Quasi-Newton Methods for Monotone Inclusions: Efficient Resolvent Calculus and Primal-Dual Algorithms

Shida Wang*, Jalal Fadili**, and Peter Ochs*

*Department of Mathematics, Saarland University, Germany
**Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France.

September 13, 2024

Abstract

We introduce ~~an inertial~~ two quasi-Newton Forward-Backward Splitting Algorithms to solve a class of monotone inclusion problems. ~~While the inertial step is computationally cheap, in general, the~~ The bottleneck is the evaluation of the resolvent operator. Changing the metric makes its computation even harder, and this is even true for a simple operator whose resolvent is known for the standard metric. To fully exploit the advantage of adapting the metric, we develop a new efficient resolvent calculus for a low-rank perturbed standard metric, which accounts exactly for quasi-Newton metrics. Moreover, we prove the convergence of our algorithms, including linear convergence rates in case one of the two considered operators is strongly monotone. As a by-product of our general monotone inclusion framework, we ~~instiante~~introduce two variants of quasi-Newton Primal-Dual Hybrid Gradient Method (PDHG) for solving saddle point problems. The favourable performance of these two quasi-Newton PDHG methods is demonstrated on several numerical experiments in image processing.

1 Introduction

Nowadays, convex optimization is prevalent in many modern disciplines, especially when dealing with large-scale datasets. There is a strong need for efficient optimization schemes to solve such large-scale problems. Unfortunately, the high dimensionality of the problems at hand makes ~~the use of~~ second-order methods intractable. A promising alternative is quasi-Newton type methods, which aim to exploit cheap and accurate first-order approximations of the second-order information. In particular, the so-called limited memory quasi-Newton method is highly effective for solving unconstrained large-scale problems. However, many practical problems in machine learning, image processing or statistics naturally have constraints or are non-smooth by ~~construction~~design.

A problem structure that can cover a broad class of non-smooth problems in these applications is the following inclusion problem in a real Hilbert space $\mathcal{H}$:

$$\text{find } x \in \mathcal{H} \ \text{ such that } \ (A + B)x \ni 0\,, \tag{1}$$

where $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is a maximally monotone operator, $B\colon \mathcal{H} \to \mathcal{H}$ is a single-valued β-co-coercive operator with $\beta > 0$. As a special case, (1) ~~comprises~~includes the setting of minimization problems of the form

$$\min_{x\in\mathcal{H}} f(x) + g(x) \tag{2}$$

with a proper lower semi-continuous convex function f and convex function g with Lipschitz continuous gradient by setting $A = \partial f$ and $B = \nabla g$.

A fundamental algorithmic scheme to tackle the problem class (1) is Forward-Backward Splitting (FBS). However, this algorithm may exhibit slow convergence for ill-conditioned problems, where exploiting the second-order information to adapt to the local geometry of the problem is desirable. As a computationally affordable approximation, in this paper, we propose a quasi-Newton variant that ~~takes advantage from~~uses a variable metric ~~that is~~ computed solely from first-order information. We ~~manage to remedy~~address the main computational bottleneck~~s~~ for this type of approaches by developing an efficient low-rank variable metric resolvent calculus.

Our approach is inspired by the proximal quasi-Newton method in [8]. We extend the framework proposed in [8] to the resolvent setting with a "M + rank-r" or "M − rank-r" symmetric positive-definite variable metric. We develop a resolvent calculus that allows for an efficient evaluation of resolvent operators with respect to ~~this type of~~ these metrics by splitting the evaluation into two computationally simple steps: ~~calculating~~evaluating a resolvent operator with respect to a simple metric M, and solving a low dimensional root-finding problem. This allows for the incorporation of popular quasi-Newton strategies, such as the limited memory SR1 or BFGS method, in our framework. Then, we study the convergence of two variants of FBS algorithm with respect to this type of metrics. One variant uses an inertial step which opens the door to acceleration. Although accelerated rates are not proved yet, numerical results show that the inertial variant yields significantly improved convergence rates. The second variant uses a relaxation technique~~which~~, enabling convergence under a weaker assumption on the variable metrics.

~~In order to exploit~~To showcase the power and to illustrate the variety of problems that can be solved via the framework in (1), the developed algorithms are instantiated for the following saddle point problem:

$$\min_{x\in\mathcal{H}_1}\max_{y\in\mathcal{H}_2} \langle Kx, y\rangle + g(x) + G(x) - f(y) - F(y)\,, \tag{3}$$

where g and f are lower semi-continuous convex functions, G and F are convex, differentiable with Lipschitz-continuous gradients, $\langle\cdot,\cdot\rangle$ denotes the innerproduct in $\mathcal{H}_2$ and K is a bounded linear operator between Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. We develop a quasi-Newton primal-dual methods that has many potential applications in image processing, machine learning or statistics [12, 19]. The numerical performance of our algorithms is tested on several experiments~~and~~, demonstrating a clear improvement when using our quasi-Newton methods.

1.1 Related Works

Smooth quasi-Newton. Quasi-Newton methods are widely studied and used for optimization with sufficiently smooth objective functions [42]. Their ~~motivation~~goal is ~~to build a cheap~~to design a computationally inexpensive approximation to Newton's method. If the approximation of the second-order information (Hessian) is given by a positive-definite matrix, quasi-Newton methods can be interpreted as iteratively and locally adapting the metric of the space to the objective function. The success of these methods ~~requires~~depends on accurately approximating the second-order information ~~of the objective~~ by using the first-order information, which ~~is still~~remains an active research area. Recently, [34] proposed a greedy strategy to select basis vectors, rather than using the difference of successive iterates for updating the Hessian approximations. Inspired by [34], [28] developed an approximation of the indefinite Hessian of a twice continuously differentiable function. However, all methods mentioned above require sufficient smoothness of the objective functions.

Non-smooth quasi-Newton. A broad class of optimization problems can be interpreted as a composition of a smooth function f and a non-smooth function h. To deal with the non-smoothness of h efficiently, many authors consider a combination of FBS with the quasi-Newton methods. By using the forward-backward envelope, [31, 39] reinterpret the FBS algorithm as a variable metric gradient method for a smooth optimization problem ~~in order to apply~~to enable the application of the classical Newton or quasi-Newton method. For a non-smooth function g as simple as an indicator function of a non-empty convex set, [36, 37] proposed an elegant method ~~named~~called projected quasi-Newton algorithm (PQN) which, however, requires either solving a subproblem or using a diagonal metric. [26] extended PQN to a more general setting as long as the proximal operator of h is simple to compute. For a class of low-rank perturbed metrics, [7, 8] developed a proximal quasi-Newton method with a root-finding problem as the subproblem which can be solved easily and efficiently. This method can be extended to the nonconvex setting [22]. Based on [7, 8], [23] incorporated a limited-memory quasi-Newton update. The authors of [24] developed a different algorithm to evaluate the proximal operator of the separable l_1 norm with respect to a low-rank metric $V = M - UU^\top$. Recently, in [25], the authors proposed a generalized damped Newton type based on second-order generalized Hessians.

In this paper, we extend the quasi-Newton approach of [7, 8] from the nonsmooth convex minimization setting to monotone inclusion problems of type (1). Our framework opens the door to new problems (e.g. saddle-point problems) and algorithms (e.g., primal-dual algorithms) that are beyond the reach of the approach initiated in [7, 8]. It is ~~generic~~common to use a variable metric for solving a monotone inclusion problem (see [15, 16]). Its convergence relies on quasi-Fejér monotonicity [15]. However, the efficient calculation of the resolvent remains an open problem. Our approach uses a variable metric to obtain a quasi-Newton method with an efficient resolvent calculus. Regularized Newton-type methods in continuous time, both for convex optimization and monotone inclusions have been studied in a series of papers by Attouch and his co-authors: [1, 3, 4, 2]. Time discretization of these dynamics gives algorithms providing insight into regularized Newton's method for solving monotone inclusions (see [1, 4, 2]). In [2], a relative error tolerance for the solution of the proximal subproblem is also allowed. However, in all these papers, the Hessian ends up being discretized.

PDHG. Primal-Dual Hybrid Gradient (PDHG) is widely used for solving saddle point problems of the form (3). PDHG can be interpreted as a proximal point algorithm [20] with a fixed metric applied to a monotone inclusion problem. Based on this idea, [30] proposed an inertial FBS method applied to the sum of set-valued operators, from which a generalization of PDHG method is derived. ~~Also based on similar ideas~~Similarly, [32] considered diagonal preconditioning to accelerate PDHG. Their method can be regarded as using a fixed blocked matrix as a metric. Later, [29] considered non-diagonal preconditioning and pointed out that if a special preconditioner is chosen, ~~that kind of~~this preconditioned PDHG method will be a special form of the linearized ADMM. Their method requires an inner loop due to the non-diagonal preconditioning. [19] introduced an adaptive PDHG scheme which can also be understood as using a variable metric with step size tuned automatically. However, [19] focuses on changes to the diagonal of the metric. Our resolvent calculus in Section 3 provides another ~~possibility~~way to ~~change~~modify metrics at elements off the diagonal to ~~deal~~handle with resolvent operators. Additionally, in [27], the author investigated inexact inertial variable proximal point algorithm with a different condition on the inertial step and distinct assumptions on error terms compared to ours.

2 Preliminaries

Let us recall some essential notations and definitions. Let $\mathcal{H}$ be a Hilbert space equipped with an inner product $\langle\cdot,\cdot\rangle$ and the induced norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$. The symbols $\rightharpoonup$ and $\to$ respectively denote weak and strong convergence. $\ell^1_+(\mathbb{N})$ is the set of all summable sequences in $[0,+\infty)$. An operator $K \in \mathcal{B}(\mathcal{D},\mathcal{H})$ is a linear bounded mapping from a Hilbert space $\mathcal{D}$ to $\mathcal{H}$. The adjoint of K is denoted by K^*. We abbreviate $\mathcal{B}(\mathcal{H},\mathcal{H})$ to $\mathcal{B}(\mathcal{H})$. We define $\mathcal{S}(\mathcal{H}) \coloneqq \{M \in \mathcal{B}(\mathcal{H}) | M = M^*\}$ and the identity operator by $I \in \mathcal{S}(\mathcal{H})$. Without ambiguity, we also use the notation $\|M\|$ for the operator norm of $M \in \mathcal{S}(\mathcal{H})$ with respect to $\|\cdot\|$. The partial ordering on $\mathcal{S}(\mathcal{H})$ is given by

$$(\forall U \in \mathcal{S}(\mathcal{H}))(\forall V \in \mathcal{S}(\mathcal{H}))\colon \quad U \succeq V \iff (\forall x \in \mathcal{H})\colon \ \langle Ux, x\rangle \geq \langle Vx, x\rangle \ . \tag{4}$$

For $\sigma \in [0,+\infty)$, we introduce $\mathcal{S}_\sigma(\mathcal{H}) \coloneqq \{U \in \mathcal{S}(\mathcal{H}) | U \succeq \sigma I\}$. Similarly, we introduce $\mathcal{S}_{++}(\mathcal{H}) \coloneqq \{U \in \mathcal{S}(\mathcal{H}) | U \succ 0\}$. In particular, $\mathcal{S}_{++}(\mathbb{R}^n)$ denotes the set of $n \times n$ real symmetric positive definite matrices. The norm $\|\cdot\|_M$ is defined by $\sqrt{\langle M\cdot,\cdot\rangle}$ for $M \in \mathcal{S}_{++}(\mathcal{H})$. We say $Q \in \mathcal{S}_0(\mathcal{H})$ has finite rank r if $r = \dim(\operatorname{im}(Q))$. Then, there are linearly independent vectors u_i such that $Q\colon \mathcal{H} \to \mathcal{H},\ x \mapsto \sum_{i=1}^r \langle x, u_i\rangle\, u_i$. As a consequence, $Q = UU^*$ where $U\colon \mathbb{R}^r \to \operatorname{im}(Q)$, $\alpha \mapsto U\alpha \coloneqq \sum_{i=1}^r \alpha_i u_i$ is an isomorphism defined by $(u_i)_{i=1,\dots,r}$.

A set valued operator $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is defined by its graph

$$\operatorname{Graph} A \coloneqq \{(x,y) \in \mathcal{H} | x \in \operatorname{Dom}(A), y \in Ax\}\,,$$

and has a domain given by

$$\operatorname{Dom}(A) \coloneqq \{x \in \mathcal{H} | Ax \neq \emptyset\}.$$

Given two set-valued operators $A, B\colon \mathcal{H} \rightrightarrows \mathcal{H}$, we define $A + B\colon \mathcal{H} \rightrightarrows \mathcal{H}$ as follows:

$$\begin{aligned} &\operatorname{Dom}(A+B) = \operatorname{Dom}(A) \cap \operatorname{Dom}(B)\,, \\ &\quad (A+B)x = Ax + Bx \coloneqq \{y \in \mathcal{H} | \exists y_1 \in Ax, \exists y_2 \in Bx \ \text{such that} \ y = y_1 + y_2\}\,. \end{aligned}$$

The inverse of A is denoted by A^{-1} given by $A^{-1}(y) \coloneqq \{x \in \mathcal{H} | y \in Ax\}$ and the zero set of A is denoted by $\operatorname{zer}(A+B) \coloneqq \{x \in \mathcal{H} | (A+B)x \ni 0\}$. We say that A is γ_A-strongly monotone with modulus $\gamma_A \geq 0$ with respect to norm $\|\cdot\|$, if $\langle x-y, u-v\rangle \geq \gamma_A \|x-y\|^2$ for any pair $(x,u)\,,(y,v) \in \operatorname{Graph} A$. A is maximally monotone, if for every $(x,u) \in \mathcal{H} \times \mathcal{H}$,

$$(x,u) \in \operatorname{Graph} A \iff (\forall (y,v) \in \operatorname{Graph} A) \quad \langle x-y, u-v\rangle \geq 0\,.$$

For a proper, lower semi-continuous, convex function f, ∂f is maximally monotone. Here, we adopt the common definition of the subdifferential ∂f [6, Definition 16.1] We say that a single valued operator B is β-co-coercive with respect to norm $\|\cdot\|$, if $\langle x-y, u-v\rangle \geq \beta\|u-v\|^2$ for any pair $(x,u)\,,(y,v) \in \operatorname{Graph} B$. The resolvent of $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ with respect to metric $M \in \mathcal{S}_{++}(\mathcal{H})$ is defined as

$$J_A^M := (I + M^{-1}A)^{-1} \text{ and we set } J_A := J_A^I \text{ for the identity mapping } I\,, \tag{5}$$

which, as shown for example in [6], enjoys the following properties.

Proposition 2.1. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, $M \in \mathcal{S}_{++}(\mathcal{H})$ and $y \in \mathcal{H}$. Then, the following holds*

$$y = J_{\gamma A}^M(x) \iff x \in y + \gamma M^{-1}Ay \iff x - y \in \gamma M^{-1}Ay \iff (y, \gamma^{-1}M(x-y)) \in \operatorname{Graph} A\,. \tag{6}$$

Proposition 2.2. *Let $A : \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone. Then, for every sequence $(x_k, u_k)_{k\in\mathbb{N}}$ in $\operatorname{Graph} A$ and every $(x,u) \in \mathcal{H}\times\mathcal{H}$, if $x_k \rightharpoonup x$ and $u_k \to u$, we have $(x,u) \in \operatorname{Graph} A$.*

Lemma 2.3. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be a maximally monotone operator and let $M \in \mathcal{S}_{++}(\mathcal{H})$. Then, for any $z \in \mathcal{H}$, we have $J_A^M(z) = M^{-1/2} \circ J_{M^{-1/2}AM^{-1/2}} \circ M^{1/2}(z)\,.$*

Proof. See [6, Proposition 23.34]. □

Lemma 2.4. *If A is strongly monotone with modulus $\gamma_A \geq 0$ and $M \in \mathcal{S}_{++}$, then J_A^M is Lipschitz continuous with respect to $\|\cdot\|_M$ with constant $1/(1+\frac{\gamma_A}{C}) \in (0,1]$ for any C satisfying $\|M\| \leq C < \infty$.*

Proof. See Appendix A.1. □

Proposition 2.5 (Variable Metric quasi-Fejér monotone sequence [15])**.** *Let $\sigma \in (0,+\infty)$, let $\varphi\colon [0,+\infty) \to [0,+\infty)$ be strictly increasing and such that $\lim_{t\to+\infty}\varphi(t) = +\infty$, let $(M_k)_{k\in\mathbb{N}}$ be in $\mathcal{S}_\sigma(\mathcal{H})$, let C be a nonempty subset of $\mathcal{H}$, and let $(x_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ such that*

$$(\exists(\eta_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathbb{N}))(\forall z \in C)(\exists(\epsilon_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathbb{N}))(\forall k \in \mathbb{N})\colon \\ \varphi(\|x_{k+1} - z\|_{M_{k+1}}) \leq (1+\eta_k)\varphi(\|x_k - z\|_{M_k}) + \epsilon_k\,. \tag{7}$$

(a) Then $(x_k)_{k\in\mathbb{N}}$ is bounded and, for every $z \in C$, $(\|z_k - z\|_{M_k})_{k\in\mathbb{N}}$ converges.

(b) If additionally, there exists $M \in \mathcal{S}_\sigma(\mathcal{H})$ such that $M_k \to M$ pointwisely, as is the case when

$$\sup_{k\in\mathbb{N}}\|M_k\| < +\infty \quad \text{and} \quad (\exists(\eta_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathbb{N}))(\forall k \in N)\colon\, (1+\eta_k)M_k \succeq M_{k+1}\,, \tag{8}$$

then $(x_k)_{k\in\mathbb{N}}$ converges weakly to a point in C if and only if every weak sequential cluster point of $(x_k)_{k\in\mathbb{N}}$ lies in C.

A key result for our resolvent calculus in Section 3 is the following Attouch-Théra abstract duality principle.

Lemma 2.6 (A duality result for operators [5])**.** *Let $T\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be an operator such that T^{-1} is single-valued and let $R\colon \mathcal{H} \to \mathcal{H}$ be a single-valued operator. Then, the following holds for $x, u \in \mathcal{H}$:*

$$\begin{cases} 0 \in Tx + Rx \\ 0 \in R^{-1}u - T^{-1}(-u) \end{cases} \iff \begin{cases} x \in R^{-1}u \\ -u \in Tx \end{cases} \iff \begin{cases} Rx = u \\ x = T^{-1}(-u) \end{cases}\,. \tag{9}$$

Moreover, if there exists $x \in \mathcal{H}$ such that $0 \in Tx + Rx$ or there exists $u \in \mathcal{H}$ such that $0 \in R^{-1}u - T^{-1}(-u)$, then there exists a unique primal-dual pair (x,u) that satisfies the equivalent conditions above.

We also need the following lemma which was stated as [33, Lemma 2.2.2].

Lemma 2.7. *Let $C_k \geq 0$ and let*

$$\begin{aligned} &C_{k+1} \leq (1+\nu_k)C_k + \zeta_k, \quad \nu_k \geq 0,\ \zeta_k \geq 0, \\ &\sum_{k\in\mathbb{N}}\nu_k < \infty, \quad \sum_{k\in\mathbb{N}}\zeta_k < \infty. \end{aligned} \tag{10}$$

Then, C_k converges to a non-negative limit.

3 Resolvent Calculus for Low-Rank Perturbed Metric

In this section, we extend the proximal calculus of [8] to the setting of resolvent operators J_A^V with a symmetric positive definite metric $V = M + \mathrm{s}Q$, where $\mathrm{s} \in \{-1, +1\}$, M is symmetric positive definite, and Q is symmetric positive semi-definite. This extension is called resolvent calculus. Then, we show the application of our resolvent calculus to a forward-backward update step.

3.1 Resolvent Calculus

This is a key result, as it enables an efficient application of quasi-Newton methods for solving monotone inclusion problems. Computing the resolvent operator $J_A^V(z)$ involves evaluating J_A^M at a shifted point $z - M^{-1}v^* \in \mathcal{H}$, where $v^* \in \mathcal{H}$ is derived from a root-finding problem, solvable by a semi-smooth Newton method (Algorithm 1) or a bisection method (Algorithm 2). In conclusion, if J_A^M can be computed efficiently, the same is true for J_A^V. The result crucially relies on the abstract duality principle of Attouch-Théra [5] (see Lemma 2.6). We first state the abstract result in a Hilbert space $\mathcal{H}$ in Theorem 3.1 and illustrate it in Corollary 3.3 with $\mathcal{H} = \mathbb{R}^n, n \in \mathbb{N}$.

Theorem 3.1. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be a maximally monotone operator, $V := M + \mathrm{s}Q \in \mathcal{S}_{++}(\mathcal{H})$ where $\mathrm{s} \in \{-1, 1\}$, $M \in \mathcal{S}_{++}(\mathcal{H})$ and $Q \in \mathcal{S}_0(\mathcal{H})$ having finite rank r. Then, the resolvent $J_A^V(z)$ can be computed as follows:*

$$x^* = J_A^V(z) \quad \iff \quad \begin{cases} x^* = J_A^M(z - \mathrm{s}M^{-1}U\alpha^*) \ \text{ and} \\ \alpha^* \in \mathbb{R}^r \ \text{ solves } \ \ell(\alpha) = 0 \\ \ \text{where } \ell(\alpha) := U^*Q^{-1}U\alpha + U^*(z - J_A^M(z - \mathrm{s}M^{-1}U\alpha)), \end{cases} \tag{11}$$

where $U\colon \mathbb{R}^r \to \mathrm{im}(Q)$, $\alpha \mapsto U\alpha := \sum_{i=1}^r \alpha_i u_i$ is an isomorphism defined by any r linearly independent $u_1, \ldots, u_r \in \mathrm{im}(Q)$. The function $\ell\colon \mathbb{R}^r \to \mathbb{R}^r$ is Lipschitz continuous with constant $\|U^*Q^{-1}U\| + \|M^{-1/2}U\|^2$ and strictly monotone.

Proof. See Appendix A.4. □

Remark 3.2. *If $Q = UU^*$, then $U^*Q^{-1}U = I$.*

We will delete it.

In finite dimensions, Theorem 3.1 simplifies to the following corollary.

Corollary 3.3. *Let $A\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximally monotone operator and consider $V := M + \mathrm{s}Q \in \mathcal{S}_{++}(\mathbb{R}^n)$ where $\mathrm{s} \in \{-1, +1\}$, $M \in \mathcal{S}_{++}(\mathbb{R}^n)$ and $Q \in \mathcal{S}_0(\mathbb{R}^n)$. Let $Q = UU^\top$ where $U\colon \mathbb{R}^r \to \mathbb{R}^n$ is a matrix of full rank r with $r \leq n$. Then, the resolvent operator J_A^V can be computed as follows:*

$$x^* = J_A^V(z) \quad \iff \quad \begin{cases} x^* = J_A^M(z - \mathrm{s}M^{-1}U\alpha^*) \ \text{ and} \\ \alpha^* \in \mathbb{R}^r \ \text{ solves } \ \ell(\alpha) = 0 \\ \ \text{where } \ \ell(\alpha) := \alpha + U^\top(z - J_A^M(z - \mathrm{s}M^{-1}U\alpha)). \end{cases} \tag{12}$$

The solution α^* is the unique root of $\ell\colon \mathbb{R}^r \to \mathbb{R}^r$ which is Lipschitz continuous with constant $1 + \|M^{-1/2}U\|^2$ and strictly monotone.

Proof. See Appendix A.5. □

If $r \ll n$, then we have a so-called low rank perturbed metric $M + \mathrm{s}UU^\top$, which leads to a root-finding problem of low dimension r. Together with the simple metric M, with respect to which the resolvent operator is easy to evaluate, this setup leads to an efficient evaluation of a resolvent operator with respect to a low rank perturbed metric.

3.2 Forward-backward update step

Proximal quasi-Newton methods also incorporate a forward step, which is adapted to the metric. In our case, the forward-backward step (the resolvent quasi-Newton step) is $J_A^V(z - V^{-1}Bz)$ where $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is a maximally monotone operator, $B\colon \mathcal{H} \to \mathcal{H}$ is a single-valued operator, $V = M + \mathrm{s}Q$ and $Q = UU^*$ for a bounded linear isomorphism $U\colon \mathbb{R}^r \to \operatorname{im}(Q)$. The resolvent calculus (Theorem 3.1) can be directly ~~used~~applied to the forward-backward step at the shifted point $z - V^{-1}Bz$. However, evaluating this point involves inverting V. Since V has a special structure, we will show that the update amounts to inverting solely M without applying the Sherman-Morrison-Woodbury formula.

Proposition 3.4. *Consider the setting of Theorem 3.1. Let B be a single-valued operator. Let $Q \in \mathcal{S}_0(\mathcal{H})$ with $Q = UU^*$ for a bounded linear isomorphism $U\colon \mathbb{R}^r \to \operatorname{im}(Q)$. Then the forward-backward step $J_A^V(z - V^{-1}Bz)$ can be equivalently expressed by*

$$J_A^V(z - V^{-1}Bz) = J_A^M(z - M^{-1}Bz - \mathrm{s}M^{-1}U\xi^*). \tag{13}$$

Here, $\xi^* \in \mathbb{R}^r$ is the unique zero of $\mathcal{J}\colon \mathbb{R}^r \to \mathbb{R}^r$,

$$\mathcal{J}(\xi^*) \coloneqq U^*(z - J_A^M(z - M^{-1}Bz - \mathrm{s}M^{-1}U\xi^*)) + \xi^* = 0\,. \tag{14}$$

The function $\mathcal{J}$ is Lipschitz continuous with constant $1 + \|M^{-1/2}U\|^2$ and strictly monotone.

Proof. See Appendix A.8. □

3.3 Solving the Root-Finding Problem

The efficiency of the reduction in Theorem 3.1 relies also on ~~the solution of~~solving a root-finding problem which we discuss thoroughly in this subsection. We consider the space $\mathbb{R}^r$ and the root-finding problem with $\ell\colon \mathbb{R}^r \to \mathbb{R}^r$. In several instances, the root-finding problem can be solved exactly, for example, with $A = \partial g$ for special functions g as enumerated in [8, Table 3.1]. In such cases, the root-finding problem simplifies to one involving the proximal operator rather than the resolvent. Similarly, when J_A^M can be represented as a composition of proximal mappings with respect to these special functions, the associated root-finding problem can be exactly solved. In situations where this subproblem cannot be exactly solved, we employ a semi-smooth Newton approach which enjoys local super-linear convergence. To narrow down the neighborhood of the sought root for $r = 1$, we complement the semi-smooth Newton strategy with a bisection method in Section 3.3.2. For cases where $r \geq 1$, a globalization strategy is available as shown in [38].

3.3.1 Semi-smooth Newton Methods

In order to solve $\ell(\alpha) = 0$ in (11) efficiently, we employ a semi-smooth Newton method. A locally Lipschitz function is called semi-smooth if its Clarke Jacobian defines a Newton approximation scheme [18, Definition 7.4.2]. If $\ell(\alpha)$ is semi-smooth and any element of the Clarke Jacobian $\partial^C \ell(\alpha^*)$ is non-singular, then the inexact semi-smooth Newton method outlined in Algorithm 1 (analogous to [8]) can be applied. Semi-smoothness may seem restrictive. However, as shown in [9], the broad class of tame locally Lipschitz functions is semi-smooth. We refer to [9] for the definition of tameness. Therefore, it is sufficient to ensure $\ell(\alpha)$ is tame, which is asserted if the monotone operator A in $\ell(\alpha)$ is a tame map [21]. In this case, the convergence result for Algorithm 1 can be adapted from [8].

Proposition 3.5. *Let $\ell(\alpha)$ be defined as in Theorem 3.1, where A is a set-valued tame mapping. Then $\ell(\alpha)$ is semi-smooth and all elements of $\partial^C \ell(\alpha^*)$ are non-singular where α^* is the unique root of $\ell(\alpha)$ from (11). In turn there exists $\bar{\eta}$ such that if $\eta_k \leq \bar{\eta}$ for every k, there exists a neighborhood of α^* such that for all α_0 in that neighborhood, the sequence generated by Algorithm 1 is well defined and converges to α^* linearly. If $\eta_k \to 0$, the convergence is superlinear.*

Proof. See Appendix A.6. □

Example 3.6. *If f is a tame function and locally Lipschitz, then by [21, Proposition 3.1], ∂f is a tame map.*

Algorithm 1 Semi-smooth Newton method to solve $\ell(\alpha) = 0$

Initialization: A point $\alpha_0 \in \mathbb{R}^r$. N is the maximal number of iterations.
  Update for $k = 0, \ldots, N$:
  if $\ell(\alpha_k) = 0$ **then**
    stop
  else
    Select $G_k \in \partial^C \ell(\alpha_k)$, compute α_{k+1} such that

$$\ell(\alpha_k) + G_k(\alpha_{k+1} - \alpha_k) = e_k\,,$$

  where $e_k \in \mathbb{R}^r$ is an error term satisfying $\|e_k\| \leq \eta_k \|G_k\|$ and $\eta_k \geq 0$.
  end if
  End

Example 3.7. *The assumption that A is a tame mapping is not restrictive. For example, in PDHG setting we have a set-valued operator $A = \begin{pmatrix} \partial g & K^* \\ -K & \partial f \end{pmatrix}$ as defined by (33). If g and f are both tame functions, then ∂g and ∂f are tame as well [21]. As a result, A is a tame mapping.*

3.3.2 Bisection

In the case where $\mathcal{H} = \mathbb{R}^n$ and $r = 1$, we set $U = u \in \mathbb{R}^{n\times 1}$ and solve the root-finding problem $\ell(\alpha) = 0$ via the bisection method in Algorithm 2. A similar bound on the range of α^* as in [8] holds.

Proposition 3.8. *Consider the setting of Corollary 3.3. For $r = 1$, the root α^* of $\ell(\alpha) = 0$ in Corollary 3.3 lies in the set $[-\zeta, \zeta]$, where*

$$\zeta = \|u\|_{V^{-1}}(2\|z\|_V + \|J_A^V(0)\|_V)\,. \tag{15}$$

Moreover, if $V \in \mathcal{S}_c(\mathbb{R}^n)$ and $\|V\| \leq C$, then

$$\zeta = \frac{C}{c}\|u\|(2\|z\| + \|J_A^V(0)\|)\,. \tag{16}$$

Proof. See Appendix A.7. □

Algorithm 2 Bisection method to solve $\ell(\alpha) = 0$ when $r = 1$

Initialization: Tolerance $\epsilon \geq 0$, number of iterates N
  Compute the bound ζ from (15), and set $k = 0$.
  Set $\alpha_- = -\zeta$ and $\alpha_+ = \zeta$.
  Update for $k = 0, \ldots, N$:
  Set $\alpha_k = \frac{1}{2}(\alpha_- + \alpha_+)$.
  if $\ell(\alpha_k) > 0$ **then**
    $\alpha_+ \leftarrow \alpha_k$
  else
    $\alpha_- \leftarrow \alpha_k$
  end if
  if $k > 1$ and $|\alpha_k - \alpha_{k-1}| < \epsilon$ **then**
    return α_k
  end if
  End

Furthermore, we can combine the semi-smooth Newton method with the bisection method. Since the semi-smooth Newton method is locally convergent, it requires a starting point in a sufficiently ~~near~~close neighborhood of the solution. Using the bisection method, we can generate a sequence of points approaching the solution. When these points reach the neighborhood required for convergence of the semi-smooth Newton method, we transition to using the semi-smooth Newton method to achieve faster convergence. In Proposition 3.5, α_0 is required to belong to a neighborhood of α^*, which can be achieved by Algorithm 2, i.e., we can assert that a point α can be found such that $|\alpha - \alpha^*| < \delta$ in $\log_2((2\zeta/\delta))$ steps, where ζ is as in (15).

4 Quasi-Newton FBS for Monotone Inclusions

We consider the monotone inclusion problem in a real Hilbert space $\mathcal{H}$:

$$\text{find } z \in \mathcal{H} \text{ such that } (A+B)z \ni 0\,, \tag{17}$$

where

1. $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is a maximally monotone operator,
2. A is strongly monotone with modulus $\gamma_A \geq 0$,
3. $B\colon \mathcal{H} \to \mathcal{H}$ is a single-valued β-co-coercive operator with $\beta > 0$,
4. B is strongly monotone with modulus $\gamma_B \geq 0$,
5. and $\operatorname{zer}(A+B) \neq \emptyset$.

Note that by setting $\gamma_A = 0$ or $\gamma_B = 0$, we include the general case of monotone operators that are not necessarily strongly monotone. The most common method to solve (17) are FBS-type methods with respect to some metric M, which generate a sequence of points $(z_k)_{k\in\mathbb{N}}$. For instance, the major update step of the classic FBS method is given by

$$z_{k+1} = J_A^{M_k}(z_k - M_k^{-1}Bz_k)\,. \tag{18}$$

Instead of a fixed $M_k \equiv M$, in this paper, we derive FBS-type methods with variable metric M_k. Before introducing our algorithms in detail, we first provide a quasi-Newton framework by which we generate the variable metrics based on the sequence $(z_k)_{k\in\mathbb{N}}$.

4.1 A General 0SR1 Quasi-Newton Metric

In this subsection, we detail a quasi-Newton framework by which we compute the variable metric M_k such that the requirement of applying resolvent calculus is satisfied. We start with the motivation. We note that if we set in the inclusion problem (17) $A \equiv 0$, and $B = \nabla f$ of some convex smooth function f, the update step (18) reduces to Gradient Descent when $M_k \equiv I$ and to the classic Newton method when $M_k = \nabla^2 f(z_k)$. Motivated by classic Newton and quasi-Newton methods, we construct M_k as an approximation of Bz_k at z_k, i.e. we generalize the quasi-Newton method 0SR1 from ∇f (SR1 method with 0-memory) to a co-coercive operator B. The approximation M_k shall satisfy the modified secant condition:

$$M_k s_k = y_k, \quad \text{where} \quad y_k = Bz_k - Bz_{k-1}, \quad s_k = z_k - z_{k-1}\,. \tag{19}$$

Choose $M \in \mathcal{S}_{++}(\mathcal{H})$ to be positive-definite. When $k = 0$, set $M_0 = M$. For $k \geq 1$, update M_k as follows. If $\langle y_k - Ms_k, s_k\rangle = 0$, skip the update of the metric. If $\langle y_k - Ms_k, s_k\rangle \neq 0$, the update is:

$$M_k = M + \mathrm{s}U_k U_k^* \quad \text{with} \quad s = \operatorname{sign}(\langle y_k - Ms_k, s_k\rangle)\,, \tag{20}$$

and $U_k\colon \mathbb{R} \to \mathcal{H}$, $\alpha \mapsto U_k\alpha := \alpha\sqrt{\gamma_k}\hat{u}_k$ is a bounded linear mapping with

$$\hat{u}_k = (y_k - Ms_k)/\sqrt{|\,\langle y_k - Ms_k, s_k\rangle\,|}\,. \tag{21}$$

The parameter $\gamma_k \in [0, +\infty)$ needs to be selected such that M_k is positive-definite.

4.2 Algorithms

We propose two variants of an efficiently implementable quasi-Newton Forward-Backward Splitting (FBS) algorithm. The main update step is a variable metric FBS step of the following form:

$$z_{k+1} = J_A^{M_k}(\bar{z}_k - M_k^{-1}B\bar{z}_k) + \epsilon_k \,,$$

i.e., a forward step with respect to the co-coercive operator B, followed by a proximal point step (computation of the resolvent) of the maximally monotone operator A, both evaluated with the iteration dependent metric M_k from Section 4.1. Both variants account for potential errors in the evaluation of the forward-backward step. In contrast to related works, as discussed in Section 1.1, we emphasize the importance of efficiently implementable resolvent operators (see Section 3).

The two variants allow for more or less flexibility for the choice of the metric. Algorithm 3 combines a FBS step with an additional inertial step which has the potential of accelerating the convergence, as we illustrate in our numerical experiments in Section 6. It is a generalization of the algorithms in [11, 30] to a quasi-Newton variant. In [30], Lorenz and Pock proposed an inertial Forward-Backward Splitting algorithm with a fixed metric, namely, $M_k \equiv \tilde{M}$ for some $\tilde{M} \in \mathcal{S}_{++}(\mathcal{H})$, which is different from our Algorithm 3 where M_k is a variable metric $M_k = M + \mathrm{s}U_kU_k^*$ with $\mathrm{s} \in \{-1, 1\}$. By setting $M = \tilde{M}$ and $U_k \equiv 0$, we can retrieve the algorithm in [30]. Algorithm 4 combines FBS with a relaxation step in (iii) which yields convergence under weak assumptions on the metric. It generalizes the correction step introduced in [20] which can be retrieved by setting $M_k \equiv M$ and $B \equiv 0$.

Algorithm 3 Inertial Quasi-Newton Forward-Backward Splitting to solve (1)

Initialization: $z_0 \in \mathcal{H}$, $N \geq 0$, $M \in \mathcal{S}_{++}(\mathcal{H})$, $(\epsilon_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ with $(\|\epsilon_k\|)_{k\in\mathbb{N}} \in \ell_+^1(\mathbb{N})$

Update for $k = 0, \dots, N$:

(i) Compute $M_k = M + \mathrm{s}U_kU_k^*$ according to the quasi-Newton framework in Section 4.1.

(ii) Compute the inertial step (extrapolation step):

$$\bar{z}_k = z_k + \alpha_k(z_k - z_{k-1}) \,, \tag{22}$$

(iii) and the forward-backward step:

$$z_{k+1} = J_A^M(\bar{z}_k - M^{-1}B\bar{z}_k - \mathrm{s}M^{-1}U_k\xi_k) + \epsilon_k \,, \tag{23}$$

where ξ_k solves $\mathcal{J}(\xi_k) = 0$ and

$$\mathcal{J}(\xi_k) := U_k^*(\bar{z}_k - J_A^M(\bar{z}_k - M^{-1}B\bar{z}_k - \mathrm{s}M^{-1}U_k\xi_k)) + \xi_k \,. \tag{24}$$

End

Algorithm 4 Relaxed Quasi-Newton Forward-Backward Splitting to solve (1)

Initialization: $z_0 \in \mathcal{H}$, $N \geq 0$, $M \in \mathcal{S}_{++}(\mathcal{H})$, $(\epsilon_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ with $(\|\epsilon_k\|)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$

Update for $k = 0, \ldots, N$

(i) Compute $M_k = M + \mathrm{s}U_k U_k^*$ according to the quasi-Newton framework in Section 4.1.

(ii) Compute the forward-backward step:

$$\tilde{z}_k = J_A^M(z_k - M^{-1}Bz_k - \mathrm{s}M^{-1}U_k\xi_k) + \epsilon_k\,, \tag{25}$$

where ξ_k solves $\mathcal{J}(\xi_k) = 0$ and

$$\mathcal{J}(\xi_k) := U_k^*(z_k - J_A^M(z_k - M^{-1}Bz_k - \mathrm{s}M^{-1}U_k\xi_k)) + \xi_k\,, \tag{26}$$

(iii) and relaxation step:

$$t_k = \frac{\langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle}{2\|(M_k - B)(z_k - \tilde{z}_k)\|^2}\,, \tag{27}$$

$$z_{k+1} = z_k - t_k[(M_k - B)(z_k - \tilde{z}_k)]\,. \tag{28}$$

End

4.3 Convergence Guarantees

In this subsection, we prove the convergence of Algorithm 3 and Algorithm 4. The implementation details for the specific quasi-Newton features were discussed in Section 3.

4.3.1 Algorithm 3: Inertial Quasi-Newton Forward-Backward Splitting

The following convergence result is a generalization of [13, Theorem 3.1] to an inertial version of variable metric Forward-Backward Splitting.

Assumption 1. *Let $\sigma \in (0, +\infty)$. $(M_k)_{k\in\mathbb{N}}$ is a sequence in $\mathcal{S}_\sigma(\mathcal{H})$ such that*

$$\begin{cases} C := \sup_{k\in\mathbb{N}}\|M_k\| < \infty\,, \\ (\exists(\eta_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N}))(\forall k \in \mathbb{N}): \quad (1+\eta_k)M_k \succeq M_{k+1}\,, \end{cases} \tag{29}$$

and $M_k - \frac{1}{2\beta}I \in \mathcal{S}_\kappa(\mathcal{H})$ for all $k \in \mathbb{N}$ and some $\kappa > 0$.

Theorem 4.1. *Consider Problem (17) and let the sequence $(z_k)_{k\in\mathbb{N}}$ be generated by Algorithm 3 where Assumption 1 holds. The sequence $(\alpha_k)_{k\in\mathbb{N}}$ is selected such that $\alpha_k \in (0, \Lambda]$ with $\Lambda < \infty$ and*

$$\sum_{k\in\mathbb{N}} \alpha_k \max\{\|z_k - z_{k-1}\|_{M_k}, \|z_k - z_{k-1}\|^2_{M_k}\} < +\infty\,.$$

Then $(z_k)_{k\in\mathbb{N}}$ is bounded and weakly converges to a point $z^* \in \mathrm{zer}(A+B)$, i.e. $z_k \rightharpoonup z^*$ as $k \to \infty$.

Furthermore, if additionally we assume $\epsilon_k \equiv 0$ for any $k \in \mathbb{N}$, $\gamma_A > 0$ or $\gamma_B > 0$ and $M_k - \frac{1}{\beta}I \in \mathcal{S}_\kappa(\mathcal{H})$ for some $\kappa > 0$, then there exist some $\xi \in (0,1)$, some $\Theta > 0$ and some $K_0 \in \mathbb{N}$ such that for any $k > K_0$,

$$\|z_k - z^*\|^2_{M_k} \leq (1-\xi)^{k-K_0}\|z_{K_0} - z^*\|^2_{M_{K_0}} + \sum_{i=K_0}^{k-1}\Theta(1-\xi)^{k-i}\alpha_i \max\{\|z_i - z_{i-1}\|_{M_i}, \|z_i - z_{i-1}\|^2_{M_i}\}\,. \tag{30}$$

Proof. See Appendix A.2. □

Remark 4.2.

- *Using Lemma 3.1 (iv) from [14], we deduce that the second term on the right hand of inequality (30) converges to 0.*
- *The linear convergence factor $1-\xi$ is chosen such that there exists $K_0 \in \mathbb{N}$ with*

$$(1+\eta_k)\left(\frac{1-\frac{\gamma_B}{C}}{1+\frac{\gamma_A}{C}}\right) \leq 1-\xi \quad \textit{for all } k \geq K_0\,.$$

- *The convergence rate for the strongly monotone setting can be influenced by the decay rate of α_k.*
  - *(i) If $\alpha_k = O(q^k)$ for $q = 1 - \xi$, we have convergence rate of $O(kq^k)$ for $k > K_0$ where K_0 is sufficiently large.*
  - *(ii) If $\alpha_k = O(\frac{1}{k^2})$, we have convergence rate of $O(\frac{1}{k})$ for $k > K_0$ where K_0 is sufficiently large.*

Remark 4.3. *In practice, Assumption 1 is hard to verify and restrictive, however, it is a common assumption for variable metric methods [16]. It can be avoided by relaxation methods, which we propose in the next section for this reason.*

In the rest of this subsection, we provide conditions under which M_k defined in (21) meets Assumption 1. We introduce a bounded linear mapping $u_k : \mathbb{R} \to \mathcal{H}$, $\alpha \mapsto \alpha \hat{u}_k$ where $\hat{u}_k \in \mathcal{H}$ is defined by (21). For convenience, we will also equate $U_k U_k^*$ with $\gamma_k u_k u_k^*$.

Lemma 4.4. *Let M be symmetric positive-definite and B is β-co-coercive. Suppose that $M - \frac{1}{\beta} \in \mathcal{S}_c(\mathcal{H})$ for $c > 0$. Let $(\eta_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$.*

(i) Case $M_k = M + \gamma_k u_k u_k^*$: take $\gamma_k = \frac{\eta_k}{\|u_k\|^2}(1/\beta + c)$, and assume that $(\eta_k)_{k\in\mathbb{N}}$ is non-increasing. Then for all $k \in \mathbb{N}$, $M_k - \frac{1}{\beta} \in \mathcal{S}_c(\mathcal{H})$ and M_k verifies Assumption 1.

(ii) Case $M_k = M - \gamma_k u_k u_k^*$: take $\gamma_k = \frac{\kappa\eta_k}{\|u_k\|^2}(1/\beta + c)$, where $\kappa \in]1/(1+\beta c), 1[$, and assume that $\sup_{k\in\mathbb{N}} \eta_k \le 1/\kappa - 1$. Then M_k is positive-definite and satisfies Assumption 1.

Proof. See Appendix A.10. □

Remark 4.5. *The proof of Lemma 4.4 does not rely on co-coercivity of B unlike that of Lemma 4.10. If one uses that property, and more precisely the bounds in (103), then the choice of γ_k can be made independent of u_k in the form $\gamma_k = C\eta_k$, where C is a positive constant that depends only on $(\beta, c, \|M\|)$.*

4.3.2 Algorithm 4: Relaxed Quasi-Newton Forward-Backward Splitting

At the expense of a relaxation instead of an inertial step, we achieve a substantial enhancement in the flexibility of selecting the metric. This method is inspired by [20]. It is worth noting that, without loss of generality, we assume $\tilde{z}_k \neq z_k$ for all $k \in \mathbb{N}$, since otherwise $\tilde{z}_k$ would already solve the inclusion problem. ~~after a finite number of iterations.~~

Assumption 2. *Let $\sigma \in (0, +\infty)$. $(M_k)_{k\in\mathbb{N}}$ is a sequence in $\mathcal{S}_\sigma(\mathcal{H})$ such that:*

(i) For all $k \in \mathbb{N}$, we have $(M_k - \frac{1}{\beta} I) \in \mathcal{S}_c(\mathcal{H})$, for some $c > 0$,

(ii) $C \coloneqq \sup_{k\in\mathbb{N}} \|M_k\| < \infty$.

Theorem 4.6. *Consider Problem (17), and let the sequence $(z_k)_{k\in\mathbb{N}}$ be generated by Algorithm 4 where Assumption 2 holds. Then $(\|z_k - z^*\|)_{k\in\mathbb{N}}$ is bounded for any $z^* \in \operatorname{zer}(A+B)$ and $(z_k)_{k\in\mathbb{N}}$ weakly converges to some $z^* \in \operatorname{zer}(A+B)$, i.e. $z_k \rightharpoonup z^*$ as $k \to \infty$.*
Moreover, if $\epsilon_k \equiv 0$, then $\|z_k - z^*\|$ decreases for any $z^* \in \operatorname{zer}(A+B)$ as $k \to \infty$. Furthermore, if $\gamma_A > 0$ or $\gamma_B > 0$, then z_k converges linearly: there exist some $\xi \in (0,1)$ such that

$$\|z_k - z^*\|^2 \le (1-\xi)^k \|z_0 - z^*\|^2 \,. \tag{31}$$

Proof. See Appendix A.3. □

Remark 4.7. *The linear convergence factor is given by*

$$\xi = \tfrac{\delta}{2} \min\{2(\gamma_A + \gamma_B), c\}\,, \quad \textit{where } \delta = \tfrac{c}{2(C+1/\beta)^2}\,.$$

Remark 4.8. *In [20], the authors studied a relaxed proximal point algorithm for primal-dual splitting with a fixed metric. Our setting and Algorithm 4 here are much ~~broader~~general.*

Remark 4.9. *It is noteworthy that Algorithm 4 can be interpreted as a closed loop system that uses the previous iterates (states) to update the relaxation parameter t_k and the quasi-Newton metric M_k, meaning that the update does not explicitly depend on k.*

In the rest of this subsection, we provide conditions under which M_k defined in (21) meets Assumption 2. We use the same notations as previous.

Lemma 4.10. *Let M be symmetric positive-definite and B is β-co-coercive. Suppose that $M - \frac{1}{\beta} \in \mathcal{S}_c(\mathcal{H})$ for $c > 0$.*

(i) If $M_k = M + \gamma_k u_k u_k^*$, where $0 < \inf_{k\in\mathbb{N}} \gamma_k \leq \sup_{k\in\mathbb{N}} \gamma_k < +\infty$, then for all $k \in \mathbb{N}$, $M_k - \frac{1}{\beta} \in \mathcal{S}_c(\mathcal{H})$ and M_k verifies Assumption 2.

(ii) If $M_k = M - \gamma_k u_k u_k^*$, where $0 < \inf_{k\in\mathbb{N}} \gamma_k \leq \sup_{k\in\mathbb{N}} \gamma_k < \frac{c^2}{(1/\beta + \|M\|)^2}$, then M_k is positive-definite and satisfies Assumption 2.

Proof. See Appendix A.9. □

5 Quasi-Newton PDHG for Saddle-point Problems

In this section, we consider a min-max problem as follows:

$$\min_{x\in\mathcal{H}_1} \max_{y\in\mathcal{H}_2} g(x) + G(x) + \langle Kx, y\rangle - f(y) - F(y) \tag{32}$$

with a linear mapping $K\colon \mathcal{H}_1 \to \mathcal{H}_2$ and proper lower semi-continuous convex functions $g\colon \mathcal{H}_1 \to \overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{+\infty\}$ and $f\colon \mathcal{H}_2 \to \overline{\mathbb{R}}$. Additionally, we consider continuously differentiable convex functions $G\colon \mathcal{H}_1 \to \mathbb{R}$ and $F\colon \mathcal{H}_2 \to \mathbb{R}$ both with Lipschitz continuous gradients. This problem can be solved using Primal-Dual Hybrid Gradient Method (PDHG). Alternatively, ~~the problem~~it can be expressed as a special monotone inclusion problem. By Fermat's rule, the optimality condition for (32) is the following inclusion problem:

$$0 \in Az + Bz \quad \text{with } z = \begin{pmatrix} x \\ y \end{pmatrix}, \text{ where } Az = \begin{pmatrix} \partial g(x) + K^* y \\ -Kx + \partial f(y) \end{pmatrix} \text{ and } Bz = \begin{pmatrix} \nabla G(x) \\ \nabla F(y) \end{pmatrix}. \tag{33}$$

Then, PDHG can be viewed as a FBS method for solving Problem (33) with a specific fixed metric M (see Proposition 5.1).

Proposition 5.1 (PDHG update step as a special FBS update step [20])**.** *The update step of PDHG at some $\bar{z}_k = \begin{pmatrix} \bar{x}_k \\ \bar{y}_k \end{pmatrix} \in \mathcal{H}_1 \times \mathcal{H}_2$ can be regarded as a forward-backward update step with special metric M:*

$$z_{k+1} = J_A^M(\bar{z}_k - M^{-1} B\bar{z}_k), \tag{34}$$

where $z_{k+1} = \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} \in \mathcal{H}_1 \times \mathcal{H}_2$, $M = \begin{pmatrix} \mathcal{T}^{-1} & -K^* \\ -K & \Sigma^{-1} \end{pmatrix}$ with two fixed operators $\mathcal{T} \in \mathcal{S}_{++}(\mathcal{H}_1)$ and $\Sigma \in \mathcal{S}_{++}(\mathcal{H}_2)$ such that $M \succ 0$. The latter is verified when $\|\Sigma^{1/2} K \mathcal{T}^{1/2}\| < 1$.

Remark 5.2. *It is straightforward to verify that*

$$M \succeq (1 - \|\Sigma^{1/2} K \mathcal{T}^{1/2}\|) \begin{pmatrix} \mathcal{T}^{-1} & 0 \\ 0 & \Sigma^{-1} \end{pmatrix}$$

If $\mathcal{T} = \tau I$ and $\Sigma = \sigma I$, we can retrieve a PDHG algorithm with constant stepsizes $\tau > 0$ and $\sigma > 0$. Moreover, $M \succeq (1 - \tau^{1/2}\sigma^{1/2}\|K\|) \min(\tau^{1/2}, \sigma^{1/2})$, and a sufficient condition for $M \succ 0$ is that $\tau\sigma\|K\|^2 < 1$.

Remark 5.3. *In Proposition 5.1, if we set $\bar{z}_k = z_k$, we obtain the classic PDHG update step [11]. Conversely, if we set $\bar{z}_k = z_k + \alpha(z_k - z_{k-1})$ for some $\alpha \in \mathbb{R}_+$, then we obtain inertial PDHG update step [30].*

The interpretation of PDHG in Proposition 5.1 allows us to develop novel quasi-Newton PDHG methods by instantiating the major update steps of Algorithms 3 and 4 with (33) and $M_k = M + \mathrm{s}U_kU_k^*$ where $\mathrm{s} \in \{-1, 1\}$ and U_k as in Section 4.1. Here, M is the fixed metric from Proposition 5.1. For $U_k \in \mathcal{B}(\mathbb{R}^r, \mathcal{H}_1 \times \mathcal{H}_2)$, we set $U_k = \begin{pmatrix} U_{k,x} \\ U_{k,y} \end{pmatrix}$ with $U_{k,x} \in \mathcal{B}(\mathbb{R}^r, \mathcal{H}_1)$ and $U_{k,y} \in \mathcal{B}(\mathbb{R}^r, \mathcal{H}_2)$. To show how to calculate the update step using Proposition 3.4 in quasi-Newton PDHG (Algorithm 5 and Algorithm 6), we introduce Proposition 5.4. In practice, this turns out to be more tractable than other ways to evaluate the variable metric proximal mapping, e.g. using coordinate descent to solve a subproblem which has the same dimension as the original problem [35].

Proposition 5.4. *The update step from* $\bar{z}_k = \begin{pmatrix} \bar{x}_k \\ \bar{y}_k \end{pmatrix}$ *(*$z_k = \begin{pmatrix} x_k \\ y_k \end{pmatrix}$*) to* $z_{k+1} = \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix}$ *(*$\tilde{z}_k = \begin{pmatrix} \tilde{x}_k \\ \tilde{y}_k \end{pmatrix}$*) in Algorithm 5 (Algorithm 6), which is the quasi-Newton PDHG update step, reduces to compute*

$$z_{k+1} = J_A^{M_k}(\bar{z}_k - M_k^{-1}B\bar{z}_k) \iff \begin{cases} x_{k+1} &= J_{\partial g}^{\mathcal{T}^{-1}}(\bar{x}_k - \mathcal{T}\nabla G(\bar{x}_k) - \mathcal{T}K^*\bar{y}_k - \mathrm{s}\mathcal{T}U_{k,x}\xi_k) \\ y_{k+1} &= J_{\partial f}^{\Sigma^{-1}}(\bar{y}_k - \Sigma\nabla F(\bar{y}_k) + \Sigma K(2x_{k+1} - \bar{x}_k) - \mathrm{s}\Sigma U_{k,y}\xi_k)\,. \end{cases} \tag{35}$$

where, $\xi_k \in \mathbb{R}^r$ *is the unique zero of* $\mathcal{J} : \mathbb{R}^r \to \mathbb{R}^r$*:*

$$\begin{aligned} \mathcal{J}(\xi) = {} & (U_{k,x})^*[\bar{x}_k - \underbrace{J_{\partial g}^{\mathcal{T}^{-1}}(\bar{x}_k - \mathcal{T}\nabla G(\bar{x}_k) - \mathcal{T}K^*\bar{y}_k - \mathrm{s}\mathcal{T}U_{k,x}\xi)}_{x_{k+1}(\xi)}] \\ & + (U_{k,y})^*[\bar{y}_k - J_{\partial f}^{\Sigma^{-1}}(\bar{y}_k - \Sigma\nabla F(\bar{y}_k) + \Sigma K(2x_{k+1}(\xi) - \bar{x}_k) - \mathrm{s}\Sigma U_{k,y}\xi)] + \xi\,. \end{aligned} \tag{36}$$

Proof. It is a direct consequence of Proposition 3.4 and Proposition 5.1. ☐

Algorithm 5 Inertial quasi-Newton PDHG to solve (33)

Initialization: $z_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \in \mathcal{H}$, $N \geq 0$, $(\epsilon_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ with $(\|\epsilon_k\|)_{k\in\mathbb{N}} \in \ell_+^1(\mathbb{N})$, $\mathcal{T} \in \mathcal{S}_{++}(\mathcal{H}_1)$, $\Sigma \in \mathcal{S}_{++}(\mathcal{H}_2)$ and $M = \begin{pmatrix} \mathcal{T}^{-1} & -K^* \\ -K & \Sigma^{-1} \end{pmatrix}$

Update for $k = 0, \dots, N$:

(i) Compute 0-SR1 metric $M_k = M + \mathrm{s}U_kU_k^*$ according to the quasi-Newton framework in Section 4.1.

(ii) Compute the inertial step with parameter α_k:

$$\begin{aligned} \bar{x}_k &= x_k + \alpha_k(x_k - x_{k-1})\,, \\ \bar{y}_k &= y_k + \alpha_k(y_k - y_{k-1})\,. \end{aligned} \tag{37}$$

(iii) Compute the main quasi-Newton PDHG step:

$$\begin{aligned} x_{k+1} &= J_{\partial g}^{\mathcal{T}^{-1}}(\bar{x}_k - \mathcal{T}\nabla G(\bar{x}_k) - \mathcal{T}K^*\bar{y}_k - \mathrm{s}\mathcal{T}U_{k,x}\xi_k) + \epsilon_{k,x}\,, \\ y_{k+1} &= J_{\partial f}^{\Sigma^{-1}}(\bar{y}_k - \Sigma\nabla F(\bar{y}_k) + \Sigma K(2x_{k+1} - \bar{x}_k) - \mathrm{s}\Sigma U_{k,y}\xi_k) + \epsilon_{k,y}, \end{aligned} \tag{38}$$

where ξ_k solves $\mathcal{J}(\xi_k) = 0$ (see (36)) and $\epsilon_k = \begin{pmatrix} \epsilon_{k,x} \\ \epsilon_{k,y} \end{pmatrix}$ is the error caused by computation at the k-th iterate.

End

Remark 5.5. *By switching* $\bar{z}_k$ *with* z_k *and* z_{k+1} *with* $\tilde{z}_k$*, we obtain the update step for Algorithm 6.*

Algorithm 6 Relaxed Quasi-Newton PDHG to solve (33)

Initialization: $z_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \in \mathcal{H}$, $N \geq 0$, $(\epsilon_k)_{k\in\mathbb{N}} \subset \mathcal{H}$ with $(\|\epsilon_k\|)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$, $\mathcal{T} \in \mathcal{S}_{++}(\mathcal{H}_1)$, $\Sigma \in \mathcal{S}_{++}(\mathcal{H}_2)$ and $M = \begin{pmatrix} \mathcal{T}^{-1} & -K^* \\ -K & \Sigma^{-1} \end{pmatrix}$

Update for $k = 0, \ldots, N$:

(i) Compute 0-SR1 metric $M_k = M + \mathrm{s}U_kU_k^*$ according to the quasi-Newton framework in Section 4.1.

(ii) Compute the main quasi-Newton PDHG step:

$$\begin{aligned} \tilde{x}_k &= J_{\partial g}^{\mathcal{T}^{-1}}(x_k - \mathcal{T}\nabla G(x_k) - \mathcal{T}K^* y_k - \mathrm{s}\mathcal{T}U_{k,x}\xi_k) + \epsilon_{k,x}\,, \\ \tilde{y}_k &= J_{\partial f}^{\Sigma^{-1}}(y_k - \Sigma\nabla F(y_k) + \Sigma K(2\tilde{x}_k - x_k) - \mathrm{s}\Sigma U_{k,y}\xi_k) + \epsilon_{k,y}\,, \end{aligned} \tag{39}$$

where ξ_k solves $\mathcal{J}(\xi_k) = 0$ and $\epsilon_k = \begin{pmatrix} \epsilon_{k,x} \\ \epsilon_{k,y} \end{pmatrix}$ is the error caused by computation at the k-th iterate.

(iii) Relaxation step:

$$v_k \coloneqq M_k \begin{pmatrix} x_k - \tilde{x}_k \\ y_k - \tilde{y}_k \end{pmatrix} + \begin{pmatrix} \nabla G(\tilde{x}_k) \\ \nabla F(\tilde{y}_k) \end{pmatrix} - \begin{pmatrix} \nabla G(x_k) \\ \nabla F(y_k) \end{pmatrix} \tag{40}$$

to compute the relaxation parameter t_k

$$t_k = \frac{\left\langle \begin{pmatrix} x_k - \tilde{x}_k \\ y_k - \tilde{y}_k \end{pmatrix}, v_k \right\rangle}{2\|v_k\|^2} \tag{41}$$

and update x_k and y_k as follows

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} \leftarrow \begin{pmatrix} x_k \\ y_k \end{pmatrix} - t_k v_k\,. \tag{42}$$

End

Remark 5.6. *In special cases, the root-finding problem can be solved exactly. For instance, when* $\mathcal{T} = \tau I$, $\Sigma = \sigma I$, $K = I$, $g \equiv 0$ *and* $f(y) = \|y\|_1$, *according to [8], the root-finding problem with* $r = 1$ *is piece-wisely linear and can be solved exactly.*

We would like to emphasize that using Proposition 5.4, we can avoid the computation of M_k^{-1} in the primal and dual setting, which is a computationally significant advantage. The convergence of Algorithms 5 and 6 is a direct consequence of Theorems 4.1 and 4.6.

Proposition 5.7 (Convergence of quasi-Newton PDHG method)**.** *Let* $M_0 = M$ *as defined in (34) and* $M_k = M + \mathrm{s}U_kU_k^*$ *for* $k \geq 1$.

(i) If $(M_k)_{k\in\mathbb{N}}$ *satisfies Assumption 1 (see e.g. Lemma 4.4), then* (x_k, y_k) *generated by Algorithm 5 converges weakly to some solution of (3). Furthermore, if* g *and* f *are both strongly convex (or* G *and* F *are both strongly convex), then we obtain the same convergence rate as in Theorem 4.1.*

(ii) If $(M_k)_{k\in\mathbb{N}}$ *satisfies Assumption 2 (see e.g. Lemma 4.10), then* (x_k, y_k) *generated by Algorithm 6 converges weakly to some solution of (3). Furthermore, if* g *and* f *are both strongly convex (or* G *and* F *are both strongly convex) , then we obtain linear convergence.*

6 Numerical experiments

The algorithms that we analyze in the experiments are summarized in Table 1.

Algorithm	Algorithm Name	Metric
FBS	Foward-Backward Primal-Dual Hybrid Gradient Method	M fixed as in (34)
IFBS	Inertial Primal-Dual Hybrid Gradient Method	M fixed as in (34)
QN-FBS	quasi-Newton Primal-Dual Hybrid Gradient Method Gradient	Variable metric as in (21) with ~~$M_0 = M$~~M from (34)
RQN-FBS	Relaxed Quasi-Newton Primal-Dual Hybrid Gradient Method	Variable metric as in (21) with ~~$M_0 = M$~~M from (34)
IQN-FBS	Inertial quasi-Newton Primal-Dual Hybrid Gradient Method	Variable metric as in (21) with ~~$M_0 = M$~~M from (34)

Table 1: Summary of algorithms used in the numerical experiments. Details are provided within each section.

Note that PDHG is used interchangeably as FBS in the later experiments since PDHG is a specialization of FBS.

6.1 TV-l_2 deconvolution

In this experiment, we solve a problem that is used for image deconvolution [10]. Given a blurry and noisy image $b \in \mathbb{R}^{N_x N_y}$ (interpreted as a vector by stacking the N_x columns of length N_y), we seek to find a clean image $x \in \mathbb{R}^{N_x N_y}$ by solving the following optimization problem:

$$\min_{0 \le x \le 255} \frac{1}{2}\|Lx - b\|_2^2 + \mu \|Dx\|_{2,1}\,, \tag{43}$$

where $L \in \mathbb{R}^{N_x N_y \times N_x N_y}$ is a linear operator that acts as a blurring operator and $\|Dx\|_{2,1}$ implements a discrete version of the isotropic total variation norm of x using simple forward differences in the horizontal and vertical directions with $\|D\| \le 2\sqrt{2}$. The parameter $\mu > 0$ stresses the influence of the regularization term $\|Dx\|_{2,1}$ versus the data fidelity term $\frac{1}{2}\|Lx - b\|^2$. ~~In order to deal with~~To address the non-smoothness, we rewrite the problem as a saddle point problem:

$$\min_x \max_y \langle Dx, y\rangle + \delta_\Delta(x) + \frac{1}{2}\|Lx - b\|_2^2 - \delta_{\{\|\cdot\|_{2,\infty} \le \mu\}}(y)\,, \tag{44}$$

where $\Delta := \{x \in \mathbb{R}^{N_x N_y} | 0 \le x_i \le 255, \forall i \in \{1, \ldots, N_x N_y\}\}$. We can cast this problem into the general class of problems (32) by setting $K = D$, $f = \delta_{\{\|\cdot\|_{2,\infty} \le \lambda\}}$, $G(x) = \frac{1}{2}\|Lx - b\|_2^2$, $F(p) = 0$ and $g = \delta_\Delta$. Here, G is $1/\beta$-smooth with $\beta = 2/3$ as we took $\|L\|^2 \le 3/2$. Let $z_k = (x_k, y_k)$ be the primal-dual iterate sequence. Choosing $\mathcal{T} = \tau I = 0.05I$ and $\Sigma = \sigma I = 0.05I$, it follows from Proposition 5.1 and Remark 5.2 that $M - \frac{1}{\beta}I \succeq (16 - 3/2)I > 0$ as desired. We compute the low-rank part $Q_k = \gamma_k u_k u_k^\top$ by (21) with $Bz_k = \begin{pmatrix} L^\top L x_k - L^\top b \\ 0 \end{pmatrix}$ which is of course β-co-coercive. This leads to a metric that affects only the primal update. In each iteration, we use the semi-smooth Newton method (Algorithm 1) to locate the root.

Figure 1 shows the primal gap against the number of iterations and against the time (seconds), where the optimal primal value was computed by running the original PDHG method for 10000 iterations. For the variable metric at iteration k, we fixed $\gamma_k = \min(0.8, 15/\|u_k\|_2^2)$. Thus, Assumption 2 is satisfied and the convergence of RQN-FBS (Algorithm 4) is guaranteed. Although Assumption 1 can be guaranteed by appropriately defining $(\eta_k)_{k\in\mathbb{N}}$ in the view of Lemma 4.4, we choose not to include the condition that $(1+\eta_k)M_k \succeq M_{k+1}$ with $(\eta_k)_{k\in\mathbb{N}} \in \ell_+^1(\mathbb{N})$ in the numerical experiments. This decision ~~comes from the challenge of setting $(\eta_k)_{k\in\mathbb{N}}$ in practice~~was made due to the practical difficulties in setting $(\eta_k)_{k\in\mathbb{N}}$ and the concern that imposing stringent conditions on $(\eta_k)_{k\in\mathbb{N}}$ in advance might negate the benefits of utilizing a variable metric. In this practical problem, we still observe the convergence of IQN-FBS (Algorithm 3). We notice that our quasi-Newton type algorithms IQN-FBS, RQN-FBS and QN-FBS are much faster than original FBS algorithm and inertial FBS (IFBS) according to the left plot in Figure 1. This can be explained by the fact that the Hessian of $G(x)$ is not the identity. By applying our quasi-Newton SR1 methods, we can adapt the metric to the local geometry of the objective. Even though we are concerned about the cost of solving the root-finding problem, the right plot illustrates ~~the additional~~that the extra iterations can pay off , as shown by the convergence results over time. Quasi-Newton type algorithms (QN-FBS, IQN-FBS) achieve higher accuracy compared to FBS and IFBS within the same amount of time.

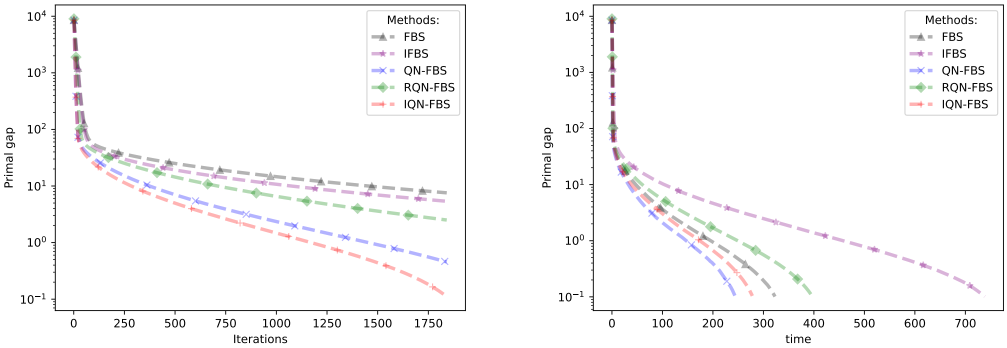

Figure 1: We compare convergence of the inertial quasi-Newton PDHG (IQN-FBS) to other algorithms in the Table 1 with $\mu = 0.0001$, $\tau = 0.05$, $\sigma = 0.05$ and inertial parameter $\alpha_0 = 10$, $\alpha_k = \frac{10}{k^{1.1}(\max\{\|z_k - z_{k-1}\|, \|z_k - z_{k-1}\|^2\})}$. The left plot ~~depicts the convergence against the number of iterations~~illustrates convergence in terms of the number of iterations, while the right plot shows the convergence with respect to time (seconds). We observe that our two quasi-Newton type algorithm QN-, and IQN-FBS significantly outperform the original FBS and IFBS algorithm.

6.2 TV-l_2 deconvolution with infimal convolution type regularization

A source of optimization problems that fits (32) is derived from the following:

$$\min_{x \in \mathbb{R}^n} g(x) + G(x) + (f \Box h)(Dx), \tag{45}$$

where $f \Box h(\cdot) \coloneqq \inf_{v \in \mathbb{R}^m} f(v) + h(\cdot - v)$ denotes the infimal convolution of f and h. As a prototypical image processing problem, we define a regularization term as the infimal convolution between the total variation norm and a weighted squared norm, i.e. $g = 0$, $G(x) = \frac{1}{2}\|Lx - b\|^2$, $h(\cdot) = \frac{1}{2}\|W \cdot\|^2$ and $f(\cdot) = \mu\| \cdot \|_{2,1}$. This yields the problem:

$$\min_x \frac{1}{2}\|Lx - b\|^2 + \mu R(Dx) \tag{46}$$

where $R(\cdot) \coloneqq \inf_{v \in \mathbb{R}^m} \|v\|_{2,1} + \frac{1}{2\mu}\|W(\cdot - v)\|^2$, W is a diagonal matrix of weights which is given to favor discontinuities along image edges and L, b, D are defined as in the first experiment. In practice, W can be computed by additional edge finding steps or by extra information. Here, we select W such that $\frac{1}{6} \leq \|W\|^2 \leq 1$. The optimization problem (46) given in primal type can be converted into the saddle point problem:

$$\min_x \max_y \langle Dx, y \rangle + \frac{1}{2}\|Lx - b\|^2 - \delta_{\{\|\cdot\|_{2,+\infty} \leq \mu\}}(y) - \frac{1}{2}\|W^{-1}y\|^2 . \tag{47}$$

We compute the low-rank part $Q_k = \gamma_k u_k u_k^\top$ by (21) with $Bz_k = \begin{pmatrix} L^\top L x_k - L^\top b \\ (W^{-1})^* W^{-1} y_k \end{pmatrix}$. This approach constructs a metric that affects both the primal and dual update. Here, B is β-co-coercive with $\beta \geq 1/6$. In each iteration, we combine the bisection method (Algorithm 2) with the semi-smooth Newton method (Algorithm 1) to locate the root.

Figure 2 also shows the primal gap where the optimal primal value was computed by running original PDHG for 10000 iterates. For the variable metric at iterate k, we fixed $\gamma_k = 0.64$, $\mathcal{T} = \tau I = 0.01I$ and $\Sigma = \sigma I = 0.01I$. Thus, by Lemma 4.10, Assumption 2 is satisfied. As for Assumption 1, we follow the same strategy as in the previous experiment and drop the condition that $(1 + \eta_k)M_k \succeq M_{k+1}$with $(\eta_k)_{k \in \mathbb{N}} \in \ell_+^1(\mathbb{N})$ in this numerical experiment. We can observe from Figure 2: IQN-FBS is still the fastest one. Moreover, the two quasi-Newton type methods (IQN-FBS and QN-FBS) converge more quickly than IFBS and FBS. Inertial methods (IQN-FBS, IFBS) are slightly faster than QN-FBS and FBS, respectively.

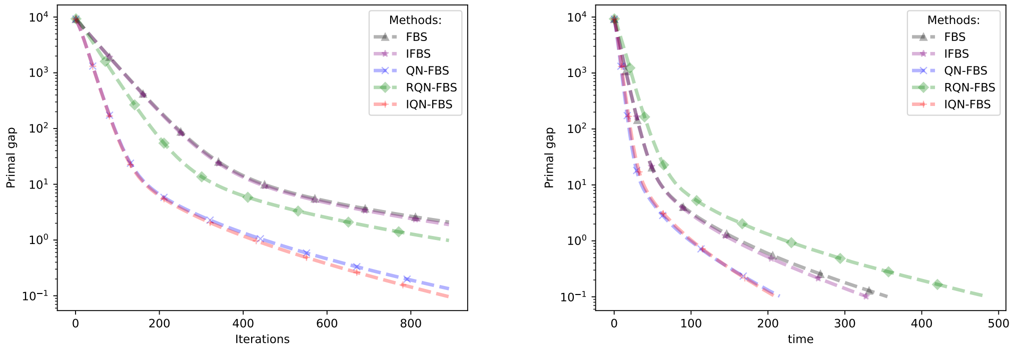

Figure 2: We compare convergence of the inertial quasi-Newton PDHG (IQN-FBS) to other algorithms in the Table 1 with $\mu = 0.5$, $\tau = 0.01I$, $\sigma = 0.01$ and the extrapolation parameter $\alpha_0 = 1$, $\alpha_k = \min\{\frac{10}{k^{1.1}(\max\{\|z_k - z_{k-1}\|, \|z_k - z_{k-1}\|^2\})}, 1\}$. Our quasi-Newton type algorithm IQN-FBS can converge faster than the original FBS and IFBS algorithm.

6.2.1 Image denoising

We consider the same setting as in (47). By setting $L = I$ in (47), we derive an image denoising problem with a special norm defined by infimal convolution of total variation and weighted norm, which has strong convexity for both primal part and dual part. Besides, due to the simple formula, we obtain the dual problem explicitly, allowing us to calculate the primal and dual gap. The dual problem reads

$$\max_{\|y\|_{2,\infty} \le 1} -\frac{1}{2}\|D^* y - b\|^2 - \frac{1}{2}\|W^{-1} y\|^2 . \tag{48}$$

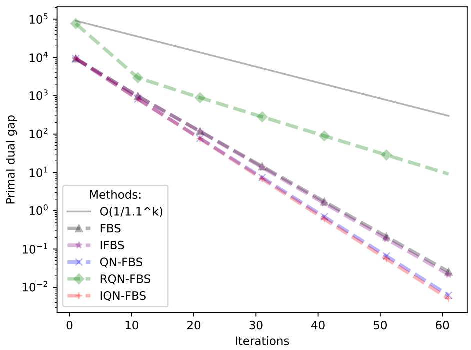

Figure 3: We compare convergence of the inertial quasi-Newton PDHG (IQN-FBS) to other algorithms in the Table 1 with $\mu = 0.1$, $\tau = 0.1$, $\sigma = 0.1$ and extrapolation parameter $\alpha_0 = 10$, $\alpha_k = \frac{10}{\max\{k^{1.1}, k^{1.1}\|z_k - z_{k-1}\|^2\}}$. The plot shows that all algorithms converge linearly and faster than $O(\frac{1}{1.1^k})$.

Figure 3 shows the convergence of the primal-dual gap. For constructing Q_k, we set $\gamma_k = \frac{1}{\|u_k\|^2}$, $\mathcal{T} = \tau I = 0.1I$ and $\Sigma = \sigma I = 0.1I$. Assumption 2 is satisfied. However, Assumption 1 is not satisfied since we do not include the condition that $(1 + \eta_k)M_k \succeq M_{k+1}$with $(\eta_k)_{k \in \mathbb{N}} \in \ell^1_+(\mathbb{N})$ in this numerical experiment. As observed in Figure 3, quasi-Newton type methods exhibit linear convergence as expected from Theorem 4.1 and 4.6. However, Figure 3 shows that in this experiment quasi-Newton type algorithms are in fact not more efficient compared to FBS or IFBS, which is plausible due to the well-conditioned $H = I$.

6.3 Conclusion

In this paper, we extend the framework of [8] for variable metrics to the setting of resolvent operators, solving efficiently the monotone inclusion problem ((1)) consisting of a set-valued operator A and a co-coercive operator B. We propose two variants of quasi-Newton Forward-Backward Splitting. We develop a general efficient resolvent calculus that applies to this quasi-Newton setting. The convergence of the variant with relaxation requires mild assumptions on the metric which are easy to satisfy, whereas the other variant implements an inertial feature and is therefore often fast. As a special case of this framework, we develop an inertial quasi-Newton primal-dual algorithm that can be flexibly applied to a large class of saddle point problems. Throughout the paper, we employ a rank-1 perturbed variable metric denoted as $M_k = M + \mathrm{s}U_k U_k^*$ with $r = 1$ and $U_k : \mathbb{R} \to \mathcal{H}$ which is generated using the 0-memory SR1 method. Alternatively, one can generate the variable metric using an m-memory quasi-Newton method (refer to [23, 41]), wherein $U_k : \mathbb{R}^m \to \mathcal{H}$. Consequently, we are able to derive an m-memory quasi-Newton primal-dual method.

Another potential application of our resolvent calculus in Theorem 3.1 lies in non-diagonal preconditioning of the primal-dual method (PDHG). Moreover, there are many directions to improve our methods. Further investigation is needed to develop an optimal sequence of variable metrics $(M_k)_{k\in\mathbb{N}}$. Given that our variable metrics are currently designed solely based on the geometry of the single valued operator B, it is reasonable to explore a metric that adapts to both operators A and B. Additionally, there remains an open question regarding how to eliminate the condition that the growth of M_k is controlled by a summable sequence $(\eta_k)_{k\in\mathbb{N}}$ while ensuring fast convergence.

Acknowledgement

We sincerely appreciate the valuable feedback provided by the Associate Editor and the two anonymous referees. Our work is supported by the ANR-DFG joint project TRINOM-DS under number ANR-20-CE92-0037-01, and OC150/5-1.

Appendices

A Appendix

A.1 Proof of Lemma 2.4

Proof. This proof is adapted from [6]. Let $(u, v) = (J_A^M(x), J_A^M(y))$ for some $x, y \in \mathcal{H}$. By the definition of resolvent operator J_A^M, we obtain

$$u = J_A^M(x) \iff M(x - u) \in Au\,.$$

Similarly, we obtain $M(y - v) \in Av$. Then γ_A-strong monotonicity of A yields

$$\begin{aligned} \langle M(x-u) - M(y-v), u - v\rangle &\geq \gamma_A \|u - v\|^2 \\ \langle M(x-y), u - v\rangle - \langle M(u-v), u-v\rangle &\geq \gamma_A \|u - v\|^2 \\ \langle M(x-y), u-v\rangle &\geq \gamma_A\|u-v\|^2 + \|u-v\|_M^2\,. \end{aligned}$$

Since $\|M\|$ is bounded by C, we obtain

$$\langle M(x-y), u-v\rangle \geq \gamma_A\|u-v\|^2 + \|u-v\|_M^2 \geq (1 + \tfrac{\gamma_A}{C})\|u-v\|_M^2\,. \tag{49}$$

Consequently, J_A^M is $(1 + \frac{\gamma_A}{C})$-co-coercive in the metric M and Lipschitz continuous with constant $1/(1 + \frac{\gamma_A}{C})$ with respect to the norm $\|\cdot\|_M$. □

A.2 Proof of Theorem 4.1

Proof. For convenience, we set $B_k := M_k^{-1}B$. Fix $z \in \operatorname{zer}(A + B)$ which is equivalent to $z = J_A^{M_k}(z - B_k z)$. We set $\hat{z}_{k+1} := J_A^{M_k}(\bar{z}_k - B_k\bar{z}_k)$, i.e., $z_{k+1} = \hat{z}_{k+1} + \epsilon_k$.
Boundedness:

First, we are going to show that $\|z_k - z\|_{M_k}$ is bounded. The following are several useful estimates we will use later. Assumption 1 yields

$$\|z_{k+1} - z\|^2_{M_{k+1}} \leq (1 + \eta_k)\|z_{k+1} - z\|^2_{M_k}. \tag{50}$$

Since $z_{k+1} = \hat{z}_{k+1} + \epsilon_k$, it follows that

$$\|z_{k+1} - z\|^2_{M_k} = \|\hat{z}_{k+1} - z + \epsilon_k\|^2_{M_k} \leq \|\hat{z}_{k+1} - z\|^2_{M_k} + 2\|\epsilon_k\|_{M_k}\|\hat{z}_{k+1} - z\|_{M_k} + \|\epsilon_k\|^2_{M_k}. \tag{51}$$

The assumption that B is β-co-coercive yields that

$$\langle \bar{z}_k - z, B_k\bar{z}_k - B_k z\rangle_{M_k} = \langle \bar{z}_k - z, B\bar{z}_k - Bz\rangle \geq \beta\|B\bar{z}_k - Bz\|^2. \tag{52}$$

The assumption that $M_k - \frac{1}{2\beta}I \in \mathcal{S}_\kappa(\mathcal{H})$ yields that $2\beta I - M_k^{-1} \in \mathcal{S}_{++}(\mathcal{H})$.
The co-coercivity of B and monotonicity of B yield the following estimates (I) and (II) respectively:

$$\begin{aligned}
\|(\bar{z}_k - B_k\bar{z}_k) - (z - B_k z)\|^2_{M_k} &= \|\bar{z}_k - z\|^2_{M_k} - 2\langle \bar{z}_k - z, B_k\bar{z}_k - B_k z\rangle_{M_k} + \|B_k\bar{z}_k - B_k z\|^2_{M_k} \\
&\overset{\text{(i)}}{=} \|\bar{z}_k - z\|^2_{M_k} - 2\langle \bar{z}_k - z, B\bar{z}_k - Bz\rangle + \|B\bar{z}_k - Bz\|^2_{M_k^{-1}} \\
&\overset{\text{(ii)}}{\leq} \|\bar{z}_k - z\|^2_{M_k} - \|B\bar{z}_k - Bz\|^2_{2\beta - M_k^{-1}} && \text{(I)} \\
&\text{or} \\
&\overset{\text{(iii)}}{\leq} \|\bar{z}_k - z\|^2_{M_k} - \|B\bar{z}_k - Bz\|^2_{\beta - M_k^{-1}} - \gamma_B\|\bar{z}_k - z\|^2, && \text{(II)}
\end{aligned}$$

where (i) uses $\|B_k\bar{z}_k - B_k z\|^2_{M_k} = \|B\bar{z}_k - Bz\|^2_{M_k^{-1}}$, (ii) uses (52) and (iii) uses strong monotonicity of B. Note that we use the shorthand $\beta - M_k^{-1}$ for $\beta I - M_k^{-1}$. The fact that $J_A^{M_k}$ is firmly non-expansive since A is maximally monotone with respect to M_k implies that

$$\begin{aligned}
\|\hat{z}_{k+1} - z\|^2_{M_k} &= \|J_A^{M_k}(\bar{z}_k - B_k\bar{z}_k) - J_A^{M_k}(z - B_k z)\|^2_{M_k} \\
&\leq \|(\bar{z}_k - B_k\bar{z}_k) - (z - B_k z)\|^2_{M_k} \\
&\quad - \|(I - J_A^{M_k})(\bar{z}_k - B_k\bar{z}_k) - (I - J_A^{M_k})(z - B_k z)\|^2_{M_k} \\
&\leq \|\bar{z}_k - z\|^2_{M_k} - \|B\bar{z}_k - Bz\|^2_{2\beta - M_k^{-1}} - \|(\bar{z}_k - \hat{z}_{k+1}) - (B_k\bar{z}_k - B_k z)\|^2_{M_k},
\end{aligned} \tag{53}$$

where the last inequality uses (I). It follows from Assumption 1 that the term $\|B\bar{z}_k - Bz\|^2_{2\beta - M_k^{-1}} \geq 0$. We continue to bound the first term on the right hand side of (53). Using [6, Lemma 2.14] and the definition of $\bar{z}_k$, we obtain the following:

$$\|\bar{z}_k - z\|^2_{M_k} = (1 + \alpha_k)\|z_k - z\|^2_{M_k} - \alpha_k\|z_{k-1} - z\|^2_{M_k} + (1 + \alpha_k)\alpha_k\|z_k - z_{k-1}\|^2_{M_k}, \tag{54}$$

and by using the triangle inequality, we also obtain another estimation:

$$\|\bar{z}_k - z\|_{M_k} \leq (1 + \alpha_k)\|z_k - z\|_{M_k} + \alpha_k\|z_{k-1} - z\|_{M_k}. \tag{55}$$

In order to address complete update step, we make the following estimation:

$$
\begin{aligned}
\|z_{k+1}-z\|^2_{M_{k+1}} &\leq (1+\eta_k)\|z_{k+1}-z\|^2_{M_k} \\
&\overset{\text{(i)}}{\leq} (1+\eta_k)\big(\|\hat{z}_{k+1}-z\|^2_{M_k} + 2\|\epsilon_k\|_{M_k}\|\hat{z}_{k+1}-z\|_{M_k} + \|\epsilon_k\|^2_{M_k}\big) \\
&\overset{\text{(ii)}}{\leq} (1+\eta_k)\big(\|\bar{z}_k-z\|^2_{M_k} + 2\|\epsilon_k\|_{M_k}\|\bar{z}_k-z\|_{M_k} + \|\epsilon_k\|^2_{M_k} \\
&\quad - \|B\bar{z}_k - Bz\|^2_{2\beta-M_k^{-1}} - \|(\bar{z}_k-\hat{z}_{k+1}) - (B_k\bar{z}_k - B_k z)\|^2_{M_k}\big) \\
&\overset{\text{(iii)}}{\leq} (1+\eta_k)\big((1+\alpha_k)\|z_k-z\|^2_{M_k} - \alpha_k\|z_{k-1}-z\|^2_{M_k} \\
&\quad + (1+\alpha_k)\alpha_k\|z_k-z_{k-1}\|^2_{M_k} \\
&\quad + 2\|\epsilon_k\|_{M_k}((1+\alpha_k)\|z_k-z\|_{M_k} + \alpha_k\|z_{k-1}-z\|_{M_k}) \\
&\quad + \|\epsilon_k\|^2_{M_k} - \|B\bar{z}_k - Bz\|^2_{2\beta-M_k^{-1}} - \|(\bar{z}_k-\hat{z}_{k+1}) - (B_k\bar{z}_k - B_k z)\|^2_{M_k}\big) \\
&\overset{\text{(iv)}}{\leq} (1+\eta_k)\big(\|z_k-z\|^2_{M_k} \\
&\quad + \alpha_k(\|z_k-z\|_{M_k} - \|z_{k-1}-z\|_{M_k})(\|z_k-z\|_{M_k} + \|z_{k-1}-z\|_{M_k}) \\
&\quad + (1+\alpha_k)\alpha_k\|z_k-z_{k-1}\|^2_{M_k} \\
&\quad + 2\|\epsilon_k\|_{M_k}((1+\alpha_k)\|z_k-z\|_{M_k} + \alpha_k\|z_{k-1}-z\|_{M_k}) \\
&\quad + \|\epsilon_k\|^2_{M_k} - \|B\bar{z}_k - Bz\|^2_{2\beta-M_k^{-1}} - \|(\bar{z}_k-\hat{z}_{k+1}) - (B_k\bar{z}_k - B_k z)\|^2_{M_k}\big) \\
&\overset{\text{(v)}}{\leq} (1+\eta_k)\big(\|z_k-z\|^2_{M_k} + \alpha_k\|z_k-z_{k-1}\|_{M_k}(\|z_k-z\|_{M_k} + \|z_{k-1}-z\|_{M_k}) \\
&\quad + (1+\alpha_k)\alpha_k\|z_k-z_{k-1}\|^2_{M_k} + 2\|\epsilon_k\|_{M_k}((1+\alpha_k)\|z_k-z\|_{M_k} + \alpha_k\|z_{k-1}-z\|_{M_k}) \\
&\quad + \|\epsilon_k\|^2_{M_k} - \|B\bar{z}_k - Bz\|^2_{2\beta-M_k^{-1}} - \|(\bar{z}_k-\hat{z}_{k+1}) - (B_k\bar{z}_k - B_k z)\|^2_{M_k}\big)\,.
\end{aligned} \tag{56}
$$

where (i) uses (51), (ii) uses (53), (iii) uses (54) and (55), (iv) uses factorization of the quadratic, and (v) uses the triangle inequality to obtain the bound $\|z_k-z\|_{M_k} - \|z_{k-1}-z\|_{M_k} \leq \|z_k-z_{k-1}\|_{M_k}$.

Now, our goal is to conclude boundedness using Lemma 2.7. For simplicity, we set:

$$
\begin{cases}
e_k &:= \alpha_k\|z_k-z_{k-1}\|^2_{M_k} \\
r_k &:= \alpha_k\|z_k-z_{k-1}\|_{M_k} \\
\theta_k &:= \|z_k-z\|_{M_k} \\
m_k &:= \|z_{k-1}-z\|_{M_k} \\
p_k &:= \|B\bar{z}_k - Bz\|^2_{2\beta-M_k^{-1}} \\
q_k &:= \|(\bar{z}_k-\hat{z}_{k+1}) - (B_k\bar{z}_k - B_k z)\|^2_{M_k}\,.
\end{cases}
$$

By Assumption 1, we have $m_k \leq (1+\eta_{k-1})\theta_{k-1}$. Without loss of generality, we can assume $0 < \eta_k < 1$ for any $k \in \mathbb{N}$. Replacing each term in (56) with new corresponding notations, we obtain:

$$
\begin{aligned}
\theta^2_{k+1} &\leq (1+\eta_k)\big(\theta_k^2 + r_k(\theta_k + m_k) \\
&\quad + (1+\alpha_k)e_k + 2\|\epsilon_k\|_{M_k}((1+\alpha_k)\theta_k + \alpha_k m_k) + \|\epsilon_k\|^2_{M_k} - p_k - q_k\big) \\
&\leq (1+\eta_k)\big(\theta_k^2 + r_k(\theta_k + (1+\eta_{k-1})\theta_{k-1}) \\
&\quad + (1+\alpha_k)e_k + 2\|\epsilon_k\|_{M_k}((1+\alpha_k)\theta_k + \alpha_k(1+\eta_{k-1})\theta_{k-1}) + \|\epsilon_k\|^2_{M_k}\big) \\
&\leq (1+\eta_k)\big(\theta_k^2 + r_k(\theta_k + 2\theta_{k-1}) + (1+\Lambda)e_k + 2\|\epsilon_k\|_{M_k}((1+\Lambda)\theta_k + 2\Lambda\theta_{k-1}) + \|\epsilon_k\|^2_{M_k}\big)\,,
\end{aligned} \tag{57}
$$

where the last inequality uses $0 < \alpha_k \leq \Lambda$ and $0 < \eta_k < 1$. Now, we claim that θ_k is bounded in two steps. We introduce an auxiliary bounded sequence $(C_k)_{k\in\mathbb{N}}$ (step 1) such that $\theta_k \leq C_k$ for any $k \in \mathbb{N}$ (step 2). The boundedness of θ_k follows from that of C_k.

Step1: We construct a sequence C_k as the following:

$$
\begin{cases}
C_0 &= \max\{\theta_0, 1\}\,, \\
C_{k+1} &= (1+\eta_k)C_k + \nu_k\,,
\end{cases} \tag{58}
$$

where $\nu_k = (1+\eta_k)((1+\Lambda)e_k + 2r_k + (1+3\Lambda)\|\epsilon_k\|_{M_k})$. From our assumptions, it holds that $(M_k)_{k\in\mathbb{N}}$ is bounded from above, $(r_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N}), (e_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$ and $(\|\epsilon_k\|)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$, which implies $(\nu_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$. Using Lemma 2.7, we obtain the convergence of C_k to some $C_\infty < +\infty$.
Step2: From the update step of (58), we observe that $(C_k)_k$ is a non-decreasing sequence and $C_k \geq 1$ for any $k \in \mathbb{N}$. We claim that for each k, $\theta_k \leq C_k$. We argument by induction. Clearly, we have $\theta_0 \leq C_0$. Assume $\theta_i \leq C_i$ holds true for $i \leq k$. Then, (57) yields that

$$\begin{aligned}\theta_{k+1}^2 &\leq (1+\eta_k)\big(C_k^2 + r_k(C_k + 2C_{k-1}) + (1+\Lambda)e_k + 2\|\epsilon_k\|_{M_k}((1+\Lambda)C_k + 2\Lambda C_{k-1}) + \|\epsilon_k\|^2_{M_k}\big)\\ &\overset{(*)}{\leq} (1+\eta_k)\big(C_k^2 + 4r_kC_k + 2(1+\Lambda)e_kC_k + 2(1+3\Lambda)\|\epsilon_k\|_{M_k}C_k + \|\epsilon_k\|^2_{M_k}\big)\\ &\leq (1+\eta_k)\big(C_k + 2r_k + (1+\Lambda)e_k + (1+3\Lambda)\|\epsilon_k\|_{M_k}\big)^2,\end{aligned} \tag{59}$$

where $(*)$ uses $C_k \geq C_{k-1} \geq 1$ and $r_k > 0$. By the definition of C_{k+1}, we obtain

$$\begin{aligned}\theta_{k+1} &\overset{\text{(i)}}{\leq} \sqrt{1+\eta_k}\big(C_k + 2r_k + (1+\Lambda)e_k + (1+3\Lambda)\|\epsilon_k\|_{M_k}\big)\\ &\overset{\text{(ii)}}{\leq} (1+\eta_k)C_k + (1+\eta_k)(2r_k + (1+\Lambda)e_k + (1+3\Lambda)\|\epsilon_k\|_{M_k})\\ &\overset{\text{(iii)}}{\leq} (1+\eta_k)C_k + \nu_k\\ &= C_{k+1},\end{aligned} \tag{60}$$

where (i) uses (59), (ii) holds true since $(1+\eta_k) > 1$ and (iii) uses definition of ν_k. This concludes the induction, and we deduce that θ_k is bounded and therefore, z_k and $\bar{z}_k$ are both bounded.
Weak convergence:
This part of the proof is adapted from the one for [13, Theorem 4.1]. Since θ_k is bounded, we set $\zeta := \sup_{k\in\mathbb{N}} \theta_k$. The last inequality in (56) implies that

$$\begin{aligned}\theta_{k+1}^2 &\leq (1+\eta_k)\big(\theta_k^2 + r_k(\zeta + 2\zeta) + (1+\Lambda)e_k + 2(1+3\Lambda)\|\epsilon_k\|_{M_k}\zeta + \|\epsilon_k\|^2_{M_k} - p_k - q_k\big)\\ &\leq (1+\eta_k)\big(\theta_k^2 + 3r_k\zeta + (1+\Lambda)e_k + 2(1+3\Lambda)\|\epsilon_k\|_{M_k}\zeta + \|\epsilon_k\|^2_{M_k} - p_k - q_k\big)\\ &\leq \theta_k^2 + \eta_k\theta_k^2 + (1+\eta_k)\big(3r_k\zeta + (1+\Lambda)e_k + (2+6\Lambda)\|\epsilon_k\|_{M_k}\zeta + \|\epsilon_k\|^2_{M_k}\big) - p_k - q_k\\ &\leq \theta_k^2 + \underbrace{\eta_k\zeta^2 + 2(3r_k\zeta + (1+\Lambda)e_k + (2+6\Lambda)\|\epsilon_k\|_{M_k}\zeta + \|\epsilon_k\|^2_{M_k})}_{\delta_k} - p_k - q_k\,.\end{aligned} \tag{61}$$

We set $\delta_k := \eta_k\zeta^2 + 2(3r_k\zeta + (1+\Lambda)e_k + (2+6\Lambda)\|\epsilon_k\|_{M_k}\zeta + \|\epsilon_k\|^2_{M_k})$ and observe that $(\delta_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$. Now, (61) yields that

$$\theta_{k+1}^2 \leq \theta_k^2 + \delta_k\,. \tag{62}$$

Using (62) and Lemma 2.7, we obtain the convergence of $\theta_k^2 = \|z_k - z\|^2_{M_k}$ for any $z \in \operatorname{zer}(A+B)$. Rearranging (61) to $p_k \leq \theta_k^2 - \theta_{k+1}^2 + \delta_k$, using Assumption 1 and summing it for $k = 0, \cdots, N$, there exists some $\varepsilon := \rho_{\min}(2\beta I - M_k^{-1}) > 0$ s.t.

$$\varepsilon\sum_{k=0}^N \|B\bar{z}_k - Bz\|^2 \leq \sum_{k=0}^N \|B\bar{z}_k - Bz\|^2_{2\beta - M_k^{-1}} = \sum_{k=0}^N p_k \leq \theta_0^2 - \theta_N^2 + \sum_{k=0}^N \delta_k \leq \zeta^2 + \sum_{k=0}^N \delta_k\,. \tag{63}$$

Since $(\delta_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$, by taking limit as $N \to +\infty$, we obtain

$$\sum_{k\in\mathbb{N}} \|B\bar{z}_k - Bz\|^2 \leq \frac{1}{\varepsilon^2}(\zeta^2 + \sum_{k\in\mathbb{N}}\delta_k) < +\infty\,. \tag{64}$$

Similarly, we obtain from (61) using $q_k \leq \theta_k^2 - \theta_{k+1}^2 + \delta_k$ that

$$\sum_{k\in\mathbb{N}} \|(\bar{z}_k - \hat{z}_{k+1}) - (B_k\bar{z}_k - B_kz)\|^2_{M_k} < +\infty\,. \tag{65}$$

Set z^* as an arbitrary weak sequential cluster point of $(z_k)_{k\in\mathbb{N}}$, namely, a subsequence $z_{k_n} \rightharpoonup z^*$ as $n \to \infty$.

In order to obtain weak convergence of z_k, by Proposition 2.5 with $\varphi(t) = t^2$, (62) and Assumption 1, it suffices to show that $z^* \in \operatorname{zer}(A+B)$. It follows from the selection of α_k that:

$$\|\bar{z}_k - z_k\| \le \alpha_k \|z_k - z_{k-1}\| \to 0\,. \tag{66}$$

Thus, (66) yields $\bar{z}_{k_n} \rightharpoonup z^*$. From (64), we obtain that $B\bar{z}_{k_n} \to Bz$ as $n \to \infty$. Since B is co-coercive, it is maximally monotone and we can use the weak strong graph closedness of B in Proposition 2.2 to infer that $(z^*, Bz) \in \operatorname{Graph} B$, i.e. $Bz \in Bz^*$. However, since B is single valued, we obtain $Bz^* = Bz$ and hence $B\bar{z}_{k_n} \to Bz^*$. Setting $u_k := M_k(\bar{z}_k - \hat{z}_{k+1}) - B\bar{z}_k$, by definition of the resolvent $J_A^{M_k}$, we have $u_k \in A(\hat{z}_{k+1})$ for all $k \in \mathbb{N}$. From (65), we obtain as $k \to +\infty$,

$$\begin{aligned}\|u_k + Bz^*\| &= \|M_k(\bar{z}_k - \hat{z}_{k+1} - B_k\bar{z}_k + B_k z^*)\| \\ &\le C\|\bar{z}_k - \hat{z}_{k+1} - B_k\bar{z}_k + B_k z^*\| \\ &\le \frac{C}{\sqrt{\sigma}}\|\bar{z}_k - \hat{z}_{k+1} - B_k\bar{z}_k + B_k z^*\|_{M_k} \to 0\,.\end{aligned} \tag{67}$$

Furthermore, from (64) and (65), we have

$$\begin{aligned}\|\bar{z}_k - \hat{z}_{k+1}\| &\le \|\bar{z}_k - \hat{z}_{k+1} - B_k\bar{z}_k + B_k z^*\| + \|B_k\bar{z}_k - B_k z^*\| \\ &\le \|\bar{z}_k - \hat{z}_{k+1} - B_k\bar{z}_k + B_k z^*\| + \frac{1}{\sqrt{\sigma}}\|B\bar{z}_k - Bz^*\| \to 0\,.\end{aligned} \tag{68}$$

Therefore, together with (68), $\bar{z}_{k_n} \rightharpoonup z^*$ implies $\hat{z}_{k_{n+1}} \rightharpoonup z^*$ as $n \to \infty$. Now we already have $u_{k_n} \to -Bz^*$ as $n \to \infty$ and

$$(\forall k \in \mathbb{N}):\quad (\hat{z}_{k_n+1}, u_{k_n}) \in \operatorname{Graph} A\,. \tag{69}$$

Since A is maximally monotone and using Proposition 2.2, we infer that $-Bz^* \in Az^*$, hence $z^* \in \operatorname{zer}(A+B)$. As mentioned above, the result follows from Proposition 2.5 with $\varphi(t) = t^2$.

Convergence rate:

In the following part, we are going to show the convergence rate of Algorithm 3: Assume $\epsilon_k \equiv 0$ for $k \in \mathbb{N}$ and either $\gamma_A > 0$ or $\gamma_B > 0$. Because of (50), Assumption 1 and Lipschitz continuity of $J_A^{M_k}$, we obtain for any $z \in \operatorname{zer}(A+B)$ that

$$\begin{aligned}\|z_{k+1} - z\|_{M_{k+1}}^2 &\le (1+\eta_k)\|z_{k+1} - z\|_{M_k}^2 \\ &\overset{\text{Lemma 2.4}}{\le} (1+\eta_k)\Big(\frac{1}{1+\frac{\gamma_A}{C}}\Big)^2 \|(\bar{z}_k - B_k\bar{z}_k) - (z - B_k z)\|_{M_k}^2 \\ &\overset{\text{(i)}}{\le} (1+\eta_k)\Big(\frac{1}{1+\frac{\gamma_A}{C}}\Big)^2 \big(\|\bar{z}_k - z\|_{M_k}^2 - \|B\bar{z}_k - Bz\|_{\beta - M_k^{-1}}^2 - \gamma_B\|\bar{z}_k - z\|^2\big) \\ &\overset{\text{(ii)}}{\le} (1+\eta_k)\Big(\frac{1}{1+\frac{\gamma_A}{C}}\Big)^2 (1 - \frac{\gamma_B}{C})\big(\|\bar{z}_k - z\|_{M_k}^2\big) \\ &\overset{\text{(iii)}}{=} (1+\eta_k)\frac{(1-\frac{\gamma_B}{C})}{(1+\frac{\gamma_A}{C})^2}\big((1+\alpha_k)\|z_k - z\|_{M_k}^2 - \alpha_k\|z_{k-1} - z\|_{M_k}^2 \\ &\quad + (1+\alpha_k)\alpha_k\|z_k - z_{k-1}\|_{M_k}^2\big) \\ &= (1+\eta_k)\frac{(1-\frac{\gamma_B}{C})}{(1+\frac{\gamma_A}{C})^2}\big(\|z_k - z\|_{M_k}^2 + \alpha_k(\|z_k - z\|_{M_k}^2 - \|z_{k-1} - z\|_{M_k}^2) \\ &\quad + (1+\alpha_k)\alpha_k\|z_k - z_{k-1}\|_{M_k}^2\big) \\ &\overset{\text{(iv)}}{\le} (1+\eta_k)\frac{(1-\frac{\gamma_B}{C})}{(1+\frac{\gamma_A}{C})^2}\big(\|z_k - z\|_{M_k}^2 + \alpha_k\|z_k - z_{k-1}\|_{M_k}(\|z_k - z\|_{M_k} \\ &\quad + \|z_{k-1} - z\|_{M_k}) + (1+\alpha_k)\alpha_k\|z_k - z_{k-1}\|_{M_k}^2\big) \\ &\overset{\text{(v)}}{\le} (1+\eta_k)\frac{(1-\frac{\gamma_B}{C})}{(1+\frac{\gamma_A}{C})^2}\big(\|z_k - z\|_{M_k}^2 + \alpha_k\|z_k - z_{k-1}\|_{M_k}(\|z_k - z\|_{M_k} \\ &\quad + (1+\eta_{k-1})\|z_{k-1} - z\|_{M_{k-1}}) + (1+\alpha_k)\alpha_k\|z_k - z_{k-1}\|_{M_k}^2\big) \\ &\overset{\text{(vi)}}{=} (1+\eta_k)\frac{(1-\frac{\gamma_B}{C})}{(1+\frac{\gamma_A}{C})^2}\big(\|z_k - z\|_{M_k}^2 + 3\alpha_k\zeta\|z_k - z_{k-1}\|_{M_k} \\ &\quad + (1+\Lambda)\alpha_k\|z_k - z_{k-1}\|_{M_k}^2\big)\end{aligned} \tag{70}$$

where (i) uses (II), (ii) uses the fact M_k is bounded uniformly and the assumption that $M_k - \frac{1}{\beta}I \in \mathcal{S}_\kappa(\mathcal{H})$, (iii) uses (54), (iv) uses factorization of the quadratic and uses the triangle inequality to obtain the bound $\|z_k - z\|_{M_k} - \|z_{k-1} - z\|_{M_k} \le \|z_k - z_{k-1}\|_{M_k}$, (v) uses Assumption 1 and (vi) uses boundedness of α_k and $\|z_k - z\|_{M_k}$. Since either $\gamma_A > 0$ or $\gamma_B > 0$, $\frac{1-\frac{\gamma_B}{C}}{(1+\frac{\gamma_A}{C})^2} < 1$. Then there exists sufficient large $K_0 > 0$ such that for any $k > K_0$, $(1+\eta_k)\Big(\frac{1-\frac{\gamma_B}{C}}{(1+\frac{\gamma_A}{C})^2}\Big) < 1 - \xi < 1$ for some $\xi \in (0,1)$. Thus, we infer that for any $k > K_0$:

$$\|z_k - z\|^2_{M_k} \le (1-\xi)^{k-K_0}\|z_{K_0} - z\|^2_{M_{K_0}} + \sum_{i=K_0}^{k-1}(1-\xi)^{k-i}\alpha_i(3\zeta\|z_i - z_{i-1}\|_{M_i} + (1+\Lambda)\|z_i - z_{i-1}\|^2_{M_i}) . \tag{71}$$

Let $\Theta = 3\zeta + (1+\Lambda)$. Therefore (71) can be simplified as the following:

$$\|z_k - z\|^2_{M_k} \le (1-\xi)^{k-K_0}\|z_{K_0} - z\|^2_{M_{K_0}} + \sum_{i=K_0}^{k-1}\Theta(1-\xi)^{k-i}\alpha_i \max\{\|z_i - z_{i-1}\|_{M_i}, \|z_i - z_{i-1}\|^2_{M_i}\} . \tag{72}$$

Since z_k is bounded and $M_k \in S_\sigma(\mathcal{H})$, it follows that

$$\|z_k - z\|^2 \le \tfrac{1}{\sigma}(1-\xi)^{k-K_0}\|z_{K_0} - z\|^2_{M_{K_0}} + O(\sum_{i=K_0}^{k-1}(1-\xi)^{k-i}\alpha_i) . \tag{73}$$

Furthermore, if $\alpha_i \equiv 0$, for $k > K_0$ and for any $z \in \operatorname{zer}(A+B)$, we obtain linear convergence:

$$\|z_{k+1} - z\|^2 \le \frac{1}{\sigma}(1-\xi)^{k-K_0}\|z_{K_0} - z\|^2_{M_{K_0}} . \tag{74}$$

If $\alpha_k \ne 0$, $\alpha_k = O(\frac{1}{k^2})$ and K_0 large enough, then $\|z_{k+1} - z\|^2$ converges in the rate of $O(\frac{1}{k})$ for $k > K_0$ according to [33, Lemma 2.2.4 (Chung)]; if $\alpha_k \ne 0$ and $\alpha_k = O(q^k)$ for $q = 1 - \xi$ and $k > K_0$, then $\|z_{k+1} - z\|^2$ converges in the rate of $O(kq^k)$ for $k > K_0$ since (73).

□

A.3 Proof of Theorem 4.6

Proof. For simplicity, we set

$$\begin{cases} \hat{z}_k := J_A^{M_k}(z_k - M_k^{-1}Bz_k), \text{and, } \tilde{z}_k = \hat{z} + \epsilon_k , \\ \delta_k := \max\{1, \rho\}\|\epsilon_k\| , \end{cases} \tag{75}$$

where $\rho = \frac{1}{2}\sqrt{\frac{C}{\sigma c}}(C + \frac{1}{\beta})$ and $(\delta_k)_{k\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$. We set z^* such that $-Bz^* \in Az^*$.
Boundedness:
We claim that by choosing proper t_k for each $k \in \mathbb{N}$, we have

$$\|z_{k+1} - z^*\| \le \|z_k - z^*\| + \delta_k. \tag{76}$$

Note that if (76) is satisfied, it follows from Lemma 2.7 that $\|z_k - z^*\|$ is bounded and converges.

To stress the relation between z_{k+1} and t_k, we define $z(t) := z_k - t[(M_k - B)(z_k - \tilde{z}_k)]$ and we will use z_{k+1} and $z(t_k)$ interchangeably. We also set $\gamma_k(t) := (\delta_k + \|z_k - z^*\|)^2 - \|z(t) - z^*\|^2$.

In order to prove (76), it is sufficient to show that for proper t_k at each iterate, $\gamma_k(t_k) > 0$. It results from the definition of $\gamma_k(t_k)$ that

$$\begin{aligned} \gamma_k(t) &= (\|z_k - z^*\| + \delta_k)^2 - \|z(t) - z^*\|^2 \\ &= \langle z_k - z^* + z(t) - z^*, z_k - z(t)\rangle + 2\delta_k\|z_k - z^*\| + \delta_k^2 \\ &= \underbrace{2t\,\langle z_k - z^*, (M_k - B)(z_k - \tilde{z}_k)\rangle}_{\text{(I)}} - \underbrace{t^2\|(M_k - B)(z_k - \tilde{z}_k)\|^2}_{\text{(II)}} \\ &\quad + 2\delta_k\|z_k - z^*\| + \delta_k^2 . \end{aligned} \tag{77}$$

We need several useful properties to estimate (I) and (II).

It follows from the definition of $\hat{z}_k$ that $-M_k(\hat{z}_k - z_k) - Bz_k \in A\hat{z}_k$, and γ_A-strong monotonicity of A implies that

$$\begin{aligned}\gamma_A\|\hat{z}_k - z^*\|^2 &\leq \langle \hat{z}_k - z^*, -M_k(\hat{z}_k - z_k) - Bz_k + Bz^*\rangle \\ &= \langle \hat{z}_k - z^*, M_k(z_k - \hat{z}_k) - Bz_k + B\hat{z}_k - B\hat{z}_k + Bz^*\rangle \\ &= \langle \hat{z}_k - z^*, M_k(z_k - \hat{z}_k) - Bz_k + B\hat{z}_k\rangle - \langle \hat{z}_k - z^*, B\hat{z}_k - Bz^*\rangle \ .\end{aligned} \tag{78}$$

We deduce by (78) and γ_B-strong monotonicity of B that

$$\begin{aligned}\langle \hat{z}_k - z^*, M_k(z_k - \hat{z}_k) - Bz_k + B\hat{z}_k\rangle &\geq \langle \hat{z}_k - z^*, B\hat{z}_k - Bz^*\rangle + \gamma_A\|\hat{z}_k - z^*\|^2 \\ &\geq \gamma_B\|\hat{z}_k - z^*\|^2 + \gamma_A\|\hat{z}_k - z^*\|^2 .\end{aligned} \tag{79}$$

$J_A^{M_k}$ is Lipschitz since A is monotone. Therefore,

$$\begin{aligned}\|\hat{z}_k - z^*\|_{M_k}^2 &\leq \|(z_k - B_k z_k) - (z^* - B_k z^*)\|_{M_k}^2 \\ &= \|z_k - z^*\|_{M_k}^2 - 2\langle z_k - z^*, B_k z_k - B_k z^*\rangle_{M_k} + \|B_k z_k - B_k z^*\|_{M_k}^2 \\ &\leq \|z_k - z^*\|_{M_k}^2 - \|Bz_k - Bz\|_{2\beta - M_k^{-1}} \\ &\overset{(*)}{\leq} \|z_k - z^*\|_{M_k}^2 ,\end{aligned} \tag{80}$$

where $(*)$ uses Assumption 2 that $M_k - \frac{1}{\beta}I \in \mathcal{S}_c(\mathcal{H})$. Using the assumption that $M_k - \frac{1}{\beta}I \in \mathcal{S}_c(\mathcal{H})$ again, we obtain

$$\langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle \geq \|z_k - \tilde{z}_k\|_{M_k - \frac{1}{\beta}I}^2 > 0 . \tag{81}$$

Combining (79), (80), the first term (I) in (77) can be estimated by the following:

$$\begin{aligned}\text{(I)} &= 2t\langle z_k - z^*, (M_k - B)(z_k - \tilde{z}_k)\rangle \\ &= 2t\langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle + 2t\langle \tilde{z}_k - \hat{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle \\ &\quad + 2t\langle \hat{z}_k - z^*, (M_k - B)(z_k - \hat{z}_k)\rangle + 2t\langle \hat{z}_k - z^*, (M_k - B)(\hat{z}_k - \tilde{z}_k)\rangle \\ &\overset{\text{(i)}}{\geq} 2t\langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle + 2t\langle \tilde{z}_k - \hat{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle \\ &\quad + 2t(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 + 2t\langle \hat{z}_k - z^*, (M_k - B)(\hat{z}_k - \tilde{z}_k)\rangle \\ &\overset{\text{(ii)}}{\geq} 2t\langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle - 2t\|\epsilon_k\|\,\|(M_k - B)(z_k - \tilde{z}_k)\| \\ &\quad + 2t(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 - 2t\|(M_k - B)\epsilon_k\|_{M_k^{-1}}\|\hat{z}_k - z^*\|_{M_k} \\ &\overset{\text{(iii)}}{\geq} 2t\langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle - \|\epsilon_k\|^2 - t^2\|(M_k - B)(z_k - \tilde{z}_k)\|^2 \\ &\quad + 2t(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 - 2\tfrac{1}{\sqrt{\sigma}}(C + \tfrac{1}{\beta})t\|\epsilon_k\|\,\|z_k - z^*\|_{M_k} \\ &\overset{\text{(iv)}}{\geq} 2t\langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle - \|\epsilon_k\|^2 - t^2\|(M_k - B)(z_k - \tilde{z}_k)\|^2 \\ &\quad + 2t(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 - 2\sqrt{\tfrac{C}{\sigma}}(C + \tfrac{1}{\beta})t\|\epsilon_k\|\,\|z_k - z^*\| ,\end{aligned} \tag{82}$$

where (i) uses (79), (ii) uses Cauchy inequality and (iii) uses (80), $2ab \leq a^2 + b^2$ and Assumption 2. We set $b_k := \langle z_k - \tilde{z}_k, (M_k - B)(z_k - \tilde{z}_k)\rangle$ and $a_k := \|(M_k - B)(z_k - \tilde{z}_k)\|^2$. The definition of δ_k yields that $\delta_k \geq \|\epsilon_k\|^2$ and it follows from (82) that:

$$\begin{aligned}\gamma_k(t) &\geq 2tb_k - 2t^2a_k + 2\delta_k\|z_k - z^*\| + \delta_k^2 - \|\epsilon_k\|^2 \\ &\quad - 2\sqrt{\tfrac{C}{\sigma}}(C + \tfrac{1}{\beta})t\|\epsilon_k\|\,\|z_k - z^*\| + 2t(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 \\ &\geq \underbrace{2tb_k - 2t^2a_k}_{\text{(III)}} + 2t(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 + \underbrace{2(\delta_k - \sqrt{\tfrac{C}{\sigma}}(C + \tfrac{1}{\beta})t\|\epsilon_k\|)\|z_k - z^*\|}_{\text{(IV)}} .\end{aligned} \tag{83}$$

We continue to find a proper t_k such that $\gamma_k(t_k) > 0$. This goal boils down to ensuring both (III) and (IV) are positive.

Let $t_k = \frac{b_k}{2a_k}$. We first show $t_k = \frac{b_k}{2a_k}$ is the proper value to make sure (III) is positive. From (81), we observe that $b_k > 0$. Since $a_k > 0$ and $b_k > 0$, the quadratic term (III) in (83) will be zero for $t_k = 0$ or $t_k = \frac{b_k}{a_k}$ and will be strictly positive for any $t_k \in (0, b_k/a_k)$ with the maximum value obtained at $t_k = \frac{b_k}{2a_k}$. As a result, (III) is strictly positive.

Second, we will show (IV) is positive when $t_k = \frac{b_k}{2a_k}$. We observe that $0 < t_k = \frac{b_k}{2a_k} < \frac{1}{2\sqrt{c}}$ for Assumption 2. Thus, the definition of δ_k and $t_k = \frac{b_k}{2a_k}$ imply that

$$
\begin{aligned}
\text{(IV)} &= 2(\delta_k - \sqrt{\tfrac{C}{\sigma}}(C + \tfrac{1}{\beta})t_k\|\epsilon_k\|)\|z_k - z^*\| \\
&\geq (2\delta_k - \sqrt{\tfrac{C}{\sigma c}}(C + \tfrac{1}{\beta})\|\epsilon_k\|)\|z_k - z^*\| \\
&\geq 0\,.
\end{aligned} \tag{84}
$$

Since (III) and (IV) both are positive when $t_k = \frac{b_k}{2a_k}$, (83) yields:

$$
\begin{aligned}
\gamma_k(t_k) &\geq 2t_k b_k - 2t_k^2 a + 2t_k(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 + (2\delta_k - \sqrt{\tfrac{C}{\sigma c}}(C + \tfrac{1}{\beta})\|\epsilon_k\|)\|z_k - z^*\| \\
&\geq \frac{b_k^2}{2a_k} + 2t_k(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2 \\
&> 0\,,
\end{aligned} \tag{85}
$$

It results from (85) and the definition of $\gamma_k(t_k)$ that for each $k \in \mathbb{N}$, $(\|z_k - z^*\| + \delta_k) \geq \|z(t_k) - z^*\| = \|z_{k+1} - z^*\|$. We conclude that if $t_k = \frac{b_k}{2a_k}$, then $\gamma_k(t_k) > 0$ for all $k \in \mathbb{N}$ and the sequence $\|z_k - z^*\|$ is bounded and converges as $k \to +\infty$ by using Lemma 2.7.

Weak convergence:

The sequence $(z_k)_{k\in\mathbb{N}}$ generated by Algorithm 4 is bounded and $\|z_k - z\|$ converges as $k \to \infty$ and $\gamma_k(t_k)$ converges to zero as $k \to \infty$ for all $z \in \operatorname{zer}(A + B)$. Set z^* as an arbitrary weak sequential cluster point of $(z_k)_{k\in\mathbb{N}}$ and there exists a subsequence $(z_{k_n})_{n\in\mathbb{N}}$ such that $z_{k_n} \rightharpoonup z^*$.

In order to obtain weak convergence of z_k, by Proposition 2.5 with $\varphi(t) = t$ and fixed metric $M_k = I$ and (76), it suffices to show that $z^* \in \operatorname{zer}(A + B)$.

Using Assumption 2, $(M_k)_{k\in\mathbb{N}}$ is bounded uniformly by C. Together with boundedness of operator B, we obtain a_k is bounded by $(C + \frac{1}{\beta})^2\|z_k - \tilde{z}_k\|^2$ for each $k \in \mathbb{N}$. Using Assumption 2, we obtain that $b_k \geq \|z_k - \tilde{z}_k\|^2_{M_k - \frac{1}{\beta}I} \geq c\|z_k - \tilde{z}_k\|^2$. By definition of t_k and (85), we have $\gamma_k(t_k) \geq \frac{b_k^2}{2a_k} \geq \frac{c^2}{2(C+\frac{1}{\beta})^2}\|z_k - \tilde{z}_k\|^2$. Since $\gamma_k(t_k) \to 0$, $\|z_k - \tilde{z}_k\| \to 0$ as $k \to \infty$. Moreover, since $\epsilon_k \to 0$, $\|z_k - \hat{z}_k\| \to 0$. We set $u_k := M_k(z_k - \hat{z}_k) + B\hat{z}_k - Bz_k$. Therefore, we obtain that $u_k \to 0$ as $k \to +\infty$ and $\hat{z}_{k_n} \rightharpoonup z^*$ as $n \to +\infty$. We observe that $u_k \in A\hat{z}_k + B\hat{z}_k$. Then, by using Proposition 2.2 and the fact that $A + B$ is maximally monotone, we conclude that $0 \in Az^* + Bz^*$. Besides, $\|z_k - z^*\|$ decreases since $z^* \in \operatorname{zer}(A + B)$. As mentioned above, the result follows from Proposition 2.5 with $\varphi(t) = t$.

Linear convergence rate:

If we assume $\epsilon_k \equiv 0$, then $\hat{z}_k = \tilde{z}_k$ and $\delta_k \equiv 0$. Therefore, from (85) we can obtain an estimation for $\gamma_k(t_k)$ when $t_k = \frac{b_k}{2a_k}$:

$$
\gamma_k(t_k) = \|z_k - z^*\|^2 - \|z_{k+1} - z^*\|^2 \geq \frac{b_k^2}{2a_k} + 2t_k(\gamma_A + \gamma_B)\|\hat{z}_k - z^*\|^2\,. \tag{86}
$$

The following part is to derive linear convergence for the case that either $\gamma_A > 0$ or $\gamma_B > 0$. The definition of b_k and that of a_k yield the following estimation for t_k:

$$
t_k = \frac{b_k}{2a_k} = \frac{\langle z_k - \hat{z}_k, (M_k - B)(z_k - \hat{z}_k)\rangle}{2\|(M_k - B)(z_k - \hat{z}_k)\|^2} \overset{\text{(i)}}{\geq} \frac{\|z_k - \hat{z}_k\|^2_{M_k - \frac{1}{\beta}}}{2(C + \frac{1}{\beta})^2\|z_k - \hat{z}_k\|^2} \overset{\text{(ii)}}{>} \frac{c}{2(C + \frac{1}{\beta})^2}, \tag{87}
$$

where both (i) and (ii) use Assumption 2. For convenience, we denote $\frac{c}{2(C+\frac{1}{\beta})^2}$ by δ. Using Assumption 2 again, we have the estimation for the first term at the right hand side of (86):

$$
\frac{b_k^2}{2a_k} \geq \delta\|z_k - \hat{z}_k\|^2_{M_k - \frac{1}{\beta}I} > c\delta\|z_k - \hat{z}_k\|^2\,. \tag{88}
$$

Furthermore, combining (86) with (87), (88) and the definition $\gamma_k(t_k) := \|z_k - z^*\|^2 - \|z_{k+1} - z^*\|^2$, we obtain

$$\begin{aligned}
\|z_{k+1} - z^*\|^2 &\le \|z_k - z^*\|^2 - 2(\gamma_A + \gamma_B)\delta\|\hat{z}_k - z^*\|^2 - c\delta\|z_k - \hat{z}_k\|^2 \\
&\le \|z_k - z^*\|^2 - \tfrac{1}{2}\min\{2(\gamma_A + \gamma_B)\delta, c\delta\}(2\|\hat{z}_k - z^*\|^2 + 2\|z_k - \hat{z}_k\|^2) \\
&\overset{\text{(i)}}{\le} \|z_k - z^*\|^2 - \tfrac{1}{2}\min\{2(\gamma_A + \gamma_B)\delta, c\delta\}(\|\hat{z}_k - z^*\| + \|z_k - \hat{z}_k\|)^2 \\
&\overset{\text{(ii)}}{\le} \|z_k - z^*\|^2 - \tfrac{1}{2}\min\{2(\gamma_A + \gamma_B)\delta, c\delta\}\|z_k - z^*\|^2 \\
&\le (1 - \tfrac{1}{2}\min\{2(\gamma_A + \gamma_B)\delta, c\delta\})\|z_k - z^*\|^2 \,,
\end{aligned} \tag{89}$$

where (i) uses inequality $2x^2 + 2y^2 \ge (x + y)^2$ and (ii) uses triangle inequality. Consequently, we obtain linear convergence if $(\gamma_A + \gamma_B) > 0$:

$$\|z_k - z^*\|^2 \le (1 - \xi)^k \|z_0 - z^*\|^2 \,, \tag{90}$$

where $\xi = \frac{1}{2}\min\{2(\gamma_A + \gamma_B)\delta, c\delta\} > 0$. □

A.4 Proof of Theorem 3.1

Now, we give the proof of Theorem 3.1. For convenience, we define translation operator $\tau_p \colon \mathcal{H} \to \mathcal{H}$ by $\tau_p(x) = x - p$ with inverse $\tau_p^{-1} = \tau_{-p}$.

Proof. Computing the resolvent operator shows the following equivalences

$$\begin{aligned}
& x^* = J_A^V(z) = (I + V^{-1}A)^{-1}(z) \\
&\iff Vz \in (V + A)(x^*) \\
&\iff Mz \in (M + A)(x^*) + \mathrm{s}Q(x^* - z) \\
[y^* = M^{1/2}x^*] &\iff Mz \in (M + A)(M^{-1/2}y^*) + \mathrm{s}Q(M^{-1/2}y^* - z) \\
&\iff Mz \in (M + A)(M^{-1/2}y^*) + \mathrm{s}QM^{-1/2}(y^* - M^{1/2}z) \\
&\iff Mz \in (M^{1/2} + AM^{-1/2})(y^*) + \mathrm{s}QM^{-1/2}(y^* - M^{1/2}z) \\
[W = M^{-1/2}QM^{-1/2}] &\iff M^{1/2}z \in (I + M^{-1/2}AM^{-1/2})(y^*) + \mathrm{s}W(y^* - M^{1/2}z) \,.
\end{aligned} \tag{91}$$

Since uniqueness and existence of x^* is guaranteed by the properties of J_A^V, Lemma 2.6 yields the existence of a unique primal-dual pair (x^*, u^*) that satisfies the equivalent relations in Lemma 2.6 with $R := +\mathrm{s}W \circ \tau_{M^{1/2}z}$ and $T := \tau_{M^{1/2}z} \circ (I + M^{-1/2}AM^{-1/2})$. The mapping R is single-valued and, as A is a maximally monotone operator and M is positive-definite, $T^{-1} = J_{M^{-1/2}TM^{-1/2}} \circ \tau_{-M^{1/2}z}$ is single valued. Therefore, the solution of J_A^V can be computed by finding $u^* \in \mathrm{im}(W)$, namely, $u^* \in \mathrm{im}(M^{-1/2}Q)$, such that

$$0 \in R^{-1}u^* - T^{-1}(-u^*) = [(\mathrm{s}W)^{-1}u^* + M^{1/2}z] - J_{M^{-1/2}AM^{-1/2}}(M^{1/2}z - u^*) \,, \tag{92}$$

and the using

$$x^* = M^{-1/2}y^* \quad \text{and} \quad y^* = T^{-1}(-u^*) = J_{M^{-1/2}AM^{-1/2}}(M^{1/2}z - u^*) \,. \tag{93}$$

Substituting $u^* = \mathrm{s}M^{-1/2}v^* \in \mathrm{im}(M^{-1/2}Q)$ in both problems, multiplying the former one from left with $M^{-1/2}$, and using $M^{-1/2}W^{-1}M^{-1/2} = Q^{-1}$ where Q^{-1} is a set-valued inverse operator of Q defined by the graph $\{(v, w) \in \mathrm{im}(Q) \times \mathcal{H} | Qv = w\}$ leads to

$$\begin{cases} 0 \in Q^{-1}v^* + z - M^{-1/2} \circ J_{M^{-1/2}AM^{-1/2}} \circ M^{1/2}(z - \mathrm{s}M^{-1}v^*) \\ x^* = M^{-1/2} \circ J_{M^{-1/2}AM^{-1/2}} \circ M^{1/2}(z - \mathrm{s}M^{-1}v^*) \,. \end{cases} \tag{94}$$

Since a unique solution to J_A^V exists, there exists $v^* \in \mathrm{im}(Q)$ that satisfies the inclusion. We notice that $\mathrm{im}(Q^{-1}) = \mathrm{im}(Q^+) + \ker Q$ where $Q^+ \colon \mathrm{im}(Q) \to \mathrm{im}(Q)$ is the inverse of Q restricted on $\mathrm{im}(Q)$ and $\ker(Q)$ denotes the kernel of Q. Given the linear mapping $U \colon \mathbb{R}^r \to \mathrm{im}(Q)$, which can be realized using r linearly independent $u_1, \cdots, u_r \in \mathcal{H}$ by $\alpha \to \sum_{i=1}^r \alpha_i u_i$, the inclusion problem is equivalent to finding the unique root $\alpha^* \in \mathbb{R}^r$ of $\ell(\alpha)$, namely,

$$\ell(\alpha) := U^*Q^+U\alpha + U^*(z - J_A^M(z - \mathrm{s}M^{-1}U\alpha)) \,, \tag{95}$$

where U^* denotes the adjoint of U and J_A^M is an abbreviation of the mapping $M^{-1/2}\circ J_{M^{-1/2}AM^{-1/2}}\circ M^{1/2}$. The following shows that $\ell(\alpha)$ is Lipschitz continuous with constant $\|U^*Q^+U\|+\|M^{-1/2}U\|^2$: (We abbreviate $J_{M^{-1/2}AM^{-1/2}}$ by J in the following)

$$\begin{aligned}
&\langle \ell(\alpha)-\ell(\beta),\alpha-\beta\rangle\\
=&\|\alpha-\beta\|^2_{U^*Q^+U}-\Big\langle J(M^{1/2}z-\mathrm{s}M^{-1/2}U\alpha)-J(M^{1/2}z-\mathrm{s}M^{-1/2}U\beta),M^{-1/2}U(\alpha-\beta)\Big\rangle\\
\leq&\|U^*Q^+U\|\,\|\alpha-\beta\|^2+\|M^{-1/2}U\|^2\|\alpha-\beta\|^2\,,
\end{aligned}\tag{96}$$

where, in the last line, we use the 1-Lipschitz continuity (non-expansive) of J.

The following shows strict monotonicity of l. We rewrite $\ell(\alpha)$ as follows:

$$\begin{aligned}
\ell(\alpha)&=U^*Q^+U\alpha+U^*M^{-1/2}(M^{1/2}z-J_{M^{-1/2}AM^{-1/2}}(M^{1/2}z-\mathrm{s}M^{-1/2}U\alpha))\\
&=U^*Q^+U\alpha+\mathrm{s}U^*M^{-1}U\alpha+U^*M^{-1/2}(I-J_{M^{-1/2}AM^{-1/2}})(M^{1/2}z-\mathrm{s}M^{-1/2}U\alpha)\\
&=U^*(Q^++\mathrm{s}M^{-1})U\alpha+U^*M^{-1/2}J_{M^{-1/2}A^{-1}M^{-1/2}}(M^{1/2}z-\mathrm{s}M^{-1/2}U\alpha)\,.
\end{aligned}\tag{97}$$

Using the 1-co-coercivity of $J_{M^{1/2}A^{-1}M^{1/2}}$, the function $\ell(\alpha)$ can be seen to be strictly monotone if $\alpha\mapsto U^*(Q^++\mathrm{s}M^{-1})U\alpha$ is strictly monotone. This fact is clear for the case $\mathrm{s}=1$. Therefore, in the remainder, we show strictly monotonicity of $\alpha\mapsto U^*(Q^+-M^{-1})U\alpha=U^*M^{-1/2}(M^{1/2}Q^+M^{1/2}-I)M^{-1/2}U\alpha$. We observe $M-Q\in\mathcal{S}_0(\mathcal{H})$ implies that $\|M^{-1/2}QM^{-1/2}\|<1$ and by $1\leq\|TT^{-1}\|\leq\|T\|\|T^{-1}\|$, we conclude that $\|M^{1/2}Q^+M^{1/2}\|_{\mathrm{im}(M^{-1/2}Q)}>1$ for the restriction of the operator norm to $\mathrm{im}(M^{-1/2}Q)$, hence, $Q^+-M^{-1}\in\mathcal{S}_{++}(\mathcal{H})$.

According to Lemma 2.3, we can replace $M^{-1/2}\circ J_{M^{-1/2}AM^{-1/2}}\circ M^{1/2}$ with J_A^M. Then we obtain the formula in the statement of Theorem 3.1. □

Remark A.1. *A priori Q^{-1} is set-valued, however it is easy to check that $U^*Q^{-1}U$ is single-valued (see Appendix A.4.1). We define $Q^+\colon\mathrm{im}(Q)\to\mathrm{im}(Q)$ as the inverse of Q restricted to $\mathrm{im}(Q)$ which is a single-valued mapping. It allows us to replace Q^{-1} by Q^+ in (11).*

A.4.1 Proof of Remark A.1

Proof. Given $y\in\mathrm{im}(Q)$, assume there exist $U^*x_1,U^*x_2\in U^*Q^{-1}y$ with $x_1,x_2\in Q^{-1}y$. For arbitrary $\beta\in\mathbb{R}^r$, we have $\langle\beta,U^*x_1-U^*x_2\rangle=\langle U\beta,x_1-x_2\rangle$. Since $U\beta\in\mathrm{im}Q$, then there exists some z such that $U\beta=Qz$. As a result, $\langle\beta,U^*x_1-U^*x_2\rangle=\langle U\beta,x_1-x_2\rangle=\langle Qz,x_1-x_2\rangle=\langle z,Qx_1-Qx_2\rangle=0$. We notice that $\langle\beta,U^*x_1-U^*x_2\rangle$ holds for arbitrary $\beta\in\mathbb{R}^r$. It implies $U^*x_1=U^*x_2$ and $U^*Q^{-1}y$ is single-valued. □

A.5 Proof of Corollary 3.3

Proof. Let $\mathcal{H}=\mathbb{R}^n$. In this case, we can identify a linear mapping $U\colon\mathbb{R}^r\to\mathcal{H}$ with a low rank matrix $U\in\mathbb{R}^{n\times r}$. Similarly, we can also identify $U^*\colon\mathcal{H}\to\mathbb{R}^r$ with $U^\top\in\mathbb{R}^{n\times r}$. Since $Q=UU^\top$ for some $U\in\mathcal{B}(\mathbb{R}^r,\mathbb{R}^n)$, we have

$$\mathrm{im}(Q)=\{UU^\top v|v\in\mathbb{R}^n\}=\{U\alpha|\alpha\in\mathbb{R}^r\}\,.\tag{98}$$

Since Q^+ is inverse of Q on $\mathrm{im}(Q)$, the following holds for arbitrary $v\in\mathbb{R}^n$:

$$QQ^+Qv=Qv\iff UU^\top Q^+UU^\top v=UU^\top v\iff UU^\top Q^+U\alpha=U\alpha\,.\tag{99}$$

Since the column vectors $\{u_i\}_{i=1,\cdots,r}$ of U are independent with each other, $UU^\top Q^+U\alpha=U\alpha$ yields that $U^\top Q^+U\alpha=\alpha$. Therefore, the root-finding problem in Theorem 3.1 simplifies to (12). □

A.6 Proof of Proposition 3.5

Proof. Our proof relies on the convergence result [18, Theorem 7.5.5]. By the same argument as the one in Appendix B.5 paper [8], we obtain $\partial^C l(\alpha^*)$ is non-singular. If $\ell(\alpha)$ is tame, then by [9, Theorem 1], $\ell(\alpha)$ is semi-smooth. In order to apply this result it remains to show that $\ell(\alpha)$ is tame. The property of definable functions is preserved by operations including the sum, composition by a linear operator, derivation and canonical projection ([40], [17]). Since A is a tame mapping, $I + M^{-1/2}AM^{-1/2}$ is tame as well as its graph. Here I is identity. Then the resolvent $J_{M^{-1/2}AM^{-1/2}} = (I + M^{-1/2}AM^{-1/2})^{-1}$ which is defined by the inverse of the same graph is tame ([21]) and single-valued. By the stability of the sum and composition by linear operator, we obtain that $\ell(\alpha)$ is tame. □

A.7 Proof of Proposition 3.8

Proof. Let $p = J_A^V(z)$. Since resolvent operator is non-expansive with respect to V, $\|p\|_V = \|J_T^V(z)\|_V \leq \|z\|_V + \|J_A^V(0)\|_V$. By duality, optimal α^* will satisfy $\alpha^* = u^\top(p - z)$. Then,

$$\begin{aligned} |\alpha^*| &= |u^*(p-z)| \\ &\leq \|u\|_{V^{-1}}(2\|z\|_V + \|J_A^V(0)\|_V)\,. \end{aligned} \tag{100}$$

If $V \in \mathcal{S}_c(\mathbb{R}^n)$ is bounded by a constant C, then

$$|\alpha^*| \leq \|u\|_{V^{-1}}(2\|z\|_V + \|J_A^V(0)\|_V) \leq \frac{C}{c}\|u\|(2\|z\| + \|J_A^V(0)\|)\,. \quad \square \tag{101}$$

□

A.8 Proof of Proposition 3.4

Proof. $\ell(\alpha)$ is as defined in Theorem 3.1. Set $\breve{z} = z - V^{-1}Bz$. Substituting $\alpha = \xi + U^*V^{-1}Bz$ in $\ell(\alpha)$, we obtain $\mathcal{J}(\xi) = \ell(\alpha)$. Then, there exists ξ^* such that $\alpha^* = \xi^* + U^*V^{-1}Bz$. In (11), we do the same substitution.

$$\begin{aligned} \hat{z} &= J_T^M(\breve{z} - \mathrm{s}M^{-1}U\alpha^*) \\ &= J_A^M(\breve{z} - \mathrm{s}M^{-1}U\xi^* - \mathrm{s}M^{-1}UU^*V^{-1}Bz) \\ &= J_A^M(z - M^{-1}MV^{-1}Bz - \mathrm{s}M^{-1}UU^*V^{-1}Bz - \mathrm{s}M^{-1}U\xi^*) \\ &= J_A^M(z - M^{-1}\underbrace{(M + \mathrm{s}UU^*)}_{=V}V^{-1}Bz - \mathrm{s}M^{-1}U\xi^*) \\ &= J_A^M(z - M^{-1}Bz - \mathrm{s}M^{-1}U\xi^*)\,. \end{aligned} \tag{102}$$

Due to Theorem 3.1, $\ell(\alpha)$ is Lipschitz with constant $1 + \|M^{-1/2}U\|^2$ and strongly monotone. Since $\mathcal{J}(\xi)$ is obtained by translation, it enjoys the same properties. □

A.9 Proof of Lemma 4.10

Proof. (i) M is symmetric positive-definite and $\gamma_k u_k u_k^*$ is symmetric positive semi-definite since $\inf_{k\in\mathbb{N}} \gamma_k > 0$. Thus, $M_k \succeq M$ showing that M_k is symmetric positive-definite. Moreover, $M_k - \frac{1}{\beta}I \succeq M - \frac{1}{\beta}I \succeq cI$, which shows part (i) of Assumption 2. We now show that M_k obeys (ii) of the assumption. Since B is β-co-coercive, we have in view of [6, Remark 4.34 and Proposition 4.35] that

$$\beta\|y_k\|^2 \leq \langle y_k, s_k\rangle \leq \|s_k\|^2/\beta. \tag{103}$$

Assume that $s_k \neq 0$ (otherwise, there is nothing to prove). Thus

$$\langle Ms_k - y_k, s_k\rangle = \|s_k\|_M^2 - \langle y_k, s_k\rangle \geq \|s_k\|_M^2 - \|s_k\|^2/\beta = \|s_k\|_{M-\beta I}^2 \geq c\|s_k\|^2.$$

Combining this with (103), we get

$$\|u_k\| = \frac{\|y_k - Ms_k\|}{\sqrt{\langle Ms_k - y_k, s_k\rangle}} \leq \frac{\|y_k\| + \|M\|\|s_k\|}{\sqrt{c}\|s_k\|} \leq \frac{1/\beta + \|M\|}{\sqrt{c}}. \tag{104}$$

This entails that

$$\sup_{k\in\mathbb{N}} \|M_k\| \leq \|M\| + \sup_{k\in\mathbb{N}} \gamma_k\|u_k\|^2 \leq \|M\| + \frac{(1/\beta + \|M\|)^2}{c}\sup_{k\in\mathbb{N}} \gamma_k < +\infty.$$

(ii) Let us focus on part (i) of Assumption 2. We have, using (104),

$$M_k - \frac{1}{\beta}I \succeq M - \frac{1}{\beta}I - \gamma_k\|u_k u_k^*\|I = (c - \gamma_k\|u_k\|^2)I \succeq \left(c - \frac{(1/\beta + \|M\|)^2}{c}\gamma_k\right)$$

and the last term is positive under the prescribed choice of γ_k. To verify part (ii), it is sufficient to observe that $\|M_k\| \leq \|M\|$.

□

A.10 Proof of Lemma 4.4

Proof. (i) As in the proof of Lemma 4.10, $M_k \succeq M \succeq (1/\beta + c)I > 0$. Moreover

$$\begin{aligned}(1+\eta_k)M_k - M_{k+1} &= \eta_k M + (1+\eta_k)\gamma_k u_k u_k^* - \gamma_{k+1}u_{k+1}u_{k+1}^* \\ &\succeq \eta_k(1/\beta + c)I - \gamma_{k+1}\|u_{k+1}\|^2 \\ &= \eta_k(1/\beta + c)I - \eta_{k+1}(1/\beta + c)I \\ &= (1/\beta + c)(\eta_k - \eta_{k+1})I \succeq 0,\end{aligned}$$

since η_k is non-increasing. The uniform boundedness of M_k is straightforward as $\|M_k\| \leq \|M\| + \eta_k(1/\beta + c)$.

(ii) We have in this case

$$M_k - \frac{1}{\beta}I \succeq M - \frac{1}{\beta}I - \gamma_k\|u_k u_k^*\|I \succeq (c - \eta_k\kappa(1/\beta + c))\, I \succeq (c - (1-\kappa)(1/\beta + c))\, I$$

Under our condition on κ, we have $c - (1-\kappa)(1/\beta + c > 0$. In addition,

$$\begin{aligned}(1+\eta_k)M_k - M_{k+1} &= \eta_k M - (1+\eta_k)\gamma_k u_k u_k^* + \gamma_{k+1}u_{k+1}u_{k+1}^* \\ &\succeq \eta_k(1/\beta + c)I - (1+\eta_k)\eta_k\kappa(1/\beta + c)I \\ &= \eta_k(1/\beta + c)(1 - (1+\eta_k)\kappa)I \succeq 0\end{aligned}$$

since $\eta_k \leq (1-\kappa)/\kappa$. M_k is also uniformly bounded with the same argument as above. This completes the proof.

□

Availability of data

In order to record the exact algorithmic details for our experiments, all code for experiments from this paper is available at https://github.com/wsdxiaohao/quasi_Newton_FBS.git.

Conflict of interest statements

This work was supported by Deutsche Forschungsgemeinschaft (Grant numbers OC150/5-1) and Agence Nationale de la Recherche (Grant numbers ANR-20-CE92-0037-01).

REFERENCES

[1] B. Abbas, H. Attouch, and B. F. Svaiter, *Newton-like dynamics and forward-backward methods for structured monotone inclusions in Hilbert spaces*, Journal of Optimization Theory and Applications, 161 (2014), pp. 331–360.

[2] H. Attouch, M. M. Alves, and B. F. Svaiter, *A dynamic approach to a proximal-Newton method for monotone inclusions in Hilbert spaces, with complexity* $O(1/n^2)$, arXiv preprint arXiv:1502.04286, (2015).

[3] H. Attouch, P. Redont, and B. F. Svaiter, *Global convergence of a closed-loop regularized Newton method for solving monotone inclusions in Hilbert spaces*, Journal of Optimization Theory and Applications, 157 (2013), pp. 624–650.

[4] H. Attouch and B. F. Svaiter, *A continuous dynamical Newton-like approach to solving monotone inclusions*, SIAM Journal on Control and Optimization, 49 (2011), pp. 574–598.

[5] H. Attouch and M. Théra, *A general duality principle for the sum of two operators*, Journal of Convex Analysis, 3 (1996), pp. 1–24.

[6] H. H. Bauschke, P. L. Combettes, H. H. Bauschke, and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, Springer, 2017.

[7] S. Becker and J. Fadili, *A quasi-Newton proximal splitting method*, Advances in Neural Information Processing Systems, (2012).

[8] S. Becker, J. Fadili, and P. Ochs, *On quasi-Newton forward-backward splitting: Proximal calculus and convergence*, SIAM Journal on Optimization, 29 (2019), pp. 2445–2481.

[9] J. Bolte, A. Daniilidis, and A. Lewis, *Tame functions are semismooth*, Mathematical Programming, 117 (2009), pp. 5–19.

[10] K. Bredies, D. Lorenz, et al., *Mathematical image processing*, Springer, 2018.

[11] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of mathematical imaging and vision, 40 (2011), pp. 120–145.

[12] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319.

[13] P. Combettes, L. Condat, J. Pesquet, and B. Vu, *A forward-backward view of some primal-dual optimization methods in image recovery*, IEEE International Conference on Image Processing, (2014).

[14] P. L. Combettes, *Quasi-Fejérian analysis of some optimization algorithms*, in Studies in Computational Mathematics, vol. 8, Elsevier, 2001, pp. 115–152.

[15] P. L. Combettes and B. C. Vũ, *Variable metric quasi-Fejér monotonicity*, Nonlinear Analysis: Theory, Methods and Applications, 78 (2013), pp. 17–31.

[16] ——, *Variable metric forward–backward splitting with applications to monotone inclusions in duality*, Optimization, 63 (2014), pp. 1289–1318.

[17] M. Coste, *An Introduction to o-minimal Geometry*, Istituti editoriali e poligrafici internazionali Pisa, 2000.

[18] F. Facchinei and J. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer, 2003.

[19] T. Goldstein, M. Li, and X. Yuan, *Adaptive primal-dual splitting methods for statistical learning and image processing*, Advances in neural information processing systems, 28 (2015).

[20] B. He and X. Yuan, *Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective*, SIAM Journal on Imaging Sciences, 5 (2012), pp. 119–149.

[21] A. D. Ioffe, *An invitation to tame optimization*, SIAM Journal on Optimization, 19 (2009), pp. 1894–1917.

[22] C. Kanzow and T. Lechner, *Globalized inexact proximal Newton-type methods for nonconvex composite functions*, Computational Optimization and Applications, 78 (2021), pp. 377–410.

[23] C. Kanzow and T. Lechner, *Efficient regularized proximal quasi-Newton methods for large-scale nonconvex composite optimization problems*, tech. rep., University of Würzburg, Institute of Mathematics, January 2022.

[24] S. Karimi and S. Vavasis, *IMRO: A proximal quasi-Newton method for solving ℓ_1-regularized least squares problems*, SIAM Journal on Optimization, 27 (2017), pp. 583–615.

[25] P. D. Khanh, B. S. Mordukhovich, V. T. Phat, and D. B. Tran, *Generalized damped Newton algorithms in nonsmooth optimization via second-order subdifferentials*, Journal of Global Optimization, 86 (2023), pp. 93–122.

[26] J. D. Lee, Y. Sun, and M. A. Saunders, *Proximal Newton-type methods for minimizing composite functions*, SIAM Journal on Optimization, 24 (2014), pp. 1420–1443.

[27] J. Liang, *Convergence rates of first-order operator splitting methods*, PhD thesis, Normandie Université; GREYC CNRS UMR 6072, 2016.

[28] C. Liu and L. Luo, *Quasi-Newton methods for saddle point problems*, Advances in Neural Information Processing Systems, 35 (2022), pp. 3975–3987.

[29] Y. Liu, Y. Xu, and W. Yin, *Acceleration of primal–dual methods by preconditioning and simple subproblem procedures*, Journal of Scientific Computing, 86 (2021), pp. 1–34.

[30] D. A. Lorenz and T. Pock, *An inertial forward-backward algorithm for monotone inclusions*, Journal of Mathematical Imaging and Vision, 51 (2015), pp. 311–325.

[31] P. Patrinos, L. Stella, and A. Bemporad, *Forward-backward truncated Newton methods for convex composite optimization*, arXiv preprint arXiv:1402.6655, (2014).

[32] T. Pock and A. Chambolle, *Diagonal preconditioning for first order primal-dual algorithms in convex optimization*, IEEE International Conference on Computer Vision, (2011).

[33] B. Polyak, *Introduction to optimization*, Optimization Software, 1987.

[34] A. Rodomanov and Y. Nesterov, *Greedy quasi-Newton methods with explicit superlinear convergence*, SIAM Journal on Optimization, 31 (2021), pp. 785–811.

[35] K. Scheinberg and X. Tang, *Practical inexact proximal quasi-Newton method with global complexity analysis*, Mathematical Programming, 160 (2016), pp. 495–529.

[36] M. Schmidt, D. Kim, and S. Sra, *Projected Newton-type methods in machine learning*, Optimization for Machine Learning, (2012).

[37] M. Schmidt, E. van den Berg, M. Friedlander, and K. Murphy, *Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm*, in AISTATS, 2009.

[38] M. V. Solodov and B. F. Svaiter, *A globally convergent inexact Newton method for systems of monotone equations*, Reformulation: Nonsmooth, piecewise smooth, semismooth and smoothing methods, (1999), pp. 355–369.

[39] L. Stella, A. Themelis, and P. Patrinos, *Forward–backward quasi-Newton methods for nonsmooth optimization problems*, Computational Optimization and Applications, 67 (2017), pp. 443–487.

[40] L. Van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Mathematical Journal, 84 (1996), pp. 497–540.

[41] S. Wang, J. Fadili, and P. Ochs, *A quasi-newton primal-dual algorithm with line search*, in International Conference on Scale Space and Variational Methods in Computer Vision, Springer, 2023, pp. 444–456.

[42] S. Wright and J. Nocedal, *Numerical optimization*, Springer Science, (1999).