxation parameter with $1 < \theta < \frac{\sqrt{5}+1}{2}$ outperform the ones obtained when $\theta = 1$;

(3) embed the investigations made in this paper in the more general framework of monotone inclusion problems, as it was recently done in [10] starting from the Proximal ADMM algorithm.

6 Conclusions

The Proximal AMA method has the advantage over the classical AMA method that, as long as the sequence of variable metrics is chosen appropriately, it performs proximal steps when calculating new iterates. In this way, it avoids the use in every iteration of minimization subroutines. In addition, it handles properly smooth and convex functions which might appear in the objective. The sequences of generated iterates converge

to a primal–dual solution in the same setting as for the classical AMA method. The fact that instead of solving of minimization subproblems one has only to make proximal steps, may lead to better numerical performances, as we show in the experiments on image processing and support vector machines classification.

Acknowledgements Open access funding provided by Austrian Science Fund (FWF). The work of SB and GW has been partially supported by DFG (Deutsche Forschungsgemeinschaft), project WA922/9-1. The work of RIB has been partially supported by FWF (Austrian Science Fund), project I 2419-N32. The work of ERC has been supported by FWF, project P 29809-N32. The authors are thankful to two anonymous referees for helpful comments and remarks which improved the presentation of the manuscript.

References

1. Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim. **29**(1), 119–138 (1991)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
3. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and Its Applications, vol. 49, pp. 185–212. Springer, New York (2011)
4. Chouzenoux, E., Pesquet, J.-C., Repetti, A.: A block coordinate variable metric forward–backward algorithm. J. Global Optim. **66**(3), 457–485 (2016)
5. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximations. Comput. Math. Appl. **2**, 17–40 (1976)
6. Attouch, H., Soueycatt, M.: Augmented Lagrangian and proximal alternating direction methods of multipliers in Hilbert spaces. Applications to games, PDE's and control. Pac. J. Optim. **5**(1), 17–37 (2009)
7. Fazel, M., Pong, T.K., Sun, D., Tseng, P.: Hankel matrix rank minimization with applications in system identification and realization. SIAM J. Matrix Anal. Appl. **34**, 946–977 (2013)
8. Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**, 269–297 (2014)
9. Banert, S., Boţ, R.I., Csetnek, E.R.: Fixing and extending some recent results on the ADMM algorithm. Preprint arXiv:1612.05057 (2017)
10. Boţ, R.I., Csetnek, E.R.: ADMM for monotone operators: convergence analysis and rates. Adv. Comput. Math. https://doi.org/10.1007/s10444-018-9619-3 **(to appear)**
11. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics. Springer, New York (2011)
12. Boţ, R.I.: Conjugate Duality in Convex Optimization. Lecture Notes in Economics and Mathematical Systems, vol. 637. Springer, Berlin (2010)
13. Bredies, K., Sun, H.: A proximal point analysis of the preconditioned alternating direction method of multipliers. J. Optim. Theory Appl. **173**(3), 878–907 (2017)
14. Bredies, K., Sun, H.: Preconditioned Douglas–Rachford splitting methods for convex-concave saddle-point problems. SIAM J. Numer. Anal. **53**(1), 421–444 (2015)
15. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
16. Esser, E., Zhang, X., Chan, T.F.: A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imaging Sci. **3**(4), 1015–1046 (2010)
17. Combettes, P.L., Vũ, B.C.: Variable metric quasi-Fejér monotonicity. Nonlinear Anal. **78**, 17–31 (2013)

18. Boţ, R.I., Hendrich, C.: Convergence analysis for a primal-dual monotone + skew splitting algorithm with applications to total variation minimization. J. Math. Imaging Vis. **49**(3), 551–568 (2014)
19. Hendrich, C.: Proximal Splitting Methods in Non-smooth Convex Optimization. Ph.D. Thesis, Technical University of Technology, Chemnitz, Germany (2014)
20. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total-variation-based noise removal algorithms. Physica D Nonlinear Phenom. **60**(1–4), 259–268 (1992)
21. Boţ, R.I., Csetnek, E.R., Heinrich, A., Hendrich, C.: On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems. Math. Program. **150**(2), 251–279 (2015)
22. Fortin, M., Glowinski, R.: On decomposition-coordination methods using an augmented Lagrangian. In: Fortin, M., Glowinski, R. (eds.) Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems. North-Holland, Amsterdam (1983)