Alternating block coordinate proximal forward-backward descent for nonconvex regularised problems with biconvex terms

Mila Nikolova, Pauline Tan

Alternating block coordinate proximal forward-backward descent for nonconvex regularised problems with biconvex terms

Mila Nikolova*, and Pauline Tan+

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

Abstract In this work we consider a broad class of smooth optimization problems composed of a biconvex data-fidelity terms and smooth, nonconvex regularisation terms. We propose a family of attractive schemes for solving this class of problems. It is based on the standard alternate proximal linearized forward-backward approach. Unlike the existing prox-based algorithms, our approach exploits the biconvex structure of the data term. Thus we use proximity operators with respect to convex functions only. The iterates are uniquely defined, independently of the form of regularization terms.

1 Introduction

In this work we consider a broad class of smooth optimization problems composed of a biconvex data-fidelity terms and regularisation terms. They consist of solving the following nonconvex minimization problem:

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

where x and y belong to real finite-dimensional spaces of appropriate dimensions. Our focus is on situations when H is a biconvex function: for any y fixed, $x \mapsto H(x,y)$ is convex and for any x fixed, $y \mapsto H(x,y)$ is convex. It worths emphasizing that such a H is generally nonconvex, and thus J can have numerous local minimizers. Bicovex data terms H arise in various important applications, such as total least squares [15], blind deblurring [8], channel source separation typically blind [11], patch-based methods [4], to cite a few. We focus on differentiable regularizers F and G for four main reasons. (a) In a large amount of cases there is no evidence that x or y belong to a dictionary that is sparse. The use of sparsity (F or G nonsmooth at the origin) then leads to adding unrealistic frequency components in the solution [1]. (b) Quite few facts on the global minimizers of J when F or G is nonsmooth are known, a pitfall in blind deblurring [8] shows that the global minimizers are trivial. (c) Numerous algorithms involving nonsmooth regularization terms are approximated by smooth regularizers [7]. In some cases the error induced by the smooth approximation is thoroughly analyzed see, e.g., [12]. (d) Nonetheless the way of smoothing, the latter definitely alterates the main features of all local minimizers, see e.g., [18]. The paramount reason (a) together with those in (b), (c), and (d), justify our choice to consider well-selected smooth regularizers F and G.

In some cases the biconvex structure is due to a bilinear / biaffine term. In [3, 21] the problem is reformulated based on the bilinear term and a branch-and-bound procedure taking into account the affine structure is proposed.

The usual approach to solve the problem in (1) is to use the Alternating Convex Minimization [16] which amounts to generate a sequence starting from (x^0, y^0) with iterates (x^k, y^k) given by

$$\begin{array}{rcl} x^k & \in & \arg\min_x J(x, y^{k-1}), \\ y^k & \in & \arg\min_y J(x^k, y), \end{array}$$

which amounts to a block-coordinate Gauss-Seidel method, also known as block-coordinate descent (BCD) method, which can be found in several studies, see e.g., [20]. For biconvex differentiable objectives, this approach was applied in [4], [2]. A key condition necessary to obtain convergence is that the minimum in each step is uniquely attained.

A way to relax this assumption is to consider the proximal regularization of the BCD scheme:

$$\begin{array}{rcl} x^k & \in & \arg\min_x \left(J(x, y^{k-1}) + \frac{1}{2\tau_1}\|x - x^{k-1}\|^2 \right), \\ y^k & \in & \arg\min_y \left(J(x^k, y) + \frac{1}{2\tau_2}\|y - y^{k-1}\|^2 \right), \end{array}$$

where $\tau_1 > 0$ and $\tau_2 > 0$ are step-sizes. In the non-convex case, the situation is much harder, see e.g., [13, 5]. In [22] a multi-convex problem with nonsmooth regularization is solved by 3 methods, BCD, proximal BCD, and prox-linear BCD where proximity operators are based on the regularization terms. In [10, 6], differentiable biconvex data-terms with nonsmooth regularization are considered using prox-linear BCD also based on the regularization terms. In the articles mentioned in this paragraph, the objectives are nonsmooth and convergence results are obtained using the Kurdyka-Łojasiewicz property.

We propose a family of attractive schemes for solving a broad class of problems of the form in (1) where F and G are differentiable, nonconvex in general. It is based on the standard alternate proximal linearized forward-backward approach. Unlike the prox-based algorithms presented before, our approach exploits the biconvex structure of the term H: each one of our proximal steps is computed with respect to the partial convex functions $x \mapsto H(x, y)$ and $y \mapsto H(x, y)$, respectively. Thus we use proximity operators with respect to convex functions only. The iterates are uniquely defined, independently of the form of F and G. Further, our stepsizes depend only on the Lipschitz constants of F and G. The only major assumption is that the iterates are bounded. However, if the initial value $J(x^0, y^0)$ is below a threshold, boundedness is ensured without coercivity requirements. The general form of the proposed algorithms can deal with any biconvex function. When H is biquadratic the proximal operators have an explicit simple form. The threshold leading to bounded iterates is easy to compute explicitly.

Outline of the paper Preliminary facts on Proximal Forward-Backward and on biconvex functions are presented in section 2. The main algorithm is given in section 3. Convergence facts are presented in the following subsections. They concern the convergence of the sequences generated by the algorithm towards stationary points, as well as the existence of limit points. Section 5 is devoted to the classical case when H is biquadratic. The implementable form of the algorithm is presented in 5.1. When the null-space of F or G is non-null, the objective J is not coercive. In subsection 5.2 we compute an initial value ensuring boundedness of the iterates. Existence of global minimizers is established as well. Under convexity assumptions on F and G, stronger convergence facts are derived in subsection 5.3.

Notations We consider that $x \in U$ and $y \in V$ where U and V are real finite-dimensional spaces of appropriate dimensions. For a positive integer n, we use the index set $\mathbb{I}_n = \{1, \ldots, n\}$. Sequences (e.g., iterates) are denoted by $\{z^k\}_k$ and their subsequences by $\{z^{k_j}\}_j$. The kth element of the vector or a matrix x (seen as a vector) reads as x_k. A vector or a matrix indexed for some purpose is denoted by $x_{(i)}$. The identity operator is denoted by I, its size is clear from the context. A vector or a matrix of zeros of arbitrary dimension is denoted by 0. We write x^{T} for the transposed of x. The subdifferential of the convex function h is denoted by ∂h. Likewise, $\partial_x H(\cdot, y)$ (reps., $\partial_y H(x, \cdot)$) denotes the subdifferential of the convex function $x \mapsto H(x, y)$ (resp., $y \mapsto H(x, y)$).

2 Preliminaries

2.1 Proximal Forward-Backward (PFB) descent

Proximity operator Let $h : U \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous (lsc) and convex function. Given $u \in U$ and $\tau > 0$, let us define [17]

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

Note that $\operatorname{prox}_{\tau h}(u)$ is uniquely defined since it is the minimizer of a strongly convex (thus strictly and coercive) function.

Lemma 1 (Proximal inequality)**.** *Let $h : U \to \mathbb{R} \cup \{+\infty\}$ be a proper, lsc and convex function. Given $u \in U$ and $\tau > 0$, let $x^+ = \operatorname{prox}_{\tau h}(u)$. Then*

$$h(x^+) - h(x) \leqslant \frac{1}{\tau} \langle x - x^+,\ x^+ - u \rangle \quad \forall\, x \in U$$

Proof. The optimality condition which characterizes x^+ yields

$$p = \frac{u - x^+}{\tau} \in \partial h(x^+)$$

Writing the subgradient inequality for $p \in \partial h(x^+)$, we have

$$\forall\, x \in U \quad h(x) \geqslant h(x^+) + \langle x - x^+, p\rangle = h(x^+) + \frac{1}{\tau}\langle x - x^+, u - x^+\rangle$$

which is the desired result. □

PFB descent Let $h : U \to \mathbb{R} \cup \{+\infty\}$ be a proper, lsc and convex function and $f : U \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient. Such functiond are said $\mathrm{L}_{\nabla f}$*-smooth*. Note that f may be nonconvex. Let us consider the Proximal Forward-Backward (PFB) descent, defined for any $u \in U$ by:

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

Once again, thanks to the convexity of h, the point x^+ is uniquely defined.

Remark 1. The PFB descent can be viewed as the minimization of $h + \tilde{f}$, where $\tilde{f}$ is a quadratic approximation of around the point u:

$$\begin{aligned} x^+ &= \mathrm{prox}_{\tau h}\left(u - \tau \nabla f(u)\right) && (2)\\ &= \arg\min_{x \in U} \left\{ h(x) + \frac{1}{2\tau}\|x - (u - \tau\nabla f(u))\|^2 \right\} && (3)\\ &= \arg\min_{x \in U} \left\{ h(x) + \underbrace{f(u) + \langle x - u, \nabla f(u)\rangle + \frac{1}{2\tau}\|x - u\|^2}_{=\tilde{f}(z)} \right\} && (4) \end{aligned}$$

Let us recall a classical result which comes from the smoothness of f.

Lemma 2 (Descent facts)**.** *Let $f : U \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient. Then [9, A. 24]*

$$f(x) \leqslant f(u) + \langle x - u, \nabla f(u)\rangle + \frac{\mathrm{L}_{\nabla f}}{2}\|x - u\|^2 \qquad \forall\, x, u \in U \tag{5}$$

The next lemma makes possible a sufficient-decrease condition.

Lemma 3 (Decrease properties)**.** *Let $f : U \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient and $h : U \to \mathbb{R}$ a convex, lsc and proper function. For any u consider x^+ defined by* (2)*. Then*

$$f(x^+) + h(x^+) \leqslant f(u) + h(u) - \left(\frac{1}{\tau} - \frac{\mathrm{L}_{\nabla f}}{2}\right)\|x^+ - u\|^2 \tag{6}$$

Proof. Applying the proximal inequality to h in (2) with $u := x - \tau\nabla f(x)$ and $x := u$ we have

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

Then, adding the descent Lemma 2 (5) for f to the obtained result yields

$$f(x^+) + h(x^+) \leqslant f(u) + h(u) - \left(\frac{1}{\tau} - \frac{\mathrm{L}_{\nabla f}}{2}\right)\|x^+ - u\|^2$$

which completes the proof. □

2.2 Biconvexity

Definition 1. *A function $H : U \times V \to \mathbb{R}$ is called biconvex, if 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.*

Similarly, a set $C \subset U \times V$ is biconvex if $C|_U$ is convex for every $x \in U$ and $C|_V$ is convex for every $y \in C|_V$. Obviously, a biconvex set is not convex in general. The following proposition was taken from [14].

Proposition 1 (Goh et al. 1994)**.** *Let $H : U \times V \to \mathbb{R}$ be biconvex. Then its level sets*

$$\operatorname{lev}_{H \leqslant r} := \{(x, y) \in U \times V \ \mid \ H(x, y) \leqslant r\}$$

are biconvex for every $r \in \mathbb{R}$.

Biconvex function are in general nonconvex and may have a large amount of local minimizers. A survey on biconvex optimization problems can be found in [16]. However, the question arises whether the convex substructures of a biconvex function can be utilized more efficiently for the minimization of objective functions involving biconvex terms than in the case of general non-convex optimization problems.

A noteworthy algorithm that exploits the partial convexity in biconvex minimization problems is the Alternate Convex Search (ACS) algorithm [16, Algorithm 4.1]. However, despite the study led in [16], there are few theoretical results about the convergence of this algorithm when applied in the general case.

3 The ABC-PFB algorithm

Let us now consider the following nonconvex and (possibly nonsmooth) minimization problem:

$$\arg \min_{(x,y) \in U \times V} \Big\{ J(x, y) := F(x) + G(y) + H(x, y) \Big\} \tag{7}$$

Assumptions (J)

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

(b) J is either biconvex or continuously differentiable.

(c) $F : U \to \mathbb{R}$ and $G : V \to \mathbb{R}$ are $\mathrm{L}_{\nabla F}$-smooth and $\mathrm{L}_{\nabla G}$-smooth, respectively ;

(d) $H : U \times V \to \mathbb{R} \cup \{+\infty\}$ is continuous and biconvex.

To solve Problem (7), we consider the following algorithm

Algorithm 1 (Alternating Block Coordinate Proximal Forward-Backward descent (ABC-PFB))**.**
Initialization: (x^0, y^0) *and* $\tau_1 > 0$, $\tau_2 > 0$
Iterations: for $k \geqslant 0$

$$x^k = \operatorname{prox}_{\tau_1 H(., y^{k-1})} \Big(x^{k-1} - \tau_1 \nabla F(x^{k-1}) \Big) \tag{8}$$

$$y^k = \operatorname{prox}_{\tau_2 H(x^k, .)} \Big(y^{k-1} - \tau_2 \nabla G(y^{k-1}) \Big) \tag{9}$$

Using Remark 1, the iterations (8) and (9) are equivalent to minimizing the following quadratic approximations of $H(\cdot, y^{k-1})$ around the point x^{k-1} and $H(x^k, \cdot)$ around the point y^{k-1}, respectively:

$$\tilde{J}_x(x, y^{k-1}) = F(x^{k-1}) + \langle x - x^{k-1}, \nabla F(x^{k-1}) \rangle + H(x, y^{k-1}) + \frac{1}{2\tau_1} \|x - x^{k-1}\|^2 \tag{10}$$

$$\tilde{J}_y(x^k, y) = G(y^{k-1}) + \langle y - y^{k-1}, \nabla G(y^{k-1}) \rangle + H(x^k, y) + \frac{1}{2\tau_2} \|y - y^{k-1}\|^2 \tag{11}$$

Hence, ignoring the constant terms, by definition of the proximity operators, x^k and y^k are defined as

$$x^k = \arg\min_{x \in U} \left\{ \langle x - x^{k-1}, \nabla F(x^{k-1}) \rangle + H(x, y^{k-1}) + \frac{1}{2\tau_1} \|x - x^{k-1}\|^2 \right\} \tag{12}$$

$$y^k = \arg\min_{y \in V} \left\{ \langle y - y^{k-1}, \nabla G(y^{k-1}) \rangle + H(x^k, y) + \frac{1}{2\tau_2} \|y - y^{k-1}\|^2 \right\} \tag{13}$$

4 General convergence facts

The assumptions below concerns the step-sizes in the iterations so that convergence can be ensured. Unlike [10, 6], our step-sizes do not need to be updated at each iteration.

Assumption (τ) We assume that $\tau_1 < \dfrac{2}{\mathrm{L}_{\nabla F}}$ and that $\tau_2 < \dfrac{2}{\mathrm{L}_{\nabla G}}$ and denote:

$$\lambda_x := \frac{1}{\tau_1} - \frac{\mathrm{L}_{\nabla F}}{2} > 0 \qquad \lambda_y := \frac{1}{\tau_2} - \frac{\mathrm{L}_{\nabla G}}{2} > 0 \qquad \lambda := \max\{\lambda_x,\ \lambda_y\} > 0$$

4.1 Convergence of the objective and its subdifferentials

Our first result states the convergence of the sequence $\{J(x^k, y^k)\}_k$ to a real number J^*, which can be proved to be the value of a critical point of J if $\{(x^k, y^k)\}_k$ has at least one limit point.

Proposition 2. *Assume that (x^k, y^k) is generated by Algorithm 1 and Assumptions (J) and (τ) hold. Then*

(a) *$\{J(x^k, y^k)\}_k$ and $\{J(x^k, y^{k-1})\}_k$ are nonincreasing and converge to the same value denoted J^*;*

(b) *$J(x^{k-1}, y^{k-1}) \geqslant J(x^k, y^{k-1}) \geqslant J(x^k, y^k) \quad \forall\, k > 1$ (interlacing inequality);*

(c) *The following holds for every k*

$$\begin{aligned} J(x^{k-1}, y^{k-1}) - J(x^k, y^{k-1}) &\geqslant \lambda_x \|x^k - x^{k-1}\|^2; \\ J(x^k, y^{k-1}) - J(x^k, y^k) &\geqslant \lambda_y \|y^k - y^{k-1}\|^2; \\ J(x^{k-1}, y^{k-1}) - J(x^k, y^k) &\geqslant \lambda \left(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2\right). \end{aligned} \tag{14}$$

(d) $\lim\limits_{k\to\infty} \|x^{k+1} - x^k\|^2 = 0$ *and* $\lim\limits_{k\to\infty} \|y^{k+1} - y^k\|^2 = 0.$

Proof. Using Lemma 3 with $f := F$, $h := H(., y^{k-1})$, $x^+ := x^k$ as defined in (12) and $u := x^{k-1}$ shows that

$$F(x^k) + H(x^k, y^{k-1}) \leqslant F(x^{k-1}) + H(x^{k-1}, y^{k-1}) - \lambda_x \|x^k - x^{k-1}\|^2$$

hence

$$J(x^{k-1}, y^{k-1}) \geqslant J(x^k, y^{k-1}) + \lambda_x \|x^k - x^{k-1}\|^2 \tag{15}$$

In a similar way, we have

$$J(x^k, y^k) \geqslant J(x^{k+1}, y^k) + \lambda_x \|x^{k+1} - x^k\|^2 \tag{16}$$

From Lemma 3 yet again but with $f := G$, $h := H(x^k, .)$, $x^+ := y^k$ as defined in (13) and $u := y^{k-1}$ one has

$$G(y^k) + H(x^k, y^k) \leqslant G(y^{k-1}) + H(x^k, y^{k-1}) - \lambda_y \|y^k - y^{k-1}\|^2$$

hence

$$J(x^k, y^{k-1}) \geqslant J(x^k, y^k) + \lambda_y \|y^k - y^{k-1}\|^2 \tag{17}$$

Taking (15), (16) and (17) together yields

$$\begin{aligned} J(x^{k-1}, y^{k-1}) &\geqslant J(x^k, y^{k-1}) + \lambda_x \|x^k - x^{k-1}\|^2 \\ &\geqslant J(x^k, y^k) + \lambda_x \|x^k - x^{k-1}\|^2 + \lambda_y \|y^k - y^{k-1}\|^2 \\ &\geqslant J(x^{k+1}, y^k) + \lambda_x \left(\|x^{k+1} - x^k\|^2 + \|x^k - x^{k-1}\|^2\right) + \lambda_y \|y^k - y^{k-1}\|^2 \end{aligned}$$

This proves (c). It follows that the sequences $\{J(x^k, y^k)\}_k$ and $\{J(x^k, y^{k-1})\}_k$ are monotonous decreasing, *interlaced* by

$$J(x^{k-1}, y^{k-1}) \geqslant J(x^k, y^{k-1}) \geqslant J(x^k, y^k) \qquad \forall\, k \geqslant 1 \tag{18}$$

Hence, (b) is shown. Moreover, these sequences are bounded from below. since J is. Therefore $\{J(x^k, y^k)\}_k$ and $\{J(x^k, y^{k-1})\}_k$ converge to the same finite number, say J^*, which proves (a). Taking the limit as $k \to \infty$ shows that

$$0 = \lim_{k\to\infty} J(x^{k-1}, y^{k-1}) - J(x^k, y^k) \geqslant \lim_{k\to\infty} \left(\lambda_x \|x^k - x^{k-1}\|^2 + \lambda_y \|y^k - y^{k-1}\|^2\right) \tag{19}$$

Since $\lambda_x > 0$ and $\lambda_y > 0$ (see Assumption (τ)), it follows that

$$\lim_{k\to\infty} \|x^{k+1} - x^k\|^2 = 0 \qquad \text{and} \qquad \lim_{k\to\infty} \|y^{k+1} - y^k\|^2 = 0.$$

The proof is complete. □

Remark 2. The *interlacing* in (18) is clearly important. In [10] the authors prove only (19), but their algorithm generates Cauchy sequences. Since our sequences are certainly not Cauchy, this *interlacing* property is welcome.

In what follows, $\partial_x J(\cdot, y^{k-1})$ stands for the subdifferential for $J(\cdot, y^{k-1})$ and $\partial_y J(x^k, \cdot)$ stands for the subdifferential for $J(x^k, \cdot)$. Since F and G are smooth, one has $\partial_x J(x, y^{k-1}) = \nabla F(x) + \partial_x J(x, y^{k-1})$ and $\partial_y J(x^k, y) = \nabla G(y) + \partial_y J(x^k, y)$ for any $(x, y) \in U \times V$.

Proposition 3. *Let (x^k, y^k) be generated by Algorithm 1 and let Assumption (J) hold. Then for any $k \geqslant 1$, there exist $p_x^k \in \partial_x J(x^k, y^{k-1})$ and $p_y^k \in \partial_y J(x^k, y^k)$ such that*

(a) $\|p_x^k\| \leqslant \left(\mathrm{L}_{\nabla F} + \dfrac{1}{\tau_1}\right) \|x^{k-1} - x^k\|$ *and* $\|p_y^k\| \leqslant \left(\mathrm{L}_{\nabla G} + \dfrac{1}{\tau_2}\right) \|y^{k-1} - y^k\|$

(b) *Let Assumption (τ) hold as well. Then*

$$\lim_{k\to\infty} \|p_x^k\| = 0 \qquad \textit{and} \qquad \lim_{k\to\infty} \|p_y^k\| = 0. \tag{20}$$

Proof. The optimality condition for x^k defined by (12) yields

$$\frac{1}{\tau_1}(x^{k-1} - x^k) \ \in \ \nabla F(x^{k-1}) + \partial_x H(x^k, y^{k-1}) \tag{21}$$

By extracting $\partial_x H(x^k, y^{k-1})$ from (21) and using its explicit expression, we have

$$p_x^k := \nabla F(x^k) - \nabla F(x^{k-1}) + \frac{1}{\tau_1}(x^{k-1} - x^k) \ \in \partial_x J(x^k, y^{k-1})$$

Using Assumption (J) yields

$$\|p_x^k\| \leqslant \mathrm{L}_{\nabla F}\|x^{k-1} - x^k\| + \frac{1}{\tau_1}\|x^{k-1} - x^k\| = \left(\mathrm{L}_{\nabla F} + \frac{1}{\tau_1}\right) \|x^{k-1} - x^k\|$$

Using Assumption (τ), we can apply the last statement of Proposition 2 which entails

$$\lim_{k\to\infty} \|p_x^k\| \leqslant \left(\mathrm{L}_{\nabla F} + \frac{1}{\tau_1}\right) \lim_{k\to\infty} \|x^{k-1} - x^k\| = 0$$

Similar computations prove that

$$\lim_{k\to\infty} \|p_y^k\| \leqslant \left(\mathrm{L}_{\nabla G} + \frac{1}{\tau_2}\right) \lim_{k\to\infty} \|y^{k-1} - y^k\| = 0$$

which ends the proof. □

Assumption (B). The sequence generated by the Algorithm 1 is bounded.

This assumption is trivially satisfied if J is coercive. Boundedness of the iterates can be proven in other less restrictive cases, see e.g., [5]. The question is examined in a more general context in subsection 4.3 and for a family of non-coercive objectives in section 5.

The set of all limit points corresponding to a starting point (x^0, y^0) of Algorithm 1 is denoted by $\mathcal{L}(x^0, y^0)$.

Proposition 4. *Let (x^k, y^k) be generated by Algorithm 1 and let Assumptions (J), (τ) and (**B**) hold. Then the limit points (x^*, y^*) of the sequence of iterates are critical points of the objective J, i.e., they satisfy*

$$0 \ \in \partial_x J(x^*, y^*) \qquad \textit{and} \qquad 0 \ \in \partial_y J(x^*, y^*)$$

Proof. Let $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$. Then from assumption (B) there exists a subsequence $\left\{\left(x^{k_j}, y^{k_j}\right)\right\}_j$ such that $\left(x^{k_j}, y^{k_j}\right) \to (x^*, y^*)$ and $\left(x^{k_j}, y^{k_{j-1}}\right) \to (x^*, y^*)$ as $j \to \infty$ (thanks to Proposition 2(d)). Since J is continuous

$$\lim_{j\to\infty} J\left(x^{k_j}, y^{k_j}\right) = J(x^*, y^*) \qquad \text{and} \qquad \lim_{j\to\infty} J\left(x^{k_j}, y^{k_{j-1}}\right) = J(x^*, y^*)$$

Proposition 3 shows that there exist two sequences $\{(p_x^k, p_y^k)\}$ such that for any $k \in \mathbb{N}$

$$p_x^k \in \partial_y J(x^k, y^{k-1}) \qquad \text{and} \qquad p_y^k \in \partial_y J(x^k, y^k)$$

and $\|p_x^k\|, \|p_y^k\| \to 0$.

Then, following assumption (J)(b) two cases may occur.

Case 1: J is biconvex. Then the subgradient inequalities yield for any $j \in \mathbb{N}$

$$J(x, y^{k_j-1}) \geqslant J(x^{k_j}, y^{k_j-1}) + \langle p_x^{k_j}, x - x^{k_j} \rangle \qquad \text{and} \qquad J(x^{k_j}, y) \geqslant J(x^{k_j}, y^{k_j}) + \langle p_y^{k_j}, y - y^{k_j} \rangle$$

Taking the limit as $j \to \infty$ and using the continuity of J, we have

$$J(x, y^*) \geqslant J(x^*, y^*) \quad \forall\, x \qquad \text{and} \qquad J(x^*, y) \geqslant J(x^*, y^*) \quad \forall\, y$$

that is, $0 \in \partial_x J(x^*, y^*)$ and $0 \in \partial_y J(x^*, y^*)$.

Case 2: J is $\mathcal{C}^1$. This implies that H is continuously differentiable. Then, using the continuity of $\nabla_x J$ and $\nabla_y J$, we have

$$\lim_{k\to\infty} \nabla_x J(x^k, y^k) = \nabla_x J(x^*, y^*) \qquad \text{and} \qquad \lim_{j\to\infty} \nabla_y J(x^{k_j}, y^{k_j}) = \nabla_y J(x^*, y^*)$$

Proposition 3 shows that $\nabla_x J(x^*, y^*) = 0$ and $\nabla_y J(x^*, y^*) = 0$. □

4.2 Convergence of the iterates

Without additional strong hypothesis about the objective J, no convergence results can be proved for the sequence of iterates $\{(x^k, y^k)\}_k$. However, some weaker yet useful results can be shown.

Proposition 5 (Fixed points)**.** *If (x^*, y^*) is a critical point of J, then $(x^k, y^k) = (x^*, y^*)$ for any k.*

Proof. We prove it by induction. Suppose that (x^k, y^k) is a critical point of J. Then

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

By definition,

$$x^{k+1} = \operatorname{prox}_{\tau_1 H(\cdot, y^k)}(x^k + \tau_1 \partial_x H(x^k, y^k))$$

so that the optimality condition yields

$$0 \in \partial_x H(x^{k+1}, y^k) + \frac{1}{\tau_1}(x^{k+1} - x^k - \tau_1 \partial_x H(x^k, y^k))$$

This implies that

$$-\frac{1}{\tau_1}(x^{k+1} - x^k) \in \partial_x H(x^{k+1}, y^k) - \partial_x H(x^k, y^k)$$

Since $H(\cdot, y^k)$ is convex, the monotonicity of its subgradient implies that

$$-\frac{1}{\tau_1}\langle (x^{k+1} - x^k), x^{k+1} - x^k \rangle \geqslant 0$$

which obviously leads to $x^{k+1} = x^k$. Same computations show that $y^{k+1} = y^k$. □

The following result ensures that the sequence generated by Algorithm 1 globally approaches critical points of J.

Proposition 6. *Let $z^k := (x^k, y^k)$ be generated by Algorithm 1 and let Assumptions (J), (τ) and (**B**) hold. Then the distance of z^k to the set $\mathcal{L} := \mathcal{L}(x^0, y^0)$ of its limit points goes to zero.*

Recalling Proposition 3(b), this proposition is a direct consequence of the following lemma:

Lemma 4. *Let (z^k) be a sequence of U such that $\|z^{k+1} - z^k\|$ goes to zero. Suppose that the set $\mathcal{L}$ of the limit points of $\{z^k\}_k$ is nonempty. Then the distance of z^k to the set $\mathcal{L}$ goes to zero as $k \to \infty$.*

Proof. Since $\mathcal{L}$ is closed and nonempty, the distance d of z^k to the set $\mathcal{L}$ is well defined:

$$d(z^k, \mathcal{L}) := \min_{z \in \mathcal{L}} \|z - z^k\|$$

Suppose that $d(z^k, \mathcal{L})$ does not go to zero as $k \to \infty$, namely that there exist $\varepsilon > 0$ and a subsequence $\{z^{k_n}\}_n$ such that

$$\forall\, n \in \mathbb{N} \qquad d(z^{k_n}, \mathcal{L}) > 2\varepsilon. \tag{22}$$

Observe that $\{\|z^{k+1} - z^k\|\}_k$ is a sequence of real non-negative numbers converging to zero, hence it is a Cauchy sequence converging to zero. This, together with the fact that $z \mapsto d(z, \mathcal{L})$ is a continuous function [19, p. 19.], entails that for each $\varepsilon > 0$ there is $K \in \mathbb{N}$ such that for any $k > K$ one has $|d(z^{k+1}, \mathcal{L}) - d(z^k, \mathcal{L})| < \varepsilon$. Therefore, there exists $N \in \mathbb{N}$ so that for any $n > N$ one has [1] $k_n > K$ and thus

$$\forall\, n > N \qquad -\varepsilon < d(z^{k_n+1}, \mathcal{L}) - d(z^{k_n}, \mathcal{L}) < \varepsilon. \tag{23}$$

From the definition of $\mathcal{L}$ as the (nonempty) set of the limit points of $\{z^k\}$, there is an infinite number of points of z^k satisfying $d(z^k, \mathcal{L}) \leqslant \varepsilon$. Consequently, there exists k_n such that $d(z^{k_n+1}, \mathcal{L}) < \varepsilon$. From (23), together with (22), one obtains

$$d(z^{k_n+1}, \mathcal{L}) > d(z^{k_n}, \mathcal{L}) - \varepsilon > \varepsilon,$$

in contradiction to the fact that $d(z^{k_n+1}, \mathcal{L}) < \varepsilon$. Thus the assumption in (22) fails. □

When the iterate convergence cannot be proved, Proposition 6 gives strong information on the sequence behaviour. In particular, it shows that for k sufficiently large, the iterates (x^k, y^k) are arbitrary close to a critical point of J. In other terms, if Algorithm 1 is stopped after a sufficient number K of iterations, (x^K, y^K) is ensured to be close enough to a critical point.

4.3 Existence of limit points

Propositions 4 and 6 rely on the existence of limit points of the sequence $\{(x^k, y^k)\}_k$ (see assumption (B)). This may not be ensured, unless one adds assumptions on the objective function J. Additional hypotheses include the coercivity of J or the boundedness of the sequence $\{(x^k, y^k)\}_k$, which are sufficient but strong hypotheses.

Proposition 7. *Suppose F and G are coercive. Then, if*

$$s := \inf_{\ker F \times \ker G} J > \inf J$$

then for any $r < s$, the set $\operatorname{lev}_{\leqslant r} J$ is compact.

Remark 3. The assumption in Proposition 7 only aims at ensuring that there exists r such that $J(x, y) > r$ for any $(x, y) \in \ker F \times \ker G$ and such that $\operatorname{lev}_{\leqslant r} J$ is nonempty.

Proof. Let $r < s$ and suppose that $\operatorname{lev}_{\leqslant r} J$ is unbounded. Then there exist $(x^k, y^k)_k \in \operatorname{lev}_{\leqslant r} J$ such that $\|(x^k, y^k)\| \to +\infty$. Since $J(x^k, y^k) < s$, this implies that $(x^k, y^k) \notin (\ker F \times \ker G)$. Hence, $F(x^k)$ or $G(y^k)$ goes to infinity (by coercivity) and since $J(x^k, y^k) \geqslant F(x^k) + G(y^k)$, we have that $J(x^k, y^k) \to +\infty$, which leads to a contradiction with $J(x^k, y^k) < s$. Thus $\operatorname{lev}_{\leqslant r} J$ is bounded. Since J is lsc, $\operatorname{lev}_{\leqslant r} J$ is also closed. □

This general result shows that, if one manages to initialize Algorithm 1 with $J(x^0, y^0) < s$, then the iterates $\{(x^k, y^k)\}_k$ remain in a compact set, thus they are bounded. In the following section, we show how this can be used in practice for some specific classes of objective functions J.

5 A family of objectives J with H biquadratic

Let us consider objective functions of general form where the biconvex term H arises from a biaffine form, e.g., $xy - w$. Then

$$J(x, y) = F(x) + G(y) + H(x, y) = f(\|Ax\|) + g(\|Bx\|) + \frac{1}{2}\|xy - w\|^2 \tag{24}$$

with $x \in U = \mathbb{R}^{m \times n}$, $y \in V = \mathbb{R}^{n \times p}$ and $w \in W = \mathbb{R}^{m \times p}$ (data or perturbation), and where A and B are two linear operators. The norm $\|\cdot\|$ should be understood as the Frobenius norm when the argument is a matrix. Following Proposition 7 let us define

$$s := \inf_{(x,y) \in \ker F \times \ker G} J(x, y). \tag{25}$$

[1] If there was no such a k_n, the definition / the existence of $\mathcal{L}$ would fail.

Assumption (f, g)

(a) f and g are $\mathcal{C}^2$, strictly increasing functions satisfying $f(t) = g(t) = 0$ iff $t = 0$;

(b) for any $z^k \in U \setminus \ker F$ (resp., $z^k \in V \setminus \ker F$) with $\|z^k\| \to \infty$, there is K such that for any $k \geqslant K$ it holds that $f(z^k) \geqslant s$ (resp., $g(z^k) \geqslant s$).

Assumption (b) is quite weak, compared to coercivity.

Remark 4. The results in this section can be generalized to the case

$$H(x, y) = h(\|(Mx)y - w\|)$$

with M a square invertible matrix and $h : U \times V \to \mathbb{R}$ symmetric, $\mathcal{C}^2$-smooth, coercive. Further

$$F(x) = \sum_{i=1}^{q_x} f_i(\|A_i x\|) \quad \text{and} \quad G(y) = \sum_{j=1}^{q_y} g_j(\|B_j y\|) \tag{26}$$

with f_i and g_j as stated above, and A_i and B_j linear operators. We denote $A = (A_1, \ldots, A_I)$ and $B = (B_1, \ldots, B_J)$. Then $\ker F = \ker A$ and $\ker G = \ker A$.

5.1 The algorithm for J in (24)

Since $H(x, y) = \frac{1}{2}\|xy - w\|_F^2$ one has

$$\nabla_x H(x, y) = (xy - w)y^{\mathrm{T}} \quad \text{and} \quad \nabla_y H(x, y) = x^{\mathrm{T}}(xy - w)$$

Then a simple calculation shows that

$$\begin{aligned} \operatorname{prox}_{\tau_1 H(.,y)}(z) &= (z + \tau_1 w y^{\mathrm{T}})(I + \tau_1 y y^{\mathrm{T}})^{-1}; \\ \operatorname{prox}_{\tau_2 H(x,.)}(z) &= (I + \tau_2 x x^{\mathrm{T}})^{-1}(z + \tau_2 x^{\mathrm{T}} w). \end{aligned} \tag{27}$$

Observe that these proximal operators are simple and well defined since $(I + \tau_1 y y^{\mathrm{T}})$ and $(I + \tau_2 x x^{\mathrm{T}})$ are positive definite. Hence, Algorithm 1 applied to J reads

Algorithm 2 (ABC-PFB with H biquadratic)**.**
Initialization: (x^0, y^0) *and* $\tau_1 > 0$, $\tau_2 > 0$
Iterations: for $k \geqslant 0$

$$\begin{aligned} x^k &= (z + \tau_1 w y^T)(I + \tau_1 y y^T)^{-1} \qquad \textit{where} \qquad z = x^{k-1} - \tau_1 \nabla F(x^{k-1}) \\ y^k &= (I + \tau_2 x x^T)^{-1}(z + \tau_2 x^T w) \qquad \textit{where} \qquad z = y^{k-1} - \tau_2 \nabla G(y^{k-1}) \end{aligned}$$

When $F(x) = f(\|Ax\|)$ and $G(y) = g(\|By\|)$, one may check that

$$\nabla F(x) = \begin{cases} f'(\|Ax\|)\frac{A^*Ax}{\|Ax\|} & \text{if } x \notin \ker A \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \nabla G(y) = \begin{cases} g'(\|By\|)\frac{B^*By}{\|By\|} & \text{if } y \notin \ker B \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

The stepsizes $\tau_1 > 0$ and $\tau_2 > 0$ are chosen according to Assumption (τ) *i.e.* such that $\tau_1 < 2/L_{\nabla F}$ and $\tau_2 < 2/L_{\nabla G}$. These Lipschitz moduli have to be estimated. If f' and g' are resp. $L_{f'}$ and $L_{g'}$-smooth, then, if $x, x' \notin \ker A$,

$$\nabla F(x) - \nabla F(x') = \frac{f'(\|Ax\|)}{\|Ax\|}(A^*Ax - A^*Ax') + \frac{f'(\|Ax\|)}{\|Ax\|}\frac{A^*Ax'}{\|Ax'\|}(\|Ax'\| - \|Ax\|) + \big(f'(\|Ax\|) - f'(\|Ax'\|)\big)\frac{A^*Ax'}{\|Ax'\|}$$

Hence,

$$\begin{aligned} \|\nabla F(x) - \nabla F(x')\| &\leqslant L_{f'}|||A|||^2\|x - x'\| + L_{f'}|||A|||\,|\|Ax'\| - \|Ax\|| + L_{f'}|\|Ax'\| - \|Ax\||\,|||A||| \\ &\leqslant 3L_{f'}|||A|||^2\|x - x'\| \end{aligned}$$

since $\|Ax'\| - \|Ax\| \leqslant \|Ax' - Ax\|$. If $x \in \ker A$ and $x' \notin \ker A$, then

$$\|\nabla F(x) - \nabla F(x')\| = \|\frac{f'(\|Ax\|)}{\|Ax\|}A^*Ax'\| \leqslant L_{f'}\|A^*Ax'\| = L_{f'}\|A^*Ax - A^*Ax'\| \leqslant L_{f'}|||A|||^2\|x - x'\|$$

This proves that $L_{\nabla F}$ is bounded by $3L_{f'}|||A|||^2$. Similarly, $L_{\nabla G}$ is bounded by $3L_{g'}|||B|||^2$.

In the case of more general regularization terms as those in (26), one has

$$\nabla F(x) = \sum_{i=1}^{q_x} \nabla F_i(x) \quad \text{and} \quad \nabla G(y) = \sum_{j=1}^{q_y} \nabla G_j(y) \tag{29}$$

with $F_i(x) = f_i(\|A_i x\|)$ and $G_j(y) = g_j(\|B_j y\|)$ and $(\nabla F_i, \nabla G_j)$ given by (28).

5.2 Facts on J in (24) and Algorithm 2

Remark 5. When $\ker A \neq \{0\}$ (resp. $\ker B \neq \{0\}$), the objective J in (24) is not coercive, since for any (x, y) of the form

$$z_{(1)} := (u, 0_n) \quad u \in \ker F \backslash \{0\} \qquad \text{and} \qquad z_{(2)} := (0_n, v) \quad v \in \ker G \backslash \{0\}$$

$J(z_{(i)}) = J(tz_{(i)}) = \|w\|^2$ for $i \in \{1, 2\}$ and all $t > 0$. The result follows from the form of J in (24).

Lemma 5. *If* $\ker A = \{0\}$ *or* $\ker B = \{0\}$ *and* $\|w\|^2 > \inf J$*, then the sequence* $(x^k, y^k)_k$ *generated by Algorithm 1 is bounded if* (x^0, y^0) *is chosen such that* $J(x^0, y^0) < \|w\|^2$.

Proof. Just remark that if $\ker A = \{0\}$ or $\ker B = \{0\}$, then $(x, y) \in \ker A \times \ker B$ implies that $xy = 0$, thus $J(x, y) = \|w\|^2$. Then, apply Proposition 7. □

Lemma 5 shows that, when $\ker A = \{0\}$ or $\ker B = \{0\}$, it suffices to initialize Algorithm 1 with $J(x^0, y^0) < \|w\|^2$ to generate a bounded sequence $\{(x^k, y^k)\}_k$.

Proposition 8. *The constant* s *in* (25) *is well defined. It can be computed and an element* $(\bar{x}, \bar{y})$ *yielding* $J(\bar{x}, \bar{y}) = s$ *can be identified.*

Proof. Let $(x, y) \in \ker F \times \ker G$. Since the latter is a finite-dimensional vector subspace and using the form of J in (24), $J(x, y) = H(x, y)$ and

$$\inf_{(x,y) \in \ker F \times \ker G} J(x, y) = \inf_{(x,y) \in \ker F \times \ker G} H(x, y).$$

We denote

$$\begin{aligned} \ker F &= \text{span}\{u_{(1)}, \ldots, u_{(n_F)}\} \quad \text{where} \quad n_F := \dim \ker F \\ \ker G &= \text{span}\{v_{(1)}, \ldots, v_{(n_G)}\} \quad \text{where} \quad n_G := \dim \ker G \end{aligned} \tag{30}$$

Any $(x, y) \in \ker F \times \ker G$ has the form

$$(x, y) = \left(\sum_{i=1}^{n_F} a_i u_{(i)}, \sum_{j=1}^{n_G} b_j v_{(j)} \right), \quad \forall\, a \in \mathbb{R}^{n_F}, \forall\, b \in \mathbb{R}^{n_G}.$$

Then $H(x, y) = \dfrac{1}{2} \left\| \sum_{i=1}^{n_F} \sum_{j=1}^{n_G} a_i b_j u_{(i)} v_{(j)} - w \right\|^2$ has a minimum that is determined by the $n_F n_G$ numbers $c_{i,j} = a_i b_j$. We can hence write down that

$$s = \frac{1}{2} \min_{\{c_{i,j}\}} \left\| \sum_{i=1}^{n_F} \sum_{j=1}^{n_G} c_{i,j} u_{(i)} v_{(j)} - w \right\|^2.$$

The sum above is composed out of $n_F n_G$ known elements $u_{(i)} v_{(j)} \in W$ and the unknown $n_F n_G$ numbers $c_{i,j}$ solve a quadratic minimization problem. A set of coefficients $\bar{c}_{i,j}$ yielding the minimum value above can be computed. [2] One can identify all $\bar{a}_i$ and $\bar{b}_j$ yielding $\bar{a}_i = \bar{b}_j = \bar{c}_{i,j}$ and hence for $\bar{x} = \sum_{i=1}^{n_F} \bar{a}_i u_{(i)}$ and $\bar{y} = \sum_{j=1}^{n_G} \bar{b}_j v_{(j)}$ one has $J(\bar{x}, \bar{y}) = s$. □

[2] A solution $\{\bar{c}_{i,j}\}_{i=1,j=1}^{n_F,n_G}$ always exists. It is unique only if the set of all $u_{(i)} v_{(j)}$ is linearly independent.

Lemma 6. *Let $w \in W$. Assume that $n_F\, n_G < \dim W$. Then the subset $\mathcal{W}$*

$$\mathcal{W} := \left\{ z \in W \;\mid\; z = \sum_{i=1}^{n_F} \sum_{j=1}^{n_G} a_i b_j u_{(i)} v_{(j)} \quad \forall\, a_i, b_j \in \mathbb{R} \right\} \tag{31}$$

is closed in W and its Lebesgue measure in W is $\mathbb{L}^W(\mathcal{W}) = 0$.

Proof. $\mathcal{W}$ is a vector space spanned by all $u_{(i)} v_{(j)}$ where $1 \leqslant i \leqslant n_F$ and $1 \leqslant j \leqslant n_G$. Then

$$\dim \mathcal{W} \leqslant n_F n_G < \dim W,$$

hence the statement. □

We emphasize that $W \setminus \mathcal{W}$ is an open dense subset of W. We can always assume that data w satisfy $w \in W \setminus \mathcal{W}$. Since typical data w is noisy, the chance that a w following a non singular probability distribution comes across $\mathcal{W}$ can be ignored in practice.

The goal of the following claim is to show that for any w, except for some $w \in \mathcal{W}$ where $\mathcal{W}$ is closed and of Lebesgue measure zero, there are points (x, y) satisfying $J(x, y) < s$.

Proposition 9. *Let $w \in W \setminus \mathcal{W}$. Then there exist (x, y) and $\varepsilon > 0$ such that $\inf\limits_{(x,y) \in U \times V} J(x, y) \leqslant s - \varepsilon$.*

Proof. Let $(\bar{x}, \bar{y})$ be such that $\bar{z} := \bar{x}\bar{y} \perp \mathcal{W}$ and that $\langle \bar{z}, w \rangle > 0$. Using the notations in Proposition 8, set

$$x(c) := \sum_{i=1}^{n_F} \bar{a}_i u_{(i)} + c\bar{x} \quad \text{and} \quad y(c) := \sum_{j=1}^{n_G} \bar{b}_j v_{(j)} + c\bar{y}. \tag{32}$$

Then $J(x(c), y(c)) = f(c\|A\bar{x}\|) + g(c\|B\bar{y}\|) + H(x(c), y(c))$ and

$$\inf_{(x,y) \in U \times V} J(x, y) \leqslant f(c\|A\bar{x}\|) + g(c\|B\bar{y}\|) + \min_{c \in \mathbb{R}} H(x(c), y(c)) \quad \forall\, c. \tag{33}$$

Note that $u_{(i)} v_{(i)} \in \mathcal{W}$ for all (i, j). Thus using that $\bar{z} \perp \mathcal{W}$ one has

$$H(x(c), y(c)) = \frac{1}{2} \left\| \sum_{i=1}^{n_F} \sum_{j=1}^{n_G} \bar{a}_i \bar{b}_j u_{(i)} v_{(i)} + c^2 \bar{z} - w \right\|^2 = \frac{1}{2} \left\| \sum_{i=1}^{n_F} \sum_{j=1}^{n_G} \bar{a}_i \bar{b}_j u_{(i)} v_{(i)} - w \right\|^2 + \frac{1}{2} c^4 \|\bar{z}\|^2 - c^2 \langle \bar{z}, w \rangle$$

Recalling that $\langle \bar{z}, w \rangle > 0$, the optimal $\tilde{c}$ obeys [3] $\tilde{c}^2 = \dfrac{\langle \bar{z}, w \rangle}{\|\bar{z}\|^2}$ and thus [4]

$$\min_{c \in \mathbb{R}} H(x(c), y(c)) = H(x(\tilde{c}), y(\tilde{c})) = s - \frac{1}{2} \frac{\langle \bar{z}, w \rangle^2}{\|\bar{z}\|^2}. \tag{34}$$

From Assumption (f, g), there exists $\bar{c} > 0$ such that for any $c \in (0, \bar{c}]$ one has $f(c\|A\bar{x}\|) \leqslant \dfrac{1}{4} \left(\dfrac{1}{2} \dfrac{\langle \bar{z}, w \rangle^2}{\|\bar{z}\|^2} \right)$ and $g(c\|B\bar{y}\|) \leqslant \dfrac{1}{4} \left(\dfrac{1}{2} \dfrac{\langle \bar{z}, w \rangle^2}{\|\bar{z}\|^2} \right)$. Let us now consider that $(x(\bar{c}), y(\bar{c}))$ in (32). This, together with (33) and (34) leads to

$$\inf_{(x,y) \in U \times V} J(x, y) \leqslant s - \frac{1}{4} \frac{\langle \bar{z}, w \rangle^2}{\|\bar{z}\|^2}.$$

Thus $\varepsilon = \dfrac{1}{4} \dfrac{\langle \bar{z}, w \rangle^2}{\|\bar{z}\|^2}$ in the statement. □

If the optimal $\tilde{c}$ used in (34) is such that $\tilde{c} \leqslant \bar{c}$, then set $\bar{c} := \tilde{c}$ and the stated inequality will hold for a smaller positive ε. Otherwise, another $(\bar{x}, \bar{y})$ should be chosen so that $\tilde{c} \leqslant \bar{c}$. This will help to find an initial point (x^0, y^0) such that $J(x^0, y^0) < s$ in order to guarantee the boundedness of the iterates.

Now we can state a result for J in (24) that reinforces Proposition 7.

[3] Differentiating wrt c and setting to zero yields $c^2 \|\bar{z}\|^2 = \langle \bar{z}, w \rangle$.

[4] One has $\dfrac{1}{2} c^4 \|\bar{z}\|^2 - c^2 \langle \bar{z}, w \rangle = \dfrac{1}{2} \dfrac{\langle \bar{z}, w \rangle^2}{\|\bar{z}\|^4} \|\bar{z}\|^2 - \dfrac{\langle \bar{z}, w \rangle}{\|\bar{z}\|^2} \langle \bar{z}, w \rangle = -\dfrac{1}{2} \dfrac{\langle \bar{z}, w \rangle^2}{\|\bar{z}\|^2}$

Theorem 1. *Suppose that f and g obey Assumption (f,g). Then J has a global minimizer $(\widehat{x},\widehat{y})$ satisfying $\widehat{J} := J(\widehat{x},\widehat{y}) < s$ and for any $r \in \left[\widehat{J}, s\right)$ the set $\operatorname{lev}_{\leqslant r} J$ is compact and nonempty.*

Proof. Let $r < s$. The set $\operatorname{lev}_{\leqslant r} J$ is closed since J is continuous. Suppose that $\operatorname{lev}_{\leqslant r} J$ is unbounded. Then there exist $(x^k, y^k)_k \in \operatorname{lev}_{\leqslant r} J$ such that $\|(x^k, y^k)\| \to +\infty$. Since $J(x^k, y^k) < s$, then $(x^k, y^k) \notin \ker F \times \ker G$. Hence, by Assumption (f,g) (b), there is K such that $F(x^k) \geqslant s$ or $G(y^k) \geqslant s$ for all $k \geqslant K$. Since $J(x^k, y^l) \geqslant F(x^k) + G(y^k)$, we have that $J(x^k, y^k) \geqslant s$, a contradiction to $J(x^k, y^k) < s$. Thus $\operatorname{lev}_{\leqslant r} J$ is bounded. Therefore, for any $r < s$, the set $\operatorname{lev}_{\leqslant r} J$ is compact.

Now from Proposition 9 we know that $\inf J < s - \varepsilon$ where $\varepsilon > 0$. It follows that J has a global minimizer $(\widehat{x},\widehat{y})$ satisfying $\widehat{J} := J(\widehat{x},\widehat{y}) \leqslant s - \varepsilon$ and $(\widehat{x},\widehat{y}) \in \operatorname{lev}_{\leqslant s-\varepsilon} J$. Noticing that $\inf J = \widehat{J}$, it follows that that $\operatorname{lev}_{\leqslant r} J \neq \varnothing$ for any $r \geqslant \widehat{J}$. □

5.3 Convergence when F and G are convex

Even if F and G are convex and coercive, the objective J is nonconvex and non-coercive. Nonetheless, some strong-convexity-like properties hold, due to the biconvex structure of H.

Lemma 7. *Let F and G are convex and $x^+ \in U$ and $y^+ \in V$. Then*

$$\forall\ (x,y) \in U \times V \quad J(x,y) \geqslant J(x^+,y) + \langle x - x^+, \nabla_x J(x^+,y)\rangle + \frac{1}{2}\|(x-x^+)y\|^2 \tag{35}$$

$$\forall\ (x,y) \in U \times V \quad J(x,y) \geqslant J(x,y^+) + \langle y - y^+, \nabla_y J(x,y^+)\rangle + \frac{1}{2}\|x(y-y^+)\|^2 \tag{36}$$

Proof. Let us prove (35). The convexity of F yields

$$\forall\ (x,y) \quad F(x) + G(y) \geqslant F(x^+) + G(y) + \langle x - x^+, \nabla F(x^+)\rangle.$$

We notice that $\nabla_x H(x^+,y) = (x^+y - w)y^{\mathrm{T}}$ so by expanding $H(x,y)$ one obtains

$$\begin{aligned} H(x,y) = \frac{1}{2}\|x^+y - w + (x-x^+)y\|^2 &= H(x^+,y) + \langle x^+y - w, (x-x^+)y\rangle + \frac{1}{2}\|(x-x^+)y\|^2 \\ &= H(x^+,y) + \langle x - x^+, \nabla_x H(x^+,y)\rangle + \frac{1}{2}\|(x-x^+)y\|^2. \end{aligned}$$

Combining the last equality with the previous inequality proves (35). □

Lemma 8. *Let (x^k, y^k) be generated by Algorithm in subsection 5.1, with limit point (x^*, y^*). Let Assumptions (J) and (τ) hold with F and G convex. If x^* has full column rank, then y^k converges to y^*.*

Proof. From Proposition 4, (x^k) has a convergent subsequence $\{x^{k_j}\}$, of limit point x^*. Lemma 3 applied to $f = G$, $h = H(x^{k_j}, \cdot)$, and $x^+ = y^{k_j}$ yields

$$\forall\ y \quad \left(\frac{1}{\tau_2} - \frac{L_{\nabla G}}{2}\right)\|y^{k_j} - y\|^2 \leqslant J(x^{k_j}, y) - J(x^{k_j}, y^{k_j}).$$

Using Proposition 2 and the continuity of J, we can take the limit in the right-hand side of the previous inequality, which converges to $J(x^*, y) - J^*$, hence it has limit points. Suppose that it has two limit points $\hat{y}$ and $\tilde{y}$. Thus, $(x^*, \hat{y})$ and $(x^*, \tilde{y})$ are limit points of (x^k, y^k). Then using Propositions 2 and 4, we have $J(x^*, \hat{y}) = J^* = J(x^*, \tilde{y})$ and $\nabla_y J(x^*, \hat{y}) = 0 = \nabla_y J(x^*, \tilde{y})$. Now apply (36) in Lemma 7 to $x = x^*$, $y = \hat{y}$ and $y_0 = \tilde{y}$:

$$J^* \geqslant J^* + \frac{\mu}{2}\|x^*(\hat{y} - \tilde{y})\|^2$$

that is, $x^*(\hat{y} - \tilde{y}) = 0$. Since x^* has full column rank, this implies that $\hat{y} = \tilde{y}$. In other terms, the sequence y^{k_j} has a unique limit point. □

6 Hadamard based biconvex term and application

We consider objective functions of the form

$$J(x,y) = F(x) + G(y) + H(x,y) = f(\|Ax\|) + g(\|Bx\|) + \frac{1}{2}\|x \circ y - w\|^2 \tag{37}$$

with $x \in U = \mathbb{R}^{m\times n}$, $y \in V = \mathbb{R}^{m\times n}$ and $w \in W = \mathbb{R}^{m\times n}$ (data or perturbation), and where $\circ$ stands for the Hadamard product. All results in section 5 hold if we replace matrix products by element-wise multiplication and matrix inverses by element-wise inverses, since $x \circ y = \mathrm{diag}(x)y$. Then

$$\nabla_x J(x,y) = \mathrm{diag}\left((x \circ y - w) \circ y\right) + \nabla F(x) \qquad \nabla_y J(x,y) = \mathrm{diag}\left(x \circ (x \circ y - w)\right) + \nabla G(y)$$

and the first order optimality conditions are

$$\mathrm{diag}(\tilde{x} \circ \tilde{y} \circ \tilde{y}) + \nabla F(\tilde{x}) = \mathrm{diag}(w \circ \tilde{y}) \qquad \text{and} \qquad \mathrm{diag}(\tilde{x} \circ \tilde{x} \circ \tilde{y}) + \nabla G(\tilde{y}) = \mathrm{diag}(w \circ \tilde{x}) \tag{38}$$

Corollary 1. *Let $w \in W \setminus \mathcal{W}$ satisfy $w_i \neq 0$ for any i. Then each critical point $(\tilde{x}, \tilde{y})$ of J obeys $\tilde{x} \notin \ker F$ and $\tilde{y} \notin \ker G$.*

Proof. Let $u \in \ker F$. The optimality condition (38) for $\tilde{x} = u$ yields $u_i \tilde{y}_i^2 = w_i \tilde{y}_i \quad \forall\, i$. Then, noticing that $w_i \neq 0$, one has

$$\begin{cases} \tilde{y}_i = 0 & \text{if} \quad u_i = 0 \\ \tilde{y}_i = \dfrac{w_i}{u_i} & \text{if} \quad u_i \neq 0 \end{cases}$$

Suppose that $\tilde{y} = v \in \ker G$, hence $w = uv \quad \forall\, u \in \ker F \quad \forall\, v \in \ker G$ and thus $w \in \mathcal{W}$. □

Then the algorithm in subsection 5.1 reads as:

Algorithm 3 (ABC-PFB with Hadamard product)**.**
Initialization: (x^0, y^0) *and* $\tau_1 > 0$, $\tau_2 > 0$
Iterations: for $k \geqslant 0$

$$\begin{aligned} x^k &= \frac{z + \tau_1 y^{k-1} \circ w}{1 + \tau_1 y^{k-1} \circ y^{k-1}} \qquad \textit{where} \qquad z = x^{k-1} - \tau_1 \nabla F(x^{k-1}) \\ y^k &= \frac{z + \tau_2 x^k \circ w}{1 + \tau_2 x^k \circ x^k} \qquad \textit{where} \qquad z = y^{k-1} - \tau_2 \nabla G(y^{k-1}) \end{aligned}$$

where division is componentwise.

Once again, the gradients are given by (28) or (29), and the stepsizes $\tau_1 > 0$ and $\tau_2 > 0$ have to satisfy Assumption (τ) *i.e.* chosen such that $\tau_1 < 2/L_{\nabla F}$ and $\tau_2 < 2/L_{\nabla G}$.

7 Acknowledgement

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

References

[1] R. Abergel and L. Moisan, *The shannon total variation*, hal-01349516, (2016).

[2] C. Aguerrebere, A. Almansa, Y. Gousseau, and P. Musé, *A hyperprior bayesian approach for solving image inverse problems*, HAL, (2016).

[3] F. A. Al-Khayyal and J. E. Falk, *Jointly constrained biconvex programming*, Math. Oper. Res., 8 (1983), pp. 273–286.

[4] 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.

[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*, Mathematics of Operations Research, 35 (2010), pp. 438–457.

[6] 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.

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

[8] A. Benichoux, E. Vincent, and R. Gribonval, *A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors*, in Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, vol. 1, May 2013, pp. 6108–6112.

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

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

[11] J. F. Cardoso, *Blind signal separation: Statistical principles*, Proceedings of the IEEE, 86 (1998), pp. 2009–2025.

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

[13] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations*, Wiley, Tokyo, 2009.

[14] K. Goh, L. Turan, M. Safonov, G. Papavassilopoulos, and J. Ly, *Biaffine matrix inequality properties and computational methods*, in Proceedings of the American Control Conference, June 1994, pp. 850–855.

[15] G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

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

[17] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bulletin de la Société mathématique de France, 93 (1965), pp. 273–299.

[18] M. Nikolova, *Energy minimzation methods*, Handbook of Mathematical Methods in Imaging, Editor: O. Scherzer, Springer, 2 ed., 2015.

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

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

[21] H. Q. Tuyen and L. D. Muu, *Biconvex programming approach to optimization over the weakly effcient set of a multiple objective affne fractional problem*, Oper Res Lett, 28 (2001).

[22] 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.