Inertial Alternating Generalized Forward-Backward Splitting for Image Colorization

Pauline Tan, Fabien Pierre, Mila Nikolova

Inertial Alternating Generalized Forward-Backward Splitting for Image Colorization

Pauline Tan[1], CMLA, CNRS, ENS Paris-Saclay, 94235 Cachan, France
Fabien Pierre[2], Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
Mila Nikolova[3], CMLA, CNRS, ENS Paris-Saclay, 94235 Cachan, France

Abstract

In this paper, we propose a novel accelerated alternating optimization scheme to solve block-biconvex nonsmooth problems whose objectives can be split into smooth (separable) regularizers and simple coupling terms. The proposed method performs a Bregman distance based generalization of the well-known forward-backward splitting for each block, along with an inertial strategy which aims at getting empirical acceleration. We discuss the theoretical convergence of the proposed scheme and provide numerical experiments on image colorization.

1 Introduction

1.1 Image Colorization

In the previous decade, the number of papers dealing with the challenging problem of image colorization has literally exploded. This technique consists in transforming a gray-scale image (also known as black-and-white image) into a color one by addition of a chromatic information. This topic has many opportunities in the entertainment industry with the success of the recolorized historical documentaries. The challenge of this technique comes from the fact that turning a color into a gray level is not an invertible operation. Indeed the color space has three dimensions whereas the gray level space has only one. In this way, turning a gray-scale image into a color one requires additional information.

In the literature, there are two ways to add color information to a gray scale image. First, the manual approaches require some color strokes given by a user, which are then propagated over the whole image with diffusion techniques guided by the gray-scale channel (see, e.g., [29, 46, 26]). The major drawback of this approach is the huge amount of work needed from the user, especially when the scene represented on the image is complex or contains a lot of textured areas.

To tackle this issue, the exemplar-based methods use a color image as a reference and transfer the color to the gray-scale one based on texture criterion (see, e.g., [44, 27, 28, 25, 24, 17, 16]), morphological mappings [33] or based on deep learning approach [47]. Obviously, the choice of the reference image is a first issue and influences the results. To take into account the various aspects of the textures for a reliable comparison, it is common in the literature to use many descriptors to transfer the colors between images [44, 38]. The main difficulty is the choice of the descriptors for textures, as well as the distance between the image patches in the case of patch-based approach. This choice can been done experimentally [44] or with metric learning approach [36].

[1]E-mail: pauline.tan@cmla.ens-cachan.fr
[2]E-mail: fabien.pierre@inria.fr
[3]E-mail: nikolova@cmla.ens-cachan.fr

The alternative consists in the computation of multiple candidates followed by a method to select the final result. These candidates can be extracted from a patch-based approach [11] or based on a learning approach [47]. Some state-of-the art methods perform the choice among these candidates without taking into account the geometrical information of the result. For instance, the authors of [11] use a median to choose one candidate among the ones obtained by their patch based approach. As another example, the authors of [47] use an annealing mean to compute the final result from the different results given by a Convolutional Neural Network (CNN).

To choose a candidate and to enhance the regularity of the result, a first approach is the method of [16] which is based on a graph-cut approach to select the final candidate. The authors of [12] have proposed a variational method to choose one candidate among given ones with a regularity hypothesis. This method has been improved then for image colorization by the authors of [34]. The aim of these last two models is to choose, for each pixel of the image, a candidate such that the total variation of the resulting image is minimized.

This kind of approach based on a multiple criterion candidate extraction, followed by a total variation regularized choice, can be applied to other problems. For instance, it can be useful for optical flow estimation [20, 21] or for video colorization [35].

These problems are usually nonconvex and often nonsmooth. To solve them, authors of [34] have proposed a primal-dual algorithm to make their problem numerically tractable but without convergence proof. Indeed, the variational model used in [34] is *biconvex*, that is, the functional depends on two block-variables, with convexity in each block. This partial convexity feature allows to apply schemes encountered in convex optimization which alternate partial (block) optimizations, since each subproblem remains convex. However, for nonconvex functionals, no convergence is guaranteed.

1.2 Biconvex Optimization

Despite some notable progress in the field of continuous nonconvex optimization, the minimization of a biconvex functional under separable biconvex constraints

$$\min_{(x,y)\in C_{\mathcal{X}}\times C_{\mathcal{Y}}} J(x,y) \qquad \text{with} \qquad C_{\mathcal{X}}, C_{\mathcal{Y}} \text{ convex, } J \text{ biconvex}$$

remains a difficult task. The partial convexity carried by the biconvex structure of the functional yet allows applying convex optimization algorithm in a block-optimization framework, but convergence may be hard to prove. Moreover, unlike in the convex case, such schemes can usually be guaranteed to converge to a critical point, which may not be a (even local) minimizer.

In 2013, [45] and [8] proposed an alternating proximal scheme (known as the PALM method in [8]) which aims at solving nonsmooth and nonconvex optimization problems, by alternating the well-known forward-backward splitting on each partial problem. They provided a convergence analysis based on the Kurdyka-Łojasiewicz property [9, 7], and the proposed scheme presents in practice similar numerical performance as when used for convex optimization. However, the application hypotheses of this algorithm only allow to consider *simple* regularizers[4], in the sense that their Moreau proximity operator [30]

$$\operatorname{prox}_F(x^0) := \arg\min_{x\in\mathcal{X}} \left\{ F(x) + \frac{1}{2}\|x - x^0\|^2 \right\}$$

must have a closed form for any $x^0 \in \mathcal{X}$, or be exactly computable in a reasonable time. In practice, this limits the class of functionals which can be optimized by this scheme. In the case where the regularizers are not simple but sufficiently smooth (or replaced

[4] In this paper, we will refer to functions which only depend on one variable x or y as *regularizers*, while those which depends on both (x, y) are called *coupling terms*. Obviously, the additive decomposition of a functional into regularizers and coupling terms is not unique, since one always may add regularizers to a coupling term.

by a smooth approximation), they can be incorporated in the coupling term, but may lead to smaller stepsizes (hence, slower convergence). Thus, in 2017, the authors in [31] proposed another alternating scheme, which is basically a mirror of PALM, in the sense that the forward and the backward steps are not applied on the same part of the functional. This led to the ASAP algorithm, whose convergence has been studied in details in [31] for a large class of nonsmooth and nonconvex functional (among others biconvex functionals). Like the PALM method, the ASAP algorithm yet presents two main drawbacks: the simplicity hypothesis for the functional (to ensure that the backward step is computable), and the convergence rate, which is empirically in $O(1/K)$ (with K the number of iterations).

A way to relax this assumption is to consider more general proximity operators, typically by using the so-called Bregman distances [10, 14]. Indeed, some functions may not be simple with respect to the usual Moreau proximity operator, but simple to a more general one defined thanks to an adapted Bregman distance. We will show in this paper that such a generalization can be carried out in the framework introduced in [31]. Besides, the convergence speed issue has already been successfully tackled in the convex optimization by using inertia [40, 23, 5, 15, 22] and for some nonconvex proximal schemes [32, 39]. We will show in this paper that ASAP and its Bregman-distance based generalization can also incorporate such a speed-up strategy, which leads to our proposed method. Numerical experiments will prove that an acceleration can be observed.

Apart from the colorization problem studied in this paper, many other applications can be concerned by the proposed accelerated scheme. Indeed, most of the joint estimation problems (*i.e.* the estimation of two or more objects during the same process) increasingly encountered in image processing result from the combination of the variational model of each variable, which may usually be chosen convex, so that the resulting functional is nonconvex but still biconvex (or multiconvex), see *e.g.* [2, 43, 13].

The paper is organized as follows: in the next section, we present the colorization method proposed in [34] and the associated variational model. In particular, we analyze its regularity. Then, in Section 3, we recall some mathematical preliminaries, mainly about subdifferential calculus and Bregman generalization of the forward-backward splitting. In Section 4, we introduce the general form of our proposed algorithm and discuss its convergence properties for large classes of nonconvex nonsmooth functions. The parameter choice is discussed in Section 5. Eventually, in Section 6, we present the numerical application of our method for the image colorization problem, and compared its performances with three other methods.

2 Variational Approach for Image Colorization

2.1 Variational Model

The model of [34], implemented in [37], settles on a multiple candidates selection strategy. The model is written in the YUV color-space, and the target image (the image to colorize) is considered to be the luminance (Y) channel of the expected color image. Since they carry the color information, the aim is to compute the U and V channels of this image from the reference (or source) image.

The method proposed in [34] estimates the U and V channels in a two-step procedure. In the first step, for each pixel of the target image, C color candidates are selected from the source image *via* a *candidate extraction method* (see Section 6). Then, a *voting process* is run, to select one candidate among the C candidates extracted at the previous step. Once the selection is done for each pixel, the resulting U and V channels are combined with the luminance channel to get the expected color image.

In [34], the voting process is done in a variational framework. For any pixel $x \in \Omega$, given C chrominance candidates, denoted by $c_i(x)$, the choice of a specific candidate is modelized thanks to a vector $w(x) \in \mathbb{R}^C$ belonging to the simplex Σ_C which represents the distribution of the votes (or the likelihood) for each candidate. Thus, at the end of

the selection process, one retains the candidate $c_i(x)$ which obtained most votes, namely such as $w_i(x)$ is the largest coefficient of $w(x)$. In [34], the vote vector (and thus, its largest value) is chosen so that the resulting color image is feasible (*i.e.* within a given range $\mathcal{R} := \{u \mid u(x) = (U(x), V(x)) \in [U_{\min}, U_{\max}] \times [V_{\min}, V_{\max}]\}$) and has a small total variation. Thus, it has been proposed to minimize the following variational model:

$$J(u, w) \quad = \quad \mathrm{TV}_{\mathfrak{C}}(u) + \chi_{\mathcal{R}}(u) + \chi_{\Sigma}(w) + \frac{\lambda}{2} \int_{\Omega} \sum_{i=1}^{C} w_i(x) \|u(x) - c_i(x)\|^2 \, \mathrm{d}x. \quad (1)$$

In this model, $u = (U, V)$ represents the chrominance channels of the sought-after color image and x is the pixel position in the image domain Ω. $\lambda > 0$ is a scalar parameter which controls the regularization of the results. The *coupled* total variation $\mathrm{TV}_{\mathfrak{C}}(u)$ is defined as follows:

$$\mathrm{TV}_{\mathfrak{C}}(u) = \int_{\Omega} \sqrt{\gamma \|\nabla Y\|^2 + \|\nabla u\|^2} \quad (2)$$

and can be seen as the smoothed by $\gamma\|\nabla Y\|^2$ total variation of the color image in the YUV space, with weight $\gamma > 0$ for the luminance channel. As shown is [34], this term enforces the correlation between the (given) luminance variations and the (estimated) chrominances variations. Thus, the optimization of the model in (1) yields the joint estimation of both the weight (votes) w and the chrominance of the sought-after color image.

As underlined in the introduction, the main difficulty of the selection process is the optimization of (1). Indeed, it is obvious that the constraints are bounded, convex and separable, the $\mathrm{TV}_{\mathfrak{C}}$ regularizer is convex, but the coupling term is solely biconvex. Thus, the whole functional J is biconvex, and in particular nonconvex.

2.2 Regularity of the Problem

Biconvex optimization can be handled by various algorithms. Among them, two methods are noteworthy, namely the PALM method and the ASAP algorithm. The application of these two methods requires simplicity and smoothness conditions of the functional J. The critical conditions in the studied problem (1) are the smoothness of the coupled total variation and the simplicity and/or smoothness of the coupling term.

A simple way to ensure the smoothness of the coupled total variation regularizer for any grayscale image Y is to introduce a threshold $\alpha > 0$ for the weighted gradient of the luminance Y, namely to replace $\gamma\|\nabla Y\|^2$ by $\max\{\alpha, \gamma\|\nabla Y\|^2\}$ in (2), so that the resulting regularized function $\mathrm{TV}^{\alpha}_{\mathfrak{C}}$ has a Lipschitz continuous gradient given by

$$\nabla \mathrm{TV}^{\alpha}_{\mathfrak{C}}(u) = -\frac{\mathrm{div}(\nabla u)}{\sqrt{\max\{\gamma\|\nabla Y\|^2, \alpha\} + \|\nabla u\|^2}}$$

with modulus $\mathrm{L}_{\nabla F} \leq |||\nabla|||^2/\sqrt{\alpha}$.

In practice, given that, except in completely flat areas in the image (which is unlikely for real noisy images), ∇Y is expected to be numerically nonnull, if α is chosen small enough, one has the identity $\mathrm{TV}^{\alpha}_{\mathfrak{C}}(u) = \mathrm{TV}_{\mathfrak{C}}(u)$ for any u. Otherwise said, for proper choices of α, this shows that for real images, the coupled total variation is smooth with a Lipschitz gradient, with a modulus bounded thanks to α.

The remaining terms – the convex constraints and the coupling term – can be easily handled with standard implicit or explicit numerical schemes. Thus, both ASAP and PALM are directly relevant for this problem. However, it is noteworthy that the computation of the proximity operator which involves a simplex constraint may be time-consuming, but can also be cleverly handled by generalized Bregman proximity operators.

3 Mathematical Preliminaries

3.1 Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ be real finite-dimensional spaces. The ith element of a vector or a matrix x (seen as a vector) reads as x_i. For an $m \times n$ real matrix w we denote

$$\|w\| := \|w\|_F = \sqrt{\sum_{i,j} w_{i,j}^2}\,, \tag{3}$$

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 $\mathcal{S} \subset \mathcal{X}$, the distance of any point $x^+ \in \mathcal{X}$ to $\mathcal{S}$ is defined by

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

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

3.2 Subdifferential and Partial Subdifferentials

Let us first recall the definition of the subdifferential of a convex function.

Définition 1 *Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper[5], convex, lower semicontinuous (l.s.c.) function and $x^+ \in \operatorname{dom} f$. The subdifferential $\partial f(x^+)$ of f at x^+ is the set of all $p \in \mathbb{R}^m$, called subgradients of f at x^+, such that*

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

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

Its main interest is to characterize the minimizers x^* (when they exist) of f, *via* the well-known Fermat's rule $0 \in \partial f(x^*)$ (which is the generalization of the first-order optimality condition for smooth functions $0 = \nabla f(x^*)$). To extend this rule to nonconvex nonsmooth function (Proposition 1), we introduce the following extensions of the subdifferential:

Définition 2 *[41, Def. 8.3] Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper function and let $x^+ \in \operatorname{dom} f$.*

1. *The* Fréchet subdifferential $\widehat{\partial} f(x^+)$ *of f at x^+ is the set of all $p \in \mathbb{R}^m$ such that*

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

2. *The* (limiting-)subdifferential $\partial f(x^+)$ *of f at x^+ is the set of all $p \in \mathbb{R}^m$ such that there exist a sequence $\{x^k\}_{k\in\mathbb{N}} \in (\mathbb{R}^m)^{\mathbb{N}}$ converging to x^+ and a sequence $\{p^k\}_{k\in\mathbb{N}} \in (\mathbb{R}^m)^{\mathbb{N}}$ converging to p satisfying*

$$\forall\, k \in \mathbb{N}, p^k \in \widehat{\partial} f(x^k) \quad \textit{and} \qquad f(x^k) \to f(x^+).$$

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

Remark 1 If f is convex, then the definitions of the subdifferential in Definitions 1 and 2 coincide, and $\widehat{\partial} f(x) = \partial f(x)$ for any $x \in \operatorname{dom} f$. If f is continuously differentiable at x, one has $\partial f(x) = \{\nabla f(x)\}$.

As for smooth / convex functions, the subdifferential gives a necessary first-order optimality condition, known as Fermat's rule.

Proposition 1 *[41, Theorem 10.1] Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper function. If f has a local minimum at x^*, then $0 \in \partial f(x^*)$. Such points are called* critical points, *and the set of the critical points of f is denoted by $\operatorname{crit} f$.*

Given a proper function $h : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $(x^+, y^+) \in \operatorname{dom} h$, we define its partial subdifferentials $\partial_x h(x^+, y^+)$ and $\partial_y h(x^+, y^+)$ at (x^+, y^+) respectively as the subdifferentials of the partial functions $x \mapsto h(x, y^+)$ at x^+ and $y \mapsto h(x^+, y)$ at y^+. Unlike the differentiable case, the links between the partial subdifferentials and the subdifferential may be quite complicated. In particular, in general, one has

$$\partial h(x, y) \neq \partial_x h(x, y) \times \partial_y h(x, y).$$

As explored in details in [31], this may make alternating schemes such as the proposed method fail to find critical points. Another issue which may arise when dealing with such schemes in the nonsmooth and nonconvex case is the following one. The notion of limiting subdifferential is introduced because it brings more robustness than that of the Fréchet subdifferential, in the sense that it satisfies the following *closedness* property: if $\{x^k\}_{k\in\mathbb{N}}$ in $\operatorname{dom} f$ converges to $x^* \in \operatorname{dom} f$ such that $f(x^k) \to f(x^*)$ and $p^k \to p^*$ with $p^k \in \partial f(x^k)$ for any $k \in \mathbb{N}$, then one has $p^* \in \partial f(x^*)$. Such a closedness property is usually necessary to preserve, at convergence, useful property for sequences generated by first-order optimization schemes. However, as shown further in the convergence analysis, the proposed scheme may need a stronger closedness property on the *partial* subdifferentials.

Proposition 2 *Let $h : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper, biconvex function continuous on its domain. Then both its x- and y-partial subdifferential are parametrically closed at any point $(x^*, y^*) \in \operatorname{dom} h$, that is, for any sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ which converges to (x^*, y^*) such that $h(x^k, y^k) \to h(x^*, y^*)$ and for any $p_x^k \to p_x$ (resp. $p_y^k \to p_y$) with $p_x^k \in \partial_x h(x^k, y^k)$ (resp. $p_y^k \in \partial_y h(x^k, y^k)$) for any $k \in \mathbb{N}$, one has $p_x \in \partial_x h(x^*, y^*)$ (resp. $p_y \in \partial_y h(x^*, y^*)$).*

Proof. Let $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence as in Proposition 2. Let $k \in \mathbb{N}$. Since $x \mapsto h(x, y^k)$ is a convex, l.s.c. and proper functions, the subgradient inequality (Definition 1) leads to

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

If we let $k \to +\infty$ in the inequality above, we get by continuity

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

which means that p_x is a subgradient of the convex function $x \mapsto h(x, y^*)$ at x^*. Similar computations for the convex function $y \mapsto h(x, y)$ completes the proof. ■

Note that, unlike the closedness of the subdifferential, the parametric closedness of the partial subdifferentials may not hold for general nonconvex functions, even for quite smooth functions (consider for instance $h(x, y) = \sqrt{xy} + \chi_{\mathbb{R}^+}(x) + \chi_{\mathbb{R}^+}(y)$ at $(0, 0)$).

3.3 Bregman Distances

Proximal methods have brought notable improvement in the field of continuous optimization, but the main issue is their applicability for general problems, due to the noncomputability of many proximity operators. To deal with this problem, we consider in this paper a generalization of the Moreau proximity operator, which aims at making this operator computable in many cases, as shown in Section 6.

Définition 3 (Bregman distance[6] [10]) *Let $b : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a strictly convex function and differentiable on* $\operatorname{int}(\operatorname{dom} b)$. *The* Bregman "distance" associated to b *is defined for any $(x, y) \in \mathbb{R}^m \times \mathbb{R}^m$ by*

$$D_b(x,y) := \begin{cases} b(y) - b(x) - \langle \nabla b(x), y - x \rangle & \text{if } (x,y) \in \operatorname{int}(\operatorname{dom} b) \times \operatorname{dom} b \\ +\infty & \text{otherwise.} \end{cases}$$

Remark 2 Although we refer to D_b as a distance, it does not satisfy the definition of a distance. Indeed, it does not fulfill the symmetry condition, since $D_b(x, y) \neq D_b(y, x)$ in general. However, the other properties hold. Since b is convex, $D_b(x, y)$ is nonnegative for any $(x, y) \in \operatorname{int}(\operatorname{dom} b) \times \operatorname{dom} b$. The strict convexity of C also guarantees that $D_b(x, y) > 0$ for $x \neq y$.

Remark 3 For any $x \in \operatorname{int}(\operatorname{dom} b)$, the function $y \mapsto D_b(x, y)$ is convex and continuous on $\operatorname{dom} b$.

Example 1 Let us give some examples.

(a) Let M be a definite positive matrix, and let $\|x\|_M^2 = \langle x, Mx \rangle$. Then, $D_{\|\cdot\|_M}(x, y) = \|x - y\|_M^2/2$ is the Bregman distance associated to $\|\cdot\|_M$. In particular, if M is the identity, one recovers the Euclidean squared distance.

(b) For any $m \in \mathbb{N}^*$, let Σ_m denotes the simplex of dimension m, defined by

$$\Sigma_m := \left\{ x = (x_i) \in \mathbb{R}^m \mid \sum_{i=1}^m x_i = 1 \text{ and } x_i \geq 0 \right\}.$$

We also consider the entropy function, given by

$$b_e(x) := \begin{cases} \displaystyle\sum_{i=1}^m x_i \log x_i & \text{if } x \in \Sigma_m \\ +\infty & \text{otherwise,} \end{cases}$$

where $0 \times \log 0 = 0$ by convention. Then, the Bregman distance associated to the entropy function is explicitly given for $(x, y) \in (\operatorname{int}\Sigma_m) \times \Sigma_m$ by

$$D_{b_e}(x,y) = \sum_{i=1}^m y_i \left(\log y_i - \log x_i \right).$$

Proposition 3 *Let $b : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a strongly convex function of modulus L, differentiable on* $\operatorname{int}(\operatorname{dom} b)$. *Then for any $(x, y) \in \mathbb{R}^m \times \mathbb{R}^m$, one has the following lower bound:*

$$D_b(x,y) \geq \frac{L}{2} \|x - y\|^2. \tag{4}$$

Proof. The proof is a direct consequence of the strong convexity. ■

The following lemma will be further used in proofs.

Lemma 1 ([4]) *Let $b : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a strictly convex function and differentiable on $\operatorname{int}(\operatorname{dom} b)$. Then, one has the following identity for any three points $x, y \in \operatorname{int}(\operatorname{dom} b)$ and $z \in \operatorname{dom} b$:*

$$D_b(x, z) + D_b(y, x) - D_b(y, z) = \langle z - x, \nabla b(y) - \nabla b(x) \rangle.$$

3.4 Bregman Proximal Gradient Descent

In this section, we will introduce a generalization of well-known Forward-Backward Splitting (FBS) based on a generalization of the Moreau proximity operator that uses the Bregman distances introduced above. This generalization was initially introduced in [14].

Définition 4 *Let $b : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a strictly convex function and differentiable on $\operatorname{int}(\operatorname{dom} b)$. Let $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper function and let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function. Then we define for any $(u, v) \in \operatorname{int}(\operatorname{dom} b) \times \mathbb{R}^m$*

$$\mathcal{P}_h^b(u; \nabla f(v)) := \arg\min_{x \in \mathbb{R}^m} \Big\{ h(x) + \langle x - u, \nabla f(v) \rangle + D_b(u, x) \Big\}. \tag{5}$$

When $b = \|\cdot\|$ is the Frobenius norm, the optimization problem (5) when $u = v$ is equivalent to the FBS or proximal gradient descent applied to $h + f$.

Remark 4 If h is a proper convex function and if b is strongly convex, $\mathcal{P}_h^b(u; \nabla f(v))$ is nonempty. Indeed, the function involved in (5) is convex and lowerbounded by a strongly convex function. Hence, it is coercive and thus admits at least a minimizer on its domain.

The following proposition is a descent-like property for the Bregman FBS.

Proposition 4 *Let $b : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a strictly convex function and differentiable on $\operatorname{int}(\operatorname{dom} b)$. Let $h : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function and let $f : \mathbb{R}^m \to \mathbb{R}$ be a continuously differentiable function, with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient. Then, for any $u \in \operatorname{int}(\operatorname{dom} b)$ and any $\tau > 0$, if $x^+ \in \mathcal{P}_{\tau h}^b(u; \tau \nabla f(v))$, one has for any $x \in \mathbb{R}^m$*

$$\begin{aligned} &h(x^+) + f(x^+) + \frac{1}{\tau}\left(D_b(u, x^+) + D_b(x^+, x)\right) \\ &\leq h(x) + f(x) + \frac{1}{\tau} D_b(u, x) + \frac{\mathrm{L}_{\nabla f} + s}{2} \|x - x^+\|^2 + \frac{\mathrm{L}_{\nabla f}^2}{2s} \|x - v\|^2. \end{aligned}$$

Before giving the proof of Proposition 4, let us recall two useful lemmas.

Lemma 2 (Descent lemma [6]) *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function with $\mathrm{L}_{\nabla f}$-Lipschitz continuous gradient. Then, for any $(x, u) \in \mathbb{R}^m \times \mathbb{R}^m$, one has*

$$f(x) \geq f(u) - \langle u - x, \nabla f(x) \rangle - \frac{\mathrm{L}_{\nabla f}}{2} \|x - u\|^2. \tag{6}$$

Lemma 3 (Young inequality) *For any $s > 0$,*

$$\langle x, y \rangle \leq \frac{s}{2} \|x\|^2 + \frac{1}{2s} \|y\|^2.$$

Proof of Proposition 4. Fermat's rule for (5) yields

$$-\nabla f(v) - \frac{\nabla b(x^+) - \nabla b(u)}{\tau} \in \partial h(x^+).$$

By definition of the subgradient of a convex function, one has for any $x \in \mathbb{R}^m$

$$h(x^+) - \left\langle x - x^+, \nabla f(v) + \frac{\nabla b(x^+) - \nabla b(u)}{\tau} \right\rangle \leq h(x).$$

Applying Lemma 1 with $(x, y, z) = (x^+, u, x)$, we get

$$h(x^+) - \langle x - x^+, \nabla f(v)\rangle + \frac{1}{\tau}\left(D_b(x^+, x) + D_b(u, x^+) - D_b(u, x)\right) \leq h(x).$$

Adding (6) with $u = x^+$, using Lemma 3 with some $s > 0$ to bound the scalar product and using the Lipschitz continuity of ∇f then yields the desired result. ■

4 Proposed Algorithm

4.1 Optimization Problem

We consider the following two-block unconstrained optimization problem

$$\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} J(x, y) := F(x) + G(y) + H(x, y). \tag{7}$$

We assume that the objective function J satisfies some of the following assumptions:

Assumptions (H1)

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

(b) $F : \mathcal{X} \to \mathbb{R}$ and $G : \mathcal{Y} \to \mathbb{R}$ are continuously differentiable, and their gradients are Lipschitz continuous, of modulus $\mathrm{L}_{\nabla F}$ and $\mathrm{L}_{\nabla G}$, respectively;

(c) $H : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is proper, l.s.c. and lower bounded.

Assumptions (H2)

(a) H is continuous on its closed domain;

(b) For any $(x, y) \in \operatorname{dom} H$, the partial subdifferentials satisfy $\partial_x H(x, y) \times \partial_y H(x, y) \subset \partial H(x, y)$;

(c) H is biconvex, *i.e.* $x \mapsto H(x, y)$ is convex for any $y \in \mathcal{Y}$ and $y \mapsto H(x, y)$ is convex for any $x \in \mathcal{X}$.

Remark 5 From Assumptions (H1) and (H2), J is continuous on its closed domain $\operatorname{dom} J = \operatorname{dom} H$, which means that for any convergent sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ in $\operatorname{dom} J$ with limit $(x^*, y^*) \in \operatorname{dom} J$, one has $J(x^k, y^k) \to J(x^*, y^*)$.

Example 2 Assumption (H2)(b) is satisfied as soon as H is differentiable, or additively separable, or a sum of such functions. Otherwise said, Assumption (H2)(b) holds true if

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

with h differentiable, f and g proper l.s.c. functions.

Assumptions (H3) $H : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ can be split into $H(x, y) = h(x, y) + f(x) + g(y)$, where

(a) $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ are continuous on their domain;

(b) $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is continuous;

(c) $x \mapsto h(x, y)$ is continuously differentiable for any $y \in \operatorname{dom} g$. Moreover, $y \mapsto \nabla_x h(x, y)$ locally Lipschitz continuous on $\operatorname{dom} f$, in the sense that for each bounded subset $\mathcal{B}_\mathcal{X} \times \mathcal{B}_\mathcal{Y} \subset \operatorname{dom} f \times \operatorname{dom} g$, there is a constant $\xi > 0$ such that for any $x \in \mathcal{B}_\mathcal{X}$ and $(y, y') \in \mathcal{B}_\mathcal{Y}^2$, one has

$$\|\nabla_x h(x, y) - \nabla_x h(x, y')\| \leq \xi \, \|y - y'\|.$$

Obviously, (H3)(c) holds for Lipschitz continuously differentiable functions h. In this case, (H2) holds as soon as (H3) is satisfied (according to Example 2).

Assumption (H4) The function J satisfies the Kurdyka-Łojasiewicz (KŁ) property on every point (x^+, y^+) of its domain, namely there exist $\eta \in (0, +\infty]$, a neighborhood $\mathcal{O}(x^+, y^+)$ of (x^+, y^+) and $\kappa > 0$ such that for any $(x, y) \in \mathcal{O}(x^+, y^+) \cap \mathcal{N}_\eta(J(x^+, y^+))$

$$\kappa \operatorname{dist}(0, \partial J(x, y)) \geq \left| J(x, y) - J(x^+, y^+) \right|^\theta, \tag{8}$$

where $\theta \in [0, 1)$ and $\mathcal{N}_\eta(t) := J^{-1}\big((t, t + \eta)\big)$.

Remark 6 Assumption (H4) is crucial to prove strong convergence for nonconvex and nonsmooth optimization schemes such as [8]. The main apparent difficulty is to verify that the functional J fulfills (8). In practice, it is sufficient to show that they are *subanalytic* [9, Theorem 3.1], e.g., sums of lower-bounded semi-algebraic or real-analytic functions [42, Chapter II.1] and/or compositions [19, Prop. 2.46] of such functions provided they preserve boundedness.

As shown in the analysis below, adding hypotheses on the optimization problem (7) leads to stronger convergence results for the proposed method. For the sake of simplicity, we will mainly focus on the case where H is biconvex[7]. However, the hypotheses (H2) and (H3) given in this paper can be relaxed while yielding same convergence results. The interested reader may see [31] for more details.

4.2 Inertial Bregman Alternating Structure-Adapted Proximal Gradient Descent

We can now introduce the general form of the proposed algorithm in Algorithm 1. It basically consists in alternating partial (block) minimization steps, each step being decomposed into two overrelaxation (inertial) steps (in the spirit of the strategy proposed by Nesterov for projected gradient descent), followed by a Bregman FBS.

When $b_\mathcal{X}$ and $b_\mathcal{Y}$ are Frobenius norm, $\alpha_1^k = \alpha_2^k = \beta_1^k = \beta_2^k = 0$ (no relaxation) and $(\tau_k, \sigma_k) = (\tau, \sigma)$ (constant stepsizes), one recovers the ASAP (Alternating Structure-Adapted Proximal gradient descent) algorithm introduced in [31]. Thus, the proposed scheme can be seen as an inertial Bregman generalization of ASAP with varying stepsizes.

Remark 7 Let us make some preliminary remarks about Algorithm 1.

(a) If $\hat{x}^k \in \operatorname{int}(\operatorname{dom} b_\mathcal{X})$, then $x^{k+1} \in \operatorname{dom} b_\mathcal{X}$, and if $\hat{y}^k \in \operatorname{int}(\operatorname{dom} b_\mathcal{Y})$, then $y^{k+1} \in \operatorname{dom} b_\mathcal{Y}$.

[7] Note that the biconvexity of the whole objective J is not required in our analysis, and that for some weak convergence results, it is even not needed for the coupling term H.

Algorithm 1 Inertial Bregman ASAP

Input: $b_{\mathcal{X}} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $b_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ strictly convex, differentiable on $\operatorname{int}(\operatorname{dom} b_{\mathcal{X}})$ and $\operatorname{int}(\operatorname{dom} b_{\mathcal{Y}})$ respectively. Choose $(x^0, y^0) \in (\operatorname{int}(\operatorname{dom} b_{\mathcal{X}}) \times \mathcal{Y}) \cap \operatorname{dom} J$. Let $x^{-1} = x^0$ and $y^{-1} = y^0$.
for all $k \geq 0$**, do**

$$\begin{aligned}
\bar{x}^k &= x^k + \alpha_1^k (x^k - x^{k-1}), && (9)\\
\hat{x}^k &= x^k + \alpha_2^k (x^k - x^{k-1}), && (10)\\
x^{k+1} &\in \mathcal{P}^{b_{\mathcal{X}}}_{\tau_k H(\cdot, y^k)}(\hat{x}^k; \tau_k \nabla F(\bar{x}^k)); && (11)\\
\bar{y}^k &= y^k + \beta_1^k (y^k - y^{k-1}), && (12)\\
\hat{y}^k &= y^k + \beta_2^k (y^k - y^{k-1}), && (13)\\
y^{k+1} &\in \mathcal{P}^{b_{\mathcal{Y}}}_{\sigma_k H(x^{k+1}, \cdot)}(\hat{y}^k; \sigma_k \nabla G(\bar{y}^k)). && (14)
\end{aligned}$$

end for

(b) If $b_{\mathcal{X}}$ and $b_{\mathcal{Y}}$ are strongly convex and H lower bounded, then (11) and (14) admit at most one solution. The solution is uniquely defined iff the function minimized in the cited problems are not identically equal to $+\infty$, that is, iff there exists $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $(x, y^k), (x^{k+1}, y) \in \operatorname{dom} H$.

In order to ensure that the iterations in Algorithm 1 are well-defined and to prove convergence results, we need to make some assumptions on the algorithm parameters, *i.e.* on the choice of the Bregman distances $(b_{\mathcal{X}}, b_{\mathcal{Y}})$ (Assumptions (A1)–(A2)), the inertial parameters $\{(\alpha_1^k, \alpha_2^k, \beta_1^k, \beta_2^k)\}_{k\in\mathbb{N}}$ (Assumptions (A2)–(A3)), and the stepsizes $\{(\tau_k, \sigma_k)\}_{k\in\mathbb{N}}$ (Assumptions (A3)).

Assumptions (A1)

(a) $b_{\mathcal{X}} : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $b_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ are continuously differentiable on $\operatorname{int}(\operatorname{dom} b_{\mathcal{X}})$ and $\operatorname{int}(\operatorname{dom} b_{\mathcal{Y}})$, respectively;

(b) $b_{\mathcal{X}}$ and $b_{\mathcal{Y}}$ are strongly convex of respective modulus $L_{\mathcal{X}}$ and $L_{\mathcal{Y}}$.

Note that all Assumptions (A1) hold when $b_{\mathcal{X}}$ and $b_{\mathcal{Y}}$ are defined thanks to a norm $\|\cdot\|_M$.

Assumptions (A2)

(a) If $\operatorname{dom} b_{\mathcal{X}} \neq \mathcal{X}$ (resp. $\operatorname{dom} b_{\mathcal{Y}} \neq \mathcal{Y}$), then $\alpha_2^k = 0$ (resp. $\beta_2^k = 0$) for any $k \in \mathbb{N}$.

(b) If $\alpha_2^k > 0$ (resp. $\beta_2^k > 0$), then $\nabla b_{\mathcal{X}}$ (resp. $\nabla b_{\mathcal{Y}}$) is Lipschitz continuous of modulus $\mathrm{L}_{\nabla b_{\mathcal{X}}}$ (resp. $\mathrm{L}_{\nabla b_{\mathcal{Y}}}$).

Assumption (A2)(a) guarantees that the overrelaxed sequences $\{\hat{x}^k\}_{k\in\mathbb{N}}$ and $\{\hat{y}^k\}_{k\in\mathbb{N}}$ stay in $\operatorname{int}(\operatorname{dom} b_{\mathcal{X}})$ and $\operatorname{int}(\operatorname{dom} b_{\mathcal{Y}})$ respectively, so that the iterates $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ are well-defined (see Remark 4).

Assumptions (A3) Let $\{s_k\}_{k\in\mathbb{N}}$ and $\{t_k\}_{k\in\mathbb{N}}$ be two sequences of positive numbers. Set for any $k \geq 1$

$$B_{\mathcal{X}}^k := \frac{\mathrm{L}^2_{\nabla F}\,(\alpha_1^k)^2}{2\,s_k} + \frac{\mathrm{L}^2_{\nabla b_{\mathcal{X}}}\alpha_2^k}{2\,\tau_k} \geq 0 \qquad \text{and} \qquad B_{\mathcal{Y}}^k := \frac{\mathrm{L}^2_{\nabla G}\,(\beta_1^k)^2}{2\,t_k} + \frac{\mathrm{L}^2_{\nabla b_{\mathcal{Y}}}\beta_2^k}{2\,\sigma_k} \geq 0$$

(if $\nabla b_{\mathcal{X}}$ is not Lipschitz continuous and $\alpha_2^k = 0$, then we set $\mathrm{L}^2_{\nabla b_{\mathcal{X}}} \alpha_2^k = 0$ by convention; same for $\mathrm{L}^2_{\nabla b_{\mathcal{Y}}} \beta_2^k$) and

$$A_{\mathcal{X}}^k := \frac{2\,L_{\mathcal{X}} - \alpha_2^{k-1}}{2\,\tau_k} - \frac{\mathrm{L}_{\nabla F} + s_{k-1}}{2} \qquad \text{and} \qquad A_{\mathcal{Y}}^k := \frac{2\,L_{\mathcal{Y}} - \beta_2^{k-1}}{2\,\sigma_k} - \frac{\mathrm{L}_{\nabla G} + t_{k-1}}{2}.$$

Assume that $\tau_k, \sigma_k, s_k, t_k > 0$ and $\alpha_1^k, \alpha_2^k, \beta_1^k, \beta_2^k \geq 0$ are chosen such that

$$\rho_{\mathcal{X}} := \inf_k \left\{ A_{\mathcal{X}}^k - B_{\mathcal{X}}^k \right\} > 0, \quad \rho_{\mathcal{Y}} := \inf_k \left\{ A_{\mathcal{Y}}^k - B_{\mathcal{Y}}^k \right\} > 0$$

and let $\rho := \min\{\rho_{\mathcal{X}}, \rho_{\mathcal{Y}}\}$.

Remark 8 For any $\alpha_1^k, \alpha_2^k, \beta_1^k, \beta_2^k \geq 0$ and $s_k, t_k > 0$, the quantities $B_{\mathcal{X}}^k$ and $B_{\mathcal{Y}}^k$ are nonnegative. Moreover, they cancel iff $\alpha_1^k = \alpha_2^k = 0$ and $\beta_1^k = \beta_2^k = 0$ respectively. Hence, when Assumptions (A3) hold, $A_{\mathcal{X}}^k$ and $A_{\mathcal{Y}}^k$ are necessarily positive and upper bounded.

Eventually, we introduce the following assumption:

Assumption (A4) Assume that $\tau_k, \sigma_k, s_k, t_k > 0$ and $\alpha_1^k, \alpha_2^k, \beta_1^k, \beta_2^k \geq 0$ are chosen such that

$$\inf_k A_{\mathcal{X}}^k > \sup_k B_{\mathcal{X}}^k \qquad \text{and} \qquad \inf_k A_{\mathcal{Y}}^k > \sup_k B_{\mathcal{Y}}^k$$

with $A_{\mathcal{X}}^k$, $A_{\mathcal{Y}}^k$, $B_{\mathcal{X}}^k$ and $B_{\mathcal{Y}}^k$ defined in Assumption (A4).

It is obvious that Assumption (A4) is a particular case of Assumptions (A3). When all the sequences involved in Assumption (A4) are chosen constant, (A3) and (A4) are equivalent.

4.3 Convergence Analysis

Proposition 5 (Objective convergence) *Let Assumptions (H1), (A1)–(A3) hold. Let $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1. Then the following assertions hold:*

(a) *For every $k \geq 1$, one has*

$$\begin{aligned} &J(x^{k+1}, y^{k+1}) + A_{\mathcal{X}}^k \|x^{k+1} - x^k\|^2 + A_{\mathcal{Y}}^k \|y^{k+1} - y^k\|^2 \\ &\quad \leq J(x^k, y^k) + B_{\mathcal{X}}^k \|x^k - x^{k-1}\|^2 + B_{\mathcal{Y}}^k \|y^k - y^{k-1}\|^2 \end{aligned}$$

(b) $\lim_{k\to\infty} \|x^{k-1} - x^k\| = 0$ *and* $\lim_{k\to\infty} \|y^{k+1} - y^k\| = 0$.

(c) *The sequence $\{J(x^k, y^k)\}_{k\in\mathbb{N}}$ is convergent.*

Proof. (a) Let $k \geq 1$. Applying Proposition 4 with $b = b_{\mathcal{X}}$, $h = H(\cdot, y^k)$, $f = F$, $x^+ = x^{k+1}$, $x = x^k$, $u = \hat{x}^k$, $v = \bar{x}^k$, and $s = s_k > 0$ yields

$$\begin{aligned} &J(x^{k+1}, y^k) + \frac{1}{\tau_k} \left(D_{b_{\mathcal{X}}}(\hat{x}^k, x^{k+1}) + D_{b_{\mathcal{X}}}(x^{k+1}, x^k) \right) \\ &\leq J(x^k, y^k) + \frac{1}{\tau_k} D_{b_{\mathcal{X}}}(\hat{x}^k, x^k) + \frac{\mathrm{L}_{\nabla F} + s_k}{2} \|x^k - x^{k+1}\|^2 + \frac{\mathrm{L}^2_{\nabla F}}{2 s_k} \|x^k - \bar{x}^k\|^2 \end{aligned}$$

By definition of $\hat{x}^k$ and $\bar{x}^k$, one has $x^k - \bar{x}^k = \alpha_1^k (x^k - x^{k-1})$. Hence, the inequality above also reads

$$\begin{aligned} &J(x^{k+1}, y^k) + \frac{1}{\tau_k} \left(D_{b_{\mathcal{X}}}(\hat{x}^k, x^{k+1}) + D_{b_{\mathcal{X}}}(x^{k+1}, x^k) \right) - \frac{\mathrm{L}_{\nabla F} + s_k}{2} \|x^k - x^{k+1}\|^2 \qquad (15) \\ &\qquad \leq J(x^k, y^k) + \frac{1}{\tau_k} D_{b_{\mathcal{X}}}(\hat{x}^k, x^k) + \frac{\mathrm{L}^2_{\nabla F} (\alpha_1^k)^2}{2 s_k} \|x^k - x^{k-1}\|^2. \end{aligned}$$

Using Proposition 1 with $x = x^k$, $y = \hat{x}^k$, and $z = x^{k+1}$, we then get

$$J(x^{k+1}, y^k) + \frac{1}{\tau_k}\left(D_{b_\mathcal{X}}(x^k, x^{k+1}) + D_{b_\mathcal{X}}(x^{k+1}, x^k)\right) - \frac{\mathrm{L}_{\nabla F} + s_k}{2}\|x^k - x^{k+1}\|^2$$
$$\leq J(x^k, y^k) + \frac{\mathrm{L}^2_{\nabla F}(\alpha_1^k)^2}{2s_k}\|x^k - x^{k-1}\|^2 + \frac{1}{\tau_k}\langle \nabla b(\hat{x}^k) - \nabla b(x^k), x^{k+1} - x^k\rangle.$$

First assume that $\alpha_2^k > 0$. Then, Lemma 3 along with the Lipschitz continuity of $\nabla b_\mathcal{X}$ yield

$$J(x^{k+1}, y^k) + \frac{1}{\tau_k}\left(D_{b_\mathcal{X}}(x^k, x^{k+1}) + D_{b_\mathcal{X}}(x^{k+1}, x^k)\right) - \frac{\mathrm{L}_{\nabla F} + s_k}{2}\|x^k - x^{k+1}\|^2$$
$$\leq J(x^k, y^k) + \frac{\mathrm{L}^2_{\nabla F}(\alpha_1^k)^2}{2s_k}\|x^k - x^{k-1}\|^2 + \frac{1}{\tau_k}\left(\frac{\mathrm{L}^2_{\nabla b_\mathcal{X}}}{2\alpha_2^k}\|\hat{x}^k - x^k\|^2 + \frac{\alpha_2^k}{2}\|x^{k+1} - x^k\|^2\right).$$

Then, noticing that $\hat{x}^k - x^k = \alpha_2^k(x^{k-1} - x^k)$ and rearranging the terms in the inequality above lead to

$$J(x^{k+1}, y^k) + \frac{1}{\tau_k}\left(D_{b_\mathcal{X}}(x^k, x^{k+1}) + D_{b_\mathcal{X}}(x^{k+1}, x^k)\right) - \left(\frac{\mathrm{L}_{\nabla F} + s_k}{2} + \frac{\alpha_2^k}{2\tau_k}\right)\|x^{k+1} - x^k\|^2$$
$$\leq J(x^k, y^k) + \left(\frac{\mathrm{L}^2_{\nabla F}(\alpha_1^k)^2}{2s_k} + \frac{\mathrm{L}^2_{\nabla b_\mathcal{X}}\alpha_2^k}{2\tau_k}\right)\|x^k - x^{k-1}\|^2. \quad (16)$$

If $\alpha_2^k = 0$, (15) and (16) are the same, so (16) holds whenever $\alpha_2^k \geq 0$. Now, since $b_\mathcal{X}$ is $L_\mathcal{X}$-convex, one can lowerbound the previous inequality using (4), which gives:

$$J(x^{k+1}, y^k) + \left(\frac{2\,L_\mathcal{X} - \alpha_2^k}{2\tau_k} - \frac{\mathrm{L}_{\nabla F} + s_k}{2}\right)\|x^{k+1} - x^k\|^2$$
$$\leq J(x^k, y^k) + \left(\frac{\mathrm{L}^2_{\nabla F}(\alpha_1^k)^2}{2s_k} + \frac{\mathrm{L}^2_{\nabla b_\mathcal{X}}\alpha_2^k}{2\tau_k}\right)\|x^k - x^{k-1}\|^2.$$

Using the notations introduced in Assumptions (A2), one finally gets

$$J(x^{k+1}, y^k) + A_\mathcal{X}^k\|x^{k+1} - x^k\|^2 \leq J(x^k, y^k) + B_\mathcal{X}^k\|x^k - x^{k-1}\|^2. \quad (17)$$

Similar computations yield

$$J(x^{k+1}, y^{k+1}) + A_\mathcal{Y}^k\|y^{k+1} - y^k\|^2 \leq J(x^{k+1}, y^k) + B_\mathcal{X}^k\|y^k - y^{k-1}\|^2. \quad (18)$$

Summing (17) and (18) then yields (a).
(b) Since $A_\mathcal{X}^k > B_\mathcal{X}^K \geq 0$ and $A_\mathcal{Y}^k > B_\mathcal{X}^K \geq 0$, summing the inequality (a) for $k = 0$ to $k = K - 1$ yields

$$J(x^K, y^K) + \sum_{k=0}^{K-1}\left(A_\mathcal{X}^k\|x^{k+1} - x^k\|^2 + A_\mathcal{Y}^k\|y^{k+1} - y^k\|^2\right) \leq J(x^0, y^0).$$

Using that $J(x^K, y^K)$ is lower bounded by $\inf J \in \mathbb{R}$ and that $A_\mathcal{X}^k, A_\mathcal{Y}^k \geq \rho > 0$ yields that the series $\sum\|x^{k+1} - x^k\|^2$ and $\sum\|y^{k+1} - y^k\|^2$ converge. Thus, the sequences $\{\|x^{k+1} - x^k\|\}_{k\in\mathbb{N}}$ and $\{\|y^{k+1} - y^k\|\}_{k\in\mathbb{N}}$ converge to zero.
(c) Since $A_\mathcal{X}^k > B_\mathcal{X}^K \geq 0$ and $A_\mathcal{Y}^k > B_\mathcal{Y}^K \geq 0$, the inequality (a) implies that the sequence $\{J(x^k, y^k) + B_\mathcal{X}^k\|x^k - x^{k-1}\|^2 + B_\mathcal{Y}^k\|y^k - y^{k-1}\|^2\}_{k\in\mathbb{N}}$ is nonincreasing. Moreover, it is lower bounded, as J is. Thus it converges. According to the previous point, $\{\|x^{k+1} - x^k\|\}_{k\in\mathbb{N}}$ and $\{\|y^{k+1} - y^k\|\}_{k\in\mathbb{N}}$ converge to zero. Thus, $\{J(x^k, y^k)\}_{k\in\mathbb{N}}$ converges as well. ■

Remark 9 Unlike the ASAP method, Algorithm 1 may generate objective-convergent sequences which are not monotone.

The set of all limit points of a sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ generated by Algorithm 1 starting from a point (x^0, y^0) is denoted by $\mathcal{L}(x^0, y^0)$.

Proposition 6 (Subsequential convergence to critical points) *Let Assumptions (H1), (H2) and (A1)–(A3) hold and let $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1 which is assumed to be bounded. Let $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$.*

(a) *there is a subsequence $(x^{k_j}, y^{k_j})_{j\in\mathbb{N}}$ converging 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.*

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

Proof. (a) is a consequence of the boundedness assumption.
(b) Since J is continuous on its closed domain, and $(x^k, y^k) \in \operatorname{dom} H = \operatorname{dom} J$ (since there are defined as minimizers), one has that $(x^{k_j}, y^{k_j}) \to (x^*, y^*) \in \operatorname{dom} J$ and

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

(c) Using Fermat's rule on the definition of x^{k+1} and y^{k+1} yields after rearrangement

$$\begin{aligned} q_x^{k+1} &:= \nabla F(x^{k+1}) - \nabla F(\bar{x}^k) - \frac{\nabla b_{\mathcal{X}}(x^{k+1}) - \nabla b_{\mathcal{X}}(\hat{x}^k)}{\tau_k} \\ &\in \nabla F(x^{k+1}) + \partial_x H(x^{k+1}, y^k) = \partial_x J(x^{k+1}, y^k) \end{aligned}$$

$$\begin{aligned} p_y^{k+1} &:= \nabla G(y^{k+1}) - \nabla G(\bar{y}^k) - \frac{\nabla b_{\mathcal{Y}}(y^{k+1}) - \nabla b_{\mathcal{Y}}(\hat{y}^k)}{\sigma_k} \\ &\in \nabla G(y^{k+1}) + \partial_y H(x^{k+1}, y^{k+1}) = \partial_y J(x^{k+1}, y^{k+1}). \end{aligned}$$

Note that, according to Proposition 5(b), one has

$$\begin{aligned} x^{k+1} - \bar{x}^k &= x^{k+1} - x^k - \alpha_1^k(x^k - x^{k-1}) \to 0 \\ x^{k+1} - \hat{x}^k &= x^{k+1} - x^k - \alpha_2^k(x^k - x^{k-1}) \to 0 \\ y^{k+1} - \bar{y}^k &= y^{k+1} - y^k - \beta_1^k(y^k - y^{k-1}) \to 0 \\ y^{k+1} - \hat{y}^k &= y^{k+1} - y^k - \beta_2^k(y^k - y^{k-1}) \to 0 \end{aligned}$$

Then, the Lipschitz continuity of ∇F, ∇G, $\nabla b_{\mathcal{X}}$, and $\nabla b_{\mathcal{Y}}$ implies that $q_x^k \to 0$ and $p_y^k \to 0$. One eventually gets the desired result by applying Proposition 2 with the subsequences $\{q_x^{k_j}\}_j$ and $\{p_y^{k_j}\}_j$, and using that, by hypothesis, $\partial_x J(x^*, y^*) \times \partial_y J(x^*, y^*) \subset \partial J(x^*, y^*)$.
(d) Suppose that $\{\operatorname{dist}((x^k, y^k), \operatorname{crit} J)\}_{k\in\mathbb{N}}$ does not go to zero as $k \to \infty$. Then there exist $M > 0$ and $(k_j)_{j\in\mathbb{N}}$ such that for any $j \in \mathbb{N}$, $\operatorname{dist}\big((x^{k_j}, y^{k_j}), \operatorname{crit} J\big) > M$. However, since $\{(x^{k_j}, y^{k_j})\}_{j\in\mathbb{N}}$ is a subsequence of the bounded sequence $\{(x^k, y^k)\}_{k\in\mathbb{N}}$, it has convergent subsequence $\{(x^{k_{j_n}}, y^{k_{j_n}})\}_n$ of limit $(x^*, y^*) \in \operatorname{crit} J$ (according to (c)). Thus,

$$\begin{aligned} M &< \operatorname{dist}\big((x^{k_{j_n}}, y^{k_{j_n}}), \operatorname{crit} J\big) \\ &\leq \|(x^{k_{j_n}}, y^{k_{j_n}}) - (x^*, y^*)\| \underset{n\to\infty}{\longrightarrow} 0 \end{aligned}$$

which leads to a contradiction. ■

Remark 10 If $\operatorname{dom} J$ is bounded, then all the sequences generated by Algorithm 1 are bounded.

To prove the strong convergence of the iterates generated by Algorithm 1, we need to make additional assumptions and to prove some preliminaries lemmas. First introduce the following notations for $\varepsilon_x, \varepsilon_y \geq 0$ and $X = (x_1, x_2) \in U^2$ and $Y = (y_1, y_2) \in V^2$:

$$\Psi_{(\varepsilon_x,\varepsilon_y)}(X,Y) := J(x_1,y_1) + \varepsilon_x \|x_1 - x_2\|^2 + \varepsilon_y \|y_1 - y_2\|^2.$$

Assume that Assumption (A4) holds. Then, one can assume that

$$\inf_k A^k_{\mathcal{X}} > \varepsilon_x > \sup_k B^k_{\mathcal{X}} \qquad \text{and} \qquad \inf_k A^k_{\mathcal{Y}} > \varepsilon_y > \inf_k B^k_{\mathcal{Y}} \tag{19}$$

Corollaire 1 *Let Assumptions (H1) and (A1)–(A4) hold. Let $\{(x^k, y^k\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1. For any $k \geq 0$, set $X^k = (x^k, x^{k-1})$ and $Y^k = (y^k, y^{k-1})$. Then, there exists $a > 0$ such that*

$$\Psi_{(\varepsilon_x,\varepsilon_y)}(X^{k+1},Y^{k+1}) + a \left\|(X^{k+1} - X^k, Y^{k+1} - Y^k)\right\|^2 \leq \Psi_{(\varepsilon_x,\varepsilon_y)}(X^k,Y^k)$$

where $(\varepsilon_x, \varepsilon_y)$ satisfies (19).

Proof. According to Proposition 5(a), one has

$$\begin{aligned} J(x^{k+1},y^{k+1}) + \varepsilon_x \|x^{k+1} - x^k\|^2 + \varepsilon_y \|y^{k+1} - y^k\|^2 + (A^k_{\mathcal{X}} - \varepsilon_x) \|x^{k+1} - x^k\|^2 \\ +(\varepsilon_x - B^k_{\mathcal{X}}) \|x^k - x^{k-1}\|^2 + (A^k_{\mathcal{Y}} - \varepsilon_y) \|y^{k+1} - y^k\|^2 + (\varepsilon_y - B^k_{\mathcal{X}}) \|y^k - y^{k-1}\|^2 \\ \leq J(x^k,y^k) + \varepsilon_x \|x^k - x^{k-1}\|^2 + \varepsilon_y \|y^k - y^{k-1}\|^2 \end{aligned}$$

Thus, one can choose $a := \inf_k\{A^k_{\mathcal{X}} - \varepsilon_x, \varepsilon_x - B^k_{\mathcal{X}}, A^k_{\mathcal{Y}} - \varepsilon_y, \varepsilon_y - B^k_{\mathcal{Y}}\}$, which is positive according to (19). ■

Corollaire 2 *Let Assumptions (H1)–(H3) and (A1)–(A3) hold. Let $\{(x^k, y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1 which is assumed to be bounded. Then for any $k \geq 0$ one has $(p_x^{k+1}, p_y^{k+1}) \in \partial J(x^{k+1}, y^{k+1})$ and $\xi > 0$ such that*

$$\begin{aligned} \|p_x^{k+1}\| &\leq \left(\mathrm{L}_{\nabla F} + \frac{\mathrm{L}_{\nabla b_{\mathcal{X}}}}{\tau_k}\right) \|x^{k+1} - x^k\| \\ &\quad + \left(\mathrm{L}_{\nabla F}\, \alpha_1^k + \frac{\mathrm{L}_{\nabla b_{\mathcal{X}}}}{\tau_k}\alpha_2^k\right) \|x^k - x^{k-1}\| + \xi \|y^k - y^{k+1}\| \\ \|p_y^{k+1}\| &\leq \left(\mathrm{L}_{\nabla G} + \frac{\mathrm{L}_{\nabla b_{\mathcal{Y}}}}{\sigma_k}\right) \|y^{k+1} - y^k\| \\ &\quad + \left(\mathrm{L}_{\nabla G}\, \beta_1^k + \frac{\mathrm{L}_{\nabla b_{\mathcal{Y}}}}{\sigma_k}\beta_2^k\right) \|y^k - y^{k-1}\|. \end{aligned}$$

Proof. Since $J(x, y^k) = F(x) + h(x, y^k) + f(x) + \text{constant}$, with $x \mapsto h(x, y^k)$ continuously differentiable, simple subdifferential calculus shows that

$$\begin{aligned} \partial_x J(x^{k+1}, y^k) &= \nabla F(x^{k+1}) + \nabla_x h(x^{k+1}, y^k) + \partial f(x^{k+1}) \\ &= \nabla F(x^{k+1}) + \nabla_x h(x^{k+1}, y^{k+1}) + \partial f(x^{k+1}) \\ &\quad + \nabla_x h(x^{k+1}, y^k) - \nabla_x h(x^{k+1}, y^{k+1}). \end{aligned}$$

Hence, using the notations introduced in the proof of Proposition 6(c), one has $(p_x^{k+1}, p_y^{k+1}) \in \partial J(x^{k+1}, y^{k+1})$ with

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

which satisfies the following bounds

$$\begin{aligned}
\|p_x^{k+1}\| &\leq \left\|\nabla F(x^{k+1}) - \nabla F(\bar{x}^k)\right\| + \frac{1}{\tau_k}\left\|\nabla b_{\mathcal{X}}(x^{k+1}) - \nabla b_{\mathcal{X}}(\hat{x}^k)\right\| \\
&\quad + \left\|\nabla_x h(x^{k+1}, y^k) - \nabla_x h(x^{k+1}, y^{k+1})\right\| \\
\|p_y^{k+1}\| &\leq \left\|\nabla G(y^{k+1}) - \nabla G(\bar{y}^k)\right\| + \frac{1}{\sigma_k}\left\|\nabla b_{\mathcal{Y}}(y^{k+1}) - \nabla b_{\mathcal{Y}}(\hat{y}^k)\right\|.
\end{aligned}$$

Using the Lipschitz continuity of ∇F, ∇G, $\nabla b_{\mathcal{X}}$, $\nabla b_{\mathcal{Y}}$, and the local Lipschitz continuity of $\nabla_x h$ ($\{(x^k, y^k)\}_{k\in\mathbb{N}}$ being bounded by assumption), one shows that there exists $\xi > 0$ such that

$$\begin{aligned}
\|p_x^{k+1}\| &\leq \mathrm{L}_{\nabla F}\|x^{k+1} - \bar{x}^k\| + \frac{\mathrm{L}_{\nabla b_{\mathcal{X}}}}{\tau_k}\|x^{k+1} - \hat{x}^k\| \\
&\quad + \xi\|y^k - y^{k+1}\| \\
\|p_y^{k+1}\| &\leq \mathrm{L}_{\nabla G}\|y^{k+1} - \bar{y}^k\| + \frac{\mathrm{L}_{\nabla b_{\mathcal{Y}}}}{\sigma_k}\|y^{k+1} - \hat{y}^k\|
\end{aligned}$$

and one completes the proof using the definition of the relaxed sequences. ■

The following corollary shows that Algorithm 1 generates a sequence of subgradient which can be bounded.

Corollaire 3 *Let Assumptions (H1)–(H3) and (A1)–(A3) hold. Let $\{(x^k, y^k\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1 which is assumed to be bounded. For any $k \geq 0$, set $X^k = (x^k, x^{k-1})$ and $Y^k = (y^k, y^{k-1})$. Then there exists $\xi > 0$ such that for any $k \geq 0$ one has $P^{k+1} \in \partial\Psi_{\varepsilon_x,\varepsilon_y}(X^{k+1}, Y^{k+1})$ such that*

$$\|P^{k+1}\| \leq \xi\left\|(X^{k+1} - X^k, Y^{k+1} - Y^k)\right\|$$

where $(\varepsilon_x, \varepsilon_y)$ satisfies (19).

Proof. First note that, for any $X = (x_1, x_2) \in \mathcal{X}^2$ and $Y = (y_1, y_2) \in \mathcal{Y}^2$, $\partial_X\Psi_{\varepsilon_x,\varepsilon_y}(X, Y)$ is the set of all $P_X = (p_{X,1}, p_{X,2}) \in \mathcal{X}^2$ such that

$$\begin{cases} p_{X,1} \in \partial_x J(x_1, y_1) + 2\,\varepsilon_x\,(x_1 - x_2) \\ p_{X,2} = -2\,\varepsilon_x\,(x_1 - x_2) \end{cases}$$

while $\partial_Y\Psi_{\varepsilon_x,\varepsilon_y}(X, Y)$ is the set of all $P_Y = (p_{Y,1}, p_{Y,2}) \in \mathcal{Y}^2$ such that

$$\begin{cases} p_{Y,1} \in \partial_y J(x_1, y_1) + 2\,\varepsilon_y\,(y_1 - y_2) \\ p_{Y,2} = -2\,\varepsilon_y\,(y_1 - y_2) \end{cases}$$

Moreover, $\Psi_{\varepsilon_x,\varepsilon_y}(X, Y)$ can be split into

$$\Psi_{\varepsilon_x,\varepsilon_y}(X, Y) = S(X, Y) + N(X, Y)$$

with $S(X, Y) = F(x_1) + G(y_1) + h(x_1, y_1) + \varepsilon_x\|x_1 - x_2\|^2 + \varepsilon_y\|y_1 - y_2\|^2$ a continuously differentiable function and $N(X, Y) = f(x_1) + g(y_1)$ which is separable. Thus, one can check that

$$\partial_X\Psi_{\varepsilon_x,\varepsilon_y}(X, Y) \times \partial_Y\Psi_{\varepsilon_x,\varepsilon_y}(X, Y) = \partial\Psi_{\varepsilon_x,\varepsilon_y}(X, Y)$$

Hence, one has $P^{k+1} \in \partial\Psi_{\varepsilon_x,\varepsilon_y}(X^{k+1}, Y^{k+1})$ when setting

$$P^{k+1} := \left(p_x^{k+1}, -2\,\varepsilon_x\,(x^{k+1} - x^k),\ p_y^{k+1}, -2\,\varepsilon_y\,(y^{k+1} - y^k)\right).$$

Let us bound P^{k+1}. Pythagora's theorem yields

$$\|P^{k+1}\|^2 \quad = \quad \|p_x^{k+1}\|^2 + 4\,\varepsilon_x^2\,\|x^{k+1} - x^k\|^2 + \|p_y^{k+1}\|^2 + 4\,\varepsilon_y^2\,\|y^{k+1} - y^k\|^2.$$

Using that $(a+b)^2 \leq a^2/2 + b^2/2$ (Lemma 3 with $s = 2$), together with the bounds in Corollary 2 yields

$$\begin{aligned}\|P^{k+1}\|^2 \quad \leq \quad & \left(\frac{1}{4}\left(\mathrm{L}_{\nabla F} + \frac{\mathrm{L}_{\nabla b_{\mathcal{X}}}}{\tau_k}\right)^2 + 4\,\varepsilon_x^2\right)\|x^{k+1} - x^k\|^2 \\ & + \frac{1}{4}\left(\mathrm{L}_{\nabla F}\,\alpha_1^k + \frac{\mathrm{L}_{\nabla b_{\mathcal{X}}}}{\tau_k}\alpha_2^k\right)^2\|x^k - x^{k-1}\|^2 \\ & + \left(\frac{1}{2}\left(\mathrm{L}_{\nabla G} + \frac{\mathrm{L}_{\nabla b_{\mathcal{Y}}}}{\sigma_k}\right)^2 + \frac{1}{2}\,\xi^2 + 4\,\varepsilon_y^2\right)\|y^{k+1} - y^k\|^2 \\ & + \frac{1}{2}\left(\mathrm{L}_{\nabla G}\,\beta_1^k + \frac{\mathrm{L}_{\nabla b_{\mathcal{Y}}}}{\sigma_k}\beta_2^k\right)^2\|y^k - y^{k-1}\|^2.\end{aligned}$$

Choosing

$$\begin{aligned}\xi^2 = \quad \max\Bigg\{ & \frac{1}{4}\left(\mathrm{L}_{\nabla F} + \frac{\mathrm{L}_{\nabla b_{\mathcal{X}}}}{\tau_k}\right)^2 + 4\,\varepsilon_x^2,\ \frac{1}{4}\left(\mathrm{L}_{\nabla F}\,\alpha_1^k + \frac{\mathrm{L}_{\nabla b_{\mathcal{X}}}}{\tau_k}\alpha_2^k\right)^2, \\ & \frac{1}{2}\left(\mathrm{L}_{\nabla G} + \frac{\mathrm{L}_{\nabla b_{\mathcal{Y}}}}{\sigma_k}\right)^2 + \frac{1}{2}\,\xi^2 + 4\,\varepsilon_y^2, \frac{1}{2}\left(\mathrm{L}_{\nabla G}\,\beta_1^k + \frac{\mathrm{L}_{\nabla b_{\mathcal{Y}}}}{\sigma_k}\beta_2^k\right)^2\Bigg\}\end{aligned}$$

completes the proof. ■

Proposition 7 (Strong convergence) *Let Assumptions (H1)–(H4) and (A1)–(A4) hold. Let $\{(x^k, y^k\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1 which is assumed to be bounded. Then $\{(x^k, y^k\}_{k\in\mathbb{N}}$ is a convergent Cauchy sequence.*

Proof. For any $k \geq 0$, set $X^k = (x^k, x^{k-1})$ and $Y^k = (y^k, y^{k-1})$. Then the sequence $\{\Psi_{\varepsilon_x,\varepsilon_y}(X^k, Y^k)\}_{k\in\mathbb{N}}$ satisfies the three conditions needed in [1, Theorem 2.9], namely

(a) sufficient decrease condition (Corollary 1);

(b) relative error condition (Corollary 3);

(c) continuity condition ($\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ is bounded).

Eventually, one may check that $\Psi_{\varepsilon_x,\varepsilon_y}$ is KŁ whenever the objective function J is. ■

5 Inertial Parameter Choice

In this section, we detail some feasible choices for the algorithm parameters so that Assumptions (A1)–(A4) are satisfied. We focus here on the constant case and also give some trick to allow large inertial parameters along with large stepsizes.

5.1 Constant Case

Consider the case when all the parameters are constant. Then, set

$$B^0_{\mathcal{X}} := \frac{\mathrm{L}^2_{\nabla F}\,(\alpha_1)^2}{2\,s} + \frac{\mathrm{L}^2_{\nabla b_{\mathcal{X}}}\alpha_2}{2\,\tau} \geq 0 \qquad \text{and} \qquad B^0_{\mathcal{Y}} := \frac{\mathrm{L}^2_{\nabla G}\,(\beta_1)^2}{2\,t} + \frac{\mathrm{L}^2_{\nabla b_{\mathcal{Y}}}\beta_2}{2\,\sigma} \geq 0$$

(we recall that if $\nabla b_{\mathcal{X}}$ is not Lipschitz continuous and $\alpha_2 = 0$, then $\mathrm{L}^2_{\nabla b_{\mathcal{X}}}\alpha_2 = 0$ by convention; same for $\mathrm{L}^2_{\nabla b_{\mathcal{Y}}}\beta_2$) and

$$A^0_{\mathcal{X}} := \frac{2\,L_{\mathcal{X}} - \alpha_2}{2\,\tau} - \frac{\mathrm{L}_{\nabla F} + s}{2} \qquad \text{and} \qquad A^0_{\mathcal{Y}} := \frac{2\,L_{\mathcal{Y}} - \beta_2}{2\,\sigma} - \frac{\mathrm{L}_{\nabla G} + t}{2}$$

In this case, Assumption (A3) reduces to find $\tau, \sigma, \alpha_1, \alpha_2,\ \beta_1, \beta_2 > 0$ and $s, t > 0$ such that

$$\rho_{\mathcal{X}} := A^0_{\mathcal{X}} - B^0_{\mathcal{X}} > 0 \qquad \text{and} \qquad \rho_{\mathcal{Y}} := A^0_{\mathcal{Y}} - B^0_{\mathcal{Y}} > 0$$

According to [39], we may first choose (s, t) and the best choice is that which maximizes the upperbound for $J(x^{k+1}, y^{k+1}) - J(x^k, y^k)$, namely that which maximizes both

$$\frac{\mathrm{L}_{\nabla F} + s}{2}\,\|x^{k+1} - x^k\|^2 + \frac{\mathrm{L}^2_{\nabla F}\,(\alpha_1)^2}{2\,s}\,\|x^k - x^{k-1}\|^2$$

and

$$\frac{\mathrm{L}_{\nabla G} + t}{2}\,\|y^{k+1} - y^k\|^2 + \frac{\mathrm{L}^2_{\nabla G}\,(\beta_1)^2}{2\,t}\,\|y^k - y^{k-1}\|^2$$

Thus, the best choice for (s, t) is given by

$$\left(\mathrm{L}_{\nabla F}\,\alpha_1\,\frac{\|x^k - x^{k-1}\|}{\|x^{k+1} - x^k\|}, \mathrm{L}_{\nabla G}\,\beta_1\,\frac{\|y^k - y^{k-1}\|}{\|y^{k+1} - y^k\|}\right)$$

According to [39], we may choose

$$s = \mathrm{L}_{\nabla F}\,\alpha_1 \qquad \text{and} \qquad t = \mathrm{L}_{\nabla G}\,\beta_1$$

Thus, one has to find $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ such that

$$\frac{2 - \alpha_2}{\tau} - \mathrm{L}_{\nabla F}\,(1 + \alpha_1) > \frac{\alpha_2}{\tau} + \mathrm{L}_{\nabla F}\,\alpha_1 \qquad \text{and} \qquad \frac{2 - \beta_2}{\sigma} - \mathrm{L}_{\nabla G}\,(1 + \beta_1) > \frac{\beta_2}{\sigma} + \mathrm{L}_{\nabla G}\,\beta_1$$

i.e.

$$\frac{2}{\tau} > \mathrm{L}_{\nabla F} + 2\left(\frac{\alpha_2}{\tau} + \mathrm{L}_{\nabla F}\,\alpha_1\right) \qquad \text{and} \qquad \frac{2}{\sigma} > \mathrm{L}_{\nabla G} + 2\left(\frac{\beta_2}{\sigma} + \mathrm{L}_{\nabla G}\,\beta_1\right)$$

Note that, since the left-hand side terms are nonnegative, this implies that τ and σ should be chosen so that $\tau < 2/\mathrm{L}_{\nabla F}$ and $\sigma < 2/\mathrm{L}_{\nabla G}$. One has

$$\rho^k_x = \frac{2}{\tau} - \mathrm{L}_{\nabla F} - 2\left(\frac{\alpha_2}{\tau} + \mathrm{L}_{\nabla F}\,\alpha_1\right) \qquad \text{and} \qquad \rho^k_y = \frac{2}{\sigma} - \mathrm{L}_{\nabla G} - 2\left(\frac{\beta_2}{\sigma} + \mathrm{L}_{\nabla G}\,\beta_1\right)$$

Let us consider two simple cases, namely the case where $\alpha_1 = \alpha_2 =: \alpha$ (resp. $\beta_1 = \beta_2 =: \beta$) and the the case where $\alpha_1 = 0$ and $\alpha_2 =: \alpha$ (resp. $\beta_1 = 0$ and $\beta_2 =: \beta$). In the first case, the feasibility conditions are

$$\rho_{\mathcal{X}} = \frac{2}{\tau} - \mathrm{L}_{\nabla F} - 2\,\alpha\left(\frac{1}{\tau} + \mathrm{L}_{\nabla F}\right) > 0 \qquad \text{and} \qquad \rho_{\mathcal{Y}} = \frac{2}{\sigma} - \mathrm{L}_{\nabla G} - 2\,\beta\left(\frac{1}{\sigma} + \mathrm{L}_{\nabla G}\right) > 0$$

which means that, for given stepsizes,

$$1 \geq \frac{2/\tau - \mathrm{L}_{\nabla F}}{2/\tau + 2\,\mathrm{L}_{\nabla F}} > \alpha \geq 0 \qquad \text{and} \qquad 1 \geq \frac{2/\sigma - \mathrm{L}_{\nabla G}}{2/\sigma + 2\,\mathrm{L}_{\nabla G}} > \beta \geq 0$$

Given the idea that the inertial parameters should be as large as possible (within a reasonable range) to have a visible effect on the algorithm convergence, one can choose for instance

$$\alpha^*(\tau) = 0.999 \times \frac{2/\tau - \mathrm{L}_{\nabla F}}{2/\tau + 2\,\mathrm{L}_{\nabla F}} \qquad \text{and} \qquad \beta^*(\sigma) = 0.999 \times \frac{2/\sigma - \mathrm{L}_{\nabla G}}{2/\sigma + 2\,\mathrm{L}_{\nabla G}}$$

Then, we see that, the larger the stepsizes, the smaller the inertial parameters, since

$$\alpha^*(\tau) \approx 0.999 \times \frac{2/\tau - \mathrm{L}_{\nabla F}}{3\,\mathrm{L}_{\nabla F}} \qquad \text{and} \qquad \beta^*(\sigma) \approx 0.999 \times \frac{2/\sigma - \mathrm{L}_{\nabla G}}{3\,\mathrm{L}_{\nabla G}}$$

In the second case, the feasibility conditions become

$$1 - \frac{\tau \mathrm{L}_{\nabla F}}{2} > \alpha \qquad \text{and} \qquad 1 - \frac{\sigma \mathrm{L}_{\nabla G}}{2} > \beta$$

One can see that, once again, these upperbounds decrease to zero as the stepsizes increase to their maximum values.

5.2 Adaptive Choice for Larger Inertial Parameters

For some problems, the feasible sets for inertial parameters leads to either small stepsizes (τ, σ) or small inertial parameters (α, β) (in the constant parameter cases). In both cases, the convergence is ensured by the results in the previous section, but the overrelaxation does not enhance the observed convergence rate.

However, it may happen that larger inertial parameters than those allowed by the feasibility set lead in practice to empirical convergence. In particular, some none-null large inertial parameters (*e.g.* $\alpha = 1$) can be chosen along with large stepsizes (*i.e.* close to their upperbound), which is theoretically not permitted by the analysis in the previous subsection. In order to guarantee convergence in such cases, the following trick may be applied.

In what follows, we consider the case when $(\alpha_1^k, \beta_1^k) = (0, 0)$ and $(\alpha_2^k, \beta_2^k) = (\alpha^k, \beta^k)$. Suppose that empirical convergence is observed for $(\alpha, \beta) = (1, 0)$. The idea is to adopt the following update rule for the inertial parameters: for any k, one set $\beta^k = 0$ and

$$\alpha^k = \begin{cases} 1 & \text{while } \|(x^{k+1}, y^{k+1}) - (x^k, y^k)\| \leq M\,(1-\varepsilon)^k \\ 0.999 \times \left(1 - \dfrac{\tau\,\mathrm{L}_{\nabla F}}{2}\right) & \text{otherwise} \end{cases} \tag{20}$$

for given $M > 0$ and $\varepsilon \in (0, 1)$. Then, two cases can occur:

(a) for any k, $\alpha^k = 1$. This means that $\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\| \leq M\,(1-\varepsilon)^k$, that is, $(x^k, y^k)_{k\in\mathbb{N}}$ is a Cauchy sequence, thus it converges.

(b) there exists a k_0 such that

$$\|(x^{k_0+1}, y^{k_0+1}) - (x^{k_0}, y^{k_0})\| \leq M\,(1-\varepsilon)^{k_0}$$

In this case, the choice for α^k is feasible for $k \geq k_0$. Since the convergence results are asymptotic, this implies that all the convergence results stated in the previous section hold. In particular, under the proper hypotheses, $(x^k, y^k)_{k\in\mathbb{N}}$ converges.

Note that if the quantities M and ε are chosen so that M is large enough and $1 - \varepsilon$ is close enough to 1, the condition $\|(x^{k+1}, y^{k+1}) - (x^k, y^k)\| \leq M\,(1-\varepsilon)^k$ may hold for a sufficient number of iterations. This means that, in practice, one can always choose $\alpha^k = 1$ at least for the first iterations, without compromising the asymptotic convergence properties of the algorithm.

6 Numerical experiments

We recall that the studied colorization method consists in two steps: the candidate extraction step, which aims at extracting from the source image C color candidates for each pixel, and the candidate selection step via a vote process, which selects the best candidate under regularity assumptions, in the variational approach described in Section 2.

6.1 Candidate Extraction

As described in [11], for each pixel of the target image, a patch-based method is used to extract color candidates. Basically, the algorithm searches in the Y channel of the source image the closest patch with respect to a given criterion. The U and V chrominance channels corresponding to the pixel center of this retained patch is considered as a possible candidate. To speed-up the search of candidates, PatchMatch algorithm [3] has been implemented in [37]. The criterion used for the comparisons of the patches have been chosen experimentally, inspired from [11]. The number and the variety of the criterion have been reduced from the original version, because it has been experimentally observed in the work of [37] that some criterion were especially low to compute and do not increase significantly the quality of the results. In the experiments, we have focused on four criterion:

- SSD (Euclidean norm) between patches with size 5×5;
- absolute difference between standard-deviation of patches with size 3×3 and 5×5;
- L1-norm between cumulative histograms (15 bins) of patches with size 11×11.

Applying these criterion provide four ($C = 4$) candidates. Let us remark that this candidate number is not restricted by the model of [34] but by computational time considerations.

6.2 Candidate Selection

As discussed in Section 2, Model (1) fits with Model (7) with the following identifications:

$$F(u) = \mathrm{TV}^{\alpha}_{\mathfrak{C}}(u), \qquad G(w) = 0$$

and

$$H(u,w) = \frac{\lambda}{2} \int_{\Omega} \sum_{i=1}^{C} w_i(x) \|u(x) - c_i(x)\|_2^2 \, \mathrm{d}x + \chi_{\mathcal{R}}(u) + \chi_{\Sigma}(w)$$

In our implementation, the discrete version of $\nabla u = (\partial_x U, \partial_y U, \partial_x V, \partial_y V)$ used in the coupled total variation is defined thanks to horizontal (resp. vertical) forward finite differences, with symmetric boundary conditions. If Y is a noisy image, there exists $\alpha > 0$ so that $\mathrm{TV}^{\alpha}_{\mathfrak{C}}(u) = \mathrm{TV}_{\mathfrak{C}}(u)$ for any u. In practical case, when choosing $\alpha = 1$, $\max\left\{\gamma\|\nabla Y\|^2, \alpha\right\} = \gamma\|\nabla Y\|^2$ in a large part of pixels (for instance about 92% for Figure 1(b)). Thus this regularization of the coupled total variation has few influence on the result.

Hence, J satisfies Assumptions (H1). Moreover, one can check that, according to Example 2, Assumptions (H2) are fulfilled, while Assumptions (H3) hold when setting $f = \chi_{\mathcal{R}}$ and $g = \chi_{\Sigma}$. Eventually, Assumption (H4) is satisfied according to Remark 6. As a consequence, the constraints Σ and $\mathcal{R}$ being bounded, one can apply Algorithm 1 with a guaranteed strong convergence, as soon as the algorithm parameters satisfy Assumptions (A1)–(A4).

In order to compute the update of w, we use the Bregman distance associated to the entropy function to avoid an optimization on the simplex. The explicit solution

is then obtained thanks to the KKT conditions. For the u-update, we consider the usual Moreau proximity operator (meaning that the associated Bregman distance is the Euclidean squared distance). With this setting, the inertial parameters associated to w are set to be null, and we choose a constant stepsize, which can be set arbitrarily large; we chose $\sigma = 100$. For the u-sequence, the inertial parameter sequence is chosen according to Subsection 5.2, namely equal to $(\alpha_1^k, \alpha_2^k) = (1, 1)$. The stepsize is then chosen to be large, *e.g.* $\tau = 1.999/L_{\nabla F}$. Hence, the updates for u and w are explicitly given in Algorithm 2.

Algorithm 2 Inertial Bregman-based proximal gradient descent for image colorization

Input: Choose $w_i^0(x) = 1/C$ for any $x \in \Omega$ and $i = 1, \cdots, C$, $u^0(x) = \sum_{i=1}^{C} w_i^0(x) c_i(x)$.

Let $u^{-1} = u^0$.

for all $k \geq 0$**, do**

$$
\begin{aligned}
\bar{u}^k &= 2\,u^k - u^{k-1}, \\
\hat{u}^k &= 2\,u^k - u^{k-1}, \\
u^{k+1} &= \mathrm{proj}_{\mathcal{R}} \left(\frac{\hat{u}^k - \tau \nabla \mathrm{TV}_{\mathfrak{C}}^{\alpha}(\bar{u}^k) + \tau\,\lambda \sum_{i=1}^{C} w_i^k\, c_i}{1 + \tau\,\lambda} \right); \\
w_i^{k+1}(x) &= \frac{w_i^k(x)\, \exp\left(-\sigma\,\lambda \sum_{j=1}^{C} \|u^{k+1}(x) - c_j(x)\|^2 \right)}{\sum_{i=1}^{C} w_i^k(x)\, \exp\left(-\sigma\,\lambda \sum_{j=1}^{C} \|u^{k+1}(x) - c_j(x)\|^2 \right)}
\end{aligned}
$$

end for

6.3 Experimental Results and Discussion

In this section, some experimental results are proposed and an alternative optimization strategy is considered.

The implementation is done in Matlab. In practical experiments, $\lambda = 10^{-3}$ and $\gamma = 25$ are reliable parameters for all images. The values of the limits of $\mathcal{R}$ are chosen as follows: $U_{\max} = -U_{\min} = 111$, and $V_{\max} = -V_{\min} = 157$.

From a visually qualitative point of view, let us remark that the results performed with Algorithm 2 are the same as with the primal-dual algorithm developed in [34]. For instance, in Figure 1, we can see a source image, a target image and the results provided by Algorithm 2 and by [34]. Visually, there is no difference and numerically, in average, the difference is pixel-wise less than 1/256 of the image range. Thus we can conclude that the local minimum computed by our proposed method is as reliable as the result computed by the method of [34].

In Figure 2 the curves show the value of the functional with respect to the iteration of the algorithm. To see the benefit of using inertial scheme and Bregman distance, we also run the ASAP algorithm [31], which is a particular instance of the proposed method without inertia (*i.e.* null inertial parameters) and with usual proximity operator (*i.e.* the Bregman distances are the Euclidean distances). We also compared our method with the PALM method [8, 45], which reduces for this problem to an alternating projected gradient scheme. In our experiments, the projection onto the simplex is done using the KKT conditions. As expected, the proposed method (in purple) performs better that

the ASAP algorithm (in yellow) in terms of energy convergence. The observed gain is entirely due to the use of inertia. Using Bregman distance to compute the update of the weights w has a notable incidence on the computational time: despite the addition of the overrelaxation steps for u, 300 iterations take 71.07s to compute for the proposed algorithm, while they take 81.32s for ASAP for the image tested in Figure 2. The PALM algorithm (in red) shows the slowest convergence of the bench, along with larger computational time (130.35s for the same image).

(a) Source image.

(b) Target image.

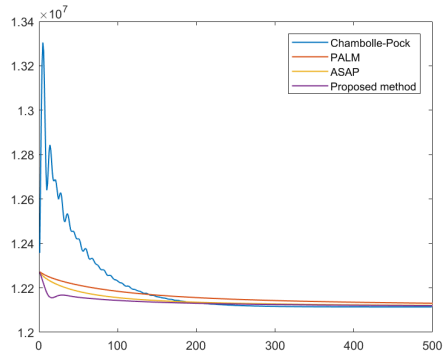
(c) Result of [34].

(d) Result of Algorithm 2.

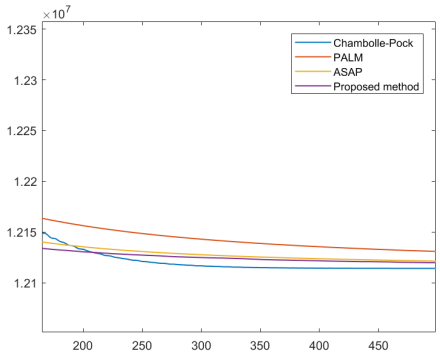
Figure 1: The result provided by the primal-dual approach of [34] is as reliable as the one computed by our Algorithm 2. In addition, with our proposed method, the convergence is guaranteed.

When comparing the energy evolution for the primal-dual approach of [34] (in blue) and the proposed method (in yellow), one can see that Algorithm 2 makes the energy to decrease fast whereas the primal-dual algorithm may increase and oscillates in the first 100 iterations. Nevertheless, it loses this advantage in comparison of the primal-dual approach after 330 iterations (see Figure 2(b)). The comparison of the histograms after 500 iterations (Figure 2(c) and (d)) shows that the local minimum is not reached by Algorithm 2 but more likely approached by the primal-dual algorithm, since it has been shown in [34] that, at convergence, the minimizer w^* contains only 0 and 1. Nevertheless, without any convergence guaranties, the primal-dual algorithm may have not converged. A way to benefit from the empirical convergence speed of the primal-dual algorithm while keeping the convergence guarantee provided by our proposed method is to consider the following strategy: first we use the primal-dual algorithm for 500 iterations and then we start Algorithm 2 from the values computed by the 500 first iterations of the primal-dual algorithm. The final result is visually the same as the one of the two previous approaches (Algorithm 2, and primal-dual) and numerically, the difference between the

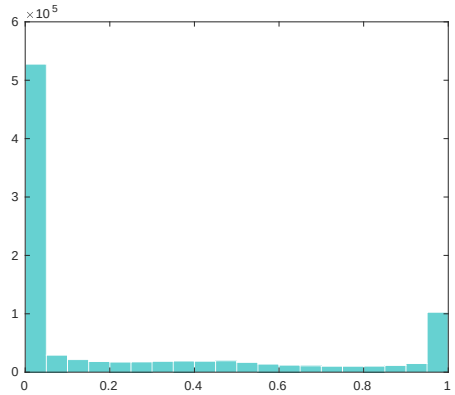
two algorithms and our strategy is pixel-wise less than 1/256 of the image range in average.



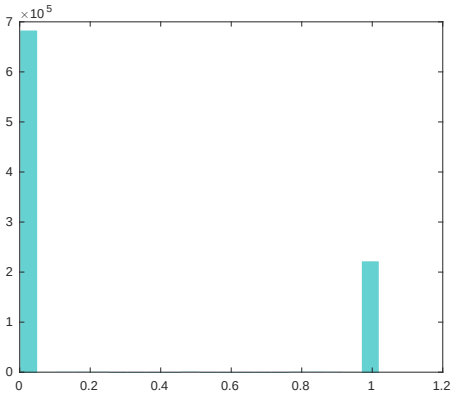
(a) Evolution of the functionnal for the different alorithms.



(b) Zoom on (a).



(c) Weights histogram of ASAP.



(d) Weights histogram of [34].

Figure 2: Value of the functional with respect to the number of iterations of the different algorithms (above) and the histograms of the weights after 500 iterations (bottom). In yellow, the energy of ASAP, in red, the one of [34] and in blue, the one of our strategy.

In Figure 3, some additional results are available. One can see the source images (left column) the target ones (middle column) and the results computed with our strategy (right column). Visually, there is no difference with the results of [34] but the convergence is guaranteed.

Some additional results can be found on http://www.fabienpierre.fr/colorisation/.

7 Conclusion

In this paper, we proposed a variant of an accelerated alternating proximal descent scheme introduced in [31] based on the Bregman distances. We provided a convergence analysis in the case when the coupling term is biconvex, along with a guide for the inertial parameter choice. This algorithm was applied to the image colorization problem. The numerical experiments confirm the empirical acceleration of the proposed scheme compared with the original one, as well as the benefit of the use of the Bregman distance in terms of computation times. Comparisons with two other optimization schemes, namely a Chambolle-Pock inspired primal-dual algorithm and the PALM algorithm prove that the proposed method gives comparable results as those obtained in the

primal-dual framework with a guaranteed convergence (unlike the algorithm proposed in [34]), while outperforming the PALM scheme, which is known as the state-of-the-art optimization scheme for nonconvex and nonsmooth problems. In future works, one could extend our algorithm to others biconvex problems such as joint image restoration and motion estimation [18].

8 Acknowledgement

This work has been partially funded by the French Research Agency (ANR) under grant No ANR-14-CE27-001 (MIRIAM) and supported by grants from Région Île-de-France.

References

[1] 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(1), Feb. 2013.

[2] Clara Barbanson, Andrés Almansa, Yann Ferrec, and Pascal Monasse. Relief computation from images of a fourier transform spectrometer for interferogram correction. In *Fourier Transform Spectroscopy*. Optical Society of America, 2016.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, 28(3):24, 2009.

[4] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[6] Dimitri P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, Massachussetts, 1995.

[7] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.*, 18(2):556–572, 2007.

[8] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.ser. A*, 146(1), 2014.

[9] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[10] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[11] Aurélie Bugeau and Vinh-Thong Ta. Patch-based image colorization. In *International Conference on Pattern Recognition*, pages 3058–3061. IEEE, 2012.

[12] Aurélie Bugeau, Vinh-Thong Ta, and Nicolas Papadakis. Variational exemplar-based image colorization. *IEEE Transactions on Image Processing*, 23(1):298–307, 2014.

[13] Martin Burger, Hendrik Dirks, and Carola-Bibiane Schonlieb. A variational model for joint motion estimation and image reconstruction. *SIAM Journal on Imaging Sciences*, 11(1):94–128, 2018.

[14] Yair Censor and Stavros Andrea Zenios. Proximal minimization algorithm withd-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

[15] Antonin Chambolle, Pauline Tan, and Samuel Vaiter. Accelerated alternating descent methods for Dykstra-like problems. *Journal of Mathematical Imaging and Vision*, 2017.

[16] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *European Conference on Computer Vision*, pages 126–139. Springer, 2008.

[17] Tongbo Chen, Yan Wang, Volker Schillings, and Christoph Meinel. Grayscale image matting and colorization. In *Proceedings of Asian Conference on Computer Vision*, volume 6. Citeseer, 2004.

[18] Sergio Conti, Janusz Ginster, and Martin Rumpf. A bv functional and its relaxation for joint motion estimation and image sequence recovery. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(5):1463–1487, 2015.

[19] Zofia Denkowska and Maciej P Denkowski. A long and winding road to definable sets. *Journal of Singularities*, 13:57–86, 2015.

[20] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Aggregation of local parametric candidates with exemplar-based occlusion handling for optical flow. *Computer Vision and Image Understanding*, 145:81–94, 2016.

[21] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. A variational aggregation framework for patch-based optical flow estimation. *Journal of Mathematical Imaging and Vision*, 56(2):280–299, 2016.

[22] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

[23] Osman Güler. On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419, 1991.

[24] Raj Kumar Gupta, Alex Yong-Sang Chia, Deepu Rajan, Ee Sin Ng, and Huang Zhiyong. Image colorization using similar images. In *ACM International Conference on Multimedia*, pages 369–378, 2012.

[25] Takahiko Horiuchi. Colorization algorithm using probabilistic relaxation. *Image and Vision Computing*, 22(3):197–202, 2004.

[26] Takahiko Horiuchi and Sayaka Hirano. Colorization algorithm for grayscale image by propagating seed pixels. In *International Conference on Image Processing*, volume 1, pages I–457. IEEE, 2003.

[27] Revital Ironi, Daniel Cohen-Or, and Dani Lischinski. Colorization by example. In *Rendering Techniques*, pages 201–210. Citeseer, 2005.

[28] Sung Ha Kang and Riccardo March. Variational models for image colorization via chromaticity and brightness decomposition. *IEEE Transactions on Image Processing*, 16(9):2251–2261, 2007.

[29] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. In *ACM Transactions on Graphics*, volume 23, pages 689–694, 2004.

[30] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[31] Mila Nikolova and Pauline Tan. Alternating proximal gradient descent for nonconvex nonsmooth block-regularised problems, 2018. Submitted.

[32] Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[33] Johannes Persch, Fabien Pierre, and Gabriele Steidl. Exemplar-based face colorization using image morphing. *Journal of Imaging*, 3(4):48, 2017.

[34] Fabien Pierre, J-F Aujol, Aurélie Bugeau, Nicolas Papadakis, and V-T Ta. Luminance-chrominance model for image colorization. *SIAM Journal on Imaging Sciences*, 8(1):536–563, 2015.

[35] Fabien Pierre, J-F Aujol, Aurélie Bugeau, and V-T Ta. Interactive video colorization within a variational framework. *SIAM Journal on Imaging Sciences*, 10(4):2293–2325, 2017.

[36] Fabien Pierre, Jean-François Aujol, Aurélie Bugeau, and Vinh-Thong Ta. Combinaison linéaire optimale de métriques pour la colorisation d'images. In *XXVème colloque GRETSI*, pages 1–4, 2015.

[37] Fabien Pierre, Jean-François Aujol, Aurélie Bugeau, and Vinh-Thong Ta. Colociel. Dépôt Agence de Protection des Programmes No IDDN.FR.001.080021.000.S.P.2016.000.2100, 2016. Available online at http://www.labri.fr/perso/fpierre/colociel_v1.zip.

[38] Fabien Pierre, Jean-François Aujol, Aurélie Bugeau, Vinh-Thong Ta, and Nicolas Papadakis. Exemplar-based colorization in RGB color space. In *International Conference on Image Processing*, pages 1–5. IEEE, 2014.

[39] Thomas Pock and Shoham Sabach. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9(4):1756–1787, 2016.

[40] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

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

[42] Masahiro Shiota. *Geometry of subanalytic and semialgebraic sets*, volume 150. Springer Science & Business Media, 2012.

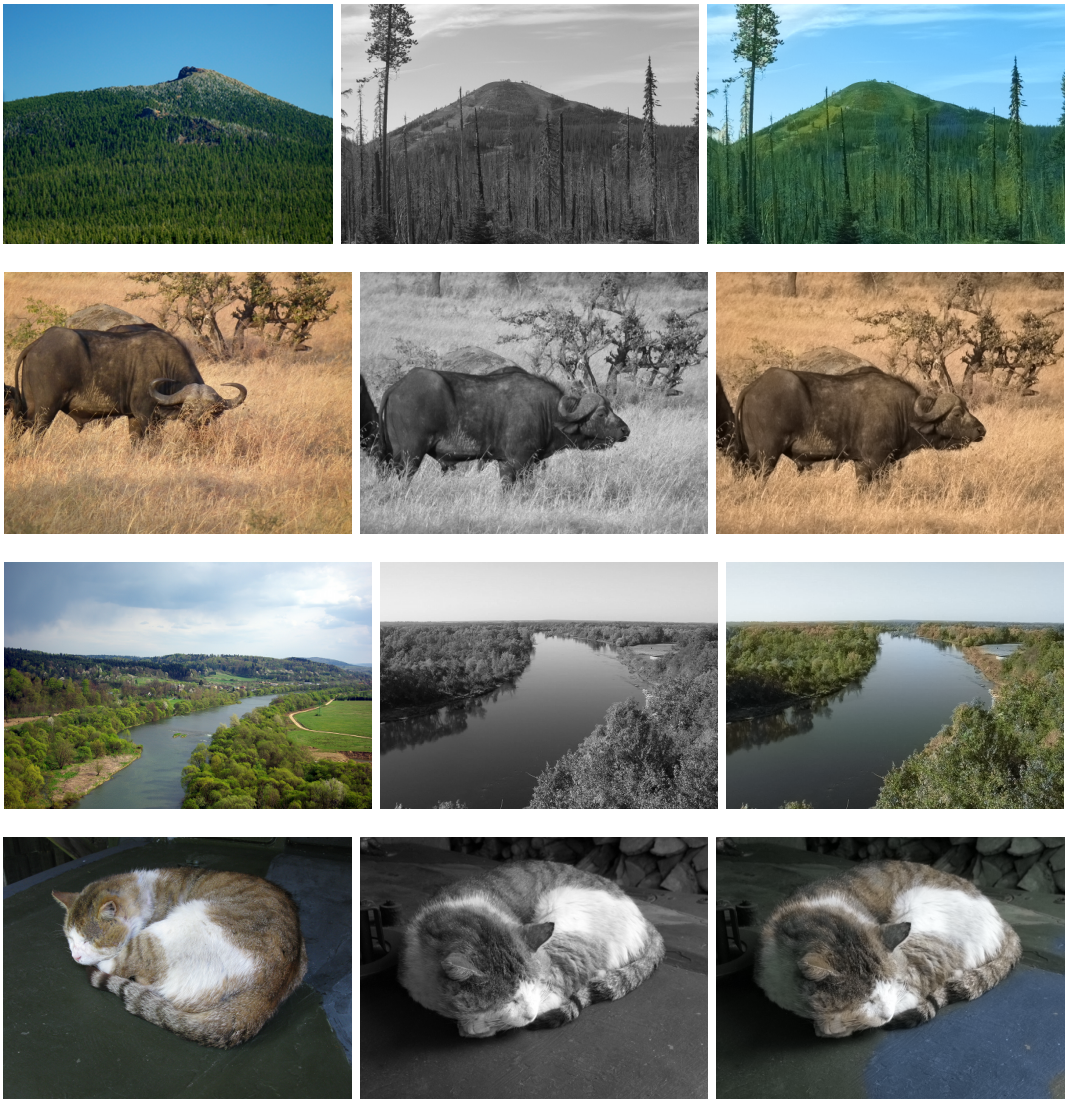
[43] Pauline Tan, Yann Ferrec, and Laurent Rousset-Rouvière. Correction par méthode variationnelle des non uniformités des détecteurs d'un interféromètre imageur. In *XXVIe colloque GRETSI 2017*, 2017.

[44] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. In *ACM Transactions on Graphics*, volume 21, pages 277–280, 2002.

[45] 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(3):1758–1789, 2013.

[46] Liron Yatziv and Guillermo Sapiro. Fast image and video colorization using chrominance blending. *IEEE Transactions on Image Processing*, 15(5):1120–1129, 2006.

[47] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.



Source image. Target image Result

Figure 3: Results with our strategy. The result is first computed with the method of [34] and refined with Algorithm 2. With this strategy, the speed of convergence is roughly the same as in [34] and the convergence is theoretically guaranteed.