SIAM J. OPTIM.
Vol. 29, No. 3, pp. 2053–2078

ALTERNATING STRUCTURE-ADAPTED PROXIMAL GRADIENT DESCENT FOR NONCONVEX NONSMOOTH BLOCK-REGULARIZED PROBLEMS[*]

MILA NIKOLOVA[†] AND PAULINE TAN[‡]

Dedicated to the memory of Mila Nikolova

Abstract. There has been increasing interest in constrained nonconvex regularized block optimization problems. We introduce an approach that enables complex application-dependent regularization terms to be used. The proposed alternating structure-adapted proximal gradient descent algorithm enjoys simple well-defined updates and is proved to be a value-convergent descent scheme in general cases. Global convergence of the algorithm to a critical point is proved using the so-called Kurdyka–Łojasiewicz property.

Key words. alternating minimization, block coordinate descent, global convergence, Kurdyka–Łojasiewicz property, nonconvex-nonsmooth optimization, forward-backward splitting, proximal gradient descent, subdifferential calculus

AMS subject classifications. 90C30, 90C26, 90C46, 65K05, 65K10, 65F22, 49M37, 47J25

DOI. 10.1137/17M1142624

1. Introduction.

Recently, there has been increasing interest in the design and analysis of regularized block optimization problems. In this work we consider problems of the form

$$\min_{x\in U, y\in V} J(x,y) := F(x) + G(y) + H(x,y), \tag{1}$$

where x and y belong to finite-dimensional real vector spaces U and V, respectively. Such an objective is also known as a two-block optimization model with blocks x and y. In our setting, the functions F and G are supposed continuously differentiable (for example, smooth approximations of nonsmooth functions). A natural extension of (1) to N blocks in finite-dimensional real vector spaces $\{U_i\}_{i=1}^N$ reads as

$$\min_{z=(z_{(1)},\dots,z_{(N)})\in U_1\times\cdots\times U_N} J(z_{(1)},\dots,z_{(N)}) := \sum_{i=1}^N F_i(z_{(i)}) + H(z), \tag{2}$$

where, in our setting, the functions $F_i : U_i \to \mathbb{R}$ are continuously differentiable. For ease of presentation, we study the two-block problem (1); the results on the multiblock problem (2) can easily be derived by the interested reader.

[*] Received by the editors August 7, 2017; accepted for publication (in revised form) May 14, 2019; published electronically July 30, 2019.
https://doi.org/10.1137/17M1142624
Funding: The work of the first author was partially supported by the French Research Agency (ANR) grant ANR-14-CE27-001 (MIRIAM). The work of the second author was supported by grants from Région Île-de-France.

[†] The author is deceased. Former address: CMLA, CNRS, ENS Cachan, Université Paris-Saclay, 61 Avenue du Président Wilson, 94230 Cachan, France.

[‡] CMLA, CNRS, ENS Cachan, Université Paris-Saclay, 61 Avenue du Président Wilson, 94230 Cachan, France. Current address: LJLL, Sorbonne Université, 4 place Jussieu, 75005 Paris, France (pauline.tan@sorbonne-universite.fr).

Optimization problems of the form (1), (2) are widely used in various areas of science and engineering. They are rich enough to cover various practical applications, such as blind source separation, blind deconvolution, nonnegative matrix factorization,[1] structured total least squares, multimodal learning for image classification [41], and patch-based methods for inpainting [3].

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 generates a sequence of iterates $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ defined by

$$x^k \in \arg\min_{x\in U} J(x, y^{k-1}), \tag{3a}$$

$$y^k \in \arg\min_{y\in V} J(x^k, y). \tag{3b}$$

This classical scheme is known as the block coordinate Gauss–Seidel method or block coordinate descent (BCD) method and was considered for various problems; see, e.g., [26, 37, 24, 3, 13, 2]. Analytical results for this approach are generally difficult to obtain. If J is continuously differentiable and if the minimum in each step is uniquely attained, convergence to a critical point holds [14, 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 (3a), (3b) is to consider the proximal BCD scheme

$$x^k \in \arg\min_{x\in U} \left\{ J(x, y^{k-1}) + \frac{1}{2\tau}\|x - x^{k-1}\|^2 \right\}, \tag{4a}$$

$$y^k \in \arg\min_{y\in V} \left\{ J(x^k, y) + \frac{1}{2\sigma}\|y - y^{k-1}\|^2 \right\}, \tag{4b}$$

which is equivalent to replacing the partial minimizations (3a) and (3b) by proximal updates. This approach was introduced for convex functions J by Auslender [9, sect. 4]. Convergence facts on (4a), (4b) for other nonconvex nonsmooth objectives can be found in [8, 5, 6, 39].

Note that the BCD and the proximal BCD schemes in (3a), (3b) and in (4a), (4b), respectively, generally entail an inner minimization problem, and hence a higher computational burden when the minimizer is not given by a closed formula.

1.2. Related literature: Prox-linearized BCD/PALM. An efficient approach to dealing with the difficulties arising with the (proximal) BCD for nonconvex and nonsmooth objectives J was proposed in 2013 by Xu and Yin [39] for block multiconvex J with H differentiable. Bolte, Sabach, and Teboulle [18] extended it for larger classes of objective functions, which include proper lower semicontinuous extended-valued functions (F, G). *Assuming the smoothness of* H, the idea was to apply a proximal linearized BCD to generate (x^k, y^k):

$$x^k \in \arg\min_{x\in U} \left\{ \langle x, \nabla_x H(x^{k-1}, y^{k-1})\rangle + F(x) + \frac{1}{2\tau_k}\|x - x^{k-1}\|^2 \right\}, \tag{5a}$$

$$y^k \in \arg\min_{y\in V} \left\{ \langle y, \nabla_y H(x^k, y^{k-1})\rangle + G(y) + \frac{1}{2\sigma_k}\|y - y^{k-1}\|^2 \right\}, \tag{5b}$$

[1] The nonsmooth nonnegativity constraint is then transferred to the coupling part H.

which is called prox-linearized BCD/PALM[2] following the names of the original algorithms. The step sizes (τ_k, σ_k) are computed according to

$$\tau_k \in (0, (\mathrm{Lip}(\nabla_x H(\cdot, y^{k-1})))^{-1}) \quad \text{and} \quad \sigma_k \in (0, (\mathrm{Lip}(\nabla_y H(x^k, \cdot)))^{-1}), \tag{6}$$

where "Lip" denotes the Lipschitz constant of the function in parentheses. The scheme needs functions (F, G) that are "simple" in the sense that their proximity operator

$$\mathrm{prox}_{\tau F}(z) := \arg\min_{x \in U} \left\{ F(x) + \frac{1}{2\tau} \|x - z\|^2 \right\} \tag{7}$$

has a closed-form expression or can be accurately calculated via a fast scheme. This approach became very popular and was further successfully used and improved in numerous works; see, e.g., [25, 38, 11, 31, 22, 40]. We recall that the proximal linearized BCD method is also known in the literature as the alternating proximal gradient descent method and the alternating forward-backward splitting method.

The advantages of this approach compared to the schemes in subsection 1.1 are tremendous. All three schemes—the BCD (3a), (3b); the proximal BCD (4a), (4b); and the prox-linearized BCD/PALM (5a), (5b)—were analyzed and compared in [39]. The conclusion is that in general the three schemes give different solutions and that prox-linearized BCD/PALM offers a larger decrease in the objective function than the other algorithms. Moreover, prox-linearized BCD/PALM needs less computation, since it involves the evaluation of the proximal operator of a single function, which in applications is usually assumed to be simple.

Remark 1.1 (choices of H in applications). In nearly all applications solved using a scheme of the form (5a), (5b), the coupling term H is quadratic, sometimes combined with a bilinear term. When restricted to two blocks, H has the form $H(x, y) = \|L_0(x \star y) - w\|^2 + \langle L_1(x), L_2(x, y)\rangle$, where "$\star$" denotes a certain product (e.g. Hadamard, convolution, direct), w is a known matrix, $\|\cdot\|$ stands for Frobenius norm, and L_i are linear forms. The latter term is used in [11] and with $L_0 = 1$ and $L_1 = L_2 = 0$ in [39, 18, 31, 40]. In all these applications, $(\nabla_x H, \nabla_y H)$ are only locally Lipschitz.

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 [7]. 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 [16, 17]. In several articles [25, 11, 31, 11] convergence is proven using the methodology proposed in [18].

1.3. Motivation and proposed ASAP algorithm.

The general idea behind the two-block alternating minimization schemes presented in subsections 1.1 and 1.2 is to generate a sequence of iterates $\{(x^k, y^k)\}_{k\in\mathbb{N}}$, where x^k is obtained by considering the partial minimization problem

$$\min_{x \in U} J(x, y^{k-1}) := F(x) + H(x, y^{k-1}), \tag{8}$$

while y^k is obtained by considering the partial minimization problem

$$\min_{y \in V} J(x^k, y) := G(y) + H(x^k, y). \tag{9}$$

[2] PALM stands for "proximal alternating linearized minimization"; see [18].

In the BCD scheme, the computed points are global minimizers of the subproblems (8) and (9). In the proximal BCD, they can be interpreted as the computation of an implicit subgradient descent (or proximal update) of (8) and (9). In this framework, considering alternating proximal linearized BCD schemes consists in applying a forward-backward splitting to each subproblem (8) and (9), namely considering a quadratic approximation of $f_1 + f_2$ at the current point (x^k or y^k, respectively). The quadratic approximation is obtained by exploiting the smoothness of one of the two functions f_1 and f_2. For instance, if f_1 is smooth, this is equivalent to an evaluation of the proximity operator of f_2 at a certain point. In the prox-linearized BCD/PALM scheme (given in (5a), (5b)), the quadratic local approximations are done on $H(\cdot, y^k)$ and $H(x^{k+1}, \cdot)$, respectively. Thus, the proximity operator is computed with respect to F at step x (and analogously, with respect to G at step y).

The other choice for a proximal linearized BCD scheme is to consider a local quadratic approximation of F and the proximity operator with respect to the partial function $H(\cdot, y^k)$ at step x and analogously, a local quadratic approximation of G and the proximity operator with respect to the partial function $H(x^k, \cdot)$ at step y. In practice this means exchanging the roles of F and $H(\cdot, y^k)$ in the x-step in (5a), (5b), and analogously, exchanging the roles of G and of $H(x^{k+1}, \cdot)$ in the y-step in (5a), (5b). Inserting these changes into (5a), (5b), and (6) gives rise to the alternating structure-adapted proximal (ASAP) gradient descent algorithm:

$$x^k \in \arg\min_{x \in U} \left\{ \left\langle x, \nabla F\left(x^{k-1}\right)\right\rangle + H(x, y^{k-1}) + \frac{1}{2\tau}\|x - x^{k-1}\|^2 \right\}, \tag{10a}$$

$$y^k \in \arg\min_{y \in V} \left\{ \left\langle y, \nabla G\left(y^{k-1}\right)\right\rangle + H(x^k, y) + \frac{1}{2\sigma}\|y - y^{k-1}\|^2 \right\}, \tag{10b}$$

along with

$$\tau \in (0, (\mathrm{Lip}(\nabla F))^{-1}) \quad \text{and} \quad \sigma \in (0, (\mathrm{Lip}(\nabla G))^{-1}). \tag{11}$$

The proposed ASAP algorithm is hence the alternative to prox-linearized BCD/PALM in (5a), (5b) when applying a proximal linearized BCD approach. Since its structure is similar to prox-linearized BCD/PALM, it shares the same algorithmic simplicity. Besides, as ASAP and prox-linearized BCD/PALM exploit the proximal linearized BCD approach from different (complementary) viewpoints, both schemes cannot be considered and analyzed within the same framework.

Let us provide some insights about the proposed ASAP algorithm.

- From (10a) and (10b), the gradients of F and G, namely ∇F and ∇G, must have uniform Lipschitz constants. This is not a paramount restriction for several reasons. Nonsmooth parts of F and G may be transferred to H as soon as the resulting partial functions $H(\cdot, y)$ and $H(x, \cdot)$ stay simple (in the sense of (7)). Further, we can assume that F and G involve smooth approximations of nonsmooth functions, which is a common approach in the design of algorithms [20, 28, 12, 19, 21] and very popular in the construction of numerical algorithms to solve some nonsmooth optimization problems in image processing [19, 27, 21, 31].
- The step sizes τ and σ in the proposed ASAP algorithm depend only on the global smoothness of ∇F and ∇G, so they can be fixed in a stable way *at initialization.* This could present an advantage compared to prox-linearized BCD/PALM when the step sizes (τ_k, σ_k) in (6) are difficult to obtain at

each iteration [40] or their values show important variations during the iterations [22].

- In many applications, the term H typically has a good structure in the sense of being simple with respect to each variable x and y in the sense of (7); hence, the proximity operators with respect to $H(\cdot, y)$ and $H(x, \cdot)$ have closed-form expressions. This is also corroborated by Remark 1.1.
- Functions F and G can have an arbitrarily complex structure in order to capture various application-dependent features, since they no longer need to be simple in the sense of (7).
- Many practical prox-linearized BCD/PALM algorithms were designed for functions H that are biconvex/multiconvex; see, e.g., [39, 38, 11, 40]. For those cases, the ASAP algorithm ensures uniqueness of the iterates (x^k, y^k) and also step sizes that are two times larger (see section 5). Its counterpart prox-linearized BCD/PALM may not benefit from these features.

Aside from these interesting features, the theoretical study of the ASAP algorithm has revealed several technical issues that were hidden in its counterpart, prox-linearized BCD/PALM. Most of them are related to subtleties of subdifferential calculus, and need to be tackled in order to establish interesting convergence results. Hence, a section of this paper is dedicated to this topic, which helps understanding of the specificity of the proposed algorithm and the assumptions made on the problem.

1.4. Outline. The paper is organized as follows: in section 2, the optimization problem we consider is introduced, along with the minimal assumptions made on it; some examples that can be encountered in signal processing are also proposed. In section 3, mathematical notions used throughout the paper are recalled. In section 4, some insights and analysis about subdifferential calculus for alternating nonconvex and nonsmooth optimization schemes are discussed. Finally, section 5 is devoted to the proposed ASAP method and its convergence properties.

2. The problem and examples.

Notation. We consider that $x \in U$ and $y \in V$, where U and V are finite-dimensional real vector spaces. The ith element of a vector or a matrix x (seen as a vector) is denoted by x_i. For an $m \times n$ real matrix w we define

$$\|w\| := \|w\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} w_{i,j}^2},$$

noticing that if w is a vector ($n = 1$), the Frobenius norm $\|\cdot\|_F$ boils down to the ℓ_2 norm. Given a nonempty set $\mathcal{S} \subset U$, the distance of a 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

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

2.1. The optimization problem. We are interested in solving nonconvex and nonsmooth minimization problems of the form $J : U \times V \to \mathbb{R} \cup \{+\infty\}$,

$$J(x, y) := F(x) + G(y) + H(x, y), \tag{12}$$

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

Assumption (H1).
(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 proper, lower semicontinuous and lower bounded.[3]

From Assumption (H1), the objective J is lower semicontinuous.

Let us introduce the following notation needed for later use:

$$\forall\ x \in U, \quad \mathcal{D}_x := \{y \in V \mid (x, y) \in \operatorname{dom} J\}, \tag{13a}$$

$$\forall\ y \in V, \quad \mathcal{D}_y := \{x \in U \mid (x, y) \in \operatorname{dom} J\}. \tag{13b}$$

These assumptions are minimal in the sense that they are sufficient to ensure the applicability of the proposed method. They will be proved to guarantee some weak convergence results (see subsection 5.2). We will also consider additional assumptions, which aim at ensuring stronger convergence results.

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 as follows:

$$J(x, y) := \underbrace{\sum_i f_i(\|A_i x\|)}_{=:F(x)} + \underbrace{\sum_j g_j(\|B_j y\|)}_{=:G(y)} + \underbrace{\sum_k h_k(b_k(x, y) - w_k) + \chi_{\mathcal{D}}(x, y)}_{=:H(x,y)}, \tag{14}$$

with $b_k : U \times V \to W$ bilinear forms, $\mathcal{D} \subset U \times V$ a closed nonempty set, A_i and B_i bounded linear operators, $w \in W$ the given data, and (f_i, g_j, h_k) real-valued functions. Functions h_k are assumed continuously differentiable and thus so is $h(x, y) = \sum_k h_k(b_k(x, y) - w_k)$. Next we provide some basic sufficient assumptions on functions (f_i, g_j, h_k), ensuring that the objective J fulfills the assumptions formulated in (H1).

Assumption (R). We consider f_i and g_j belonging to a family of functions $\psi : \mathbb{R} \to \mathbb{R}$ that have the following three properties:
(a) ψ is twice differentiable, symmetric, and strictly increasing on $\mathbb{R}_+$;
(b) $\psi'(t)/t$ goes to 0 as t approaches 0;
(c) ψ'' is bounded on $\mathbb{R}$.

LEMMA 2.1 (smoothness of F and G). *Let F and G be as in* (14)*. Assume that each f_i and g_j satisfies Assumption* (R)*. Then F and G are continuously differentiable and their gradients are Lipschitz continuous.*

The proof is in given in Appendix A. Some relevant choices for functions (f_i, g_j) satisfying the properties in Lemma 2.1 are listed below.

Example 2.2. The functions f_i, g_j, and h_k in (14) are of the form $\psi : \mathbb{R} \to \mathbb{R}$ or $(\psi)^p : \mathbb{R} \to \mathbb{R}$, where p is a rational number, defined for $t \in \mathbb{R}$ by the following (note

[3] Thus, H never goes to $-\infty$ and its domain $\operatorname{dom} H = \{(x, y) \in U \times V \mid H(x, y) < +\infty\}$ is nonempty.

that ψ may depend on a parameter $\alpha > 0$):

(i) $\psi(t) := |t|^2$ and $(\psi(t))^p$ for $p > 1/2$;

(ii) $\psi(t) := \begin{cases} |t| - \alpha/2 & \text{if } |t| > \alpha, \\ t^2/(2\,\alpha) & \text{if } |t| \leq \alpha, \end{cases}$ and $(\psi(t))^p$ for $p \in (0,1]$;

(iii) $\psi(t) := \sqrt{t^2 + \alpha}$ and $(\psi(t))^p$ for $p \in (0,1]$;

(iv) $\psi(t) := |t| - \alpha \log(1 + |t|/\alpha)$;

(v) $\psi(t) := \log(1 + t^2/\alpha)$;

(vi) $\psi(t) := t^2/(\alpha + t^2)$;

(vii) $\psi(t) := 1 - \exp(-t^2/\alpha)$.

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 be used to 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 [20, 12], whereas ψ^p for $p \in (0,1]$ are used to approximate the corresponding ℓ_p "norm" for "sparse recovery" [28, 27, 21].

Practical examples with the functions in Example 2.2. A first typical example of functions (14) that may be encountered in image processing is the smoothed total variation (TV) regularization. It can be obtained with $\{A_i\}_i$ providing the first-order differences of x at pixel i and $f_i = \psi$ with ψ as in (ii)–(iv). In particular, (iv) is used in [35, 36] for applications of the proposed method to interferometric imagery. If $f_i = (\psi)^p$ in (ii), (iii), then the resulting F is a nonconvex TV model; see [27]. In [35] and [36], respectively, the fidelity term H is defined by setting $b_k : (x,y) \mapsto x_k y_k$ and $h_k : t \mapsto t^2$ for all k, yielding $h(x,y) = \sum_k (x_k y_k - w_k)^2$.

3. Mathematical preliminaries.

3.1. Subdifferential.

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

$$\operatorname{dom} f := \{x \in U \mid f(x) < +\infty\},$$

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

DEFINITION 3.1 (subdifferential of convex functions). *Let $f : U \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, and lower semicontinuous, and $x^+ \in \operatorname{dom} f$. The* subdifferential *$\partial f(x^+)$ of f at x^+ is the set of $p \in U^*$, called* subgradients *of f at x^+, such that*

$$\forall\, x \in U, \quad f(x) \geq f(x^+) + \langle p, x - x^+ \rangle.$$

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

The subdifferential for nonconvex nonsmooth functions is defined below.

DEFINITION 3.2 (see [32, Def. 8.3]). *Let $f : U \to \mathbb{R} \cup \{+\infty\}$ be a function.*

(a) *The* Fréchet subdifferential *of f at $x^+ \in \operatorname{dom} f$, denoted by $\widehat{\partial} f(x^+)$, is the set of vectors $p \in U$ such that one has*

$$f(x) \geq f(x^+) + \langle p, x - x^+ \rangle + o(\|x - x^+\|).$$

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

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

$$\partial f(x^+) := \{p \in U \mid \exists\, x^k \to x^+,\ f(x^k) \to f(x^+),\ p^k \to p,\ p^k \in \widehat{\partial} f(x^k)\}.$$

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

It is obvious from the definition that $\widehat{\partial} f(x^+) \subset \partial f(x^+)$ for any $x^+ \in \operatorname{dom} f$. If f is convex, then for any $x^+ \in \operatorname{dom} f$ the sets $\widehat{\partial} f(x^+)$ and $\partial f(x^+)$ defined in Definition 3.2 coincide with the set $\partial f(x^+)$ defined in Definition 3.1. If f is differentiable at x^+, $\widehat{\partial} f(x^+) = \{\nabla f(x^+)\}$, and if f is continuously differentiable around x^+, then $\partial f(x^+) = \{\nabla f(x^+)\}$.

Fermat's rule, which gives a first-order necessary optimality condition, is given next.

PROPOSITION 3.3 (see [32, Thm. 10.1]). *Let $f : U \to \mathbb{R} \cup \{+\infty\}$ be a proper function. If f has a local minimum at x^*, then $0 \in \partial f(x^*)$.*

DEFINITION 3.4. *We say that x^* is a* critical point *of f if $0 \in \partial f(x^*)$.*

The set of critical points of f will be denoted by $\operatorname{crit} f$.

3.2. Proximal gradient descent. Proximity operators were inaugurated by Moreau in 1962 as a generalization of convex projection operators [30]. They were studied in numerous works; see [10] for an overview in the convex setting. The extension to nonconvex functions was investigated in [4] and used by many authors afterwards; see, e.g., [18, 31, 22, 40].

DEFINITION 3.5. *Let $h : U \to \mathbb{R} \cup \{+\infty\}$ be proper and lower semicontinuous. Given $z \in U$ and $\tau > 0$, the* proximity operator *of h at $z \in U$ is defined as*

$$\operatorname{prox}_{\tau h}(z) = \arg\min_{x \in U} \left\{ h(x) + \frac{1}{2\tau}\|z - x\|^2 \right\}. \tag{15}$$

Remark 3.6. Some important consequences of this definition are given below.

(a) For any $\tau > 0$, the set $\operatorname{prox}_{\tau h}(z)$ is nonempty and compact if h, e.g., is lower bounded [32, sect. 7].

(b) For any $z \in U$, one has $\operatorname{prox}_{\tau h}(z) \subset \operatorname{dom} h$, which follows from Fermat's rule.

(c) If h is convex, the minimizer $\operatorname{prox}_{\tau h}(z)$ is uniquely defined as being the minimizer of a strictly convex coercive function.

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

DEFINITION 3.7. *Let $h : U \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous function and $f : U \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient. The main iteration to minimize $f + h$ using the proximal gradient method starting from any $u \in U$ is given by*

$$x^+ \in \operatorname{prox}_{\tau h}(u - \tau \nabla f(u)). \tag{16}$$

The point x^+ satisfies $x^+ \in \operatorname{dom} h$; see Remark 3.6.

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

LEMMA 3.8. *Let $f : U \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient and $h : U \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous and proper function. For $\tau > 0$ fixed, consider x^+ defined by (16). Then*

$$\forall\, u \in U, \quad f(u) + h(u) \geq f(x^+) + h(x^+) + \frac{1}{2}\left(\frac{Q}{\tau} - \mathrm{L}_{\nabla f}\right)\|x^+ - u\|^2, \tag{17}$$

where $Q = 2$ if h is convex and $Q = 1$ otherwise.

Proof. The proof for $Q = 1$ can be found in [18, Lem. 2], while the case when $Q = 2$ is considered in [18, Rem. 4(iii)]. ∎

3.3. The Kurdyka–Łojasiewicz (KŁ) property.

The Kurdyka–Łojasiewicz (KŁ) property was initially introduced to analyze the behavior of smooth optimization algorithms. In a nonsmooth context, it was studied originally in [16] and later in [4, 6, 7].

DEFINITION 3.9 (KŁ property). *Let $f : U \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. The function f is said to have* the Kurdyka–Łojasiewicz 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*

$$\forall\, x \in \widetilde{\mathcal{O}}(x^*), \quad \kappa \ \operatorname{dist}(0, \partial f(x)) \geq |f(x) - f(x^*)|^{\theta}, \tag{18}$$

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

$$\widetilde{\mathcal{O}}(x^*) := \mathcal{O}(x^*) \cap \{x \in U \mid f(x^*) < f(x) < f(x^*) + \eta\}. \tag{19}$$

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.

Usual examples of functions which satisfy the KŁ property are *semialgebraic* and *real-analytic* functions. 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 [29, p. 92]. However, objective functions encountered in practical applications are usually built from such functions, but are neither semialgebraic nor real-analytic, since the union of these two classes of functions is not stable by sum nor composition. On the other hand, checking that a given function satisfies the KŁ property may be a hard task. A way to overcome this difficulty while benefiting from the KŁ properties of *semialgebraic* and *real-analytic* functions is to consider the larger class of *subanalytic functions* [15]. Indeed, this class is partially stable by sum and composition, with the restrictions given in the following proposition.

PROPOSITION 3.10. *Let f and g be two subanalytic functions. Then the following results hold:*

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

(b) *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* [23, Prop. 2.46].

The interest of the class of subanalytic functions is that it contains both semialgebraic functions and real-analytic functions (see [15]). Moreover, one has the following.

THEOREM 3.11 (see [16, Thm. 3.1]). *Let $f : U \to \mathbb{R} \cup \{+\infty\}$ be a function that is subanalytic with a closed domain and continuous on its domain. Let x^* be a critical point of f. Then f has the KŁ property at x^*.*

Together with Proposition 3.10, this proposition proves that most of the objectives built by summing and composing semialgebraic and real-analytic functions satisfy the KŁ property at their critical points. Moreover, according to Definition 3.9, it is obvious that if a function $\widetilde{J}$ satisfies the KŁ property at a point belonging to some set $\mathcal{S}$, then so does $J = \widetilde{J} + \chi_{\mathcal{S}}$. In particular, this proves that the functions proposed in subsection 2.2 satisfy the KŁ property at their critical points.

4. Subdifferential calculus for alternating schemes.

Given $H : U \times V \to \mathbb{R} \cup \{+\infty\}$, its subdifferential at (x^+, y^+) is denoted by $\partial H(x^+, y^+)$. In addition, $\partial_x H(x^+, y^+)$ and $\partial_y H(x^+, y^+)$ are the subdifferentials of $x \mapsto J(x, y^+)$ at y^+ and $y \mapsto J(x^+, y)$ at x^+, respectively.

4.1. Partial subdifferentials.

4.1. Partial subdifferentials. The major interest of alternating schemes is when each step can be computed using an explicit formula or via a low-cost scheme. In first-order alternating methods, each iteration is equivalent to finding a minimizer of two auxiliary functions, which are characterized by Fermat's rule. Since Fermat's rule is defined on subgradients, applying alternating schemes in this framework leads us to consider *partial* subgradients at each computation step, whereas the desired limit point (when it can be reached) is characterized by its *whole* subgradient, namely $(0,0) \in \partial J(x^*, y^*)$, since the set of critical points of J contains its minimizers (when they exist). Hence, a crucial point is the link between the *partial subdifferentials* and the *whole subdifferential* at the same point. In most cases, a critical point (x^*, y^*) can be found using alternating schemes only if the objective function J satisfies

$$p \in \partial_x J(x,y) \times \partial_y J(x,y) \quad \Longrightarrow \quad p \in \partial J(x,y) \tag{20}$$

at least for $p = (0,0)$ (but we will see that this property is generally needed for larger sets of points, namely the whole domain of the objective J). As shown later, in Example 4.3, this implication holds true when J is differentiable or additively separable or is the sum of such functions. In other cases (in particular for nonseparable nonsmooth functions), (20) is generally not satisfied. More specifically, a point $(\bar{x}, \bar{y})$ satisfying $0 \in \partial_x J(\bar{x}, \bar{y})$ and $0 \in \partial_y J(\bar{x}, \bar{y})$ can generally yield $(0,0) \notin \partial J(\bar{x}, \bar{y})$, so it is not a critical point of J. Let us give some examples.

The coupling function H in the example below is taken from [33, p. 12].

Example 4.1 (rotated scaled ℓ_1 with regularization). Let $J : \mathbb{R}^2 \to \mathbb{R}$ be defined for $\beta \geq 0$ by

$$J(x,y) := \underbrace{|x+y| + 2|x-y|}_{=:H(x,y)} + \beta(x-y)^2. \tag{21}$$

The function $H : (x,y) \mapsto |x+y| + 2|x-y|$ is the $\pi/4$ radian counterclockwise-rotated version of the separable function $f : (x,y) \mapsto \sqrt{2}(|x| + 2|y|)$. The functions H and J are convex and nonsmooth. Hence, their limiting subdifferentials coincide with their subdifferentials as defined in Definition 3.1 for convex functions. For any $\beta \geq 0$, the point $(0,0)$ is the unique global minimizer of J, which, from Fermat's rule, is characterized[4] by

$$(0,0) \in \partial J(x^*, y^*) \quad \Longleftrightarrow \quad (x^*, y^*) = (0,0).$$

In particular, this implies that, for any $u \neq 0$, one has $(0,0) \notin \partial J(u,u)$. However, for any $u > 0$, one has $\partial_x J(u,u) = \partial_y J(u,u) = [-1, 3]$, and for any $u < 0$, one has $\partial_x J(u,u) = \partial_y J(u,u) = [-3, 1]$. Thus, for any $u \neq 0$,

$$(0,0) \in \partial_x J(u,u) \times \partial_y J(u,u) \quad \text{and} \quad (0,0) \notin \partial J(u,u). \tag{22}$$

A consequence of the result above is that, for any $u \neq 0$,

$$J(u,u) = \min_{y \in \mathbb{R}} J(u,y) = \min_{x \in \mathbb{R}} J(x,u),$$

that is, u is separately a global minimizer of $J(u,\cdot)$ and of $J(\cdot,u)$. Therefore, starting from any initial point $(x^0, y^0) \neq (0,0)$ the alternating minimization scheme will move

[4] For strictly convex functions, Fermat's rule is a sufficient and necessary optimality condition.

to the point (y^0, y^0) (minimization along x) or to the point (x^0, x^0) (minimization along y) and will stop because there is no further decrease in the objective J along x or y. Thus, the block-coordinate descent converges to noncritical points of the form (u, u). Similar behaviors may be observed for proximal gradient descent schemes.

It is noteworthy that the failure of (20) in Example 4.1 comes from the nonseparability of the subdifferential ∂H entailed by the nonsmoothness of H.

Below we give two more examples with simple nonseparable constraints illustrating the possibility of ensuring (20) at critical points.

Example 4.2 (block nonseparable constraint). Let us consider two different examples, which show that the nonseparability of the constraint may lead to different situations regarding the link between the subdifferential and the partial subdifferential. The proofs are given in Appendix A.

(a) Let $J = \chi_{\mathcal{B}}$ with $\mathcal{B} := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$. Its subdifferential is given by

$$\forall\, (x, y) \in \mathcal{B}, \quad \partial J(x, y) = \begin{cases} \{(0, 0)\} & \text{if } x^2 + y^2 < 1, \\ \{(\lambda\, x, \lambda\, y) \mid \lambda \geq 0\} & \text{if } x^2 + y^2 = 1, \end{cases}$$

while its partial subdifferentials are

$$\forall\, (x, y) \in \mathcal{B},$$
$$\partial_x J(x, y) \times \partial_y J(x, y) = \begin{cases} \{(0, 0)\} & \text{if } x^2 + y^2 < 1, \\ \{(\lambda_x\, x, \lambda_y\, y) \mid \lambda_x, \lambda_y \geq 0\} & \text{if } x^2 + y^2 = 1. \end{cases}$$

As a consequence, $\partial J(x, y) \subsetneq \partial_x J(x, y) \times \partial_y J(x, y)$ for any $x^2 + y^2 = 1$. Thus, (20) does not hold for such points.

(b) Let us now consider

$$J(x, y) = \begin{cases} 0 & \text{if } (x, y) \in ([0, 1] \times [0, 1]) \cup ([1, 2] \times [0, 2]), \\ +\infty & \text{otherwise.} \end{cases}$$

Hence, J is the indicator function of the union of two rectangles $\mathcal{D}_1 := [0, 2] \times [0, 1]$ and $\mathcal{D}_2 := [1, 2] \times [0, 2]$. Then, (20) holds for any $(x, y) \in \operatorname{dom} J$.

Note that in both (a) and (b) the function J is biconvex. Hence, biconvexity does not ensure the separability-like condition (20).

Condition (20), even though it is fundamental in alternating schemes, is seldom explicitly required in them. Indeed, it is often automatically satisfied in such algorithms, where H is assumed to be differentiable, as shown in the following example, which gives a sufficient condition for a function J to satisfy (20).

Example 4.3. Let $J : U \times V \to \mathbb{R} \cup \{+\infty\}$ be of the form

$$J(x, y) = h(x, y) + f(x) + g(y),$$

where $h : U \times V \to \mathbb{R}$ is a continuously differentiable function and $f : U \to \mathbb{R} \cup \{+\infty\}$ and $g : V \to \mathbb{R} \cup \{+\infty\}$ are two lower semicontinuous functions. Then, for any $(x, y) \in \operatorname{dom} J = \operatorname{dom} f \times \operatorname{dom} g$,

$$\partial J(x, y) = \nabla h(x, y) + \partial (f + g)(x, y) = (\nabla_x h(x, y), \nabla_y h(x, y)) + \partial (f + g)(x, y).$$

Using subdifferential calculus for separable functions [32, proof of Prop. 10.6] yields

$$\partial(f+g)(x,y) = \partial f(x) \times \partial g(y).$$

These equalities show that J satisfies (20).

4.2. Parametric closedness of the partial subdifferentials.

Besides condition (20), another property on the subdifferential of J, closely linked to the *closedness of the subdifferential*, is crucial.

Definition 4.4 (parametric closedness of the partial subdifferentials). *Let* $J : U \times V \to \mathbb{R} \cup \{+\infty\}$ *be a function,* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$*, and* $(x^+, y^+) \in \operatorname{dom} J$ *such that*

$$(x^k, y^k) \xrightarrow[k\to\infty]{} (x^+, y^+) \quad \textit{and} \quad J(x^k, y^k) \xrightarrow[k\to\infty]{} J(x^+, y^+). \tag{23}$$

The x-partial subdifferential of J is said to be parametrically closed at (x^+, y^+) with respect to the sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *if, for any sequence* $\{p_x^k\}_{k\in\mathbb{N}}$*,*

$$\left(p_x^k \xrightarrow[k\to\infty]{} p_x^+, \quad p_x^k \in \partial_x J(x^k, y^k)\right) \quad \Longrightarrow \quad p_x^+ \in \partial_x J(x^+, y^+), \tag{24}$$

while the y-partial subdifferential of J is said to be parametrically closed at (x^+, y^+) with respect to the sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *if, for any sequence* $\{q_y^k\}_{k\in\mathbb{N}}$*,*

$$\left(q_y^k \xrightarrow[k\to\infty]{} q_y^+, \quad q_y^k \in \partial_y J(x^k, y^k)\right) \quad \Longrightarrow \quad q_y^+ \in \partial_y J(x^+, y^+). \tag{25}$$

The x-partial and y-partial subdifferentials of J are said to be parametrically closed *if they are parametrically closed at any point* $(x^+, y^+) \in \operatorname{dom} J$ *with respect to any sequence satisfying* (23).

This notion is crucial for alternating schemes as soon as one lacks information about the whole subdifferential $\partial J(x^k, y^k)$ at the current point (x^k, y^k). For the class of methods presented in the introduction of this paper, this is typically the case as, in general, one can only derive an x-subgradient in $\partial_x J(x^k, y^{k-1})$ and a y-subgradient in $\partial_y J(x^k, y^k)$. Unlike the closedness of the subdifferential, the parametric closedness of the partial subdifferentials may not hold in general.

Example 4.5. Let us consider the following function:

$$\forall\,(x,y) \in \mathbb{R}^2, \quad J(x,y) = \begin{cases} \sqrt{x}\sqrt{y} & \text{if } x, y \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

J is obviously a continuous function on its domain. Moreover, for any $y > 0$, the partial function $x \mapsto J(x,y)$ is continuously differentiable on $(0, +\infty)$, while $x \mapsto J(x, 0)$ is identically null. Let $\{x^k\}_{k\in\mathbb{N}}$ and $\{y^k\}_{k\in\mathbb{N}}$ be two sequences that converge to zero ($x^+ = y^+ = 0$) with $x^k = y^{k-1} > 0$ for any $k \in \mathbb{N}^*$. Then, thanks to the smoothness of $x \mapsto J(x, y^{k-1})$ around $x^k > 0$, one has

$$\partial_x J(x^k, y^{k-1}) = \{\nabla_x J(x^k, y^{k-1})\} = \left\{\frac{\sqrt{y^{k-1}}}{2\sqrt{x^k}}\right\} = \left\{\frac{1}{2}\right\}$$

with $\lim_{k\to\infty} 1/2 = 1/2$. Besides, since $x \mapsto J(x, y^+) = \chi_{\mathbb{R}^+}$, which is convex, one has

$$\partial_x J(x^+, y^+) = (-\infty, 0].$$

As a consequence, $\partial_x J(x^+, y^+) \not\ni 1/2$. In other words, $\partial_x J$ is not parametrically closed with respect to $\{(y^{k-1}, y^k)\}_{k\in\mathbb{N}}$ for any sequence $\{y^k\}_{k\in\mathbb{N}}$ that converges to 0.

Remark 4.6. It is noteworthy that the parametric closedness of the partial subdifferential is quite independent of the separability of the subdifferential as defined by (20). Indeed, the objective J in the example above does obey (20) for any $x, y \geq 0$ but fails to satisfy Definition 4.4.

Let us give a general example of functions that satisfy the parametric closedness of the partial subdifferentials.

Example 4.7. Let $J : U \times V \to \mathbb{R} \cup \{+\infty\}$ be a biconvex function continuous on its domain. Then its partial subdifferentials are parametrically closed at any point $(x^+, y^+) \in \operatorname{dom} J$. Indeed, let $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ and $\{(\widetilde{x}^k, \widetilde{y}^k)\}_{k\in\mathbb{N}}$ be two sequences of $\operatorname{dom} J$ that converge to (x^+, y^+) and satisfy (24) and (25) for sequences $\{p_x^k\}_{k\in\mathbb{N}}$ and $\{\widetilde{q}_y^k\}_{k\in\mathbb{N}}$, respectively. Let $k \in \mathbb{N}$. Since $x \mapsto J(x, y^k)$ and $y \mapsto J(\widetilde{x}^k, y)$ are convex, lower semicontinuous, and proper functions, the subgradient inequality (Definition 3.1) can be applied, which leads to the following:

$$\begin{aligned} &\forall\, x \in U, \quad J(x, y^k) \geq J(x^k, y^k) + \langle p_x^k, x - x^k\rangle,\\ &\forall\, y \in V, \quad J(\widetilde{x}^k, y) \geq J(\widetilde{x}^k, \widetilde{y}^k) + \langle \widetilde{q}_y^k, y - \widetilde{y}^k\rangle. \end{aligned}$$

We can evaluate the limit in the inequalities above when $k \to +\infty$:

$$\begin{aligned} &\forall\, x \in U, \quad J(x, y^+) \geq J(x^+, y^+) + \langle p_x^+, x - x^+\rangle,\\ &\forall\, y \in V, \quad J(x^+, y) \geq J(x^+, y^+) + \langle q_y^+, y - y^+\rangle. \end{aligned}$$

Then, by the definition of the partial subgradients of the biconvex function J, this implies that $p_x^+ \in \partial_x J(x^+, y^+)$ and $q_y^+ \in \partial_y J(x^+, y^+)$.

Remark 4.8. Although the parametric closedness of the partial subdifferentials is a crucial property for alternating schemes, this notion has not been investigated in detail in this context since, in most cases, the nonseparable term H is assumed to be sufficiently smooth (typically, with Lipschitz-continuous gradient). Indeed, in this case, from the knowledge of an x-subgradient of J at (x^k, y^{k-1}) one can derive knowledge of an x-subgradient of J at (x^k, y^k). Hence, one gets a whole subgradient at the current point (x^k, y^k) thanks to Example 4.3. Thus, when the parametric closedness of the partial subdifferentials cannot be ensured, but the problem is sufficiently smooth, one can use the closedness of the subdifferential instead (see Assumption (H2)(c)(ii) and Lemma 5.8).

5. Proposed method.

Many algorithms in the prox-linearized BCD/PALM family were designed for coupling functions H that are convex with respect to some of the blocks; see, e.g., [39, 38, 11, 40]. For such coupling terms, our ASAP algorithm can provide step sizes that are two times larger for the blocks in which H is convex. To this end, for each block $z_{(i)}$ we introduce a constant $Q_{z_{(i)}}$ defined for the N-block case by

$$Q_{z_{(i)}} = \begin{cases} 2 & \text{if } z_{(i)} \mapsto H(z_{(1)}, \ldots, z_{(i-1)}, z_{(i)}, z_{(i+1)}, \ldots, z_{(N)}) \text{ is convex,} \\ 1 & \text{otherwise.} \end{cases} \tag{26}$$

For the two-block case, we define $Q_X = Q_{z_{(1)}}$ and $Q_Y = Q_{z_{(2)}}$.

5.1. Alternating structure-adapted proximal (ASAP) gradient descent algorithm.

Using the definition of proximal gradient descent (Definition 3.7), our algorithm sketched in subsection 1.3 (see (10a), (10b)) takes the compact form stated in Algorithm 1.

Algorithm 1 ASAP: alternating structure-adapted proximal gradient descent.

```
Initialization: $(x^0, y^0) \in \operatorname{dom} J$ and $0 < \tau < Q_X/\mathrm{L}_{\nabla F}$,
$0 < \sigma < Q_Y/\mathrm{L}_{\nabla G}$;
General step: for $k = 1, 2, \ldots$, compute
    $x^k \in \operatorname{prox}_{\tau H(\cdot, y^{k-1})}(x^{k-1} - \tau \nabla F(x^{k-1}))$,
    $y^k \in \operatorname{prox}_{\sigma H(x^k, \cdot)}(y^{k-1} - \sigma \nabla G(y^{k-1}))$.
```

In Algorithm 1, Q_X and Q_Y are set according to (26). The step sizes τ and σ are set according to Lemma 3.8 to ensure a sufficient decrease in the objective J at each step (see subsection 5.2). They depend only on the Lipschitz constants of ∇F and ∇G and on (Q_X, Q_Y). In the case when H is (partially) convex, one can always set the pessimistic (general) rule $Q_X = Q_Y = 1$; we will see in Corollary 5.3 that this would entail a slower convergence of the value of J.

Remark 5.1. The ASAP algorithm confirms immediately two attractive points that are direct consequences of Remark 3.6:

(a) the iterates $(x^k, y^k) \in \operatorname{dom} J$ are well defined and one also has $(x^k, y^{k-1}) \in \operatorname{dom} J$ (in particular, for any $k \in \mathbb{N}$, the functions $x \mapsto J(x, y^{k-1})$ and $y \mapsto J(x^k, y)$ are proper);

(b) when H is biconvex, iterates (x^k, y^k) are uniquely defined (even if F or G are nonconvex), and step sizes can be chosen to be two times larger than in the nonbiconvex case.

5.2. Basic convergence facts. A sufficient condition to have well-defined iterations is to choose the initial point (x^0, y^0) in $\operatorname{dom} J$. In this paper, we assume that such an initialization can be done. There may be some cases where this feasibility problem cannot be solved, but this is beyond the scope of this paper.

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.

Based on the step-size initialization of the algorithm, we set

$$\rho := \frac{1}{2} \min\left\{ \frac{Q_X}{\tau} - \mathrm{L}_{\nabla F}, \frac{Q_Y}{\sigma} - \mathrm{L}_{\nabla G} \right\} > 0, \tag{27}$$

where Q_X and Q_Y are defined as in (26).

PROPOSITION 5.2. *Let Assumption* (H1) *hold and let* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by ASAP.*

(a) *For every* $k \geq 1$*, the following sufficient decrease properties hold:*

$$\begin{aligned} J(x^{k-1}, y^{k-1}) &\geq J(x^k, y^{k-1}) + \rho \, \|x^k - x^{k-1}\|^2, \\ J(x^k, y^{k-1}) &\geq J(x^k, y^k) + \rho \, \|y^k - y^{k-1}\|^2, \\ J(x^{k-1}, y^{k-1}) &\geq J(x^k, y^k) + \rho(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2). \end{aligned} \tag{28}$$

Hence,

$$J(x^{k-1}, y^{k-1}) \geq J(x^k, y^{k-1}) \geq J(x^k, y^k). \tag{29}$$

(b) *The sequences* $\{J(x^k,y^k)\}_{k\in\mathbb{N}}$ *and* $\{J(x^k,y^{k-1})\}_{k\in\mathbb{N}^*}$ *converge to the same finite value, denoted by* J^*.

(c) *We have*

$$\sum_{k=1}^{+\infty}\left(\|y^k-y^{k-1}\|^2+\|x^k-x^{k-1}\|^2\right)<+\infty,$$

and hence

$$\lim_{k\to\infty}\|x^{k+1}-x^k\|=0 \quad \textit{and} \quad \lim_{k\to\infty}\|y^{k+1}-y^k\|=0.$$

Proof. According to Remark 5.1(a), one can apply Lemma 3.8 with $f:=F$, $h:=H(\cdot,y^{k-1})$, $x^+:=x^k$, and $x:=x^{k-1}$, which, along with the definition of ρ in (27), shows that

$$\begin{aligned}J(x^{k-1},y^{k-1})&=F(x^{k-1})+G(y^{k-1})+H(x^{k-1},y^{k-1})\\&\geq F(x^k)+G(y^{k-1})+H(x^k,y^{k-1})+\rho\,\|x^k-x^{k-1}\|^2\\&=J(x^k,y^{k-1})+\rho\,\|x^k-x^{k-1}\|^2.\end{aligned}\tag{30}$$

Lemma 3.8 with $f:=G$, $h:=H(x^k,\cdot)$, $x^+:=y^k$, and $x:=y^{k-1}$, and (27) yield

$$\begin{aligned}J(x^k,y^{k-1})&=F(x^k)+G(y^{k-1})+H(x^k,y^{k-1})\\&\geq F(x^k)+G(y^k)+H(x^k,y^k)+\rho\,\|y^k-y^{k-1}\|^2\\&=J(x^k,y^k)+\rho\,\|y^k-y^{k-1}\|^2.\end{aligned}\tag{31}$$

Bringing (30) and (31) together leads to

$$\begin{aligned}J(x^{k-1},y^{k-1})&\geq J(x^k,y^{k-1})+\rho\,\|x^k-x^{k-1}\|^2\\&\geq J(x^k,y^k)+\rho\,\|x^k-x^{k-1}\|^2+\rho\,\|y^k-y^{k-1}\|^2,\end{aligned}\tag{32}$$

which completes the proof of (a). It follows that the sequences $\{J(x^k,y^k)\}_{k\in\mathbb{N}}$ and $\{J(x^k,y^{k-1})\}_{k\in\mathbb{N}^*}$ are nonincreasing and interlaced by (29), 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 (32) we also have

$$\forall\, k\geq 1,\quad \|x^k-x^{k-1}\|^2+\|y^k-y^{k-1}\|^2\leq\frac{1}{\rho}\left(J(x^{k-1},y^{k-1})-J(x^k,y^k)\right).\tag{33}$$

For $K\geq 1$, summing (33) from $k=1$ to K and using statement (b) yields

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

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

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

which establishes statement (c). □

From the proof of Proposition 5.2(c), we can derive a global $O(1/k)$ convergence rate for $\{\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2\}_{k\in\mathbb{N}^*}$.

Corollary 5.3 (convergence rate). *Let Assumption* (H1) *hold. Further, let* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by ASAP. Then, for any* $K \geq 1$*, it holds that*

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

This corollary shows that the smaller ρ, the higher the convergence rate. From the definition of ρ in (27), one can see that large step sizes (τ, σ) lead to better convergence rates in terms of the value of the objective function.

The following result, together with Proposition 5.2(c), shows that ASAP generates two sequences of partial subgradients of J that vanish when k goes to infinity.

Proposition 5.4. *Let Assumption* (H1) *hold and let* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by ASAP. For any integer* $k \geq 1$*, define*

$$p_x^k := \nabla F(x^k) - \nabla F(x^{k-1}) + \frac{1}{\tau}(x^{k-1} - x^k), \tag{34a}$$

$$q_y^k := \nabla G(y^k) - \nabla G(y^{k-1}) + \frac{1}{\sigma}(y^{k-1} - y^k). \tag{34b}$$

Then $p_x^k \in \partial_x J(x^k, y^{k-1})$ *and* $q_y^k \in \partial_y J(x^k, y^k)$*, and furthermore they obey*

$$\|p_x^k\| \leq \left(\mathrm{L}_{\nabla F} + \frac{1}{\tau}\right)\|x^{k-1} - x^k\| \quad \textit{and} \quad \|q_y^k\| \leq \left(\mathrm{L}_{\nabla G} + \frac{1}{\sigma}\right)\|y^{k-1} - y^k\|. \tag{35}$$

Proof. Fermat's rule for x^k in (10a) yields

$$\frac{1}{\tau}(x^{k-1} - x^k) \in \nabla F(x^{k-1}) + \partial_x H(x^k, y^{k-1}). \tag{36}$$

From the expression for J given in (12) one has

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

Inserting this expression into (36) shows that p_x^k in (34a) obeys $p_x^k \in \partial_x J(x^k, y^{k-1})$. In a similar way, one can prove that q_y^k in (34b) satisfies $q_y^k \in \partial_y J(x^k, y^k)$.

Using that ∇F is Lipschitz continuous of constant $\mathrm{L}_{\nabla F}$ (see Assumption (H1)(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. Finally, the bounds in (35) follow from Proposition 5.2(c). □

5.3. The limit point set of ASAP. According to the discussion in section 4, we now specialize the assumptions on the coupling term H in the main model (12) in order to establish stronger convergence results.

Assumption (H2).

(a) The subdifferential of H obeys

$$\forall\,(x, y) \in \operatorname{dom} H, \quad \partial_x H(x, y) \times \partial_y H(x, y) \subset \partial H(x, y).$$

(b) The domain of H is closed and we set $\mathcal{D} := \operatorname{dom} H \subset U \times V$.

(c) Moreover, H satisfies one of the following conditions.

(i) H is continuous on its domain and biconvex.

(ii) $H : U \times V \to \mathbb{R} \cup \{+\infty\}$ can be cast in the form

$$H(x, y) = h(x, y) + f(x), \tag{37}$$

where $f : U \to \mathbb{R} \cup \{+\infty\}$ is continuous on its domain; $h : U \times V \to \mathbb{R} \cup \{+\infty\}$ is a continuous function on $\operatorname{dom} H$ such that for any $y \in V$ obeying $\mathcal{D}_y \neq \varnothing$ the partial function $h(\cdot, y)$ is differentiable on U of gradient $\nabla_x h$, and at least one of the following holds:

(A) for each bounded subset $\mathcal{B}_U \times \mathcal{B}_V \subset \operatorname{dom} H$, there exists $\xi > 0$ such that, for any $x' \in \mathcal{B}_U$ and any $(y, y') \in \mathcal{B}_V^2$,

$$\|\nabla_x h(x', y) - \nabla_x h(x', y')\| \leq \xi \, \|y - y'\|; \tag{38}$$

(B) $\nabla_x h$ is continuous on $\operatorname{dom} H$.

Remark 5.5. From Assumptions (H1) and (H2), the function J is continuous on its domain $\operatorname{dom} J = \operatorname{dom} H$, which is closed and nonempty. In other words, for any sequence $\{z^k\}_{k\in\mathbb{N}}$ in $\operatorname{dom} J$ converging to $z^* \in \operatorname{dom} J$, one has $J(z^k) \xrightarrow[k\to+\infty]{} J(z^*)$.

Let us present some comments on Assumption (H2).

Remark 5.6.

(a) Condition (H2)(a) is a generic assumption for the convergence of alternating schemes; see subsection 4.1.

(b) Since F and G are continuously differentiable, Assumption (H2)(a) is equivalent to having $\partial_x J(x, y) \times \partial_y J(x, y) \subset \partial J(x, y)$ for any $(x, y) \in \operatorname{dom} J$.

(c) Conditions (H2)(c) are motivated by the discussion in subsection 4.2 (especially Remark 4.8). More specifically, they aim at guaranteeing that the limit points reached by the algorithm are critical points of the objective J (see Proposition 5.9). However, they are sufficient but not necessary, and the results presented in this paper can be established for a broader class of functions H.

(d) If H is assumed biconvex, then no differentiability is necessary at this stage.

(e) Note that, for the nonbiconvex case in Assumption (H2)(c)(ii), the regularity assumptions on H are not symmetric in x and y. In particular, $h(x, \cdot)$ does not need to be differentiable.

(f) Assumption (H2)(c)(ii)(A) holds for various h that are nonsmooth with respect to the y-term (see Example 5.7). It can also be derived if h is twice continuously differentiable or if ∇h is Lipschitz continuous on bounded sets.

Indeed, let us consider some examples.

Example 5.7.

- Let $h(x, y) = x^2\,|y|/2$ for any $(x, y) \in \mathbb{R}^2$. Then one has $\nabla_x h(x, y) = x\,|y|$, where $y \mapsto x\,|y|$ is (globally) Lipschitz continuous for any $x \in \mathbb{R}$.
- Let $h(x, y) = (xy)^2\,|y|/2$ for any $(x, y) \in \mathbb{R}^2$. Then $\nabla_x h(x, y) = y^2|y|x$, where $y \mapsto |y|^3 x$ is locally Lipschitz continuous for any $x \in \mathbb{R}$.

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 $\mathcal{L}(x^0, y^0)$; it reads as follows:

$$\mathcal{L}(x^0, y^0) := \left\{ (x^*, y^*) \;\middle|\; \exists\, \{k_j\}_{j\in\mathbb{N}} \text{ strictly increasing s.t. } (x^{k_j}, y^{k_j}) \xrightarrow[j\to+\infty]{} (x^*, y^*) \right\}.$$

This set is nonempty as soon as the sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ is assumed to be *bounded*. Thus, from now on, we will make this assumption. The boundedness of the iterates is a common assumption; see, e.g., [6, 39, 18, 25, 11]. This assumption holds, for instance, when the level sets of J are bounded, i.e., when J is coercive, or when the domain of J is bounded.

Before stating the main result of this subsection, let us prove the following lemma for the nonbiconvex case in Assumption (H2)(c)(ii).

LEMMA 5.8. *Let Assumptions* (H1) *and* (H2)(c)(ii) *hold and let* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by ASAP, which is assumed to be bounded. For any* $k \in \mathbb{N}^*$*, set*

$$q_x^k := p_x^k + \nabla_x h(x^k, y^k) - \nabla_x h(x^k, y^{k-1})$$

with p_x^k *defined in Proposition* 5.4*. Then, for any* $k \in \mathbb{N}$*, one has* $q_x^k \in \partial_x J(x^k, y^k)$*. Moreover, for any convergent subsequence* $\{(x^{k_j}, y^{k_j})\}_{j\in\mathbb{N}}$*,* $\{q_x^{k_j}\}_{j\in\mathbb{N}}$ *converges to* 0*.*

Proof. Let $k \in \mathbb{N}^*$. From (12), (37), and the smoothness assumptions made on J, one has

$$\partial_x J(x^k, y^k) = \partial_x J(x^k, y^{k-1}) + \nabla_x h(x^k, y^k) - \nabla_x h(x^k, y^{k-1}),$$

which proves in particular that

$$q_x^k \in \partial_x J(x^k, y^k).$$

Since $\{p_x^k\}_{k\in\mathbb{N}^*}$ goes to 0 according to Proposition 5.4, it follows that $\{q_x^{k_j}\}_{j\in\mathbb{N}}$ goes to 0 as soon as $\{\nabla_x h(x^{k_j}, y^{k_j}) - \nabla_x h(x^{k_j}, y^{k_j-1})\}_{j\in\mathbb{N}}$ converges to 0. Given $\{(x^{k_j}, y^{k_j})\}_{j\in\mathbb{N}}$ is a convergent sequence and, by Proposition 5.2((c)), $\{(y^{k_j-1} - y^{k_j})\}_{j\in\mathbb{N}}$ converges to 0, it follows that $\{(x^{k_j}, y^{k_j-1})\}_{j\in\mathbb{N}}$ has the same limit as $\{(x^{k_j}, y^{k_j})\}_{j\in\mathbb{N}}$. Thus, $\{\nabla_x h(x^{k_j}, y^{k_j}) - \nabla_x h(x^{k_j}, y^{k_j-1})\}_{j\in\mathbb{N}}$ converges to 0 as soon as h satisfies Assumption (H2)(c)(ii)(A) (thanks to the local Lipschitz continuity of $\nabla_x h$) or Assumption (H2)(c)(ii)(B) (thanks to the global continuity of $\nabla_x h$). □

The next proposition states that, under Assumptions (H1) and (H2), all limit points of a sequence generated by ASAP are critical points of J.

PROPOSITION 5.9. *Let Assumptions* (H1) *and* (H2) *hold and let* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by ASAP, which is assumed to be bounded. Let* $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$*. Then the following hold:*

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

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

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

Proof. (a) This follows from the definition of $\mathcal{L}(x^0, y^0)$.

(b) From Remark 5.1(a), (x^{k_j}, y^{k_j}) belongs to $\operatorname{dom} J$, which is closed, by Assumption (H2)(b). Hence, $(x^*, y^*) \in \operatorname{dom} J$, and by continuity (see Remark 5.5) one has

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

Since $(J(x^k, y^k))_{k\in\mathbb{N}}$ is a convergent sequence by Proposition 5.2(b), this proves assertion (b).

(c) For the case in Assumption (H2)(c)(ii): Lemma 5.8, along with Proposition 5.4 and Assumption (H2)(a), ensures that there exists a sequence of subgradient

$(q_x^{k_j}, q_y^{k_j}) \in \partial J(x^{k_j}, y^{k_j})$ which converges to 0 as j goes to infinity. Then, one can then use the closedness of the subdifferential,[5] which guarantees that

$$(q_x^{k_j}, q_y^{k_j}) \in \partial J(x^{k_j}, y^{k_j}) \xrightarrow[j\to+\infty]{} (0,0) \quad \Longrightarrow \quad (0,0) \in \partial J(x^*, y^*).$$

For the case in Assumption (H2)(c)(i): according to Proposition 5.4 and the continuity of J on its domain, it suffices to establish that $\partial_x J$ is parametrically closed at (x^*, y^*) with respect to $\{(x^k, y^{k-1})\}_{k\in\mathbb{N}}$ and that $\partial_y J$ is parametrically closed at (x^*, y^*) with respect to $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ (see Definition 4.4). To this end, one can use the parametric closedness of the partial subdifferentials as proved in Example 4.7. This implies that $0 \in \partial_x J(x^*, y^*)$ and $0 \in \partial_y J(x^*, y^*)$, which, thanks to Assumption (H2)(a), ensures that $(0,0) \in \partial J(x^*, y^*)$. ∎

The following lemma is proved in Appendix C.

LEMMA 5.10. *Let $\{z^k\}_{k\in\mathbb{N}}$ be a bounded sequence of $U \times V$. Then*

$$\lim_{k\to+\infty} \operatorname{dist}(z^k, \mathcal{L}(z^0)) = 0.$$

Below, we establish two other facts on the limit point set of ASAP.

PROPOSITION 5.11. *Let Assumptions* (H1) *and* (H2) *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) $\mathcal{L}(x^0, y^0) \subset \operatorname{crit}(J)$,

(b) $\lim_{k\to+\infty} \operatorname{dist}((x^k, y^k), \operatorname{crit} J) = 0$.

Proof. Part (a) follows from Proposition 5.9(c).

Part (b) is a consequence of the fact that

$$\operatorname{dist}((x^k, y^k), \operatorname{crit} J) \leq \operatorname{dist}((x^k, y^k), \mathcal{L}(x^0, y^0))$$

and of Lemma 5.10. ∎

According to Proposition 5.11(a), there may be critical points of J that cannot be reached when ASAP starts from a given initial (x^0, y^0). This may emphasize the role of the initial estimate. Statement (b) guarantees that, for k large enough, the iterates (x^k, y^k) are arbitrarily close to a critical point of J. We conclude this subsection with the following remark.

Remark 5.12. Proposition 5.9 ensures only subsequential convergence of ASAP to critical points of J. In practice, this result might be sufficient thanks to Proposition 5.11. 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 $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ toward a critical point of J.

5.4. Subgradient convergence.

A subgradient convergence can be proved when H satisfies Assumption (H2)(c)(ii)(A).

PROPOSITION 5.13. *Let Assumptions* (H1) *and* (H2) *with* (H2)(c)(ii)(A) *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 $\beta > 0$ such that, for any $k \geq 1$, one has*

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

[5] The closedness of the subdifferential always holds for a function J continuous on its domain (it is a direct consequence of the definition of the (limiting) subdifferential).

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 (H2)(c)(ii)(A), there exists ξ such that

$$\forall\, k \in \mathbb{N}^*, \quad \|\nabla_x h(x^k, y^k) - \nabla_x h(x^k, y^{k-1})\| \leq \xi\, \|y^k - y^{k-1}\|. \tag{39}$$

From Lemma 5.8, there exists $(q_x^k, q_y^k) \in \partial J(x^k, y^k)$ such that

$$q_x^k = \nabla_x h(x^k, y^k) - \nabla_x h(x^k, y^{k-1}) + p_x^k$$

with p_x^k and q_y^k defined in Proposition 5.4. Using the first inequality in (35), one has the following bound for q_x^k:

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

This bound, together with the second inequality in (35) in Proposition 5.4, and using that $(a+b)^2 \leq 2\,a^2 + 2\,b^2$, shows that

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

Setting

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

completes the proof. □

Remark 5.14. Assumption (H2)(c)(ii)(A) is a sufficient but not necessary condition to have the subgradient convergence of ASAP. Indeed, one can consider the following objective function:

$$J(x,y) = H(x,y) = \begin{cases} x^2\sqrt{y} & \text{if } y > 0, \\ 0 & \text{otherwise,}\end{cases} \quad \text{with} \quad \nabla_x H(x,y) = \begin{cases} 2x\sqrt{y} & \text{if } y > 0, \\ 0 & \text{otherwise,}\end{cases}$$

which satisfies Assumptions (H1) and (H2), but not (H2)(c)(ii)(A) as $y \mapsto \nabla_x H(x,y)$ is not locally Lipschitz at 0 (because $t \mapsto \sqrt{t}$ is not). In this case, the ASAP iterates (for $F = G = 0$) are, for any $k \in \mathbb{N}$,

$$x^{k+1} = \frac{x^k}{1 + \tau y^k} \quad \text{and} \quad y^{k+1} = y^k.$$

Thus, one can find $(q_x^k, q_y^k) \in \partial J(x^k, y^k)$ such that $(q_x^k, q_y^k) \to 0$, since $\nabla_x h(x^k, y^k) = \nabla_x h(x^k, y^{k-1})$, which leads to $q_x^k = p_x^k$.

5.5. Convergence of ASAP to critical points under the KŁ property.

We summarize that, under Assumptions (H1) and (H2) with (H2)(c)(ii)(A), we have proved that any *bounded* sequence $\{z^k := (x^k, y^k)\}_{k\in\mathbb{N}}$ generated by ASAP satisfies the assumptions in [7, Thm. 2.9]:

(1) there exists $\rho \in (0, \infty)$ such that for any $k \geq 1$ the *sufficient decrease condition* holds (Proposition 5.2(a)):

$$J(z^k) + \rho \|z^k - z^{k-1}\|^2 \leq J(z^{k-1});$$

(2) there exist a subsequence $\{z^{k_j}\}_{j\in\mathbb{N}}$ and a critical point $z^* := (x^*, y^*)$ of J such that the *continuity condition* holds (Proposition 5.9):

$$z^{k_j} \to z^* \quad \text{and} \quad J(z^{k_j}) \to J(z^*) \quad \text{as } j \to +\infty;$$

(3) there exists $\beta \in (0, +\infty)$ such that for any $k \geq 1$ the *relative error condition* holds (Proposition 5.13):

$$\exists\, q^k := (q_x^k, q_y^k) \in \partial J(z^k) \quad \text{obeying} \quad \|q^k\| \leq \beta \|z^k - z^{k-1}\|.$$

Now we can prove the strong convergence of ASAP for objective functions satisfying the KŁ property at their critical points.

THEOREM 5.15. *Let Assumptions* (H1) *and* (H2) *with* (H2)(c)(ii)(A) *hold and let* $\{z^k :=(x^k, y^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by ASAP which is assumed to be bounded. Let* $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$ *be a limit point of* $\{(x^k, y^k)\}_{k\in\mathbb{N}}$. *Assume also that* J *satisfies the KŁ property at* (x^*, y^*). *Then the following hold:*

(a) $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ *is a Cauchy sequence converging to* (x^*, y^*), *which is a critical point of* J;

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

Proof. This theorem is a direct consequence of [7, Thm. 2.9]. □

Theorem 5.15 proves that, for objectives J satisfying Assumptions (H1) and (H2) with (H2)(c)(ii)(A), any bounded sequence generated by ASAP converges to a limit point $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$ provided J satisfies the KŁ property at (x^*, y^*). This includes various nonsmooth and nonconvex objective functions J that are continuous on their closed domain (see subsections 2.2 and 3.3 and especially Theorem 3.11, which provides a *sufficient* condition for critical points to satisfy the KŁ property).

Appendix A. Proof of Lemma 2.1. All terms composing functions F and G in (14) are of the form $\theta : U \to \mathbb{R}_+$

$$\theta(z) = \psi(\|Lz\|),$$

where U is a finite-dimensional vector space, $L : U \to \mathbb{R}^n$ is a bounded linear operator and $\|\cdot\|$ stands for Frobenius norm. The function θ is continuous as being the composition of continuous functions. By Assumptions (R)(a) and (R)(b), its gradient is

$$\forall\, z \in U, \quad \nabla\theta(z) = \begin{cases} \dfrac{\psi'(\|Lz\|)}{\|Lz\|} L^T L z & \text{if } Lz \neq 0, \\ 0 & \text{if } Lz = 0. \end{cases}$$

Since ψ satisfies (R)(b), $\nabla\theta$ is continuous. The second derivative of θ (i.e., its Hessian) reads, for any $z \in U$, as

$$\nabla^2\theta(z) = \left(\psi''(\|Lz\|) - \frac{\psi'(\|Lz\|)}{\|Lz\|}\right) L^T \frac{Lz}{\|Lz\|}\,\frac{(Lz)^T}{\|Lz\|} L + \frac{\psi'(\|Lz\|)}{\|Lz\|}\, L^T L,$$

where $\nabla^2\theta(z)$ is well defined for any $z \in U$. Thus, θ is a continuously differentiable function. Noticing that L and $|\psi''|$ are bounded, the Lipschitz constant of $\nabla\theta$ is upper bounded as

$$\|\nabla^2\theta\|_\infty \leq 3\,\|\psi''\|_\infty \|L\|^2,$$

where the last result is due to Assumption (R)(c).

Appendix B. Proof of Example 4.2. (a) Let $J = \chi_{\mathcal{B}}$ with

$$\mathcal{B} := \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}.$$

As a proper lower semicontinuous closed[6] convex function, the subdifferential of J satisfies

$$(p_x, p_y) \in \partial J(x,y) \quad \iff \quad (x,y) \in \partial J^*(p_x, p_y)$$

with $J^* : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\}$ the convex conjugate of J given by

$$J^*(p_x,p_y) := \sup_{(x,y)\in\mathbb{R}^2} \{\langle x, p_x\rangle + \langle y, p_y\rangle - J(x,y)\}.$$

It is easy to check that $J^* = \|\cdot\|$. Then,

$$\partial J^*(p_x,p_y) = \begin{cases} \left\{ \dfrac{(p_x,p_y)}{\|(p_x,p_y)\|} \right\} & \text{if } (p_x,p_y) \neq (0,0), \\ \mathcal{B} & \text{if } (p_x,p_y) = (0,0). \end{cases}$$

Hence, we obtain the following results from the above:

(i) $(0,0) \in \partial J(x,y)$ iff $(x,y) \in \partial J^*(0,0) = \mathcal{B}$,

(ii) $(0,0) \neq (p_x,p_y) \in \partial J(x,y)$ iff $\|(x,y)\| = 1$ and $(p_x,p_y)/\|(p_x,p_y)\| = (x,y)$.

As a consequence, one has

$$\forall\,(x,y) \in \mathcal{B}, \quad \partial J(x,y) = \begin{cases} \{(0,0)\} & \text{if } x^2+y^2 < 1, \\ \{(\lambda\, x, \lambda\, y) \mid \lambda \neq 0\} & \text{if } x^2+y^2 = 1. \end{cases}$$

Let us now compute the partial subdifferentials. Let $y \in [-1,1]$. Then,

$$J(x,y) = \begin{cases} 0 & \text{if } |x| \leq \sqrt{1-y^2}, \\ +\infty & \text{otherwise} \end{cases}$$
$$= \chi_{[-\sqrt{1-y^2},\sqrt{1-y^2}]}(x).$$

Since $x \mapsto J(x,y)$ is proper, lower semicontinuous, closed, and convex, one may again use the convex conjugate to prove that

$$p_x \in \partial_x J(x,y) \quad \iff \quad x \in \partial\chi^*_{[-\sqrt{1-y^2},\sqrt{1-y^2}]}(p_x).$$

[6] Meaning that the graph of J is closed.

Using the definition of the convex conjugate, one gets

$$\chi^*_{[-\sqrt{1-y^2},\sqrt{1-y^2}]}(p_x) = \sup_{x\in[-\sqrt{1-y^2},\sqrt{1-y^2}]} \langle x, p_x\rangle = \sqrt{1-y^2}\,|p_x|$$

so that, for any $p_x \in [-\sqrt{1-y^2}, \sqrt{1-y^2}]$,

$$\partial\chi^*_{[-\sqrt{1-y^2},\sqrt{1-y^2}]}(p_x) = \begin{cases} \left\{\sqrt{1-y^2}\,\dfrac{p_x}{|p_x|}\right\} & \text{if } p_x \neq 0,\\ [-\sqrt{1-y^2}, \sqrt{1-y^2}] & \text{if } p_x = 0. \end{cases}$$

Hence, one eventually gets that

$$\forall\,(x,y)\in\mathcal{B},\quad \partial_x J(x,y) = \begin{cases} \{0\} & \text{if } x^2+y^2<1,\\ \{\lambda_x\, x \mid \lambda_x \geq 0\} & \text{if } x^2+y^2=1. \end{cases}$$

Similarly, one can prove that

$$\forall\,(x,y)\in\mathcal{B},\quad \partial_y J(x,y) = \begin{cases} \{0\} & \text{if } x^2+y^2<1,\\ \{\lambda_y\, y \mid \lambda_y \geq 0\} & \text{if } x^2+y^2=1. \end{cases}$$

Hence, this shows that

$$\forall\,(x,y)\in\mathcal{B},\quad \partial_x J(x,y)\times\partial_y J(x,y) = \begin{cases} \{(0,0)\} & \text{if } x^2+y^2<1,\\ \{(\lambda_x\, x, \lambda_y\, y) \mid \lambda_x,\lambda_y \geq 0\} & \text{if } x^2+y^2=1. \end{cases}$$

As a consequence, $\partial J(x,y) \subsetneq \partial_x J(x,y)\times\partial_y J(x,y)$ for any $x^2+y^2=1$.

(b) Let us now consider, for any $(x,y)\in\mathbb{R}^2$,

$$J(x,y) = \begin{cases} 0 & \text{if } (x,y)\in([0,1]\times[0,1])\cup([1,2]\times[0,2]),\\ +\infty & \text{otherwise.} \end{cases}$$

The domain of J is clearly nonseparable and consists in the union of two (convex) rectangles $\mathcal{D}_1 := [0,2]\times[0,1]$ and $\mathcal{D}_2 := [1,2]\times[0,2]$. Let us first compute $\widehat{\partial} J(x,y)$. It is obvious that, for any point $(x,y)\in \operatorname{int}(\operatorname{dom} J)$, J is locally equal to zero, so that $\widehat{\partial} J(x,y) = \{(0,0)\}$. Similarly, for any point (x,y) on the boundary $\partial(\operatorname{dom} J)$ of $\operatorname{dom} J$ except $(1,1)$, J is locally equal to the (convex) indicator function of $\mathcal{D}_1$ for $(x,y)\in\mathcal{D}_1\setminus(2,1)$ and equal to that of $\mathcal{D}_2$ when $(x,y)\in\mathcal{D}_2\setminus(1,0)$. Let $(x,y)\in\partial(\operatorname{dom} J)\cap(\mathcal{D}_1\setminus\{(1,1),(2,1)\})$ (solid red line in Figure 1).[7] As noted above, one has $\widehat{\partial} J(x,y) = \partial\chi_{\mathcal{D}_1}(x,y)$. The separability of $\mathcal{D}_1$ implies that $\partial\chi_{\mathcal{D}_1}(x,y) = \partial_x\chi_{\mathcal{D}_1}(x,y)\times\partial_y\chi_{\mathcal{D}_1}(x,y)$. It is also easy to check that

$$\partial_x J(x,y) = \partial\chi_{[0,2]}(x) = \partial_x\chi_{\mathcal{D}_1}(x,y) \quad\text{and}\quad \partial_y J(x,y) = \partial\chi_{[0,1]}(y) = \partial_y\chi_{\mathcal{D}_1}(x,y)$$

so that $\partial_x J(x,y)\times\partial_y J(x,y) = \widehat{\partial} J(x,y) \subset \partial J(x,y)$. The same inclusion holds for any $(x,y)\in\partial(\operatorname{dom} J)\cap(\mathcal{D}_2\setminus\{(1,1),(1,0)\})$ (dashed blue line in Figure 1) with the same arguments. Besides, one can check that $\partial_x J(1,1)\times\partial_y J(1,1) = (0,0)$. Since one has $J(x,y)\geq 0$ for any $(x,y)\in\mathbb{R}^2$, it yields $(0,0)\in\widehat{\partial} J(1,1)$. As a consequence, we have proved that

$$\forall\,(x,y)\in\operatorname{dom} J,\quad \partial_x J(x,y)\times\partial_y J(x,y) \subset \partial J(x,y).$$

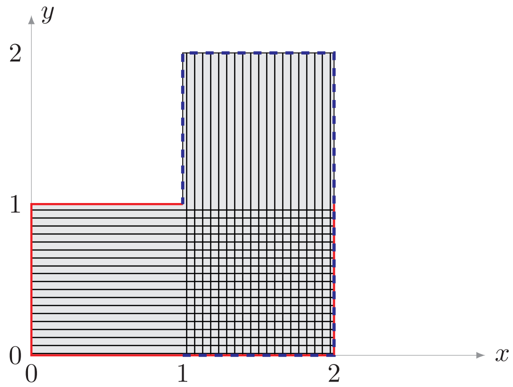

FIG. 1. *Illustration for Example* 4.2(b)*. The gray shaded domain is* dom H*, where the horizontal lines represent* $\mathcal{D}_1 := [0,2] \times [0,1]$ *and the vertical lines represent* $\mathcal{D}_2 := [1,2] \times [0,2]$.

Appendix C. Proof of Lemma 5.10. Suppose that $\{\operatorname{dist}(z^k, \mathcal{L}(z^0))\}_{k\in\mathbb{N}}$ does not converge to zero as $k \to +\infty$. Then there exist $M > 0$ and a strictly increasing sequence $\{k_j\}_{j\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}}$ such that, for any $j \in \mathbb{N}$, $\operatorname{dist}(z^{k_j}, \mathcal{L}(z^0)) > M$. However, since $(z^{k_j})_{j\in\mathbb{N}}$ is a subsequence of the bounded sequence $\{z^k\}_{k\in\mathbb{N}}$, it has a convergent subsequence $\{z^{k_{j_n}}\}_{n\in\mathbb{N}}$ of limit $z^* \in \mathcal{L}(z^0)$. Thus, in particular,

$$M < \operatorname{dist}(z^{k_{j_n}}, \mathcal{L}(z^0)) \leq \|z^{k_{j_n}} - z^*\| \xrightarrow[n\to+\infty]{} 0,$$

which leads to a contradiction.

Acknowledgments. This work is dedicated to the memory of Mila Nikolova, who sadly passed away in June 2018. She devoted precious time to continuously enhance this paper, even when her health condition became critical. Her contribution was invaluable, and one could not help but admire the energy and insights she brought to this work under very tough circumstances. The authors would like to thank Jérôme Bolte for his comments about semialgebraicity, as well as Jalal Fadili for his helpful remarks on the introduction.

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 (2005), pp. 531–547.

[2] C. Aguerrebere, A. Almansa, Y. Gousseau, and P. Musé, *A Bayesian hyperprior approach for joint image denoising and interpolation, with an application to HDR imaging*, IEEE Trans. Comput. Imaging, 3 (2017), pp. 633–646.

[3] P. Arias, V. Caselles, and G. Facciolo, *Analysis of a variational framework for exemplar-based image inpainting*, 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, *Alternating proximal algorithms for weakly coupled convex minimization problems: Applications to dynamical games and PDE's*, J. Convex Anal., 15 (2008), pp. 485–506.

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

[7] Color figures are available in the online version of this paper.

[7] 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), pp. 91–129.

[8] H. Attouch, P. Redont, and A. Soubeyran, *A new class of alternating proximal minimization algorithms with costs-to-move*, SIAM J. Optim., 18 (2007), pp. 1061–1081.

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

[10] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books Math., Springer, Cham, 2011.

[11] 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.

[12] A. Beck and M. Teboulle, *Smoothing and first order methods: A unified framework*, SIAM J. Optim., 22 (2012), pp. 557–580.

[13] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060.

[14] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1995.

[15] E. Bierstone and P. D. Milman, *Semianalytic and subanalytic sets*, Publ. Math. Inst. Hautes Etudes Sci., 67 (1988), pp. 5–42.

[16] 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.

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

[18] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program. 146 (2014), pp. 459–494.

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

[20] 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.

[21] X. Chen and W. Zhou, *Penalty methods for a class of non-Lipschitz optimization problems*, SIAM J. Optim., 26 (2016), pp. 1465–1492.

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

[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. Methods Oper. Res., 66 (2007), pp. 373–407.

[25] 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.

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

[27] 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. 1385–1415.

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

[29] S. Łojasiewicz, *Ensembles semi-analytiques*, preprint, Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette, France, 1965.

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

[31] 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.

[32] R. T. Rockafellar and J. B. Wets, *Variational analysis*, Grundlehren Math. Wiss. 317, Springer, New York, 1998.

[33] H.-J. M. Shi, S. Tu, Y. Xu, and W. Yin, *A primer on coordinate descent algorithms*, Techical report, UCLA, Los Angeles, CA, 2017.

[34] M. Shiota, *Geometry of subanalytic and semialgebraic sets*, Prog. Math. 150, Birkhäuser, Basel, 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.

Comput. Imaging, 3 (2017), pp. 187–201.

[36] P. Tan, Y. Ferrec, and L. Rousset-Rouvière, *Correction par méthode variationnelle des non uniformités des détecteurs d'un interféromètre imageur*, in XXVIe colloque GRETSI 2017, 2017; available at https://hal.archives-ouvertes.fr/hal-01622265.

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

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

[39] 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.

[40] Y. Xu and W. Yin, *A globally convergent algorithm for nonconvex optimization based on block coordinate update*, J. Sci. Comput., 72 (2017), pp. 700–734.

[41] X. Zhang, X. Zhang, X. Li, Z. Li, and S. Wang, *Classify social image by integrating multimodal content*, Multimed. Tools Appl., 77 (2017), pp. 7469–7485.