A survey on operator splitting and decomposition of convex programs

Arnaud Lenoir* Philippe Mahey†

July 30, 2015

Abstract

Many structured convex minimization problems can be modeled by the search of a zero of the sum of two monotone operators. Operator splitting methods have been designed to decompose and regularize at the same time these kind of models. We review here these models and the classical splitting methods. We focus on the numerical sensitivity of these algorithms with respect to the scaling parameters that drive the regularizing terms, in order to accelerate convergence rates for different classes of models.

1 Introduction

We survey here some classical monotone operator splitting methods and discuss technical issues which address the difficult question of the acceleration of the numerical performance of these techniques when dealing with the decomposition of large-scale convex programs.

Monotone operator theory has been developed for Hilbert spaces in the seventies by different people, where we first distinguish the monograph by Brezis [10] and the seminal research results of J.J. Moreau [56] on the proximal mapping. The Proximal Point Method (PPM) became popular in the Mathematical Programming community with the seminal papers of Rockafellar [63, 62] who showed precisely its link with the Augmented Lagrangian algorithm. Operator splitting is generally referred when dealing with the sum of maximal monotone operators and aiming at decomposing the numerical computations on each operator separately. Early splitting algorithms were analyzed by Lieutaud in his thesis [48] using the term *Fractional Steps* method, borrowed from early works by russian researchers (Denidov, Marchuk, Samarskii and Yanenko, see Temam [71] for a convergence analysis). The first application of the celebrated Alternate Direction Method

*EDF R& D, Clamart, France

†Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, (L.I.M.O.S), Clermont Université, France. e-mail:Philippe.Mahey@isima.fr

of Multipliers (ADMM) stems back to the work of Glowinski and Marocco [33] who solved some heat conduction equations, see too Gabay and Mercier [31] and the theoretical analysis by Gabay [30]. In 1979, Lions and Mercier [49] analyzed the convergence of a family of splitting methods (combining forward and backward steps like in the earlier methods of Douglas-Rachford and Peaceman-Rachford for linear operators), for solving a general monotone inclusion problem involving the sum of two maximal monotone operators. In parallel, Cohen [15] introduced the Auxiliary Problem principle to define a very general family of decomposition algorithms, generalizing the role of the Proximity step in Augmented Lagrangian functions. The necessity of reformulation to decompose separable convex programs was the motivation of Spingarn's Partial Inverse method [68]. The Partial Inverse method was then shown to be closely related to ADMM and to the Douglas-Rachford method by J. Eckstein in his thesis and a collection of related papers (see [24, 27]). In parallel, Chen and Teboulle [14] derived a decomposition algorithm adapted from Rockafellar's Proximal Method of Multipliers (PMM, see [62]) and Mahey et al [51] analyzed the convergence rate of the Proximal Decomposition Algorithm (PDA), generalizing Spingarn's block-decomposition algorithm [69].

In the recent years, the research on the subject has been mainly focused on the following issues :

- Revisiting forward-backward schemes and extending to some classes of non convex models ([67, 2, 19, 66]);
- Extending splitting methods to composite operators ([11, 8, 1]);
- Exploring new strategies adapted to new models coming from Signal Processing and Classification theory ([18, 12, 35, 9]);
- Studying worst-case complexity bounds ([34, 40, 66]);
- Introducing scaling parameters to accelerate convergence with self-adaptive update ([23, 39]).

We will focus on the last point in our presentation as it is a critical issue to derive efficient algorithms to decompose large optimization problems. For instance, it is well-known that the rate of convergence of the Douglas-Rachford method is highly sensitive to the choice of the parameter introduced in the proximal step. More worrying is the fact that the control of the rate of the primal and dual sequences are conflicting, thus limiting the best expected linear rate of convergence to the value 0.5 and provoking nasty spiralling effects (see [24, 37]).

On the other hand, worst-case complexity bounds have been studied in the spirit of Nesterov's smoothing techniques [58] to better understand the limits of performance of splitting methods for decomposing the sum of

general convex functions. Some results may look rather frustrating as commented in a recent study [20] : the global convergence of Douglas-Rachford splitting scheme can be as fast as the Proximal iteration in the ergodic sense and as slow as a subgradient method in the non ergodic sense.

The present study does not pretend to cover all research around operator splitting methods and applications. Since the original work of Gabay thirty years ago [30], a few interesting tutorials and surveys have been proposed in the literature, see [26, 4, 66].

We first present the main splitting techniques to find a zero of the sum of two maximal monotone operators. We apply the most promising techniques to the decomposition of separable convex programs in section 3, focusing on constrained optimization problems in finite dimension. In section 4, we explore theoretical and practical issues on the convergence of operator splitting methods.

2 Splitting the sum of two monotone operators

2.1 Forward and backward steps

Let recall first a few useful definitions about monotone operators. We will use the notation $\langle \cdot, \cdot \rangle$ to denote the dot product in a Hilbert space X.

Definition 1 *$T : X \mapsto X$ is monotone if*

$$\langle T(x) - T(x'), x - x' \rangle \geq 0, \forall x, x' \in X$$

It is maximal if its graph is not strictly contained in the graph of any monotone operator. It is strongly monotone with constant $a > 0$ if

$$\langle T(x) - T(x'), x - x' \rangle \geq a\|x - x'\|^2, \forall x, x'$$

and co-coercive with module $a > 0$ if

$$\langle T(x) - T(x'), x - x' \rangle \geq a\|T(x) - T(x')\|^2, \forall x, x'$$

For example, the subdifferential of a closed convex function on a convex subset of $\mathbb{R}^n$ is a maximal monotone operator. A symmetric linear operator which is positive definite (thus the subdifferential of a strongly convex function) is strongly monotone (the constant will be the smallest eigenvalue). Co-coercivity is the dual property, so that an operator is co-coercive if its inverse is strongly monotone.

Let T be a maximal monotone operator on X. Let consider first the following inclusion :

$$\text{Find} \qquad x^* \qquad \text{such that} \qquad 0 \in T(x^*) \tag{1}$$

When T is the subdifferential operator of a closed convex function, it corresponds naturally to the global minimization of that function on X.
The *forward* step is the following iteration :

$$x^{t+1} \in (I\!I - \lambda_t T)(x^t)$$

which is not completely defined unless $T(x^t)$ is a singleton. The parameter $\lambda_t > 0$ is the stepsize and it is generally expected to decrease at each iteration like in the so-called *subgradient* algorithm for minimizing nonsmooth convex functions.
The *backward* step is the following iteration :

$$x^{t+1} = (I\!I + \lambda T)^{-1}(x^t)$$

where $J_\lambda^T = (I\!I + \lambda T)^{-1}$ is the *resolvent* operator of T which is indeed defined on the whole space so that the backward iteration is now uniquely defined for any $\lambda > 0$. It is too a firmly non expansive operator (or co-coercive with modulus 1, see Minty [55] and the definitions of averaged operators below) which means that :

$$\forall x, x' \in X, \langle x - x', J_\lambda^T(x) - J_\lambda^T(x') \rangle \geq \|J_\lambda^T(x) - J_\lambda^T(x')\|^2$$

This suggests that the backward step behaves like a fixed-point iteration to solve (1) which can be derived directly by :

$$0 \in T(x) \Longleftrightarrow 0 \in \lambda T(x) \Longleftrightarrow x \in (I\!I + \lambda T)(x) \Longleftrightarrow x = (I\!I + \lambda T)^{-1}(x)$$

To say more about the parameter $\lambda > 0$, the fixed-point equation above says that we have substituted the operator T by its Moreau-Yosida approximation $T_\lambda = \frac{1}{\lambda}(I\!I - J_\lambda^T)$. It is shown in Brézis [10] that T_λ is maximal monotone and Lipschitz with constant $1/\lambda$. When T is the subdifferential of a closed convex function f, the iteration becomes :

$$x^{t+1} = \operatorname{arginf} f(x) + \frac{1}{2\lambda}\|x - x^t\|^2 \tag{2}$$

This implicit step can be viewed as an explicit gradient step to minimize the regularized Moreau-Yoshida function $f_\lambda(x) = \inf\{f(z) + \frac{1}{2\lambda}\|z - x\|^2\}$ which is indeed smooth with gradient $T_\lambda(x)$. Thus, (2) is equivalent to $x^{t+1} = x^t - \lambda \nabla f_\lambda(x^t)$.

Implementing iteration (2) gives the celebrated *Proximal Point Method* (PPM) the convergence of which was first analyzed by Martinet [53]. Rockafellar in his analysis of (PPM) concluded that the rate of convergence improves when λ increases [63].

The former subproblem is generally implementable when f corresponds to the dual function associated with a constrained optimization problem,

yielding in the primal space the famous Augmented Lagrangian algorithm (the first detailed analysis of this relation is due to Rockafellar [62]).

To complete our introduction on the basic iterative schemes, we review the notion of *averaged operators*, introduced by Baillon et al [3] (see too [4]). Let S be a non expansive operator on X (such that $\|Sx - Sx'\| \leq \|x - x'\|, \forall x, x'$). A ρ-averaged operator is, for any $\rho \in (0,1)$, defined by :

$$S_\rho = \rho S + (1-\rho)I\!I$$

Averaged operators are too non expansive, but they are more amenable to fixed-point iterations than general non expansive ones as they share the following property :

$$\|S_\rho x - S_\rho x'\|^2 \leq \|x - x'\|^2 - \frac{1-\rho}{\rho}\|(I\!I - S_\rho)(x) - (I\!I - S_\rho)(x')\|^2$$

S_ρ shares the same set of fixed-points with S and the convergence of the general fixed-point iteration $x^{t+1} = S_{\rho_t}(x^t)$ (referred as the Krasnosel'ski-Mann algorithm) converges to a solution x^* in that set (see [17]). Moreover, it was shown in [20] that the sequence $\|x^t - x^*\|$ is nonincreasing and that $\|Sx^t - x^t\|^2 = o(1/t)$, assuming only that the sequence $\tau_t = \rho_t(1-\rho_t)$ is bounded away from 0.

Conveniently, compositions of averaged operators are easily seen to be averaged. Indeed, if S_a and S_b are averaged operators with constant a and b respectively, then $S_\rho = S_a \circ S_b$ is averaged too with constant $1-(1-a)(1-b) < 1$.

It is immediate to observe that the resolvent operator J_λ^T of a maximal monotone operator T is 1/2-averaged which is incidentally equivalent to be firmly non expansive. In consequence, the backward iteration in the Proximal Point method converges to a fixed-point of J_λ^T, i.e. a zero of T. These convergence properties of the averaged operator iterations will apply to most splitting schemes studied in the next section.

Finally, we introduce the *reflector operator* associated with a monotone operator T :

Definition 2 *Let $T : I\!R^n \mapsto I\!R^n$ maximal monotone, $P = (I\!I + T)^{-1}$ and $Q = I\!I - P$. The operator*

$$N = P - Q = 2P - I\!I = I\!I - 2Q$$

is the generalized reflector associated with T.

That notion truly generalizes a symmetry corresponding to the case when $T = \mathcal{N}_\mathcal{A}$ is the normal cone of a linear subspace $\mathcal{A}$ so that $\text{Graph}(T) = \mathcal{A} \times \mathcal{A}^\perp$. Generalized reflectors correspond exactly to the set of non expansive operators. The correspondence between the graphs of T and N appears clearly in the following construction:

Proposition 1 *Let $T : I\!R^n \mapsto I\!R^n$ maximal monotone and N its generalized reflector. Then $y \in T(x) \Longleftrightarrow d = N(s)$ with :*

$$\left\{ \begin{array}{ll} s & = x + y \\ d & = x - y \end{array} \right. \quad or \quad \left\{ \begin{array}{ll} x & = \frac{1}{2}(d+s) \\ y & = \frac{1}{2}(d-s) \end{array} \right.$$

which leads to the following decomposition:

Proposition 2 *Let $T : I\!R^n \mapsto I\!R^n$ maximal monotone and $u \in I\!R^n$. Then the following assertions are equivalent :*

i) $y \in Tx$

ii) $x = P(s)$ *and* $s = x + y$

iii) $y = Q(s)$ *and* $s = x + y$

In other words, there exists a unique pair $(x, y) \in \text{Graph}(T)$ such that $x+y = s$. That decomposition on the graph of the maximal monotone operator T is called the *Moreau-Minty decomposition* [55]. It will be at the heart of the *Proximal Decomposition method* presented in the next section.

2.2 Main splitting methods

We consider now the basic model of interest to derive decomposition methods, i.e. the case of $T = T_1 + T_2$ where T_1 and T_2 are two maximal monotone operators on X. The basic problem is then :

$$\text{Find } x^* \in X \text{ such that } 0 \in T_1(x^*) + T_2(x^*) \qquad (P)$$

One generally defines an operator splitting method as one which combines forward and backward steps applied separately to operators T_1 and T_2 but never to $T_1 + T_2$.

We will use throughout the following trivial reformulation in the primal-dual space $X \times X$:

x^* is a solution of (1) if and only if there exists $y^* \in X$ such that (x^*, y^*) solves :

$$y^* \in T_1(x^*) \qquad (3)$$

$$-y^* \in T_2(x^*) \qquad (4)$$

Before getting into the description of the main splitting algorithms, let introduce a generic example for (P) that will be extensively used in the remainder :

$$\text{Find } x^* \in \mathcal{A} \quad \text{that minimizes } f(x) \qquad (P0)$$

where f is closed convex on X and $\mathcal{A}$ is a subspace of X. Typically, f is a separable function and $\mathcal{A}$ is the coupling subspace. The optimality conditions for (P0), assuming the existence of an optimal solution x^*, are :

$$(x^*, y^*) \in \mathcal{A} \times \mathcal{A}^{\perp} \bigcap \text{Graph}\,(T) \qquad (5)$$

where $T = \partial f$.

2.2.1 Double Backward splitting

As suggested by its name, that splitting scheme uses two sequential proximal steps on each operator :

$$x^{t+1} = J_{\lambda_t}^{T_2} \circ J_{\lambda_t}^{T_1}(x^t) \tag{6}$$

In general, the zeroes of the composed operator $J_\lambda^{T_2} \circ J_\lambda^{T_1}$ do not correspond to the zeroes of $T_1 + T_2$. A possibility to solve (1) is to force the scaling parameter λ_t to decrease to zero and characterize convergence in the following *ergodic* sense :

Assuming $\sum_t \lambda_t = +\infty, \sum_t \lambda_t^2 < +\infty$, the sequence $\{z_t = \frac{\sum_{\tau=0}^t \lambda_\tau x^\tau}{\sum_{\tau=0}^t \lambda_\tau}\}$ converges to a zero of T (see Passty [59]).

2.2.2 Forward-Backward splitting

Here we use a forward step associated with (3), i.e. :

$$x - \lambda y \in (I - \lambda T_1)(x)$$

composed with a backward step on (4) :

$$x = J_\lambda^{T_2}(x - \lambda y)$$

so that the forward-backward iteration is given by :

$$x^{t+1} = J_\lambda^{T_2} \circ (I - \lambda T_1)(x^t) \tag{7}$$

Observe that the same parameter λ should be used in both forward and backward steps at each iteration.

If $T_1 = \nabla F$ where F is smooth convex and $T_2 = \mathcal{N}_C$, the normal cone of a closed convex set, this is in fact the Projected Gradient method originally proposed by Goldstein [36]. Indeed, the resolvent $J_\lambda^{T_2}$ is the projection operator on C.

Convergence was analyzed first by Passty [59] who proved ergodic convergence of the sequence $\{z^t\}$ like in the former section. To get the convergence of the whole sequence, one needs additional properties like the co-coercivity of T_1 with modulus a and a controlled stepsize λ in the interval $[0, 1/2a]$(see Mercier [54] and related extensions in [13]).

Figure 1 illustrates the convergence on the example of two operators T_1, T_2 with values in $\mathbb{R}$.

2.2.3 Peaceman- and Douglas-Rachford splitting methods

Peaceman-Rachford method was originally applied to linear operators [60] and we consider here its extension to monotone operators as studied by Lions and Mercier in [49].

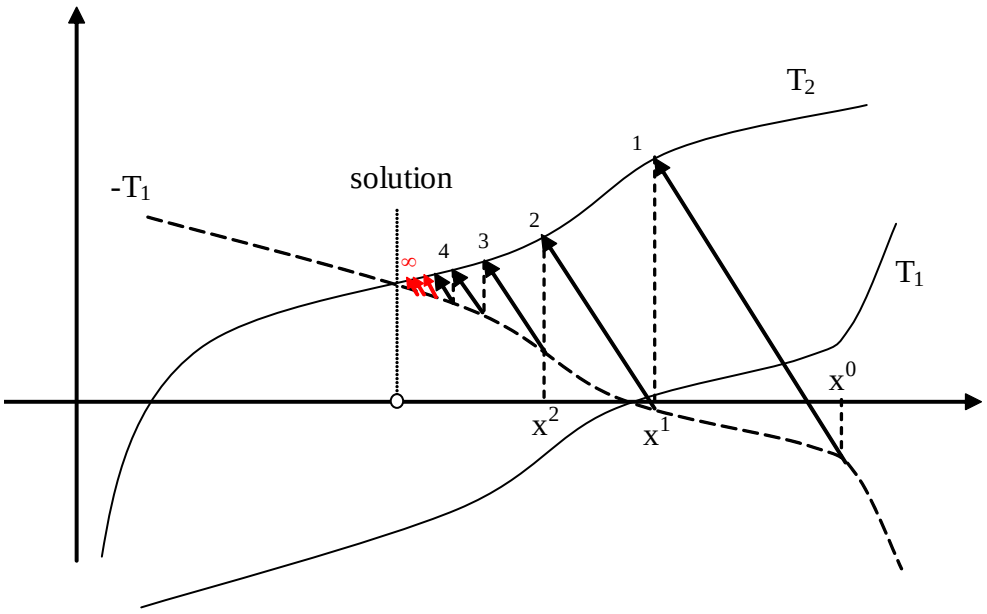

Figure 1: Sequence of iterates for Forward-backward algorithm

Observe first that the optimality conditions (3),(4) can be rewritten in the following way :

$$x^* + y^* = [N_2 \circ N_1](x^* + y^*)$$

where N_1 and N_2 are the reflectors associated with operators T_1 and T_2 respectively. Indeed, using Prop. 1, elementary calculations transform (3) in $x^* - y^* = N_1(x^* + y^*)$ and (4) in $x^* + y^* = N_2(x^* - y^*)$. The operator $N_2 \circ N_1$ is non expansive, inducing the following iteration :

$$s^{t+1} = [N_2 \circ N_1](s^t)$$

so that the primal and dual sequences of iterates are obtained using the construction of Prop. (2). A scaling parameter can be introduced again using the following change of scale of both operators substituting T_1 by $aT_1(b^{-1}\cdot)$ and T_2 by $aT_2(b^{-1}\cdot)$ with $a, b > 0$. It corresponds to a simple change of variables, $x \leftarrow b^{-1}x$ for the primal variables, and $y \leftarrow a^{-1}y$ for the dual variables. The iteration is now detailed using the new primal and dual variables with the single parameter $\lambda = \frac{b}{a}$:

Algorithm 1 Peaceman Rachford algorithm (PRA)

Require: $\lambda > 0, t = 0$, choose (x^0, y^0)

1: $x^{t+\frac{1}{2}} = J_\lambda^{T_1}(x^t + \lambda y^t)$
2: $y^{t+\frac{1}{2}} = \lambda^{-1}(x^t - x^{t+\frac{1}{2}}) + y^t$
3: $x^{t+1} = J_\lambda^{T_2}(x^{t+\frac{1}{2}} - \lambda y^{t+\frac{1}{2}})$
4: $y^{t+1} = \lambda^{-1}(x^{t+1} - x^{t+\frac{1}{2}}) + y^{t+\frac{1}{2}}$
5: $t \leftarrow t + 1$; Go To step 1

Convergence of algorithm (PRA) is not guaranteed in the general case but can be stated with some additional hypotheses (the following proposition is proved in [49]) :

Proposition 3
- *If T_1 (respectively T_2) is strongly monotone, then $x^{t+\frac{1}{2}}$ (resp. x^t) converges to the unique optimal primal solution x^*.*
- *If T_1 (respectively T_2) is co-coercive, then $y^{t+\frac{1}{2}}$ (resp. y^t) converges to the unique optimal dual solution y^*.*

Many authors, following Varga [73], have considered a natural underrelaxation of (PRA) associated with a new parameter $\alpha > 0$:

$$s^{t+1} = [(1-\alpha_t)I\!I + \alpha_t(N_2 \circ N_1)](s^t) \tag{8}$$

Its convergence is now guaranteed for any $0 < \alpha_t < 1$ as shown by Lions and Mercier in [49]. Indeed, the iteration operator is now averaged and, moreover, the case $\alpha = 1/2$ is of interest because it is exactly the Douglas-Rachford method for linear inclusions [22] which is known to be intimately linked with the Alternate Direction method of Multipliers (ADMM) ([33, 30]) and the Partial Inverse method of Spingarn [68]. We will not detail here the fine correspondence between these now classical splitting methods. For a complete overview of this material, Eckstein's PhD thesis [24] is a very accurate and partly unexploited reading.

We give below the classical form of Douglas-Rachford algorithm (DRA) for finding a zero of the sum of two maximal monotone operators.

Algorithm 2 Douglas Rachford algorithm (DRA)

Require: $\lambda > 0, t = 0$, choose (x^0, y^0)

1: $x^{t+\frac{1}{3}} = J_\lambda^{T_1}(x^t + \lambda y^t)$
2: $y^{t+\frac{1}{3}} = \lambda^{-1}(x^t - x^{t+\frac{1}{3}}) + y^t$
3: $x^{t+\frac{2}{3}} = J_\lambda^{T_2}(x^{t+\frac{1}{3}} - \lambda y^{t+\frac{1}{3}})$
4: $y^{t+\frac{2}{3}} = \lambda^{-1}(x^{t+\frac{2}{3}} - x^{t+\frac{1}{3}}) + y^{t+\frac{1}{3}}$
5: $(x^{t+1}, y^{t+1}) = \frac{1}{2}[(x^{t+\frac{2}{3}}, y^{t+\frac{2}{3}}) + (x^t, y^t)]$
6: $t \leftarrow t+1$; Go To step 1

The only difference with (PRA) is the addition of step 5 where we substitute the last primal-dual estimate in (PRA) by the mean of the two successive estimates during the iteration. Observe that the second proximal step on T_2 flips the sign of the dual counterpart y. To be more precise as observed in [24], we have the following proposal :

Proposition 4 *At each step of algorithm (DRA), the sequence $\{z^t\}$ given by $z^t = x^t + \lambda y^t$ satisfies :*

$$J_\lambda^{T_2}(z^t) = x^t \tag{9}$$

$$(I\!I - J_\lambda^{T_2})(z^t) = \lambda y^t \tag{10}$$

$$N_\lambda^{T_2}(z^t) = x^t - \lambda y^t \tag{11}$$

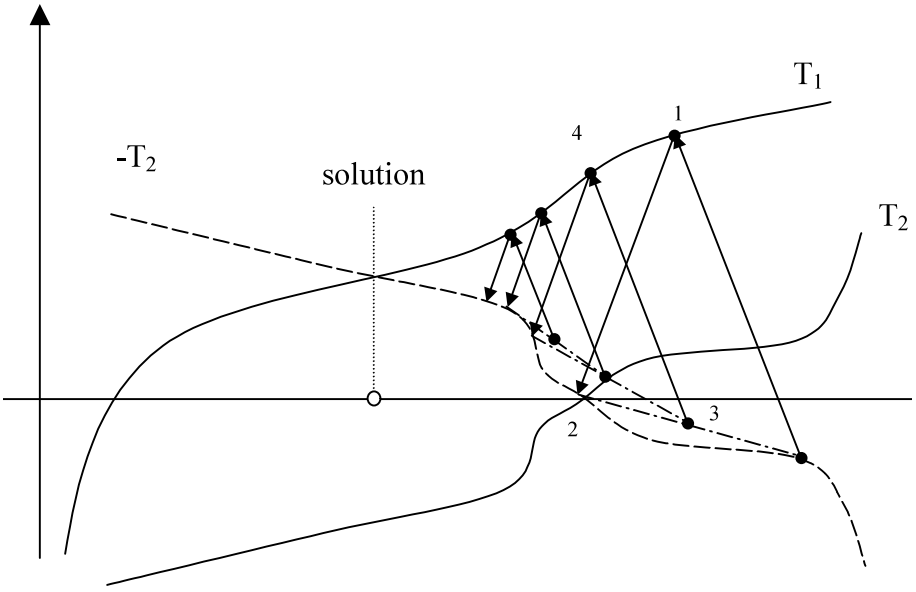

Figure 2: Douglas-Rachford algorithm

so that $(x^t, y^t) \in \text{Graph}(T_2)$. The role of the new variable $z^t = x^t + \lambda y^t$, which combines the primal and dual iterates, is central as the sequence generated by (DRA) satisfies :

$$z^{t+1} = [J_\lambda^{T_1} \circ N_\lambda^{T_2} + I\!I - J_\lambda^{T_2}](z^t)$$

Lenoir [46] has shown that the averaging step 5 of algorithm 2 could be performed between the proximal steps without changing the convergence properties of the method, thus allowing to play with the modeling of the two operators composing the inclusion problem.

The sequence of iterates for (DRA) is illustrated on Figure 2.

2.2.4 Proximal decomposition method

We close now this section with the presentation of the *Proximal Decomposition on the graph of a maximal monotone operator*, motivated by Prop. (2). The main idea is that the two proximal steps used in (DRA) can be performed in parallel, followed by the averaging step which is itself a proximal step performed on an appropriate coupling subspace. Having in mind the generic model ($P0$), this can be done by duplicating the space X and creating two copies x_1 and x_2 of the original variables x to get an equivalent

formulation :

$$\text{Find } (x_1, x_2) \in X \times X \mid 0 \in T_1(x_1) + T_2(x_2) \text{ and } x_1 = x_2$$

Let denote by $\mathcal{A} = \{(x_1, x_2) \in X \times X \mid x_1 = x_2\}$ the coupling subspace. Observe that the dual variables can again be introduced to write the optimality condition in the following way :

$$\begin{aligned} y_1^* &\in T_1(x_1^*) \\ y_2^* &\in T_2(x_2^*) \\ (y_1^*, y_2^*) &\in \mathcal{A}^\perp \end{aligned}$$

which is equivalent to (3)-(4) as $\mathcal{A}^\perp = \{(y_1, y_2) \mid y_1 + y_2 = 0\}$. The optimality conditions above induce the following fixed-point equations $x_1^* = J_\lambda^{T_1}(x_1^* + \lambda y_1^*)$ and $x_2^* = J_\lambda^{T_2}(x_2^* + \lambda y_2^*)$, inducing the alternate proximal steps of the following algorithm proposed in Mahey et al [51] under the name 'Proximal Decomposition method' :

Algorithm 3 Proximal Decomposition algorithm (PDA)

Require: $\lambda > 0, t = 0$, choose (x^0, y^0) and set $x_1^0 = x_2^0 = x^0$ and $y_1^0 = -y_2^0 = y^0$

1: $x_1^{t+\frac{1}{2}} = J_\lambda^{T_1}(x_1^t + \lambda y_1^t)$
2: $y_1^{t+\frac{1}{2}} = \lambda^{-1}(x_1^t - x_1^{t+\frac{1}{2}}) + y_1^t$
3: $x_2^{t+\frac{1}{2}} = J_\lambda^{T_2}(x_2^t + \lambda y_2^t)$
4: $y_2^{t+\frac{1}{2}} = \lambda^{-1}(x_2^t - x_2^{t+\frac{1}{2}}) + y_1^t$
5: $x_1^{t+1} = x_2^{t+1} = \frac{1}{2}(x_1^{t+\frac{1}{2}} + x_2^{t+\frac{1}{2}})$
6: $y_1^{t+1} = -y_2^{t+1} = \frac{1}{2}(y_1^{t+\frac{1}{2}} - y_2^{t+\frac{1}{2}})$
7: $t \leftarrow t + 1$; Go To step 1

To resume, (DRA) alternates proximal steps on T_1 and T_2 in a *Gauss-Seidel* fashion, whereas (PDA) produces the same proximal steps in parallel in a *Jacobi* fashion. But (PDA) can also be interpreted as the application of (DRA) to the following inclusion in $X \times X$:

$$\text{Find } (x_1^*, x_2^*) \in X \times X \text{ such that } 0 \in A(x_1^*, x_2^*) + B(x_1^*, x_2^*)$$

where $A = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}$ and $\text{Graph}(B) = \mathcal{A} \times \mathcal{A}^\perp$.
The sequence of iterates for (PDA) is illustrated on Figure 3.

The name 'Proximal Decomposition' is justified by the application of the scheme above to model (P0). There is no need to introduce the copies of primal variables, simplifying the above steps in :

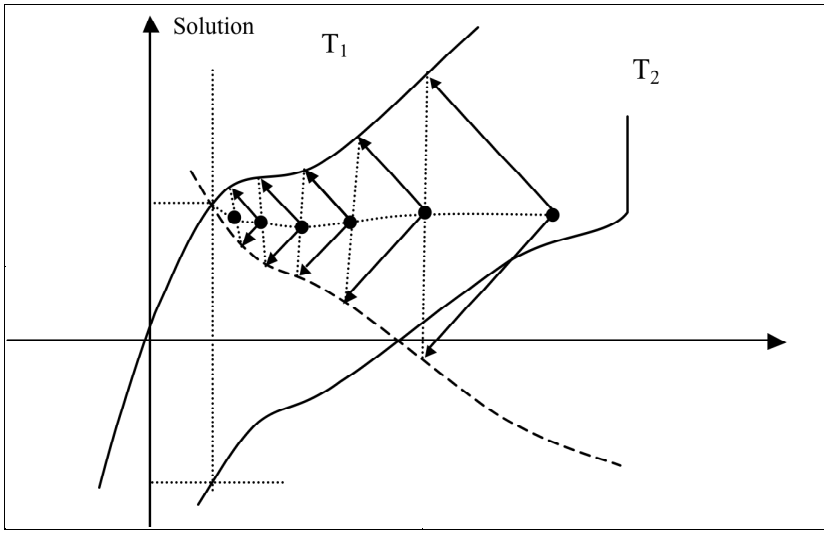

Figure 3: Proximal decomposition algorithm

Step 1 : $(x^{t+\frac{1}{2}}, \lambda y^{t+\frac{1}{2}})$ is the Moreau-Minty Decomposition of $x^t + \lambda y^t$ on the graph of T.
Step 2 : $(x^{t+1}, y^{t+1}) = (x_{\mathcal{A}}^{t+\frac{1}{2}}, y_{\mathcal{A}^\perp}^{t+\frac{1}{2}})$
where $x_{\mathcal{A}}$ denotes the projection of x on the subspace $\mathcal{A}$.

(PDA) shares a strong link with a former method introduced by Spingarn, the Partial Inverse method [68]. Motivated by model (P0), he introduced the Partial Inverse operator associated with the maximal monotone operator T and the subspace $\mathcal{A}$, defined by its graph :

$$\text{Graph}\,(T_{\mathcal{A}}) = \{(x_{\mathcal{A}} + y_{\mathcal{A}^\perp}, x_{\mathcal{A}^\perp} + y_{\mathcal{A}}) \mid y \in T(x)\}$$

$T_{\mathcal{A}}$ is maximal monotone if and only if T is so, and, moreover, we have

$$0 \in T_{\mathcal{A}}(s) \Longleftrightarrow s_{\mathcal{A}^\perp} \in T(s_{\mathcal{A}})$$

In other words, the projections of a zero of $T_{\mathcal{A}}$ on the orthogonal subspaces $\mathcal{A}$ and $\mathcal{A}^\perp$ are the primal and dual solutions of (P0) respectively. Thus, Spingarn proposed to apply (PPM) to operator $T_{\mathcal{A}}$ to solve (P0). But the backward iteration applied to the Partial Inverse operator reduces to :

$$s^{t+1} = (I\!I + T_{\mathcal{A}})^{-1} \Longleftrightarrow s^t - s^{t+1} \in T_{\mathcal{A}}(s^{t+1})$$

Using the definition of the graph of $T_{\mathcal{A}}$, this implies that there exists $(x, y) \in \text{Graph}\,(T)$ such that $s^{t+1} = x_{\mathcal{A}} + y_{\mathcal{A}^\perp}$ and $s^t - s^{t+1} = x_{\mathcal{A}^\perp} + y_{\mathcal{A}}$, or equivalently : $s^t = x + y$ (so that (x, y) is the Proximal Decomposition of s^t on the graph of T) and $s^{t+1} = x_{\mathcal{A}} + y_{\mathcal{A}^\perp}$.

Observe that the relationship between (PDA) and Spingarn's Partial Inverse needs to set the λ parameter to 1. Eckstein has derived too the direct relationship between (DRA) and the Partial Inverse method in [24] using appropriate changes of variables.

The main advantage of (PDA) is that it can easily be extended to inclusions with more than two operators. Again, if the inclusion involves p maximal monotone operators $T_1, \ldots, T_p$ on X :

$$\text{Find } x^* \in X \text{ such that } 0 \in T_1(x^*) + \cdots + T_p(x^*)$$

we can create p copies of space X and apply (PDA) to the cross-product $\mathcal{T} = T_1 \times \ldots \times T_p$ which is indeed a maximal monotone operator on X^p over the coupling subspace $\mathcal{A} = \{(x_1, \ldots, x_p) \in X^p \mid x_1 = \cdots = x_p\}$:

$$\text{Find } \xi^* = (x_1^*, \ldots, x_p^*), \zeta^* = (y_1^*, \ldots, y_p^*) \text{ such that } (\xi^*, \zeta^*) \in \text{Graph}(\mathcal{T}) \bigcap \mathcal{A} \times \mathcal{A}^\perp$$

3 Application to the decomposition of convex programs

Decomposition methods are designed to answer two different objectives :

- To reduce the dimension of large-scale optimization problems with different interconnected subsystems; the challenge here is to identify the coupling variables and/or constraints.
- To exploit hidden 'easy' submodels or, equivalently, to isolate the 'hard' features of the model without which the model is solvable by ad hoc software.

Reformulations of the model are frequently necessary to identify these situations. A typical example for the first case is block-angular linear programs which gave rise to the first decomposition algorithms for Operations Research like Dantzig-Wolfe's and Benders' decomposition methods (see Lasdon's textbook [44] for example). In this case, the separable coupling constraints (or variables) contain the hard features of the model as they prevent to try solving the block subproblems separately.

These strategies lead frequently to implicit value functions (like the dual function with Lagrangian Relaxation) which are typically nonsmooth. As was mentioned before, the inherent non smoothness of the dual function is the main motivation of the Proximal Point method, leading to Augmented Lagrangian subproblems. The problem is that separability which allows decomposing the Lagrangian subproblems is destroyed by the quadratic terms introduced in the Augmented Lagrangian function. Operator splitting methods will be able to address that issue.

3.1 Separable Augmented Lagrangian

We consider first a general convex minimization problem in $I\!R^n$:

$$\left\{ \begin{array}{ll} \text{Minimize} & \sum_{i=1}^{p} f_i(x) \\ & x \in S \end{array} \right. \tag{P1}$$

where the f_i are extended real valued convex functions supposed to be proper and lower-semi-continuous on the closed convex set $S \subset I\!R^n$. Additional local constraints may be present in the model, here modeled inside the functions f_i. A convenient special case for illustrating the next methods is the problem of finding $x \in \bigcap_{i=1,\ldots,p} C_i$ where the C_i are closed convex sets. A simple way to decouple the p pieces of the objective function is to introduce p copies of the variable x denoted $\xi_i, i = 1, \ldots, p$ and reformulate problem $(P1)$ in the product space $(I\!R^n)^p$:

$$\left\{ \begin{array}{ll} \text{Minimize} & \sum_{i=1}^{p} f_i(\xi_i) \\ & \xi_i \in S, i = 1, \ldots, p \\ & \xi = (\xi_1, \ldots, \xi_p) \in \mathcal{A} \end{array} \right. \tag{12}$$

where $\mathcal{A} = \{\xi \in (I\!R^n)^p | \quad \xi_1 = \cdots = \xi_p\}$ is the coupling subspace.

Thus, $(P1)$ has been reformulated into the generic model $(P0)$. That reformulation has been early used by Pierra [61] who applied the double-backward splitting to (12) (recall that this requires the parameter to decrease to zero).

The application of algorithm (PDA) to (12) is straightforward, working in $I\!R^{np}$ with primal $\xi = (\xi_1, \ldots, \xi_p)$ and dual variables $\zeta = (\zeta_1, \ldots, \zeta_p)$. The monotone operator in the product space will be the cartesian product of the subdifferential operators $\partial f_1 \times \cdots \times \partial f_p$:

Algorithm 4 PDA-separable

Require: $t = 0, \lambda > 0, \epsilon > 0, \xi^0 \in \mathcal{A}, \zeta^0 \in \mathcal{A}^\perp$

repeat
  for all $i = 1, \ldots, p$ **do**
    $\xi_i^{t+\frac{1}{2}} := \arg\min_{\xi_i \in S} f_i(\xi_i) + \frac{1}{2\lambda}\|\xi_i - \xi_i^t - \lambda\zeta_i^t\|^2$
    $\zeta_i^{t+\frac{1}{2}} := \lambda^{-1}(\xi_i^t - \xi_i^{t+\frac{1}{2}}) + \zeta_i^t$
    $\xi_i^{t+1} \leftarrow \frac{1}{p}\sum_i \xi_i^{t+\frac{1}{2}}$
    $\zeta_i^{t+1} \leftarrow \zeta_i^{t+\frac{1}{2}} - \frac{1}{p}\sum_i \zeta_i^{t+\frac{1}{2}}$
  end for
  $t \leftarrow t + 1$
until $\|\xi^{t+1} - \xi^t\| + \|\zeta^{t+1} - \zeta^t\| < \epsilon$

The algorithm (PDA-separable) can be applied to many structured models, but a typical situation which we will describe now is the case of a separable objective with separable coupling constraints.

We are interested in solving the following convex program, called hereafter the *S-Model*, defined on the product space $\mathbb{R}^n = \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p}$.

$$\left\{ \begin{array}{ll} \text{Minimize} & \sum_{i=1}^{p} f_i(x_i) \\ & \sum_{i=1}^{p} g_i(x_i) = 0 \\ & x_i \in S_i, i = 1, \ldots, p \end{array} \right. \qquad (S - Model)$$

where f_i are extended real valued convex functions on closed convex sets S_i, supposed to be proper and lower-semi-continuous (l.s.c) and g_i are affine:

$$\begin{aligned} g_i : \mathbb{R}^{n_i} &\to \mathbb{R}^m \\ x_i &\mapsto g_i(x_i) = G_i x_i - b_i \end{aligned}$$

where G_i are $(n_i \times m)$ not necessarily full-rank matrices and $b_i \in \mathbb{R}^m$.

We will now apply (PDA-separable) to the Lagrangian dual of the S-Model, i.e.

$$\text{Maximize}_{u \in \mathbb{R}^m} \quad \sum_{i=1}^{p} h_i(u)$$

where $h_i(u) = \inf_{x_i \in S_i} f_i(x_i) + \langle u, g_i(x_i) \rangle$. The concave dual problem is thus in the form of problem (P1), so that it can be reformulated by creating p copies of the dual variables $u = \xi_1 = \cdots = \xi_p$ to get an equivalent model in $\mathbb{R}^{mp}$:

$$\text{Maximize}_{\xi=(\xi_1,\ldots,\xi_p) \in \mathcal{A}} \quad \sum_{i=1}^{p} h_i(\xi_i) \qquad (13)$$

where $\mathcal{A} = \{(\xi_1, \ldots, \xi_p) \in \mathbb{R}^{mp} \mid \xi_1 = \cdots = \xi_p\}$ is the coupling subspace. Let $\zeta = (\zeta_1, \ldots, \zeta_p)$ be the corresponding variables in duality relation with ξ. The application of (PDA) will give the following algorithm :

Algorithm 5 PDA-dualseparable

Require: $t = 0, \lambda > 0, \epsilon > 0, \xi^0 \in \mathcal{A}, \zeta^0 \in \mathcal{A}^\perp$

1: **repeat**
2: **for all** $i = 1, \ldots, p$ **do**
3: $\xi_i^{t+\frac{1}{2}} := \arg\max h_i(\xi_i) - \frac{1}{2\lambda}\|\xi_i - \xi_i^t - \lambda\zeta_i^t\|^2$
4: $\zeta_i^{t+\frac{1}{2}} := \lambda^{-1}(\xi_i^t - \xi_i^{t+\frac{1}{2}}) + \zeta_i^t$
5: $\xi_i^{t+1} \leftarrow \frac{1}{p}\sum_i \xi_i^{t+\frac{1}{2}}$
6: $\zeta_i^{t+1} \leftarrow \zeta_i^{t+\frac{1}{2}} - \frac{1}{p}\sum_i \zeta_i^{t+\frac{1}{2}}$
7: **end for**
8: $k \leftarrow k + 1$
9: **until** $\|\xi^{t+1} - \xi^t\| + \|\zeta^{t+1} - \zeta^t\| < \epsilon$

Observe that the quadratic term added in the proximal step 3 of algorithm 5 is here subtracted as the dual problem is a maximization problem.

The algorithm can be developed in the primal setting yielding a *separable Augmented Lagrangian* algorithm early proposed by Spingarn [69]. Indeed, the optimality conditions of the proximal step are :

$$\xi_i^{t+\frac{1}{2}} = \xi_i^t + \lambda\zeta_i^t + \lambda g_i(x_i^{t+\frac{1}{2}})$$

where $g_i(x_i^{t+\frac{1}{2}}) \in \partial h_i(\xi_i^{t+\frac{1}{2}})$ and $x_i^{t+\frac{1}{2}}$ minimizes the Augmented Lagrangian obtained by substituting ξ_i by $\xi_i^{t+\frac{1}{2}}$ in the ordinary Lagrangian function $L_i(x_i, \xi_i) = f_i(x_i) + \langle \xi_i, g_i(x_i)\rangle$, which gives the following subproblem :

$$\text{Minimize}_{x_i \in S_i}\ f_i(x_i) + \langle \xi_i^t, g_i(x_i)\rangle + \frac{\lambda}{2}\|g_i(x_i) + \zeta_i^t\|^2$$

The presentation of the algorithm can be simplified, avoiding the use of auxiliary variables ξ as $\xi^t = (u^t, \ldots, u^t)$ after the projection on $\mathcal{A}$ and integrating the intermediate step $t + \frac{1}{2}$ by observing that

$$\begin{aligned} u_i^{t+1} &= u^t + \tfrac{1}{p} r(x^{t+1}) \\ y_i^{t+\frac{1}{2}} &= -g_i(x_i^{t+1}) \end{aligned}$$

where $r(x) = \sum_i g_i(x_i)$ is the residual of the relaxed coupling constraints.

The complete algorithm, called (SALA) for Separable Augmented Lagrangian Algorithm, is then :

Algorithm 6 SALA : a Separable Augmented Lagrangian Algorithm

Require: $t = 0, \lambda > 0, \epsilon > 0, u^0 \in I\!R^m, \zeta^0 = (y_1^0, \ldots, y_p^0) \in \mathcal{A}^\perp$

repeat
  for all $i = 1, \ldots, p$ **do**
    $x_i^{t+1} :\in \arg\min_{x_i \in S_i} f_i(x_i) + \langle u^t, g_i(x_i)\rangle + \frac{\lambda}{2}\|g_i(x_i) + y_i^t\|^2$
  end for
  $r^{t+1} \leftarrow \sum_{i=1}^p g_i(x_i^{t+1})$
  for all $i = 1, \ldots, p$ **do**
    $y_i^{t+1} \leftarrow -g_i(x_i^{t+1}) + \frac{1}{p} r^{t+1}$
  end for
  $u^{t+1} \leftarrow u^t + \frac{\lambda}{p} r^{t+1}$
  $t \leftarrow t + 1$
until $\|r(x^{t+1})\| < \epsilon$

Observe that the variables ζ must lie in the orthogonal subspace $\mathcal{A}^\perp = \{\zeta = (\zeta_1, \ldots, \zeta_p) \mid \sum_i \zeta_i = 0\}$. They are exactly the right-hand side allocations $\zeta_i = -g_i(x_i)$ used in resource-directive primal decomposition (see Lasdon for example [44]). Indeed, the primal separable counterpart of (13) is :

$$\text{Minimize}_{\zeta=(\zeta_1,\ldots,\zeta_p)\in\mathcal{A}^\perp} \quad \sum_{i=1}^p v_i(\zeta_i) \tag{14}$$

where $v_i(\zeta_i) = \inf\{f_i(x_i) \mid g_i(x_i) = -\zeta_i, x_i \in S_i\}$ is the implicit primal block function (convex on a convex domain with the current hypotheses).

In other words, (SALA) is exactly the application of an Augmented Lagrangian algorithm to the resource-directive reformulation (14) followed by the projection steps on the respective subspaces.

We observe here that Separable Augmented Lagrangian algorithms have been early proposed in the literature (see [70, 6, 16] for instance), but these methods do not rely on splitting schemes, rather linearizing the non separable terms to give rise to three-levels decomposition schemes with extensions to the non convex case.

3.2 Alternate direction method of multipliers

We consider now another situation involving two different convex functions. The functions can be smooth or not and are generally composite as they include linear transformations of the variables, so that many authors have analyzed the following *M-model* :

$$\text{Minimize } f(x) + g(Mx) \tag{P2}$$

with $f : I\!R^n \mapsto I\!R$ strongly convex, $g : I\!R^m \mapsto I\!R$ simply convex and M is a (generally full-rank) $(m \times n)$ matrix. A dual formulation is also convenient as

the conjugate function of f (i.e. $f^*(y) = \sup\{\langle y, x\rangle - f(x)\}$) is differentiable and the dual problem associated with $(P2)$ presents the same nice structure as the primal :

$$\text{Minimize } f^*(M^T u) + g^*(-u) \tag{15}$$

Even if this model can be reformulated as $\{\min\ f^*(v) + g^*(-u) \mid v + M^T u = 0\}$, thus minimizing the sum of two convex functions on a linear subspace, the direct application of the Forward-Backward splitting is natural using $T_1 = \partial f^*$ and $T_2 = \partial g^* \circ (-M^T)$. The resulting algorithm (see [30] and [72]) is given below :

Algorithm 7 FB-M

Require: $t = 0, \lambda > 0, \epsilon > 0, x^0, z^0, u^0$

1: $x^{t+1} := \arg\min f(x) - \langle u^t, Mx\rangle$
2: $z^{t+1} := \arg\min g(z) + \langle u^t, z\rangle + \frac{\lambda}{2}\|z - Mx^{t+1}\|^2$
3: $u^{t+1} := u^t - \lambda(Mx^{t+1} - z^{t+1})$
4: $t \leftarrow t + 1$

Convergence properties of the FB-M scheme have been analyzed by Chen and Rockafellar [13].

Application of (DRA) to the M-Model leads naturally to what is generally referred to as the *Alternate Direction Method of Multipliers* (ADMM) originally studied by Gabay [30] (see too [7]).

Algorithm 8 ADMM

Require: $t = 0, \lambda > 0, \epsilon > 0, x^0, z^0, u^0$

1: $x^{t+1} := \arg\min f(x) - \langle u^t, Mx\rangle + \frac{\lambda}{2}\|z^t - Mx\|^2$
2: $z^{t+1} := \arg\min g(z) + \langle u^t, z\rangle + \frac{\lambda}{2}\|z - Mx^{t+1}\|^2$
3: $u^{t+1} := u^t - \lambda(Mx^{t+1} - z^{t+1})$
4: $t \leftarrow t + 1$

The structural links between (ADMM), (DRA) and the Proximal Point method (PPM) has been extensively detailed by Eckstein (see [27, 28]). One recurrent question which remains partly open with (ADMM) is the extension to more than two blocks of variables. As mentioned early by Lions and Mercier [49], the Douglas-Rachford splitting method does not naturally generalize to more than two operators. Even if an extension of (DRA) with n operators have been early proposed by Douglas and Gunn [21], the sequential steps in (ADMM) turn the convergence analysis more intricate. Inspired by multi-block Gauss-Seidel algorithm, Goldfarb and Ma [34] proposed a generalized ADMM which converges in $O(1/\sqrt{\epsilon})$ iterations to obtain an approximation within ϵ of the optimal value but requires that all functions are smooth. Of course, the question can be turned around as commented before, by including the separable functions into a single dual operator and

forcing the dual copies in a coupling subspace playing the role of the second operator, as in the (PDA) algorithm.

There is still a comment on ADMM assumptions and rates of convergence. Many authors have studied the convergence properties of ADMM (which assumptions on the functions and, with which rates of convergence ?) as shown in the recent monograph [9], beyond the natural results inherited from (DRA) as early studied by Lions and Mercier [49]. It appears that linear convergence is possible when at least one of the block function is strongly convex. A recent study by Hong and Luo [40] made new progress in relaxing these assumptions and extending too to more than two blocks. Their separable model extends the M-model to include the S-model used in the former section. It also uses block functions of the form $f_i(x_i) = f_{0i}(A_i x_i) + f_{1i}(x_i)$ where f_{0i} are strictly convex and smooth and f_{1i} are typically polyhedral. Some steering matrices A_i can be zero allowing not strongly convex blocks. The coupling constraints should be linearly independent, i.e. the matrix $[G_1|\cdots|G_p]$ is full-rank, but some G_i may not. These assumptions allow to prove linear convergence of the primal and dual sequences as well as the sequence of the coupling constraints violation. Similar results in earlier papers exploiting the strongly convex parts in the objective function can be cited here, like [38, 57, 42].

We have seen in this section the impact of problem reformulation to get constructive variants of the basic splitting schemes. The literature has focussed mainly on the M-Model and its dual version (15). An alternative and interesting primal-dual setting has been recently proposed by Chambolle and Pock [12] who used a saddle-point formulation :

$$\inf_x \sup_y f(x) - g^*(y) + \langle y, Mx \rangle$$

and proved convergence of the following splitting algorithm :

Algorithm 9 CP

Require: $t = 0, \lambda > 0, \epsilon > 0, x^0, y^0$

1: $x^{t+1} := \arg\min f(x) + \frac{1}{2\lambda}\|x - x^t + \lambda M^T y^t\|^2$
2: $y^{t+1/2} := \arg\min g^*(y) + \frac{1}{2\mu}\|y - y^t - \mu M x^{t+1}\|^2$
3: $y^{t+1} := y^{t+1/2} + \theta(y^{t+1/2} - y^t)$
4: $t \leftarrow t + 1$

where $0 \leq \theta \leq 1$ and scaling parameters chosen such that $\lambda\mu\|M\|^2 < 1$. Convergence to a saddle-point (x^*, y^*) was proved in the ergodic sense with rate $O(1/t)$.

4 Convergence results and complexity issues

We close this survey by inspecting practical issues concerning operator splitting techniques. Before relating the different extensions which have been investigated, and they are quite numerous, it is of interest to overview the areas of applications where splitting methods have been used successfully. Historically, R. Temam [71], R. Glowinski [32], Gabay and Mercier [31] among many others, have used (DRA) to solve evolution equations in Mechanics, referring the approach as a penalty-dualization method. As the formal setting involves inclusion problems with general monotone operators, the application of splitting methods to *Variational Inequalities* is still an attractive subject for applied mathematicians. Besides these applications, the report of Bensoussan et al [5] and more recently, the textbook by Bertsekas and Tsitsiklis [7] have largely contributed to disseminate these techniques to the areas of Mathematical Programming and Operations Research where decomposition techniques are very popular since the sixties. Among many different areas of applications, we can cite Multicommodity Flow problems with convex arc costs ([52], [29]), Stochastic Programming (adapting (DRA) to two-stage stochastic optimization with recourse leads to the *Progressive Hedging* method of Rockafellar and Wets [65]), Fermat-Weber problems (the Partial Inverse of Spingarn was applied to a polyhedral operator splitting model in [41]). More recently, new models received a lot of interest in the areas of Image Reconstruction and Signal Processing ([12, 19]), with similar models in Classification problems [35, 9]. These models involve in general the combination of two norms, one smooth and not the other, inducing the use of the M-model. For instance, the *lasso* or compressed sensing problem considers $f(x) = \|x\|_1$ and $g(x) = \frac{1}{2\mu}\|y - Ax\|^2$, where A is a huge sparse matrix, with applications in deblurring images [12] or classifying big data [35].

Motivated by these applications, new versions of the classical splitting schemes have been proposed. As a striking example, we note the recent primal-dual splitting, inspired by the Forward-Backward scheme, on inclusion problems involving composite operators like $0 \in Ax + K^*BKx$ where A and B are general maximal monotone operators and K is a linear continuous operator with adjoint K^*. The convergence of the explicitly composite case was first analyzed by Briceño-Arias and Combettes in [11] (see too [8] and [1] for an extension of Spingarn's Partial Inverse method).

We will present first some algorithmic enhancements relative to convergence issues and, in the second part, discuss numerical scaling issues.

4.1 Algorithmic enhancements

Many variants of the basic schemes have been analyzed in the literature and we focus here on a few important choices that can produce new decomposi-

tion methods rather than on the variety of models to which these schemes may be applied. Most of these enhancements correspond to known variants of the Proximal Point method itself.

We will discuss the following issues :

- the introduction of additional regularizing terms;
- approximate solutions in the proximal steps;

A recent survey by Shefi and Teboulle [66] recalled the complexity issues to improve global convergence ratios, in particular for regularized versions of the splitting schemes using ideas from the Proximal method of Multipliers of Rockafellar [62]. The corresponding extension of the (FB-M) algorithm was proposed by Chen and Teboulle [14] and a similar idea was used by Eckstein [25] to extend (ADMM). The Proximal-ADMM can be sketched as the following algorithm :

Algorithm 10 Prox-ADMM

Require: $t = 0, \lambda > 0, \epsilon > 0, x^0, z^0, u^0$

1: $x^{t+1} := \arg\min f(x) - \langle u^t, Mx\rangle + \frac{\lambda}{2}\|z^t - Mx\|^2 + \frac{1}{2\mu}\|x - x^t\|^2$
2: $z^{t+1} := \arg\min g(z) + \langle u^t, z\rangle + \frac{\lambda}{2}\|z - Mx^{t+1}\|^2 + \frac{1}{2\mu}\|z - z^t\|^2$
3: $u^{t+1} := u^t - \lambda(Mx^{t+1} - z^{t+1})$
4: $t \leftarrow t + 1$

Global convergence of the Proximal-ADMM algorithm and variants are considered in [66] with refined convergence rates. In the general case without further assumptions on f and g, typical ergodic convergence is exhibited with $O(1/\sqrt{t})$ global rate. Assuming g is convex Lipschitz continuous, then the whole primal-dual sequence converges with $O(1/t)$ global rate.

Approximate solutions in the proximal steps have been considered early by Rockafellar [63] and further introduced in many splitting schemes (see [17]). A typical inexact version of (DRA) will be :

Algorithm 11 Inexact-ADMM

Require: $t = 0, \lambda > 0, \epsilon_0 > 0, x^0, z^0, u^0$

1: $x^{t+1} :\approx_{\epsilon_{1t}} \arg\min f(x) - \langle u^t, Mx\rangle + \frac{\lambda}{2}\|z^t - Mx\|^2$
2: $z^{t+1} :\approx_{\epsilon_{2t}} \arg\min g(z) + \langle u^t, z\rangle + \frac{\lambda}{2}\|z - Mx^{t+1}\|^2$
3: $u^{t+1} := u^t - \lambda(Mx^{t+1} - z^{t+1})$
4: $t \leftarrow t + 1$

where $a :\approx_\epsilon b$ is a shorthand for $\|a - b\| \leq \epsilon$. As in the inexact version of the Proximal Point method, convergence is maintained if the errors satisfy $\sum_{t=0}^{\infty} \epsilon_t < +\infty$ (see [27]).

4.2 Scaling and numerical enhancements

Many authors have revisited the basic splitting methods discussed in the former sections to improve convergence results and obtain implementable algorithms with optimized performance. As mentioned before, the introduction of a second parameter to generate a family of Peaceman-Rachford pure iteration appeared in Lions and Mercier's paper [49]. Later, Spingarn in [69] suggested to use a weighted scalar product to scale the projection step, but the difficulty to adjust many parameters during the convergence process reduced the attempts to implement these ideas. By the way, many authors have considered the introduction of different parameters associated with scaling matrices in the same spirit of what has been investigated for Augmented Lagrangian algorithms. The particular sensitivity to parameter λ of the supposedly most efficient splitting, i.e. (DRA) or (PDA), deserves special attention.

4.2.1 Convergence rates for a single scaling parameter

To analyze the role of the scaling parameter, we use here the Proximal Decomposition setting where $T_1 = T$ is a maximal monotone operator, coercive with constant ρ and Lipschitz with constant L, and T_2 is the subdifferential of the indicator of a linear subspace, i.e. $\mathrm{Gr}(T_2) = \{(x, y) \in X \times X \mid (x, y) \in \mathcal{A} \times \mathcal{A}^{\perp}\}$. Recall that this corresponds to minimizing a strongly convex function (possibly separable) on a coupling subspace. The scaled proximal decomposition on the graph of T with parameter $\lambda > 0$ corresponds to the following step : given $(x, y) \in X \times X$, find the unique pair $(u, v) \in \mathrm{Gr}(T)$ such that $u + \lambda v = x + \lambda y = z$. The corresponding proximal steps on the primal and dual variables are :

$$\begin{aligned} u &= (I\!I + \lambda T)^{-1}(z) \\ v &= \lambda^{-1}(z - u)) \end{aligned}$$

which are followed by the projection steps on $\mathcal{A} \times \mathcal{A}^{\perp}$. It was proved by Lions and Mercier [49] for the general case (T_2 is a maximal monotone operator) and revisited by Mahey et al [51] for the proximal decomposition that linear convergence is guaranteed with a convergence rate

$$r_u(\lambda) = \sqrt{1 - \frac{2\lambda\rho}{(1 + \lambda L)^2}}$$

inducing the best choice for $\lambda = 1/L$ with optimal rate $\sqrt{1 - \frac{\rho}{2L}}$. Observe now that the update of the dual variable y is equivalent to :

$$v = (I\!I + \mu T^{-1})^{-1}(\mu z)$$

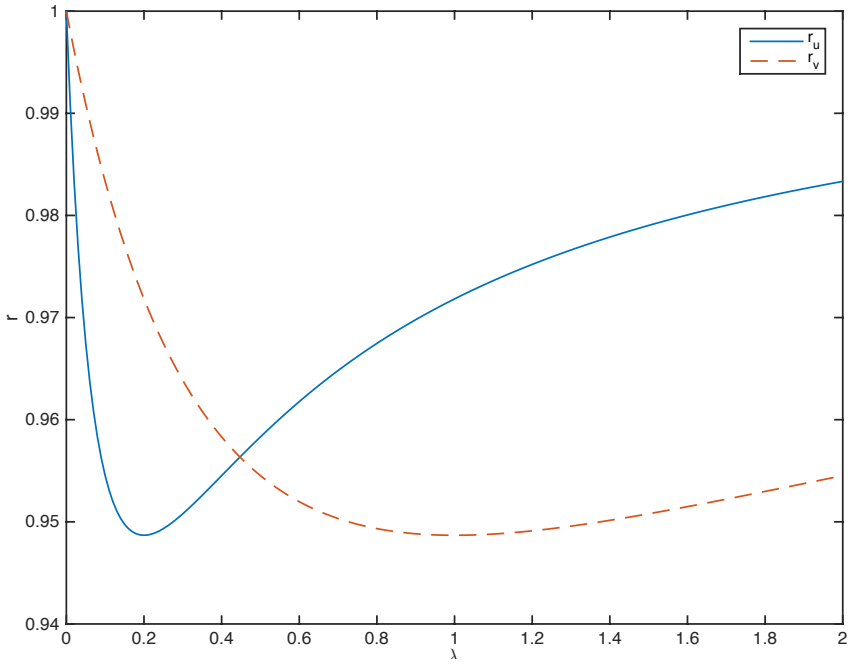

Figure 4: Primal and dual optimal convergence rates

with $\lambda\mu = 1$. We can then derive an upper bound for the rate of the dual sequence as we did for the primal one, noting that the coercivity constant for T^{-1} is $1/L$ and its Lipschitz constant is $1/\rho$. Straightforward calculations give the convergence rate (of the dual sequence) :

$$r_v(\lambda) = \sqrt{1 - \frac{2\lambda\rho^2}{L(1+\lambda\rho)^2}}$$

with the best choice $\lambda_v = 1/\rho$ giving the optimal rate $\sqrt{1 - \frac{\rho}{2L}}$.

Then the optimal rates are equal but for different values of the parameter. Figure 4 (taken from [50]) shows the two rates as function of λ. The best compromise is to minimize $\max(r_u(\lambda), r_v(\lambda)$ which yields

$$\lambda^* = \frac{1}{\sqrt{\rho L}}$$

Following the analysis in [50], we can obtain lower and upper bounds for both primal and dual sequences using the properties of operator T :

$$\begin{aligned} \frac{1}{1+\lambda L}\|z - z'\|^2 &\leq \|u - u'\|^2 \leq \frac{1}{1+\lambda\rho}\|z - z'\|^2 \\ \frac{\lambda\rho}{1+\lambda\rho}\|z - z'\|^2 &\leq \|u - u'\|^2 \leq \frac{\lambda L}{1+\lambda L}\|z - z'\|^2 \end{aligned}$$

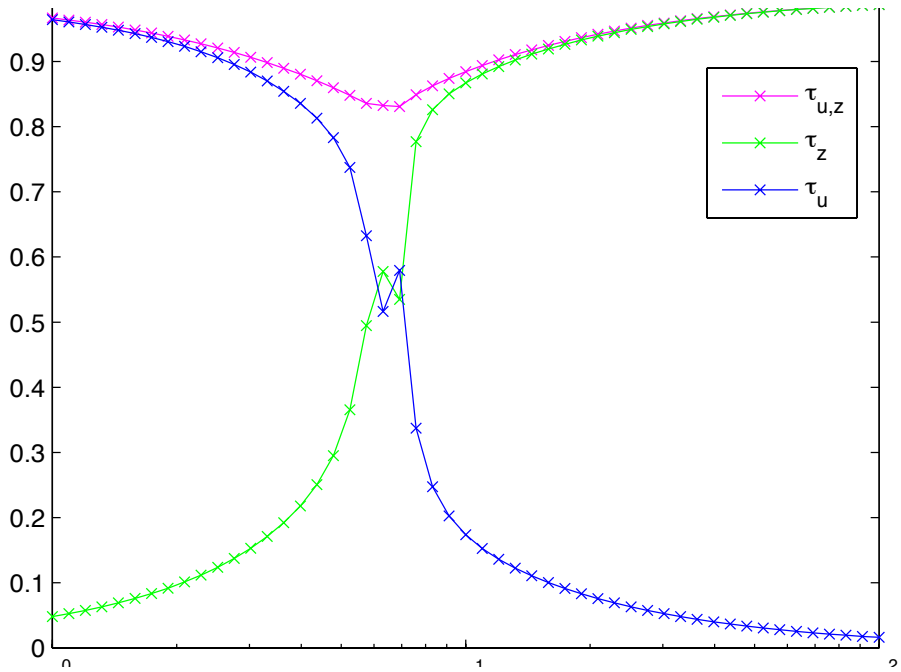

Figure 5: Primal-dual residual rates

Setting $\lambda = \lambda^*$, we get the same bounds (LB for the lower bound and UB for the upper bound) for both sequences :

$$\begin{aligned} LB &= \frac{1}{1+\sqrt{\frac{L}{\rho}}} \leq 0.5 \\ UB &= \frac{1}{1+\sqrt{\frac{\rho}{L}}} \geq 0.5 \end{aligned}$$

Ideally, we would expect a rate of 0.5 for both sequences when λ is tuned to its compromise value λ^*. Figure 5 compares the behaviour of primal and dual rates (estimated during the tail of iterations) τ_u, and τ_z respectively, with respect to parameter λ, and the primal-dual rate $\tau_{uz} = \lim \frac{\| \begin{pmatrix} u^{t+1} \\ z^{t+1} \end{pmatrix} - \begin{pmatrix} u^* \\ z^* \end{pmatrix} \|}{\| \begin{pmatrix} u^t \\ z^t \end{pmatrix} - \begin{pmatrix} u^* \\ z^* \end{pmatrix} \|}$, the latter keeping a deceivingly slow pace.

This rather frustrating limiting rate of 0.5 induces us to explore the possibility of multidimensional parameters, as discussed below.

Another drawback, early observed by Eckstein [24] is spiralling which tends to slow down the iterates in the neighborhood of a fixed point (a typical behaviour is shown in Figure 7). This phenomenon was currently observed when splitting is applied to polyhedral (thus not strongly monotone) models.

4.2.2 Multidimensional scaling

It is easy to extend the former splitting methods to a multidimensional scaling strategy. To understand the transformation, for a positive definite matrix M, let consider a variable change $z = Mx$ and substitute the monotone operator T by $\mathcal{T} = M^{-T} \circ T \circ M^{-1}$. $\mathcal{T}$ is indeed maximal monotone if T is so and their graphs correspond in the following way :

$$y \in Tx \Longleftrightarrow u \in \mathcal{T}(z) \text{ for } u = M^{-T}y \text{ and } z = Mx$$

Consequently, a backward step with the scaled operator $\mathcal{T}$ derived back in the original x-space using the inverse transformation $x = M^{-1}z$, corresponds to the resolvent operator $(I\!I + \Lambda T)^{-1}$ where $\Lambda = M^T M$.

Setting the scaling parameter $\lambda = 1$, it appears that the matrix Λ plays the role of the scaling parameter in the splitting method. Convergence result are identical but different convergence rates are expected if the scaling matrix Λ is adequately chosen. Most practical applications of multidimensional scaling suggest to use a diagonal scaling to turn the estimation of the parameters easier. The choice of the scaling matrix depends on the model (or its reformulation) to maintain the decomposition features of the method. For instance, considering the S-Model, the scaling corresponds to substituting each resource allocation in $(S - Model)$ by $Mg_i(x_i) + y_i = 0$. It is shown on a simple quadratic example with two coupling constraints that two different parameters can drastically improve convergence (see Figure 6 illustrating the rate of convergence of the primal-dual sequences with a two-dimensional scaling).

Many applications of operator splitting techniques include preconditioning techniques which is indeed equivalent to multidimensional scaling, the exhaustive list of references is far too long to be included here (see for instance [9]).

4.2.3 Spiralling and foldings

Analyzing now the spiralling effect observed in many practical situations where strong monotonicity is not present, it was shown in [47] an acceleration strategy based on iterative foldings. The idea takes roots in a former study by Lawrence and Spingarn [45] on folding operators. In the case where $T_1 = \partial F$ and $T_2 = \partial C$ with F and C polyhedral, they introduced a family of averaged operators $\mathcal{F}$, called foldings, which are piecewise isometric, non expansive and positively homogeneous, and showed that the iteration

$$s^{t+1} = \frac{1}{2}s^t + \frac{1}{2}\mathcal{F}(s^t)$$

produces successive directions which generate linear subspaces, themselves converging to a limit subspace where spiralling is expected. In [47], the authors used the following hypotheses on the operators :

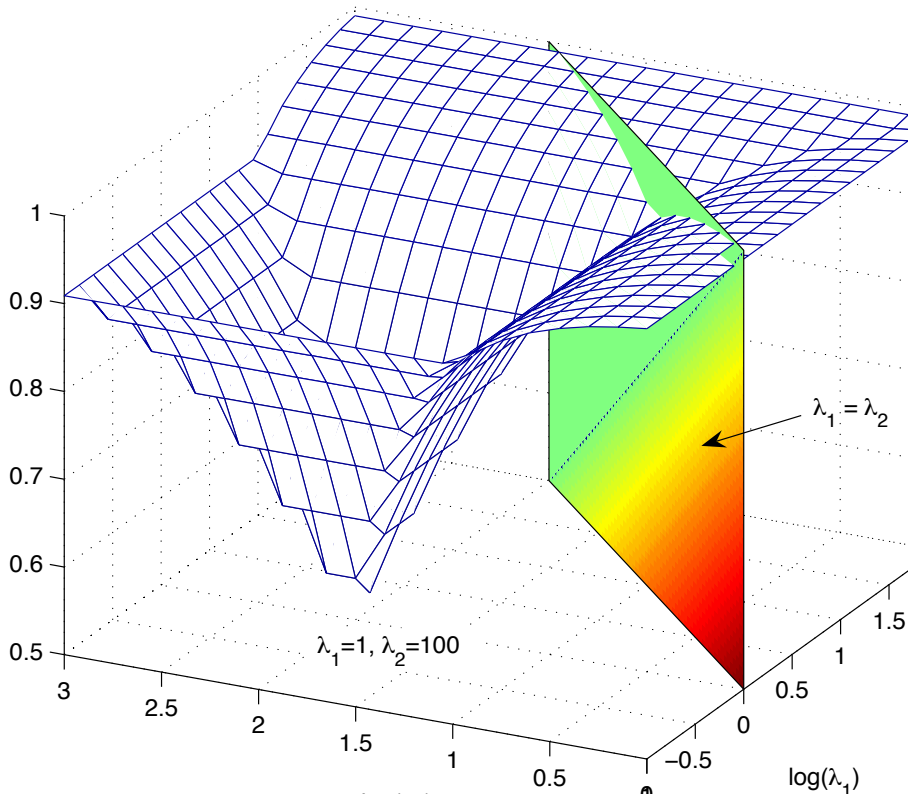

Figure 6: Multidimensional scaling

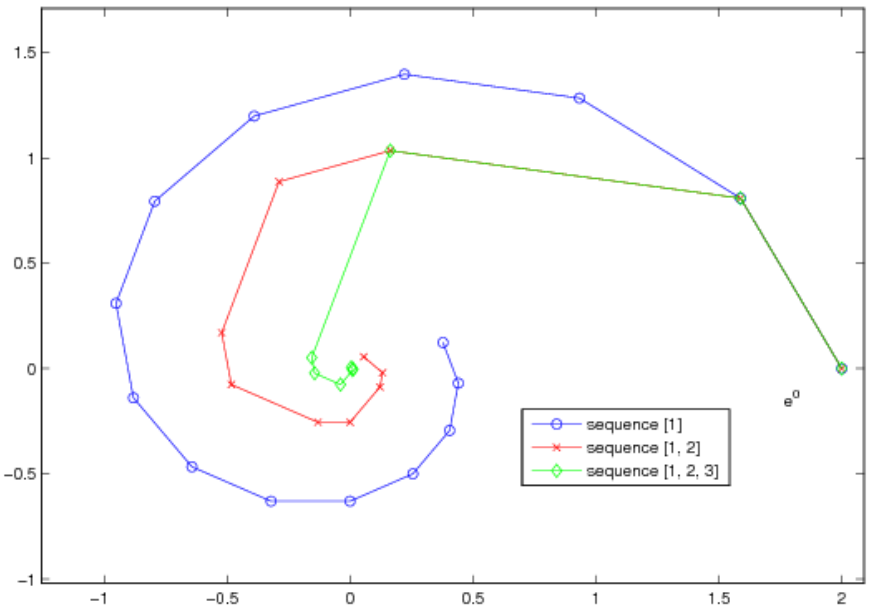

Figure 7: Breaking spiralling by iterative folding

- The subdifferential operator $T_1 = \partial F$ is *proto-differentiable*, which means that the graph of the directional variations of T_1 converges graphically to a limit operator, the proto-derivative of T_1, a weaker notion than differentiability (see [64] for details).
- C is the the indicator function of a coupling subspace.

In this case, they show how to accelerate convergence by application of averaged sequence of foldings to break spiralling (see Figure 7).

4.2.4 Variable scaling and parameter updates

The high sensitivity to the value of the scaling parameters turns their estimation a difficult issue in practice. So, many researchers have considered the generalization of splitting methods with varying scaling parameters at each iteration. The earliest proposal to our knowledge is a variant of the Forward-Backward splitting proposed by Tseng [72]. In his analysis, he gave conditions on the scaling parameter to guarantee linear convergence of the algorithm, based on the previous knowledge of the co-coercivity radius of one of the two operators. This is of course difficult to check in practice. The use of adaptive updates of parameters in the splitting algorithms has been early studied by Kontogyorgis and Meyer in [43] with the additional difficulty of revising the theoretical convergence results. Typically, in the case of a single parameter λ, convergence will be maintained if the sequence

$\{\lambda_t\}$ of parameters converges to a limit value and satisfies

$$\sum_{t=0}^{+\infty} |\lambda_{t+1} - \lambda_t| < +\infty$$

Practical choices for the implementation of these updates appeared in [39] for (ADMM) and in [23] for (SALA), taking the following form :

$$\lambda_{t+1} = \begin{cases} \theta\lambda_t & \text{if } 0 \leq t \leq 100 \text{ and } t \equiv 0 \pmod{10} \\ \lambda_t & \text{else} \end{cases}$$

where $\theta \in [0.5\ 1]$.

More sophisticated updates that take in consideration the relative behaviour of primal and dual sequences are proposed in [50]. Based on the S-Model, their analysis estimates the primal and dual rates by computing $\tau_x = \frac{r^{t+1}}{r^t}$ where $r^t = \sum_i g_i(x_i^t)$ is the primal residual at iteration t, and $\tau_u = \frac{\delta^{t+1}}{\delta^t}$ where $\delta^t = \nabla L(x^t, u^t)$ the gradient of the ordinary Lagrangian of the S-Model. The proposed update of the parameter is such that both sequences are kept at a similar pace and is implemented by :

$$\lambda_{t+1} = \left(\frac{\tau_x}{\tau_u}\right)^{\alpha} \lambda_t$$

with $0 < \alpha < 1$.

5 Conclusion

We have surveyed the main monotone operator splitting methods and their applications to the decomposition of separable convex problems. This is a still very active research area where recent motivations concerning large-scale problems in signal processing and statistical learning have induced many new adaptations of these relatively old methods which appeared as early as the fifties with Douglas- and Peaceman-Rachford algorithms for linear operators. As these techniques can be interpreted as separable versions of the Augmented Lagrangian dual methods, their main benefit is the regularization effect of the proximal steps which induce numerical stability and smoothness of the implicit primal and dual value functions. We have seen the importance of reformulation to better exploit the decomposition features of each splitting scheme, for example by opposing the S-Model and the M-model. This reveals that the operator splitting techniques are potential candidates for the decomposition of nonconvex problems, even if this has been relatively little explored in the applications (see [2] for theoretical extensions dealing with semi-algebraic functions). The main drawback which has slowed down their practical use is the difficulty to reach better convergence rates, like superlinear convergence which is unlikely to occur even in the strongly monotone

models, neither theoretically nor practically. Nevertheless, we have discussed that question in the last section showing that multidimensional scaling along with adaptive updates of the parameters can significantly improve the speed of convergence in many concrete applications.

Finally, as for most decomposition methods, parallel implementations are natural issues that have been tested by different authors which were not surveyed in the present paper (see Bertsekas and Tsitsiklis [7] or Eckstein's thesis [24]).

References

[1] M.A. Alghamdi, A. Alotaibi, P.L. Combettes, and N. Shahzad. A primal-dual method of partial inverses for composite inclusions. *Optimization Letters*, 8:2271–2284, 2014.

[2] 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. Programming*, 137:91–129, 2013.

[3] J.B. Baillon, R.E. Bruck, and S. Reich. On the asymptotic behavior of nonexpansive mappings and semi-groups. *Houston J. of Mathematics*, 4:1–9, 1978.

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

[5] A. Bensoussan, J.L. Lions, and R. Temam. Sur les méthodes de décomposition, décentralisation et de coordination et applications. Technical report, Cahiers de l'INRIA, 1972.

[6] D.P. Bertsekas. Convexification procedures and decomposition methods for nonconvex optimization problems. *J. Optimization Theory and Appl.*, 29:169–197, 1979.

[7] D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[8] R.I. Bot, E.R. Csetnek, and A. Heinrich. A primal-dual splitting for finding zeros of sums of maximally monotone operators. *SIAM J. Optimization*, 23:2011–2036, 2013.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning with the alternating direction method of multipliers. In M. Jordan, editor, *Foundations and Trends in Machine Learning*, volume 3, pages 1–122. 2011.

[10] H. Brezis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert.* Lecture Notes 5. North-Holland, 1973.

[11] L.M. Briceno-Arias and P.L. Combettes. A monotone+skew splitting model for composite monotone inclusions in duality. *SIAM J. Optimization*, 21:1230–1250, 2011.

[12] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex programs with applications to imaging. *J. of Math. Imaging and Vision*, 40:1–26, 2011.

[13] G.H.G Chen and R.T. Rockafellar. Convergence rates in forward-backward splitting. *SIAM J. Optimization*, 7:421–444, 1997.

[14] G.H.G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Math. Programming*, 64:81–101, 1994.

[15] G. Cohen. Auxiliary problem principle and decomposition of optimization problems. *J. Opt. Theory Appl.*, 32:277–305, 1980.

[16] G. Cohen and D.L. Zhu. Decomposition coordination methods in large-scale optimization problems. the nondifferentiable case and the use of augmented lagrangians. In J.B. Cruz Junior, editor, *Advances in Large-Scale Systems*, volume 1. JAI Press Inc., 1983.

[17] P.L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5):475–504, October 2004.

[18] P.L. Combettes and J.C. Pesquet. Proximal splitting methods in signal processing. In P.L. Combettes V. Elser D.R. Luke H. Wolkowicz H.H. Bauschke, R.S. Burachik, editor, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer V., 2011.

[19] P.L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modelling and Simulation*, 4(4):1168–1200, 2005.

[20] D. Davis and W. Yin. Convergence rate analysis of several splitting schemes. Technical report, UCLA CAM report 14-51, 2014.

[21] J. Douglas and J.E. Gunn. A general formulation of alternating direction methods. *Numer. Math.*, (6):428–453, 1964.

[22] J. Douglas and H. H. Rachford. On the numerical solution of the heat conduction problem in 2 and 3 space variables. *Trans. Amer. Math. Soc.*, (82):421–439, 1956.

[23] J.P. Dussault, O.M. Gueye, and P. Mahey. Separable augmented lagrangian algorithm with multidimensional scaling for monotropic programming. *Journal of Optimization Theory and Application*, 127:329–345, 2005.

[24] J. Eckstein. *Splitting methods for monotone operators with applications to parallel optimization.* PhD thesis, Massachusetts institute of technology, cambridge, June 1989.

[25] J. Eckstein. Some saddle-point splitting method for convex programming. *Optimization Methods and Software*, 4:75–83, 1994.

[26] J. Eckstein. Augmented lagrangian and alternating direction method of multipliers : A tutorial and some illustrative computational examples. Technical report, RUTCOR Research Rept - RR32-2012, December 2012.

[27] J. Eckstein and D. P. Bertsekas. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Math. Programming*, 55:293–318, 1992.

[28] J. Eckstein and M. Fukushima. Some reformulations and applications of the alternating direction method of multipliers. *Large Scale Optimization : State of the Art*, pages 119–138, 1993.

[29] M. Fukushima. Application of the alternating direction method of multipliers to separable convex programming. *Comput. Optimization and Appl.*, 1:93–112, 1992.

[30] D. Gabay. *Applications of the method of multipliers to variational inequalities*, volume 15 of *Studies in Mathematics and its Applications.* 1983.

[31] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computer Math. Appl.*, 2:17–40, 1976.

[32] R. Glowinski and P. Le Tallec. *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics.* SIAM Philadelphia, 1989.

[33] R. Glowinski and A. Marocco. Sur l'approximation par éléments finis d'ordre 1 et la résolution par pénalisation-dualité d'une classe de problèmes de dirichlet. *RAIRO*, 2:41–76, 1975.

[34] D. Goldfarb and S. Ma. Fast multiple splitting algorithms for convex optimization. *SIAM J. Optimization*, 22:533–556, 2012.

[35] D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming*, 2009.

[36] A.A. Goldstein. Convex programming in hilbert space. *Bulletin of the Amer. Math. Society*, 70:709–710, 1964.

[37] A. Hamdi, P. Mahey, and J.P Dussault. A new decomposition method in nonconvex programming via a separable augmented lagrangian. *Lecture Notes in Economics and Mathematical Systems*, 452:90–104, 1997.

[38] S.P. Han and G. Lou. A parallel algorithm for a class of convex programs. *SIAM J. Control and Optim.*, 26:345–355, 1988.

[39] B.S. He, H. Yang, and S.L. Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *J. of Optimization Theory and Applications*, 106:349–368, 2000.

[40] M. Hong and Luo Z.Q. Convergence of the alternating method of multipliers. *SIAM J.Control and optimization*, 2013.

[41] H. Idrissi, Odile Lefebvre, and Christian Michelot. Applications and numerical convergence of the partial inverse method. *Lecture notes in mathematics*, 1405:39–54, 1989.

[42] K.C. Kiwiel, C.H. Rosa, and A. Ruszczynski. Proximal decomposition via alternating linearization. *SIAM J. Optim.*, 9:668–689, 1999.

[43] S. Kontogiorgis and R.R. Meyer. A variable-penalty alternating directions method for convex optimization. *Mathematical Programming*, 83:29–53, 1998.

[44] L.S. Lasdon. *Optimization for Large Systems.* Mac Millan, 1970.

[45] J. Lawrence and J.E. Spingarn. On fixed points of nonexpansive piecewise isometric mappings. *Proceedings of the London Mathematical Society*, 55:605–624, 1987.

[46] A. Lenoir. *Modéles et algorithmes pour la planification de production á moyen terme en environnement incertain.* PhD thesis, Université Blaise Pascal, Clermont-Ferrand, 2008.

[47] A. Lenoir and P. Mahey. Accelerating a class of splitting algorithms by iterative foldings. *Acta Mathematica Vietnamica*, 39:49–65, 2009.

[48] J. Lieutaud. *Approximation d'opérateurs par les méthodes de décomposition.* PhD thesis, Université de Paris, 1969.

[49] P.L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16:964–979, 1979.

[50] P. Mahey, J.P. Dussault, A. Benchakroun, and A. Hamdi. Adaptive scaling and convergence rates of a separable augmented lagrangian algorithm. In V.H. Nguyen, J.J. Strodiot, and P. Tossings, editors, *Optimization*, volume 481 of *Lecture Notes in Economics and Mathematical Systems*, pages 278–287. Springer, 2000.

[51] P. Mahey, S. Oualibouch, and D.T. Pham. Proximal decomposition on the graph of a maximal monotone operator. *SIAM J. Optimization*, 5:454–466, 1995.

[52] P. Mahey, A. Ouorou, L. Leblanc, and J. Chifflet. A new proximal decomposition algorithm for routing in telecommunication networks. *Networks*, 31:227–238, 1998.

[53] B. Martinet. Régularistion d'inéquations variationnelles par approximations successives. *Revue Française d'Informatique et de Recherche Opérationnelle*, pages 154–159, 1970.

[54] B. Mercier. *Topics in finite-element solution of elliptic problems.* Lectures on Mathematics 63. Tata Institute of Fundamental Research, Bombay, 1979.

[55] G.J. Minty. Monotone operators in hilbert spaces. *Duke Math. Journal*, 29:341–346, 1962.

[56] Jean-Jaccques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique Française*, 93:273–299, 1965.

[57] K. Mouallif, V.H. Nguyen, and J-J. Strodiot. A perturbed parallel decomposition method for a class of nonsmooth convex minimization problems. *SIAM J. Control and Optim.*, 29:829–847, 1991.

[58] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

[59] G.B. Passty. Ergodic convergence to a zero of the sum of monotone operators in hilbert spaces. *J. Math. Anal. Appl.*, 72:383–390, 1979.

[60] D.H. Peaceman and H. H. Rachford. The numerical solution of parabolic elliptic differential equations. *J. Soc. Indust. Appl. Math.*, (3):28–41, 1955.

[61] G. Pierra. Decomposition through formalization on a product space. *Mathematical Programming*, 28:96–115, 1984.

[62] R. T. Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1:97–116, 1976.

[63] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J.Control and optimization*, 14:877–898, 1976.

[64] R. T. Rockafellar. Proto-differentiability of set-valued mappings and its application in optimization. *Annales de l'IHP*, S6:449–482, 1989.

[65] R. T. Rockafellar and R. J-B Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16:119–147, 1991.

[66] R. Shefi and M. Teboulle. Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. *SIAM Journal on Optimization*, 2014.

[67] M.V. Solodov. A class of decomposition methods for convex optimization and monotone variational inclusions via the hybrid inexact proximal point framework. *Optim. Methods and Software*, 19:557–575, 2004.

[68] J.E. Spingarn. Partial inverse of a monotone operator. *Applied mathematics and optimization*, 10:247–265, 1983.

[69] J.E. Spingarn. Applications of the method of partial inverses to convex programming:decomposition. *Mathematical Programming*, 32:199–223, 1985.

[70] G. Stephanopoulos and A.W. Westerberg. The use of hestenesÕs method of multipliers to re- solve dual gaps in engineering system optimization. *J. of Optimization Theory and App.*, 15:285Ñ309, 1975.

[71] Roger Temam. Sur la stabilité et la convergence de la méthode des pas fractionnaires. *Annali di Matematica Pura ed Applicata*, 79(1):191–379, 1968.

[72] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138, 1991.

[73] R.S. Varga. *Matrix Iterative Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1966.