On the convergence of warped proximal iterations for solving nonmonotone inclusions and applications

Dimitri Papadimitriou PAPADIMITRIOU.DIMITRI.BE@GMAIL.COM
Bang Cong Vu BANGCVVN@GMAIL.COM
MCO Institute (3nLab) & Belgium Research Center, Leuven, Belgium

Abstract

In machine learning, tackling fairness, robustness, and safeness requires to solve nonconvex optimization problems with various constraints. In this paper, we investigate the warped proximal iterations for solving the nonmonotone inclusions and its application to nonconvex QP with equality constraints.

1. Introduction

In general, a monotone inclusion problem in a real Hilbert space $\mathcal{H}$ refers to finding a zero of a set-valued maximally monotone operator A, i.e., the zero set $\mathrm{zer}(A)$ [1]. The term "warped proximal iteration" was recently introduced in [3] as generalization of the proximal point algorithm for finding a zero point of a maximally monotone operator A acting on $\mathcal{H}$. The maximal monotonicity of A restricted its applicability to the class of convex optimization problems as well as operator splitting methods for composite monotone inclusions [7] [2] [6] [12] [3] [9]. On the other hand, solving a nonmonotone inclusion, i.e., an inclusion where the operator A is nonmonotone, is generally open and challenging [4]. Consequently, over the last two decades, only few papers were published where the notion of ρ-(co)hypomonotonicity has been used [11] [5] to guarantee the convergence of the generated sequence. More precisely, the proximal point algorithm converges weakly to a solution to the following problem:

$$\text{Find } x \in \mathcal{H} \text{ such that } 0 \in Ax, \text{ where } A^{-1} \text{ is } \rho\text{-hypomonotone}, \rho > 0. \tag{1.1}$$

Its motivation stems from the observation that compared to the case where the operator A is (maximally) monotone, the solving of this problem becomes considerably more complicated when the operator A fails to be monotone. From this perspective, the first objective of this work is to extend the notion of ρ- to $(\rho V, W)$-hypomotonicity and investigate the weak convergence property of the warped proximal iteration as well as its various applications to (constrained) nonconvex optimization problems. In particular, we place our attention to finding a KKT point of a class of quadratic programming problems with equality constraints.

In general, the finding of $\mathrm{zer}(A)$ when A is nonmonotone is by nature more complex [4] but fundamental from the optimization perspective when convexity assumptions are relaxed. Indeed, the application of Augmented Lagrangian methods (ALM) for the minimization of nonconvex functions can be defined as a particular instance of the proximal point method for finding zeroes of nonmonotone operators. Recent developments of proximal ALM methods when applied to convex problems have been shown to be equivalent to resolvent iterations on the monotone operator encoding the KKT optimality conditions.

From the machine learning perspective, various requirements are often imposed using penalties, i.e., by integrating constraint violation costs in the empirical risk minimization (ERM) objective. While it is straightforward to find penalty terms leading to optimal solutions when the objective of the ERM problem is convex, training of neural networks leads to nonconvex optimization problems. Enforcing their behavior to tackle fairness, robustness, and safety requires to solve instead constrained nonconvex optimization problems. Incorporating constraints into the model enables to account for prior knowledge as well as smoothness or sparsity.

2. Background, Notation and Preliminary Results

Let $\mathcal{H}$ be a real Hilbert space. Its inner product and induced norm are denoted $\langle \cdot \mid \cdot \rangle$ and $\|\cdot\|$, respectively. The power set of $\mathcal{H}$ is denoted $2^{\mathcal{H}}$. $\mathcal{B}(\mathcal{H})$ refers to the algebra of all bounded linear operators acting on $\mathcal{H}$. The graph of an operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ is defined by $\operatorname{gra}(A) = \{(x, u) \in \mathcal{H} \times \mathcal{H} \mid u \in Ax\}$ where Ax denotes that A operates on x, i.e. $Ax = A(x)$. The range of an operator A writes as $\operatorname{ran}(A)$. The set of zero points of the operator A is defined by $\operatorname{zer}(A) = \{x \in \operatorname{dom}(A) | Ax \ni 0\}$. Id denotes the Identity operator. An operator A is monotone if $\langle x - y \mid u - v\rangle \geq 0\ \forall (x, u) \in \operatorname{gra}(A)$ and $(y, v) \in \operatorname{gra}(A)$. Further, it is maximal(ly) monotone if no extension of $\operatorname{gra}(A)$ exists that preserves monotonicity. The operator A is said strongly monotone with constant $\theta \in \left]0, +\infty\right[$ if $A - \theta\,\mathrm{Id}$ is monotone. More specific definitions including (maximally) hypermonotone operator, $\rho-$ and $(\rho V, W)-$hypomonotone operator, nonexpansive operator as well as the $W-$resolvent of an operator that are used in the context of this paper are detailed in Appendix A.

The following Lemma plays a central role in the demonstration of the main Theorem of this paper as it provides a necessary condition on the iterates.

Lemma 1 *Let $W : \mathcal{H} \to \mathcal{H}$ be a single-valued (nonlinear) operator and $V : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator with $\operatorname{dom}(V) = \operatorname{dom}(W) = \mathcal{H}$. Let $A \colon \mathcal{H} \to 2^{\mathcal{H}}$ be such that A^{-1} is $(\rho V, W)$-hypomonotone with ρ in $\left]0, +\infty\right[$. Then, the following hold.*

(i) *If $x \in \operatorname{dom}(J^W_{\gamma A})$ and $\gamma \in \left]\rho\|V\|^2, +\infty\right[$; then, $W J^W_{\gamma A} x$ is at most single-valued.*

(ii) *Moreover,*

$$\left(\forall \gamma \in \left]\rho\|V\|^2, +\infty\right[\right)\ \left(\forall (x, y) \in (\operatorname{dom}(J^W_{\gamma A}))^2\right)$$
$$\|W J^W_{\gamma A} x - W J^W_{\gamma A} y\|^2 \leq \|x - y\|^2 - \left(1 - 2\rho\|V\|^2/\gamma\right) \|x - y - W J^W_{\gamma A} x + W J^W_{\gamma A} y)\|^2. \tag{2.1}$$

As a consequence, if $\gamma > 2\rho\|V\|^2$, then, $W(W + \gamma A)^{-1}$ is $\frac{1}{2}\frac{\gamma}{\gamma - \rho\|V\|^2}$-averaged whenever $\operatorname{dom}((W + \gamma A)^{-1}) = \mathcal{H}$.

Proof. See Appendix B.1. □

Remark 2 *Lemma 1 can be found in [11] [5] for the case $W = V = \mathrm{Id}$. When $V = 0$, the notion of $(\rho V, W)$-hypomotonocity reduces to the notion of hypermonotonicity with respect to W. Several examples of this case can be found in [17].*

3. Main Algorithm

In this section, we investigate the convergence properties of the following warped proximal iterations for solving the nonmonotone inclusion (1.1). Let $W\colon \operatorname{dom}(W) = \mathcal{H} \to \mathcal{H}$ be a nonzero, single-valued (nonlinear) operator such that $(\forall n \in \mathbb{N})\ \operatorname{dom}(J^W_{\gamma_n A}) = \mathcal{H}$. Assume $(\lambda_n, \gamma_n)_{n\in\mathbb{N}}$ is a sequence in $]0, +\infty[^2$. Select $x_0 \in \mathcal{H}$ and iterate, for $n \in \mathbb{N}$,

$$\begin{cases} y_n \in (W + \gamma_n A)^{-1}(Wx_n) \\ Wx_{n+1} = Wx_n + \lambda_n(Wy_n - Wx_n). \end{cases} \tag{3.1}$$

Here are some main connections to existing work.

(i) When A is a maximally monotone operator and $W = \mathrm{Id}$, the iterative Algorithm (3.1) reduces to the classical proximal point algorithm. When W is a linear self-adjoint positive definite or semi-definite operator, the iteration (3.1) is known to as the *preconditioned proximal point algorithm*. The proper choice of W and A in suitable space enables recovering various primal- and primal-dual splitting algorithms such as [7] [2] [6] [12] [3] [9]. When W is a nonlinear operator, the Algorithm (3.1) is referred to as the *warped proximal iteration* [3]; it was also investigated independently in [9]. In these references, we can also find several choices of the nonlinear operator W.

(ii) When A is nonmonotone, solving inclusions is challenging in general, cf. [4]. However, when A^{-1} is hypomonotone and $W = \mathrm{Id}$, the iteration (3.1) first appeared in [11] where the weak convergence of the generated sequence to a point in $\operatorname{zer}(A)$ was proved. A general proximal point framework for finding a common zero point of the sequence $(A_i)_{i\in I}$ of co-hypomonotone operators was also presented in [5]. When A is the sum of two operators, two Nesterov's accelerated variants were recently proposed in [14] although without proving convergence of the iterations. Instead, when W is self-adjoint, and positive semidefinite, (3.1) has been recently studied under the V-oblique weak Minty solutions condition on A [8].

(iii) When A is hypermonotone w.r.t. W [15] [16], the Algorithm (3.1) was investigated in [16] and its extension in [17]. In particular, some special nonlinearly composed inclusions were also investigated in [17].

(iv) When A^{-1} is hypomonotone and W nonlinear, the convergence of (3.1) is still open. In this work, we show that weak convergence can be obtained when A^{-1} is $(\rho V, W)$-hypomonotone.

When A is maximally monotone or maximally ρ-hypermonotone, the graph of A, $\operatorname{gra}(A)$, is closed in $\mathcal{H}^{weak} \times \mathcal{H}^{strong}$, i.e., for every sequence $(x_n, u_n)_{n\in\mathbb{N}}$ such that $u_n \in Ax_n$, if $u_n \to u$ (strong convergence) and $x_n \rightharpoonup x$ (weak convergence); then, $u \in Ax$. In the following Proposition, we extend this property to the case where A^{-1} is $(\rho V, W)$-maximally hypomonotone.

Proposition 3 *Suppose that A^{-1} is $(\rho V, W)$-maximally hypomonotone for some $\rho \in\ ]0, +\infty[$, $V \in \mathcal{B}(\mathcal{H})$ and $W\colon \mathcal{H} \to \mathcal{H}$ is sequentially weakly continuous. Then, the graph of A, $\operatorname{gra}(A)$, is closed in $\mathcal{H}^{weak} \times \mathcal{H}^{strong}$.*

Proof. See Appendix C.1 □

Theorem 4 *Suppose that the operator A^{-1} is $(\rho V, W)$-hypomonotone for some $\rho \in]0,+\infty[$, the operator V belongs to $\mathcal{B}(\mathcal{H})$, i.e., the space of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$ with domain $\mathcal{H}$, and the single-valued operator $W\colon \mathcal{H} \to \mathcal{H}$ with $\operatorname{dom}(W) = \operatorname{dom}(V) = \operatorname{dom}(J^W_{\gamma A}) = \mathcal{H}$ for all $\gamma \in]0,+\infty[$. Let $n \in \mathbb{N}, \lambda_n \in [\varepsilon, 1]$, and $1 - 2\rho\|V\|^2/\gamma_n \geq \varepsilon$ for some $\varepsilon \in]0,1[$. Suppose either one of the following conditions is satisfied.*

(c1) A^{-1} is maximally hypomonotone and W is weakly continuous.

(c2) The graph of A, $\operatorname{gra}(A)$, is closed in $\mathcal{H}^{strong} \times \mathcal{H}^{weak}$.

Then,

(i) *The sequence $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point in $\operatorname{zer}(A)$ if W is injective, weakly continuous, and $\lim_{\|x\|\to\infty} \|Wx\| = \infty$.*

(ii) *The sequence $(y_n)_{n\in\mathbb{N}}$ converges weakly to a point in $\operatorname{zer}(A)$ if $J^W_{\gamma_n A}$ is singled-valued, and for all $(x, u) \in \operatorname{gra}(J^W_{\gamma_n A})$ and $(y, v) \in \operatorname{gra}(J^W_{\gamma_n A})$,*

$$\|u - v\| \leq \mu\|x - y\|, \tag{3.2}$$

where μ is a positive constant that is independent of n.

Proof. See Appendix C.2. □

Remark 5 *The assumption that the graph of A is closed in $\mathcal{H}^{weak} \times \mathcal{H}^{strong}$ was also used in [8].*

4. Applications

In this section, we apply Algorithm (3.1) to quadratic programming (QP) with equality constraints [10] and with (at least) one negative eigenvalue [13].

4.1. Nonconvex Quadratic programming

Let $A\colon \mathcal{H} \to \mathcal{H}$ be a self-adjoint, linear operator with closed range ($\operatorname{ran}(A)$ is closed). Let $b \in \mathcal{H}$ and $c \in \mathbb{R}$. Consider the following (unconstrained) problem:

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ \frac{1}{2}\langle Ax \mid x\rangle + \langle b \mid x\rangle + c. \tag{4.1}$$

This problem is nonconvex in general. We aim at finding a stationary point of this problem by solving the non-monotone inclusion of the form

$$0 \in Ax + b. \tag{4.2}$$

Since the range of A, $\operatorname{ran}(A)$, is closed and $A = A^*$, in view of [1, Fact 2.25, Fact 2.26], for each $u \in \operatorname{ran}(A)$, there exists only one $x \in \operatorname{ran}(A)$ such that $y = Ax$. Let denote by $x = A_r^{-1}y$; then, A_r^{-1} is a bounded linear operator as well.

Lemma 6 *Let W be an unitary operator on $\mathcal{H}$, and $A\colon \mathcal{H} \to \mathcal{H}$ a self-adjoint, linear operator. Then, $A^{-1}\colon \mathcal{H} \to 2^{\mathcal{H}}$ is $(\rho \operatorname{Id}, W)$-maximally hypomonotone with $\rho = \|A_r^{-1}\|$.*

Proof. See Appendix D.1. □

The main result of this section can now be stated.

Theorem 7 *Let $W\colon \mathcal{H} \to \mathcal{H}$ be a unitary operator on $\mathcal{H}$, and $\gamma \in \left]2\|A_r^{-1}\|, +\infty\right[$. Choose $x_0 \in \mathcal{H}$ and iterate*

$$x_{n+1} \in (W + \gamma A)^{-1}(Wx_n - \gamma b). \tag{4.3}$$

Then, $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point $x \in zer(A + b)$.

Proof. See Appendix D.2. □

4.2. Constrained quadratic programming

Let $\mathcal{H} = \mathbb{R}^n$. Suppose the operator $A\colon \mathcal{H} \to \mathcal{H}$ is self-adjoint, $b \in \mathcal{H}$ and $c \in \mathbb{R}$. Consider the following constrained minimization problem, where X is a closed convex set of $\mathcal{H}$, $C\colon \mathcal{H} \to \mathbb{R}^m$ is a linear operator and $d \in \mathbb{R}^m$

$$\underset{x\in X, Cx=d}{\text{minimize}}\ \frac{1}{2}\langle Ax \mid x\rangle + \langle b \mid x\rangle + c. \tag{4.4}$$

Let us consider the Lagrangian function $\mathcal{L}(x, \lambda)$ defined by

$$\mathcal{L}\colon (x, \lambda) \mapsto \iota_X(x) + \frac{1}{2}\langle Ax \mid x\rangle + \langle b \mid x\rangle + c - \langle \lambda \mid Cx - d\rangle\,, \tag{4.5}$$

where $\iota_X(x) = \iota_{\{x\in X\}}$ is the indicator function of the set X. The first order optimality conditions are $\nabla_x\mathcal{L}(x, \lambda) = 0$ and $\nabla_\lambda\mathcal{L}(x, \lambda) = 0$, which lead to the KKT system

$$\begin{cases} \partial\iota_X(x) + Ax + b - C^\intercal\lambda = 0 \\ Cx - d = 0, \end{cases}$$

where $\partial\iota_X(x) = N_X(x)$ is the normal cone operator of the set X at x. This system can be formulated as follows

$$\underbrace{\begin{pmatrix} \partial\iota_X & 0 \\ 0 & 0 \end{pmatrix}}_{\boldsymbol{B}} \underbrace{\begin{pmatrix} x \\ \lambda \end{pmatrix}}_{\boldsymbol{x}} + \underbrace{\begin{pmatrix} A & -C^\intercal \\ C & 0 \end{pmatrix}}_{\boldsymbol{A}} \begin{pmatrix} x \\ \lambda \end{pmatrix} + \underbrace{\begin{pmatrix} b \\ -d \end{pmatrix}}_{\boldsymbol{b}} \ni 0 \tag{4.6}$$

Corollary 8 *Let $W\colon \mathcal{H} \to \mathcal{H}$ be a single-valued operator. Set*

$$\boldsymbol{\mathcal{W}} := \begin{pmatrix} W & 0 \\ 0 & V = \mathrm{Id} \end{pmatrix}, \tag{4.7}$$

and $\boldsymbol{M} := \boldsymbol{A} + \boldsymbol{B}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined per (4.6). Suppose that $\boldsymbol{M} + \boldsymbol{b}$ is $(\rho\,\mathrm{Id}, \boldsymbol{\mathcal{W}})$-hypomonotone with closed graph. Let $\gamma \in \left]2\rho, +\infty\right[$. Choose $x_0 \in \mathbb{R}^n$ and $\lambda_0 \in \mathbb{R}^m$. Then, the iteration

$$\begin{cases} x_{n+1} = \big(W + \gamma(\partial\iota_X + A) + \gamma^2 C^\intercal C\big)^{-1}\big(Wx_n - \gamma(b - C^\intercal\lambda_n) + \gamma^2 C^\intercal d\big) \\ \lambda_{n+1} = \lambda_n - \gamma(Cx_{n+1} - d), \end{cases} \tag{4.8}$$

converges to a point $\boldsymbol{x} = (x, \lambda)$ solving (4.6).

Proof. See Appendix D.3. □ The problem is convex when the matrix of the (linear) operator A is positive semi-definite (PSD). On the other hand, the problem is nonconvex when the matrix of the (linear) operator A is indefinite. Similar condition applies to the reduced problem of the original linearly constrained QP defined by (4.1). Consider the constrained problem (4.9) is feasible for $x = (x_1, x_2)$.

$$\underset{C^\intercal x = d}{\text{minimize}}\ \frac{1}{2}\langle Ax \mid x\rangle + \langle b \mid x\rangle + c. \tag{4.9}$$

Assume the constraint matrix $C(m \times n), m \leq n$ is decomposable into $[C_1, C_2]$ with nonsingular $C_1(m \times m)$ and $C_2(m \times (n-m))$ such that the constraint (4.9) can be written as $C_1 x_1 + C_2 x_2 = d$. Then, the constrained problem can be reformulated as the reduced unconstrained problem (4.10) by eliminating $x_1 = C_1^{-\intercal}(d - C_2^\intercal x_2)$

$$\underset{x_2}{\text{minimize}}\ \frac{1}{2} x_2^\intercal \tilde{A} x_2 + \tilde{b}^\intercal x_2, \tag{4.10}$$

Note that if the matrix $\tilde{A}$ is not PSD (i.e., has negative eigenvalues); then, the reduced QP is unbounded. On the other hand, the reduced QP admits a unique solution if and only if the reduced Hessian $\tilde{A} = Z^\intercal A Z$ is positive definite, where Z is a basis of the null-space, i.e., $Z^\intercal C = 0$.

4.3. Numerical results

In this section, we report the numerical results obtained when solving i) the constrained problem (4.9) by means of iteration (4.3) named W-Hypo, and ii) the constrained problem (4.4) by means of iteration (4.8) named Hypo-Lag. Two types of datasets are considered: sparse matrices (nnz elements $< 20\%$) and dense matrices (nnz elements $> 80\%$). All datasets are generated such that C has full rank (FR), C_1 is nonsingular and $A + C^\intercal C$ is positive semi-definite. Results are compared against the QP solver of `Octave 8.3`, where we set the execution time limit (TL) to $10^4 s$.

4.3.1. SPARSE MATRICES (NNZ $< 20\%$)

Method	(n, m)	A	C	min. eigen (A)	Feasibility	Best objective	Nbr of iterations	Time (s)
	(1000,100)	NS:ND	FR	-0.1092				
W-Hypo	*	*	*	*	1e-15	-120.13	5	0.6899
Hypo-Lag	*	*	*	*	1e-15, 1e-14	-120.13	168	0.1282
QP	*	*	*	*	1e-14, 1e-14	-120.13	2	1.3985
	(5000,500)	NS:ND	FR	-0.5068				
W-Hypo	*	*	*	*	1e-15	-599.08	6	60.839
Hypo-Lag	*	*	*	*	1e-15, 1e-14	-599.08	459	8.4778
QP	*	*	*	*	1e-14, 1e-14	-599.08	2	144.54
	(10000,1000)	NS:ND	FR	-0.1181				
W-Hypo	*	*	*	*	1e-15	-1225.0	6	463.18
Hypo-Lag	*	*	*	*	1e-14, 1e-14	-1225.0	164	34
QP	*	*	*	*	1e-14, 1e-14	-1225.0	2	1213.2
	(15000,1000)	NS:ND	FR	-0.5133				
W-Hypo	*	*	*	*	1e-15	-2060.3	6	2295.3
Hypo-Lag	*	*	*	*	1e-13, 1e-13	-2060.3	159	102.75
QP	*	*	*	*	1e-14, 1e-14	-2060.3	2	5745.4
	(15000,1500)	NS:ND	FR	-0.010099				
W-Hypo	*	*	*	*	1e-15	-1988.2	6	2937.6
Hypo-Lag	*	*	*	*	1e-13, 1e-12	-1988.2	144	115.46
QP	*	*	*	*	1e-14, 1e-14	-1988.2	2	6630.23

Table 1: Results: Hypo(Lag) vs. QP Solver (NS = nonsingular, ND = indefinite, FR = Full Rank)

4.3.2. DENSE MATRICES (NNZ $> 80\%$)

Method	(n, m)	A	C	min. eigen (A)	Feasibility	Best objective	Nbr of iterations	Time (s)
	(1000,100)	NS:ND	FR	-0.1522				
W-Hypo	*	*	*	*	1e-15	-40.038	5	0.7434
Hypo-Lag	*	*	*	*	1e-14, 1e-14	-40.038	180	0.1638
QP	*	*	*	*	1e-14, 1e-14	-40.038	2	1.6882
	(5000,500)	NS:ND	FR	-3.8495e-03				
W-Hypo	*	*	*	*	1e-14	-158.20	6	63.456
Hypo-Lag	*	*	*	*	1e-14, 1e-13	-158.20	188	6.3901
QP	*	*	*	*	1e-14, 1e-13	-158.20	2	206.10
	(10000,1000)	NS:ND	FR	-0.012110				
W-Hypo	*	*	*	*	1e-14	-311.68	6	506.48
Hypo-Lag	*	*	*	*	1e-15, 1e-13	-311.68	557	66.670
QP	*	*	*	*	1e-13, 1e-13	-311.68	2	2102.1
	(15000,1000)	NS:ND	FR	-0.017933				
W-Hypo	*	*	*	*	1e-14	-625.71	6	2215.3
Hypo-Lag	*	*	*	*	1e-15, 1e-13	-625.71	319	181.46
QP	*	*	*	*	1e-13, 1e-13	-625.71	2	9688.2
	(15000,1500)	NS:ND	FR	-0.3705				
W-Hypo	*	*	*	*	1e-13	-668.57	6	2756.4
Hypo-Lag	*	*	*	*	1e-13, 1e-13	-668.57	1046	243.95
QP	*	*	*	*	1e-13, 1e-13	-668.57	2	TL

Table 2: Results: Hypo(Lag) vs. QP Solver (NS = nonsingular, ND = indefinite, FR = Full Rank)

References

[1] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 2nd edition, 2017.

[2] K. Bredies, E. Chenchene, D. A. Lorenz, and E. Naldi. Degenerate preconditioned proximal point algorithms. *SIAM Journal on Optimization*, 32(3):2376–2401, 2022.

[3] M. N. Bùi and P. L. Combettes. Warped proximal iterations for monotone inclusions. *Journal of Mathematical Analysis and Applications*, 49(1):124315, 2020.

[4] P. L. Combettes. Monotone operator theory in convex optimization. *Mathematical Programming*, 8170(1):177–206, 2018.

[5] P. L. Combettes and T. Pennanen. Proximal methods for cohypomonotone operators. *SIAM Journal on Control and Optimization*, 43(4):731–742, 2004.

[6] P. L. Combettes and B. C. Vu. Variable metric forward-backward splitting with application to monotone inclusions in duality. *Optimization*, 63(5):1289–1318, 2014.

[7] P. L. Combettes, L. Condat, J.-C. Pesquet, and B. C. V u. A forward-backward view of some primal-dual optimization methods in image recovery. *Proc. IEEE Int. Conf. Image Processing*, pages 4141–4145, Oct. 27-30 2014.

[8] B. Evens, P. Pas, P. Latafat, and P. Patrinos. Convergence of the preconditioned proximal point method and douglas-rachford splitting in the absence of monotonicity. arXiv preprint arXiv:2305.03605, 2023.

[9] P. Giselsson. Nonlinear forward-backward splitting with projection correction. *SIAM Journal on Optimization*, 31:2199–2226, 2021.

[10] N. I. Gould, M. E. Hribar, and J. Nocedal. On the solution of equality constrained quadratic programming problems arising in optimization. *SIAM Journal on Scientific Computing*, 23: 1376–1395, 2021.

[11] A. N. Iusem, T. Pennanen, and B. F. Svaiter. Inexact variants of the proximal point algorithm without monotonicity. *SIAM Journal on Optimization*, 13:947–1244, 2003.

[12] A. Contreras A. Hirabayashi L. Condat, D. Kitahara. Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists. *SIAM Review*, 65:375–435, 2023.

[13] P. M. Pardalos and S. A. Vavasis. Quadratic programming with one negative eigenvalue is np-hard. *Journal of Global Optimization*, 1:15–22, 1991.

[14] Q. Tran-Dinh. Extragradient-type methods with $\mathcal{O}(1/k)$-convergence rates for co-hypomonotone inclusions, 2023. arxiv:2302.04099.

[15] R. U. Verma. A-monotone nonlinear relaxed cocoercive variational inclusions. *Central European Journal of Mathematics*, 5:386–396, 2007.

[16] R. U. Verma. General proximal point algorithmic models and nonlinear variational inclusions involving rmm mappings. *Journal of Informatics and Mathematical Sciences*, 1:15–25, 2009.

[17] B. C. Vu and D. Papadimitriou. A nonlinearly preconditioned forward-backward splitting method and applications. *Numerical Functional Analysis and Optimization*, 42:1880–1895, 2022.

Appendix A. Definitions

Definition 9 *Let $W\colon \mathcal{H} \to \mathcal{H}$ be a single-valued operator. We say that the operator $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ is*

(i) *hypermonotone with respect to (w.r.t.) W if*

$$\big(\forall (x,u) \in \operatorname{gra}(A)\big)\,\big(\forall (y,v) \in \operatorname{gra}(A)\big) \quad \langle Wx - Wy \mid u - v\rangle \geq 0, \tag{A.1}$$

(ii) *maximally hypermonotone w.r.t. W if it is hypermonotone w.r.t. W and it follows from*

$$\big(\exists (x,u) \in \mathcal{H}^2\big)\,\big(\forall (y,v) \in \operatorname{gra}(A)\big) \quad \langle Wx - Wy \mid u - v\rangle \geq 0, \tag{A.2}$$

that $(x,u) \in \operatorname{gra}(A)$.

Definition 10 *Let ρ be in $]0,+\infty[$. An operator $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ is ρ-hypomonotone if*

$$\big(\forall (x,u) \in \operatorname{gra}(A)\big)\,\big(\forall (y,v) \in \operatorname{gra}(A)\big)\ \ \langle x - y \mid u - v\rangle + \rho\|x-y\|^2 \geq 0. \tag{A.3}$$

Equivalently, an operator A is ρ-hypomonotone if and only if $A + \rho\operatorname{Id}$ is monotone. Next, A is maximally ρ-hypomonotone if A is ρ-hypomonotone and there exists no ρ-hypomonotone operator $B\colon \mathcal{H} \to 2^{\mathcal{H}}$ such that $\operatorname{gra}(B)$ properly contains $\operatorname{gra}(A)$.

Definition 11 *Let ρ be in $]0,+\infty[$. Let $W\colon \mathcal{H} \to \mathcal{H}$ and $V\colon \mathcal{H} \to \mathcal{H}$ be single-valued operators with full domains, i.e, $\operatorname{dom}(W) = \operatorname{dom}(V) = \mathcal{H}$. An operator $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ is $(\rho V, W)$-hypomonotone if*

$$\big(\forall (x,u) \in \operatorname{gra}\,\big)\,\big(\forall (y,v) \in \operatorname{gra}(A)\big)\ \ \langle x - y \mid Wu - Wv\rangle + \rho\|Vx - Vy\|^2 \geq 0. \tag{A.4}$$

Next, A is maximally $(\rho V, W)$-hypomonotone if A is $(\rho V, W)$-hypomonotone and if there exists $(y,v) \in \mathcal{H}^2$ such that

$$\big(\forall (x,u) \in \operatorname{gra}(A)\big)\ \ \langle x - y \mid Wu - Wv\rangle + \rho\|Vx - Vy\|^2 \geq 0, \tag{A.5}$$

then $(y,v) \in \operatorname{gra}(A)$.

Definition 12

(i) *An operator $T\colon \mathcal{H} \to \mathcal{H}$ is nonexpansive if*

$$(\forall x \in \mathcal{H})\,(\forall y \in \mathcal{H})\ \|Tx - Ty\| \leq \|x - y\|. \tag{A.6}$$

(ii) *Let $\alpha \in \,]0,1[$, an operator $T\colon \mathcal{H} \to \mathcal{H}$ is α-averaged if there exists an nonexpansive operator $R\colon \mathcal{H} \to \mathcal{H}$ such that*

$$T = (1-\alpha)\operatorname{Id} + \alpha R. \tag{A.7}$$

Equivalently,

$$(\forall x \in \mathcal{H})\,(\forall y \in \mathcal{H})\ \|Tx - Ty\|^2 \leq \|x-y\|^2 - \frac{1-\alpha}{\alpha}\|(x-y) - (Tx - Ty)\|^2. \tag{A.8}$$

Our algorithm relies on the $W-$resolvent of operator A.

Definition 13 *The $W-$resolvent of the operator $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ of index $\gamma > 0$ is defined by*

$$\big(\gamma \in \,]0,+\infty[\,\big)\ J^{W}_{\gamma A}\colon \mathcal{H} \to \mathcal{H}\colon x \mapsto (W + \gamma A)^{-1}x. \tag{A.9}$$

Appendix B. Proof of Section 2

B.1. Proof of Lemma 1

Proof. (i) Let $x \in \mathrm{dom}(J^W_{\gamma A})$, and let p and q be two points in $J^W_{\gamma A}x = (W + \gamma A)^{-1}x$. Then,

$$\begin{cases}(x - Wp)/\gamma & \in Ap \\ (x - Wq)/\gamma & \in Aq\end{cases} \Longleftrightarrow \begin{cases}p & \in A^{-1}((x - Wp)/\gamma) \\ q & \in A^{-1}((x - Wq)/\gamma).\end{cases} \tag{B.1}$$

Since A^{-1} is $(\rho V, W)$-hypomonotone, it follows from (B.1) and (A.4) that

$$\langle (x - Wp)/\gamma - (x - Wq)/\gamma \mid Wp - Wq\rangle + \rho\|V(x - Wp)/\gamma - V(x - Wq)/\gamma\|^2 \geq 0. \tag{B.2}$$

Since V is linear, we obtain

$$\rho\|V\|^2\|Wp - Wq\|^2 \geq \rho\|V(Wp - Wq)\|^2 \geq \gamma\|Wp - Wq\|^2, \tag{B.3}$$

which implies that $Wp = Wq$ whenever $\gamma > \rho\|V\|^2$. Hence, $WJ^W_{\gamma A}$ is single-valued.
(ii) For each $x \in \mathrm{dom}(J^W_{\gamma A})$ and each $y \in \mathrm{dom}(J^W_{\gamma A})$, define $p \in (W + \gamma A)^{-1}x$ and $q \in (W + \gamma A)^{-1}y$. Then, $(x - Wp)/\gamma \in Ap$ and $(y - Wq)/\gamma \in Aq$. Hence, $p \in A^{-1}((x - Wp)/\gamma)$ and $q \in A^{-1}((y - Wq)/\gamma)$. Since A^{-1} is $(\rho V, W)$-hypomonotone, we obtain

$$\begin{aligned}
0 &\leq \gamma\,\langle Wp - Wq \mid x - Wp - y + Wq\rangle + \rho\|V(x - Wp - y + Wq)\|^2 \\
&\leq \gamma\,\langle Wp - Wq \mid x - y - (Wp - Wq)\rangle + \rho\|V\|^2\|x - y - (Wp - Wq)\|^2 \\
&= -(\gamma - \rho\|V\|^2)\|Wp - Wq\|^2 + \rho\|V\|^2\|x - y\|^2 - (2\rho\|V\|^2 - \gamma)\,\langle x - y \mid Wp - Wq\rangle \\
&= -(\gamma - \rho\|V\|^2)\|Wp - Wq\|^2 + \rho\|V\|^2\|x - y\|^2 \\
&\quad + (\rho\|V\|^2 - \gamma/2)[\|x - y - Wp + Wq\|^2 - \|Wp - Wq\|^2 - \|x - y\|^2] \\
&= (-\gamma/2)\|Wp - Wq\|^2 + (\gamma/2)\|x - y\|^2 + (\rho\|V\|^2 - \gamma/2)\|x - y + Wq - Wp\|^2.
\end{aligned} \tag{B.4}$$

Expression (B.4) implies that

$$\|Wp - Wq\|^2 \leq \|x - y\|^2 + (2\rho\|V\|^2/\gamma - 1)\|x - y + Wq - Wp\|^2, \tag{B.5}$$

which proves (2.1). In view of (A.8), a given operator is α-averaged if there exists $\alpha \in\,]0, 1[$ such that $(1 - \alpha)/\alpha = (1 - 2\rho\|V\|^2/\gamma)$. Hence, with $\alpha = \frac{1}{2}\frac{\gamma}{\gamma - \rho\|V\|^2}$ and $\gamma > 2\rho\|V\|^2$, the operator $W(W + \gamma A)^{-1}$ is α-averaged. □

Appendix C. Proof of Section 3

C.1. Proof of Proposition 3

Proof. Suppose that A^{-1} is $(\rho V, W)$-hypomonotone. Let $(x_n, u_n)_{n\in\mathbb{N}}$ be a sequence in $\mathrm{gra}(A)$ such that $u_n \to u$ and $x_n \rightharpoonup x$ for some $(x, u) \in \mathcal{H} \times \mathcal{H}$. For every $(v, y) \in \mathrm{gra}(A^{-1})$, we have

$$\langle u_n - v \mid Wx_n - Wy\rangle + \rho\|Vu_n - Vv\|^2 \geq 0. \tag{C.1}$$

Let $n \to \infty$, we obtain

$$\langle Wu - Wv \mid x - y\rangle + \rho\|Vu - Vv\|^2 \geq 0. \tag{C.2}$$

Since this inequality holds for every $(v, y) \in \mathrm{gra}(A^{-1})$, by definition, we obtain $(u, x) \in \mathrm{gra}(A^{-1})$. Hence, $(x, u) \in \mathrm{gra}(A)$. □

C.2. Proof of Theorem 4

Proof. Let $x \in \operatorname{zer}(A)$. Then, following (1.1),

$$x \in (W + \gamma_n A)^{-1} W x. \tag{C.3}$$

Since $\gamma_n > 2\rho\|V\|^2 > \rho\|V\|^2$. The operator $W(W + \gamma_n A)^{-1}$ is single-valued by Lemma 1(i). Hence,

$$Wx = W(W + \gamma_n A)^{-1} W x \ \text{ and } \ W y_n = W(W + \gamma_n A)^{-1} W x_n. \tag{C.4}$$

We also obtain following Lemma 1(ii) that

$$\begin{aligned}\|Wy_n - Wx\|^2 &= \|W(W + \gamma_n A)^{-1}(Wx_n) - W(W + \gamma_n A)^{-1}(Wx)\|^2 \\ &\leq \|Wx_n - Wx\|^2 + (2\rho\|V\|^2/\gamma_n - 1)\|Wx_n - Wx - Wy_n + Wx||^2 \\ &\leq \|Wx_n - Wx\|^2 - \varepsilon\|Wx_n - Wy_n||^2. \end{aligned} \tag{C.5}$$

Following the update rule, we have

$$Wx_{n+1} - Wx = (1 - \lambda_n)(Wx_n - Wx) + \lambda_n(Wy_n - Wx), \tag{C.6}$$

which together with the convexity of $\|\cdot\|^2$, yields

$$\|Wx_{n+1} - Wx\|^2 = (1 - \lambda_n)\|Wx_n - Wx\|^2 + \lambda_n\|Wy_n - Wx\|^2. \tag{C.7}$$

Hence,

$$\|Wx_{n+1} - Wx\|^2 \leq \|Wx_n - Wx\|^2 - \varepsilon\lambda_n|Wx_n - Wy_n||^2 \tag{C.8}$$

$$\leq \|Wx_n - Wx\|^2 - \varepsilon^2\|Wx_n - Wy_n||^2, \tag{C.9}$$

which implies that

$$\begin{cases}\exists \lim \|Wx_n - Wx\| = \xi(x) \in \mathbb{R}_+ \\ Wx_n - Wy_n \to 0.\end{cases} \tag{C.10}$$

Hence, $(\|Wx_n - Wx\|)_{n\in\mathbb{N}}$ is a bounded sequence. Since $(\|Wx_n - Wy_n\|)_{n\in\mathbb{N}}$ is also bounded, the sequence $(\|Wy_n - Wx\|)_{n\in\mathbb{N}}$ is bounded.

(i) Since $\lim_{\|x\|\to\infty}\|Wx\| = +\infty$, it follows that $(x_n, y_n)_{n\in\mathbb{N}}$ is bounded. Hence, the set of weak cluster points of $(y_n)_{n\in\mathbb{N}}$ is non-empty. Let y^* be a weak cluster point of $(y_n)_{n\in\mathbb{N}}$. Then, there exists a subsequence $y_{k_n} \rightharpoonup y^*$. Note that by using (3.1), we can also deduce that

$$(Wx_n - Wy_n)/\gamma_n \in Ay_n. \tag{C.11}$$

Suppose that the condition (c1) is verified. Henceforth, by using Proposition 3, we obtain $0 \in Ay^*$, i.e., $y^* \in \operatorname{zer}(A)$. Instead, if the condition (c2) is satisfied; then, we derive directly from (C.10) and (C.11) that $y^* \in \operatorname{zer}(A)$. Since W is injective, in view of [17, Lemma 2.8], $(x_n)_{n\in\mathbb{N}}$ converges weakly to some point $\overline{x} \in \operatorname{zer}(A)$.

(ii) Using (3.2), we have

$$\|y_n - x\| = \|(W + \gamma_n A)^{-1}(Wx_n) - (W + \gamma_n A)^{-1}(Wx)\| \leq \mu\|Wx_n - Wx\|. \tag{C.12}$$

Hence, $(y_n)_{n\in\mathbb{N}}$ is bounded.

Next, we prove that $(y_n)_{n\in\mathbb{N}}$ has only one weak cluster point. Indeed, suppose that $y_{k_n} \rightharpoonup y_1$ and $y_{k_m} \rightharpoonup y_2$. In view of (C.10), there exists $\xi(x) = \lim \|Wy_n - Wx\|$. If $Wy_1 \neq Wy_2$, the Opial property would give a contradiction. Indeed, the proof of this last statement follows from

$$\begin{cases} \xi(y_1) = \underline{\lim}\,\|Wy_{k_n} - Wy_1\| < \underline{\lim}\,\|Wy_{k_n} - Wy_2\| = \xi(y_2) \\ \xi(y_2) = \underline{\lim}\,\|Wy_{k_m} - Wy_2\| < \underline{\lim}\,\|Wy_{k_m} - Wy_2\| = \xi(y_1). \end{cases} \tag{C.13}$$

Therefore, $Wy_1 = Wy_2$. If either of the conditions $(c1)$ or $(c2)$ is satisfied; then, as shown above, every weak cluster point of $(y_n)_{n\in\mathbb{N}}$ is in $\mathrm{zer}(A)$. Hence,

$$y_1 = J^W_{\gamma_n A} W y_1 \ \text{ and } \ y_2 = J^W_{\gamma_n A} W y_2. \tag{C.14}$$

By using (3.2) again, we obtain $y_1 = y_2$. Thus, the sequence $(y_n)_{n\in\mathbb{N}}$ possesses at most one weak sequential cluster point. In view of [1, Lemma 2.46], since bounded, $(y_n)_{n\in\mathbb{N}}$ converges weakly to a point $\overline{x} \in \mathrm{zer}(A)$. □

Appendix D. Proof of Section 4

D.1. Proof of Lemma 6

Proof. We first note that $\mathcal{H} = \mathrm{ran}(A) \oplus \ker(A)$. Hence, for each $x \in \mathcal{H}$, there exist $x_1 \in \mathrm{ran}(A)$ and $x_2 \in \ker(A)$ (thus, by definition, $Ax_2 = 0$) such that $x = x_1 + x_2$. Since A is linear, we have thus

$$Ax = Ax_1 + Ax_2 = Ax_1 = A_r x_1. \tag{D.1}$$

Hence, $x \in A^{-1}y$ implies that $y = Ax = A_r x_1$; thus, $x_1 = A_r^{-1}y$. In turn, $x = x_1 + x_2 \in A_r^{-1}y + \ker(A)$, i.e.,

$$A^{-1}y = A_r^{-1}y + \ker(A). \tag{D.2}$$

Next, we let $(u,x) \in \mathrm{gra}(A^{-1})$ and $(v,y) \in \mathrm{gra}(A^{-1})$. Then, $x = A_r^{-1}u + w_1$ and $y = A_r^{-1}v + w_2$ for some $(w_1,w_2) \in \ker(A)^2$. Therefore, since W is unitary, we obtain

$$\begin{aligned} \langle u - v \mid Wx - Wy\rangle = \langle u - v \mid x - y\rangle &= \left\langle u - v \mid A_r^{-1}u - A_r^{-1}v\right\rangle \\ &\geq -\|u-v\|\|A_r^{-1}u - A_r^{-1}v\| \geq -\|A_r^{-1}\|\|u-v\|^2, \end{aligned} \tag{D.3}$$

which implies that A is $(\rho\,\mathrm{Id}, W)$-hypomonotone.

Next, we prove that A^{-1} is maximally hypomonotone. Suppose that there exists $(v,y) \in \mathcal{H}$ such that for every $(u,x) \in \mathrm{gra}(A^{-1})$,

$$\langle u - v \mid Wx - Wy\rangle + \rho\|u - v\|^2 \geq 0. \tag{D.4}$$

Let us write $v = v_1 + v_2 \in \mathrm{ran}(A) \oplus \ker(A)$. It follows that there exists $\bar{y} \in \mathrm{ran}(A)$ such that $v_1 = A\bar{y}$. For each $\lambda \in \mathbb{R}$, set $x_\lambda = \bar{y} + \lambda v_2$ and $u_\lambda = A\bar{y} = Ax_\lambda = v_1$; thus, $Ax_\lambda = u_\lambda$. Hence,

$(u_\lambda, x_\lambda)_{\lambda\in\mathbb{R}} \subset \operatorname{gra}(A^{-1})$. Using (D.4), we obtain

$$\begin{aligned}0 &\leq \langle u_\lambda - v \mid Wx_\lambda - Wy\rangle + \rho\|u_\lambda - v\|^2 \\ &= \langle u_\lambda - v \mid x_\lambda - y\rangle + \rho\|u_\lambda - v\|^2 \\ &= \langle -v_2 \mid \bar{y} - y + \lambda v_2\rangle + \rho\|v_2\|^2 \\ &= (\rho - \lambda)\|v_2\|^2 + \langle v_2 \mid y - \bar{y}\rangle \\ &= (\rho - \lambda)\|v_2\|^2 + \langle v_2 \mid y\rangle \end{aligned} \tag{D.5}$$

If $v_2 \neq 0$, let $\lambda \to +\infty$, we get a contraction. Therefore, $v_2 = 0$ and (D.4) becomes

$$(\forall x \in \mathcal{H}) \ \langle Ax - A\bar{y} \mid x - y\rangle + \rho\|Ax - A\bar{y}\|^2 \geq 0. \tag{D.6}$$

By taking $x = y$, it follows that $Ay = A\bar{y} = v$; thus, $(v, y) \in \operatorname{gra}(A^{-1})$. Consequently, A^{-1} is maximally hypomonotone. □

D.2. Proof of Theorem 7

Proof. Let define $B\colon x \mapsto Ax + b$. Then, $\operatorname{zer}(B) = \operatorname{zer}(A + b)$ and B^{-1} is also $(\|A_r^{-1}\| \operatorname{Id}, W)$-maximally hypomonotone. Moreover, the iteration (4.3) can be rewritten as

$$x_{n+1} \in (W + \gamma B)^{-1}(Wx_n), \tag{D.7}$$

which is an instance of (3.1) with $\lambda_n := 1$. Moreover, by Lemma 6, all the conditions set on B are satisfied. In addition, since W is unitary, it is injective, weakly continuous and coercive $\|Wx\| \to +\infty$ as $\|x\| \to \infty$. Therefore, by Theorem 4(i), $x_n \rightharpoonup \overline{x} \in \operatorname{zer}(B) = \operatorname{zer}(A + b)$. □

D.3. Proof of Corollary 8

Proof. Let $n \in \mathbb{N}$. It follows from the definition of x_{n+1} in (4.8) that

$$Wx_n - Wx_{n+1} - \gamma b + \gamma^2 C^\intercal d + \gamma C^\intercal \lambda_n \in \gamma\partial\iota_X(x_{n+1}) + \gamma Ax_{n+1} + \gamma^2 C^\intercal C x_{n+1}, \tag{D.8}$$

which is equivalent to

$$Wx_n - Wx_{n+1} - \gamma b \in \gamma\partial\iota_X(x_{n+1}) + \gamma Ax_{n+1} + \gamma C^\intercal(\gamma C x_{n+1} - \lambda_n - \gamma d). \tag{D.9}$$

By (4.8), we also have $\lambda_{n+1} = \lambda_n - \gamma C x_{n+1} + \gamma d$. Hence, we can deduce from (D.10) that

$$Wx_n - Wx_{n+1} - \gamma b \in \gamma\partial\iota_X(x_{n+1}) + \gamma Ax_{n+1} - \gamma C^\intercal \lambda_{n+1}. \tag{D.10}$$

Next, we define $\boldsymbol{x}_n := (x_n, \lambda_n)^\intercal$. Using the definition of $\boldsymbol{M}$, we then obtain

$$\boldsymbol{\mathcal{W}}\boldsymbol{x}_n - \boldsymbol{\mathcal{W}}\boldsymbol{x}_{n+1} - \gamma\boldsymbol{b} \in \gamma\boldsymbol{M}\boldsymbol{x}_{n+1}. \tag{D.11}$$

Therefore, $\boldsymbol{x}_{n+1} = (\boldsymbol{\mathcal{W}} + \gamma\boldsymbol{M})^{-1}(\boldsymbol{\mathcal{W}}\boldsymbol{x}_n - \gamma\boldsymbol{b})$. Hence, $\boldsymbol{x}_n \to \boldsymbol{x}$ solves (4.6). □