SIAM J. OPTIM.
Vol. 30, No. 2, pp. 1451–1472

A FORWARD-BACKWARD SPLITTING METHOD FOR MONOTONE INCLUSIONS WITHOUT COCOERCIVITY*

YURA MALITSKY† AND MATTHEW K. TAM†‡

Abstract. In this work, we propose a simple modification of the forward-backward splitting method for finding a zero in the sum of two monotone operators. Our method converges under the same assumptions as Tseng's forward-backward-forward method, namely, it does not require cocoercivity of the single-valued operator. Moreover, each iteration only uses one forward evaluation rather than two as is the case for Tseng's method. Variants of the method incorporating a linesearch, relaxation and inertia, or a structured three operator inclusion are also discussed.

Key words. forward-backward algorithm, Tseng's method, operator splitting

AMS subject classifications. 49M29, 90C25, 47H05, 47J20, 65K15

DOI. 10.1137/18M1207260

1. Introduction.

In this work, we propose an algorithm for finding a zero in the sum of two monotone operators in a (real) Hilbert space $\mathcal{H}$. Specifically, we consider the monotone inclusion problem

$$\text{find } x \in \mathcal{H} \text{ such that } 0 \in (A+B)(x), \tag{1.1}$$

where $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ and $B\colon \mathcal{H} \to \mathcal{H}$ are (maximally) monotone operators with B (locally) Lipschitz continuous such that $(A+B)^{-1}(0) \neq \varnothing$. Inclusions of the form specified by (1.1) arise in numerous problems of fundamental importance in mathematical optimization, either directly or through an appropriate reformulation. In what follows, we provide some motivating examples.

Convex minimization. Consider the minimization problem

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

where $f\colon \mathcal{H} \to (-\infty, +\infty]$ is proper, lower semicontinuous (lsc), convex and $g\colon \mathcal{H} \to \mathbb{R}$ is convex with (locally) Lipschitz continuous gradient denoted ∇g. The solutions to this minimization problem are precisely the points $x \in \mathcal{H}$ which satisfy the *first order optimality condition:*

$$0 \in (\partial f + \nabla g)(x), \tag{1.2}$$

where ∂f denotes the *subdifferential* of f. Clearly (1.2) is of the form specified by (1.1).

*Received by the editors August 13, 2018; accepted for publication (in revised form) March 17, 2020; published electronically May 21, 2020.
https://doi.org/10.1137/18M1207260
Funding: The first author's research was supported by German Research Foundation grant SFB755-A4. The second author's research was supported in part by a fellowship from the Alexander von Humboldt Foundation and in part by a Discovery Early Career Research Award from the Australian Research Council.

†Institute for Numerical and Applied Mathematics, University of Göttingen, 37083 Göttingen, Germany (y.malitsky@gmail.com).
‡School of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3010, Australia (matthew.tam@unimelb.edu.au).

General monotone inclusions. Consider the inclusion problem

$$\text{find } x \in \mathcal{H}_1 \text{ such that } 0 \in (A + K^*BK)(x), \tag{1.3}$$

where $A\colon \mathcal{H}_1 \rightrightarrows \mathcal{H}_1$ and $B\colon \mathcal{H}_2 \rightrightarrows \mathcal{H}_2$ are maximally monotone operators, and $K\colon \mathcal{H}_1 \to \mathcal{H}_2$ is a linear, bounded operator with adjoint K^*. As was observed in [8, 9], solving (1.3) can be equivalently cast as the following monotone inclusion posed in the product space:

$$\text{find } \begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{H}_1 \times \mathcal{H}_2 \text{ such that } \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \left(\begin{bmatrix} A & 0 \\ 0 & B^{-1} \end{bmatrix} + \begin{bmatrix} 0 & K^* \\ -K & 0 \end{bmatrix} \right) \begin{pmatrix} x \\ y \end{pmatrix}. \tag{1.4}$$

Notice that the first operator in (1.4) is maximally monotone whereas the second is bounded and linear (in particular, it is Lipschitz continuous with full domain). Consequently, (1.4) is also of the form specified by (1.1).

Another variant of (1.1) is the three operator inclusion

$$\text{find } x \in \mathcal{H} \text{ such that } 0 \in (A + B + C)(x), \tag{1.5}$$

where the operators A and B are as before and $C\colon \mathcal{H} \to \mathcal{H}$ is β-cocoercive. Problems with this structure have been studied in [10, 17].

Saddle point problems and variational inequalities. Many convex optimization problems can be formulated as the *saddle point problem*

$$\min_{x \in \mathcal{H}} \max_{y \in \mathcal{H}} g(x) + \Phi(x, y) - f(y), \tag{1.6}$$

where $f, g\colon \mathcal{H} \to (-\infty, +\infty]$ are proper, lsc, convex functions and $\Phi\colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is a smooth convex-concave function. Problems of this form naturally arise in machine learning, statistics, etc., where the dual (maximization) problem comes either from dualizing the constraints in the primal problem or from using the Fenchel–Legendre transform to leverage a nonsmooth composite part. Through its first order optimality condition, the saddle point problem (1.6) can expressed as the monotone inclusion

$$\text{find } \begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{H} \times \mathcal{H} \text{ such that } \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial g(x) \\ \partial f(y) \end{pmatrix} + \begin{pmatrix} \nabla_x \Phi(x, y) \\ -\nabla_y \Phi(x, y) \end{pmatrix}, \tag{1.7}$$

which is of the form specified by (1.1). By using the definitions of the respective subdifferentials, (1.7) can also be expressed in terms of the *variational inequality*: find $z^* = (x^*, y^*)^\top \in \mathcal{H} \times \mathcal{H}$ such that

$$\langle B(z^*), z - z^* \rangle + g(x) - g(x^*) - f(y) + f(y^*) \geq 0 \quad \forall z = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{H} \times \mathcal{H}, \tag{1.8}$$

where $B(x, y) := (\nabla_x \Phi(x, y), -\nabla_y \Phi(x, y))^\top$.

Splitting algorithms are a class of methods which can be used to solve (1.1) by only invoking each operator individually rather than their sum directly. The individual steps within each iteration of these methods can be divided into two categories: *forward evaluations* in which the value of a single-valued operator is computed, and *backward evaluations* in which the *resolvent* of an operator computed. Recall that the resolvent of an operator A is given by $J_A := (I + A)^{-1}$, where $I\colon \mathcal{H} \to \mathcal{H}$ denotes the identity operator.

When the resolvents of both of the involved operators can be easily computed, there are various algorithms in the literature which are suitable for solving (1.1) with B not necessarily single-valued. The best known example of such an algorithm is the *Douglas–Rachford method* [24, 39]. In practice, however, it is usually not the case that both resolvents can be readily computed and thus in order to efficiently deal with realistic problems, it is often necessary to impose further structure on the operators in (1.1). Splitting methods which do not require computation of two resolvents are therefore of practical interest.

The best-known splitting method for solving the inclusion (1.1) when B is single-valued is the *forward-backward method*, so called because each iteration combines one forward evaluation of B with one backward evaluation of A. More precisely, the method generates a sequence according to

$$x_{k+1} = J_{\lambda A}(x_k - \lambda B(x_k)) \quad \forall k \in \mathbb{N} \tag{1.9}$$

and converges weakly to a solution provided the operator $B\colon \mathcal{H} \to \mathcal{H}$ is $1/L$-*cocoercive* and $\lambda \in (0, 2/L)$. Recall that $B\colon \mathcal{H} \to \mathcal{H}$ is β-*cocoercive* if

$$\langle x - y, B(x) - B(y)\rangle \geq \beta \left\|B(x) - B(y)\right\|^2 \quad \forall x, y \in \mathcal{H}.$$

Cocoercivity of an operator is a stronger property than Lipschitz continuity and hence can be difficult to satisfy for general monotone inclusions. For instance, apart from the trivial case when $K = 0$, the skew-symmetric operator in (1.4) is never cocoercive. Furthermore, without cocoercivity, convergence of (1.9) can be guaranteed only in the presence of similarly strong assumptions such as strong monotonicity of $A + B$ [12], or at the cost of incorporating a backtracking strategy [6] (even when the Lipschitz constant is known).

In order to relax the cocoercivity assumption, Tseng [40] proposed a modification of the forward-backward algorithm, known as the *Tseng's method* or the *forward-backward-forward method*, which only requires Lipschitzness of B at the expense of an additional forward evaluation. Applied to (1.1), Tseng's method generates sequences according to

$$\begin{cases} y_k = J_{\lambda A}(x_k - \lambda B(x_k)) \\ x_{k+1} = y_k - \lambda B(y_k) + \lambda B(x_k) \end{cases} \quad \forall k \in \mathbb{N} \tag{1.10}$$

and converges weakly provided B is L-Lipschitz and $\lambda \in (0, 1/L)$.

In this work, we introduce and analyze a new method for solving (1.1) which converges under the same assumptions as Tseng's method but whose implementation requires only one forward evaluation per iteration instead of two. For a fixed stepsize $\lambda > 0$, the proposed scheme can be simply described as

$$x_{k+1} = J_{\lambda A}\big(x_k - 2\lambda B(x_k) + \lambda B(x_{k-1})\big) \quad \forall k \in \mathbb{N} \tag{1.11}$$

and converges weakly if B is L-Lipschitz and the stepsize is chosen to satisfy $\lambda < \frac{1}{2L}$. We refer to this scheme as the *forward-reflected-backward method*. It is worth noting that the analysis of our method is entirely different from existing schemes and hence is of interest in its own right. In particular, the sequence generated by the method is not Fejér monotone, although it does satisfy a quasi-Fejér property [13]. Moreover, there are relatively few fundamentally different alternatives to Tseng's forward-backward-forward algorithm for solving inclusions in the form of (1.1) without cocoercivity [15, 22, 34].

We also remark that our method is of particular interest in the setting of the saddle point problem (1.7). Indeed, one of the first splitting techniques for solving (1.6) is the famous *Arrow–Hurwicz algorithm* [3], which suffers from the shortcoming of requiring strict assumptions to ensure convergence. This was remedied in late 1970s when various modifications of the algorithm were proposed [2, 23, 33] which turned out to be applicable not only to saddle point problems, but also to more general variational inequalities. Note also that the simplest case of (1.6) occurs when Φ is a bilinear form and gives rise to the popular *primal-dual algorithm*, first analyzed by Chambolle and Pock [11]. In a recent preprint [20], a variant of this algorithm, which can be applied when Φ is not necessarily bilinear, was considered. Such an extension is a significant improvement as it provides an approach to the saddle point problem that is different from variational inequality methods. An interesting common feature of the methods in [11, 15, 20, 27] as well as the one presented here is that their respective iterations include a "reflection term" in which the value of an operator at the previous point is subtracted from twice its value at the current point.

In addition to general interest in monotone inclusions from optimization community described above, a new surge has appeared in machine learning research; see [16, 18, 28, 29, 38] and the references therein. In these works, the authors design algorithms for training *generative adversarial networks (GANs)* [19]. Although this takes the form of a nonconvex-nonconcave min-max problem, the main workhorses are based on classical algorithms for solving monotone variational inequalities. Thus, we believe that new algorithmic ideas, even for the monotone case, may have some impact in this field as well.

The remainder of this paper is organized as follows. In section 2, we introduce our method and prove its convergence (Theorem 2.5). In section 2.1, this result is refined to show that convergence is linear whenever one of the operators is strongly monotone. In section 3, we incorporate a linesearch procedure into the method (Theorem 3.4). In section 4, we consider a relaxed inertial version (Theorem 4.3), and in section 5, we propose a variant which solves the three operator inclusion (1.5). Finally, in section 6, we analyze a version of the stochastic algorithm which can be considered in between the forward-backward method and our proposed method.

2. Forward-reflected-backward splitting. Recall that a set-valued operator $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is *monotone* if

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

where $\operatorname{gra} A = \{(x,y) \in \mathcal{H} \times \mathcal{H} : y \in A(x)\}$ denotes the *graph* of A. A monotone operator is *maximally monotone* if its graph is not properly contained in the graph of any other monotone operator. The *resolvent* of a maximally monotone operator $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$, defined by $J_A := (I + A)^{-1}$, is an everywhere single-valued operator [5]. A single-valued operator $B\colon \mathcal{H} \to \mathcal{H}$ is *L-Lipschitz* if $\|B(x) - B(y)\| \leq L\|x - y\|$ for all $x, y \in \mathcal{H}$.

In this section, we consider the problem of finding a point $x \in \mathcal{H}$ such that

$$0 \in (A + B)(x), \tag{2.1}$$

where $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is maximal monotone, and $B\colon \mathcal{H} \to \mathcal{H}$ is monotone and L-Lipschitz. Given initial points $x_0, x_{-1} \in \mathcal{H}$, we consider the scheme

$$x_{k+1} = J_{\lambda_k A}\big(x_k - \lambda_k B(x_k) - \lambda_{k-1}(B(x_k) - B(x_{k-1})\big) \quad \forall k \in \mathbb{N}, \tag{2.2}$$

where $(\lambda_k) \subseteq \mathbb{R}_+$ is a sequence of stepsizes (starting with from $k = -1$). Note that each iteration of this scheme requires one forward evaluation and one backward evaluation. Using the definition of the resolvent $J_{\lambda_k A} = (I + \lambda_k A)^{-1}$, (2.2) can be equivalently expressed as the inclusion

$$x_{k+1} - x_k + \lambda_k B(x_k) + \lambda_{k-1}\left(B(x_k) - B(x_{k-1})\right) \in -\lambda_k A(x_{k+1}) \quad \forall k \in \mathbb{N}. \tag{2.3}$$

Before turning our attention to the convergence analysis of this method, we first note some special cases in which it recovers known methods.

Remark 2.1 (special cases of (2.2)). We consider three cases in which the proposed algorithm reduces or is equivalent to known methods. For simplicity, we only consider the fixed stepsize case (i.e., $\exists \lambda > 0$ such that $\lambda_k = \lambda$ for all k). In this case, (2.2) can be expressed compactly as

$$x_{k+1} = J_{\lambda A}\big(x_k - 2\lambda B(x_k) + \lambda B(x_{k-1})\big). \tag{2.4}$$

(a) If $B = 0$, then (2.4) simplifies to the *proximal point algorithm* [36], that is, (2.4) becomes

$$x_{k+1} = J_{\lambda A}(x_k) \quad \forall k \in \mathbb{N}.$$

(b) If $A = N_C$ is the normal cone to a set C and B is an affine operator, then (2.4) can be expressed as

$$x_{k+1} = P_C\big(x_k - \lambda B(2x_k - x_{k-1})\big) \quad \forall k \in \mathbb{N}, \tag{2.5}$$

which coincides with the *projected reflected gradient method* [26] for VIs.

(c) If $A = N_{\mathcal{H}} = 0$, then the projected reflected gradient method (2.5) becomes

$$x_{k+1} = x_k - \lambda B(2x_k - x_{k-1}) \quad \forall k \in \mathbb{N}.$$

Under the change of variables $\overline{x}_k = 2x_k - x_{k-1}$, this becomes

$$\overline{x}_{k+1} = \overline{x}_k - 2\lambda B(\overline{x}_k) + (x_{k-1} - x_k) = \overline{x}_k - 2\lambda B(\overline{x}_k) + \lambda B(\overline{x}_{k-1}) \quad \forall k \in \mathbb{N},$$

which is precisely (2.2) with $A = 0$. Alternatively, (2.4) can be expressed as the two step recursion

$$\begin{cases} y_{k+1} = y_k - \lambda B(x_k), \\ x_{k+1} = y_{k+1} - \lambda B(x_k). \end{cases} \tag{2.6}$$

This is exactly Popov's algorithm [33] for unconstrained VIs. In this sense, the three methods coincide in this case up to a change of variable. Furthermore, in the GANs literature, both (2.5) and (2.6) are also known to be equivalent to the *optimistic gradient* method. For details, see the discussion in [21].

Before establishing convergence of the method, we require some preparatory results.

LEMMA 2.2. *Let $(z_k) \subseteq \mathcal{H}$ be a bounded sequence and suppose $\lim_{k\to\infty} \|z_k - z\|$ exists whenever z is a cluster point of (z_k). Then (z_k) is weakly convergent.*

Equation (2.7) in the following proposition conforms, in particular, to our proposed method given by

$$x_{k+1} = J_{\lambda_k A}\big(x_k - \lambda_k B(x_k) - \lambda_{k-1}(B(x_k) - B(x_{k-1})\big).$$

PROPOSITION 2.3. *Let $F\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, and let $d_1, v_2, u_1, v_1, u_0 \in \mathcal{H}$ be arbitrary. Define d_2 as*

$$d_2 = J_F(d_1 - u_1 - (v_1 - u_0)). \tag{2.7}$$

Then, for all $x \in \mathcal{H}$ and $u \in -F(x)$, we have

$$\begin{aligned}\|d_2 - x\|^2 + 2\langle v_2 - u_1, x - d_2\rangle &\le \|d_1 - x\|^2 + 2\langle v_1 - u_0, x - d_1\rangle \\ &\quad + 2\langle v_1 - u_0, d_1 - d_2\rangle - \|d_1 - d_2\|^2 - 2\langle v_2 - u, d_2 - x\rangle.\end{aligned} \tag{2.8}$$

Proof. By definition of the resolvent, $d_1 - u_1 - (v_1 - u_0) \in d_2 + F(d_2)$ and hence, by monotonicity of F,

$$\begin{aligned}0 &\le \langle d_2 - d_1 + u_1 + (v_1 - u_0) - u, x - d_2\rangle \\ &= \langle d_2 - d_1, x - d_2\rangle + \langle u_1 - u, x - d_2\rangle + \langle v_1 - u_0, x - d_2\rangle.\end{aligned}$$

The first term can expressed as

$$\langle d_2 - d_1, x - d_2\rangle = \frac{1}{2}\left(\|d_1 - x\|^2 - \|d_2 - x\|^2 - \|d_2 - d_1\|^2\right),$$

and the second and third terms can be rewritten, respectively, as

$$\begin{aligned}\langle u_1 - u, x - d_2\rangle &= \langle v_2 - u, x - d_2\rangle + \langle u_1 - v_2, x - d_2\rangle, \\ \langle v_1 - u_0, x - d_2\rangle &= \langle v_1 - u_0, x - d_1\rangle + \langle v_1 - u_0, d_1 - d_2\rangle.\end{aligned}$$

The claimed inequality follows by combining these expressions. ∎

Apart from using the monotonicity of F, the proof of Proposition 2.3 only uses simple algebraic manipulations involving $u_1, v_1, u_0, v_2, d_1, d_2$. Nevertheless, the resulting inequality (2.8) already provides some insight into how our subsequence analysis of (1.11) proceeds. For instance, the first line of (2.8) suggests terms for telescoping so long as the second line can be appropriately estimated.

LEMMA 2.4. *Let $x \in (A+B)^{-1}(0)$ and let (x_k) be given by (2.2). Suppose $(\lambda_k) \subseteq [\varepsilon, \frac{1-2\varepsilon}{2L}]$ for some $\varepsilon > 0$. Then, for all $k \in \mathbb{N}$, we have*

$$\begin{aligned}\|x_{k+1} - x\|^2 + 2\lambda_k\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + \left(\frac{1}{2} + \varepsilon\right)\|x_{k+1} - x_k\|^2 \\ \le \|x_k - x\|^2 + 2\lambda_{k-1}\langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2.\end{aligned} \tag{2.9}$$

Proof. By applying Proposition 2.3 with

$$\begin{aligned}&F := \lambda_k A && d_1 := x_k && u_0 := \lambda_{k-1}B(x_{k-1}) && v_1 := \lambda_{k-1}B(x_k) \\ &u := \lambda_k B(x) && d_2 := x_{k+1} && u_1 := \lambda_k B(x_k) && v_2 := \lambda_k B(x_{k+1}),\end{aligned}$$

we obtain the inequality

$$\begin{aligned}&\|x_{k+1} - x\|^2 + 2\lambda_k\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + \|x_{k+1} - x_k\|^2 \\ &\quad \le \|x_k - x\|^2 + 2\lambda_{k-1}\langle B(x_k) - B(x_{k-1}), x - x_k\rangle \\ &\qquad + 2\lambda_{k-1}\langle B(x_k) - B(x_{k-1}), x_k - x_{k+1}\rangle - 2\lambda_k\langle B(x_{k+1}) - B(x), x_{k+1} - x\rangle.\end{aligned}$$

Since B is monotone, the last term is nonnegative. Using Lipschitzness of B, the second-last term can be estimated as

$$\begin{aligned}\langle B(x_k) - B(x_{k-1}), x_k - x_{k+1}\rangle &\leq L\,\|x_k - x_{k-1}\|\,\|x_k - x_{k+1}\| \\ &\leq \frac{L}{2}\left(\|x_k - x_{k-1}\|^2 + \|x_k - x_{k+1}\|^2\right).\end{aligned} \tag{2.10}$$

Thus, altogether, we obtain

$$\begin{aligned}&\|x_{k+1} - x\|^2 + 2\lambda_k\,\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + (1 - \lambda_{k-1}L)\,\|x_{k+1} - x_k\|^2 \\ &\quad\leq \|x_k - x\|^2 + 2\lambda_{k-1}\,\langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \lambda_{k-1}L\,\|x_k - x_{k-1}\|^2.\end{aligned}$$

The claimed inequality follows since $\lambda_{k-1}L < \frac{1}{2}$ and $1 - \lambda_{k-1}L \geq 1 - \frac{1-2\varepsilon}{2} = \frac{1}{2} + \varepsilon$. ∎

We are now ready for the first main result regarding convergence of the proposed method.

THEOREM 2.5. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, let $B\colon \mathcal{H} \to \mathcal{H}$ be monotone and L-Lipschitz, and suppose that $(A+B)^{-1}(0) \neq \varnothing$. Suppose $(\lambda_k) \subseteq [\varepsilon, \frac{1-2\varepsilon}{2L}]$ for some $\varepsilon > 0$. Given $x_0, x_{-1} \in \mathcal{H}$, define the sequence (x_k) according to*

$$x_{k+1} = J_{\lambda_k A}\big(x_k - \lambda_k B(x_k) - \lambda_{k-1}(B(x_k) - B(x_{k-1}))\big) \quad \forall k \in \mathbb{N}.$$

Then (x_k) converges weakly to a point contained in $(A+B)^{-1}(0)$.

Proof. Let $x \in (A+B)^{-1}(0)$. By Lemma 2.4, we have

$$\begin{aligned}&\|x_{k+1} - x\|^2 + 2\lambda_k\,\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + \left(\frac{1}{2} + \varepsilon\right)\|x_{k+1} - x_k\|^2 \\ &\quad\leq \|x_k - x\|^2 + 2\lambda_{k-1}\,\langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2,\end{aligned} \tag{2.11}$$

which telescopes to yield

$$\begin{aligned}&\|x_{k+1} - x\|^2 + 2\lambda_k\,\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + \frac{1}{2}\|x_{k+1} - x_k\|^2 \\ &+ \varepsilon\sum_{i=0}^{k}\|x_{i+1} - x_i\|^2 \leq \|x_0 - x\|^2 + 2\lambda_{-1}\,\langle B(x_0) - B(x_{-1}), x - x_0\rangle + \frac{1}{2}\,\|x_0 - x_{-1}\|^2.\end{aligned} \tag{2.12}$$

Using Lipschitzness of B, we can estimate

$$\begin{aligned}2\lambda_k\,\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle &\geq -2\lambda_k L\,\|x_{k+1} - x_k\|\,\|x - x_{k+1}\| \\ &\geq -\lambda_k L\left(\|x_{k+1} - x_k\|^2 + \|x - x_{k+1}\|^2\right).\end{aligned} \tag{2.13}$$

Since $\lambda_k L \leq (1-2\varepsilon)/2 < 1/2$, substituting the previous equation back into (2.12) gives

$$\begin{aligned}&\frac{1}{2}\|x_{k+1} - x\|^2 + \varepsilon\sum_{i=0}^{k}\|x_{i+1} - x_i\|^2 \\ &\quad\leq \|x_0 - x\|^2 + 2\lambda_{-1}\,\langle B(x_0) - B(x_{-1}), x - x_0\rangle + \frac{1}{2}\,\|x_0 - x_{-1}\|^2,\end{aligned}$$

from which we deduce that (x_k) is bounded and that $\|x_k - x_{k+1}\| \to 0$.

Let $\overline{x}$ be a sequential weak cluster point of the bounded sequence (x_k). From (2.3),

$$\frac{1}{\lambda_{k-1}}\big(x_{k-1} - x_k + \lambda_{k-1}\left(B(x_k) - B(x_{k-1})\right) + \lambda_{k-2}\left(B(x_{k-2}) - B(x_{k-1})\right)\big) \in (A+B)(x_k) \qquad \forall k \geq 1. \tag{2.14}$$

Since $A + B$ is maximally monotone [5, Corollaries 24.4(i) and 20.25], its graph is demiclosed (i.e., sequentially closed in the weak-strong topology on $\mathcal{H} \times \mathcal{H}$) [5, Proposition 20.33]. Thus, by taking the limit along a subsequence of (x_k) which converges to $\overline{x}$ in (2.14) and noting that $\lambda_k \geq \varepsilon$ for all $k \in \mathbb{N}$, we deduce that $0 \in (A+B)(\overline{x})$. To show that (x_k) is weakly convergent, first note that, by combining (2.11) and (2.13), we deduce existence of the limit

$$\lim_{k\to\infty} \left(\|x_k - \overline{x}\|^2 + 2\lambda_{k-1}\left\langle B(x_k) - B(x_{k-1}), \overline{x} - x_k \right\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2 \right). \tag{2.15}$$

Since (x_k) and (λ_k) are bounded, $\|x_k - x_{k+1}\| \to 0$, and B is continuous, it then follows that the limit (2.15) is equal to $\lim_{k\to\infty} \|x_k - \overline{x}\|^2$. Since the cluster point $\overline{x}$ of (x_k) was chosen arbitrarily, the sequence (x_k) is weakly convergent by Lemma 2.2 and the proof is complete. ∎

As an immediate consequence of Theorem 2.5, we obtain the following corollary when the stepsize sequence (λ_k) is constant.

COROLLARY 2.6. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, let $B\colon \mathcal{H} \to \mathcal{H}$ be monotone and L-Lipschitz, and suppose that $(A+B)^{-1}(0) \neq \varnothing$. Choose $\lambda \in (0, \frac{1}{2L})$. Given $x_0, x_{-1} \in \mathcal{H}$, define the sequence (x_k) according to*

$$x_{k+1} = J_{\lambda A}\big(x_k - 2\lambda B(x_k) + \lambda B(x_{k-1})\big) \quad \forall k \in \mathbb{N}.$$

Then (x_k) converges weakly to a point contained in $(A+B)^{-1}(0)$.

Remark 2.7. In practice, it can be desirable to analyze an algorithm with respect to an auxiliary metric to encourage faster convergence. This is done by considering the metric induced by the inner product $\langle \cdot, \cdot \rangle_M$ corresponding to a symmetric positive definite operator $M\colon \mathcal{H} \to \mathcal{H}$. In the case of saddle point problems, for example, choosing the operator M to be a diagonal scaling matrix gives different weights to primal and dual variables. To keep our presentation as simple and as clear as possible, we present our analysis only for the case when $M = I$. Nevertheless, the more general case can be easily obtained through a straightforward modification of the proof. In particular, instead of (2.2), we can consider the iteration

$$x_{k+1} = J^M_{\lambda_k A}\left(x_k - M^{-1}\left[\lambda_k B(x_k) + \lambda_{k-1}(B(x_k) - B(x_{k-1}))\right]\right) \quad \forall k \in \mathbb{N},$$

where $J^M_A = (I + M^{-1}A)^{-1}$ denotes the *generalized resolvent* of A.

Remark 2.8. Since the main focus of this work lies in the development and analysis of new methods, we delay a more thorough computation comparison for future investigation. Nevertheless, the following example provides a specific problem for which the forward-reflected-backward method is faster than Tseng's method. We make no claims about the performance of the proposed method in general.

Consider (1.1) with $\mathcal{H} = \mathbb{R}^n \times \mathbb{R}^n$, $A(z_1, z_2) = (0, 0)$, and $B(z_1, z_2) = (z_2, -z_1)$. Note that zero is the unique solution to this problem and that the operator B is 1-Lipschitz. Let us also denote the identity operators on $\mathcal{H}$ and $\mathbb{R}^n$ by $I_{\mathcal{H}}$ and I_n, respectively. This is a classical example of a monotone inclusion, where the forward-backward method fails.

Tseng's method. By eliminating y_k from (1.10) and using the identity $B^2 = -I_{\mathcal{H}}$, Tseng's method can be expressed as

$$x_{k+1} = T(x_k) = T^{k+1}(x_0) \text{ where } T := (1-\lambda^2)I_{\mathcal{H}} - \lambda B.$$

Since $\langle x_k, B(x_k)\rangle = 0$ and $\|x_k\| = \|B(x_k)\|$, we have

$$\|x_{k+1}\|^2 = \|T(x_k)\|^2 = \big((1-\lambda^2)^2 + \lambda^2\big)\|x_k\|^2.$$

Let $\lambda \in (0,1)$. The sequence (x_k) therefore converges Q-linearly to zero with rate

$$\rho := \sqrt{(1-\lambda^2)^2 + \lambda^2} = \sqrt{1 - \lambda^2 + \lambda^4} < 1.$$

In fact, this shows the optimal stepsize is $\lambda = 1/\sqrt{2}$ which gives a rate of $\sqrt{3}/2$. (Note that the optimal rate does not occur for the largest possible stepsize.)

Forward-reflected-backward splitting. The forward-reflected-backward method with constant stepsize $\lambda \in (0, 1/2)$ can be expressed as

$$\begin{pmatrix} x_{k+1} \\ x_k \end{pmatrix} = T\begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix} = T^{k+1}\begin{pmatrix} x_0 \\ x_{-1} \end{pmatrix}, \quad \text{where } \; T := \begin{bmatrix} I_{\mathcal{H}} - 2\lambda B & \lambda B \\ I_{\mathcal{H}} & 0 \end{bmatrix}.$$

The eigenvalues of T are given by $\frac{1}{2} \pm \frac{1}{2} i \sqrt{8\,\lambda^2 - 1 - 4i\,\lambda\sqrt{1-4\,\lambda^2}}$. By choosing the stepsize $\lambda \approx 1/2$, we deduce that (x_k) converges R-linearly with a rate that be made arbitrarily close to $|\frac{1}{2} \pm \frac{1}{2}i| = \frac{1}{\sqrt{2}}$.

Since $1/\sqrt{2} < \sqrt{3}/2$, we conclude that the forward-reflected-backward method is faster than Tseng's method for this particular problem. Note that this comparison is in terms of the number of iterations.

2.1. Linear convergence. In this section, we establish R-linear convergence of the sequence generated by the forward-reflected-backward method when A is *strongly monotone*. Recall that $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is *m-strongly monotone* if $m > 0$ and

$$\langle x - y, u - v\rangle \geq m\,\|x - y\|^2 \quad \forall (x,u),(y,v) \in \operatorname{gra} A.$$

Strong monotonicity is a standard assumption for proving linear convergence of first order methods. We also note that there is no loss of generality in assuming that A is strongly monotone. For if B is m-strongly monotone, we can always augment the operators by the identity, i.e., $A + B = (A + mI) + (B - mI)$, without destroying monotonicity and Lipschitz continuity. Notice this does not complicate computing the resolvent of $(A + mI)$, as we have $J_{A+mI}(x) = J_{\frac{A}{1+m}}(\frac{x}{1+m})$ for all $x \in \mathcal{H}$.

THEOREM 2.9. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone and m-strongly monotone, let $B\colon \mathcal{H} \to \mathcal{H}$ be monotone and L-Lipschitz, and suppose $(A+B)^{-1}(0) \neq \varnothing$. Let $\lambda \in (0, \frac{1}{2L})$. Given $x_0, x_{-1} \in \mathcal{H}$, define the sequence (x_k) according to*

$$x_{k+1} = J_{\lambda A}\big(x_k - 2\lambda B(x_k) + \lambda B(x_{k-1})\big) \quad \forall k \in \mathbb{N}.$$

Then (x_k) converges R-linearly to the unique element of $(A+B)^{-1}(0)$.

Proof. Let $x \in (A+B)^{-1}(0)$. Using strong monotonicity of A (in place of monotonicity) in Proposition 2.3 and propagating the resulting inequality through the proof of Lemma 2.4 gives the inequality

$$\begin{aligned}(1+2m\lambda)\,\|x_{k+1} - x\|^2 + 2\lambda\,\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + (1-\lambda L)\,\|x_{k+1} - x_k\|^2 \\ \leq \|x_k - x\|^2 + 2\lambda\,\langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \frac{1}{2}\,\|x_k - x_{k-1}\|^2.\end{aligned}$$

By denoting $\varepsilon := \min\{\frac{1}{2} - \lambda L, 5m\lambda\} > 0$, this inequality implies

$$(1 + 4m\lambda)a_{k+1} + b_{k+1} + \varepsilon\|x_{k+1} - x_k\|^2 \le a_k + b_k, \tag{2.16}$$

where the nonnegative sequences (a_k) and (b_k) are given by

$$\begin{aligned} a_k &:= \frac{1}{2}\|x_k - x\|^2 \ge 0, \\ b_k &:= \frac{1}{2}\|x_k - x\|^2 + 2\lambda\langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2 \\ &\ge \frac{1}{2}\|x_k - x\|^2 - 2\lambda L\|x_k - x_{k-1}\|\,\|x_k - x\| + \frac{1}{2}\|x_k - x_{k-1}\|^2 \ge 0. \end{aligned}$$

Using Lipschitzness of B, we have

$$\begin{aligned} &(1 + 4m\lambda)a_{k+1} + b_{k+1} + \varepsilon\|x_{k+1} - x_k\|^2 \\ &\quad= \left(1 + 4m\lambda - \frac{\varepsilon}{2}\right)a_{k+1} + \left(1 + \frac{\varepsilon}{2}\right)b_{k+1} + \frac{3\varepsilon}{4}\|x_{k+1} - x_k\|^2 \\ &\qquad- \varepsilon\lambda\langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle \\ &\quad\ge \left(1 + 4m\lambda - \frac{\varepsilon}{2}\right)a_{k+1} + \left(1 + \frac{\varepsilon}{2}\right)b_{k+1} + \frac{3\varepsilon}{4}\|x_{k+1} - x_k\|^2 \\ &\qquad- \varepsilon\lambda L\|x_{k+1} - x_k\|\|x_{k+1} - x\| \\ &\quad\ge \left(1 + 4m\lambda - \frac{3\varepsilon}{4}\right)a_{k+1} + \left(1 + \frac{\varepsilon}{2}\right)b_{k+1} + \frac{\varepsilon}{2}\|x_{k+1} - x_k\|^2. \end{aligned} \tag{2.17}$$

Denote $\alpha := \min\{1 + 4m\lambda - 3\varepsilon/4, 1 + \varepsilon/2\} > 1$, which is true due to $\varepsilon \le 5m\lambda$. Combining (2.16) and (2.17) yields $\alpha(a_{k+1} + b_{k+1}) \le a_k + b_k$. Iterating this inequality gives

$$a_{k+1} \le a_{k+1} + b_{k+1} \le \frac{1}{\alpha}(a_k + b_k) \le \cdots \le \frac{1}{\alpha^{k+1}}(a_0 + b_0),$$

which establishes that $x_k \to x$ with R-linear rate. Since x was chosen arbitrarily from $(A + B)^{-1}(0)$, it must be unique. ∎

3. Forward-reflected-backward splitting with linesearch. The algorithm presented in the previous section required information about the single-valued operator's Lipschitz constant in order to select an appropriate stepsize. In practice, this requirement is undesirable for several reasons. First, obtaining the Lipschitz constant (or an estimate) is usually nontrivial and often a computationally expensive problem itself. Second, as a global constant, the (global) Lipschitz constant can often lead to overconservative stepsizes although local properties (around the current iterate) may permit the use of larger stepsizes and ultimately lead to faster convergence. Finally, when the single-valued operator is not Lipschitz continuous, any fixed stepsize scheme based on Lipschitz continuity will potentially fail to converge.

To address these shortcomings, most known methods can incorporate an additional procedure called *linesearch* (or *backtracking*) which is run in each iteration. It is worth noting, however, that in the more restrictive context of variational inequalities, the method proposed in [27] overcomes the aforementioned difficulties without resorting to a linesearch procedure.

In what follows, we show that the forward-reflected-backward method with such a linesearch procedure converges whenever the single-valued operator is *locally Lipschitz*.

Algorithm 3.1. The forward-reflected-backward method with linesearch.

Initialization: Choose $x_0, x_{-1} \in \mathcal{H}$, $\lambda_0, \lambda_{-1} > 0$, $\delta \in (0,1)$, and $\sigma \in (0,1)$.

Iteration: Having x_k, λ_{k-1}, and $B(x_{k-1})$, choose $\rho \in \{1, \sigma^{-1}\}$ and compute

$$x_{k+1} := J_{\lambda_k A}\big(x_k - \lambda_k B(x_k) - \lambda_{k-1}(B(x_k) - B(x_{k-1}))\big), \tag{3.1}$$

where $\lambda_k = \rho\lambda_{k-1}\sigma^i$ with i being the smallest nonnegative integer satisfying

$$\lambda_k \|B(x_{k+1}) - B(x_k)\| \leq \frac{\delta}{2}\|x_{k+1} - x_k\|. \tag{3.2}$$

Remark 3.1. The parameter ρ in Algorithm 3.1 has been introduced to allow for greater flexibility in the choice of possible stepsizes. Indeed, there are two possible scenarios for the value of λ_k in the first iteration of the linesearch procedure (i.e., when $i = 0$): either $\rho = \sigma^{-1}$ and $\lambda_k = \sigma^{-1}\lambda_{k-1} > \lambda_{k-1}$, or $\rho = 1$ and $\lambda_k = \lambda_{k-1}$. The former, more aggressive scenario allows for the possibility of larger stepsizes at the price of a potential increase in the number of linesearch iterations.

The following lemma shows that the linesearch procedure described in Algorithm 3.1 is well-defined so long as the operator B is locally Lipschitz continuous.

LEMMA 3.2. *Suppose $B\colon \mathcal{H} \to \mathcal{H}$ is locally Lipschitz. Then the linesearch procedure in* (3.1)–(3.2) *always terminates, i.e., (λ_k) is well-defined.*

Proof. Denote $x_{k+1}(\lambda) := J_{\lambda A}(x_k - \lambda B(x_k) - \lambda_{k-1}(B(x_k) - B(x_{k-1})))$. From [5, Theorem 23.47], we have that $J_{\lambda A}(x_{k+1}(0)) \to P_{\overline{\operatorname{dom} A}}(x_{k+1}(0))$ as $\lambda \searrow 0$ which, together with the nonexpansivity of $J_{\lambda A}$, yields

$$\begin{aligned}
&\big\|x_{k+1}(\lambda) - P_{\overline{\operatorname{dom} A}}\, x_{k+1}(0)\big\| \\
&\quad \leq \|x_{k+1}(\lambda) - J_{\lambda A}(x_{k+1}(0))\| + \big\|J_{\lambda A}(x_{k+1}(0)) - P_{\overline{\operatorname{dom} A}}(x_{k+1}(0))\big\| \\
&\quad \leq \lambda\|B(x_k)\| + \big\|J_{\lambda A}(x_{k+1}(0)) - P_{\overline{\operatorname{dom} A}}(x_{k+1}(0))\big\|.
\end{aligned}$$

By taking the limit as $\lambda \searrow 0$, we deduce that $x_{k+1}(\lambda) \to P_{\overline{\operatorname{dom} A}}(x_{k+1}(0))$.

Now, by way of a contradiction, suppose that the linesearch procedure in Algorithm 3.1 fails to terminate at the kth iteration. Then, for all $\lambda = \rho\lambda_{k-1}\sigma^i$ with $i = 0, 1, \ldots$, we have

$$\rho\lambda_{k-1}\sigma^i \|B(x_{k+1}(\lambda)) - B(x_k)\| > \frac{\delta}{2}\|x_{k+1}(\lambda) - x_k\|. \tag{3.3}$$

On one hand, taking the limit as $i \to \infty$ in (3.1) gives $P_{\overline{\operatorname{dom} A}}(x_{k+1}(0)) = x_k$. On the other hand, since B is locally Lipschitz at x_k there exists $L > 0$ such that for i sufficiently large, we have

$$\rho\lambda_{k-1}\sigma^i \|B(x_{k+1}(\lambda)) - B(x_k)\| > \frac{\delta}{2}\|x_{k+1}(\lambda) - x_k\| \geq \frac{\delta L}{2}\|B(x_{k+1}(\lambda)) - B(x_k)\|.$$

Dividing both sides by $\|B(x_{k+1}(\lambda)) - B(x_k)\|$ gives $\delta L/2 < \rho\lambda_{k-1}\sigma^i$. Since $\sigma^i \to 0$ as $i \to \infty$, this inequality gives a contradiction which completes the proof. ∎

The next lemma is a direct extension of Lemma 2.4.

LEMMA 3.3. *Let $x \in (A+B)^{-1}(0)$ and let (x_k) be generated by Algorithm* 3.1. *Then there exists $\varepsilon > 0$ such that, for all $k \in \mathbb{N}$, we have*

$$\begin{aligned}(3.4)\qquad &\|x_{k+1}-x\|^2 + 2\lambda_k \langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + \left(\frac{1}{2}+\varepsilon\right)\|x_{k+1}-x_k\|^2 \\ &\leq \|x_k - x\|^2 + 2\lambda_{k-1}\langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2.\end{aligned}$$

Proof. The proof is exactly the same as for Lemma 2.4 with the only change being that instead of using Lipschitzness of B to deduce the inequality (2.10), we use (3.2), which is well-defined due to Lemma 3.2. ∎

THEOREM 3.4. *Let $\mathcal{H}$ be finite dimensional, $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, and $B\colon \mathcal{H} \to \mathcal{H}$ be monotone and locally Lipschitz continuous, and suppose that $(A+B)^{-1}(0) \neq \varnothing$. Then the sequence (x_k) generated by Algorithm* 3.1 *converges to a point contained in $(A+B)^{-1}(0)$.*

Proof. We argue similarly to Theorem 2.5 but using Lemma 3.3 in place of Lemma 2.4, and (3.2) in place of Lipschitzness of B. This yields (2.13) from which we deduce that (x_k) is bounded and $\|x_k - x_{k+1}\| \to 0$. As a locally Lipschitz operator on finite dimensional space, B is Lipschitz on bounded sets. Thus, since (x_k) is bounded, there exists a constant $L > 0$ such that

$$(3.5)\qquad \|B(x_{k+1}) - B(x_k)\| \leq L\|x_{k+1} - x_k\| \quad \forall k \in \mathbb{N}.$$

By combining (3.2) and (3.5), we see that (λ_k) is bounded away from zero. The remainder of the proof is the same as Theorem 2.5. ∎

4. Relaxed inertial forward-reflected-backward splitting. In this section, we consider a relaxed inertial variant of the forward-reflected-backward splitting algorithm. Such variants are of interest in practice because they have the potential to improve performance as well as the range of admissible stepsizes. A treatment of a relaxed inertial variant of the forward-backward method with B cocoercive and its relation to Nesterov-type acceleration techniques can be found in [4].

Consider the monotone inclusion

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

where $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is maximally monotone, and $B\colon \mathcal{H} \to \mathcal{H}$ is either monotone and L-Lipschitz continuous or $1/L$-cocoercive. The relaxed inertial algorithm is given by

$$(4.1)\qquad \begin{cases} z_{k+1} := J_{\lambda A}\left(x_k - \lambda B(x_k) - \dfrac{\lambda}{\beta}\left(B(x_k) - B(x_{k-1})\right) + \dfrac{\alpha}{\beta}\left(x_k - x_{k-1}\right)\right), \\ x_{k+1} := (1-\beta)x_k + \beta z_{k+1} \end{cases}$$

for all $k \in \mathbb{N}$ and for appropriately chosen parameters $\alpha \geq 0$ and $\beta, \lambda > 0$ whose precise form depends on the properties of B. By denoting $B' := B - \frac{\alpha}{\lambda} I$, the scheme can be expressed as

$$(4.2)\qquad \begin{cases} z_{k+1} := J_{\lambda A}\left(x_k - \lambda B(x_k) - \dfrac{\lambda}{\beta}\left(B'(x_k) - B'(x_{k-1})\right)\right) \\ x_{k+1} := (1-\beta)x_k + \beta z_{k+1} \end{cases} \qquad \forall k \in \mathbb{N}.$$

To prove convergence of this scheme, we first prove two lemmas.

LEMMA 4.1. *Suppose $B\colon \mathcal{H} \to \mathcal{H}$ is monotone and $\rho \geq 0$. Then the operator $B' := B - \rho I$ is L'-Lipschitz with L' given by*

$$L' := \begin{cases} L + \rho & \text{if } B \text{ is } L\text{-Lipschitz}, \\ L - \rho & \text{if } B \text{ is } 1/L\text{-cocoercive and } \rho \leq \frac{L}{2}, \\ \rho & \text{if } B \text{ is } 1/L\text{-cocoercive and } \rho > \frac{L}{2}. \end{cases} \tag{4.3}$$

Proof. Let $x, y \in \mathcal{H}$. When B is L-Lipschitz, we have

$$\|(B - \rho I)x - (B - \rho I)y\| \leq \|Bx - By\| + \rho \|x - y\| \leq (L + \rho) \|x - y\|,$$

which establishes the first case. For the second and third cases, first observe that $1/L$-cocoercivity of B yields

$$\begin{aligned} \|(B - \rho I)x - (B - \rho I)y\|^2 &= \|Bx - By\|^2 - 2\rho \langle Bx - By, x - y\rangle + \rho^2 \|x - y\|^2 \\ &\leq \left(1 - \frac{2\rho}{L}\right) \|Bx - By\|^2 + \rho^2 \|x - y\|^2. \end{aligned}$$

On one hand, if $\rho > \frac{L}{2}$, then $1 - \frac{2\rho}{L} < 0$ and ρ-Lipschitzness of B' follows. On the other hand, if $\rho \leq \frac{L}{2}$, then

$$\begin{aligned} \|(B - \rho I)x - (B - \rho I)y\|^2 &\leq \left(L^2 - \frac{2\rho}{L} L^2 + \rho^2\right) \|x - y\|^2 \\ &= (L - \rho)^2 \|x - y\|^2, \end{aligned}$$

which shows that B' is $(L - \rho)$-Lipschitz. The proof is now complete. ∎

Note that it is possible to slightly improve the estimate of L' in the case when B is L-Lipschitz and monotone to $L' = \sqrt{L^2 + \rho^2}$. However, the benefits of using a new bound are minimal since, in our case, ρ will take small values relative to L.

In the following lemma, we use the following form of (2.8) from Proposition 2.3:

$$\begin{aligned} \|d_2 - x\|^2 + 2\langle u - u_1, x - d_2\rangle \leq \|d_1 - x\|^2 + 2\langle v_1 - u_0, x - d_1\rangle \\ + 2\langle v_1 - u_0, d_1 - d_2\rangle - \|d_1 - d_2\|^2. \end{aligned} \tag{4.4}$$

LEMMA 4.2. *Let $x \in (A + B)^{-1}(0)$, let (x_k) be given by (4.2), and consider constants $\alpha \geq 0$ and $\beta, \lambda > 0$. Then*

$$\begin{aligned} &(1 - \alpha) \|x_{k+1} - x\|^2 + 2\lambda \langle B'(x_{k+1}) - B'(x_k), x - x_{k+1}\rangle + b_{k+1} \\ &\quad \leq (1 - \alpha)\|x_k - x\|^2 + 2\lambda \langle B'(x_k) - B'(x_{k-1}), x - x_k\rangle + b_k \\ &\qquad + \frac{\lambda L'}{\beta} \|x_k - x_{k-1}\|^2 - \left(\frac{2 - \beta - \lambda L'}{\beta} - \alpha\right) \|x_{k+1} - x_k\|^2, \end{aligned}$$

where $b_k := 2\lambda \langle B(x) - B(x_k), x - x_k\rangle \geq 0$ and L' is the Lipschitz constant of B'.

Proof. Let $u := \lambda B(x)$. As $u \in -\lambda A(x)$, applying (4.4) with

$$\begin{aligned} F &:= \lambda A & d_1 &:= x_k & u_0 &:= (\lambda/\beta) B'(x_{k-1}) \\ u_1 &:= \lambda B(x_k) & d_2 &:= z_{k+1} & v_1 &:= (\lambda/\beta) B'(x_k) \end{aligned}$$

yields the inequality

$$\begin{aligned}(4.5)\quad &\|z_{k+1}-x\|^2+2\lambda\langle B(x)-B(x_k),x-z_{k+1}\rangle \le \|x_k-x\|^2-\|x_k-z_{k+1}\|^2\\ &+\frac{2\lambda}{\beta}\langle B'(x_k)-B'(x_{k-1}),x-x_k\rangle+\frac{2\lambda}{\beta}\langle B'(x_k)-B'(x_{k-1}),x_k-z_{k+1}\rangle.\end{aligned}$$

Using the identity $z_{k+1}=\big(x_{k+1}-(1-\beta)x_k\big)/\beta$ and the definition of B', the second term in (4.5) can be expressed as

$$\begin{aligned}&2\lambda\langle B(x)-B(x_k),x-z_{k+1}\rangle+\frac{1-\beta}{\beta}b_k-\frac{1}{\beta}b_{k+1}\\ &=\frac{2\lambda}{\beta}\langle B'(x_{k+1})-B'(x_k),x-x_{k+1}\rangle+\frac{2\alpha}{\beta}\langle x_{k+1}-x_k,x-x_{k+1}\rangle\\ &=\frac{2\lambda}{\beta}\langle B'(x_{k+1})-B'(x_k),x-x_{k+1}\rangle+\frac{\alpha}{\beta}\big(\|x_k-x\|^2-\|x_{k+1}-x_k\|^2-\|x_{k+1}-x\|^2\big).\end{aligned}$$

Substituting this back into (4.5) and using $z_{k+1}-x_k=(x_{k+1}-x_k)/\beta$ gives

$$\begin{aligned}(4.6)\quad &\|z_{k+1}-x\|^2-\frac{\alpha}{\beta}\|x_{k+1}-x\|^2+\frac{2\lambda}{\beta}\langle B'(x_{k+1})-B'(x_k),x-x_{k+1}\rangle+\frac{1}{\beta}b_{k+1}\\ &\le\left(1-\frac{\alpha}{\beta}\right)\|x_k-x\|^2+\frac{2\lambda}{\beta}\langle B'(x_k)-B'(x_{k-1}),x-x_k\rangle+\frac{1-\beta}{\beta}b_k\\ &\quad+\frac{2\lambda}{\beta}\langle B'(x_k)-B'(x_{k-1}),x_k-z_{k+1}\rangle+\left(\frac{\alpha}{\beta}-\frac{1}{\beta^2}\right)\|x_{k+1}-x_k\|^2.\end{aligned}$$

By using the identity

$$\|z_{k+1}-x\|^2=\frac{1}{\beta}\|x_{k+1}-x\|^2-\frac{1-\beta}{\beta}\|x_k-x\|^2+\frac{1-\beta}{\beta^2}\|x_{k+1}-x_k\|^2$$

in (4.6) and multiplying both sides by β, we obtain

$$\begin{aligned}(4.7)\quad &(1-\alpha)\|x_{k+1}-x\|^2+2\lambda\langle B'(x_{k+1})-B'(x_k),x-x_{k+1}\rangle+b_{k+1}\\ &\le(1-\alpha)\|x_k-x\|^2+2\lambda\langle B'(x_k)-B'(x_{k-1}),x-x_k\rangle+(1-\beta)b_k\\ &\quad+2\lambda\langle B'(x_k)-B'(x_{k-1}),x_k-z_{k+1}\rangle-\left(\frac{2-\beta}{\beta}-\alpha\right)\|x_{k+1}-x_k\|^2.\end{aligned}$$

Since B' is L'-Lipschitz, the second-last term can be estimated by

$$\begin{aligned}2\lambda\langle B'(x_k)-B'(x_{k-1}),x_k-z_{k+1}\rangle&=\frac{2\lambda}{\beta}\langle B'(x_k)-B'(x_{k-1}),x_k-x_{k+1}\rangle\\ &\le\frac{\lambda L'}{\beta}\left(\|x_k-x_{k-1}\|^2+\|x_{k+1}-x_k\|^2\right).\end{aligned}$$

The claimed inequality follows by substituting this estimate back into (4.7). ∎

THEOREM 4.3. *Let $A\colon\mathcal{H}\rightrightarrows\mathcal{H}$ be maximally monotone and let $B\colon\mathcal{H}\to\mathcal{H}$ be monotone with $(A+B)^{-1}(0)\neq\varnothing$. Suppose $\alpha\in[0,1)$, $\beta\in(0,1]$, $\lambda>0$ and either*

(a) *B is L-Lipschitz and*

$$(4.8)\qquad \lambda<\min\left\{\frac{2-\beta-\alpha\beta-2\alpha}{2L},\frac{1-\alpha-\alpha\beta}{\beta L}\right\},\quad \textit{or}$$

(b) *B is $(1/L)$-cocoercive, $\alpha < \frac{2-\beta}{2+\beta}$, and*

$$\lambda < \min\left\{\frac{2-\beta-\alpha\beta+2\alpha}{2L}, \frac{1-\alpha+\alpha\beta}{\beta L}\right\}. \tag{4.9}$$

Given $x_0, x_{-1} \in \mathcal{H}$, define the sequences (x_k) and (z_k) according to (4.1) *Then (x_k) converges weakly to a point in $(A+B)^{-1}(0)$.*

Proof. Let L' denote the Lipschitz constant of $B' := B - \frac{\alpha}{\lambda} I$. By combining Lemma 4.1 with the assumptions in each of the respective cases, we obtain

$$\varepsilon := \min\left\{\frac{2-\beta(1+\alpha)}{2}, \frac{1-\alpha}{\beta}\right\} - \lambda L' > 0. \tag{4.10}$$

Then Lemma 4.2 together with (4.10) gives

$$\begin{aligned}
&(1-\alpha)\left\|x_{k+1}-x\right\|^2 + 2\lambda\left\langle B'(x_{k+1}) - B'(x_k), x - x_{k+1}\right\rangle + b_{k+1} \\
&\quad \le (1-\alpha)\|x_k - x\|^2 + 2\lambda\left\langle B'(x_k) - B'(x_{k-1}), x - x_k\right\rangle + b_k \\
&\qquad + \frac{\lambda L'}{\beta}\left\|x_k - x_{k-1}\right\|^2 - \left(\frac{\lambda L'}{\beta} + \varepsilon\right)\left\|x_{k+1} - x_k\right\|^2,
\end{aligned}$$

which telescopes to yield

$$\begin{aligned}
&(1-\alpha)\left\|x_{k+1}-x\right\|^2 + 2\lambda\left\langle B'(x_{k+1}) - B'(x_k), x - x_{k+1}\right\rangle + b_{k+1} \\
&\quad \le (1-\alpha)\|x_0 - x\|^2 + 2\lambda\left\langle B'(x_0) - B'(x_{-1}), x - x_0\right\rangle + b_0 \\
&\qquad + \frac{\lambda L'}{\beta}\left\|x_0 - x_{-1}\right\|^2 - \frac{\lambda L'}{\beta}\left\|x_{k+1} - x_k\right\|^2 - \varepsilon\sum_{i=1}^{k}\left\|x_{i+1} - x_i\right\|^2.
\end{aligned} \tag{4.11}$$

The L'-Lipschitz continuity of B' together with (4.10) gives

$$\begin{aligned}
2\lambda\left\langle B'(x_{k+1}) - B'(x_k), x - x_{k+1}\right\rangle &\ge -2\lambda L'\left\|x_{k+1} - x_k\right\|\left\|x_{k+1} - x\right\| \\
&\ge -\frac{\lambda L'}{\beta}\left\|x_{k+1} - x_k\right\|^2 - \beta\lambda L'\left\|x_{k+1} - x\right\|^2 \\
&\ge -\frac{\lambda L'}{\beta}\left\|x_{k+1} - x_k\right\|^2 - (1-\alpha-\beta\varepsilon)\left\|x_{k+1} - x\right\|^2.
\end{aligned}$$

This, together with (4.11) and the fact that $b_{k+1} \ge 0$, yields the inequality

$$\begin{aligned}
&\beta\varepsilon\left\|x_{k+1} - x\right\|^2 + \varepsilon\sum_{i=1}^{k}\left\|x_{i+1} - x_i\right\|^2 \\
&\quad \le (1-\alpha)\|x_0 - x\|^2 + 2\lambda\left\langle B'(x_0) - B'(x_{-1}), x - x_0\right\rangle + b_0 + \frac{\lambda L'}{\beta}\left\|x_0 - x_{-1}\right\|^2,
\end{aligned}$$

which shows that (x_k) is bounded and $\|x_k - x_{k+1}\| \to 0$. The remainder of the proof follows a similar argument to Theorem 2.5. ∎

The admissible values of α and β for the two cases in Theorem 4.3 are shown in Figure 1. By setting $\beta = 1$ in Theorem 4.3, we obtain the following inertial algorithm.

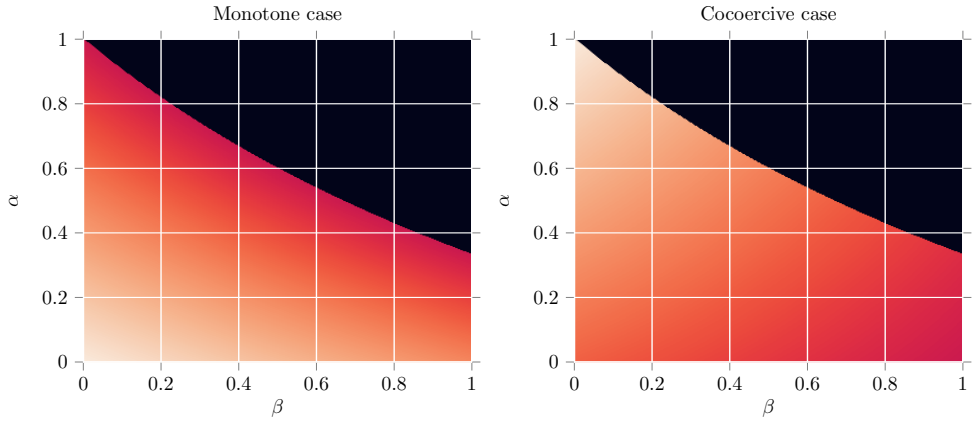

FIG. 1. *The upper bound for admissible values for λL as a function of α and β according to* (4.8) *and* (4.9). *Black regions denote infeasible combinations. Lighter colors indicate a higher admissible value.*

COROLLARY 4.4. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone and let $B\colon \mathcal{H} \to \mathcal{H}$ be monotone with $(A+B)^{-1}(0) \neq \varnothing$. Suppose $\alpha \in [0, 1/3)$, $\lambda > 0$, and either*

(a) *B is L-Lipschitz and $\lambda < \frac{1-3\alpha}{2L}$, or*

(b) *B is $(1/L)$-cocoercive and $\lambda < \frac{1+\alpha}{2L}$.*

Given $x_0, x_{-1} \in \mathcal{H}$, define the sequence (x_k) according to

$$x_{k+1} := J_{\lambda A}\big(x_k - 2\lambda B(x_k) + \lambda B(x_{k-1}) + \alpha(x_k - x_{k-1})\big). \tag{4.12}$$

Then (x_k) converges weakly to a point contained in $(A+B)^{-1}(0)$.

Remark 4.5. Although Corollary 4.4 establishes that inertia increases the range of admissible stepsizes when B is cocoercive, it has the opposite effect when B is merely monotone. A similar phenomena with Tseng's method was observed in [7].

Remark 4.6. By setting $B = 0$ in Corollary 4.4, the scheme (4.12) reduces to the classical *inertial proximal algorithm* first considered in [1]. It is interesting to note that the proof presented here does not follow the technique from [1], which is used in the analysis of most other first order inertial operator splitting methods [7, 25, 30].

5. Three operator splitting. In this section, we consider a structured three operator monotone inclusion. Specifically, we consider the inclusion

$$\text{find } x \in \mathcal{H} \text{ such that } 0 \in (A+B+C)(x), \tag{5.1}$$

where $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ is maximal monotone, $B\colon \mathcal{H} \to \mathcal{H}$ is monotone and L_1-Lipschitz, and $C\colon \mathcal{H} \to \mathcal{H}$ is $1/L_2$-cocoercive. This problem could be solved using the two operator splitting algorithm in section 2 applied to A and $(B+C)$, where we note that $(B+C)$ is L-Lipschitz continuous with $L = L_1 + L_2$. Consequently, to apply Theorem 2.5, the stepsize λ should satisfy

$$\lambda < \frac{1}{2L} = \frac{1}{2L_1 + 2L_2}.$$

In this section, we show that this can be improved by exploiting the additional structure in (5.1). Indeed, we propose a modification which only requires $\lambda > 0$ to satisfy

$$\lambda < \frac{2}{4L_1 + L_2} = \frac{1}{2L_1 + \frac{1}{2}L_2}.$$

Given initial points $x_0, x_{-1} \in \mathcal{H}$, our modified scheme is given by

$$x_{k+1} = J_{\lambda A}\big(x_k - 2\lambda B(x_k) + \lambda B(x_{k-1}) - \lambda C(x_k)\big) \quad \forall k \in \mathbb{N}. \tag{5.2}$$

In other words, the algorithm only uses a standard forward step of the operator C, as is employed in the forward-backward method (1.9). For an algorithm for the case when C is Lipschitz but not necessarily cocoercive, see [37].

We begin our analysis with the three operator analogue of Lemma 2.4.

LEMMA 5.1. *Let $x \in (A+B+C)^{-1}(0)$ and let the sequence (x_k) be given by (5.2). Suppose $\lambda \in (0, \frac{2}{4L_1+L_2})$. Then there exists an $\varepsilon > 0$ such that, for all $k \in \mathbb{N}$, we have*

$$\begin{aligned}\|x_{k+1} - x\|^2 + 2\lambda \langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + (\lambda L_1 + \varepsilon)\|x_{k+1} - x_k\|^2 \\ \leq \|x_k - x\|^2 + 2\lambda \langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \lambda L_1 \|x_k - x_{k-1}\|^2.\end{aligned}$$

Proof. Since $0 \in (A + B + C)(x)$, we have $-(B + C)(x) \in A(x)$. Combined with the monotonicity of A, this gives

$$0 \leq \langle x_{k+1} - x_k + \lambda(B + C)(x_k) + \lambda(B(x_k) - B(x_{k-1})) - \lambda(B + C)(x), x - x_{k+1}\rangle,$$

which we rewrite as

$$\begin{aligned}0 \leq\ & \langle x_{k+1} - x_k, x - x_{k+1}\rangle + \lambda \langle B(x_k) - B(x), x - x_{k+1}\rangle \\ & + \lambda \langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \lambda \langle B(x_k) - B(x_{k-1}), x_k - x_{k+1}\rangle \\ & + \lambda \langle C(x_k) - C(x), x - x_k\rangle + \lambda \langle C(x_k) - C(x), x_k - x_{k+1}\rangle.\end{aligned} \tag{5.3}$$

The first through fourth terms can be estimated as in Lemma 2.4. Using $1/L_2$-cocoercivity of C, the fifth term can be estimated as

$$\langle C(x_k) - C(x), x - x_k\rangle \leq -\frac{1}{L_2}\|C(x_k) - C(x)\|^2,$$

and the final term can be estimated as

$$\begin{aligned}\langle C(x_k) - C(x), x_k - x_{k+1}\rangle &\leq \|C(x_k) - C(x)\|\,\|x_{k+1} - x_k\| \\ &\leq \frac{1}{L_2}\|C(x_k) - C(x)\|^2 + \frac{L_2}{4}\|x_{k+1} - x_k\|^2.\end{aligned}$$

Thus, altogether, (5.3) implies that

$$\begin{aligned}0 \leq\ & \|x_k - x\|^2 - \|x_{k+1} - x_k\|^2 - \|x_{k+1} - x\|^2 - 2\lambda \langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle \\ & + 2\lambda \langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \lambda L_1 \left(\|x_k - x_{k-1}\|^2 + \|x_{k+1} - x_k\|^2\right) \\ & - \frac{2\lambda}{L_2}\|C(x_k) - C(x)\|^2 + \lambda\left(\frac{2}{L_2}\|C(x_k) - C(x)\|^2 + \frac{L_2}{2}\|x_{k+1} - x_k\|^2\right),\end{aligned}$$

which, on rearranging, gives

$$\|x_{k+1}-x\|^2 + 2\lambda\langle B(x_{k+1})-B(x_k), x-x_{k+1}\rangle + \left(1-\lambda L_1-\frac{\lambda L_2}{2}\right)\|x_{k+1}-x_k\|^2$$
$$\leq \|x_k-x\|^2 + 2\lambda\langle B(x_k)-B(x_{k-1}), x-x_k\rangle + \lambda L_1\|x_k-x_{k-1}\|^2.$$

The claimed inequality follows with $\varepsilon := (1-\lambda L_1-\frac{\lambda L_2}{2}) - \lambda L_1 = 1-2\lambda L_1 - \frac{\lambda L_2}{2} > 0$. ∎

The following theorem is our main result regarding convergence of the three operator splitting scheme.

THEOREM 5.2. *Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, let $B\colon \mathcal{H}\to\mathcal{H}$ be monotone and L_1-Lipschitz, and let $C\colon \mathcal{H}\to\mathcal{H}$ be $1/L_2$-cocoercive. Suppose that $(A+B+C)^{-1}(0)\neq\varnothing$ and $\lambda\in(0,\frac{2}{4L_1+L_2})$. Given $x_0, x_{-1}\in\mathcal{H}$, define the sequence (x_k) according to*

$$x_{k+1} = J_{\lambda A}\big(x_k - 2\lambda B(x_k) + \lambda B(x_{k-1}) - \lambda C(x_k)\big) \quad \forall k\in\mathbb{N}.$$

Then (x_k) converges weakly to a point contained in $(A+B+C)^{-1}(0)$.

Proof. The proof is more or less the same as Theorem 2.5 but uses Lemma 5.1 in place of Lemma 2.4. The only other thing to check is that

$$\|x_{k+1}-x\|^2 + 2\lambda\langle B(x_{k+1})-B(x_k), x-x_{k+1}\rangle + \lambda L_1\|x_{k+1}-x_k\|^2$$

is bounded from below by zero. To see this, observe that

$$2\lambda\langle B(x_{k+1})-B(x_k), x-x_{k+1}\rangle \geq -2\lambda L_1\|x_{k+1}-x_k\|\,\|x_{k+1}-x\|$$
$$\geq -\lambda L_1\left(\|x_{k+1}-x_k\|^2 + \|x_{k+1}-x\|^2\right),$$

and $1-\lambda L_1 > 0$ since $\lambda < \frac{2}{4L_1+L_2} < \frac{1}{L_1}$. ∎

6. Between forward-backward and forward-reflected-backward. In this section, we consider a variant of the forward-reflected-backward method for a structured version of the monotone inclusion (1.1) in a separable Hilbert space $\mathcal{H}$. Precisely, we assume that the second operator $B\colon \mathcal{H}\to\mathcal{H}$ is monotone and decomposable in the form

$$B = \frac{1}{n}\sum_{i=1}^{n} B_i, \tag{6.1}$$

where $B_i\colon \mathcal{H}\to\mathcal{H}$ is L-Lipschitz continuous for $i=1,\dots,n$. In what follows, we analyze the following iteration:

$$\begin{cases} \text{Choose } i_k \text{ uniformly at random from } \{1,\dots,n\} \\ x_{k+1} = J_{\lambda A}\big(x_k - \lambda B(x_k) - \lambda(B_{i_k}(x_k) - B_{i_k}(x_{k-1}))\big) \end{cases} \quad \forall k\in\mathbb{N}. \tag{6.2}$$

In other words, the term $B(x_k)-B(x_{k-1})$ inside the resolvent is replaced by $B_{i_k}(x_k)-B_{i_k}(x_{k-1})$. Although it is unclear if (6.2) has any practical value as it still requires one full evaluation of B in every iteration, it is surprising that such a small random perturbation still ensures its (almost sure) convergence without cocoercivity. Indeed, without this correction, the algorithm reduces to the forward-backward method which, in general, need not converge in this setting. The fact that (6.2) converges suggests that the exact form of the correction to values of B may not be important.

In what follows, given a random variable X, $\mathbb{E}[X]$ denotes its expectation and $\mathbb{E}_k[X]$ denotes its conditional expectation with respect to the σ-algebra generated by the random variables $x_1, x_2, \dots, x_k$.

LEMMA 6.1. *Let $x \in (A+B)^{-1}(0)$ and let (x_k) be given by (6.2). Suppose $\lambda \in (0, \frac{1}{2L})$. Then there exists an $\varepsilon > 0$ such that, for all $k \in \mathbb{N}$, we have*

$$\|x_{k+1} - x\|^2 + 2\lambda \langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + \left(\frac{1}{2} + \varepsilon\right)\|x_{k+1} - x_k\|^2 \leq \|x_k - x\|^2 + 2\lambda \langle B_{i_k}(x_k) - B_{i_k}(x_{k-1}), x - x_k\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2. \tag{6.3}$$

Proof. By applying Proposition 2.3 with

$$F := \lambda A \qquad d_1 := x_k \qquad u_0 := \lambda B_{i_k}(x_{k-1}) \qquad v_1 := \lambda B_{i_k}(x_k)$$
$$u := \lambda B(x) \qquad d_2 := x_{k+1} \qquad u_1 := \lambda B(x_k) \qquad v_2 := \lambda B(x_{k+1}),$$

we obtain the inequality

$$\begin{aligned}\|x_{k+1} - x\|^2 + 2\lambda \langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + \|x_{k+1} - x_k\|^2 \\ \leq \|x_k - x\|^2 + 2\lambda \langle B_{i_k}(x_k) - B_{i_k}(x_{k-1}), x - x_k\rangle \\ + 2\lambda \langle B_{i_k}(x_k) - B_{i_k}(x_{k-1}), x_k - x_{k+1}\rangle - 2\lambda \langle B(x_{k+1}) - B(x), x_{k+1} - x\rangle.\end{aligned}$$

Since B is monotone, the last term is nonnegative. Using Lipschitzness of B_{i_k}, the second-last term can be estimated as

$$\begin{aligned}\langle B_{i_k}(x_k) - B_{i_k}(x_{k-1}), x_k - x_{k+1}\rangle &\leq L\|x_k - x_{k-1}\|\,\|x_k - x_{k+1}\| \\ &\leq \frac{L}{2}\left(\|x_k - x_{k-1}\|^2 + \|x_k - x_{k+1}\|^2\right).\end{aligned}$$

Thus, altogether, we obtain

$$\begin{aligned}\|x_{k+1} - x\|^2 + 2\lambda \langle B(x_{k+1}) - B(x_k), x - x_{k+1}\rangle + (1 - \lambda L)\|x_{k+1} - x_k\|^2 \\ \leq \|x_k - x\|^2 + 2\lambda \langle B_{i_k}(x_k) - B_{i_k}(x_{k-1}), x - x_k\rangle + \lambda L\|x_k - x_{k-1}\|^2.\end{aligned}$$

The claimed inequality follows since $\lambda L < \frac{1}{2}$ and $1 - \lambda L < \frac{1}{2}$. ∎

THEOREM 6.2. *Suppose $\mathcal{H}$ is separable. Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, and let $B\colon \mathcal{H} \to \mathcal{H}$ be monotone with $B = \sum_{i=1}^n B_i$ for L-Lipschitz continuous operators $B_i\colon \mathcal{H} \to \mathcal{H}$. Suppose that $(A+B)^{-1}(0) \neq \varnothing$ and that $\lambda \in (0, \frac{1}{2L})$. Given $x_0, x_{-1} \in \mathcal{H}$, define the sequence (x_k) according to (6.2). Then (x_k) converges weakly almost surely to a point contained in $(A+B)^{-1}(0)$.*

Proof. Let $x \in (A+B)^{-1}(0)$ and let $(\varphi_k) \subseteq \mathbb{R}$ denote the sequence of random variables given by

$$\varphi_k := \|x_k - x\|^2 + 2\lambda \langle B(x_k) - B(x_{k-1}), x - x_k\rangle + \frac{1}{2}\|x_k - x_{k-1}\|^2 \geq \frac{1}{2}\|x_k - x\|^2, \tag{6.4}$$

where the latter inequality is due to (2.13). Taking conditional expectation in Lemma 6.1 gives

$$\mathbb{E}_k[\varphi_{k+1}] + \varepsilon \mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right] \leq \varphi_k.$$

The supermartingale convergence theorem [35, Theorem 1] then implies that, almost surely, (φ_k) converges to a nonnegative-valued random variable φ and that $\sum_{k=1}^\infty \mathbb{E}_k[\|x_{k+1} - x_k\|^2] < \infty$. The latter implies that $\|x_{k+1} - x_k\|^2 \to 0$ almost

surely. From (6.4), it then follows that (x_k) is bounded almost surely and that $(\|x_k - x\|^2)$ converges almost surely to φ.

Now, consider a realization $(x_k(\omega))$ of (x_k) such that $\|x_{k+1}(\omega) - x_k(\omega)\| \to 0$ and $\varphi_k(\omega) \to \varphi(\omega)$ for some $\varphi(\omega) \geq 0$ (where $\varphi_k(\omega)$ denotes the corresponding realization of φ_k). Let $\overline{x}(\omega)$ be a sequential weak cluster point of the bounded sequence $(x_k(\omega))$. From (6.2), we have

$$\begin{aligned}(6.5)\quad &\frac{1}{\lambda}\big(x_{k-1}(\omega) - x_k(\omega) + \lambda\left(B(x_k(\omega)) - B(x_{k-1}(\omega))\right)\\ &\quad + \lambda\left(B_{i_{k-1}}(x_{k-2}(\omega)) - B_{i_{k-1}}(x_{k-1}(\omega))\right)\big) \in (A+B)(x_k(\omega)) \qquad \forall k \geq 1.\end{aligned}$$

Since the graph of $A + B$ is demiclosed and $B_1, \ldots, B_n$ are Lipschitz, taking the limit along a subsequence of $(x_k(\omega))$ which converges to $\overline{x}(\omega)$ in (6.5) yields $\bar{x}(\omega) \in (A + B)^{-1}(0)$. Altogether, we have that the weak sequential cluster points of (x_k) are almost surely contained in $(A+B)^{-1}(0)$. An argument analogous to [14, Proposition 2.3] then shows that (x_k) converges weakly almost surely to a $(A+B)^{-1}(0)$-valued random variable. ∎

7. Concluding remarks. In this work, we have proposed a modification of the forward-backward algorithm for finding a zero in the sum of two monotone operators which does not require cocoercivity. To conclude, we outline three possible directions for further research into the method.

Fixed point interpretations. As the proof of the forward-reflected-backward method does not conform to the usual Krasnoselskii–Mann framework, it would be interesting to see if the method can be analyzed from the perspective of fixed point theory. To this end, consider the two operators $M, T \colon \mathcal{H} \times \mathcal{H} \to \mathcal{H} \times \mathcal{H}$ given by

$$M := \begin{bmatrix} J_{\lambda A} & 0 \\ 0 & I \end{bmatrix}, \quad T := \begin{bmatrix} I - 2\lambda B & \lambda I \\ B & 0 \end{bmatrix}.$$

By introducing the auxiliary variable $u_{k+1} := B(x_k)$, it is easy to see that (2.4) may be expressed as the fixed point iteration in $\mathcal{H} \times \mathcal{H}$ given by

$$\begin{pmatrix} x_{k+1} \\ u_{k+1} \end{pmatrix} = (M \circ T) \begin{pmatrix} x_k \\ u_k \end{pmatrix}.$$

From the perspective of fixed point theory, it is not clear what properties the operator $M \circ T$ possesses which can be used to deduce convergence. For instance, although M is *firmly nonexpansive*, the operator T need not be. A similar question regarding interpretations of the *golden ratio algorithm*, for which the operator M is of the same form, was posed in [27].

Stochastic and coordinate extensions. In large-scale problems, it is not always possible to evaluate the operator B owing to its high computational cost. Two possibilities for reducing the computational requirements are *stochastic approximations* of $B(x_k)$ and *block coordinate* variants of the algorithm. Both approaches work by employing low-cost approximation of $B(x_k)$ in each iteration. It would be interesting to consider stochastic and coordinate extensions of the method proposed here.

Acceleration schemes. As explained in section 1, the forward-reflected-backward method can be specialized to solve a minimization problem involving the sum of two convex functions, one of which is smooth. In 2007, in [31], Nesterov exploited his original idea from [32] to derive accelerated proximal gradient methods that enjoy

better complexity rates than the standard forward-backward method. It therefore seems reasonable that the forward-reflected-backward method could be adapted to incorporate a Nesterov-type acceleration.

REFERENCES

[1] F. Alvarez and H. Attouch, *An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping*, Set-Valued Anal., 9 (2001), pp. 3–11.

[2] A. Antipin, *On a method for convex programs using a symmetrical modification of the Lagrange function*, Ekon. Mat. Metody, 12 (1976), pp. 1164–1173.

[3] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*, Stanford Mathematical Studies in the Social Sciences, Stanford University Press, Stanford, CA, 1958.

[4] H. Attouch and A. Cabot, *Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions*, Appl. Math. Optim., (2019), pp. 1–52.

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

[6] J. Bello Cruz and R. Díaz Millán, *A variant of forward-backward splitting method for the sum of two monotone operators with a new search strategy*, Optimization, 64 (2015), pp. 1471–1486.

[7] R. I. Boţ and E. R. Csetnek, *An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems*, Numer. Algorithms, 71 (2016), pp. 519–540.

[8] R. I. Boţ, E. R. Csetnek, and A. Heinrich, *A primal-dual splitting algorithm for finding zeros of sums of maximal monotone operators*, SIAM J. Optim., 23 (2013), pp. 2011–2036.

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

[10] L. M. Briceño-Arias and D. Davis, *Forward-backward-half forward algorithm for solving monotone inclusions*, SIAM J. Optim., 28 (2018), pp. 2839–2871.

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

[12] G. H. Chen and R. T. Rockafellar, *Convergence rates in forward–backward splitting*, SIAM J. Optim., 7 (1997), pp. 421–444.

[13] P. L. Combettes, *Quasi-Fejérian analysis of some optimization algorithms*, in Inherently Parallel Algortihms in Feasibility and Optimization and Their Applications, Stud. Comput. Math. 8, Elsevier, Amsterdam, 2001, pp. 115–152.

[14] P. L. Combettes and J.-C. Pesquet, *Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping*, SIAM J. Optim., 25 (2015), pp. 1221–1248.

[15] E. R. Csetnek, Y. Malitsky, and M. K. Tam, *Shadow Douglas–Rachford splitting for monotone inclusions*, Appl. Math. Optim., 80 (2019), pp. 665–678.

[16] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, *Training GANs with optimism*, in Proceedings of ICLR, 2018.

[17] D. Davis and W. Yin, *A three-operator splitting scheme and its optimization applications*, Set-Valued Var. Anal., 25 (2017), pp. 829–858.

[18] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien, *A variational inequality perspective on generative adversarial networks*, in Proceedings of ICLR, 2019.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[20] E. Y. Hamedani and N. S. Aybat, *A Primal-Dual Algorithm for General Convex-Concave Saddle Point Problems*, arXiv:1803.01401, 2018.

[21] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, *On the convergence of single-call stochastic extra-gradient methods*, in Proceedings of NeurIPS, 2019, pp. 6936–6946.

[22] P. R. Johnstone and J. Eckstein, *Projective Splitting with Forward Steps: Asynchronous and Block-Iterative Operator Splitting*, arXiv:1803.07043, 2018.

[23] G. M. Korpelevich, *The extragradient method for finding saddle points and other problems*, Ekon. Mat. Metody, 12 (1976), pp. 747–756.

[24] P. L. Lions and B. Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964–979.

[25] D. Lorenz and T. Pock, *An inertial forward-backward algorithm for monotone inclusions*, J. Math. Imaging Vis., 51 (2015), pp. 311–325.
[26] Y. Malitsky, *Reflected projected gradient method for solving monotone variational inequalities*, SIAM J. Optim., 25 (2015), pp. 502–520.
[27] Y. Malitsky, *Golden ratio algorithms for variational inequalities*, Math. Program., (2019), https://doi.org/10.1007/s10107-019-01416-w.
[28] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras, *Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile*, in Proceedings of ICLR, 2019.
[29] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky, *Revisiting Stochastic Extragradient*, arXiv:1905.11373, 2019.
[30] A. Moudafi and M. Oliny, *Convergence of a splitting inertial proximal method for monotone operators*, J. Comput. Appl. Math., 155 (2003), pp. 447–454.
[31] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program., 140 (2013), pp. 125–161.
[32] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate* $O(1/k^2)$, Dokl. Akad. Nauk SSSR, 269 (1983), pp. 543–547.
[33] L. D. Popov, *A modification of the Arrow–Hurwicz method for finding saddle points*, Math. Notes, 28 (1980), pp. 845–848.
[34] J. Rieger and M. K. Tam, *Backward-forward-reflected-backward splitting for three operator monotone inclusions*, Appl. Math. Comput., 381 (2020), 125248, https://doi.org/10.1016/j.amc.2020.125248.
[35] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing Methods in Statistics, Academic Press, New York, 1971, pp. 233–257.
[36] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[37] E. K. Ryu and B. C. Vu, *Finding the forward-Douglas–Rachford-forward method*, J. Optim. Theory. Appl., 184 (2000), pp. 858–876.
[38] E. K. Ryu, K. Yuan, and W. Yin, *ODE Analysis of Stochastic Gradient Methods with Optimism and Anchoring for Minimax Problems and GANs*, arXiv:1905.10899, 2019.
[39] B. F. Svaiter, *On weak convergence of the Douglas–Rachford method*, SIAM J. Control Optim., 49 (2011), pp. 280–287.
[40] P. Tseng, *A modified forward-backward splitting method for maximal monotone mappings*, SIAM J. Control Optim., 38 (2000), pp. 431–446.