Minimax Theory and its Applications
Volume 6 (2021), No. 1, 1–24

Fast Inertial Proximal ADMM Algorithms for Convex Structured Optimization with Linear Constraint

Hedy Attouch
IMAG, Université Montpellier, CNRS, 34095 Montpellier, France
hedy.attouch@umontpellier.fr

Received: March 6, 2020
Accepted: May 5, 2020

In a Hilbert space setting, we analyze the convergence properties of a new class of proximal ADMM algorithms with inertial features. They aim to quickly solve convex structured minimization problems with linear constraint. As a basic ingredient, we use the maximally monotone operator M which is associated with the Lagrangian formulation of the problem. We specialize to this operator the inertial proximal algorithm recently introduced by Attouch and Peypouquet [*Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators*, Math. Programming 174 (2019) 391–432] to resolve general monotone inclusions. This gives an inertial proximal ADMM algorithm where the extrapolation step takes into account recent advances concerning the accelerated gradient method of Nesterov. Based on an appropriate adjustment of the viscosity and proximal parameters, we analyze the fast convergence properties of the algorithm, as well as the convergence of the iterates to saddle points of the Lagrangian function. Among the perspectives, we outline a new direction of research, linked to the introduction of the Hessian damping in the algorithms.

Keywords: Convex structured optimization, linear constraint, Lagrange multipliers, maximally monotone operators, proximal ADMM, inertial methods, Nesterov accelerated method, Hessian damping.

2010 Mathematics Subject Classification: 37N40, 46N10, 49M30, 65K05, 65K10, 90B50, 90C25.

1. Introduction

Our study concerns the resolution by accelerated (ADMM) algorithms of the convex structured minimization problem with linear constraint

$$(\mathcal{P}) \qquad \min_{x\in X,\, y\in Y} \{f(x)+g(y) : Ax-By=0\}\,,$$

where the following standing assumptions are made on the data of $(\mathcal{P})$:

$$(H) \qquad \begin{cases} X, Y, Z \text{ are real Hilbert spaces,} \\ f\colon X\to \mathbb{R}\cup\{+\infty\},\ g\colon Y\to \mathbb{R}\cup\{+\infty\} \\ \quad \text{are convex lower semicontinuous proper functions,} \\ A\colon X\to Z \text{ and } B\colon Y\to Z \text{ are linear continuous operators.} \end{cases}$$

We set $\|x\|^2=\langle x,x\rangle$, $\|y\|^2=\langle y,y\rangle$, $\|z\|^2=\langle z,z\rangle$ for $x\in X, y\in Y, z\in Z$.[1]

[1] Without ambiguity, we don't use indexes to specify which space, scalar product or norm is considered.

ISSN 2199-1413 (printed), ISSN 2199-1421 (electronic) / $ 2.50

1.1. Lagrangian function and maximally monotone operator attached to $(\mathcal{P})$

Classically, we can reformulate $(\mathcal{P})$ as a saddle value problem

$$\min_{(x,y)\in X\times Y}\max_{z\in Z}\left\{f(x)+g(y)+\langle z, Ax-By\rangle\right\}. \tag{1}$$

The Lagrangian $L\colon X\times Y\times Z\to\mathbb{R}\cup\{+\infty\}$ associated to (1) is the extended-real-valued function

$$L(x,y,z)=f(x)+g(y)+\langle z, Ax-By\rangle.$$

It is convex with respect to (x,y), and affine (and hence concave) with respect to z. A pair (x,y) is optimal for $(\mathcal{P})$, and z is a corresponding Lagrange multiplier if and only if (x,y,z) is a saddle point of the Lagrangian function L. We denote by S the set of saddle points of L. Equivalently

$$(x,y,z)\in S \iff \begin{cases} \partial f(x)+A^t(z)\ni 0\\ \partial g(y)-B^t(z)\ni 0\\ B(y)-A(x)=0\end{cases}$$

where we use the classical notions and notations of the linear and convex analysis: ∂f is the subdifferential of f, ∂g is the subdifferential of g, $A^t\colon Z\to X$ is the transposed operator of A, and $B^t\colon Z\to Y$ is the transposed operator of B. The above system can be written equivalently as the monotone inclusion

$$M_P(x,y,z)\ni 0 \tag{2}$$

where $M_P\colon X\times Y\times Z\to 2^{X\times Y\times Z}$ is the set-valued mapping defined by

$$\begin{aligned} M_P(x,y,z)&=\left(\partial_{x,y}L,-\partial_z L\right)(x,y,z)\\ &=\left(\partial f(x)+A^t z,\ \partial g(y)-B^t z,\ By-Ax\right). \end{aligned} \tag{3}$$

The crucial point is that the operator M_P is maximally monotone on $X\times Y\times Z$. Indeed, this follows simply from the following observation. The operator M_P can be splitted as

$$M_P=M_1+M_2 \tag{4}$$

where $M_1(x,y,z)=(\partial f(x),\ \partial g(y),0)$, and $M_2(x,y,z)=(A^t z,\ -B^t z,\ By-Ax)$. Thanks to the classical rules of the subdifferential calculus, it can be observed that $M_1=\partial\Phi$ is the subdifferential of the convex lower semicontinuous proper function $\Phi(x,y,z)=f(x)+g(y)$, and therefore is maximally monotone. The operator M_2 is linear continuous and skew symmetric, and therefore it is also maximally monotone. This immediately implies that M_P is maximally monotone as the sum of two maximally monotone operators, one of them being Lipschitz continuous ([37, Lemma 2.4, p. 34]).[2] This can also be achieved as a consequence of the general properties relying convexity and monotonicity, see Rockafellar [68]. Thus, S can be interpreted as the set of zeros of a maximally monotone operator. As such, it is a closed convex subset of $X\times Y\times Z$. The application of the proximal algorithm to the maximally monotone operator M_P gives the proximal (ADMM) algorithm. This will be the guiding idea of our approach, with a non-trivial adaptation for the inertial versions.

[2] The above structure was used by Briceno-Combettes [38] to develop a primal-dual splitting method based on the forward-backward-forward algorithm.

1.2. Examples of problems $(\mathcal{P})$

The optimization problem $(\mathcal{P})$ intervenes in the modeling of a wide range of situations possibly involving spaces of infinite dimension. Let's briefly describe some of them:

(1) In the case of the finite dimension, the Regularized Least Squares method (RLS) is the composite optimization problem on $\mathbb{R}^n$,

$$\text{(RLS)} \qquad \min_{y\in\mathbb{R}^n}\left\{\frac{1}{2}\|By-b\|^2+g(y)\right\}$$

where B is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, $m\leq n$, $g\colon \mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ is a proper lsc. convex function which acts as a regularizer. Problem (RLS) falls within the setting of $(\mathcal{P})$ by taking $f(x)=\frac{1}{2}\|x\|^2$, and the affine coupling equation $x-By=-b$ (the operator A is the identity of $\mathbb{R}^m$).[3] Problem (RLS) is used intensively in a variety of fields ranging from inverse problems in signal/image processing, to machine learning and statistics. Typical examples of function g include the ℓ_1 norm (Lasso), the $\ell_1-\ell_2$ norm (group Lasso), the total variation, or the nuclear norm (the ℓ_1 norm of the singular values of $x\in\mathbb{R}^{N\times N}$ identified with a vector in $\mathbb{R}^n$ with $n=N^2$). These examples induce a sparsity property of the solution. For a recent account of these methods and an extended bibliography, see Liang-Fadili-Peyré [53].

(2) For optimal control of linear systems, $Ax-By=0$ is the state equation which links the state variable x to the control variable y. The functions f and g are respectively the criteria to be minimized and the cost of the control. Note that the convex constraints on x and y can be integrated respectively in the function f and g via their indicator functions. Recall that f and g are authorized to take extended real values. In this context, the variable z is often called the adjoint state. It plays a central role in the numerical solution of the problem, and in the study of its properties of sensitivity and stability. One can consult Allaire [2] for an in-depth presentation of the optimal control of distributed systems and PDE's in mechanics, and the use of the adjoint state.

(3) Game theory gives rise to the slightly more general form of the problem $(\mathcal{P})$

$$\min\{f(x)+g(y)+Q(x,y) : Ax-By=0\}$$

where Q is a coupling function between the respective actions x and y of the two players, and $Ax-By=0$ takes account of the limitation of resources. In this setting, the proximal (ADMM) algorithm has a natural interpretation as a best response dynamic with cost to change (to move), see Attouch-Redont-Soubeyran [24].

(4) Domain decomposition for PDE's. One of the main objectives of domain decomposition is to solve partial differential equations and the associated boundary value problems on complex geometries by partitioning the original domain Ω in smaller and simpler subdomains. Let us consider the case of two subdomains, *i.e.* $\Omega=\Omega_1\cup\Omega_2\cup\Sigma$ where Ω_1 and Ω_2 are two open sets which do not intersect, and Σ is the interface

[3] passing from the linear constraint to the affine constraint is a straightforward extension.

between the two subdomains. This leads to consider the problem $(\mathcal{P})$ where f and g are the respective energy functions on the two subdomains, and A and B are the respective trace operators relative to the interface Σ. The condition $Ax - By = 0$ reflects the continuity of the solution at the interface of the two subdomains. This approach was developed by Attouch-Briceno-Combettes in [7] to provide a full splitting primal-dual algorithm.

1.3. Presentation of the results

To guide our study, we use the close link between continuous evolution equations governed by maximally monotone operators and the proximal algorithms which result from their implicit temporal discretization, see Peypouquet-Sorin [65] for first order evolution systems. Indeed, continuous inertial systems are very useful in our context: they provide a deep mechanical intuition and suggest Lyapunov functions of energy type. More precisely, we rely on the second order evolution equations with inertial features and the associated relaxed proximal algorithms recently introduced by Attouch-Peypouquet [21] for the resolution of inclusions governed by maximally monotone operators. As a main feature, these algorithms involve an extrapolation term in the line of the Nesterov method, and an appropriate tuning of the proximal and relaxation parameters. Before introducing them, we need some preparatory results. In section 2 we review the classical results concerning the convergence properties of the proximal (ADMM) algorithm (without inertia). In section 3 we present the abstract results of [21] concerning the convergence properties of a relaxed inertial proximal algorithm for general monotone inclusions. Then in section 4 we specialize these results to the maximally monotone operator M_P which is attached to problem $(\mathcal{P})$, and present our main results concerning the convergence properties of a new class of inertial proximal (ADMM) algorithms. Section 5 concerns a variant of the inertial proximal algorithm which is completely splitted. The last section presents some new research directions, notably the introduction into these algorithms of correcting terms associated with the Hessian driven damping.

Our work is part of an active research trend. Each of the following recent articles follows a specific approach. In [33] Bot-Csetnek develop an inertial alternating direction method of multipliers which is based on the abstract inertial Douglas-Rachford algorithm which was previously developed in [34]. But the extrapolation coefficient is not allowed to go to one, as is the case in the accelerated gradient method of Nesterov. A closely related approach has been developed by Marques Alves-Eckstein-Geremia-De Melo [57] who consider an inexact variant of the Douglas-Rachford splitting method for maximally monotone operators, thus allowing to cover a large class of algorithms. In [49], Kim develops an accelerated proximal point method for maximally monotone operators based on the performance estimation problem (PEP) approach of Drori-Teboulle [45]. The convergence rate of the method is evaluated in terms of the fixed-point residual. But no convergence of the iterates is obtained. Another approach by Combettes-Glaudin [43] uses a new iterative scheme in which the update is obtained by applying a composition of quasi-nonexpansive operators to a point in the affine hull of the orbit generated up to the current iterate. In [41], Chen-Chan-Ma-Yang analyze the convergence properties of an inertial proximal (ADMM) algorithm where the proximal terms are calculated relative to general weighting ma-

trices which can be positive semidefinite, thus unifying many existing results. But the damping parameter is fixed, which does not take into account the Nesterov acceleration. In [48] Goldstein-O'Donoghue-Setzer-Baraniuk develop fast alternating direction optimization methods based on the Nesterov accelerated method, and obtain fast convergence rates using strong convexity assumptions and/or restarting methods. A similar approach was followed by Goldfarb-Ma-Scheinberg in [47].

In our approach, we will get rid of some of the limitations present in the above articles, by considering an inertial proximal algorithm (ADMM) which takes advantage of recent advances in the accelerated gradient method of Nesterov (see [10], [15], [20], [40], [72] and the references contained therein). In general Hilbert spaces, we get both the convergence of the iterates and fast convergence rates.

2. The (prox-ADMM) algorithm

Let us review some classical facts concerning the maximally monotone approach to the proximal (ADMM) algorithm. The operator $M_P\colon X\times Y\times Z\to 2^{X\times Y\times Z}$ which is defined by

$$M_P(x,y,z)=(\partial_{x,y}L,-\partial_z L)\,(x,y,z)=\big(\partial f(x)+A^tz,\ \partial g(y)-B^tz,\ By-Ax\big)$$

is maximally monotone on $X\times Y\times Z$. When the proximal algorithm is applied to the maximally monotone operator M_P, one obtains the so-called proximal method of multipliers. This method was initiated by Rockafellar [67, 68, 69] (1976). A comprehensive presentation on the subject can be found in Chen-Teboulle [42] (1994). This approach is described below: By applying the proximal algorithm to M_P with positive proximal parameter λ, we obtain the iteration $(x_k,y_k,z_k)\longrightarrow(x_{k+1},y_{k+1},z_{k+1})$ where $(x_{k+1},y_{k+1},z_{k+1})=(I+\lambda M_P)^{-1}(x_k,y_k,z_k)$ is the solution of the following system

$$\begin{cases}\frac{1}{\lambda}\,(x_{k+1}-x_k)+\partial f(x_{k+1})+A^t(z_{k+1})\ni 0;\\ \frac{1}{\lambda}\,(y_{k+1}-y_k)+\partial g(y_{k+1})-B^t(z_{k+1})\ni 0;\\ \frac{1}{\lambda}\,(z_{k+1}-z_k)+B(y_{k+1})-A(x_{k+1})=0.\end{cases}\tag{5}$$

Equivalently

$$\begin{cases}\frac{1}{\lambda}\,(x_{k+1}-x_k)+\partial f(x_{k+1})+A^t\big(z_k+\lambda(A(x_{k+1})-B(y_{k+1}))\big)\ni 0;\\ \frac{1}{\lambda}\,(y_{k+1}-y_k)+\partial g(y_{k+1})-B^t\big(z_k+\lambda(A(x_{k+1})-B(y_{k+1}))\big)\ni 0;\\ \frac{1}{\lambda}\,(z_{k+1}-z_k)+B(y_{k+1})-A(x_{k+1})=0.\end{cases}\tag{6}$$

The two first equations can be interpreted as the optimality conditions of the convex optimization problem

$$\begin{aligned}(x_{k+1},y_{k+1})=\operatorname{argmin}_{(\xi,\eta)\in X\times Y}\Big\{&f(\xi)+g(\eta)+\langle z_k,A\xi-B\eta\rangle\\&+\frac{\lambda}{2}\|A\xi-B\eta\|^2+\frac{1}{2\lambda}\|\xi-x_k\|^2+\frac{1}{2\lambda}\|\eta-y_k\|^2\Big\}.\end{aligned}\tag{7}$$

Thus, the proximal method of multipliers can be naturally interpreted with the help of the augmented Lagrangian function

$$L_\lambda(x,y,z) := f(x) + g(y) + \langle z, Ax - By\rangle + \tfrac{\lambda}{2}\|Ax - By\|^2$$

in the following way: at each iteration of the algorithm, given (x_k, y_k, z_k), one performs a proximal minimization step of the augmented Lagrangian L_λ with respect to (x, y) to obtain the next iterate (x_{k+1}, y_{k+1}). Then, one updates the multiplier by the iteration $z_{k+1} = z_k + \lambda(Ax_{k+1} - By_{k+1})$, which is nothing but a proximal maximization step of the augmented Lagrangian with respect to z. Note that the Lagrangian function is a convex-concave function. In this convex setting, the Lagrangian formulation is equivalent to the augmented Lagrangian formulation. As a consequence of the convergence properties of the proximal algorithm for general maximally monotone operators, this algorithm generates sequences that always (weakly) converges to a saddle point of L, and hence an optimal solution of $(\mathcal{P})$. One just needs to assume that the set of saddle points of L is non empty. The main disadvantage of this method is that, when performing the proximal minimization step to find (x_{k+1}, y_{k+1}), one is faced with the minimization problem (7) which is not separable, because of the presence of the quadratic coupling term $\|Ax - By\|^2$.

Indeed, by combining this method with the alternating proximal algorithms for weakly coupled minimization problems, see Attouch-Bolte-Redont-Soubeyran ([5], 2008), a fully split method is obtained. This approach, developed by Attouch-Soueycatt in [25], is described below.

Starting with an initial arbitrary triple $(x_0, y_0, z_0) \in X \times Y \times Z$, the sequence $(x_k, y_k, z_k) \in X \times Y \times Z$ is generated by the iterative scheme:

$$(x_k, y_k, z_k) \to (x_{k+1}, y_{k+1}, z_{k+1}), \quad k = 0, 1, 2, ...$$

$$\begin{cases} x_{k+1} = \operatorname{argmin}\left\{f(\xi) + \langle z_k, A\xi\rangle + \frac{\lambda}{2}\|A\xi - By_k\|_Z^2 + \frac{1}{2\lambda}\|\xi - x_k\|_X^2 : \xi \in X\right\} \\ y_{k+1} = \operatorname{argmin}\left\{g(\eta) - \langle z_k, B\eta\rangle + \frac{\lambda}{2}\|B\eta - Ax_{k+1}\|_Z^2 + \frac{1}{2\lambda}\|\eta - y_k\|_Y^2 : \eta \in Y\right\} \\ z_{k+1} = z_k + (Ax_{k+1} - By_{k+1}). \end{cases} \tag{8}$$

Because of the proximal quadratic terms, the two above convex minimization problems have unique respective solutions, x_{k+1} and y_{k+1}. The above algorithm can be seen as performing alternate proximal minimization (consecutive) steps on the augmented Lagrangian. It is called the "proximal Alternating Direction Method of Multipliers" (prox-ADMM) in short. Writing optimality conditions gives the equivalent form of the algorithm:

$$\text{(prox-ADMM)} \quad \begin{cases} \frac{1}{\lambda}(x_{k+1} - x_k) + \partial f(x_{k+1}) + A^t\left[z_k + \lambda(Ax_{k+1} - By_k)\right] \ni 0; \\ \frac{1}{\lambda}(y_{k+1} - y_k) + \partial g(y_{k+1}) + B^t\left[-z_k + \lambda(By_{k+1} - Ax_{k+1})\right] \ni 0; \\ z_{k+1} = z_k + (Ax_{k+1} - By_{k+1}). \end{cases} \tag{9}$$

2.1. Convergence properties of (prox-ADMM)

The following result has been obtained in [25, Theorem 2.1]. It extends the seminal convergence result obtained by Eckstein [46] for this algorithm.[4]

[4] The approach developed by Eckstein in [46] is different from ours. It relies on the application of the Douglas-Rachford algorithm to the dual formulation of $(\mathcal{P})$.

Recall that the set of equilibria $S = M_P^{-1}(0)$ is a closed convex subset of $X \times Y \times Z$.

Theorem 2.1. *Assume that the set S of equilibria is non empty. Let us start from an arbitrary point $(x_0, y_0, z_0) \in X \times Y \times Z$, and consider the corresponding sequence $(x_k, y_k, z_k) \in X \times Y \times Z$ generated by the "proximal alternating direction method of multipliers" algorithm* (prox-ADMM). *Then, the following properties are satisfied:*

(i) *(x_k, y_k, z_k) converges weakly in $X \times Y \times Z$ to an equilibrium $(x_\infty, y_\infty, z_\infty) \in S$ as $k \to +\infty$.*

(ii) *(x_k, y_k) is a minimizing sequence for problem $(\mathcal{P})$.*

(iii) *$Ax_k - By_k$ converges strongly to zero in Z as $k \to +\infty$.*

(iv) *$\|x_{k+1} - x_k\| \to 0$, $\|y_{k+1} - y_k\| \to 0$, $\|z_{k+1} - z_k\| \to 0$ as $k \to +\infty$.*

Let us now introduce some dynamical aspects of this algorithm.

2.2. Dynamical system attached to (prox-ADMM)

It is known for long that the proximal method is obtained by the implicit discretization of the evolution system governed by the maximally monotone operator. Thus, in our setting we are led to consider the evolution system in the product space $X \times Y \times Z$ governed by the maximally monotone operator $M_P \colon X \times Y \times Z \to 2^{X \times Y \times Z}$. We obtain the following system of first-order differential inclusions

$$\begin{cases} \dot{x}(t) + \partial f(x(t)) + A^t(z(t)) \ni 0; \\ \dot{y}(t) + \partial g(y(t)) - B^t(z(t)) \ni 0; \\ \dot{z}(t) + B(y(t)) - A(x(t)) = 0. \end{cases} \tag{10}$$

The general theory for semi-groups of contractions generated by maximally monotone operators applies to this system. Following Brezis [37], the Cauchy problem for (10) is well posed. Precisely, for any initial data $(x_0, y_0, z_0) \in \mathrm{dom} M_P$, there exists a unique strong solution $(x, y, z) : [0, +\infty[\to X \times Y \times Z$ of (10) that satisfies $x(0) = x_0, y(0) = y_0, z(0) = z_0$. The solution trajectories of (10) converge weakly in an ergodic sense to equilibria, which are the zeros of the operator M_P. The implicit temporal discretization of the above evolution equation with step size $\lambda > 0$ gives the (prox-ADMM) algorithm. Note that taking λ fixed makes an important difference between the continuous and the discrete dynamic. For the continuous dynamic there is only ergodic convergence, while for the discrete one (algorithm) there is convergence of the iterates. The close link between proximal algorithms and continuous dynamics generated by maximally monotone operators will serve us as a guideline for introducing the corresponding inertial systems and algorithms.[5] This is analyzed in the next section.

3. Inertial dynamics and algorithms for solving monotone inclusions

In this section, we consider the case of a general maximally monotone operator M. In this abstract setting, we will describe the inertial dynamics and the algorithms

[5] In the case of first order evolution equations, a detailed study can be found in Peypouquet-Sorin [65]

that support our approach. Then, in the next section, we will particularize these results to the operator M_P which is attached to the problem $(\mathcal{P})$.

3.1. Inertial dynamics for solving monotone inclusions

3.1.1. The cocoercive case

The starting point of our approach is the work of Álvarez-Attouch [3] and Attouch-Maingé [19] who studied the second-order evolution equation ($\dot{x}(\cdot)$ and $\ddot{x}(\cdot)$ stand respectively for the velocity and acceleration)

$$\ddot{x}(t) + \gamma \dot{x}(t) + M(x(t)) = 0, \tag{11}$$

where M is a maximally monotone operator which is supposed to be λ-cocoercive, for some positive parameter λ. The positive parameter γ is a (viscous) damping coefficient. Recall that $M\colon \mathcal{H} \to \mathcal{H}$ is λ-cocoercive ($\lambda > 0$) if it satisfies

$$\forall x, y \in \mathcal{H} \quad \langle My - Mx, y - x\rangle \geq \lambda \|My - Mx\|^2.$$

$M\colon \mathcal{H} \to \mathcal{H}$ λ-cocoercive implies that M is maximally monotone, and Lipschitz continuous with Lipschitz constant $\frac{1}{\lambda}$. Assuming that the cocoercivity parameter λ and the damping coefficient γ satisfy $\lambda\gamma^2 > 1$, it is shown in [19] that each trajectory of (11) converges weakly to an element of S.

Moreover, the condition $\lambda\gamma^2 > 1$ is sharp, as shown by the following example: Take $\mathbb{C}$ endowed with the standard real Hilbert structure $\langle u, v\rangle = Re(\bar{u}v)$. Consider the Heavy Ball with Friction equation

$$(\mathrm{HBF})_\gamma \qquad \ddot{z}(t) + \gamma \dot{z}(t) + M(z(t)) = 0, \quad t \geq 0, \tag{12}$$

where γ is a positive damping parameter, and $M\colon \mathbb{C} \to \mathbb{C}$ is given by

$$Mz := (w^2 - i\gamma w)z \ \text{ with } w > 0.$$

The operator M is λ-cocoercive with $\lambda = 1/(w^2+\gamma^2)$. A solution of $(\mathrm{HBF})_\gamma$ is given by the harmonic oscillator $z(t) = e^{iwt}$. We can observe that $z(\cdot)$ is bounded but not convergent for any $w > 0$. By letting $w \to 0^+$ we get $\lambda\gamma^2 \to 1^-$. Consequently, $\lambda\gamma^2 < 1$ is not a sufficient condition for the convergence of $(\mathrm{HBF})_\gamma$ for a general λ-cocoercive operator.

3.1.2. The general maximally monotone case

Let now suppose that M is a general maximally monotone operator acting on a real Hilbert space $\mathcal{H}$. The development of fast inertial methods to solve the inclusion governed by a general maximally monotone operator M allows us to consider in a unifying way different classes of problems. To cite some of the most important cases:

- $M = \partial\Phi$: convex minimization.
- $M = (\partial_x L, -\partial_y L)$: convex-concave saddle value problem, (augmented) Lagrangian methods.
- $M = I - T$: fixed point of nonexpansive operator.

Consider a maximally monotone operator M which is not supposed to be cocoercive (for example, a skew symmetric linear operator). To reduce ourselves to the cocoercive case, we use the Yosida approximation of M. Recall that, for each $\lambda > 0$, the resolvent of index λ of M, $J_{\lambda M}\colon \mathcal{H} \to \mathcal{H}$, is given by $J_{\lambda M} = (I + \lambda M)^{-1}$, where I is the identity operator. We will use indifferently the two notations $J_{\lambda M}$ and $(I + \lambda M)^{-1}$ in the case of a general maximally monotone operator M[6], and the proximal notation $\mathrm{prox}_{\lambda\Phi}$ in the case of the subdifferential of a convex function $M = \partial\Phi$. The resolvent is everywhere defined (that's Minty Theorem), and firmly nonexpansive. The Yosida approximation of index λ of M is the operator $M_\lambda\colon \mathcal{H} \to \mathcal{H}$ defined by

$$M_\lambda = \frac{1}{\lambda}\left(I - J_{\lambda M}\right).$$

The following properties of the Yosida approximation play a central role in our analysis (see [26], [37]):

(i) M_λ is $\frac{1}{\lambda}$-Lipschitz continuous.

(ii) $M_\lambda^{-1}(0) = M^{-1}(0)$ (preservation of the solution set).

(iii) M_λ is λ-cocoercive.

(iv) $(M_\lambda)_\mu = M_{\lambda+\mu}$ for all $\lambda, \mu > 0$ (resolvent equation).

(v) $M_\lambda(x) \in M(J_{\lambda M}(x))$ for all $x \in \mathcal{H}$, for all $\lambda > 0$.

Based on the convergence results in the cocoercive case [3], [19], and property (iii) of the Yosida approximation, we immediately deduce that, under the condition $\lambda\gamma^2 > 1$, each trajectory of

$$\ddot{x}(t) + \gamma\dot{x}(t) + M_\lambda(x(t)) = 0 \tag{13}$$

converges weakly to a zero of M. It turns out that taking a fixed damping coefficient γ induces too much friction, which prevents the inertial effect to be fully effective. In the quest for a faster convergence, we follow the dynamic interpretation given by Su-Boyd-Candès in [72] of the Nesterov acceleration method. This leads us to replace in (13) the fixed damping coefficient γ by the vanishing damping coefficient $\frac{\alpha}{t}$, where α is a positive parameter (Nesterov method corresponds to $\alpha = 3$). To preserve the condition $\lambda\gamma^2 > 1$ which links the damping and the cocoercive parameters, we are led to introduce a time-dependent regularization parameter $\lambda(\cdot)$ satisfying the condition

$$\lambda(t) \times \frac{\alpha^2}{t^2} > 1.$$

This leads us to introduce the continuous non-autonomous evolution equation

$$\text{(FIRST)} \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + M_{\lambda(t)}(x(t)) = 0, \quad t > t_0 > 0.$$

We call it the Fast Inertial Regularized SysTem, (FIRST) for short. One can show that the corresponding Cauchy problem is well posed (see [20, Appendix]). In accordance with the above approach, the following convergence result for the trajectories of (FIRST) has been obtained in [21].

[6] Indeed the notation $(I+\lambda M)^{-1}$ is sometimes more clear when there are several varying parameters in the formula

Theorem 3.1. *Let $M\colon \mathcal{H} \to 2^{\mathcal{H}}$ be a maximally monotone operator such that $S = M^{-1}(0) \neq \emptyset$. Let us suppose that the parameters entering the evolution equation* (FIRST) *satisfy the conditions $\alpha > 2$ and*

$$\lambda(t) = (1+\epsilon)\frac{t^2}{\alpha^2} \quad \text{for some} \quad \epsilon > \frac{2}{\alpha - 2}.$$

Then, for each solution trajectory $x\colon [t_0, +\infty[\to \mathcal{H}$ of (FIRST)*, we have that $x(t)$ converges weakly to an element of S, as $t \to +\infty$. Moreover $\lim_{t\to+\infty} \|\dot{x}(t)\| = \lim_{t\to+\infty} \|\ddot{x}(t)\| = 0$.*

Proof. We just sketch the beginning of the proof, which shows the role played by the tuning of the proximal parameter $\lambda(t)$. The proof is based on the Opial lemma A.2 (see the Appendix). Let us show that for all $z \in S = M^{-1}(0)$, the limit of the anchoring function $h_z(t) := \frac{1}{2}\|x(t) - z\|^2$ exists when $t \to +\infty$.
According to the classical derivation chain rule, we have

$$\dot{h}_z(t) = \langle x(t) - z, \dot{x}(t)\rangle,\ \ddot{h}_z(t) = \langle x(t) - z, \ddot{x}(t)\rangle + \|\dot{x}(t)\|^2.$$

Using the constitutive equation $(\text{FIRST})_{\alpha,\lambda}$, we deduce that

$$\ddot{h}_z(t) + \frac{\alpha}{t}\dot{h}_z(t) + \langle M_{\lambda(t)}(x(t)), x(t) - z\rangle = \|\dot{x}(t)\|^2.$$

Since $M_{\lambda(t)}$ is $\lambda(t)$-cocoercive and $z \in S$, we have

$$\langle M_{\lambda(t)}(x(t)), x(t) - z\rangle \geq \lambda(t)\|M_{\lambda(t)}(x(t))\|^2.$$

Combining this inequality with the above equation, we obtain

$$\ddot{h}_z(t) + \frac{\alpha}{t}\dot{h}_z(t) + \lambda(t)\|M_{\lambda(t)}(x(t))\|^2 \leq \|\dot{x}(t)\|^2. \tag{14}$$

According to $(\text{FIRST})_{\alpha,\lambda}$, we have $M_{\lambda(t)}(x(t)) = -\ddot{x}(t) - \frac{\alpha}{t}\dot{x}(t)$. By replacing $M_{\lambda(t)}(x(t))$ with this expression in (14), and after developing, we obtain

$$\ddot{h}_z(t) + \frac{\alpha}{t}\dot{h}_z(t) + \left(\lambda(t)\frac{\alpha^2}{t^2} - 1\right)\|\dot{x}(t)\|^2 + \alpha\frac{\lambda(t)}{t}\frac{d}{dt}\|\dot{x}(t)\|^2 + \lambda(t)\|\ddot{x}(t)\|^2 \leq 0.$$

By assumption, $\lambda(t) = (1+\epsilon)\frac{t^2}{\alpha^2}$. Therefore,

$$\ddot{h}_z(t) + \frac{\alpha}{t}\dot{h}_z(t) + \epsilon\|\dot{x}(t)\|^2 + \frac{\alpha\,\lambda(t)}{t}\frac{d}{dt}\|\dot{x}(t)\|^2 + \lambda(t)\|\ddot{x}(t)\|^2 \leq 0. \tag{15}$$

Integration of (15) gives

$$\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2\,dt < +\infty \ \text{ and } \ \int_{t_0}^{+\infty} (\dot{h_z})_+(t)\,dt < +\infty.$$

From this last estimate, we classically obtain that $\lim_{t\to+\infty} h_z(t)$ exists. Returning to (14) we obtain the estimate $\int_{t_0}^{+\infty} t\lambda(t)\|M_{\lambda(t)}(x(t))\|^2\,dt < +\infty$, which plays a key role in the continuation of the proof. □

In the subdifferential case $M = \partial\Phi$, the rate of convergence of the Nesterov accelerated method is achieved by the above dynamic, which justifies the "fast" terminology for (FIRST). More precisely, the following result has been obtained by Attouch-Cabot in [9]:

Theorem 3.2. *Suppose that $M = \partial\Phi$, where $\Phi\colon \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ is a convex lower semicontinuous proper function, and $\lambda(\cdot)$ is a nondecreasing function of class $\mathcal{C}^1$ such that $\lambda(t) \le Ct^2$ for some positive constant C. Then, for each solution trajectory $x\colon [t_0,+\infty[\to\mathcal{H}$ of* (FIRST)*, we have*

(i) *Case $\alpha \ge 3$:* $\Phi(p(t)) - \min_{\mathcal{H}}\Phi = \mathcal{O}(\frac{1}{t^2})$, *where* $p(t) = \operatorname{prox}_{\lambda(t)\Phi}x(t)$.

(ii) *Case $\alpha > 3$:* $x(t)$ *converges weakly to an element of S , and* $\lim_{t\to+\infty}\|x(t) - p(t)\| = 0$.

3.2. Inertial algorithms for solving monotone inclusions

Let us introduce the inertial proximal algorithms resulting from the temporal discretization of the continuous dynamic (FIRST). We choose to discretize it implicitly in order to follow closely the continuous-time trajectories. Moreover, the implicit scheme does not imply more complicated computation than the explicit one: they have the same iteration complexity (they each need a computation of resolvent per iteration). Taking a fixed time step $h > 0$, and setting $t_k = kh$, $x_k = x(t_k)$, $\lambda_k = \lambda(t_k)$, an implicit finite-difference scheme for (FIRST) with centered second-order variation gives

$$\frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{kh^2}(x_k - x_{k-1}) + M_{\lambda_k}(x_{k+1}) = 0. \tag{16}$$

After developing (16), we obtain

$$x_{k+1} + h^2 M_{\lambda_k}(x_{k+1}) = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}). \tag{17}$$

Setting $s = h^2$, we equivalently have

$$x_{k+1} = (I + sM_{\lambda_k})^{-1}\left(x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1})\right), \tag{18}$$

where $(I + sM_{\lambda_k})^{-1}$ is the resolvent of index $s > 0$ of the maximally monotone operator M_{λ_k}. This gives the algorithm

$$\begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= (I + sM_{\lambda_k})^{-1}(y_k). \end{cases} \tag{19}$$

As a key property, the resolvents of the Yosida approximation M_λ can simply be expressed in terms of the resolvents of M. Hence, using the resolvent equation

$(M_\lambda)_s = M_{\lambda+s}$, we obtain the two following equivalent formulations for $(I + sM_\lambda)^{-1}$:

$$(I + sM_\lambda)^{-1} = \frac{\lambda}{\lambda+s} I + \frac{s}{\lambda+s} (I + (\lambda+s)M)^{-1} \tag{20}$$

$$= I - sM_{\lambda+s}. \tag{21}$$

Using (20), we can reformulate (19) as follows

$$\text{(RIPA)} \qquad \begin{cases} y_k = x_k + \left(1 - \dfrac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} = \dfrac{\lambda_k}{\lambda_k + s} y_k + \dfrac{s}{\lambda_k + s} J_{(\lambda_k+s)M}(y_k), \end{cases}$$

where (RIPA) stands for the Regularized Inertial Proximal Algorithm. Convergence of (RIPA) algorithm has been established by Attouch-Peypouquet in [20, Theorem 3.4], see Attouch-Cabot [11] f or the extension to general extrapolation coefficients. We recall it below.

Theorem 3.3. *Let $M\colon \mathcal{H} \to 2^{\mathcal{H}}$ be a maximally monotone operator such that $S = M^{-1}(0) \neq \emptyset$. Let (x_k) be a sequence generated by the Regularized Inertial Proximal Algorithm* (RIPA) *where $\alpha > 2$ and*

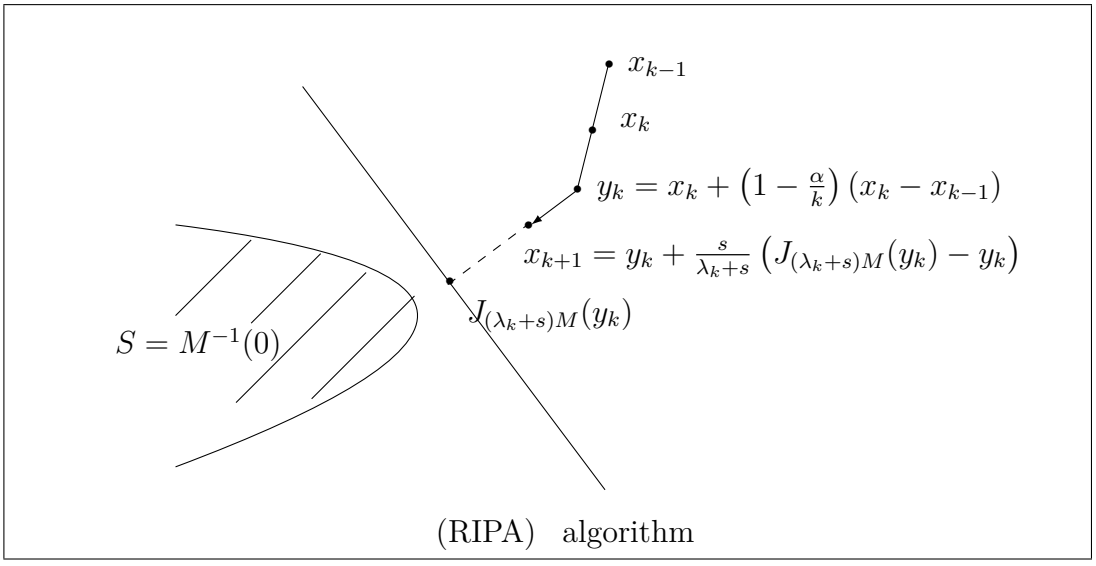
$$\lambda_k = (1+\epsilon)\frac{s}{\alpha^2}k^2 \quad \textit{for some } \epsilon > \frac{2}{\alpha - 2} \textit{ and all } k \geq 1.$$

Then,

(i) *The speed tends to zero.*
More precisely, $\|x_{k+1} - x_k\| = \mathcal{O}(\frac{1}{k})$ and $\sum_k k\|x_k - x_{k-1}\|^2 < +\infty$.

(ii) *The sequences (x_k) and (y_k) converge weakly to the same limit $\hat{x} \in S$ as $k \to +\infty$.*

This is illustrated in the following picture:



(RIPA) algorithm

As $k \to +\infty$, let us observe that $\lambda_k = (1+\epsilon)\frac{s}{\alpha^2}k^2 \to +\infty$, and $\frac{s}{\lambda_k+s} \to 0$. Therefore, $J_{(\lambda_k+s)M}(y_k) \sim \mathrm{proj}_S(y_k)$ which is an excellent direction. But we can only take a small step in this direction.

Let us make the link with the classical inertial proximal algorithm which corresponds to a discrete version of the heavy ball method.

Remark 3.4. Letting $\lambda_k \to 0$ in (RIPA) gives the classical form of the inertial proximal algorithm

$$\text{(Inertial-Prox)} \qquad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = J_{sM}(y_k). \end{cases}$$

The case $0 \leq \alpha_k \leq \bar{\alpha} < 1$ was considered by Álvarez-Attouch in [3], who proved that, under the summability assumption

$$\sum_k \alpha_k \|x_{k+1} - x_k\|^2 < +\infty, \tag{22}$$

then, for any sequence (x_k) generated by (Inertial-Prox), (x_k) converges weakly to some $\hat{x} \in S$, as $k \to +\infty$. The assumption (22) can be enforced by applying an appropriate on-line rule, for example

$$\alpha_k \in [0, \bar{\alpha}_k] \quad \text{with} \quad \bar{\alpha}_k = \min\left\{\bar{\alpha}, \tfrac{1}{k\|x_k - x_{k-1}\|^2}\right\}.$$

But the hypothesis $0 \leq \alpha_k \leq \bar{\alpha} < 1$ is rather restrictive. In the line of the Nesterov accelerated gradient method, the case $\alpha_k \to 1$ is the most interesting for obtaining fast methods, see the rich literature on the subject in the case of convex minimization [8], [10], [15], [20], [27], [40], [50, 51], [72]. Our approach, which relies on the Yosida approximation of the operator M, will allow us to get rid of this restrictive hypothesis. □

In the subdifferential case $M = \partial\Phi$, the rate of convergence of the Nesterov accelerated method (which is optimal for first order methods in the general convex case) is achieved by the above algorithm. Precisely, the following result has been obtained by Attouch-Cabot [13] and Attouch-Peypouquet [22].

Theorem 3.5. *Suppose that $M = \partial\Phi$, where $\Phi\colon \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a convex lower semicontinuous proper function with $S = \operatorname{argmin}\Phi \neq \emptyset$. Suppose that (λ_k) is a nondecreasing sequence, and $s > 0$ is a positive parameter.*
Let (x_k) be a sequence generated by the algorithm

$$\text{(RIPA)} \qquad \begin{cases} y_k = x_k + \left(1 - \dfrac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} = \dfrac{\lambda_k}{\lambda_k + s} y_k + \dfrac{s}{\lambda_k + s}\ \mathrm{prox}_{(\lambda_k+s)\Phi}(y_k). \end{cases}$$

Then, the following properties are satisfied:

- *Case $\alpha \geq 3$:* $\quad \Phi_{\lambda_k+s}(x_k) - \min_{\mathcal{H}}\Phi = \mathcal{O}(k^{-2})$.

As a consequence, setting $p_k = \operatorname{prox}_{(\lambda_k+s)\Phi}(x_k)$, *we have*

$$\Phi(p_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}(k^{-2}), \quad \text{and } \|x_k - p_k\|^2 = \mathcal{O}\left(\tfrac{\lambda_k}{k^2}\right).$$

- *Case* $\alpha > 3$*: Suppose moreover that* $\sup_k \frac{\lambda_k}{k^2} < +\infty$.

Then $x_k \rightharpoonup \hat{x} \in S$, $\Phi(p_k) - \min_{\mathcal{H}} \Phi = o(k^{-2})$, $\lim_{k\to+\infty} \|p_k - x_k\| = 0$.

3.3. Perturbation, errors, Tikhonov regularization

The following variant of the (RIPA) algorithm has been introduced in [21]. It involves additive errors (e_k):

$$\text{(RIPA-pert)} \qquad \begin{cases} y_k &= x_k + \left(1 - \dfrac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= \dfrac{\lambda_k}{\lambda_k + s}(y_k + se_k) + \dfrac{s}{\lambda_k + s} J_{(\lambda_k+s)M}(y_k + se_k). \end{cases}$$

The convergence of (RIPA-pert) algorithm is analyzed in the following theorem.

Theorem 3.6. *Let* $M\colon \mathcal{H} \to 2^{\mathcal{H}}$ *be a maximally monotone operator such that* $S = M^{-1}(0) \neq \emptyset$. *Let* (x_k) *be a sequence generated by the algorithm* (RIPA-pert)

where $\alpha > 2$ *and* $\lambda_k = (1 + \frac{s}{2} + \epsilon)\frac{2s}{\alpha^2}k^2$ *for some* $\epsilon > \frac{2+s}{\alpha-2}$ *and all* $k \geq 1$.

Suppose that $\sum_k k\|e_k\| < +\infty$ *and* $\sum_k k^3\|e_k\|^2 < +\infty$. *Then,*

(i) *The speed tends to zero.*
More precisely, $\|x_{k+1} - x_k\| = \mathcal{O}(\frac{1}{k})$ *and* $\sum_k k\|x_k - x_{k-1}\|^2 < +\infty$.

(ii) *The sequences* (x_k) *and* (y_k) *converge weakly to the same element* $\hat{x} \in S$, *as* $k \to +\infty$.

Remark 3.7. In connection with Theorem 3.6, there is the case where e_k comes from a Tikhonov regularization with vanishing coefficient ϵ_k. As a general rule, when the Tikhnonov coefficient ϵ_k does not tend to zero to quickly, we asymptotically obtain the solution of minimum norm. This result was proven in the case of the accelerated gradient method of Nesterov by Attouch-Chbani-Riahi in [16]. It is probable that such a phenomenon occurs within the framework of Theorem 3.6. This is an interesting subject to explore.

4. An inertial (ADMM) algorithm

Our program will consist in applying the convergence results obtained for the regularized inertial proximal algorithm (RIPA) to the specific operator M_P described in the sections 1 and 2. The next step will be to completely decompose the method, and thus obtain an inertial proximal ADMM algorithm. According to (5), the resolvent of the operator M_P, which is a basic ingredient of (RIPA), is given by the following formula: for any $(u, v, w) \in X \times Y \times Z$, $J_{\lambda M_P}(u, v, w) = (x, y, z)$ is the unique solution of the system

$$\begin{cases} \frac{1}{\lambda}(x-u) + \partial f(x) + A^t\left(w + \lambda(Ax - By)\right) \ni 0; \\ \frac{1}{\lambda}(y-v) + \partial g(y) - B^t\left(w + \lambda(Ax - By)\right) \ni 0; \\ \frac{1}{\lambda}(z-w) + By - Ax = 0. \end{cases} \tag{23}$$

In consequence, (RIPA) writes as follows: Starting with an initial arbitrary triple $(x_0, y_0, z_0) \in X \times Y \times Z$, the sequence $(x_k, y_k, z_k) \in X \times Y \times Z$ is generated by the iterative scheme (Proximal Inertial Method of Multiplier):

$$\text{(PIMM)} \qquad \begin{cases} u_k = x_k + \left(1 - \dfrac{\alpha}{k}\right)(x_k - x_{k-1}) \\ v_k = y_k + \left(1 - \dfrac{\alpha}{k}\right)(y_k - y_{k-1}) \\ w_k = z_k + \left(1 - \dfrac{\alpha}{k}\right)(z_k - z_{k-1}) \\ \dfrac{1}{\lambda_k + s}(p_k - u_k) + \partial f(p_k) + A^t\left(w_k + (\lambda_k + s)(Ap_k - Bq_k)\right) \ni 0 \\ \dfrac{1}{\lambda_k + s}(q_k - v_k) + \partial g(q_k) - B^t\left(w_k + (\lambda_k + s)(Ap_k - Bq_k)\right) \ni 0 \\ \dfrac{1}{\lambda_k + s}(r_k - w_k) + Bq_k - Ap_k = 0 \\ x_{k+1} = \dfrac{\lambda_k}{\lambda_k + s}u_k + \dfrac{s}{\lambda_k + s}p_k \\ y_{k+1} = \dfrac{\lambda_k}{\lambda_k + s}v_k + \dfrac{s}{\lambda_k + s}q_k \\ z_{k+1} = \dfrac{\lambda_k}{\lambda_k + s}w_k + \dfrac{s}{\lambda_k + s}r_k \end{cases}$$

A direct application of Theorem 3.3 gives the following convergence properties of (PIMM).

Theorem 4.1. *We assume the assumption (H) and that the set S of equilibria is non empty. Let's consider the Proximal Inertial Method of Multiplier* (PIMM) *where $\alpha > 2$ and*

$$\lambda_k = (1+\epsilon)\frac{s}{\alpha^2}k^2$$

for some $\epsilon > \frac{2}{\alpha-2}$ and all $k \geq 1$. Then, starting from an arbitrary point $(x_0, y_0, z_0) \in X \times Y \times Z$, the corresponding sequence $(x_k, y_k, z_k) \in X \times Y \times Z$ generated by (PIMM) *satisfies the following properties:*

(i) *The speed tends to zero.*
More precisely, $\|x_{k+1} - x_k\| = \mathcal{O}(\frac{1}{k})$ and $\sum_k k\|x_k - x_{k-1}\|^2 < +\infty$.

(ii) *The sequences (x_k) and (y_k) converge weakly to the same element $\hat{x} \in S$, as $k \to +\infty$.*

Yet, as for the proximal method of multipliers, (PIMM) is not completely decomposed. This appears clearly when writing the equivalent variational formulation

$$\begin{cases} u_k = x_k + \left(1 - \dfrac{\alpha}{k}\right)(x_k - x_{k-1}) \\ v_k = y_k + \left(1 - \dfrac{\alpha}{k}\right)(y_k - y_{k-1}) \\ w_k = z_k + \left(1 - \dfrac{\alpha}{k}\right)(z_k - z_{k-1}) \\ (p_k, q_k) = \underset{(\xi,\eta)\in X\times Y}{\operatorname{argmin}} \Big\{ f(\xi) + g(\eta) + \langle w_k, A\xi - B\eta\rangle + \dfrac{\lambda_k + s}{2}\|A\xi - B\eta\|^2 \\ \qquad\qquad + \dfrac{1}{2(\lambda_k + s)}(\|\xi - u_k\|^2 + \|\eta - v_k\|^2) \Big\} \\ r_k = w_k + (\lambda_k + s)(Ap_k - Bq_k) \\ x_{k+1} = \dfrac{\lambda_k}{\lambda_k + s}u_k + \dfrac{s}{\lambda_k + s}p_k. \\ y_{k+1} = \dfrac{\lambda_k}{\lambda_k + s}v_k + \dfrac{s}{\lambda_k + s}q_k. \\ z_{k+1} = \dfrac{\lambda_k}{\lambda_k + s}w_k + \dfrac{s}{\lambda_k + s}r_k. \end{cases}$$

Remark 4.2. Take into account the fact that the operator M_P is associated with an optimization problem, it would be interesting to study the convergence rate of the values $f(x_k) + g(y_k) - \inf \mathcal{P}$ as $k \to +\infty$.

Remark 4.3. In the general approach developed in [11] and [20], it is considered the case where the operator M satisfies a a quadratic growth property (it contains the strongly monotone case). Adapting this result to (PIMM) is also an interesting question.

5. A full splitting algorithm

Let us follow the strategy which has been developed in [25] in order to completely decompose the problem. It consists in applying one step of the alternating proximal minimization algorithms for weakly coupled minimization problems, see [5], [24]. Other strategies can be developed, for example based on the forward-backward splitting method, see [7].

We obtain the following algorithm called the inertial proximal ADMM algorithm, (ip-ADMM) for short.

$$
\text{(ip-ADMM)}\quad \left\{\begin{array}{rcl}
u_k &=& x_k + \left(1-\dfrac{\alpha}{k}\right)(x_k - x_{k-1}) \\
v_k &=& y_k + \left(1-\dfrac{\alpha}{k}\right)(y_k - y_{k-1}) \\
w_k &=& z_k + \left(1-\dfrac{\alpha}{k}\right)(z_k - z_{k-1}) \\
p_k &=& \operatorname{argmin}_{\xi\in X}\{f(\xi) + \langle w_k, A\xi - Bv_k\rangle + \dfrac{\lambda_k+s}{2}\|A\xi - Bv_k\|^2 \\
&& \qquad\qquad + \dfrac{1}{2(\lambda_k+s)}\|\xi - u_k\|^2\} \\
q_k &=& \operatorname{argmin}_{\eta\in Y}\{g(\eta) + \langle w_k, Ap_k - B\eta\rangle + \dfrac{\lambda_k+s}{2}\|Ap_k - B\eta\|^2 \\
&& \qquad\qquad + \dfrac{1}{2(\lambda_k+s)}\|\eta - v_k\|^2\} \\
r_k &=& w_k + (\lambda_k+s)(Ap_k - Bq_k) \\
x_{k+1} &=& \dfrac{\lambda_k}{\lambda_k+s}u_k + \dfrac{s}{\lambda_k+s}p_k. \\
y_{k+1} &=& \dfrac{\lambda_k}{\lambda_k+s}v_k + \dfrac{s}{\lambda_k+s}q_k. \\
z_{k+1} &=& \dfrac{\lambda_k}{\lambda_k+s}w_k + \dfrac{s}{\lambda_k+s}r_k.
\end{array}\right.
$$

Equivalently

$$
\text{(ip-ADMM)}\quad \left\{\begin{array}{l}
u_k = x_k + \left(1-\dfrac{\alpha}{k}\right)(x_k - x_{k-1}) \\
v_k = y_k + \left(1-\dfrac{\alpha}{k}\right)(y_k - y_{k-1}) \\
w_k = z_k + \left(1-\dfrac{\alpha}{k}\right)(z_k - z_{k-1}) \\
\dfrac{1}{\lambda_k+s}(p_k-u_k) + \partial f(p_k) + A^t\left(w_k + (\lambda_k+s)(Ap_k - Bv_k)\right) \ni 0 \\
\dfrac{1}{\lambda_k+s}(q_k-v_k) + \partial g(q_k) - B^t\left(w_k + (\lambda_k+s)(Ap_k - Bq_k)\right) \ni 0 \\
r_k = w_k + (\lambda_k+s)(Ap_k - Bq_k) \\
x_{k+1} = \dfrac{\lambda_k}{\lambda_k+s}u_k + \dfrac{s}{\lambda_k+s}p_k. \\
y_{k+1} = \dfrac{\lambda_k}{\lambda_k+s}v_k + \dfrac{s}{\lambda_k+s}q_k. \\
z_{k+1} = \dfrac{\lambda_k}{\lambda_k+s}w_k + \dfrac{s}{\lambda_k+s}r_k.
\end{array}\right.
$$

The crucial point is to show that taking one step of this inner loop instead of performing the whole loop induces an error which does not affect the convergence process. It is a difficult and open question.

6. Conclusion, perspectives

Obtaining fast convergent alternating direction methods of multipliers is an active research subject due to its numerous applications, either as a numerical method or for modeling purposes in decision sciences. In this article, we propose an algorithm that involves both inertial and relaxation aspects. We have proven that it generates convergent iterates with fast convergence properties. By comparison with other related approaches, it it is based on the recent improved versions of the accelerated gradient method of Nesterov, and is therefore optimal in the case without constraint. It opens the door to new directions of research. We have listed a few below.

1. Besides the many questions that have been raised throughout the paper, the main problem that remains to be solved is to show that the completely splitted algorithm (ip-ADMM) inherits the convergence properties of the Proximal Inertial Method of Multiplier (PIMM).

2. According to the dynamical interpretation of the inertial optimization algorithms, their convergence properties come from the damping term. Much progress has been made recently to explore this aspect, and choose a damping term with favorable properties. In this article, we used the asymptotic vanishing viscous damping with the coefficient $\frac{\alpha}{t}$ which is naturally attached to the acceleration of Nesterov, see Su-Boyd-Candés [72]. Following Attouch-Chbani-Fadili-Riahi [14], it would be interesting to combine this damping with the Hessian damping which takes account of the geometry of the functions which enter the constrained minimization problem $(\mathcal{P})$. In the case without linear constraint, there is theoretical and numerical evidence that the introduction of the corresponding correcting terms notably improves the convergence properties of the algorithms, especially in the case of poorly conditioned problems. In this direction, in the case of general monotone inclusions, see the recent contribution of Kim [49].
   Other types of damping are also of great numerical interest, such as the dry friction combined with the Hessian damping considered by Adly-Attouch [1]: in this case, we can expect to obtain a geometric convergence rate, and that, generically, there is finite convergence of the sequences generated by the algorithms to approximate equilibria.

3. A classical alternative approach to the Lagrangian method is the penalization method. A major advantage of the penalization methods is that they can handle nonlinear problems. Indeed, when we consider the penalized formulation of problem $(\mathcal{P})$

$$(\mathcal{P}_r) \qquad \min_{x\in X,\, y\in Y} \left\{ f(x) + g(y) + r\|Ax - By\|^2 \right\},$$

   it has been proven by Attouch-Bolte-Redont-Soubeyran [6] that, when f and g are semialgebraic functions, then the proximal alternating minimization algorithm (applied to $(\mathcal{P}_r)$) generates sequences which converge towards equilibria. Moreover, by a diagonal argument, we can combine this type of algorithm with the penalization method obtained by letting $r \to +\infty$, so as to solve $(\mathcal{P})$, see Attouch-Czarnecki-Peypouquet [18]. It is therefore natural to conjecture that similar convergence properties hold for the Lagrangian approach when f and

g are nonconvex tame functions, see Magnusson-Weeraddana-Rabbat [55] for some first results in this direction.

4. As already mentioned, various approaches have recently been proposed to accelerate the (ADMM) algorithm. It would be interesting to compare them from a theoretical and numerical point of view, as well as with the inertial primal-dual methods.
5. In Theorem 3.6, the abstract convergence theorem which supports our analysis was considered with the presence of perturbations, or errors. It is natural to consider the corresponding results for the associated inertial proximal (ADMM) algorithms. Besides taking into account noise and errors, it is a central question to obtain completely splitted algorithms.

A. Auxiliary results

A.1. Yosida regularization of an operator M

Given a maximally monotone operator M acting on a Hilbert space $\mathcal{H}$, and given λ a positive parameter, the resolvent of M with index λ and the Yosida regularization of M with parameter λ are defined by

$$J_{\lambda M} = (I + \lambda M)^{-1} \qquad \text{and} \qquad M_\lambda = \frac{1}{\lambda}\left(I - J_{\lambda M}\right),$$

respectively. The operator $J_{\lambda M}\colon \mathcal{H} \to \mathcal{H}$ is nonexpansive and everywhere defined (indeed it is firmly non-expansive). Moreover, M_λ is λ-cocoercive: for all $x, y \in \mathcal{H}$ we have

$$\langle M_\lambda y - M_\lambda x, y - x\rangle \geq \lambda \|M_\lambda y - M_\lambda x\|^2.$$

This property immediately implies that $M_\lambda\colon \mathcal{H} \to \mathcal{H}$ is $\frac{1}{\lambda}$-Lipschitz continuous. Also note that for any $x \in \mathcal{H}$, and any $\lambda > 0$

$$M_\lambda(x) \in M(J_{\lambda M}(x)) = M(x - \lambda M_\lambda(x)).$$

Moreover, for any $\lambda > 0$, M and M_λ have the same solution set $S := M_\lambda^{-1}(0) = M^{-1}(0)$.

Another property that proves useful is the resolvent equation (see, for example, [37, Proposition 2.6] or [26, Proposition 23.6]): for any $\lambda, \mu > 0$

$$(M_\lambda)_\mu = M_{(\lambda+\mu)}. \tag{24}$$

This property allows to compute simply the resolvent of M_λ: for any $\lambda, \mu > 0$ we have

$$J_{\mu M_\lambda} = \frac{\lambda}{\lambda+\mu} I + \frac{\mu}{\lambda+\mu} J_{(\lambda+\mu)M}.$$

As a consequence of the resolvent equation we have the following continuity property of the resolvents with respect to the proximal parameter. This property plays a key role in the proof of the convergence of the iterates in Theorem 3.1.

Lemma A.1. *Let $\gamma, \delta > 0$, and $x, y \in \mathcal{H}$. Then, for each $z \in S = M^{-1}(0)$, we have*

$$\|\gamma M_\gamma x - \delta M_\delta y\| \leq 2\|x - y\| + 2\|x - z\| \frac{|\gamma - \delta|}{\gamma}. \tag{25}$$

For a detailed presentation of the properties of the maximally monotone operators and the Yosida approximation, the reader can consult [26] or [37].

A.2. Opial's Lemma

Lemma A.2. (Opial) *Let S be a nonempty subset of $\mathcal{H}$ and let $x\colon [t_0, +\infty[\to \mathcal{H}$. Assume that*

(i) *for every $z \in S$, $\lim_{t\to\infty} \|x(t) - z\|$ exists;*

(ii) *every weak sequential limit point of $x(t)$, as $t \to \infty$, belongs to S.*

Then $x(t)$ converges weakly as $t \to \infty$ to a point in S.

References

[1] S. Adly, H. Attouch: *Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping* (2019) hal-02423584.

[2] G. Allaire: *Optimal Design of Structures; optimization of distributed systems: Computing a gradient by the adjoint method (MAP 562)*, Dept. of Applied Mathematics, Ecole Polytechnique, Paris (2015).

[3] F. Álvarez, H. Attouch: *An inertial proximal method for maximally monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Analysis 9(1-2) (2001) 3–11.

[4] F. Álvarez, H. Attouch, J. Bolte, P. Redont: *A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics*, J. Math. Pures Appl. 81(8) (2002) 747–779.

[5] H. Attouch, J. Bolte, P. Redont, A. Soubeyran: *Alternating proximal algorithms for weakly coupled convex minimization problems. Applications to dynamical games and PDE's*, J. Convex Analysis 15(3) (2008) 485–506.

[6] H. Attouch, J. Bolte, P. Redont, A. Soubeyran: *Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Lojasiewicz inequality*, Math. Oper. Research 35(2) (2010) 438–457.

[7] H. Attouch, L. M. Briceño-Arias, P. L. Combettes: *A strongly convergent primal-dual method for nonoverlapping domain decomposition*, Numerische Mathematik 133(3) (2016) 443–470.

[8] H. Attouch, A. Cabot: *Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity*, J. Differential Equations 263(9) (2017) 5412–5458.

[9] H. Attouch, A. Cabot: *Convergence of damped inertial dynamics governed by regularized maximally monotone operators*, J. Differential Equations 264(12) (2018) 7138–7182.

[10] H. Attouch, A. Cabot: *Convergence rates of inertial forward-backward algorithms*, SIAM J. Optimization 28(1) (2018) 849–874.

[11] H. Attouch, A. Cabot: *Convergence of a relaxed inertial proximal algorithm for maximally monotone operators*, Math. Programming, published online 2019, https://doi.org/10.1007/s10107-019-01412-0.

[12] H. Attouch, A. Cabot: *Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions*, Appl. Math. Optimization 80(3) (2019) 547–598.

[13] H. Attouch, A. Cabot: *Convergence rate of a relaxed inertial proximal algorithm for convex minimization*, Optimization, published online 2019, https://doi.org/10.1080/02331934.2019.1696337.

[14] H. Attouch, Z. Chbani, J. Fadili, H. Riahi: *First-order algorithms via inertial systems with Hessian driven damping*, (2019) hal-02193846.

[15] H. Attouch, Z. Chbani, J. Peypouquet, P. Redont: *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing damping*, Math. Programming, Ser. B 168 (2018) 123–175.

[16] H. Attouch, Z. Chbani, H. Riahi: *Combining fast inertial dynamics for convex optimization with Tikhonov regularization*, J. Math. Anal. Appl. 457 (2018) 1065–1094.

[17] H. Attouch, Z. Chbani, H. Riahi: *Convergence rate of inertial proximal algorithms with general extrapolation and proximal coefficients*, Vietnam J. Math., published online 2020, https://doi.org/10.1007/s10013-020-00399-y.

[18] H. Attouch, M.-O. Czarnecki, J. Peypouquet: *Coupling forward-backward with penalty schemes and parallel splitting for constrained variational inequalities*, SIAM J. Optimization 21(4) (2011) 1251–1274.

[19] H. Attouch, P. E. Maingé: *Asymptotic behavior of second order dissipative evolution equations combining potential with non-potential effects*, ESAIM Control Optimization Calc. Var. 17(3) (2011) 836–857.

[20] H. Attouch, J. Peypouquet: *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $\frac{1}{k^2}$, SIAM J. Optimization 26(3) (2016) 1824–1834.

[21] H. Attouch, J. Peypouquet: *Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators*, Math. Programming 174(1-2) (2019) 391–432.

[22] H. Attouch, J. Peypouquet: *Convergence rate of proximal inertial algorithms associated with Moreau envelopes of convex functions*, in: *Splitting Algorithms, Modern Operator Theory, and Applications*, H. Bauschke, R. Burachik, D. Luke (eds.), Springer, Cham (2019) 1–44.

[23] H. Attouch, J. Peypouquet, P. Redont: *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations 261(10) (2016) 5734–5783.

[24] H. Attouch, P. Redont, A. Soubeyran: *A new class of alternating proximal minimization algorithms with costs-to-move*, SIAM J. Optimization 18(3) (2007) 1061–1081.

[25] H. Attouch, M. Soueycatt: *Augmented Lagrangian and proximal alternating direction methods of multipliers in Hilbert spaces. Applications to games, PDE's and control*, Pacific J. Optimization 5(1) (2009) 17–37.

[26] H. Bauschke, P. L. Combettes: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, Berlin (2011).

[27] A. Beck, M. Teboulle: *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Science 2(1) (2009) 183–202.

[28] P. Bégout, J. Bolte, M. A. Jendoubi: *On damped second-order gradient systems*, J. Differential Equations 259 (2015) 3115–3143.

[29] R. I. Bot, E. R. Csetnek: *Approaching monotone inclusion problems via second order dynamical systems with linear and anisotropic damping*, in: *New Trends in Differential Equations, Control Theory and Optimization*, V. Barbu, C. Lefter, I. I. Vrabie (eds.), Proc. 8th Congress of Romanian Mathematicians, World Scientific, Singapore (2016) 53–72.

[30] R. I. Bot, E. R. Csetnek: *An inertial Tseng's type proximal algorithm for nonsmooth and nonconvex optimization problems*, J. Optimization Theory Appl. 171(2) (2016) 600–616.

[31] R. I. Bot, E. R. Csetnek: *Second order forward-backward dynamical systems for monotone inclusion problems*, SIAM J. Control Optimization 54(3) (2016) 1423–1443.

[32] R. I. Bot, E. R. Csetnek: *An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems*, Numerical Algorithms 71(3) (2016) 519–540.

[33] R. I. Bot, E. R. Csetnek: *An inertial alternating direction method of multipliers*, Minimax Theory Appl. 1(1) (2016) 29–49.

[34] R. I. Bot, E. R. Csetnek, C. Hendrich: *Inertial Douglas-Rachford splitting for monotone inclusion problems*, Appl. Math. Computation 256(1) (2015) 472–487.

[35] R. I. Bot, E. R. Csetnek, S. C. László: *Tikhonov regularization of a second order dynamical system with Hessian driven damping*, Math. Programming (2020).

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein: *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Machine Learning 3 (2010) 1–12.

[37] H. Brézis: *Opérateurs Maximaux Monotones dans les Espaces de Hilbert et Équations d'Évolution*, Lecture Notes 5, North Holland, Amsterdam (1972).

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

[39] C. Castera, J. Bolte, C. Févotte, E. Pauwels: *An inertial Newton algorithm for deep learning*, (2019) hal-02140748.

[40] A. Chambolle, Ch. Dossal: *On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm*, J. Optimization Theory Appl. 166 (2015) 968–982.

[41] C. Chen, R. H. Chan, S. Ma, J. Yang: *Inertial proximal ADMM for linearly constrained separable convex optimization*, SIAM J. Imaging Sciences 8(4) (2015) 2239–2267.

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

[43] P. L. Combettes, L. Glaudin: *Quasi-nonexpansive iterations on the affine hull of orbits: from Mann's mean value algorithm to inertial methods*, SIAM J. Optimization 27(4) (2017) 2356–2380.

[44] Q. L. Dong, J. Huang, X. H. Li, Y. J. Cho, Th. M. Rassias: *MiKM: Multi–step inertial Krasnosel'skiĭ–Mann algorithm and its applications*, J. Global Optimization 73(4) (2019) 801–824.

[45] Y. Drori, M. Teboulle: *Performance of first-order methods for smooth convex minimization: A novel approach*, Math. Programming 145(1-2) (2014) 451–482.

[46] J. Eckstein: *Some saddle-function splitting methods for convex programming*, Optimization Methods and Software 4 (1994) 75–83.

[47] D. Goldfarb, S. Ma, K. Scheinberg: *Fast alternating linearization methods for minimizing the sum of two convex functions*, Math. Programming 141 (2013) 349–382.

[48] T. Goldstein, B. O'Donoghue, S. Setzer, R. Baraniuk: *Fast alternating direction optimization methods*, SIAM J. Imaging Sciences 7(3) (2014) 1588–1623.

[49] D. Kim: *Accelerated proximal point method for maximally monotone operators*, arXiv: 1905.05149v3 (2020).

[50] D. Kim, J. A. Fessler: *Optimized first-order methods for smooth convex minimization*, Math. Programming 159(1) (2016) 81–107.

[51] D. Kim, J. A. Fessler: *Another look at the Fast Iterative Shrinkage/Thresholding Algorithm (FISTA)*, SIAM J. Optimization 28(1) (2018) 223–250.

[52] R. Laraki, P. Mertikopoulos: *Inertial games dynamics and applications to constrained optimization*, SIAM J. Control Optimization 53(5) (2015) 3141–3170.

[53] J. Liang, J. Fadili, G. Peyré: *Local linear convergence of forward-backward under partial smoothness*, in: NIPS'14, Proc. 27th Int. Conf. on Neural Information Processing Systems – Volume 2, ACM Digital Library (2014) 1970–1978.

[54] D. A. Lorenz, T. Pock: *An inertial forward-backward algorithm for monotone inclusions*, J. Math. Imaging Vision 51 (2015) 311–325.

[55] S. Magnusson, P. C. Weeraddana, M. G. Rabbat, C. Fischione: *On the convergence of alternating direction Lagrangian methods for nonconvex structured optimization problems*, IEEE Trans. Control Network Systems 3(3) (2016) 296–309.

[56] P.-E. Maingé, A. Moudafi: *A proximal method for maximal monotone operators via discretization of a first-order dissipative dynamical system*, J. Convex Analysis 14(4) (2007) 869–878.

[57] M. Marques Alves, J. Eckstein, M. Geremia, J. G. Melo: *Relative-error inertial-relaxed inexact versions of Douglas-Rachford and ADMM splitting algorithms*, Comp. Optimization Appl. 75(2) (2020) 389–422.

[58] A. Moudafi, M. Oliny: *Convergence of a splitting inertial proximal method for monotone operators*, J. Comput. Appl. Math. 155(2) (2003) 447–454.

[59] Y. Nesterov: *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Math. Doklady 27 (1983) 372–376.

[60] Y. Nesterov: *Introductory Lectures on Convex Optimization: A Basic Course*, Applied Optimization 87, Kluwer Academic Publishers, Boston (2004).

[61] Y. Nesterov: *Gradient methods for minimizing composite objective function*, CORE Discussion Paper 2007/76, Catholic University of Louvain (2007).

[62] P. Ochs, Y. Chen, T. Brox, T. Pock: *iPiano: Inertial proximal algorithm for nonconvex optimization*, SIAM J. Imaging Sciences 7(2) (2014) 1388–1419.

[63] Z. Opial: *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc. 73 (1967) 591–597.

[64] N. Parikh, S. Boyd: *Proximal algorithms*, Foundations and Trends in Optimization 1(3) (2013) 123–231.

[65] J. Peypouquet, S. Sorin: *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*, J. Convex Analysis 17(3-4) (2010) 1113–1163.

[66] B. T. Polyak: *Introduction to Optimization*, Optimization Software, New York (1987).

[67] R. T. Rockafellar: *Monotone operators and the proximal point algorithm*, SIAM J. Control Optimization 14(5) (1976) 877–898.

[68] R. T. Rockafellar: *Monotone operators associated with saddle-functions and minimax problems*, in: *Nonlinear Functional Analysis, Vol. 1*, 18th Proceedings of Symposia in Pure Mathematics, F. E. Browder (ed.), American Mathematical Society, Providence (1976) 241–250.

[69] R. T. Rockafellar: *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Operations Research 1(2) (1976) 97–116.

[70] B. Shi, S. S. Du, M. I. Jordan, W. J. Su: *Understanding the acceleration phenomenon via high-resolution differential equations*, arXiv: 1810.08907 (2018).

[71] M. V. Solodov, B. F. Svaiter: *A unified framework for some inexact proximal point algorithms*, Numer. Funct. Anal. Optimization 22(7-8) (2001) 1013–1035.

[72] W. Su, S. Boyd, E. J. Candès: *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*, Neural Inform. Processing Systems 27 (2014) 2510–2518.