Alternating proximal gradient descent for nonconvex regularised problems with multiconvex coupling terms

Mila Nikolova, Pauline Tan

Alternating proximal gradient descent for nonconvex regularised problems with multiconvex coupling terms

Mila Nikolova*, and Pauline Tan+

* CMLA, CNRS, ENS Cachan, Université Paris-Saclay, nikolova@cmla.ens-cachan.fr
+ ONERA - The French Aerospace Lab, pauline.tan@cmap.polytechnique.fr

Abstract

There has been an increasing interest in constrained nonconvex regularized block biconvex / multiconvex optimization problems. We introduce an approach that effectively exploits the biconvex / multiconvex structure of the coupling term and enables complex application-dependent regularization terms to be used. The proposed Alternating Structure-Adapted Proximal gradient descent algorithm enjoys simple well defined updates. Global convergence of the algorithm to a critical point is proved using the so-called Kurdyka-Łojasiewicz property for subanalytic functions. Moreover, we prove that a large class of useful objective functions obeying our assumptions are subanalytic and thus satisfy the Kurdyka-Łojasiewicz property.

Keywords Alternating minimization, Block coordinate descent, Global convergence, Kurdyka-Łojasiewicz property, Nonconvex-nonsmooth optimization, Proximal forward-backward, Proximal gradient descent, Subanalytic functions

1 Introduction

Recently, there has been an increasing interest in the design and the analysis of regularized block biconvex (multiconvex) optimization problems. In the first part of this work we consider problems of the form:

$$\text{minimize}_{x,y}\ J(x,y) = F(x) + G(y) + H(x,y), \tag{1}$$

where x and y belong to real finite-dimensional real spaces. Such an objective is also known as a two block optimization model with blocks x and y. The coupling term H is block biconvex: for any y fixed, $x \mapsto H(x,y)$ is convex and for any x fixed, $y \mapsto H(x,y)$ is convex. The feasible set of J is block biconvex. The regularization functions F and G can be nonconvex and are supposed continuously differentiable inside the domain of J (for example, smooth approximations of nonsmooth functions). It worths noting that J is generally nonconvex even if F and G are convex.

Then we consider the extension of (1) to N blocks living in finite-dimensional real spaces $\{U_i\}_{i=1}^N$, in which case $J : U_1 \times \cdots \times U_N \to \mathbb{R} \cup \{+\infty\}$ reads as

$$\text{minimize}_x\ J\left(x_{(1)}, \ldots, x_{(N)}\right) := \sum_{i=1}^{N} F_i\left(x_{(i)}\right) + H(x). \tag{2}$$

The coupling function H is block multiconvex, i.e., $x_{(i)} \mapsto H(x)$ is convex for any i, and the possibly nonconvex regularizers $F_i : U_i \to \mathbb{R} \cup \{+\infty\}$ are continuously differentiable on the domain of J.

Optimization problems of the form (1) - (2) are widely used in engineering, science and finance. They are rich enough to cover various practical applications, such as blind source separation, blind deconvolution, nonnegative matrix factorization, structured total least squares, multi-modal learning for image classification [40], patch-based methods for inpainting [3], to list a few.

In order to simplify the reading of the paper, most of the presentation is on the two block problem (1); the results on the multi block problem (2) are given later on.

1.1 General alternating minimization schemes

The most intuitive way to solve problems of the form given in (1) is to use alternating minimization which which generates a sequence of iterates $\{x^k, y^k\}_{k\in\mathbb{N}}$ defined by

$$\begin{array}{rcll} & & \text{Choose algorithm ALG to solve the } x\text{- and } y\text{-updates;} & \\ & & \text{for each } k = 1, 2, \ldots, \text{compute} & \\ x^k & \in & \arg\min_x J(x, y^{k-1}) & \text{using ALG,} \\ y^k & \in & \arg\min_y J(x^k, y) & \text{using ALG.} \end{array} \tag{3}$$

This classic scheme is known as block coordinate Gauss-Seidel method and as block coordinate descent (BCD). It was introduced in [27] (1957) and further developed for various problems, see e.g., [36, 24, 3, 11, 2]. If J is continuously differentiable and if the minimum in each step is uniquely attained, convergence to a critical point holds [12, Prop. 2.7.1]. A general convergence result on descent methods for *real-analytic* (possibly nonconvex) objectives was obtained by Absil, Mahony, and Andrews in [1].

A way to relax the requirements for convergence of the BCD in (3) is to consider the proximal BCD scheme:

$$\begin{array}{rcll} & & \text{Choose algorithm ALG to solve the } x\text{- and } y\text{-updates;} & \\ & & \text{Select step-sizes } \tau > 0 \text{ and } \sigma > 0; & \\ & & \text{for each } k = 1, 2, \ldots, \text{compute} & \\ x^k & \in & \arg\min_x \left\{ J(x, y^{k-1}) + \frac{1}{2\tau}\|x - x^{k-1}\|^2 \right\} & \text{using ALG,} \\ y^k & \in & \arg\min_y \left\{ J(x^k, y) + \frac{1}{2\sigma}\|y - y^{k-1}\|^2 \right\} & \text{using ALG.} \end{array} \tag{4}$$

This approach was introduced for convex functions J by Auslender [7, sec. 4]. An extension to J strictly quasi-convex with respect to its blocks was proposed in [25, sec. 7] and to nonsmooth objectives in [36]. Convergence facts on (4) for other nonconvex nonsmooth objectives were found in [38].

Note that the BCD and the proximal BCD schemes in (3) and in (4), respectively, employ a minimization algorithm ALG which heavily determines the numerical issue.

1.2 Review of related literature

An efficient approach to solve the difficulties arising with the proximal BCD for nonconvex and nonsmooth objectives J was proposed in 2013 by Xu and Yin [38] for block multiconvex differentiable H and by Bolte, Sabach, and Teboulle [17] for two block continuously differentiable H. Using the smoothness of H, the idea was to apply a proximal linearized BCD for (x^k, y^k):

$$\begin{array}{rcl} & & \text{for each } k = 1, 2, \ldots, \text{compute} \\ \text{find } \tau_k > 0 & & \text{according to Lip}\left(\nabla_x H\left(\cdot, y^{k-1}\right)\right) \\ x^k & \in & \arg\min_x \left\{ \left\langle x, \nabla_x H\left(x^{k-1}, y^{k-1}\right)\right\rangle + F(x) + \frac{1}{2\tau_k}\|x - x^{k-1}\|^2 \right\}, \\ \text{find } \sigma_k > 0 & & \text{according to Lip}\left(\nabla_y H\left(x^k, \cdot\right)\right) \\ y^k & \in & \arg\min_y \left\{ \left\langle y, \nabla_y H\left(x^k, y^{k-1}\right)\right\rangle + G(y) + \frac{1}{2\sigma_k}\|y - y^{k-1}\|^2 \right\}, \end{array} \tag{5}$$

where the Lipschitz constant is computed (or estimated) at each step. The scheme needs regularizers (F, G) that are "simple" in the sense that their proximity operator

$$\arg\min_x \left\{ F(x) + \frac{1}{2\tau}\|x - z\|^2 \right\} \tag{6}$$

has a closed-form expression. Since then, this approach became very popular. It was further successfully used and improved in numerous works, see e.g., [26, 37, 9, 32, 21, 39]. We recall that a proximal linearized BCD is equivalent to an alternating proximal gradient descent and to an alternating proximal forward-backward.

The advantages of this approach compared to the schemes in subsection 1.1 are tremendous. All the three schemes – the BCD (3), the proximal BCD (4) and the proximal linearized BDC (5) – were analyzed and compared in [38]. The conclusion [38, p. 1795] is that in general the three schemes give different solutions, and that the proximal linearized BDC (5) needs less computation and offers a better decrease of the objective

function than the other algorithms. One however remarks the cost to compute or to estimate the step-sizes (τ_k, σ_k) at each iteration and the possible variability of $\{\tau_k\}_{k\geq 0}$ and $\{\sigma_k\}_{k\geq 0}$ during the iterations. An attempt to fix this issue is the preconditioning proposed by Chouzenoux, Pesquet, and Repetti [21]. One can also remark that if F or G is truly nonconvex, the solution in (6) can be composed out of non-connected sets, for an example see [17, p. 487].

Remark 1. [Choices for H in applications] In nearly all applications solved using a scheme of the form (5), the coupling term H is biquadratic (multiquadratic), sometimes combined with a bilinear term. When restricted to two blocks, one finds $H(x,y) = \|L_0(x\cdot y) - w\|^2 + \langle L_1(x), L_2(x,y)\rangle$ where "$\cdot$" denotes a product, w is a known matrix, $\|\cdot\|$ stands for Frobenius norm and L_i are linear forms. Such a term is used in [9] and with $L_0 = 1$ and $L_1 = L_2 = 0$ in [38, 17, 32, 39]. In all these applications, $(\nabla_x H, \nabla_y H)$ are only locally Lipschitz.

Remark 2. [Choices for (F, G) in applications] The most typical form of F (resp. G) is the non-negativity constraint $x \geq 0$ (resp., $y \geq 0$) – see [38, 26, 32, 39] and [17, sec. 4.2, p. 486]. We note that in those cases F and G are smooth inside their domains. Other forms for F or G are the ℓ_2 norm (possibly squared) [32, 9], the counting function ℓ_0 [17] and ℓ_p, $0 \leq p < 1$ [39].

A unified approach for proving the convergence of proximal splitting methods for nonconvex and nonsmooth problems was developed by Attouch, Bolte, and Svaiter in their seminal work [6]. A central assumption in order to prove global convergence of the iterates to a critical point is that the objective function J satisfies the so-called Kurdyka-Łojasiewicz (KŁ) property [15, 16]. In several articles [26, 9, 32, 9] convergence is proven using the methodology proposed in [17].

1.3 Motivation and proposed algorithm ASAP

The biconvexity (block multiconvexity) of H is a very strong structural property of the objective function. The proximal linearized gradient approach in (5) (subsection 1.2) does not benefit from this feature. Our motivation is to build an algorithm that exploits the biconvexity (block multiconvexity) of H. This means reversing the splitting used in (5), i.e., using the proximity operators with respect to $H(\cdot, y)$ and $H(x, \cdot)$ – "simple" in practice, see Remark 1 – instead of those of F and G. The argument behind the splitting in (5) is to use F and G nonsmooth. Alternatively, we attach to H the convex constraints on F and G and assume that F and G are differentiable on their domains.

Remark 3. Smooth approximations to solve nonsmooth optimization problems are recommended in a unified framework by Beck and Teboulle [10]. Very numerous methods with nonsmooth regularizers use their smooth approximations in the numerical algorithms [18, 28, 20, 32]. Smooth (stiff) functions are customarily used for sparse recovery [19, 29, 20]. A study by Chen of smoothing methods giving rise to efficient calculations in numerical schemes for nonsmooth optimization can be found in [18]. In other cases the estimates of x, y must not be sparse and hence *the regularizers must be smooth*; a practical application on fringe separation is presented in subsection 7.2.

For simplicity, here we present the case of biconvex coupling terms H. We reformulate the problem so that H contains the biconvex constraints and let F and G be continuously differentiable. Thus we propose the simple Alternating Structure-Adapted Proximal gradient descent (ASAP) algorithm[1] sketched below:

$$
\begin{array}{lll}
\text{ASAP} & \text{Set } \tau \in (0, 2/\mathrm{Lip}(\nabla F)) \text{ and } \sigma \in (0, 2/\mathrm{Lip}(\nabla G)); & \\
& \text{for each } k = 1, 2, \ldots, \text{compute:} & \\
x^k & = \arg\min_x \left\{ \left\langle x, \nabla F\left(x^{k-1}\right)\right\rangle + H(x, y^{k-1}) + \frac{1}{2\tau}\|x - x^{k-1}\|^2 \right\}; & \\
y^k & = \arg\min_y \left\{ \left\langle y, \nabla G\left(y^{k-1}\right)\right\rangle + H(x^k, y) + \frac{1}{2\sigma}\|y - y^{k-1}\|^2 \right\}. &
\end{array}
\tag{7}
$$

Let us provide some details about the proposed algorithm.

[1] A particular form of this algorithm designed for a specific application was proposed by Soncco, Barbanson, Nikolova et al. [35] without convergence proof.

– The step-sizes τ and σ in the proposed ASAP algorithm depend only on F and G, so they are fixed in advance.

– The facts that H is biconvex and that (7) does not involve directly F and G show that x^k and y^k are given by the unique minimizer of a strictly convex coercive function. Thus the iterates (x^k, y^k) are uniquely defined (even if F or G are nonconvex). Furthermore, Remark 1 shows that in many important applications (x^k, y^k) are the explicit solutions of well-posed quadratic problems. In general, data-fidelity terms are defined using functions $H(x, \cdot)$ and $H(\cdot, y)$ that are "simple" in the sense of (6).

– Regularizers F and G can have complex structure in order to capture various application-dependent features; one example is presented in subsection 7.2.

The bi / multi convex structure of the coupling term H and its particular use in the ASAP algorithm requires a specific proof to obtain subsequential convergence of the iterates. Our assumptions are very simple and easy to verify. Besides, a series of additional results help understanding the algorithm. We exhibit important classes of optimization problems that can be solved using the proposed ASAP and prove that they are *subanalytic*. They hence enjoy the KŁ property. The global convergence of the iterates produced by ASAP is established using a result from [6].

The ASAP algorithm addresses in a simple way a wide range of problems. Two particular applications of the algorithm to big-data air-born sequences of images are actually used by our industrial partner ONERA (the French aerospace lab).

1.4 Outline

Section 2 presents the optimization model. Section 3 is devoted to the prerequisites for understanding the ASAP algorithm and its convergence. The ASAP algorithm is outlined in section 4. Section 5 establishes the convergence of ASAP for biconvex coupling terms; a generic family of objective functions is proven to satisfy the KŁ property in subsection 5.3. Section 6 is on block multiconvex coupling terms. Applications for Hadamard based coupling terms are presented in section 7.

2 The problem and examples

Notations We consider that $x \in U$ and $y \in V$ where U and V are real finite-dimensional spaces. The ith element of a vector or a matrix x (seen as a vector) reads as x_i. A vector (a matrix) indexed for some purpose is denoted by $x_{(i)}$. For an $m \times n$ real matrix w we systematically denote

$$\|w\| := \|w\|_F = \sqrt{\sum_{i,j} w_{i,j}^2}\ ,$$

noticing that if w is a vector ($n = 1$), the Frobenius norm $\|\cdot\|_F$ boils down to the ℓ_2 norm. For subsets of real finite-dimensional spaces, capital italic letters are used. Given a nonempty set $\mathcal{S} \subset U$, the distance of any point $x^+ \in U$ to $\mathcal{S}$ is defined by

$$\operatorname{dist}(x^+, \mathcal{S}) := \inf\{\|x - x^+\| \ \mid \ x \in \mathcal{S}\},$$

while the indicator function of $\mathcal{S}$ is given by

$$\chi_{\mathcal{S}}(x) := \begin{cases} 0 & \text{if} \quad x \in \mathcal{S}, \\ +\infty & \text{if} \quad x \notin \mathcal{S}. \end{cases}$$

The subdifferential of a convex function h is denoted by ∂h.

2.1 The optimization model for biconvex coupling term

We are interested in solving nonconvex minimization problems of the form $J : U \times V \to \mathbb{R}$

$$J(x, y) := F(x) + G(y) + H(x, y), \tag{8}$$

where U and V are *real finite-dimensional spaces*.[2] According to what was said in the introduction, we adopt the following blanket *model assumption* on the objective J:

[2]For instance, in section 7 an application with $U = V = \mathbb{R}^{m \times n}$ is presented.

Assumption (M)

(a) $J : U \times V \to \mathbb{R} \cup \{+\infty\}$ is lower bounded;
(b) $F : U \to \mathbb{R}$ and $G : V \to \mathbb{R}$ are continuously differentiable and their gradients ∇F and ∇G are Lipschitz continuous with constants $\mathrm{L}_{\nabla F}$ and $\mathrm{L}_{\nabla G}$, respectively;
(c) $H : U \times V \to \mathbb{R} \cup \{+\infty\}$ is biconvex[3] and differentiable on its domain.

The cases $F = 0$ or $G = 0$ are considered as well. From Assumption (M), the objective J is lower-semicontinuous. Further, H can be cast into the form

$$H(x, y) = \widetilde{H}(x, y) + \chi_{\mathcal{D}_x}(x) + \chi_{\mathcal{D}_y}(y),$$

where $\widetilde{H}$ is differentiable on $U \times V$ and $\mathcal{D}_x \subseteq U$ and $\mathcal{D}_y \subseteq V$ are nonempty closed convex sets. The optimization problem equivalently reads as

$$J(x, y) = \widetilde{H}(x, y) + F(x) + \chi_{\mathcal{D}_x}(x) + G(y) + \chi_{\mathcal{D}_y}(y). \tag{9}$$

Typically, the biconvex structure of the coupling term H is due to a bilinear mapping (see Remark 1 for examples). For clarity, we recall that

Definition 1. *A mapping $b : U \times V \to W$ is bilinear if $x \mapsto b(x, y)$ and $y \mapsto b(x, y)$ are linear applications. If $W = \mathbb{R}$, then b is called a bilinear form.*

2.2 Illustration: a general family of objective functions

A generic family of objective functions that can be minimized using the proposed algorithm is described next for the case of H biconvex:

$$J(x, y) := \underbrace{\sum_i f_i(\|A_i x\|)}_{=:F(x)} + \underbrace{\sum_j g_j(\|B_j y\|)}_{=:G(y)} + \underbrace{h(\|b(x, y) - w\|) + \chi_{\mathcal{D}_x}(x) + \chi_{\mathcal{D}_y}(y)}_{=:H(x,y)} \tag{10}$$

with $b : U \times V \to W$ a bilinear mapping, $(\mathcal{D}_x, \mathcal{D}_y)$ closed nonempty convex sets, A_i and B_i linear operators, and $w \in W$ given data. Function h is convex and differentiable; the most often, $h(t) = t^2$. Some relevant choices for functions (f_i, g_j) in accordance with Assumption (M)(b) are listed below.

Example 1. [Choices for (f_i, g_j) in (10)] These functions are of the form $\psi : \mathbb{R} \to \mathbb{R}$ or $(\psi)^p : \mathbb{R} \to \mathbb{R}$ where p is a rational number and include a parameter $\alpha > 0$:

(i) $\psi(t) := |t|^2$ and $(\psi(t))^p$ for $p > \frac{1}{2}$;
(ii) $\psi(t) := \begin{cases} |t| - \alpha/2 & \text{if} \quad |t| > \alpha \\ t^2/(2\,\alpha) & \text{if} \quad |t| \le \alpha \end{cases}$, $\alpha > 0$ and $(\psi(t))^p$ for $p \in (0, 1]$;
(iii) $\psi(t) := \sqrt{t^2 + \alpha}$, $\alpha > 0$ and $(\psi(t))^p$ for $p \in (0, 1]$;
(iv) $\psi(t) := |t| - \alpha \log(1 + |t|/\alpha)$, $\alpha > 0$;
(v) $\psi(t) := \log(1 + t^2/\alpha)$, $\alpha > 0$;
(vi) $\psi(t) := t^2/(\alpha + t^2)$, $\alpha > 0$;
(vii) $\psi(t) = 1 - \exp\left(-t^2/\alpha\right)$.

When α is small, functions (ii)-(vi) are stiff near the origin and they provide smooth approximations of nonsmooth functions. In particular, (v), (vi) and (vii) can approximate the counting function ℓ_0. Functions ψ in (iii)-(iv) are convex and for α small enough they are used to approximate the ℓ_1 norm [19, 10] whereas ψ^p for $p \in (0, 1]$ are used to approximate the corresponding ℓ_p "norm" for "sparse recovery" [29, 28, 20].

[3] for fixed $y \in V$, $x \mapsto H(x, y)$ is convex, and for fixed $x \in U$, $y \mapsto H(x, y)$ is convex.

3 Preliminary facts

3.1 Elements of subdifferential calculus

Here we recall some facts on subdifferential calculus in relation with the objective J. Given a function $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, its domain is

$$\operatorname{dom} f := \{x \in \mathbb{R}^m \ \mid \ f(x) < +\infty\},$$

and f is proper if and only if $\operatorname{dom} f \neq \varnothing$.

Definition 2. *[Subgradients of convex functions] Let $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper convex lower semicontinuousnn function and $x^+ \in \operatorname{dom} h$. The subdifferential $\partial h(x^+)$ of h at x^+ is the set of all $p \in \mathbb{R}^m$, called subgradients of h at x^+, such that*

$$\forall\, x \in \mathbb{R}^m \quad h(x) \geq h(x^+) + \langle p, x - x^+ \rangle$$

If $x^+ \notin \operatorname{dom} h$, then $\partial h(x^+) = \varnothing$.

The subdifferential for nonconvex nonsmooth functions is defined below.

Definition 3. *[33, Def. 8.3] Let $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a function.*

(a) *The* Fréchet subdifferential *of h at $x^+ \in \operatorname{dom} h$, denoted $\widehat{\partial} h(x^+)$, is the set of vectors $p \in \mathbb{R}^m$ such that $\forall\, x \in \mathbb{R}^m$ one has $h(x) \geq h(x^+) + \langle p, x - x^+ \rangle + o(\|x - x^+\|)$. If $x^+ \notin \operatorname{dom} h$, then $\widehat{\partial} h(x^+) = \varnothing$.*

(b) *The* (limiting-)subdifferential *of h at $x^+ \in \operatorname{dom} h$, written $\partial h(x^+)$, is defined by*

$$\partial h(x^+) := \left\{ p \in \mathbb{R}^m \ \mid \ \exists\, x^k \to x^+,\ h(x^k) \to h(x^+),\ p^k \to p,\ p^k \in \widehat{\partial} h(x^k) \right\}.$$

Then $\widehat{\partial} h(x) \subset \partial h(x)$ and both subsets are closed [33, Theorem 8.6]. If h is convex, then $\widehat{\partial} h(x) = \partial h(x)$ as in Definition 2. If h is differentiable, $\partial h(x^+) = \{\nabla h(x^+)\}$.

The next remark simplifies several parts in the analysis of our algorithm.

Remark 4. From Assumption (M), one has

(a) $\operatorname{dom} J := \{(x, y) \in U \times V \ \mid \ J(x, y) < +\infty\} = \mathcal{D}_x \times \mathcal{D}_y$ is biconvex and closed;

(b) J is continuous at any point $(x, y) \in \operatorname{int}(\operatorname{dom} J)$ and obeys $J(x, y) = \widetilde{J}(x, y)$ where the continuous function $\widetilde{J}$ is given by $\widetilde{J}(x, y) := F(x) + G(y) + \widetilde{H}(x, y)$.

We can note that (b) holds true for the assumptions in [4, 38, 39].

For y fixed, the partial subdifferential of $J(\cdot, y)$ at x is denoted by $\partial_x J(x, y)$; for x fixed, $\partial_y J(x, y)$ is defined in a similar way.

Proposition 1. *Let J obey Assumption (M). Then, for any $(x, y) \in \operatorname{dom} J$,*

$$\partial J(x, y) = \partial_x J(x, y) \times \partial_y J(x, y) = (\nabla F(x) + \partial_x H(x, y)) \times (\nabla G(y) + \partial_y H(x, y))$$

where the symbol "$\times$" stands for Cartesian product.

Proof. From (9), the function $\widetilde{J}(x, y) := F(x) + G(y) + \widetilde{H}(x, y)$ is differentiable. This shows that for any $(x, y) \in \operatorname{dom} J$ one has

$$\partial J(x, y) = \nabla \widetilde{J}(x, y) + \partial \left(\chi_{\mathcal{D}_x}(x) + \chi_{\mathcal{D}_y}(y) \right).$$

Using subdifferential calculus for separable functions [33, Prop. 10.6] yields

$$\partial \left(\chi_{\mathcal{D}_x}(x) + \chi_{\mathcal{D}_y}(y) \right) = \partial \chi_{\mathcal{D}_x}(x) \times \partial \chi_{\mathcal{D}_y}(y).$$

Since $\nabla \widetilde{J}(x, y) = (\nabla F(x) + \nabla_x \widetilde{H}(x, y), \nabla G(y) + \nabla_y \widetilde{H}(x, y))$, and using the last equality together with Remark 4(b), leads to the stated result. □

The Fermat's rule, extended to nonconvex/nonsmooth functions is given next.

Proposition 2. *[33, Theorem 10.1] Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper function. If f has a local minimum at x^*, then $0 \in \partial f(x^*)$.*

For nonconvex functions, the Fermat's rule is only a necessary conditions. First-order methods can find only points satisfying this rule.

Definition 4. *We say that (x^*, y^*) is a critical point of J if $(0,0) \in \partial J(x^*, y^*)$.*

The set of the critical points of J will be denoted by $\mathrm{crit}(J)$.

3.2 Two facts on limit point sets

Any bounded sequence $\{z^k\}_{k\in\mathbb{N}}$ has a convergent subsequence; thus the set of all its limit points

$$\mathcal{L}(z^0) := \left\{ z^* \;\middle|\; \exists\, \{k_j\}_{j\in\mathbb{N}} \text{ strictly increasing such that } z^{k_j} \to z^* \text{ as } j \to \infty \right\}. \tag{11}$$

is nonempty and closed. The next two claims are proven in Appendix.

Lemma 1. *Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence of $\mathbb{R}^m$ such that $\|z^{k+1} - z^k\| \to 0$ as k goes to infinity. Suppose that the set $\mathcal{L}(z^0)$ of the limit points of $\{z^k\}_{k\in\mathbb{N}}$ is nonempty and bounded. Then $\lim_{k\to\infty} \mathrm{dist}\left(z^k, \mathcal{L}(z^0)\right) = 0$.*

An immediate useful consequence is the following:

Corollary 1. *Let $\{z^k\}_{k\in\mathbb{N}}$ be a sequence of $\mathcal{R}^m$ such that $\|z^{k+1} - z^k\| \to 0$ as k goes to infinity. The following statements are equivalent:*

(a) *the set $\mathcal{L}(z^0)$ of the limit points of $\{z^k\}_{k\in\mathbb{N}}$ is nonempty and bounded;*
(b) *the sequence $\{z^k\}_{k\in\mathbb{N}}$ is bounded.*

3.3 Elements of proximal gradient descent

Proximity operators were inaugurated by Moreau in 1962 [31] as a generalization of convex projection operators. They were studied in numerous works, see the monograph of Bauschke and Combettes [8] for an overview.

Definition 5. *Let $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous and convex function. Given $z \in \mathbb{R}^m$ and $\tau > 0$. The proximity operator of h at $z \in \mathbb{R}^m$ is defined as [31, 8]*

$$\mathrm{prox}_{\tau h}(z) = \arg\min_{x\in\mathbb{R}^m} \left\{ h(x) + \frac{1}{2\tau}\|z - x\|^2 \right\}. \tag{12}$$

Remark 5. Two important consequences of this definition are given below:

(a) the minimizer $\mathrm{prox}_{\tau h}(z)$ always exists and is uniquely defined as being the minimizer of a strictly convex coercive function;
(b) for any $z \in \mathbb{R}^m$ one has $\mathrm{prox}_{\tau h}(z) \in \mathrm{dom}\, h$, which follows from Fermat's rule.

A fundamental inequality associated with the proximity operator is given next.

Lemma 2 (Proximal inequality)**.** *Let $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous convex function. Given $z \in \mathbb{R}^m$ and $\tau > 0$, if $x^+ = \mathrm{prox}_{\tau h}(z)$, then*

$$h(x) \geq h(x^+) - \frac{1}{\tau}\langle x - x^+,\, x^+ - z\rangle \quad \forall\, x \in \mathbb{R}^m.$$

The proximal gradient descent, known also as proximal forward-backward, has a crucial role when minimising the sum of a convex and a smooth function.

Definition 6. *Let $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous and convex function and $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable (not necessarily convex) function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient. The main iteration to minimize $f + h$ using the proximal gradient method starting from any $u \in \mathbb{R}^m$ is given by*

$$x^+ = \mathrm{prox}_{\tau h}\left(u - \tau \nabla f(u)\right). \tag{13}$$

The point x^+ is uniquely defined and satisfies $x^+ \in \mathrm{dom}\, h$, see Remark 5.

Remark 6. The proximal gradient descent can be seen as the minimization of $h + \widetilde{f}$, where $\widetilde{f}$ is a quadratic approximation of f around the point u:

$$\begin{aligned} x^+ &= \mathrm{prox}_{\tau h}\left(u - \tau \nabla f(u)\right) \\ &= \arg\min_{x \in \mathbb{R}^m} \left\{ h(x) + \frac{1}{2\tau}\|x - (u - \tau \nabla f(u))\|^2 \right\} \\ &= \arg\min_{x \in \mathbb{R}^m} \Big\{ h(x) + \underbrace{f(u) + \langle x - u, \nabla f(u)\rangle + \frac{1}{2\tau}\|x - u\|^2}_{=:\widetilde{f}(z)} \Big\}. \end{aligned}$$

We recall a classical result whose proof can be found in the textbook [12, A. 24].

Lemma 3 (Descent lemma)**.** *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient. Then*

$$f(u) \geq f(x) - \langle x - u, \nabla f(u)\rangle - \frac{\mathrm{L}_{\nabla f}}{2}\|x - u\|^2 \qquad \forall\, x, u \in \mathbb{R}^m. \tag{14}$$

The next lemma warrants a sufficient decrease of the objective after a proximal step.

Lemma 4. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient and $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ a convex, lower semicontinuous and proper function. Fix τ such that $\tau < 2/\mathrm{L}_{\nabla f}$. If x^+ is defined by* (13)*, then*

$$f(u) + h(u) \geq f(x^+) + h(x^+) + \left(\frac{1}{\tau} - \frac{\mathrm{L}_{\nabla f}}{2}\right)\|x^+ - u\|^2 \qquad \forall\, u \in \mathbb{R}^m \tag{15}$$

where $(1/\tau - \mathrm{L}_{\nabla f}/2)$ is a positive number.

Proof. Apply the proximal inequality to h with $x := u$ and $z := u - \tau \nabla f(u)$:

$$\begin{aligned} h(u) &\geq h(x^+) - \frac{1}{\tau}\left\langle u - x^+,\ x^+ - u + \tau \nabla f(u)\right\rangle \\ &= h(x^+) + \frac{1}{\tau}\|x^+ - u\|^2 + \left\langle x^+ - u, \nabla f(u)\right\rangle. \end{aligned}$$

Then, adding equation (14) for f and $x = x^+$ in the descent Lemma 3 to the result obtained above leads to the desired inequality (15). Finally, $(1/\tau - \mathrm{L}_{\nabla f}/2) > 0$ because of the choice of τ. □

4 A simple algorithm adapted to multiconvex coupling terms

4.1 Alternating Structure-Adapted Proximal gradient descent algorithm (ASAP)

According to Remark 6, the iterations of our algorithm sketched in subsection 1.3, see (7), are equivalent to minimizing the following quadratic approximations of F around the point x^{k-1} and of G around the point y^{k-1}, respectively:

$$x^k = \arg\min_{x \in U} \left\{ F(x^{k-1}) + \langle x - x^{k-1}, \nabla F(x^{k-1})\rangle + H(x, y^{k-1}) + \frac{1}{2\tau}\|x - x^{k-1}\|^2 \right\}, \tag{16}$$

$$y^k = \arg\min_{y \in V} \left\{ G(y^{k-1}) + \langle y - y^{k-1}, \nabla G(y^{k-1})\rangle + H(x^k, y) + \frac{1}{2\sigma}\|y - y^{k-1}\|^2 \right\}. \tag{17}$$

Algorithm 1 ASAP: Alternating Structure-Adapted Proximal gradient descent

Initialization: $(x^0, y^0) \in U \times V$ and $0 < \tau < 2/\mathrm{L}_{\nabla F}$, $0 < \sigma < 2/\mathrm{L}_{\nabla G}$

General Step: for $k = 1, 2, \ldots$, compute

$$x^k = \mathrm{prox}_{\tau H(\cdot, y^{k-1})}\left(x^{k-1} - \tau \nabla F(x^{k-1})\right), \tag{18}$$

$$y^k = \mathrm{prox}_{\sigma H(x^k, \cdot)}\left(y^{k-1} - \sigma \nabla G(y^{k-1})\right). \tag{19}$$

Hence, ignoring the terms that are constant at iteration k, namely $F(x^{k-1})$, $\langle x^{k-1}, \nabla F(x^{k-1})\rangle$, $G(y^{k-1})$ and $\langle y^{k-1}, \nabla G(y^{k-1})\rangle$, and using the definition of proximal gradient descent (Definition 6), the proposed ASAP algorithm takes the compact form stated below.

The step-sizes τ and σ are set according to Lemma 4 so that they ensure a sufficient decrease of the objective J at each step. They depend only on the Lipschitzians of ∇F and ∇G. According to the initialization, if $F = 0$ (resp., $G = 0$), then τ (resp., σ) is a positive number.

Remark 7. The ASAP algorithm confirms immediately two attractive points that are direct consequences of Remark 5:

(a) iterates (x^k, y^k) are always uniquely defined, even if F or G are nonconvex;

(b) for any $k \geq 1$, it holds that $(x^k, y^k) \in \mathrm{dom}\, J$.

4.2 ASAP algorithm with N blocks

The ASAP algorithm can be applied to the case when H is block multiconvex. Then the variable x is split into N blocks $x = \left(x_{(1)}, \ldots, x_{(N)}\right)$ and the optimization problems has the form

$$\begin{aligned} J\left(x_{(1)}, \ldots, x_{(N)}\right) &:= H\left(x_{(1)}, \ldots, x_{(N)}\right) + \textstyle\sum_{i=1}^N F_i\left(x_{(i)}\right), \\ H\left(x_{(1)}, \ldots, x_{(N)}\right) &= \widetilde{H}\left(x_{(1)}, \ldots, x_{(N)}\right) + \textstyle\sum_{i=1}^N \chi_{\mathcal{D}_i}(x_{(i)}). \end{aligned} \tag{20}$$

Analogously to Assumption (M), J is lower-bounded, and for any $i = 1, \ldots, N$, functions F_i obeys (M)(b) and $x_{(i)} \mapsto \widetilde{H}\left(x_{(1)}, \ldots, x_{(N)}\right)$ is convex and differentiable, and the set $\mathcal{D}_i \neq \varnothing$ is closed and convex, see (M)(c). The corresponding algorithm is:

Algorithm 2 ASAP for Multiconvex coupling term

Initialization: Updating rule (UR), $\left(x^0_{(i)}\right)_i$ and $0 < \tau_i < 2/\mathrm{L}_{\nabla F_i}$ for $i = 1, \ldots, N$

General Step: for $k = 1, 2, \ldots$

pick $i \in \{1, 2, \ldots, N\}$ according to UR and compute

$$x^k_{(i)} = \mathrm{prox}_{\tau_i H(x^k_{(1)}, \ldots, x^k_{(i-1)}, \cdot, x^{k-1}_{(i+1)}, \ldots, x^{k-1}_{(N)})}\left(x^{k-1}_{(i)} - \tau_i \nabla F_i(x^{k-1}_{(i)})\right).$$

If the update is sequential, the alternating system has the form

$$(x^{k-1}_{(1)}, \cdots, x^{k-1}_{(N)}) \to (x^k_{(1)}, x^{k-1}_{(2)}, \cdots, x^{k-1}_{(N)}) \to \cdots \to (x^k_{(1)}, \cdots, x^k_{(N)}). \tag{21}$$

Other block updates can also be used; in [39, sec. 3] a random shuffling was shown to provide better numerical results for a scheme of the form (5).

5 Convergence analysis of ASAP for biconvex coupling term

5.1 Essential convergence facts

Given an initial (x^0, y^0), the alternating system we have to study is of the form $(x^{k-1}, y^{k-1}) \to (x^k, y^{k-1}) \to (x^k, y^k)$. Our first result states the convergence of the sequence $\{J(x^k, y^k)\}_{k\in\mathbb{N}}$ to a real number J^* with a guaranteed decrease.

Proposition 3. *Let Assumption (M) hold and let $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by ASAP. Based on the initialization of the algorithm we set*

$$\rho_x := \frac{1}{\tau} - \frac{\mathrm{L}_{\nabla F}}{2} > 0, \quad \rho_y := \frac{1}{\sigma} - \frac{\mathrm{L}_{\nabla G}}{2} > 0, \quad \text{and} \quad \rho := \min\{\rho_x,\ \rho_y\} > 0. \tag{22}$$

(a) *For every $k \geq 1$ the following sufficient decrease property holds:*

$$\begin{aligned} J(x^{k-1}, y^{k-1}) &\geq J(x^k, y^{k-1}) + \rho\,\|x^k - x^{k-1}\|^2; \\ J(x^k, y^{k-1}) &\geq J(x^k, y^k) + \rho\,\|y^k - y^{k-1}\|^2; \\ J(x^{k-1}, y^{k-1}) &\geq J(x^k, y^k) + \rho\left(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2\right); \end{aligned} \tag{23}$$

and hence

$$J(x^{k-1}, y^{k-1}) \geq J(x^k, y^{k-1}) \geq J(x^k, y^k). \tag{24}$$

(b) *The sequences $\{J(x^k, y^k)\}_{k\in\mathbb{N}}$ and $\{J(x^k, y^{k-1})\}_{k\in\mathbb{N}}$ are non-increasing and interlaced by (24), hence they converge to a value denoted by J^*;*

(c) *We have* $\displaystyle\sum_{k=1}^{+\infty}\left(\|y^k - y^{k-1}\|^2 + \|x^k - x^{k-1}\|^2\right) < +\infty$
and hence $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$ *and* $\lim_{k\to\infty} \|y^{k+1} - y^k\| = 0$.

Proof. Applying Lemma 4 with $f := F$, $h := H(\cdot, y^{k-1})$, $x^+ := x^k$ and $x := x^{k-1}$, along with the definition of ρ in (22), shows that

$$\begin{aligned} J(x^{k-1}, y^{k-1}) &= F(x^{k-1}) + G(y^{k-1}) + H(x^{k-1}, y^{k-1}) \\ &\geq F(x^k) + G(y^{k-1}) + H(x^k, y^{k-1}) + \rho\,\|x^k - x^{k-1}\|^2 \\ &= J(x^k, y^{k-1}) + \rho\|x^k - x^{k-1}\|^2. \end{aligned} \tag{25}$$

Lemma 4 with $f := G$, $h := H(x^k, \cdot)$, $x^+ := y^k$ and $x := y^{k-1}$, and (22) yield

$$\begin{aligned} J(x^k, y^{k-1}) &= F(x^k) + G(y^{k-1}) + H(x^k, y^{k-1}) \\ &\geq F(x^k) + G(y^k) + H(x^k, y^k) + \rho\,\|y^k - y^{k-1}\|^2 \\ &= J(x^k, y^k) + \rho\,\|y^k - y^{k-1}\|^2. \end{aligned} \tag{26}$$

Bringing (25) and (26) together leads to

$$\begin{aligned} J(x^{k-1}, y^{k-1}) &\geq J(x^k, y^{k-1}) + \rho\,\|x^k - x^{k-1}\|^2 \\ &\geq J(x^k, y^k) + \rho\,\|x^k - x^{k-1}\|^2 + \rho\,\|y^k - y^{k-1}\|^2 \end{aligned} \tag{27}$$

which completes the proof of (a). It follows that the sequences $\{J(x^k, y^k)\}_k$ and $\{J(x^k, y^{k-1})\}_k$ are non-increasing and interlaced by (24), and bounded from below because J is lower bounded. Therefore, they converge to the same finite number J^*, which proves (b).

Using the inequalities in (27) we also have

$$\forall\, k \geq 1, \qquad J(x^{k-1}, y^{k-1}) - J(x^k, y^k) \geq \rho\left(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2\right). \tag{28}$$

For $K \geq 1$, summing (28) from $k = 1$ to K yields

$$\forall\, K \geq 1, \qquad J(x^0, y^0) - J(x^K, y^K) \geq \rho\sum_{k=1}^{K}\left(\|y^k - y^{k-1}\|^2 + \|x^k - x^{k-1}\|^2\right).$$

Statement (b) entails that

$$\forall\, K \geq 1, \qquad \frac{1}{\rho}\left(J(x^0, y^0) - J^*\right) \geq \sum_{k=1}^{K}\left(\|y^k - y^{k-1}\|^2 + \|x^k - x^{k-1}\|^2\right).$$

Taking the limit as $K \to \infty$ leads to

$$\sum_{k=1}^{+\infty} \left(\|y^k - y^{k-1}\|^2 + \|x^k - x^{k-1}\|^2 \right) \leq \frac{1}{\rho} \left(J(x^0, y^0) - J^* \right),$$

which establishes statement (c). □

From the proof of Proposition 3(c) we can derive a global $O(1/k)$ convergence rate for $\left\{ \|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2 \right\}_{k \in \mathbb{N}}$.

Corollary 2 (Convergence rate)**.** *Let Assumption (M) hold and $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by ASAP. Then for any $K \in \mathbb{N}$ it holds that*

$$\inf_{k \geq K} \left\{ \|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2 \right\} \leq \frac{1}{\rho K} \left(J(x^0, y^0) - J^* \right).$$

The following result, together with Proposition 3(c), shows that the partial subgradients of J vanish when k goes to infinity.

Proposition 4. *Let Assumption (M) hold and let $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by ASAP. Then for any $k \geq 1$, one has*

$$p_x^k := \nabla F(x^k) - \nabla F(x^{k-1}) + \frac{1}{\tau}(x^{k-1} - x^k) \in \partial_x J(x^k, y^{k-1}), \tag{29}$$

$$q_y^k := \nabla G(y^k) - \nabla G(y^{k-1}) + \frac{1}{\sigma}(y^{k-1} - y^k) \in \partial_y J(x^k, y^k), \tag{30}$$

such that

$$\|p_x^k\| \leq \left(\mathrm{L}_{\nabla F} + \frac{1}{\tau} \right) \|x^{k-1} - x^k\| \quad \textit{and} \quad \|q_y^k\| \leq \left(\mathrm{L}_{\nabla G} + \frac{1}{\sigma} \right) \|y^{k-1} - y^k\|. \tag{31}$$

Proof. The Fermat's rule for x^k in (16) yields

$$\frac{1}{\tau}(x^{k-1} - x^k) \in \nabla F(x^{k-1}) + \partial_x H(x^k, y^{k-1}). \tag{32}$$

From the original formula for J in (8) one has

$$\partial_x J(x^k, y^{k-1}) = \nabla F(x^k) + \partial_x H(x^k, y^{k-1}).$$

Subtracting (32) from this expression yields the result in (29). Similarly, using Fermat's rule for y^k in (17) – the y-step of the algorithm – and (8) leads to (30).

Using that ∇F has Lipschitzian $\mathrm{L}_{\nabla F}$ (see Assumption (M)(b)), it follows that

$$\|p_x^k\| \leq \mathrm{L}_{\nabla F} \|x^{k-1} - x^k\| + \frac{1}{\tau} \|x^{k-1} - x^k\| = \left(\mathrm{L}_{\nabla F} + \frac{1}{\tau} \right) \|x^{k-1} - x^k\|,$$

while the Lipschitz-continuity of ∇G leads to the bound on q_y^k. □

The set of all limit points of a sequence $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ generated by ASAP starting from a point (x^0, y^0) is denoted by $\mathcal{L}(x^0, y^0)$; see (11) in subsection 3.2.

Remark 8. According to Corollary 1, assuming that $\mathcal{L}(x^0, y^0)$ is nonempty and bounded is equivalent to assume that the iterates $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ are bounded.

Boundedness of the iterates is a usual assumption, see e.g., [5, 38, 17, 26, 9]. This assumption holds for instance when the level sets of J are bounded or if J is coercive.

Below we summarize several facts on the limit point set of ASAP.

Proposition 5. *Let Assumption (M) hold and let $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by ASAP which is assumed to be bounded. Let $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$.*

(a) *there is a subsequence* $\left(x^{k_j}, y^{k_j}\right)_{j\in\mathbb{N}}$ *such that* $(x^{k_j}, y^{k_j}) \to (x^*, y^*)$ *as* $j \to \infty$*;*

(b) $\lim\limits_{k\to\infty} J(x^k, y^k) = J(x^*, y^*)$*;*

(c) $(0,0) \in \partial J(x^*, y^*)$ *and thus* (x^*, y^*) *is a critical point of* J*.*

Proof. (a) follows from the definition of $\mathcal{L}(x^0, y^0)$.
(b) From Remark 7(b), (x^{k_j}, y^{k_j}) belongs to dom J. From Remark 4(b), $J(x^{k_j}, y^{k_j}) = \widetilde{J}(x^{k_j}, y^{k_j})$ where $\widetilde{J}$ is a continuous function. Therefore

$$\lim_{j\to\infty} J(x^{k_j}, y^{k_j}) = \lim_{j\to\infty} \widetilde{J}(x^{k_j}, y^{k_j}) = \widetilde{J}(x^*, y^*).$$

Since dom J is closed (see Remark 4(a)) one has $(x^*, y^*) \in$ dom J and $J(x^*, y^*) = \widetilde{J}(x^*, y^*)$. From Proposition 3(a)-(b) the sequence $\left\{J(x^k, y^k)\right\}_{k\in\mathbb{N}}$ is nonincreasing, interlaced by (24) and converges to J^*. This, together with the above results completes statement (b).
(c) The formula for J in (8), together with (29) and (30) in Proposition 4 yield

$$p_x^{k_j} - \nabla F(x^{k_j}) \in \partial_x H(x^{k_j}, y^{k_j-1}) \quad \text{and} \quad q_y^{k_j} - \nabla G(y^{k_j}) \in \partial_y H(x^{k_j}, y^{k_j}).$$

Since $x \mapsto H(x, y^{k_j-1})$ and $y \mapsto H(x^{k_j}, y)$ are convex, lower semicontinuous and proper functions (see Assumption (M)(c)) the subgradient inequality (Definition 2) can be applied which leads to

$$\forall\, x \in U, \quad H(x, y^{k_j-1}) \geq H(x^{k_j}, y^{k_j-1}) + \langle p_x^{k_j} - \nabla F(x^{k_j}), x - x^{k_j}\rangle,$$

$$\forall\, y \in V, \quad H(x^{k_j}, y) \geq H(x^{k_j}, y^{k_j}) + \langle q_y^{k_j} - \nabla G(y^{k_j}), y - y^{k_j}\rangle.$$

From (31) in Proposition 4 we have $(p_x^{k_j})_j \to 0$ and $(q_y^{k_j})_j \to 0$. By Proposition 3(c), $\left\{(y^{k_j-1} - y^{k_j})\right\}_{j\in\mathbb{N}}$ converges to 0. Since $\{y^{k_j}\}_{j\in\mathbb{N}}$ converges to y^*, it follows that $\{y^{k_j-1}\}_{j\in\mathbb{N}}$ converges to y^* as well. Using the facts that ∇F, ∇G and $H = \widetilde{H}$ are continuous functions on int(dom J) and that ∂J is closed (see Assumption (M) and Remark 4) we can evaluate the limit in the inequalities above when $j \to +\infty$:

$$\forall\, x \in U, \quad H(x, y^*) \geq H(x^*, y^*) + \langle 0 - \nabla F(x^*), x - x^*\rangle,$$

$$\forall\, y \in V, \quad H(x^*, y) \geq H(x^*, y^*) + \langle 0 - \nabla G(y^*), y - y^*\rangle.$$

Then, by the definition of the (partial) subgradients of H (see Definition 2),

$$0 \in \nabla F(x^*) + \partial_x H(x^*, y^*) \qquad \text{and} \qquad 0 \in \nabla G(y^*) + \partial_y H(x^*, y^*).$$

Invoking Proposition 1, it holds that $(0,0) \in \partial J(x^*, y^*)$. □

Proposition 6. *Let Assumption (M) hold and let* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by ASAP which is assumed to be bounded. Then the following properties hold:*

(a) $\mathcal{L}(x^0, y^0) \subset \mathrm{crit}(J)$*;*

(b) $\lim\limits_{k\to\infty} \mathrm{dist}\left((x^k, y^k),\, \mathcal{L}(x^0, y^0)\right) = 0$*;*

(c) $\mathcal{L}(x^0, y^0)$ *is connected;*

(d) J *is finite and constant on* $\mathcal{L}(x^0, y^0)$*.*

Proof. (a) follows from Proposition 5(c).
(b) is a direct consequence of Lemma 1 and of Proposition 3(c).
(c) is a consequence of Proposition 3(c) and the previous claim (b).
(d) The sequences $J(x^k, y^{k-1})$ and $J(x^k, y^k)$ decrease and converge to a finite value J^* by Proposition 3. □

According to (a), there may be critical points of J that cannot be reached when ASAP starts from a given initial (x^0, y^0) which emphasizes the role of this initial guess. Statement (b) guarantees that for k large enough, the iterates (x^k, y^k) are arbitrary close to a critical point of J. We conclude this subsection with the following remark:

Remark 9. Proposition 5 ensures only subsequential convergence of ASAP to critical points of J. In practice this result might be enough thanks to Proposition 6. Yet, if J shares some special properties (e.g., all critical points are isolated), one may be able to conclude with convergence of the whole sequence (x^k, y^k) towards a critical point of J, without considering the development presented in the following subsections 5.2–5.4. For an example, see [38, Cor. 2.4].

5.2 Subgradient convergence

Subgradient convergence can be proved if one considers an additional assumption on the regularity of the smooth part $\widetilde{H}$ of the coupling term H (see Assumption (M)(c) and (9)):

Assumption (H) $\nabla\widetilde{H}$ is one-sided locally Lipschitz continuous on bounded subsets of $U \times V$ in the sense that for each bounded subset $\mathcal{B}_U \times \mathcal{B}_V \subset U \times V$ there is a constant $\xi > 0$ such that

$$\forall\ (x,y),(x',y') \in \mathcal{B}_U \times \mathcal{B}_V, \qquad \|\nabla_x\widetilde{H}(x,y) - \nabla_x\widetilde{H}(x,y')\| \leq \xi\ \|y - y'\|.$$

A stronger assumption, namely that $\nabla\widetilde{H}$ is globally Lipschitz continuous on bounded subsets is often used in other papers, see e.g., [5, 17, 32, 9, 21]. Assumption (H) can also be derived if $\widetilde{H}$ is twice continuously differentiable.

Proposition 7. *Let Assumptions (M) and (H) hold. Let $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by ASAP which is assumed to be bounded. Then there exists $\xi \in (0,\infty)$ such that for any $k \geq 1$ one has*

$$\exists\ \left(q_x^k, q_y^k\right) \in \partial J(x^k, y^k) \quad \textit{obeying} \quad \left\|(q_x^k, q_y^k)\right\| \leq \beta \left\|\left(x^k - x^{k-1}, y^k - y^{k-1}\right)\right\|,$$

where

$$\beta := \max\left\{\sqrt{2}\left(\mathrm{L}_{\nabla F} + \frac{1}{\tau}\right), \sqrt{\left(\mathrm{L}_{\nabla G} + \frac{1}{\sigma}\right)^2 + 2\,\xi^2}\right\}.$$

Proof. Since the iterates are bounded, there exist $\mathcal{B}_U \times \mathcal{B}_V \subset U \times V$ bounded such that $(x^k, y^k) \in \mathcal{B}_U \times \mathcal{B}_V$ for any $k \in \mathbb{N}$. Using assumption (H), the constant ξ below is finite:

$$\forall\ k \in \mathbb{N}, \qquad \|\nabla_x\widetilde{H}(x^k, y^k) - \nabla_x\widetilde{H}(x^k, y^{k-1})\| \leq \xi\ \|y^k - y^{k-1}\|. \tag{33}$$

From the equivalent formula for J in (9), $p_x^k = \nabla F(x^k) + \nabla_x\widetilde{H}(x^k, y^{k-1}) + g_x^k \in \partial_x J(x^k, y^{k-1})$, where by (29) in Proposition 4 one has

$$g_x^k = -\nabla F(x^{k-1}) + \frac{1}{\tau}(x^{k-1} - x^k) - \nabla_x\widetilde{H}(x^k, y^{k-1}) \in \partial\chi_{\mathcal{D}_x}(x^k).$$

This result, together with (9) and (29) shows that $q_x^k \in \partial_x J(x^k, y^k)$ reads as

$$\begin{aligned} q_x^k &:= \nabla_x\widetilde{H}(x^k, y^k) + \nabla F(x^k) + g_x^k \\ &= \nabla_x\widetilde{H}(x^k, y^k) + \nabla F(x^k) - \nabla F(x^{k-1}) + \frac{1}{\tau}(x^{k-1} - x^k) - \nabla_x\widetilde{H}(x^k, y^{k-1}) \\ &= \nabla_x\widetilde{H}(x^k, y^k) - \nabla_x\widetilde{H}(x^k, y^{k-1}) + p_x^k. \end{aligned}$$

Using (33) and the first inequality in (31) in Proposition 4 yields

$$\begin{aligned} \|q_x^k\| &\leq \|\nabla_x\widetilde{H}(x^k, y^k) - \nabla_x\widetilde{H}(x^k, y^{k-1})\| + \|p_x^k\| \\ &\leq \xi\ \|y^k - y^{k-1}\| + \left(\mathrm{L}_{\nabla F} + \frac{1}{\tau}\right)\|x^k - x^{k-1}\|. \end{aligned}$$

Thus $(q_x^k, q_y^k) \in \partial J(x^k, y^k)$ where q_y^k is given in (30). The obtained result on $\|q_x^k\|$ together with the second inequality in (31) in Proposition 4, and using that $(a+b)^2 \leq 2\,a^2 + 2\,b^2$, shows that

$$\begin{aligned}
\|(q_x^k, q_y^k)\|^2 &= \|q_x^k\|^2 + \|q_y^k\|^2 \\
&\leq \left(\xi\|y^k - y^{k-1}\| + \left(\mathrm{L}_{\nabla F} + \frac{1}{\tau}\right)\|x^k - x^{k-1}\|\right)^2 + \left(\mathrm{L}_{\nabla G} + \frac{1}{\sigma}\right)^2 \|y^k - y^{k-1}\|^2 \\
&\leq 2\left(\mathrm{L}_{\nabla F} + \frac{1}{\tau}\right)^2 \|x^k - x^{k-1}\|^2 + \left(\left(\mathrm{L}_{\nabla G} + \frac{1}{\sigma}\right)^2 + 2\xi^2\right)\|y^k - y^{k-1}\|^2.
\end{aligned}$$

This gives the value of β in the statement which completes the proof. □

5.3 The Kurdyka-Łojasiewicz (KŁ) property and the objective J

The Kurdyka-Łojasiewicz (KŁ) property was studied for the optimization of nonsmooth functions originally in [15] and further on in [4, 5, 6].

Definition 7 (KŁ property)**.** *Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. The function f is said to have the Kurdyka-Łojasiewicz (KŁ) property at $x^* \in \operatorname{dom} \partial f$ if there exist $\eta \in (0, +\infty]$, a neighborhood $\mathcal{O}(x^*)$ of x^* and a constant $\kappa > 0$ such that*

$$\forall\, x \in \widetilde{\mathcal{O}}(x^*), \qquad \kappa\, \operatorname{dist}(0, \partial f(x)) \geq |f(x) - f(x^*)|^\theta\,, \tag{34}$$

where $\theta \in [0, 1)$ and

$$\widetilde{\mathcal{O}}(x^*) := \mathcal{O}(x^*) \cap \{x \in U \;\;|\;\; f(x^*) < f(x) < f(x^*) + \eta\}\,. \tag{35}$$

Observe that the KŁ property implies that all critical points belonging to $\widetilde{\mathcal{O}}(x^*)$ share the same critical value $f(x^*)$ since otherwise (34) would fail. The KŁ property requires assumptions only on the shape of the function around its critical points. The KŁ property *does not require* that the critical points are strict or connected. We recall that if f is a real analytic function on $\mathcal{O}(x^*)$, then the KŁ property holds with $\theta \in [1/2, 1)$ thanks to the Łojasiewicz gradient inequality [30, p. 92].

It is instructive to provide some examples and counter-examples.

Example 2.

(a) The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = (x-1)^2/2$ satisfies $|f'(x)|/\sqrt{2} = |f(x) - f(x^*)|^{1/2}$ for any $x \in \mathbb{R} \setminus \{x^*\}$ where $x^* = 1$ is its minimizer. Thus f satisfies the KŁ property at x^* with $\theta = 1/2$ and $\kappa = \sqrt{\theta}$.

(b) The function $h(x, y) = (xy - 1)^2$ is analytic, has a nonstrict minimizer at $(x^*, y^*) = (1, 1)$, and satisfies the KŁ property at (x^*, y^*).

(c) The functions below are infinitely differentiable but fail the KŁ property:

$$f(x) = \begin{cases} \exp\left(-1/x^2\right) & \text{if } x > 0, \\ 0 & \text{if } x \leq 0, \end{cases} \qquad f(x) = \begin{cases} x^2 \sin(1/x) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

The first one is "infinitely flat" [5, p. 453] while the second presents "wild oscillations" [16, p. 569].

We recall the following fundamental fact:

Remark 10. Both real-analytic and semi-algebraic functions are semi-analytic functions and thus they are subanalytic functions, according to [13].

Functions that are all together real-analytic and semi-algebraic form the class of Nash functions [14]; see, e.g., functions ψ in (i) and (iii) in Example 1. The objectives J we suggest in subsection 2.2 contain real-analytic and semi-algebraic functions and involve compositions and sums of such functions, so they belong to the wider class of extended-real-valued subanalytic functions. We will use the result obtained by Bolte, Daniilidis, and Lewis in [15] saying that all subanalytic functions enjoy the KŁ property at their critical points:

Theorem 1. *[15, Theorem 3.1] Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a function which is subanalytic with a closed domain and continuous on its domain. Let x^* be a critical point of f. Then there is an exponent $\theta \in [0, 1)$, a neighborhood $\widetilde{\mathcal{O}}(x^*)$ of x^* and a constant $\kappa > 0$ such that*

$$\forall\, x \in \widetilde{\mathcal{O}}(x^*), \qquad \kappa \operatorname{dist}(0, \partial f(x)) \geq |f(x) - f(x^*)|^\theta. \tag{36}$$

The following properties of subanalytic functions are taken from the monograph of Shiota [34] and from the work of Denkowska and Denkowski [23], respectively.

Proposition 8. *Let f and g be two subanalytic functions. Then the following results hold:*

(a) *[34, Chapter II.1] If f and g are lower-bounded, then $f + g$ is subanalytic.*

(b) *[23, Proposition 2.46] If g maps bounded sets on bounded sets or if $f^{-1}(\mathcal{X})$ is bounded for any bounded subset $\mathcal{X}$, then $f \circ g$ is a subanalytic function.*

Thanks to these results we can derive some useful properties of the objectives J in subsection 2.2. We begin with the terms composing J in (10).

Proposition 9. *Let $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be a subanalytic function and U a finite dimensional real space. Then:*

(a) *If A is a linear operator on U then $x \mapsto f(\|Ax\|)$ is a subanalytic function.*

(b) *If (V, W) are finite dimensional real spaces and $b := (b_{(1)}, \ldots, b_{(N)}) : U \times V \to W$ a bilinear mapping, then $(x, y) \mapsto f(\|b(x, y) - w\|)$ is a subanalytic function.*

Proof. (a) Since $\|\cdot\|$ is a distance function, the application $x \mapsto \|Ax\|$ is semi-algebraic and thus subanalytic (Remark 10), and it also maps bounded sets on bounded sets. Then its composition with f is subanalytic according to Proposition 8(b).
(b) For any $n = 1, \ldots, N$, $b_{(n)} : U \times V \to \mathbb{R}$ is a bilinear form. Since (U, V) are of finite dimension, $b_{(n)}$ can be represented using a real matrix A_n so that $b_{(n)}(x, y) = \langle A_n x, y\rangle$. Hence, $b_{(n)}$ is real polynomial. Consequently, $(x, y) \mapsto \|b(x, y) - w\|^2 = \sum_n (b_{(n)}(x, y) - w_{(n)})^2$ is semi-algebraic, and so is $(x, y) \mapsto \|b(x, y) - w\|$ since $x \mapsto x^{1/2}$ is a semi-algebraic function. Using Proposition 8(b), it follows that $(x, y) \mapsto f(\|b(x, y) - w\|)$ is subanalytic. □

Theorem 2. *For the family of J in subsection 2.2, the following hold:*

(a) *all functions in Example 1 are subanalytic;*

(b) *the objective function J in (10) is subanalytic on its domain. Hence, J satisfy the KŁ property at its critical points.*

Proof. (a) Functions ψ in (i), (ii) and (iii) in Example 1 are semi-algebraic and thus subanalytic. Further, since p is rational, then all $(\psi)^p$ in (i)-(iii) are semi-algebraic by composition [17, Ex. 4] and hence subanalytic. Functions ψ in (v), (vi) and (vii) are real-analytic, and thus subanalytic. Since $t \mapsto |t|$ maps bounded sets on bounded sets, and since $t \mapsto \log(1 + t/\alpha)$ is real-analytic, $t \mapsto \alpha\, \log(1 + |t|/\alpha)$ is subanalytic according to Proposition 8(b) Thus, function (iv) is the sum of the lower-bounded semi-algebraic function $t \mapsto |t|$ and a subanalytic function, hence it is subanalytic according to Proposition 8(a).
(b) According to Proposition 9, all functions $x \mapsto f_i(\|A_i x\|)$, $y \mapsto g_j(\|B_i y\|)$ and $(x, y) \mapsto h(\|b(x, y) - w\|)$ are subanalytic since f_i, g_j and h are subanalytic functions by (a). Thus J is a finite sum of lower-bounded subanalytic functions, hence J is subanalytic according to Proposition 8(a). Noticing also that J has a closed domain and is continuous on its domain, see Remark 4(b), Theorem 1 shows that J has the KŁ property at its critical points. □

5.4 Convergence of ASAP to critical points under the KŁ property

We summarize that, under Assumptions (M) and (H), we have proved that any *bounded* sequence generated by ASAP satisfy the assumptions in [6, Theorem 2.9]:

1) there exists $\rho \in (0, \infty)$ such that for any $k \geq 1$ the *sufficient decrease property* holds (Proposition 3(a)):

$$J(x^k, y^k) + \rho \left(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2 \right) \leq J(x^{k-1}, y^{k-1});$$

2) there exists a subsequence (x^{k_j}, y^{k_j}) and a critical point (x^*, y^*) of J such that the *subsequential continuity towards a critical point* holds (Proposition 5):

$$\lim_{j \to \infty} \left(x^{k_j}, y^{k_j} \right) = (x^*, y^*) \qquad \text{and} \qquad \lim_{j \to \infty} J \left(x^{k_j}, y^{k_j} \right) = J(x^*, y^*);$$

3) there exists $\beta \in (0, \infty)$ such that $\forall\, k \geq 1$ the *subgradient relative error condition* holds (Proposition 7):

$$\exists \left(q_x^k, q_y^k \right) \in \partial J(x^k, y^k) \quad \text{obeying} \quad \|(q_x^k, q_y^k)\| \leq \beta \|(x^k - x^{k-1}, y^k - y^{k-1})\|.$$

Theorem 3. *Let Assumptions (M) and (H) hold and let $\{(x^k, y^k)\}_{k \in \mathbb{N}}$ be a sequence generated by ASAP which is assumed to be bounded. Assume also that J has the KŁ property at a limit point $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$. Then the following hold:*

(a) *$\{(x^k, y^k)\}_{k \in \mathbb{N}}$ is a Cauchy sequence that converges to a critical point (x^*, y^*) of J as k goes to infinity;*

(b) *moreover,* $\displaystyle\sum_{k=0}^{+\infty} \left(\|x^{k+1} - x^k\| + \|y^{k+1} - y^k\| \right) < +\infty.$

Proof. This theorem is a direct consequence of [6, Theorem 2.9]. Our Definition 7 endowed with the function $\varphi : \mathbb{R} \to \mathbb{R}$ given by $\varphi(t) = \kappa t^{1-\theta}/(1-\theta)$ fulfills [6, Def. 2.4]. All conditions needed in [6, Theorem 2.9] are the three statements given before the theorem, along with the KŁ property as stated. □

Theorem 3 proves that for objectives J satisfying Assumptions (M) and (H), any bounded sequence generated by ASAP converges to a limit point $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$ provided J has the KŁ property at (x^*, y^*). This includes various *nonconvex* objective functions J that are continuous on their closed domain. A large family of nonconvex objective functions J was given in subsection 2.2. More generally, the objective functions J can be composed out of differentiable components that are real-analytic and semi-algebraic (nonconvex) functions, subanalytic (nonconvex) functions, strongly and uniformly convex functions.

6 Convergence of ASAP with H multiconvex

The variable x is split into N blocks, i.e., $x = (x_{(1)}, \ldots, x_{(N)}) \in U_1 \times \cdots \times U_N$ with U_i finite-dimensional real spaces, and the objective J reads as

$$J\left(x_{(1)}, \ldots, x_{(N)}\right) := \sum_{i=1}^{N} F_i\left(x_{(i)}\right) + \underbrace{\sum_{i=1}^{N} \chi_{\mathcal{D}_i}(x_{(i)}) + \widetilde{H}\left(x_{(1)}, \ldots, x_{(N)}\right)}_{=: H\left(x_{(1)}, \ldots, x_{(N)}\right)} \tag{37}$$

The objective and the full algorithm were presented in subsection 4.2 where the main assumptions were briefly described. In this section we adopt the following hypothesis:

Assumption (U) The update in Algorithm 2 is sequential, see (21).

Then all statements in subsections 5.1 and 5.2 have a straightforward extension to the case when J has $N > 2$ blocks, using analogous arguments and proofs, but at the expense of heavier notation. Only the main statements are presented.

The next assumption is a direct extension of Assumption (M):

Assumption (M')

(a) $J : U_1 \times \cdots \times U_N \to \mathbb{R} \cup \{+\infty\}$ is lower bounded;

for any $i = 1, \ldots, N$, (b) and (c) hold:

(b) $F_i : U_i \to \mathbb{R}$ has a gradient which is Lipschitz continuous with constant $\mathrm{L}_{\nabla F_i}$;

(c) $x_{(i)} \mapsto H(x)$ is convex and differentiable on its domain.

Proposition 10. *Let Assumptions (M') and (U) hold and let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by ASAP in Algorithm 2. Then the following holds:*

(a) *for every $k \geq 1$ one has $J(x^{k-1}) \geq J(x^k) + \rho\|x^k - x^{k-1}\|^2$ where*

$$\rho := \min_{i=1,\ldots,N} \left\{ \frac{1}{\tau_i} - \frac{L_{\nabla F_i}}{2} \right\} > 0;$$

(b) *assuming that $\{x^k\}_{k\in\mathbb{N}}$ is bounded and denoting $x^* \in \mathcal{L}(x^0)$ a limit-point of $\{x^k\}_{k\in\mathbb{N}}$, one has $\lim_{k\to\infty} J(x^k) = J(x^*)$ and x^* is a critical point of J.*

Proof. (a) As in the proof of Proposition 3(a), applying Lemma 4 with $f := F_i$ and $h := H(x^k_{(1)}, \ldots, x^k_{(i-1)}, \cdot, x^{k-1}_{(i+1)}, \ldots, x^{k-1}_{(N)})$ yields for any $i = 1, \ldots, N$

$$\begin{aligned} &J(x^k_{(1)}, \cdots, x^k_{(i-1)}, x^{k-1}_{(i)}, \cdots, x^{k-1}_{(N)}) \\ &\qquad \geq J(x^k_{(1)}, \cdots, x^k_{(i-1)}, x^k_{(i)}, x^{k-1}_{(i+1)}, \cdots, x^{k-1}_{(N)}) + \rho\, \|x^k_{(i)} - x^{k-1}_{(i)}\|^2. \end{aligned}$$

The result is obtained by summation over $i = 1, \cdots, N$.
(b) By Remark 6, $x^k_{(i)}$ is defined for any $i = 1, \ldots, N$ and for all $k \geq 1$ by:

$$\arg\min_u \left\{ \langle u, \nabla F_i(x^{k-1}_{(i)})\rangle + H(x^k_{(1)}, \ldots, x^k_{(i-1)}, u, x^{k-1}_{(i+1)}, \ldots, x^{k-1}_{(N)}) + \frac{1}{2\tau_i}\|u - x^{k-1}_{(i)}\|^2 \right\}.$$

By Fermat's rule, for any $i = 1, \ldots, N$ it holds that

$$\frac{x^{k-1}_{(i)} - x^k_{(i)}}{\tau_i} \in \nabla F_i(x^{k-1}_{(i)}) + \nabla_{x_i} \widetilde{H}(x^k_{(1)}, \ldots, x^k_{(i-1)}, x^k_{(i)}, x^{k-1}_{(i+1)}, \ldots, x^{k-1}_{(N)}) + \partial \chi_{\mathcal{D}_i}(x^k_{(i)}). \tag{38}$$

Then the claim is proven using the pipeline established in the proofs of Propositions 4 and 5. □

As in the biconvex case, strong subgradient convergence can be stated using an additional assumption on the smooth part $\widetilde{H}$ of the coupling term:

Assumption (H') $\nabla\widetilde{H}$ is partially locally Lipschitz continuous on bounded subsets in the sense that for each bounded subset $\mathcal{B}_1 \times \cdots \times \mathcal{B}_N \subset U_1 \times \cdots \times U_N$ there is a constant $\xi > 0$ such that for any $x, x' \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_N$ such that $x_{(j)} = x'_{(j)}$ if $j \leq i$,

$$\left\| \nabla_{x_{(i)}} \widetilde{H}(x) - \nabla_{x_{(i)}} \widetilde{H}(x') \right\| \leq \xi\, \|x - x'\|.$$

The papers dealing with multiconvex coupling terms use a slightly stronger assumption to prove subgradient convergence [38, 39].

Proposition 11. *Let Assumptions (M'), (U) and (H') hold. Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by ASAP (Algorithm 2) which is assumed to be bounded. Then there exists $\xi \in (0, \infty)$ such that for any $k \geq 1$ one has*

$$\exists\, q^k \in \partial J(x^k) \quad \textit{obeying} \quad \left\| q^k \right\| \leq \beta \left\| x^k - x^{k-1} \right\|,$$

where

$$\beta := \max\left\{ \sqrt{2}\gamma_1,\ \max_{i=2,\ldots,N-1} \left\{ \sqrt{2\left(\gamma_i^2 + (i-1)\,\xi^2\right)} \right\},\ \sqrt{\gamma_N^2 + 2\,(N-1)\,\xi^2} \right\}$$

and $\gamma_i := \mathrm{L}_{\nabla F_i} + 1/\tau_i$ for any $i = 1, \ldots, N$.

Proof. Using that the iterates are bounded and Assumption (M'), for any $k \in \mathbb{N}$ and for any $i = 1, \cdots, N$

$$\left\| \nabla_{x_{(i)}} \widetilde{H}(x^k) - \nabla_{x_{(i)}} \widetilde{H}(x^k_{(1)}, \cdots, x^k_{(i)}, x^{k-1}_{(i+1)}, \cdots, x^{k-1}_{(N)}) \right\| \leq \xi \sqrt{\sum_{j=i+1}^{N} \|x^k_{(j)} - x^{k-1}_{(j)}\|^2}.$$

Following the proof of Proposition 7, the idea is to find $g_i^k \in \partial \chi_{\mathcal{D}_i(x^k)}$ in order to obtain $q_i^k \in \partial_{x_{(i)}} J(x^k)$. The expression for g_i^k folllows from (38). Using that $H = \widetilde{H} + \sum_{i=1}^n \chi_{\mathcal{D}_i}$ with $\widetilde{H}$ a differentiable function, $q^k_{(i)} \in \partial_{x_{(i)}} J(x^k)$ is given by

$$q^k_{(i)} = \nabla_{x_{(i)}} \widetilde{H}(x^k) + \nabla F_i \left(x^k_{(i)} \right) + g^k_{(i)}.$$

Inserting the expression for g_i^k in the above equation shows that

$$\begin{aligned} q^k_{(i)} &= \nabla_{x_{(i)}} \widetilde{H}(x^k) - \nabla_{x_{(i)}} \widetilde{H}(x^k_{(1)}, \cdots, x^k_{(i)}, x^{k-1}_{(i+1)}, \cdots, x^{k-1}_{(N)}) \\ &\quad + \nabla F_i(x^k_{(i)}) - \nabla F_i(x^{k-1}_{(i)}) + \frac{1}{\tau_i}(x^{k-1}_{(i)} - x^k_{(i)}). \end{aligned}$$

Using Assumption (H') and the Lipschitz continuity of ∇F_i, the last equation yields

$$\|q^k_{(i)}\| \leq \xi \sqrt{\sum_{j=i+1}^{N} \|x^{k-1}_{(j)} - x^k_{(j)}\|^2} + (\mathrm{L}_{\nabla F_i} + 1/\tau_i)\|x^{k-1}_{(i)} - x^k_{(i)}\|,$$

which in particular gives $\|q^k_{(N)}\| \leq (\mathrm{L}_{\nabla F_N} + 1/\tau_i)\|x^{k-1}_{(N)} - x^k_{(N)}\|$. Then, using yet again that $(a+b)^2 \leq 2\,a^2 + 2\,b^2$, the sum of $\|q^k_{(i)}\|^2$ over $i = 1, \ldots, N$ is computed. Its square root is the desired $\|q^k\|$ and β follows from the obtained inequality. □

We have proven that ASAP for multiconvex coupling terms satisfies all the three conditions in [6, Theorem 2.9]: sufficient decrease by Proposition 10(a), subgradient relative error by Proposition 11 and subsequential continuity towards a critical point by Proposition 10(b). Therefore we can state the following result:

Theorem 4. *Let Assumptions (M'), (U) and (H') hold. Let $\{x^k\}_{k\in\mathbb{N}}$ be a sequence generated by ASAP (Algorithm 2) which is assumed to be bounded. Assume also that J has the KŁ property at a limit point $(x^*) \in \mathcal{L}(x^0)$. Then the following assertions hold:*

(a) *$\{(x^k)\}_{k\in\mathbb{N}}$ is a Cauchy sequence that converges to a critical point (x^*) of J as k goes to infinity;*

(b) *moreover,* $\displaystyle\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty$.

Let us consider general objectives of form (see subsection 2.2)

$$J(x) := \sum_{i,j} f_{ij}(\|A_{ij} x_{(j)}\|) + h(\|b(x) - w\|) + \sum_j \chi_{\mathcal{D}_j}(x_{(j)}), \tag{39}$$

where $b : U_1 \times \cdots \times U_N \to W$ is a multilinear form, A_{ij} are linear mappings and $f_{i,j}$ are as in Example 1. The results in subsection 5.3 show that J in (39) is also a subanalytic function and thus fulfils the KŁ property.

7 An application for Hadamard based coupling terms

Here we focus on objective functions J of the form presented in subsection 2.2 for $U = V = W = \mathbb{R}^{m\times n}$ and a bilinear mapping b given by $b(x, y) = x \circ y$ where "$\circ$" denotes the Hadamard (componentwise) product. Following Assumption (M), the coupling term H reads as

$$H(x, y) := \frac{1}{2}\|x \circ y - w\|^2 + \chi_{\mathcal{D}_x}(x) + \chi_{\mathcal{D}_y}(y).$$

Hadamard matrix products arise in lossy compression, tensor factorization, image processing, statistics, among others.

7.1 The ASAP algorithm with Hadamard product

Here x, y and w are seen as mn-length vectors. We focus on constraints $(\mathcal{D}_x, \mathcal{D}_y)$ of the form

$$\mathcal{D}_x := \{x \in U \ \mid \ a_x \leq x_i \leq b_x\} \quad \text{and} \quad \mathcal{D}_y := \{y \in U \ \mid \ a_y \leq x_i \leq b_y\}$$

for some constants $-\infty \leq a_x < b_x \leq +\infty$ and $-\infty \leq a_y < b_y \leq +\infty$. We define the function $h : \mathbb{R}^2 \to \mathbb{R}$ by

$$h(x_i, y_i) = \frac{1}{2}(x_i\, y_i - w_i)^2$$

so that $\widetilde{H}(x, y) = \sum_{i=1}^{mn} h(x_i, y_i)$. Then the proximity operators are computed componentwisely $\mathrm{prox}_{\tau H(\cdot,y)}(z) = \Big(\mathrm{prox}_{\tau h(\cdot,y_i)}(z_i)\Big)_{i=1}^{mn}$ where

$$\mathrm{prox}_{\tau h(\cdot,y_i)}(z_i) = \arg\min_{x_i \in [a_x, b_x]} \left\{ \frac{1}{2}(x_i y_i - w_i)^2 + \frac{1}{2\tau}(x_i - z_i)^2 \right\} \quad \forall\, i.$$

The computation of $\mathrm{prox}_{\sigma H(x,\cdot)}(z)$ is done in a similar way.

The full algorithms is particularly simple:

Algorithm 3 ASAP with Hadamard product

Initialization: $(x^0, y^0) \in U \times V$ and $0 < \tau < 2/\mathrm{L}_{\nabla F}$, $0 < \sigma < 2/\mathrm{L}_{\nabla G}$

General Step: for $k = 1, 2, \ldots$, compute

$$x^k = \min\left\{ \max\left\{ \frac{\tau w \circ y^{k-1} + z}{\tau(y^{k-1} \circ y^{k-1} + 1)},\, a_x \right\},\, b_x \right\} \quad \text{with } z = x^{k-1} - \tau \nabla F(x^{k-1});$$

$$y^k = \min\left\{ \max\left\{ \frac{\sigma w \circ x^k + z}{\sigma(x^k \circ x^k + 1)},\, a_y \right\},\, b_y \right\} \quad \text{with } z = y^{k-1} - \sigma \nabla G(y^{k-1}).$$

7.2 Fringe Separation in Interferometric Images

Sieleters [22] is a nonconventional infrared spectro-imaging device, which is based on interferometric imaging. Its purpose is to obtain hyperspectral images from a temporal image sequence, provided the images can be accurately registered. Due to the imaging system, the images provided by Sieleters have so-called interference fringes, which have a particular structure (see Fig. 7.2 bottom left). However, for many applications (such as stereo-matching, which is necessary to compute accurate image registration), these fringes have to be removed, so that one can recover the so-called panchromatic images. The latter can then be handled by conventional image processing techniques.

According to [35], a multiplicative model for the fringe formation was shown to be much more physically accurate than previous additive decomposition models. According to this model, an observed image $w \in \mathbb{R}^{m \times n}$ is given by

$$w = x \circ y + \text{perturbations},$$

where x is the panchromatic image to recover, y – the fringe oscillations.and $\mathtt{1}$ stand for an image composed of ones. Two facts known from the physics of the observation device are that

- the fringe oscillations have a range constraint, $|y_{i,j}| \leq 1$;
- the 1D Fourier transform of the columns of the fringe y, denoted by $\mathcal{F}(y)$, belongs to a known interval.

Applying these constraints enables the obtention of a good initialization (x^0, y^0). Other observations are that the gradients of the panchromatic image and of the fringes are *not sparse* and that the perturbations under the multiplicative model are negligible. we focus on an optimization problem of the form

$$J(x, y) = \underbrace{\sum_i \psi(D_i^v x)}_{=:F(x)} + \underbrace{\mu \sum_j \psi(D_j^h y) + \frac{\nu}{2}\|\mathcal{F}^{-1} M \mathcal{F}(y)\|^2}_{=:G(y)} + \underbrace{\frac{\eta}{2}\|x \circ y - w\|^2 + \chi_{\mathcal{D}_y}(y)}_{=:H(x,y)}$$

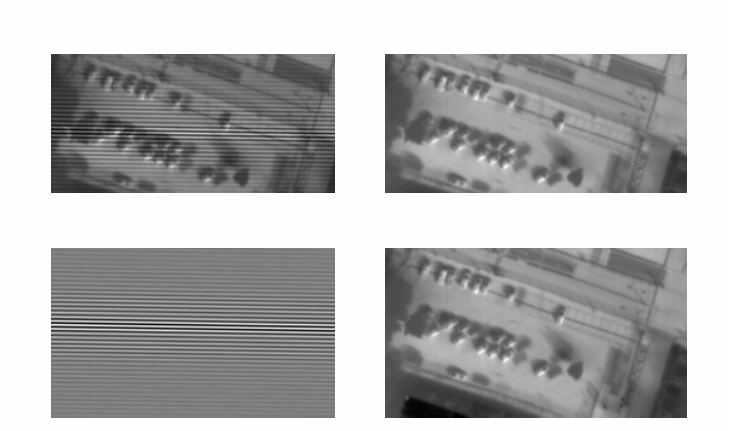

Figure 1: Result with Algorithm 3 (zooms). Clockwise: data, recovered panchromatic image, recovered fringe, reference

where ψ is function (iv) in Example 1, $\{D_i^v\}_i$ and $\{D_j^h\}_j$ are vertical and horizontal first-order finite difference operators, respectively, M is a binary mask ensuring the spectral constraint, and $\mathcal{D}_y = \{y \mid -1 \leq y_{i,j} \leq 1\}$. Finally, μ, ν and η are positive tuning parameters and thus $\mathrm{L}_{\nabla F} = 4/\alpha$ and $\mathrm{L}_{\nabla G} = 4\mu/\alpha + \nu$.

We were given 9 "ground-truth" 424×1000 images computed from an ONERA's Sieleters image sequence using a protocol based on the physical model. Fig. 7.2 shows 140×250 zooms. Algorithm 3 was successfully applied on Sieleters sequences, which typically consist in thousands of such airborne images.

Appendix

7.3 Proof of Lemma 1

For simplicity, set $\mathcal{L} := \mathcal{L}(z^0)$. Since $\mathcal{L}$ is closed and nonempty, the distance $\mathrm{dist}(z^k, \mathcal{L})$ of z^k to the set $\mathcal{L}$ is well defined. Suppose that $\mathrm{dist}(z^k, \mathcal{L})$ does not go to zero as $k \to \infty$, i.e., that there exist $\varepsilon > 0$ and a subsequence $\{z^{k_j}\}_j$ such that

$$\forall\, j \in \mathbb{N} \qquad \mathrm{dist}(z^{k_j}, \mathcal{L}) > 2\varepsilon. \tag{40}$$

Observe that $\{\|z^{k+1} - z^k\|\}_{k\in\mathbb{N}}$ is a Cauchy sequence of non-negative numbers converging to zero. This, together with the fact that $z \mapsto \mathrm{dist}(z, \mathcal{L})$ is a continuous function [33, p. 19.], entails that for $\varepsilon > 0$ there is $K \in \mathbb{N}$ such that for any $k > K$ one has $\big|\mathrm{dist}(z^{k+1}, \mathcal{L}) - \mathrm{dist}(z^k, \mathcal{L})\big| < \varepsilon$. Therefore, there exists $K' \in \mathbb{N}$ so that for any $j > K'$ one has $k_j > K$ and thus

$$\forall\, j > K' \qquad \mathrm{dist}(z^{k_j+1}, \mathcal{L}) - \mathrm{dist}(z^{k_j}, \mathcal{L}) < \varepsilon.$$

This, together with (40), yields

$$\mathrm{dist}(z^{k_j+1}, \mathcal{L}) > \mathrm{dist}(z^{k_j}, \mathcal{L}) - \varepsilon > \varepsilon.$$

From the definition of $\mathcal{L}$, see (11), there is an infinite number of points of z^k satisfying $\mathrm{dist}(z^k, \mathcal{L}) \leq \varepsilon$. Hence there exists $k_j + 1 \leq k_{j+1}$ such that $\mathrm{dist}(z^{k_j+1}, \mathcal{L}) < \varepsilon$; thus, a contradiction.

7.4 Proof of Corollary 1

If $\mathcal{L}(z^0)$ is nonempty and bounded, there exists $M > 0$ such that any $z^* \in \mathcal{L}(z^0)$ satisfies $\|z^*\| \leq M$. From Lemma 1, the distance of z^k to $\mathcal{L}(z^0)$ goes to zero as k goes to infinity, hence there exists $k_0 \in \mathbb{N}$ such that $k \geq k_0$ implies that $\min_{z\in\mathcal{L}(z^0)} \|z^k - z\| < M$. Therefore, for any $k \geq k_0$, one has $\|z^k\| \leq \min_{z\in\mathcal{L}(z^0)} \|z^k - z\| + \|z^*\| \leq 2M$, hence (a) $\Rightarrow$ (b). The converse claim is obvious.

Acknowledgements

The authors would like to thank Jérôme Bolte for helpful comments on Remark 10.

This work has been partially funded by the French Research Agency (ANR) under grant No ANR-14-CE27-001 (MIRIAM).

References

[1] P.-A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM J. Optim., 16 (2006), pp. 531–547.

[2] C. Aguerrebere, A. Almansa, Y. Gousseau, and P. Musé, *A hyperprior bayesian approach for solving image inverse problems*, IEEE Trans. on Comput Imaging, (2017).

[3] P. Arias, V. Caselles, and G. Facciolo, *Analysis of a variational framework for exemplar-based image inpainting*, SIAM J. Multiscale Model Simul, 10 (2012), pp. 473–514.

[4] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Math. Program., 116 (2009), pp. 5–16.

[5] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.

[6] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Math. Program., 137 (2013).

[7] A. Auslender, *Asymptotic properties of the Fenchel dual functionaland applications to decomposition problems*, J. Optim. Theory Appl., 73 (1992), pp. 427–449.

[8] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011.

[9] A. Beck, S. Sabach, and M. Teboulle, *An alternating semiproximal method for nonconvex regularized structured total least squares problems*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 1129–1150.

[10] A. Beck and M. Teboulle, *Smoothing and first order methods: a unified framework*, SIAM J. Optim., 22 (2012), pp. 667–580.

[11] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 28 (2013), pp. 2037–2060.

[12] D. P. Bertsekas, *Nonlinear programming*, Athena Scientific, Belmont, Massachussetts, 1995.

[13] E. Bierstone and P. D. Milman, *Semianalytic and subanalytic sets*, Publ. Math. Inst. Hautes Études Sci., 1988 (1988).

[14] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, Ergeb. Math. Grenzgeb. (3) 36, Springer-Verlag, Berlin, 1998.

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

[16] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM J. Optim., 18 (2007), pp. 556–572.

[17] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program. ser. A, 146 (2014).

[18] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134 (2012), pp. 71–99.

[19] X. Chen and W. Zhou, *Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization*, SIAM J. Imaging Sci., 3 (2010), pp. 765–790.

[20] ——, *Penalty methods for a class of non-Lipschitz optimization problems*, SIAM J. Optim., 26 (2016), pp. 1465–1492.

[21] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, *A block coordinate variable metric forwardbackward algorithm*, J. Global Optim., 66 (2016), pp. 457–485.

[22] C. Coudrain, S. Bernhardt, M. Caes, R. Domel, Y. Ferrec, R. Gouyon, D. Henry, M. Jacquart, A. Kattnig, P. Perrault, L. Poutier, L. Rousset-Rouvire, M. Tauvy, S. Thétas, and J. Primot, *SIELETERS, an airborne infrared dual-band spectro-imaging system for measurement of scene spectral signatures*, Optics express, 23 (2015), pp. 16164–16176.

[23] Z. Denkowska and M. P. Denkowski, *A long and winding road to definable sets*, J. Singul., 13 (2015), pp. 57–86.

[24] J. Gorski, F. Pfeuffer, and K. Klamroth, *Biconvex sets and optimization with biconvex functions: a survey and extensions*, Math. Meth. Oper. Res., 66 (2007).

[25] L. Grippo and M. Sciandrone, *On the convergence of the block nonlinear Gauss-Seidel method under convex constraints*, Oper Res Lett, 26 (2000), pp. 127–136.

[26] R. Hesse, D. R. Luke, S. Sabach, and M. K. Tam, *Proximal heterogeneous block implicit-explicit method and application to blind ptychographic diffraction imaging*, SIAM J. Imaging Sci., 8 (2015), pp. 426–457.

[27] C. Hildreth, *A quadratic programming procedure*, Naval Res. Logist., 4 (1957), pp. 79–85.

[28] M. Hintermüller and T. Wu, *Nonconvex TV^q-models in image restoration: Analysis and a trust-region regularization–based superlinearly convergent solver*, SIAM J. Imaging Sci., 6 (2013), pp. 1385–1515.

[29] M.-J. Lai and J. Wang, *An unconstrained ℓ^q minimization with $0 < q \leq 1$ for sparse solution of underdetermined linear systems*, SIAM J. Optim., 21 (2011), pp. 82–101.

[30] S. Łojasiewicz, *Ensembles semi-analytiques*, preprint IHES, France, 1965.

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

[32] T. Pock and S. Sabach, *Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems*, SIAM J. Imaging Sci., 9 (2016), pp. 1756–1787.

[33] R. T. Rockafellar and J. B. Wets, *Variational analysis*, Springer-Verlag, New York, 1998.

[34] M. Shiota, *Geometry of subanalytic and semialgebraic sets*, vol. 150, Springer Science & Business Media, 2012.

[35] D.-C. Soncco, C. Barbanson, M. Nikolova, A. Almansa, and Y. Ferrec, *Fast and accurate multiplicative decomposition for fringe removal in interferometric images*, IEEE Trans. on Comput Imaging, 3 (2017), pp. 187–201.

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

[37] Y. Xu, *Alternating proximal gradient method for sparse nonnegative Tucker decomposition*, Math. Program. Comput., 7 (2015), pp. 39–70.

[38] Y. Xu and W. Yin, *A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion*, SIAM J. Imaging Sci., 6 (2013), pp. 1758–1789.

[39] ——, *A globally convergent algorithm for nonconvex optimization based on block coordinate update*, J. Sci. Comput., (2017).

[40] X. Zhang, X. Zhang, X. Li, Z. Li, and S. Wang, *Classify social image by integrating multi-modal content*, Multimedia Tools and Applications, (2017).