Optimization Letters (2020) 14:1071–1088
https://doi.org/10.1007/s11590-019-01388-y

Alternating forward–backward splitting for linearly constrained optimization problems

Cesare Molinari[1] · Juan Peypouquet[2] · Fernando Roldan[3]

Received: 26 December 2017 / Accepted: 7 January 2019 / Published online: 17 January 2019

Abstract
We present an alternating forward–backward splitting method for solving linearly constrained structured optimization problems. The algorithm takes advantage of the separable structure and possibly asymmetric regularity properties of the objective functions involved. We also describe some applications to the study of non-Newtonian fluids and image reconstruction problems. We conclude with a numerical example, and its comparison with Condat's algorithm. An acceleration heuristic is also briefly outlined.

Keywords Convex optimization · Forward–backward splitting · Structured problems

1 Introduction

The gradient method [19] and the proximal point algorithm [45,54] are cornerstones in convex programming. Qualitatively, they have analogous asymptotic properties but have noticeable differences in terms of their convergence analysis, implementation, stability and several quantitative aspects, which result in distinct potential application frameworks. Gradient-type methods are defined by an explicit formula, and it is fairly

✉ Juan Peypouquet
jpeypou@dim.uchile.cl

Cesare Molinari
cecio.molinari@gmail.com

Fernando Roldan
fdoroldan94@gmail.com

1 ENSICAEN, CNRS, GREYC, Normandie Université, 6, boulevard du Maréchal Juin - CS, 45053 - 14050 Caen Cedex 4, France

2 Departamento de Ingeniería Matemática, Centro de Modelamiento Matemático (CNRS UMI2807), FCFM, Universidad de Chile, Beauchef 851, Santiago, Chile

3 Departamento de Matemática, Universidad Técnica Federico Santa María, Avenida España, 1680 Valparaiso, Chile

straightforward to implement line-search routines to determine efficient step sizes. The price to pay is that, typically, convergence can only be proved under restrictive smoothness assumptions. In particular, the step sizes are limited by a function of the Lipschitz constant of the gradient, and it is not surprising that small steps may lead to slower convergence. In turn, proximal-type methods enjoy a stable evolution and tend to converge under much weaker assumptions, which makes them suitable for nonsmooth functions. In general, their implementation relies on solving an auxiliary subproblem at each iteration. However, the latter can be solved explicitly in several relevant and frequent situations. Over the last decades, gradient and proximal iterations have been refined, improved and combined in order to produce sophisticated first-order methods, which are both versatile and efficient. A sufficiently thorough but very accessible introduction to gradient and proximal algorithms can be found in [51, Chapter 6] (see also [28,50] for a more detailed discussion).

A number of convex programming problems arising, for instance, in signal/image processing, optimal control, statistics and mechanics, involves the minimization of functions with a *smooth* + *nonsmooth* structure. In this scenario, a reasonable strategy is to apply a *proximal-gradient splitting* paradigm to take advantage of the particular properties of each component, using gradient-type iterations for the smooth part, and proximal-type iterations for the nonsmooth one. The *projected gradient method* (see, for instance [38,44]) and the *iterative shrinkage thresholding algorithm* (ISTA) (see [30,32]) are well-known examples.

Another frequent situation is the presence of an additive structure in a product space: the objective function can be expressed as the sum of two (or more) functions, each depending on different sets of variables, all of which are coupled by some relation. This leads to a natural *componentwise splitting*, where minimization with respect to each set of variables can be treated independently (either in series or in parallel). The coupling is interpreted as a constraint, which can be dealt with by means of projection, penalization or Lagrange multiplier techniques, among others. To the best of our knowledge, most existing algorithms treat the different summands in the objective function in a symmetric manner. For instance, if no differentiability is assumed, one may be inclined to using a proximal approach for each set of variables (see [4,22,25, 27,35], to cite a few). With this strategy possible differences in terms of regularity are not taken into account and, therefore, not exploited. A framework allowing to take advantage of possible dissymmetries will be useful in order to design efficient, versatile and easily implementable algorithms. A first step in this direction was given in [29], but the decissive leap was given by in [31] (Condat's algorithm, in the sequel), and further developed in [60] and also [15]. Their algorithm and its convergence analysis are in the line of (relaxed) fixed point iterations, following the Krasnoselskii–Mann scheme. This work provides an alternative approach, which is intrinsically different. We focus on the case of two functions whose variables are coupled by an affine constraint, for which we use a penalization procedure, as in [9] (see also [5,6,16]). Our method is inspired by the (purely proximal) method studied in [4], which is closely connected to the ones considered in [1,2,8,10], but in a time-varying context. Our method shares some features with the one in [31], namely: (1) exploits the fact that the components involved in the objective function may be of a different nature; (2) uses the structured character of the problem, reducing the size and complexity of the sub-problem solved at each iteration; (3) avoids unnecessary operator inversions; and (4) combines gradient and proximal iterations with a penalization scheme in a simple way. An alternative approach using Lagrange multipliers is given in [46]. It is important to say that, in its full generality, the computation of the proximal subiteration requires an inner loop, for which some approximation techniques are available (see [26,33,36,41,42,52,56]).

This paper is organized as follows: in Sect. 2, we describe the precise framework and the proposed algorithm, and state the main theoretical result. Section 3 contains the convergence analysis, showing that the sequences generated by the algorithm converge to solutions of the problem. Moreover, in spite of the penalization scheme involved, we are able to obtain a worst-case convergence rate in terms of a Bregman-type divergence. In Sect. 4, we illustrate the application of our method. First, we briefly comment some examples in the study of non-Newtonian fluids and image/signal processing (readers interested in dynamical games are referred to [2]). Next, we provide a numerical implementation, along with a comparison with the algorithm in [31], in a medium size image restoration problem, which is a basic, yet quite informative, example. Although there are cases where the presence of parameters tending to ∞ might lead to numerical instabilities, we did not witness this in our tests. We finish by outlining a possible acceleration heuristic, whose convergence proof can be a subject for future research.

2 Algorithm, hypotheses and main result

Let X and Y and Z be real Hilbert spaces. We are interested in the numerical approximation of solutions for the following problem:

$$\text{Find } (x^*, y^*) \in \operatorname{Argmin}\{f(x) + g(y) : (x, y) \in C\}, \tag{$\mathcal{P}$}$$

where $f : X \to \mathbf{R}$ is a smooth convex function; $g : Y \to \mathbf{R} \cup \{+\infty\}$ is proper, lower-semicontinuous and convex; and the set $C \subset X \times Y$ is defined by

$$C = \{(x, y) \in X \times Y : Ax + By = c\},$$

for bounded linear operators $A : X \to Z$ and $B : Y \to Z$, and a vector $c \in Z$.

The method described here is a proximal-gradient algorithm with two-step iterations based on the alternating minimization of the function

$$\mathcal{L}_\lambda(x, y) = f(x) + g(y) + \frac{\gamma}{2\lambda}\|Ax + By - c\|^2$$

with respect to the x and y variables at each iteration: Given (x_n, y_n), start by performing a gradient iteration with respect to the x variable using the function $\mathcal{L}_{\lambda_n}$ along with λ_n as a step size. This is simply

$$x_{n+1} = x_n - \lambda_n \nabla f(x_n) - \gamma A^*(Ax_n + By_n - c) \tag{1}$$

or, equivalently,

$$x_{n+1} - x_n + \gamma A^*(Ax_n + By_n - c) = -\lambda_n \nabla f(x_n).$$

Next, the idea is to apply a proximal iteration with respect to the y variable, namely

$$y_{n+1} = \operatorname{Argmin}_{y \in Y} \left\{ \mathcal{L}_{\lambda_n}(x_{n+1}, y) + \frac{1}{2\lambda_n} \|y - y_n\|^2 \right\}, \tag{2}$$

whose optimality condition can be written as

$$y_{n+1} - y_n + \gamma B^*(Ax_{n+1} + By_{n+1} - c) \in -\lambda_n \partial g(y_{n+1}).$$

However, although the auxiliary subproblem given by (2) is strongly convex, it must be solved inexactly in practice. Computational errors in this iteration can take a number of forms: approximate minimization in (2), relaxed optimality condition based on the approximate subdifferential, or proximity to the exact solution. Here, we use the second option, which is intermediate in terms of computational complexity, but has sufficiently strong geometric properties. More precisely, we require y_{n+1} to satisfy

$$y_{n+1} - y_n + \gamma B^*(Ax_{n+1} + By_{n+1} - c) \in -\lambda_n \partial_{\eta_n} g(y_{n+1}), \tag{3}$$

where

$$\partial_\eta g(\bar{y}) = \{ \upsilon \in Y : g(y) \geq g(\bar{y}) + \langle \upsilon, y - \bar{y} \rangle - \eta \quad \text{for all} \quad y \in Y \} \quad \text{for } \eta \geq 0.$$

Problem $(\mathcal{P})$ can also be solved, for instance, by means of Douglas–Rachford [35] or forward-Douglas–Rachford methods [14] (see also [31]). Their implementation relies on the computation of the pseudo-inverses of the operators involved, a procedure that may be costly for large problems. Our method does not involve operator inversion, unless the proximal step were to be computed exactly, which would result in a standard proximal iteration with respect to an auxiliary metric involving B. It is purely primal, so the dimension of the problem is not increased, as it is in dual, primal-dual or Lagrangian methods. Although it includes a penalization scheme, the worst-case convergence rate is consistent with that of standard first-order methods (see Sect. 3.1). A simple numerical experiment, reported in Sect. 4, suggests that it admits an accelerated variant, but we shall not address the theoretical aspects here.

The main convergence result relies on the following set of assumptions:

Hypothesis (H):

- $(\mathbf{H}_1)$ The function $f : X \to \mathbf{R}$ is convex, differentiable and its gradient ∇f is Lipschitz-continuous with constant L. The function $g : Y \to \mathbf{R} \cup \{+\infty\}$ is proper, lower-semicontinuous and convex. The operators $A : X \to Z$ and $B : Y \to Z$ are linear and continuous.
- $(\mathbf{H}_2)$ The solution set $\mathcal{S}$ is nonempty, and $(x^*, y^*) \in \mathcal{S}$ if, and only if, there exists $z^* \in Z$ such that (x^*, y^*, z^*) satisfies

$$\begin{cases} -A^*z^* = \nabla f(x^*) \\ -B^*z^* \in \partial g(y^*) \\ c = Ax^* + By^*. \end{cases} \tag{4}$$

($\mathbf{H}_3$) The step sizes (λ_n) form a nonincreasing sequence in ℓ^2 such that $\sup_{n\in\mathbf{N}} \left(\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right) < +\infty$.

($\mathbf{H}_4$) The relaxation parameter verifies $\gamma \in \left(0, \frac{2}{\|A\|^2}\right)$.

($\mathbf{H}_5$) The error sequence (η_n) belongs to ℓ^1.

In ($\mathbf{H}_1$), a line search method could be used as an alternative if the Lipschitz constant is not available, but we shall not follow this line of research here. A sufficient condition for ($\mathbf{H}_2$) is that $R(A) + R(B)$ be closed in Z, in view of the Moreau–Rockafellar theorem [47,53]. It can also be obtained using the Attouch–Brézis qualification condition [3]. Hypothesis ($\mathbf{H}_3$) implies the sequence (λ_n) is not in ℓ^1 and, therefore, it does not converge to zero too fast. It holds for $\lambda_n = n^{-q}$ with $1/2 < q \leq 1$. In ($\mathbf{H}_4$), as one should expect, the relaxation parameter does not depend on B, which enters in the proximal subiteration.

Our main theoretical result is the following:

Theorem 1 *Let hypothesis* ($\mathbf{H}$) *hold. Every sequence* (x_n, y_n) *satisfying* (1) *and* (3) *converges weakly, as* $n \to +\infty$, *to a point in* $\mathcal{S}$.

3 Convergence analysis and rate

Fix (x^*, y^*, z^*) satisfying (4) and set

$$\begin{cases} D_n = \|x_n - x^*\|^2 + \|y_{n-1} - y^*\|^2 \\ \alpha_n = \|B\|^2\,|\lambda_{n-1} - \lambda_n| \\ \varepsilon_n = \left[\frac{\lambda_n^2}{\gamma} + |\lambda_{n-1} - \lambda_n|\right] \|z^*\|^2 + 2\lambda_{n-1}\eta_{n-1}. \end{cases} \tag{5}$$

If (λ_n) is a nonincreasing sequence in ℓ^2 then (α_n) and (ε_n) belong to ℓ^1.

Lemma 2 *Let hypothesis* ($\mathbf{H}_1$) *hold and let* (x^*, y^*, z^*) *satisfy* (4). *We have*

$$D_{n+1} - D_n + \gamma\|Ax_n + By_n - c\|^2 \leq \|x_{n+1} - x_n\|^2 + \alpha_n\|y_n - y^*\|^2 + \varepsilon_n. \tag{6}$$

Proof From the monotonicity of ∇f we have

$$\langle x_{n+1} - x_n + \gamma A^*(Ax_n + By_n - c) - \lambda_n A^*z^*, x_n - x^*\rangle \leq 0.$$

This is equivalent to

$$\begin{aligned} &\|x_{n+1} - x^*\|^2 - \|x_n - x^*\|^2 - \|x_{n+1} - x_n\|^2 + 2\gamma\langle Ax_n + By_n - c, A(x_n - x^*)\rangle \\ &\quad \leq 2\lambda_n\langle z^*, A(x_n - x^*)\rangle. \end{aligned} \tag{7}$$

In a similar fashion, the quasi-monotonicity of $\partial_{\eta_n} g$ gives

$$\begin{aligned}&\|y_n - y^*\|^2 - \|y_{n-1} - y^*\|^2 + \|y_n - y_{n-1}\|^2 + 2\gamma\langle Ax_n + By_n - c, B(y_n - y^*)\rangle \\ &\quad \leq 2\lambda_{n-1}\langle z^*, B(y_n - y^*)\rangle + 4\lambda_{n-1}\eta_{n-1}. \end{aligned} \tag{8}$$

Keeping in mind that $Ax^* + By^* = c$, add inequalities (8) and (7) to deduce that

$$\begin{aligned}&\|x_{n+1} - x^*\|^2 - \|x_n - x^*\|^2 + \|y_n - y^*\|^2 - \|y_{n-1} - y^*\|^2 \\ &\quad +2\gamma\|Ax_n + By_n - c\|^2 - \|x_{n+1} - x_n\|^2 \\ &\quad \leq 2\lambda_n\langle z^*, A(x_n - x^*)\rangle + 2\lambda_{n-1}\langle z^*, B(y_n - y^*)\rangle + 4\lambda_{n-1}\eta_{n-1}. \end{aligned} \tag{9}$$

But

$$\begin{aligned}&2\lambda_n\langle z^*, A(x_n - x^*)\rangle + 2\lambda_{n-1}\langle z^*, B(y_n - y^*)\rangle \\ &\quad = 2\lambda_n\langle z^*, Ax_n + By_n - c\rangle + 2(\lambda_{n-1} - \lambda_n)\langle z^*, B(y_n - y^*)\rangle \\ &\quad \leq \left[\frac{\lambda_n^2}{\gamma} + |\lambda_{n-1} - \lambda_n|\right]\|z^*\|^2 + \gamma\|Ax_n + By_n - c\|^2 \\ &\qquad +\|B\|^2\,|\lambda_{n-1} - \lambda_n|\,\|y_n - y^*\|^2. \end{aligned} \tag{10}$$

Inequality (6) is obtained by combining (9) and (10) and using the definitions in (5). □

Now set

$$E_n = f(x_n) + g(y_n) + \frac{\gamma}{2\lambda_n}\|Ax_n + By_n - c\|^2$$

and

$$L_n = L + \frac{\gamma\|A\|^2}{\lambda_n}.$$

Lemma 3 *Let hypothesis* $(\mathbf{H}_1)$ *hold. We have*

$$\begin{aligned}E_{n+1} - E_n + \frac{1}{\lambda_n}\|y_{n+1} - y_n\|^2 &\leq \left[\frac{L_n}{2} - \frac{1}{\lambda_n}\right]\|x_{n+1} - x_n\|^2 \\ &\quad +\frac{\gamma}{2}\left[\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right]\|Ax_{n+1} + By_{n+1} - c\|^2 + \eta_n. \end{aligned} \tag{11}$$

Proof Since ∇f is Lipschitz-continuous with constant L, we have (see [51, Lemma 1.30])

$$f(z) \leq f(x) + \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|^2$$

for all $x, z \in X$. Using this inequality with $z = x_{n+1}$ and $x = x_n$ we deduce that

$$
\begin{aligned}
& f(x_{n+1}) - f(x_n) \\
& \quad \le \frac{1}{\lambda_n}\langle x_{n+1} - x_n + \gamma A^*(Ax_n + By_n - c), x_n - x_{n+1}\rangle + \frac{L}{2}\|x_{n+1} - x_n\|^2 \\
& \quad = \left[\frac{L}{2} - \frac{1}{\lambda_n}\right]\|x_{n+1} - x_n\|^2 + \frac{\gamma}{\lambda_n}\langle Ax_n + By_n - c, A(x_n - x_{n+1})\rangle. \qquad (12)
\end{aligned}
$$

But

$$
\begin{aligned}
& 2\langle Ax_n + By_n - c, A(x_n - x_{n+1})\rangle \\
& \quad = \|Ax_n + By_n - c\|^2 + \|A(x_{n+1} - x_n)\|^2 - \|Ax_{n+1} + By_n - c\|^2 \\
& \quad \le \|Ax_n + By_n - c\|^2 + \|A\|^2\|x_{n+1} - x_n\|^2 - \|Ax_{n+1} + By_n - c\|^2 \qquad (13)
\end{aligned}
$$

and so

$$
\begin{aligned}
f(x_{n+1}) - f(x_n) \le & \left[\frac{L}{2} + \frac{\gamma\|A\|^2}{2\lambda_n} - \frac{1}{\lambda_n}\right]\|x_{n+1} - x_n\|^2 \\
& + \frac{\gamma}{2\lambda_n}\|Ax_n + By_n - c\|^2 - \frac{\gamma}{2\lambda_n}\|Ax_{n+1} + By_n - c\|^2. \qquad (14)
\end{aligned}
$$

On the other hand, the subdifferential inequality for g gives

$$
g(y_n) \ge g(y_{n+1}) + \frac{1}{\lambda_n}\langle y_{n+1} - y_n - \gamma B^*(Ax_{n+1} + By_{n+1} - c), y_{n+1} - y_n\rangle - \eta_n,
$$

which we rewrite as

$$
\begin{aligned}
& g(y_{n+1}) - g(y_n) + \frac{1}{\lambda_n}\|y_{n+1} - y_n\|^2 \\
& \quad \le \frac{\gamma}{\lambda_n}\langle Ax_{n+1} + By_{n+1} - c, B(y_n - y_{n+1})\rangle + \eta_n. \qquad (15)
\end{aligned}
$$

Since

$$
\begin{aligned}
& 2\langle Ax_{n+1} + By_{n+1} - c, B(y_n - y_{n+1})\rangle \\
& \quad = \|Ax_{n+1} + By_n - c\|^2 - \|Ax_{n+1} + By_{n+1} - c\|^2 - \|B(y_{n+1} - y_n)\|^2 \\
& \quad \le \|Ax_{n+1} + By_n - c\|^2 - \|Ax_{n+1} + By_{n+1} - c\|^2 \qquad (16)
\end{aligned}
$$

we obtain

$$
\begin{aligned}
& g(y_{n+1}) - g(y_n) + \frac{1}{\lambda_n}\|y_{n+1} - y_n\|^2 \\
& \quad \le \frac{\gamma}{2\lambda_n}\|Ax_{n+1} + By_n - c\|^2 - \frac{\gamma}{2\lambda_n}\|Ax_{n+1} + By_{n+1} - c\|^2 + \eta_n. \qquad (17)
\end{aligned}
$$

Finally add (14) and (17) to obtain (11). □

Remark 4 For (11), it suffices that ∇f be L-Lipschitz-continuous on the segment joining x_n and x_{n+1}.

Corollary 5 *Let hypothesis* $(\mathbf{H}_1)$ *hold and let* $K > 0$. *If* $\left[\frac{L_n}{2} - \frac{1}{\lambda_n}\right] \leq -K$ *and* $\left[\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right] \leq K$, *then*

$$E_{n+1} - E_n + \frac{1}{\lambda_n}\|y_{n+1} - y_n\|^2 + K\|x_{n+1} - x_n\|^2 \leq \frac{\gamma K}{2}\|Ax_{n+1} + By_{n+1} - c\|^2 + \eta_n. \tag{18}$$

Define

$$P_n = D_n + \frac{2}{K}\left[f(x_n) + g(y_n)\right] + \gamma\left[\frac{1}{K\lambda_n} - 1\right]\|Ax_n + By_n - c\|^2 + \frac{2}{K\lambda_{n-1}}\|y_n - y_{n-1}\|^2.$$

Multiplying (18) by $2/K$ and adding the result to (6) we obtain

Corollary 6 *Let hypotheses* $(\mathbf{H}_1)$ *hold and let* $K > 0$. *If* $\left[\frac{L_n}{2} - \frac{1}{\lambda_n}\right] \leq -K$ *and* $\left[\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right] \leq K$, *then*

$$P_{n+1} - P_n + \|x_{n+1} - x_n\|^2 \leq \alpha_n\|y_n - y^*\|^2 + \varepsilon_n + \frac{2\eta_n}{K}. \tag{19}$$

Notice that the condition $\left[\frac{L_n}{2} - \frac{1}{\lambda_n}\right] \leq -K$ can be rewritten as $\lambda_n \leq \frac{2-\gamma\|A\|^2}{L+2K}$. Then, the existence of $K > 0$ satisfying the hypotheses of Corollaries 5 and 6 is implied by $(\mathbf{H}_1)$, $(\mathbf{H}_3)$ and $(\mathbf{H}_4)$.

The following result concerning real sequences will be useful to prove Proposition 8 below.

Lemma 7 *Let* (p_n), (r_n), (a_n) *and* (e_n) *be nonnegative sequences such that* (a_n) *and* (e_n) *belong to* ℓ^1, *and*

$$(1 - a_n)p_{n+1} - p_n + r_n \leq e_n \tag{20}$$

for all sufficiently large n. *Then* (p_n) *is convergent and* (r_n) *belongs to* ℓ^1.

Proof Inequality (20) can be written as

$$p_{n+1} - p_n + r_n \leq e_n + a_n p_{n+1},$$

so it suffices to show that (p_n) is bounded from above. But, $1 - a_n > 0$ for large n, because $\lim_{n\to\infty} a_n = 0$. We have

$$p_{n+1} \leq \frac{1}{1 - a_n}\left[p_n + e_n\right].$$

By induction, and observing that $(1-a_n)^{-1} \le \prod_{k=1}^{n}(1-a_k)^{-1}$, we obtain

$$p_{n+1} \le \left[\prod_{k=1}^{n}(1-a_k)^{-1}\right]\left[p_1 + \sum_{k=1}^{n} e_k\right].$$

Since (a_n) and (e_n) belong to ℓ^1, the right-hand side is bounded, and so is (p_n) . □

We are now in a position to establish the asymptotic behavior, as $n \to \infty$, of the auxiliary sequences (D_n), (E_n) and (P_n), namely:

Proposition 8 *Let hypotheses* $(\mathbf{H}_1)$, $(\mathbf{H}_3)$ *and* $(\mathbf{H}_4)$ *hold, and let* (x^*, y^*, z^*) *satisfy* (4). *Then*

(i) *The sequence* (P_n) *is convergent and* $\sum \|x_{n+1} - x_n\|^2 < +\infty$.
(ii) *The sequence* (D_n) *is convergent and* $\sum \|Ax_n + By_n - c\|^2 < +\infty$.
(iii) *The sequence* (x_n, y_n) *is bounded and every weak cluster point lies in* C.
(iv) *The sequence* (E_n) *is convergent and* $\sum \frac{1}{\lambda_n}\|y_n - y_{n-1}\|^2 < +\infty$.

Proof Let $K > 0$ such that $\left[\frac{1}{\lambda_{n+1}} - \frac{1}{\lambda_n}\right] \le K$ for all $n \in \mathbf{N}$. Since $\lim_{n\to\infty} \lambda_n = 0$, for all sufficiently large n we have $\left[\frac{L_n}{2} - \frac{1}{\lambda_n}\right] \le -K$. Hence, inequalities (18) and (19) hold for all sufficiently large n.

First observe that $\frac{2}{K} f(x_{n+1}) + \|x_{n+1} - x^*\|^2$ is bounded from below by a constant, say P_1. In turn, since

$$\frac{1}{4}\|y_{n+1} - y^*\|^2 \le \frac{1}{2}\|y_{n+1} - y_n\|^2 + \frac{1}{2}\|y_n - x^*\|^2$$

and $\lim_{n\to\infty} \lambda_n = 0$, eventually $\lambda_n \le \frac{4}{K}$ and we have

$$\begin{aligned}
&\frac{2}{K} g(y_{n+1}) + \|y_n - x^*\|^2 + \frac{2}{K\lambda_n}\|y_{n+1} - y_n\|^2 \\
&\quad \ge \frac{2}{K} g(y_{n+1}) + \frac{1}{2}\|y_n - x^*\|^2 + \frac{1}{2}\|y_{n+1} - y_n\|^2 + \frac{1}{2}\|y_n - x^*\|^2 \\
&\quad \ge \frac{2}{K} g(y_{n+1}) + \frac{1}{4}\|y_{n+1} - x^*\|^2 + \frac{1}{2}\|y_n - x^*\|^2.
\end{aligned}$$

The sum of the first two terms is bounded from below by a constant P_2. Finally, once λ_n is sufficiently small, we have $\gamma\left[\frac{1}{K\lambda_n} - 1\right]\|Ax_n + By_n - c\|^2 \ge 0$. Writing $P = P_1 + P_2$ we see that

$$P_{n+1} \ge P + \frac{1}{2}\|y_n - x^*\|^2 \ge P$$

for all sufficiently large n. As a consequence, inequality (19) gives

$$(1 - 2\alpha_n)(P_{n+1} - P) - (P_n - P) + \|x_{n+1} - x_n\|^2 \le \varepsilon_n + \frac{2\eta_n}{K}.$$

To prove (i), it suffices to use Lemma 7 with $p_n = P_n - P$, $r_n = \|x_{n+1} - x_n\|^2$, $a_n = 2\alpha_n$ and $e_n = \varepsilon_n + \frac{2\eta_n}{K}$.

For (ii), observe that inequality (6) implies

$$(1-\alpha_n)D_{n+1} - D_n + \gamma\|Ax_n + By_n - c\|^2 \le \|x_{n+1} - x_n\|^2 + \varepsilon_n.$$

Use Lemma 7 with $p_n = D_n$, $r_n = \gamma\|Ax_n + By_n - c\|^2$, $a_n = \alpha_n$ and $e_n = \|x_{n+1} - x_n\|^2 + \varepsilon_n$.

Part (iii) follows from (ii) along with the weak lower-semicontinuity of the function $(x, y) \mapsto \|Ax + By - c\|^2$.

Finally, in order to verify (iv), notice that inequality (18) gives

$$E_{n+1} - E_n + \frac{1}{\lambda_n}\|y_{n+1} - y_n\|^2 \le \frac{\gamma K}{2}\|Ax_{n+1} + By_{n+1} - c\|^2 + \eta_n.$$

Apply Lemma 7 with $p_n = E_n$, $r_n = \frac{1}{\lambda_n}\|y_{n+1} - y_n\|^2$, $a_n = 0$ and $e_n = \frac{\gamma K}{2}\|Ax_{n+1} + By_{n+1} - c\|^2 + \eta_n$. □

Proposition 9 *Let hypotheses* ($\mathbf{H}_1$), ($\mathbf{H}_3$) *and* ($\mathbf{H}_4$) *hold, and let* (x^*, y^*, z^*) *satisfy* (4). *Then*

$$\sum_{n=1}^{\infty} \lambda_n \left[E_n - f(x^*) - g(y^*)\right] < \infty \qquad (\textit{possibly } -\infty). \tag{21}$$

Proof The subdifferential inequality for f gives

$$f(x^*) \ge f(x_n) + \langle x_{n+1} - x_n + \gamma A^*(Ax_n + By_n - c), x_n - x\rangle.$$

Therefore,

$$\begin{aligned}\lambda_n\left[f(x_n) - f(x^*)\right] \le \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2 + \|x_{n+1} - x_n\|^2 \\ -\gamma\langle Ax_n + By_n - c, A(x_n - x^*)\rangle.\end{aligned}$$

In a similar fashion, the subdifferential inequality for g produces

$$\begin{aligned}\lambda_{n-1}\left[g(y_n) - g(y^*)\right] \le \|y_{n-1} - y^*\|^2 - \|y_n - y^*\|^2 - \|y_n - y_{n-1}\|^2 \\ -\gamma\langle Ax_n + By_n - c, B(y_n - y^*)\rangle + \lambda_{n-1}\eta_{n-1}.\end{aligned}$$

Adding these two inequalities, recalling the definition of E_n, using the fact that $Ax^* + By^* = c$, and neglecting the nonpositive terms on the right-hand side, we deduce that

$$\begin{aligned}\lambda_n\left[E_n - f(x^*) - g(y^*)\right] \le D_n - D_{n+1} + \|x_{n+1} - x_n\|^2 \\ +(\lambda_{n-1} - \lambda_n)(g(y^*) - g(y_n)) + \lambda_{n-1}\eta_{n-1}.\end{aligned}$$

But (y_n) is bounded because (D_n) converges. Since g is convex, $(g(y_n))$ must be bounded from below by some $\mu \in \mathbf{R}$. The fact that (λ_n) is nonincreasing implies

$$\lambda_n \left[E_n - f(x^*) - g(y^*)\right] \leq D_n - D_{n+1} + \|x_{n+1} - x_n\|^2 + (\lambda_{n-1} - \lambda_n)(g(y^*) - \mu) + \lambda_{n-1}\eta_{n-1}. \quad (22)$$

The summability of the terms on the right-hand side immediately gives the result. □

Proof of Theorem 1 Assume (x_{k_n}, y_{k_n}) converges weakly to (x_∞, y_∞). From part (iii) of Proposition 8 we have $(x_\infty, y_\infty) \in C$. The weak lower-semicontinuity of f and g, along with Proposition 9, together imply

$$f(x_\infty) + g(x_\infty) \leq \lim_{n\to\infty} E_n \leq f(x) + g(y)$$

for all $(x, y) \in C$. That is, weak cluster points belong to $\mathcal{S}$. Since the sequence (D_n) is convergent by part (ii) of Proposition 8, the result follows by Opial's Lemma [49] (as stated, for instance, in [51, Lemma 5.2]). □

Remark 10 In Theorem 1, the Lipschitz continuity assumption on ∇f is used only along the polygonal line Q connecting the iterates.

3.1 Convergence rate

Let (x^*, y^*, z^*) be a primal-dual solution and consider the following Bregman-type divergence:

$$d_{z^*}(x, y) = [f(x) + g(y)] - \left[f(x^*) + g(y^*)\right] + \langle z^*, Ax + By - c\rangle.$$

From the optimality conditions in (4), we know that $d_{z^*}(x, y) \geq 0$ for every $(x, y) \in X \times Y$. Moreover, a feasible point (x, y) is a (primal) solution if and only if $d_{z^*}(x, y) = 0$. So the quantity can be interpreted as a measure for the quality of the approximate solution (x, y). We have the following result about the convergence rate for the quantity $d_{z^*}(x_n, y_n)$:

Theorem 11 *Given a primal-dual solution (x^*, y^*, z^*), set*

$$C = \sum_{k=1}^{\infty} \lambda_k \left[E_k - f(x^*) - g(y^*)\right] + \frac{\|z^*\|^2}{2\gamma} \sum_{k=1}^{+\infty} \lambda_k^2 < +\infty^1 \quad \textit{and} \quad \sigma_n = \sum_{k=1}^{n} \lambda_k.$$

For each $n \geq 1$, the average value of d_{z^*} along the first n iterations satisfies

$$\frac{1}{\sigma_n} \sum_{k=1}^{n} \lambda_k d_{z^*}(x_k, y_k) \leq \frac{C}{\sigma_n}.$$

[1] This quantity is finite, in view of Proposition 9, and the fact that $(\lambda_n) \in \ell^2$ [an upper bound can be obtained from (22)].

In particular, the best and averaged iterates satisfy, respectively,

$$\min_{1\le k\le n} d_{z^*}(x_k, y_k) \le \frac{C}{\sigma_n} \qquad \textit{and} \qquad d_{z^*}\left(\frac{1}{\sigma_n}\sum_{k=1}^{n}\lambda_k(x_k, y_k)\right) \le \frac{C}{\sigma_n}.$$

Proof First, fix $k \in \mathbf{N}$. From the definition of E_k, to we have

$$\begin{aligned} d_{z^*}(x_k, y_k) &\le [f(x_k) + g(y_k)] - \left[f(x^*) + g(y^*)\right] + \frac{\lambda_k}{2\gamma}\|z^*\|^2 \\ &\quad + \frac{\gamma}{2\lambda_k}\|Ax_k + Bx_k - c\|^2 \\ &= E_k - f(x^*) - g(y^*) + \frac{\lambda_k}{2\gamma}\|z^*\|^2. \end{aligned}$$

Multiply by λ_k, and take the sum over $k = 1, \ldots, n$, to deduce that

$$\sum_{k=1}^{n}\lambda_k d_{z^*}(x_k, y_k) \le \sum_{k=1}^{n}\lambda_k\left[E_k - f(x^*) - g(y^*) + \frac{\lambda_k}{2\gamma}\|z^*\|^2\right] = C.$$

The rest of the proof is straightforward. □

4 Illustration

In this section, we specialize and test the behavior of the proposed method. As examples, we describe briefly a variational problem used in the analysis of non-Newtonian fluids and a classical family of models used in image/signal processing. We perform a numerical simulation on a medium-size problem, and compare the outcome with a well-known algorithm with similar features, namely Condat's method [31]. We finish by outlining a possible acceleration heuristic, whose convergence proof can be a subject for future research.

4.1 Two simple keynote examples

4.1.1 TV-minimization and the *p*-Laplace equation

Let Ω be a bounded domain in $\mathbf{R}^d$, and take $p \ge 1$ and $r > 0$. In the study of non-Newtonian fluids, one encounters the following variational problem:

$$\min_{u\in W_0^{1,p}(\Omega)}\left\{\frac{1}{p}\int_\Omega |\nabla u|^p\,dx - \int_\Omega hu\,dx + r\int_\Omega |\nabla u|\,dx\right\} \tag{23}$$

(see [40] and the references therein), where a *total variation* regularization term appears. A standard numerical approach to this problem is the following: given a suitable approximation matrix $D \in \mathbf{R}^{M \times N}$ for the gradient (typically, $M = N - 1$ or less), the corresponding discrete version is given by

$$\min_{x \in \mathbf{R}^N} \left\{ \frac{1}{p} \|Dx\|_p^p - \langle h, x \rangle + r\|y\|_1 \;:\; Dx = y \right\},$$

which is Problem $(\mathcal{P})$ with $f(x) = \frac{1}{p}\|Dx\|_p^p - \langle h, x \rangle$, $g(y) = r\|y\|_1$, $A = D$ and $B = -I$. If $p \geq 2$, the function f has a Lipschitz-continuous gradient on each bounded subset of $\mathbf{R}^N$. If $p \in [1, 2)$, the same is true provided one withdraws a subset of the form $\{x \in \mathbf{R}^N : \|Dx\|_p < \varepsilon\}$ (roughly, a neighborhood of the constant functions). The closedness condition for $R(A) + R(B)$ is clearly satisfied, and the algorithm is reduced to

$$\begin{aligned} x_{n+1} &= x_n - \lambda_n \|Dx_n\|^{p-2} D^* Dx_n - \gamma D^*(Dx_n - y_n) \\ y_{k+1} &= \operatorname{Argmin}_{y \in \mathbf{R}^N} \left\{ \|y\|_1 + \frac{1+\gamma}{2r\lambda_n} \left\| y - \frac{y_n + \gamma Dx_{n+1}}{1+\gamma} \right\|_2^2 \right\}, \end{aligned}$$

which is fully explicit, in terms of the *soft thresholding* operator.

4.1.2 $\ell^1 + \ell^2$ minimization

Consider the problem

$$\min_{x \in \mathbf{R}^N} \left\{ \frac{1}{2} \|Kx - h\|_2^2 + r\|Qx\|_1 \right\}, \tag{24}$$

where $r > 0$, $h \in \mathbf{R}^{M_1}$, $K \in \mathbf{R}^{M_1 \times N}$ and $Q \in \mathbf{R}^{M_2 \times N}$. In image/signal processing, the first term represents the *fidelity* of the observation Kx of the variable x to a given measurement h. The second term induces sparsity in a *representation* Qx of x. If $M_2 = N$ and Q is the identity matrix, this is a sparse linear regression, an unconstrained version of LASSO. In turn, if Q is a (full) wavelet transform, it induces sparsity in a given wavelet basis. In both cases, Q is invertible, and standard proximal-gradient algorithms can be applied. However, Q can also account for a partial wavelet transform (typically with $M_2 \ll N$), which enforces a minimal level of compression in the given basis. Alternatively, Qx can represent x in a redundant dictionary (see [17] and the references therein), where $M_2 > N$. On the other hand, if, as in the previous paragraph, Q is a discrete approximation of the gradient, we obtain the ROF model, introduced by Rudin et al. [55]. The total variation term admits solutions with high variations, only if these are also highly localized in space. In the last three cases, Q is not invertible, and the term $r\|Qx\|_1$ poses computational or implementation difficulties. A standard trick is to introduce an auxiliary variable $y = Qx \in \mathbf{R}^{M_2}$. As before, the proposed method is fully explicit when applied to this formulation.

4.2 Benchmark: Condat's algorithm

Condat's algorithm [31] (further extended by Vu [60]) is among the few methods combining the following two features, which are the core of the motivation for our research: (1) it exploits the fact that the components involved in the objective function may be of a different nature in terms of differentiability; and (2) uses the structured character of the problem, reducing the size and complexity of the sub-problem solved at each iteration. Thus, it is a natural benchmark for our proposed method. Problem ($\mathcal{P}$) corresponds to taking $F = f$, $G = g$, $L = [A, B]$ and $H = \iota_{\{c\}}$, according to the notation in [31].

4.3 Implementation and comparison

We implement our algorithm in an instance of the ROF model, which is a simple but relevant framework. It is important to mention that, for this specific model, more sophisticated and better-performing algorithms have been designed (see [11–13,18, 21,23,24,34,39,43,58,59,61,62]). We do not expect our algorithm to outperform all of those methods. Instead, we shall carry out a comparison with Condat's algorithm, which can handle problems with the same (and even higher) level of generality.

In the experiments, we considered the 256×256-pixel *cameraman* test image (see Fig. 1a). As image corruption K, we used a Gaussian blur of size 9×9 and standard deviation 4 (applied by MatLab function `fspecial`), followed by an additive zero-mean white Gaussian noise with standard deviation 10^{-3}. The corrupted image is shown in Fig. 1b. We set the regulation parameter at $r = 10^{-4}$. The starting point is the corrupted image x_0, and we set $y_0 = Dx_0$, along with $\mu_0 = 0$ for the dual variable in Condat's algorithm. As comparison criterion, we used the energy of the original (unconstrained) problem (24) (with $Q = D$), namely

$$E(x_k) := \frac{1}{2}\|Kx_k - h\|_2^2 + r\|Dx_k\|_1. \tag{25}$$

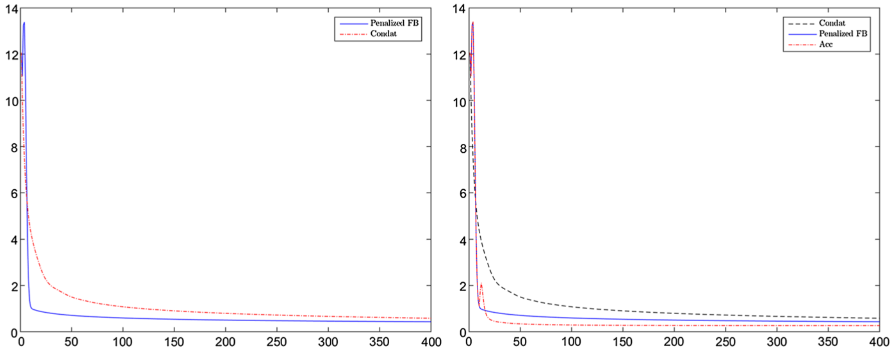
We first investigated different choices for the parameters Condat's algorithm without relaxation ($\rho_n = 1$). The best-performing parameters we found, in terms of the energy given by (25) after 1000 iterations, are $\sigma = 0.25$ and $\tau = \left(\frac{\|K\|^2}{2} + \sigma\|L\|^2\right)$, where $L = [D, -I]$. For our algorithm (1)–(3), we chose the step-sizes $\lambda_n = 5n^{-0.51}$ and penalization parameter $\gamma = 0.1$.

The final results for the image reconstruction are shown in Fig. 1c, d. For the original image the energy in (25) assumes the value 0.475, while for the corrupted one it is equal to 12.067. The final values energy after 400 iterations are respectively 0.435 for our algorithm and 0.586 for Condat's (after 1000 iterations, we obtain 0.370 and 0.400, respectively). In Fig. 2a, we show a comparison between the behaviour of the two algorithms, in terms of (25). Except for a few erratic iterations at the beginning, the values corresponding to our algorithm are always below those of Condat's scheme.

Remark 12 The proximal-gradient algorithm can be accelerated by means of Nesterov's scheme [48] (see also [11]). Inspired by the latter, we introduce a heuristically

(a) Original image **(b)** Corrupted image **(c)** Condat **(d)** Penalized FB

Fig. 1 Debluring of the cameraman image



(a) Condat's algorithm vs (1)-(3) **(b)** Acceleration heuristic

Fig. 2 Energy comparison after 400 iterations

accelerated version of our algorithm, in which the iterations (1) and (2) are not applied to (x_n, y_n), but to $(\tilde{x}_n, \tilde{y}_n)$, defined as an extrapolation of the segment joining (x_{n-1}, y_{n-1}) to (x_n, y_n), namely

$$(\tilde{x}_n, \tilde{y}_n) = (x_n, y_n) + \left(1 - \frac{\alpha}{\lfloor n + \alpha \rfloor}\right)(x_n - x_{n-1}, y_n - y_{n-1}),$$

with $\alpha > 3$ (see [7,20]). We have performed a preliminary computational experiment, in the same setting as before, and with $\alpha = 3.1$. The final energy after 400 iterations is 0.2671 (see Fig. 2b). A deeper numerical analysis and the theoretical proof of convergence and acceleration properties are left for future development, as well as the possibility to use this tool together with a restarting technique [57] or early stopping rules (see [37]).

Acknowledgements Supported by Fondecyt Grant 1181179 and Basal Project CMM Universidad de Chile. The first author was also supported by CONICYT-PCHA/Doctorado Nacional/2016 scholarship.

References

1. Acker, F., Prestel, M.A.: Convergence d'un schéma de minimisation alternée. Ann. Faculté des Sci. Toulouse V Sér. Math. **2**, 1–9 (1980)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Alternating proximal algorithms for weakly coupled convex minimization problems. Applications to dynamical games and PDE's. J. Convex Anal. **15**(3), 485–506 (2008)
3. Attouch, H., Brézis, H.: Duality for the sum of convex functions in general Banach spaces. Asp. Math. Appl. J. Barroso Ed. **34**, 125–133 (1986)
4. Attouch, H., Cabot, A., Frankel, P., Peypouquet, J.: Alternating proximal algorithms for constrained variational inequalities. Application to domain decomposition for PDE's. Nonlinear Anal. **74**(18), 7455–7473 (2011)
5. Attouch, H., Czarnecki, M.-O., Peypouquet, J.: Prox-penalization and splitting methods for constrained variational problems. SIAM J. Optim. **21**, 149–173 (2011)
6. Attouch, H., Czarnecki, M.-O., Peypouquet, J.: Coupling forward–backward with penalty schemes and parallel splitting for constrained variational inequalities. SIAM J. Optim. **21**(4), 1251–1274 (2011)
7. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward–backward method is actually faster than 1/k2. SIAM J. Optim. **26**(3), 1824–1834 (2016)
8. Attouch, H., Redont, P., Soubeyran, A.: A new class of alternating proximal minimization algorithms with costs-to-move. SIAM J. Optim. **18**, 1061–1081 (2007)
9. Auslender, A., Crouzeix, J.P., Fedit, P.: Penalty-proximal methods in convex programming. J. Optim. Theory Appl. **55**, 1–21 (1987)
10. Bauschke, H.H., Combettes, P.L., Reich, S.: The asymptotic behavior of the composition of two resolvents. Nonlinear Anal. **60**, 283–301 (2005)
11. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. **18**(11), 2419–2434 (2009)
12. Bildhauer, M., Fuchs, M.: A variational approach to the denoising of images based on different variants of the TV-regularization. Appl. Math. Optim. **66**, 331–361 (2012)
13. Bertels, S.: Total variation minimization with finite elements: convergence and iterative solution. SIAM J. Numer. Anal. **50**(3), 1162–1180 (2012)
14. Briceño-Arias, L.: Forward-Douglas–Rachford splitting and forward-partial inverse method for solving monotone inclusions. Optimization **64**(5), 1239–1261 (2015)
15. Briceño-Arias, L., López, S.: A projected primal-dual splitting for solving constrained monotone inclusions. arXiv:1805.11687v1 [math.OC]
16. Cabot, A.: Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization. SIAM J. Optim. **15**(2), 555–572 (2005)
17. Candès, E.J., Eldar, Y.C., Needell, D., Randall, P.: Compressed sensing with coherent and redundant dictionaries. Appl. Comput. Harmon. Anal. **31**(1), 59–73 (2011)
18. Carter, J.L.: Dual methods for total variation-based image restoration. Ph.D. Thesis, UCLA (2001)
19. Cauchy, A.L.: Méthode générale pour la résolution des systèmes d'équations simultanées. C. R. Acad. Sci. Paris **25**, 536–538 (1847)
20. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". J. Optim. Theory Appl. **166**(3), 968–982 (2015)
21. Chambolle, A., Lions, P.L.: Image recovery via total variation minimization and related problems. Numer. Math. **76**, 167–188 (1997)
22. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
23. Chan, T., Esedoglu, S., Park, F., Yip, A.: Total variation image restoration: overview and recent developments. In: Paragios, N., Chen, Y., Faugeras, O.D. (eds.) Handbook of mathematical models in computer vision, pp. 17–31. Springer, New York (2006)
24. Chan, T.F., Golub, G.H., Mulet, P.: A nonlinear primal-dual method for total variation-based image restoration. SIAM J. Sci. Comput. **20**(6), 1964–1977 (1999)
25. Chen, G., Teboulle, M.: A proximal-based decomposition method for convex minimization problems. Math. Program. **64**(1–3), 81–101 (1994)
26. Combettes, P.L.: Iterative construction of the resolvent of a sum of maximal monotone operators. J. Convex Anal. **16**(3–4), 727–748 (2009)

27. Combettes, P.L., Pesquet, J.C.: A proximal decomposition method for solving convex variational inverse problems. Inverse Probl. **24**(26) (2008)
28. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. Fixed-point algorithms for inverse problems in science and engineering. In: Bauschke, H.H., Burachik, R.S.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Springer Optimization and Its Applications, vol. 49, pp. 185–212. Springer, New York (2011)
29. Combettes, P.L., Pesquet, J.C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set Valued Var. Anal. **20**(2), 307–330 (2012)
30. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward–backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
31. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
32. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
33. De Oliveira, W., Sagastizábal, C., Lemaréchal, C.: Convex proximal bundle methods in depth: a unified analysis for inexact oracles. Math. Program. Ser. B **148**(1–2), 241–277 (2014)
34. Dobson, D.C., Vogel, C.R.: Convergence of an iterative method for total variation denoising. SIAM J. Numer. Anal. **34**(5), 1779–1791 (1997)
35. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. **82**(2), 421–439 (1956)
36. Eckstein, J., Silva, P.J.S.: Proximal methods for nonlinear programming: double regularization and inexact subproblems. Comput. Optim. Appl. **46**(2), 279–304 (2010)
37. Garrigos, G., Rosasco, L., Villa, S.: Iterative regularization via dual diagonal descent. J. Math. Imaging Vis. **60**, 189–215 (2017). https://doi.org/10.1007/s10851-017-0754-0
38. Goldstein, A.A.: Convex programming in Hilbert space. Bull. Am. Math. Soc. **70**(5), 709–710 (1964)
39. Goldstein, T., Osher, S.: The split Bregman method for L1-regularized problems. SIAM J. Imaging Sci. **2**(2), 323–343 (2009)
40. González-Andrade, S.: A preconditioned descent algorithm for variational inequalities of the second kind involving the p-Laplacian operator. Comput. Optim. Appl. **66**(1), 123–162 (2017)
41. Hare, W., Planiden, C.: Computing proximal points of convex functions with inexact subgradients. Set Valued Var. Anal. **26**(3), 469–492 (2018)
42. Hare, W., Sagastizábal, C., Solodov, M.: A proximal bundle method for nonsmooth nonconvex functions with inexact information. Comput. Optim. Appl. **63**(1), 1–28 (2016)
43. Lai, M.J., Messi, L.M.: Piecewise linear approximation of the continuous Rudin–Osher–Fatemi model for image denoising. SIAM J. Numer. Anal. **50**(5), 2446–2466 (2012)
44. Levitin, E.S., Polyak, B.T.: Constrained minimization problems. USSR Comput. Math. Math. Phys. **6**, 1–50 (1966) (English translation of original paper in Zh. Vychisl. Mat. Fiz. 6, 787–823, 1965)
45. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat. Recherche Opérationnelle **4**, 154–158 (1970)
46. Molinari, C., Peypouquet, J.: Lagrangian penalization algorithm with parallel forward–backward splitting. J. Optim. Theory Appl. **177**(2), 413–447 (2018)
47. Moreau, J.J.: Fonctionnelles Convexes. Cours Collège de France, Paris (1967)
48. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Dokl. **27**, 372–376 (1983)
49. Opial, Z.: Weak convergence of the sequence of successive approximations for nonexpansive mappings. Bull. Am. Math. Soc. **73**, 591–597 (1967)
50. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**(3), 123–231 (2014)
51. Peypouquet, J.: Convex Optimization in Normed Spaces. Theory, Methods and Examples. With a Foreword by Hedy Attouch. Springer, Berlin (2015)
52. Reem, D., Reich, S.: Solutions to inexact resolvent inclusion problems with applications to nonlinear analysis and optimization. Rend. Circ. Mat. Palermo (2) **67**(2), 337–371 (2018)
53. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
54. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)
55. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**, 259–268 (1992)

56. Shen, J., Xia, Z.Q., Pang, L.P.: A proximal bundle method with inexact data for convex nondifferentiable minimization. Nonlinear Anal. **66**(9), 2016–2027 (2007)
57. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. NIPS **27**, 2510–2518 (2014)
58. Tai, X.C., Wu, C.: Augmented Lagrangian method, dual methods and split Bregman iteration for ROF model. In: Tai, X.C., Moerken, K., Lysaker, M., Lie, K.A. (eds.) Scale space and variational methods in computer vision, pp. 502–513 (2009)
59. Vogel, C.R., Oman, M.E.: Iterative methods for total variation denoising. SIAM J. Sci. Comput. **17**(1), 227–238 (1996)
60. Vu, B.C.: A splitting algorithm for coupled system of primal-dual monotone inclusions. J. Optim. Theory Appl. **164**(3), 993–1025 (2015)
61. Wu, C., Zhang, J., Tai, X.C.: Augmented Lagrangian method for total variation restoration with non-quadratic fidelity. Inverse Probl. Imaging **5**(1), 237–261 (2011)
62. Zhu, M., Wright, S.J., Chan, T.F.: Duality-based algorithms for total-variation-regularized image restoration. Comput. Optim. Appl. **47**, 377–400 (2010)