Journal of Machine Learning Research 24 (2023) 1-37 Submitted 3/20; Revised 12/22; Published 1/23

A Relaxed Inertial Forward-Backward-Forward Algorithm for Solving Monotone Inclusions with Application to GANs

Radu I. Boţ RADU.BOT@UNIVIE.AC.AT
Faculty of Mathematics
University of Vienna
Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

Michael Sedlmayer MICHAEL.SEDLMAYER@UNIVIE.AC.AT
Research Network Data Science @ Uni Vienna
University of Vienna
Kolingasse 14-16, 1090 Vienna, Austria

Phan Tu Vuong T.V.PHAN@SOTON.AC.UK
Mathematical Sciences
University of Southampton
Southampton SO17 1BJ, United Kingdom

Editor: Suvrit Sra

Abstract

We introduce a relaxed inertial forward-backward-forward (RIFBF) splitting algorithm for approaching the set of zeros of the sum of a maximally monotone operator and a single-valued monotone and Lipschitz continuous operator. This work aims to extend Tseng's forward-backward-forward method by both using inertial effects as well as relaxation parameters. We formulate first a second order dynamical system that approaches the solution set of the monotone inclusion problem to be solved and provide an asymptotic analysis for its trajectories. We provide for RIFBF, which follows by explicit time discretization, a convergence analysis in the general monotone case as well as when applied to the solving of pseudo-monotone variational inequalities. We illustrate the proposed method by applications to a bilinear saddle point problem, in the context of which we also emphasize the interplay between the inertial and the relaxation parameters, and to the training of Generative Adversarial Networks (GANs).

Keywords: forward-backward-forward algorithm, inertial effects, relaxation parameters, continuous time approach, application to GANs;

1. Introduction

1.1 Motivation

The main motivation for the investigation of monotone inclusions and variational inequalities governed by monotone and Lipschitz continuous operators is represented by convex-concave minimax problems. It is well-known that determining primal-dual pairs of optimal solutions of convex optimization problems means actually solving convex-concave minimax problems (Bauschke and Combettes, 2017), nevertheless, minimax problems are of own interest. Minimax problems arise traditionally in game theory and, more recently, in the

training of Generative Adversarial Networks (GANs), as we will see in Section 4. The general convex-concave minimax, or saddle point, problem is of the form

$$\min_{u\in H}\max_{v\in G}\Psi(u,v) := f(u) + \Phi(u,v) - g(v), \tag{1}$$

where H and G are real Hilbert spaces, $\Phi : H \times G \to \mathbb{R}$ is a coupling function, and $f : H \to \mathbb{R} \cup \{+\infty\}$ and $g : G \to \mathbb{R} \cup \{+\infty\}$ are regularizers. The coupling function Φ is assumed to be convex-concave, i.e., $\Phi(\cdot, v)$ is convex for all $v \in G$ and $\Phi(u, \cdot)$ is concave for all $u \in H$, and differentiable with L-Lipschitz continuous gradient ($L > 0$), which means that for all $u, u' \in H$ and all $v, v' \in G$ we have

$$\|\nabla\Phi(u,v) - \nabla\Phi(u',v')\| \leq L\|(u,v) - (u',v')\| = L\sqrt{\|u-u'\|^2 + \|v-v'\|^2}.$$

The regularizers f and g are assumed to be proper, convex and lower semicontinuous. A solution of (1) is given by a so-called saddle point (u^*, v^*), fulfilling for all $u \in H$ and all $v \in G$

$$\Psi(u^*, v) \leq \Psi(u^*, v^*) \leq \Psi(u, v^*)$$

or, equivalently, the system of optimality conditions

$$\begin{aligned} 0 &\in \partial\Psi(\cdot, v^*)(u^*) = \partial f(u^*) + \nabla_u\Phi(u^*, v^*), \\ 0 &\in \partial(-\Psi(u^*, \cdot))(v^*) = \partial g(v^*) - \nabla_v\Phi(u^*, v^*), \end{aligned}$$

where $\partial f : H \to 2^H$ and $\partial g : G \to 2^G$ denote the convex subdifferential of the functions f and g, respectively. This system can be rewritten as the following monotone inclusion

$$0 \in A(u^*, v^*) + B(u^*, v^*), \tag{2}$$

where $A = \partial f \times \partial g : H \times G \to 2^{H\times G}$ is a maximally monotone operator and $B : H \times G \to H \times G$ with $B(u, v) = (\nabla_u\Phi(u, v), -\nabla_v\Phi(u, v))$ is monotone and Lipschitz continuous with constant $L > 0$.

Monotone inclusions governed by maximal monotone operators are oftenly seen as generalizations of systems of optimality conditions for convex optimization problems, however, (2) shows that they can also be of own interest. In order to solve (2) one cannot rely on solution methods developed for optimization problems, but needs to design and develop specific algorithms in the framework of the theory of monotone operators. This will be one of our goals in this paper and in this way we will provide further evidence for the importance of monotone operators beyond the convex optimization setting.

1.2 Problem Formulation

Let H be a real Hilbert space endowed with inner product $\langle\cdot,\cdot\rangle$ and corresponding norm $\|\cdot\|$, $A : H \to 2^H$ a maximally monotone set-valued operator, and $B : H \to H$ a monotone and Lipschitz continuous operator with Lipschitz constant $L > 0$, such that $\text{Zeros}(A+B) := \{x \in H : 0 \in Ax + Bx\} \neq \emptyset$. We are interested in solving the following inclusion problem:

Find $x^* \in H$ such that

$$0 \in Ax^* + Bx^*. \tag{3}$$

We recall that the operator $A : H \to 2^H$ is called monotone if

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

where $\text{Graph}(A) = \{(x,u) \in H \times H : u \in Ax\}$ denotes its graph. The operator A is called maximally monotone if it is monotone and its graph is not properly included in the graph of another monotone operator. The single-valued operator $B : H \to H$ is said to be Lipschitz continuous with Liphschitz constant $L > 0$ if

$$\|Bx - By\| \leq L\|x - y\| \quad \forall x, y \in H.$$

An important special case is when $A = N_C$, the normal cone of a nonempty closed convex subset C of H. Then (3) reduces to a variational inequality (VI):

Find $x^* \in C$ such that

$$\langle Bx^*, x - x^*\rangle \geq 0 \quad \forall x \in C. \tag{4}$$

1.3 Related Literature

Solution methods for solving (3), when the operator B is cocoercive, have been developed intensively in the last decades (Abbas and Attouch, 2015; Attouch and Cabot, 2019a; Bauschke and Combettes, 2017; Boţ and Csetnek, 2016a, 2018). Recall that the operator $B : H \to H$ is cocoercive if there is a constant $L > 0$ such that

$$L\,\langle Bx - By, x - y\rangle \geq \|Bx - By\|^2 \quad \forall x, y \in H.$$

Notice that every cocoercive operator is Lipschitz continuous and that the gradient of a convex and Fréchet differentiable function is a cocoercive operator if and only if it is Lipschitz continuous (Bauschke and Combettes, 2017).

The simplest method for solving (3), when B is cocoercive, is the forward-backward (FB) method, which generates an iterative sequence $(x_k)_{k\geq 0}$ via

$$(\forall k \geq 0) \quad x_{k+1} := J_{\lambda A}(I - \lambda B)x_k, \tag{5}$$

where $x_0 \in H$ is the starting point and $J_{\lambda A} := (I + \lambda A)^{-1} : H \to H$ is the resolvent of the operator A. Here, I denotes the identity operator of H and λ is a positive stepsize chosen in $(0, \frac{2}{L})$. The resolvent of a maximally monotone operator is a single-valued and cocoercive operator with constant $L = 1$. The iterative scheme (5) results by time-discretizing with time stepsize equal 1 the first order dynamical system

$$\begin{cases} \dot{x}(t) + x(t) = J_{\lambda A}\left(I - \lambda B\right) x(t), \\ x(0) = x_0. \end{cases}$$

For more about dynamical systems of implicit type associated to monotone inclusions and convex optimization problems, see for example the work by (Abbas and Attouch, 2015), (Antipin, 1994), (Bolte, 2003), and (Boţ and Csetnek, 2018).

Recently, the following second order dynamical system associated with the monotone inclusion problem (3), when B is cocoercive,

$$\begin{cases} \ddot{x}(t) + \gamma(t)\dot{x}(t) + \tau(t)\left[x(t) - J_{\lambda A}(I - \lambda B)x(t)\right] = 0, \\ x(0) = x_0, \quad \dot{x}(0) = v_0, \end{cases}$$

where $\gamma, \tau : [0, +\infty) \to [0, +\infty)$, was proposed and studied by (Boţ and Csetnek, 2016a) (see also the work of Alvarez, 2000; Antipin, 1994; Boţ and Csetnek, 2018). Explicit time discrerization of this second order dynamical system gives rise to so-called relaxed inertial forward-backward algorithms, which combine inertial effects and relaxation parameters.

In the last years, Attouch and Cabot have promoted in a series of papers relaxed inertial algorithms for monotone inclusions and convex optimization problems, as they combine the advantages of both inertial effects and relaxation techniques. More precisely, they addressed the relaxed inertial proximal method (RIPA) (Attouch and Cabot, 2019b,c) and the relaxed inertial forward-backward method (RIFB) (Attouch and Cabot, 2019a). A relaxed inertial Douglas-Rachford algorithm for monotone inclusions has been proposed by (Boţ et al., 2015). (Iutzeler and Hendrickx, 2019) investigated the influence inertial effects and relaxation techniques have on the numerical performances of optimization algorithms. The interplay between relaxation and inertial parameters for relative-error inexact under-relaxed algorithms has been addressed by (Alves and Marcavillaca, 2020; Alves et al., 2020).

Relaxation techniques are essential ingredients in the formulation of algorithms for monotone inclusions, as they provide more flexibility to the iterative schemes (Bauschke and Combettes, 2017; Eckstein and Bertsekas, 1992). Inertial effects have been introduced in order to accelerate the convergence of the numerical methods. This technique traces back to the pioneering work of (Polyak, 1964), who introduced the heavy ball method in order to speed up the convergence behavior of the gradient algorithm and allow the detection of different critical points. This idea was employed and refined by (Nesterov, 1983) and by (Alvarez, 2000) and by (Alvarez and Attouch, 2001) in the context of solving smooth convex minimization problems and monotone inclusions/nonsmooth convex minimization problems, respectively. When applied to convex minimization problems, the acceleration techniques introduced by Nesterov in (Nesterov, 1983) or those defined via asymptotically similar constructions lead to an improved convergence rate for the sequence of objective function values. When applied to monotone inclusions, the same lead, as seen in (Attouch and Cabot, 2019a,b,c) for the relaxed inertial proximal and forward-backward methods, to improved convergence rates for the sequences of discrete velocities and Yosida regularizations of the governing operator.

In this paper we will focus on solving the monotone inclusion (3) in the case when B is merely monotone and Lipschitz continuous. To this end we will formulate a relaxed inertial forward-backward-forward (RIFBF) algorithm, which we obtain through the time discretization of a second order dynamical system approaching the solution set of (3). The forward-backward-forward (FBF) method was proposed by (Tseng, 2000) and it generates an iterative sequence $(x_k)_{k\geq 0}$ via

$$(\forall k \geq 0) \quad \begin{cases} y_k = J_{\lambda_k A}(I - \lambda_k B)x_k, \\ x_{k+1} = y_k - \lambda_k(By_k - Bx_k), \end{cases}$$

where $x_0 \in H$ is the starting point. The sequence $(x_k)_{k\geq 0}$ converges weakly to a solution of (3) if the sequence of stepsizes $(\lambda_k)_{k\geq 0}$ is chosen in the interval $\left(0, \frac{1}{L}\right)$, where $L > 0$ is the Lipschitz constant of B. An inertial version of FBF, however in the absence of relaxation variables, was proposed by (Boţ and Csetnek, 2016b). The convergence of the sequence of generated iterates has been proved in a very restricting setting that requires that the inertial

parameters take very small values. This is in accordance to the inertial parameters choices in the earliest papers on inertial algorithms (Alvarez, 2000) and (Alvarez and Attouch, 2001). It is one of our aims to show that relaxation parameters allow more flexibility in the choice of the inertial ones, which leads to better convergence results.

Recently, a forward-backward algorithm for solving (3), when B is monotone and Lipschitz continuous, was proposed by (Malitsky and Tam, 2020). This method requires in every iteration only one forward step instead of two, however, the sequence of stepsizes has to be chosen constant in the interval $\left(0, \frac{1}{2L}\right)$, which slows the algorithm down in comparison to FBF. A popular algorithm used to solve the variational inequality (4), when B is monotone and Lipschitz continuous, is Korpelevich's extragradient method (Korpelevich, 1976). The stepsizes are to be chosen in the interval $\left(0, \frac{1}{L}\right)$, however, this method requires two projection steps on C and two forward steps.

1.4 Outline

First, we will approach the solution set of (3) from a continuous perspective by means of the trajectories generated by a second order dynamical system of FBF type. We will prove an existence and uniqueness result for the generated trajectories and provide a general setting in which these converge to a zero of $A + B$ as time goes to infinity. In addition, we will show that explicit time discretization of the dynamical system gives rise to an algorithm of forward-backward-forward type with inertial and relaxation parameters (RIFBF).

In Section 3 we will discuss the convergence of RIFBF and investigate the interplay between the inertial and the relaxation parameters. It is of certain relevance to notice that both the standard FBF method, the algorithm by (Malitsky and Tam, 2020) and the extragradient method require to know the Lipschitz constant of B, which is not always available. This can be avoided by performing a line-search procedure, which usually leads to additional computation costs. On the contrary, we will use an adaptive stepsize rule which does not require knowledge of the Lipschitz constant of B. We will also comment on the convergence of RIFBF when applied to the solving of the variational inequality (4) in the case when the operator B is pseudo-monotone but not necessarily monotone. Pseudo-monotone operators appear in the consumer theory of mathematical economics (Hadjisavvas et al., 2012) and as gradients of pseudo-convex functions (Cottle and Ferland, 1971), such as ratios of convex and concave functions in fractional programming (Borwein and Lewis, 2006).

Concluding, we treat two different numerical experiments supporting our theoretical results in Section 4. On the one hand we deal with a bilinear saddle point problem which can be understood as a two-player zero-sum constrained game. In this context, we emphasize the interplay between the inertial and the relaxation parameters. On the other hand we employ variants of RIFBF for training Generative Adversarial Networks (GANs), which is a class of machine learning systems where two opposing artificial neural networks compete in a zero-sum game. GANs have achieved outstanding results for producing photorealistic pictures and are typically known to be difficult to optimize. We show that our method outperforms "Extra Adam", a GAN training approach inspired by the extra-gradient algorithm, which recently achieved state-of-the-art results (Gidel et al., 2019).

2. A Second Order Dynamical System of FBF Type

In this section we will focus on the study of the dynamical system

$$\begin{cases} y(t) = J_{\lambda A}(I - \lambda B)x(t), \\ \ddot{x}(t) + \gamma(t)\dot{x}(t) + \tau(t)\left[x(t) - y(t) - \lambda\left(Bx(t) - By(t)\right)\right] = 0, \\ x(0) = x_0, \quad \dot{x}(0) = v_0, \end{cases} \tag{6}$$

where $\gamma, \tau : [0, +\infty) \to [0, +\infty)$ are Lebesgue measurable functions, $0 < \lambda < \frac{1}{L}$ and $x_0, v_0 \in H$, in connection with the monotone inclusion problem (3).

We define $M : H \to H$ by

$$Mx = x - J_{\lambda A}(I - \lambda B)x - \lambda\left[Bx - B \circ J_{\lambda A}(I - \lambda B)x\right]. \tag{7}$$

Then (6) can be equivalently written as

$$\begin{cases} \ddot{x}(t) + \gamma(t)\dot{x}(t) + \tau(t)Mx(t) = 0, \\ x(0) = x_0, \quad \dot{x}(0) = v_0. \end{cases} \tag{8}$$

The following result collects some properties of M.

Proposition 1 *Let M be defined as in (7). Then the following statements are true:*

(i) $\mathrm{Zeros}(M) = \mathrm{Zeros}(A + B)$*;*

(ii) M is Lipschitz continuous;

(iii) for all $x^* \in \mathrm{Zeros}(M)$ and all $x \in H$ it holds

$$\langle Mx, x - x^* \rangle \geq \frac{1 - \lambda L}{(1 + \lambda L)^2}\|Mx\|^2. \tag{9}$$

Proof (i) For $x \in H$ we set $y := J_{\lambda A}(I - \lambda B)x$, thus $Mx = x - y - \lambda\left(Bx - By\right)$. Using the Lipschitz continuity of B we have

$$(1 - \lambda L)\|x - y\| \leq \|Mx\| = \|x - y - \lambda\left(Bx - By\right)\| \leq (1 + \lambda L)\|x - y\|. \tag{10}$$

Therefore, $x \in \mathrm{Zeros}(M)$ if and only if $x = y = J_{\lambda A}(I - \lambda B)x$, which is further equivalent to $x \in \mathrm{Zeros}(A + B)$.

(ii) Let $x, x' \in H$ and $y := J_{\lambda A}(I - \lambda B)x$, and $y' := J_{\lambda A}(I - \lambda B)x'$. The Lipschitz continuity of B yields

$$\begin{aligned} \|Mx - Mx'\| &= \|x - y - \lambda\left(Bx - By\right) - x' + y' + \lambda\left(Bx' - By'\right)\| \\ &\leq (1 + \lambda L)\left(\|x - x'\| + \|y - y'\|\right). \end{aligned}$$

In addition, by the nonexpansiveness (Lipschitz continuity with Lipschitz constant 1) of $J_{\lambda A}$ and again by the Lipschitz continuity of B we obtain

$$\begin{aligned} \|y - y'\| &= \|J_{\lambda A}(I - \lambda B)x - J_{\lambda A}(I - \lambda B)x'\| \\ &\leq \|(I - \lambda B)x - (I - \lambda B)x'\| \leq (1 + \lambda L)\|x - x'\|. \end{aligned}$$

Therefore,

$$\|Mx - Mx'\| \leq (1 + \lambda L)(2 + \lambda L)\|x - x'\|,$$

which shows that M is Lipschitz continuous with Lipschitz constant $(1 + \lambda L)(2 + \lambda L) > 0$.

(iii) Let $x^* \in H$ be such that $0 \in (A + B)x^*$ and $x \in H$. We denote $y := J_{\lambda A}(I - \lambda B)x$ and can write $(I - \lambda B)x \in (I + \lambda A)y$ or, equivalently,

$$\frac{1}{\lambda}(x - y) - (Bx - By) \in (A + B)y. \tag{11}$$

Using the monotonicity of $A + B$ we obtain

$$\left\langle \frac{1}{\lambda}(x - y) - (Bx - By), y - x^* \right\rangle \geq 0,$$

which is equivalent to

$$\langle x - y - \lambda(Bx - By), x - x^* \rangle \geq \langle x - y - \lambda(Bx - By), x - y \rangle.$$

This means that

$$\begin{aligned} \langle Mx, x - x^* \rangle &\geq \langle x - y - \lambda(Bx - By), x - y \rangle = \|x - y\|^2 - \lambda \langle Bx - By, x - y \rangle \\ &\geq \|x - y\|^2 - \lambda \|Bx - By\| \|x - y\| \geq (1 - \lambda L)\|x - y\|^2 \\ &\geq \frac{1 - \lambda L}{(1 + \lambda L)^2} \|Mx\|^2, \end{aligned}$$

where the last inequality follows from (10). ■

The following definition makes explicit which kind of solutions of the dynamical system (6) we are looking for. We recall that a function $x : [0, b] \to H$ (where $b > 0$) is said to be absolutely continuous if there exists an integrable function $y : [0, b] \to H$ such that

$$x(t) = x(0) + \int_0^t y(s)ds \quad \forall t \in [0, b].$$

This is nothing else than x is continuous and its distributional derivative $\dot{x}$ is Lebesgue integrable on $[0, b]$.

Definition 2 *We say that $x : [0, +\infty) \to H$ is a strong global solution of (6) if the following properties are satisfied:*

(i) x, $\dot{x} : [0, +\infty) \to H$ are locally absolutely continuous, in other words, absolutely continuous on each interval $[0, b]$ for $0 < b < +\infty$;

(ii) $\ddot{x}(t) + \gamma(t)\dot{x}(t) + \tau(t)Mx(t) = 0$ for almost every $t \in [0, +\infty)$;

(iii) $x(0) = x_0$ and $\dot{x}(0) = v_0$.

Since $M : H \to H$ is Lipschitz continuous, the existence and uniqueness of the trajectory of (6) follows from the Cauchy-Lipschitz Theorem for absolutely continuous trajectories.

Theorem 3 (see Boţ and Csetnek, 2016a, Theorem 4) *Let $\gamma, \tau : [0, +\infty) \to [0, +\infty)$ be Lebesgue measurable functions such that $\gamma, \tau \in L^1_{loc}([0, +\infty))$ (that is $\gamma, \tau \in L^1([0, b])$ for all $0 < b < +\infty$). Then for each $x_0, v_0 \in H$ there exists a unique strong global solution of the dynamical system* (6).

We will prove the convergence of the trajectories of (6) in a setting which requires the damping function γ and the relaxation function τ to fulfil the assumptions below. We refer to (Boţ and Csetnek, 2016a) for examples of functions which fulfil this assumption and want also to emphasize that when the two functions are constant we recover the conditions from (Attouch and Maingé, 2011).

Assumption 1 *$\gamma, \tau : [0, +\infty) \to [0, +\infty)$ are locally absolutely continuous and there exists $\nu > 0$ such that for almost every $t \in [0, +\infty)$ it holds*

$$\dot{\gamma}(t) \leq 0 \leq \dot{\tau}(t) \quad \text{and} \quad \frac{\gamma^2(t)}{\tau(t)} \geq (1+\nu)\frac{(1+\lambda L)^2}{1-\lambda L}.$$

The result which states the convergence of the trajectories is adapted from the results by (Boţ and Csetnek, 2016a, Theorem 8). Though, it cannot be obtained as a direct consequence of it, since the operator M is not cocoercive as it is required there. However, as seen in Proposition 1 (iii), M has a property, sometimes called "cocoercive with respect to its set of zeros", which is by far weaker than cocoercivity, but strong enough in order to allow us to partially apply the techniques used to prove Theorem 8 by (Boţ and Csetnek, 2016a).

Theorem 4 *Let $\gamma, \tau : [0, +\infty) \to [0, +\infty)$ be functions satisfying Assumption 1 and $x_0, v_0 \in H$. Let $x : [0, +\infty) \to H$ be the unique strong global solution of* (6). *Then the following statements are true:*

(i) the trajectory x is bounded and $\dot{x}, \ddot{x}, Mx \in L^2([0, +\infty); H)$;

(ii) $\lim_{t\to+\infty} \dot{x}(t) = \lim_{t\to+\infty} \ddot{x}(t) = \lim_{t\to+\infty} Mx(t) = \lim_{t\to+\infty} [x(t) - y(t)] = 0$;

(iii) $x(t)$ converges weakly to an element in $\mathrm{Zeros}(A + B)$ as $t \to +\infty$.

Proof Take an arbitrary $x^* \in \mathrm{Zeros}(A+B) = \mathrm{Zeros}(M)$ and define for all $t \in [0, +\infty)$ the Lyapunov function $h(t) = \frac{1}{2}\|x(t) - x^*\|^2$. For almost every $t \in [0, +\infty)$ we have

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

Taking into account (8) we obtain for almost every $t \in [0, +\infty)$ that

$$\ddot{h}(t) + \gamma(t)\dot{h}(t) + \tau(t)\langle x(t) - x^*, Mx(t)\rangle = \|\dot{x}(t)\|^2,$$

which, together with Proposition 1 (iii), implies

$$\ddot{h}(t) + \gamma(t)\dot{h}(t) + \frac{1-\lambda L}{(1+\lambda L)^2}\tau(t)\|Mx(t)\|^2 \leq \|\dot{x}(t)\|^2.$$

From this point, we can proceed as in the proof of (Boţ and Csetnek, 2016a, Theorem 8) – for a complete explanation see Appendix A. Consequently, we obtain the statements in (i) and (ii) and the fact that the limit $\lim_{t\to+\infty} \|x(t) - x^*\| \in \mathbb{R}$ exists, which is the first assumption in the continuous version of the Opial Lemma as it was used, for example, by (Boţ and Csetnek, 2016a, Lemma 7). In order to show that the second assumption of the Opial Lemma is fulfilled, which means actually proving that every weak sequential cluster point of the trajectory x is a zero of M, one cannot use the arguments from (Boţ and Csetnek, 2016a, Theorem 8), since M is not maximal monotone. We have to use, instead, different arguments relying on the maximal monotonicity of $A + B$.

Indeed, let $\bar{x}$ be a weak sequential cluster point of x, which means that there exists a sequence $t_k \to +\infty$ such that $(x(t_k))_{k\geq 0}$ converges weakly to $\bar{x}$ as $k \to +\infty$. Since, according to (ii), $\lim_{t\to+\infty} Mx(t) = \lim_{t\to+\infty} [x(t) - y(t)] = 0$, we conclude that $(y(t_k))_{k\geq 0}$ also converges weakly to $\bar{x}$. According to (11) we have

$$\frac{1}{\lambda}(x(t_k) - y(t_k)) - (Bx(t_k) - By(t_k)) \in (A + B)y(t_k) \quad \forall k \geq 0. \tag{12}$$

Since B is Lipschitz continuous and $\lim_{k\to+\infty} \|x(t_k) - y(t_k)\| = 0$, the left hand side of (12) converges strongly to 0 as $k \to +\infty$. Since $A + B$ is maximal monotone, its graph is sequentially closed with respect to the weak-strong topology of the product space $H \times H$. Therefore, taking the limit as $t_k \to +\infty$ in (12) we obtain $\bar{x} \in \text{Zeros}(A + B)$.

Thus, the continuous Opial Lemma implies that $x(t)$ converges weakly to an element in $\text{Zeros}(A + B)$ as $t \to +\infty$. ■

Remark 5 (explicit discretization) *Explicit time discretization of* (6) *with stepsize* $s_k > 0$*, relaxation variable* $\tau_k > 0$*, damping variable* $\gamma_k > 0$*, and initial points* x_0 *and* x_1 *yields for all* $k \geq 1$ *the following iterative scheme:*

$$\frac{1}{s_k^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma_k}{s_k}(x_k - x_{k-1}) + \tau_k M z_k = 0, \tag{13}$$

where z_k *is an extrapolation of* x_k *and* x_{k-1} *that will be chosen later. The Lipschitz continuity of* M *provides a certain flexibility to this choice. We can write* (13) *equivalently as*

$$(\forall k \geq 1) \quad x_{k+1} = x_k + (1 - \gamma_k s_k)(x_k - x_{k-1}) - s_k^2 \tau_k M z_k.$$

Setting $\alpha_k := 1 - \gamma_k s_k$*,* $\rho_k := s_k^2 \tau_k$ *and choosing* $z_k := x_k + \alpha_k(x_k - x_{k-1})$ *for all* $k \geq 1$*, we can write the above scheme as*

$$(\forall k \geq 1) \begin{cases} z_k = x_k + \alpha_k(x_k - x_{k-1}) \\ y_k = J_{\lambda A}(I - \lambda B)z_k \\ x_{k+1} = (1 - \rho_k)z_k + \rho_k \left(y_k - \lambda(By_k - Bz_k)\right), \end{cases}$$

which is a relaxed version of the FBF algorithm with inertial effects.

3. A Relaxed Inertial FBF Algorithm

In this section we investigate the convergence of the relaxed inertial algorithm derived in the previous section via the time discretization of (6), however, we also assume that the stepsizes $(\lambda_k)_{k\geq 1}$ are variable. More precisely, we will address the following algorithm

$$(RIFBF) \qquad (\forall k \geq 1) \begin{cases} z_k = x_k + \alpha_k(x_k - x_{k-1}) \\ y_k = J_{\lambda_k A}(I - \lambda_k B)z_k \\ x_{k+1} = (1-\rho_k)z_k + \rho_k \left(y_k - \lambda_k(By_k - Bz_k)\right), \end{cases}$$

where $x_0, x_1 \in H$ are starting points, $(\lambda_k)_{k\geq 1}$ and $(\rho_k)_{k\geq 1}$ are sequences of positive numbers, and $(\alpha_k)_{k\geq 1}$ is a sequence of nonnegative numbers. The following iterative schemes can be obtained as particular instances of RIFBF:

- $\rho_k = 1$ for all $k \geq 1$: inertial Forward-Backward-Forward algorithm

$$(IFBF) \qquad (\forall k \geq 1) \begin{cases} z_k = x_k + \alpha_k(x_k - x_{k-1}) \\ y_k = J_{\lambda_k A}(I - \lambda_k B)z_k \\ x_{k+1} = y_k - \lambda_k(By_k - Bz_k) \end{cases}$$

- $\alpha_k = 0$ for all $k \geq 1$: relaxed Forward-Backward-Forward algorithm

$$(RFBF) \qquad (\forall k \geq 1) \begin{cases} y_k = J_{\lambda_k A}(I - \lambda_k B)x_k \\ x_{k+1} = (1-\rho_k)x_k + \rho_k \left(y_k - \lambda_k(By_k - Bx_k)\right) \end{cases}$$

- $\alpha_k = 0$, $\rho_k = 1$ for all $k \geq 1$: Forward-Backward-Forward algorithm

$$(FBF) \qquad (\forall k \geq 1) \begin{cases} y_k = J_{\lambda_k A}(I - \lambda_k B)x_k \\ x_{k+1} = y_k - \lambda_k(By_k - Bx_k). \end{cases}$$

Stepsize rules: Depending on the availability of the Lipschitz constant L of B , we have two different options for the choice of the sequence of stepsizes $(\lambda_k)_{k\geq 1}$:

- **constant stepsize**: $\lambda_k := \lambda \in \left(0, \frac{1}{L}\right)$ for all $k \geq 1$;
- **adaptive stepsize**: let $\mu \in (0,1)$ and $\lambda_1 > 0$. The stepsizes for $k \geq 1$ are adaptively updated as follows

$$\lambda_{k+1} := \begin{cases} \min\left\{\lambda_k, \frac{\mu\|y_k - z_k\|}{\|By_k - Bz_k\|}\right\}, & \text{if } By_k - Bz_k \neq 0, \\ \lambda_k, & \text{otherwise.} \end{cases} \tag{14}$$

If the Lipschitz constant L of B is known in advance, then a constant stepsize can be chosen. Otherwise, the adaptive stepsize rule (14) is highly recommended. In the following, we provide the convergence analysis for the adaptive stepsize rule, as the constant stepsize rule can be obtained as a particular of it by setting $\lambda_1 := \lambda$ and $\mu := \lambda L$.

Proposition 6 *Let $\mu \in (0,1)$ and $\lambda_1 > 0$. The sequence $(\lambda_k)_{k\geq 1}$ generated by (14) is nonincreasing and*

$$\lim_{k\to+\infty} \lambda_k = \lambda \geq \min\left\{\lambda_1, \frac{\mu}{L}\right\}.$$

In addition,

$$\|By_k - Bz_k\| \leq \frac{\mu}{\lambda_{k+1}}\|y_k - z_k\| \quad \forall k \geq 1. \tag{15}$$

Proof It is obvious from (14) that $\lambda_{k+1} \leq \lambda_k$ for all $k \geq 1$. Since B is Lipschitz continuous with Lipschitz constant L, it yields

$$\frac{\mu\|y_k - z_k\|}{\|By_k - Bz_k\|} \geq \frac{\mu}{L}, \text{ if } By_k - Bz_k \neq 0,$$

which together with (14) yields

$$\lambda_{k+1} \geq \min\left\{\lambda_1, \frac{\mu}{L}\right\} \quad \forall k \geq 1.$$

Thus, there exists

$$\lambda := \lim_{k\to+\infty} \lambda_k \geq \min\left\{\lambda_1, \frac{\mu}{L}\right\}.$$

The inequality (15) follows directly from (14). ■

Proposition 7 *Let $w_k := y_k - \lambda_k(By_k - Bz_k)$ and $\theta_k := \frac{\lambda_k}{\lambda_{k+1}}$ for all $k \geq 1$. Then for all $x^* \in \text{Zeros}(A+B)$ it holds*

$$\|w_k - x^*\|^2 \leq \|z_k - x^*\|^2 - \left(1 - \mu^2\theta_k^2\right)\|y_k - z_k\|^2 \quad \forall k \geq 1. \tag{16}$$

Proof Let $k \geq 1$. Using (15) we have that

$$\begin{aligned}
\|z_k - x^*\|^2 &= \|z_k - y_k + y_k - w_k + w_k - x^*\|^2 \\
&= \|z_k - y_k\|^2 + \|y_k - w_k\|^2 + \|w_k - x^*\|^2 \\
&\quad +2\langle z_k - y_k, y_k - x^*\rangle + 2\langle y_k - w_k, w_k - x^*\rangle \\
&= \|z_k - y_k\|^2 - \|y_k - w_k\|^2 + \|w_k - x^*\|^2 + 2\langle z_k - w_k, y_k - x^*\rangle \\
&= \|z_k - y_k\|^2 - \lambda_k^2\|By_k - Bz_k\|^2 + \|w_k - x^*\|^2 + 2\langle z_k - w_k, y_k - x^*\rangle \\
&\geq \|z_k - y_k\|^2 - \frac{\mu^2\lambda_k^2}{\lambda_{k+1}^2}\|y_k - z_k\|^2 + \|w_k - x^*\|^2 + 2\langle z_k - w_k, y_k - x^*\rangle. \qquad (17)
\end{aligned}$$

On the other hand, since

$$(I + \lambda_k A)y_k \ni (I - \lambda_k B)z_k,$$

we have

$$z_k \in y_k + \lambda_k Ay_k + \lambda_k Bz_k = y_k - \lambda_k(By_k - Bz_k) + \lambda_k(A+B)y_k = w_k + \lambda_k(A+B)y_k.$$

Therefore,

$$\frac{1}{\lambda_k}(z_k - w_k) \in (A+B)y_k,$$

which, together with $0 \in (A+B)x^*$ and the monotonicity of $A+B$, implies

$$\langle z_k - w_k, y_k - x^* \rangle \geq 0.$$

Hence,

$$\|z_k - x^*\|^2 \geq \|z_k - y_k\|^2 - \frac{\mu^2\lambda_k^2}{\lambda_{k+1}^2}\|y_k - z_k\|^2 + \|w_k - x^*\|^2,$$

which is (16). ■

The following result introduces a discrete Lyapunov function for which a decreasing property is established.

Proposition 8 *Let $(\alpha_k)_{k\geq 1}$ be a nondecreasing sequence of nonnegative numbers and $(\rho_k)_{k\geq 1}$ be such that for some $k_\rho \geq 1$ we have*

$$0 < \rho_k \leq \frac{2}{1+\mu\theta_k} \quad \forall k \geq k_\rho, \tag{18}$$

let $x^* \in \operatorname{Zeros}(A+B)$ and for all $k \geq 1$ define

$$H_k := \|x_k - x^*\|^2 - \alpha_k\|x_{k-1} - x^*\|^2 + 2\alpha_k\left(\alpha_k + \frac{1-\alpha_k}{\rho_k(1+\mu\theta_k)}\right)\|x_k - x_{k-1}\|^2. \tag{19}$$

Then there exists $k_0 \geq k_\rho$ such that for all $k \geq k_0$ it holds

$$H_{k+1} - H_k \leq -\delta_k\|x_{k+1} - x_k\|^2, \tag{20}$$

where

$$\delta_k := (1-\alpha_k)\left(\frac{2}{\rho_k(1+\mu\theta_k)} - 1\right) - 2\alpha_{k+1}\left(\alpha_{k+1} + \frac{1-\alpha_{k+1}}{\rho_{k+1}(1+\mu\theta_{k+1})}\right) \quad \forall k \geq 1.$$

Proof From Proposition 7, with $w_k = y_k - \lambda_k(By_k - Bz_k)$, we have for all $k \geq 1$

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|(1-\rho_k)z_k + \rho_k w_k - x^*\|^2 \\
&= \|(1-\rho_k)(z_k - x^*) + \rho_k(w_k - x^*)\|^2 \\
&= (1-\rho_k)\|z_k - x^*\|^2 + \rho_k\|w_k - x^*\|^2 - \rho_k(1-\rho_k)\|w_k - z_k\|^2 \\
&\leq (1-\rho_k)\|z_k - x^*\|^2 + \rho_k\|z_k - x^*\|^2 \\
&\quad -\rho_k\left(1 - \mu^2\theta_k^2\right)\|y_k - z_k\|^2 - \frac{1-\rho_k}{\rho_k}\|x_{k+1} - z_k\|^2 \\
&= \|z_k - x^*\|^2 - \rho_k\left(1 - \mu^2\theta_k^2\right)\|y_k - z_k\|^2 - \frac{1-\rho_k}{\rho_k}\|x_{k+1} - z_k\|^2. \quad (21)
\end{aligned}$$

Using (15) we obtain for all $k \geq 1$

$$\begin{aligned}\frac{1}{\rho_k}\|x_{k+1}-z_k\| = \|w_k - z_k\| &\leq \|w_k - y_k\| + \|y_k - z_k\| = \lambda_k\|By_k - Bz_k\| + \|y_k - z_k\| \\ &\leq \left(1 + \frac{\mu\lambda_k}{\lambda_{k+1}}\right)\|y_k - z_k\| = (1+\mu\theta_k)\|y_k - z_k\|.\end{aligned}$$

Since $\lim_{k\to+\infty}(1-\mu\theta_k) = 1-\mu > 0$, there exists $k_1 \geq 1$ such that

$$1 - \mu\theta_k > 0 \quad \forall k \geq k_1.$$

This means that for all $k \geq k_1$, we have

$$\frac{1-\mu\theta_k}{\rho_k(1+\mu\theta_k)}\|x_{k+1}-z_k\|^2 \leq \rho_k\left(1-\mu^2\theta_k^2\right)\|y_k - z_k\|^2.$$

We obtain from (21) that for all $k \geq k_1$

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &\leq \|z_k - x^*\|^2 - \left(\frac{1-\mu\theta_k}{\rho_k(1+\mu\theta_k)} + \frac{1-\rho_k}{\rho_k}\right)\|x_{k+1}-z_k\|^2 \\ &= \|z_k - x^*\|^2 - \left(\frac{2}{\rho_k(1+\mu\theta_k)} - 1\right)\|x_{k+1}-z_k\|^2. \end{aligned} \tag{22}$$

Now we will estimate the right-hand side of (22). For all $k \geq 1$ we have

$$\begin{aligned}\|z_k - x^*\|^2 &= \|x_k + \alpha_k(x_k - x_{k-1}) - x^*\|^2 \\ &= \|(1+\alpha_k)(x_k - x^*) - \alpha_k(x_{k-1} - x^*)\|^2 \\ &= (1+\alpha_k)\|x_k - x^*\|^2 - \alpha_k\|x_{k-1} - x^*\|^2 + \alpha_k(1+\alpha_k)\|x_k - x_{k-1}\|^2, \end{aligned} \tag{23}$$

and

$$\begin{aligned}\|x_{k+1}-z_k\|^2 &= \|(x_{k+1}-x_k) - \alpha_k(x_k - x_{k-1})\|^2 \\ &= \|x_{k+1}-x_k\|^2 + \alpha_k^2\|x_k - x_{k-1}\|^2 - 2\alpha_k\langle x_{k+1}-x_k, x_k - x_{k-1}\rangle \\ &\geq (1-\alpha_k)\|x_{k+1}-x_k\|^2 + \left(\alpha_k^2 - \alpha_k\right)\|x_k - x_{k-1}\|^2. \end{aligned} \tag{24}$$

Combining (22) with (23) and (24), together with (18) and using that $(\alpha_k)_{k\geq 1}$ is nondecreasing, we obtain for all $k \geq k_0 := \max\{k_1, k_\rho\}$

$$\begin{aligned}&\|x_{k+1}-x^*\|^2 - \alpha_{k+1}\|x_k - x^*\|^2 \\ \leq\ &\|x_{k+1}-x^*\|^2 - \alpha_k\|x_k - x^*\|^2 \\ \leq\ &\|x_k - x^*\|^2 - \alpha_k\|x_{k-1} - x^*\|^2 + \alpha_k(1+\alpha_k)\|x_k - x_{k-1}\|^2 \\ &- \left(\frac{2}{\rho_k(1+\mu\theta_k)} - 1\right)\left[(1-\alpha_k)\|x_{k+1}-x_k\|^2 + \left(\alpha_k^2 - \alpha_k\right)\|x_k - x_{k-1}\|^2\right] \\ =\ &\|x_k - x^*\|^2 - \alpha_k\|x_{k-1} - x^*\|^2 + 2\alpha_k\left(\alpha_k + \frac{1-\alpha_k}{\rho_k(1+\mu\theta_k)}\right)\|x_k - x_{k-1}\|^2 \\ &- (1-\alpha_k)\left(\frac{2}{\rho_k(1+\mu\theta_k)} - 1\right)\|x_{k+1}-x_k\|^2, \end{aligned}$$

which is nothing else than (20). ■

In order to further proceed with the convergence analysis, we have to choose the sequences $(\alpha_k)_{k\geq 1}$ and $(\rho_k)_{k\geq 1}$ such that

$$\liminf_{k\to+\infty} \delta_k > 0.$$

This is a manageable task, since we can choose for example the two sequences such that $\lim_{k\to+\infty} \alpha_k = \alpha \geq 0$, and $\lim_{k\to+\infty} \rho_k = \rho > 0$. Recalling that $\lim_{k\to+\infty} \theta_k = 1$, we obtain

$$\lim_{k\to+\infty} \delta_k = (1-\alpha)\left(\frac{2}{\rho(1+\mu)} - 1\right) - 2\alpha\left(\alpha + \frac{1-\alpha}{\rho(1+\mu)}\right) = \frac{2(1-\alpha)^2}{\rho(1+\mu)} - 1 + \alpha - 2\alpha^2,$$

thus, in order to guarantee $\lim_{k\to+\infty} \delta_k > 0$ it is sufficient to choose ρ such that

$$0 < \rho < \frac{2}{(1+\mu)}\frac{(1-\alpha)^2}{(2\alpha^2-\alpha+1)}. \tag{25}$$

Moreover, for $\alpha \geq 0$ we have

$$\frac{(1-\alpha)^2}{2\alpha^2-\alpha+1} \leq 1,$$

and thus

$$\varepsilon := \frac{1}{2}\left(\frac{2}{1+\mu} - \rho\right) > 0.$$

Then there exist k_0, $k_1 \geq 1$ such that

$$\rho_k \leq \rho + \varepsilon \quad \forall k \geq k_0, \qquad \frac{2}{1+\mu} - \varepsilon \leq \frac{2}{1+\mu\theta_k} \quad \forall k \geq k_1.$$

We obtain that for all $k \geq k_\rho := \max\{k_0, k_1\}$ we have

$$0 < \rho_k \leq \rho + \varepsilon = \frac{1}{2}\left(\rho + \frac{2}{1+\mu}\right) = \frac{2}{1+\mu} - \varepsilon \leq \frac{2}{1+\mu\theta_k},$$

hence (18) is fulfilled naturally with the choice (25).

Remark 9 (inertia versus relaxation) *Inequality* (25) *represents the necessary trade-off between inertia and relaxation (see Figure 1 for two particular choices of μ). The expression is similar to the one obtained by (Attouch and Cabot, 2019b, Remark 2.13), the exception being an additional factor incorporating the stepsize parameter μ. This means that for given $0 \leq \alpha < 1$ the upper bound for the relaxation parameter is $\overline{\rho}(\alpha,\mu) = \frac{2}{(1+\mu)}\frac{(1-\alpha)^2}{(2\alpha^2-\alpha+1)}$. We further see that $\alpha \mapsto \overline{\rho}(\alpha,\mu)$ is a decreasing function on the interval $[0,1]$. Hence, the maximal value for the limit of the sequence of relaxation parameters is obtained when $\alpha = 0$ and is $\rho_{\max}(\mu) := \overline{\rho}(0,\mu) = \frac{2}{1+\mu}$. On the other hand, when $\alpha \nearrow 1$, then $\overline{\rho}(\alpha,\mu) \searrow 0$. In addition, the function $\mu \mapsto \rho_{\max}(\mu)$ is also decreasing on $[0,1]$ with limiting values 2 as $\mu \searrow 0$, and 1 as $\mu \nearrow 1$.*

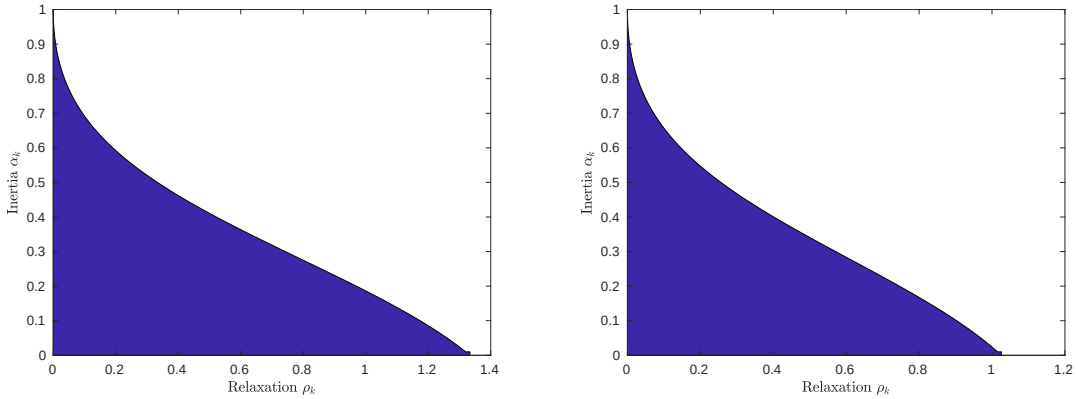


Figure 1: Trade-off between inertia and relaxation for $\mu = 0.5$ (left) and $\mu = 0.95$ (right).

Proposition 10 *Let $(\alpha_k)_{k\geq 1}$ be a nondecreasing sequence of nonnegative numbers with $0 \leq \alpha_k \leq \alpha < 1$ for all $k \geq 1$ and $(\rho_k)_{k\geq 1}$ a sequence of positive numbers such that (18) holds, with the property that $\liminf_{k\to+\infty} \delta_k > 0$. For $x^* \in \mathrm{Zeros}(A+B)$ we define the sequence $(H_k)_{k\geq 1}$ as in (19). Then the following statements are true:*

(i) The sequence $(x_k)_{k\geq 0}$ is bounded.

(ii) There exists $\lim_{k\to+\infty} H_k \in \mathbb{R}$.

(iii) $\sum_{k=1}^{+\infty} \delta_k \|x_{k+1} - x_k\|^2 < +\infty$.

Proof (i) From (20) and $\liminf_{k\to+\infty} \delta_k > 0$ we can conclude that there exists $k_0 \geq 1$ such that the sequence $(H_k)_{k\geq k_0}$ is nonincreasing. Therefore for all $k \geq k_0$ we have

$$H_{k_0} \geq H_k \geq \|x_k - x^*\|^2 - \alpha_k \|x_{k-1} - x^*\|^2 \geq \|x_k - x^*\|^2 - \alpha \|x_{k-1} - x^*\|^2$$

and, from here,

$$\begin{aligned} \|x_k - x^*\|^2 &\leq \alpha \|x_{k-1} - x^*\|^2 + H_{k_0} \\ &\leq \alpha^{k-k_0} \|x_{k_0} - x^*\|^2 + H_{k_0}(1 + \alpha + ... + \alpha^{k-k_0-1}) \\ &= \alpha^{k-k_0} \|x_{k_0} - x^*\|^2 + H_{k_0} \frac{1-\alpha^{k-k_0}}{1-\alpha}, \end{aligned}$$

which implies that $(x_k)_{k\geq 0}$ is bounded and so is $(H_k)_{k\geq 1}$.

(ii) Since $(H_k)_{k\geq k_0}$ is nonincreasing and bounded, it converges to a real number.

(iii) Follows from (ii) and Proposition 8. ■

We are now in the position to prove the main result of this section. In order to do so, we first recall two useful lemmas.

Lemma 11 (Alvarez and Attouch, 2001) *Let $(\varphi_k)_{k\geq 0}$, $(\alpha_k)_{k\geq 1}$, and $(\psi_k)_{k\geq 1}$ be sequences of nonnegative real numbers satisfying*

$$\varphi_{k+1} \leq \varphi_k + \alpha_k(\varphi_k - \varphi_{k-1}) + \psi_k \quad \forall k \geq 1, \quad \sum_{k=1}^{+\infty} \psi_k < +\infty,$$

and such that $0 \leq \alpha_k \leq \alpha < 1$ for all $k \geq 1$. Then the limit $\lim_{k\to+\infty} \varphi_k \in \mathbb{R}$ exists.

Lemma 12 (discrete Lemma of Opial, 1967) *Let C be a nonempty subset of H and $(x_k)_{k\geq 0}$ be a sequence in H such that the following two conditions hold:*

(i) For every $x \in C$, $\lim_{k\to+\infty} \|x_k - x\|$ exists.

(ii) Every weak sequential cluster point of $(x_k)_{k\geq 0}$ is in C.

Then $(x_k)_{k\geq 0}$ converges weakly to an element in C.

Theorem 13 *Let $(\alpha_k)_{k\geq 1}$ be a nondecreasing sequence of nonnegative numbers with $0 \leq \alpha_k \leq \alpha < 1$ for all $k \geq 1$ and $(\rho_k)_{k\geq 1}$ a sequence of positive numbers such that (18) holds, with the property that $\liminf_{k\to+\infty} \delta_k > 0$. Then the sequence $(x_k)_{k\geq 0}$ converges weakly to some element in $\mathrm{Zeros}(A+B)$ as $k \to +\infty$.*

Proof The result will be a consequence of the discrete Opial Lemma. To this end we will prove that the conditions (i) and (ii) in Lemma 12 for $C := \mathrm{Zeros}(A+B)$ are satisfied.

Let $x^* \in \mathrm{Zeros}(A+B)$. Indeed, it follows from (22) and (23) that for k large enough

$$\|x_{k+1} - x^*\|^2 \leq (1+\alpha_k)\|x_k - x^*\|^2 - \alpha_k\|x_{k-1} - x^*\|^2 + \alpha_k(1+\alpha_k)\|x_k - x_{k-1}\|^2.$$

Therefore, according to Lemma 11 and Proposition 10 (iii), $\lim_{k\to+\infty} \|x_k - x^*\|$ exists.

Let $\bar{x}$ be a weak limit point of $(x_k)_{k\geq 0}$ and a subsequence $(x_{k_l})_{l\geq 0}$ which converges weakly to $\bar{x}$ as $l \to +\infty$. From Proposition 10 (iii) we have

$$\lim_{k\to+\infty} \delta_k \|x_{k+1} - x_k\|^2 = 0,$$

which, as $\liminf_{k\to+\infty} \delta_k > 0$, yields $\lim_{k\to+\infty} \|x_{k+1} - x_k\| = 0$. Recall that for $k \geq 1$ we have $w_k = y_k - \lambda_k(By_k - Bz_k)$. Since

$$\|w_k - z_k\| = \frac{1}{\rho_k}\|x_{k+1} - z_k\| = \frac{1}{\rho_k}\|x_{k+1} - x_k + \alpha_k(x_k - x_{k-1})\| \quad \forall k \geq 1,$$

we have $\lim_{k\to+\infty} \|w_k - z_k\| = 0$. On the other hand, for all $k \geq 1$ holds

$$\begin{aligned} \|w_k - z_k\| &= \|y_k - z_k - \lambda_k(By_k - Bz_k)\| \geq \|y_k - z_k\| - \lambda_k\|By_k - Bz_k\| \\ &\geq (1 - \mu\theta_k)\,\|y_k - z_k\|. \end{aligned}$$

Since $\lim_{k\to+\infty} (1 - \mu\theta_k) = 1 - \mu > 0$, we can conclude that $\|y_k - z_k\| \to 0$ as $k \to +\infty$. Furthermore, for all $k \geq 1$ we have $z_k = x_k + \alpha_k(x_k - x_{k-1})$ and since $\|x_{k+1} - x_k\| \to 0$ as $k \to +\infty$ we deduce that $(z_{k_l})_{l\geq 0}$ converges weakly to $\bar{x}$ as $l \to +\infty$. Combining this with

the fact that $\|y_k - z_k\| \to 0$ as $k \to +\infty$ then shows that $(y_{k_l})_{l\geq 0}$ and $(z_{k_l})_{l\geq 0}$ converges weakly to $\bar{x}$ as $l \to +\infty$, too. The definition of $(y_{k_l})_{l\geq 0}$ gives

$$\frac{1}{\lambda_{k_l}}(z_{k_l} - y_{k_l}) + By_{k_l} - Bz_{k_l} \in (A+B)y_{k_l} \quad \forall l \geq 0.$$

Using that $\left(\frac{1}{\lambda_{k_l}}(z_{k_l} - y_{k_l}) + By_{k_l} - Bz_{k_l}\right)_{l\geq 0}$ converges strongly to 0 and that the graph of the maximal monotone operator $A+B$ is sequentially closed with respect to the weak-strong topology of the product space $H \times H$, we obtain $0 \in (A+B)\bar{x}$, thus $\bar{x} \in \text{Zeros}(A+B)$. ■

Remark 14 In the particular case of the variational inequality (4), which corresponds to the case when A is the normal cone of a nonempty closed convex subset C of H, by taking into account that $J_{\lambda N_C} = P_C$ is for all $\lambda > 0$ the projection operator onto C, the relaxed inertial FBF algorithm reads

$$(RIFBF-VI) \qquad (\forall k \geq 1) \begin{cases} z_k = x_k + \alpha_k(x_k - x_{k-1}) \\ y_k = P_C(I - \lambda_k B)z_k \\ x_{k+1} = (1-\rho_k)z_k + \rho_k\left(y_k - \lambda_k(By_k - Bz_k)\right), \end{cases}$$

where $x_0, x_1 \in H$ are starting points, $(\lambda_k)_{k\geq 1}$ and $(\rho_k)_{k\geq 1}$ are sequences of positive numbers, and $(\alpha_k)_{k\geq 1}$ is a sequence of nonnegative numbers.

The algorithm converges weakly to a solution of (4) when B is a monotone and Lipschitz continuous operator in the hypotheses of Theorem 13.

As it has been shown in (Boţ et al., 2020), it also converges when B is pseudo-monotone on H, Lipschitz continuous, and sequentially weak-to-weak continuous, and also when H is finite dimensional, and B is pseudo-monotone on C and Lipschitz continuous.

We recall that B is said to be pseudo-monotone on C (on H) if for all $x, y \in C$ ($x, y \in H$) it holds

$$\langle Bx, y - x\rangle \geq 0 \;\Rightarrow\; \langle By, y - x\rangle \geq 0.$$

Denoting $w_k := y_k - \lambda_k(By_k - Bz_k)$ and $\theta_k := \frac{\lambda_k}{\lambda_{k+1}}$ for all $k \geq 1$, then for all $x^* \in \text{Zeros}(N_C + B)$ it holds

$$\|w_k - x^*\|^2 \leq \|z_k - x^*\|^2 - \left(1 - \mu^2\theta_k^2\right)\|y_k - z_k\|^2 \quad \forall k \geq 1$$

which is nothing else than relation (16).

Indeed, since $y_k \in C$, we have $\langle Bx^*, y_k - x^*\rangle \geq 0$, and, further, the pseudo-monotonicity of B gives $\langle By_k, y_k - x^*\rangle \geq 0$ for all $k \geq 1$. On the other hand, since $y_k = P_C(I - \lambda_k B)z_k$, we have $\langle x^* - y_k, y_k - z_k + \lambda_k Bz_k\rangle \geq 0$ for all $k \geq 1$. The two inequalities yield

$$\langle y_k - x^*, z_k - w_k\rangle \geq 0 \quad \forall k \geq 1,$$

which, combined with (17), lead as in the proof of Proposition 7 to the conclusion.

Now, since (16) holds, the statements in Proposition 8 and Proposition 10 remain true and, as seen in the proof of Theorem 13, they guarantee that the limit $\lim_{k\to+\infty}\|x_k - x^*\| \in$

$\mathbb{R}$ exists, and that $\lim_{k\to+\infty} \|y_k - z_k\| = \lim_{k\to+\infty} \|By_k - Bz_k\| = 0$. Having that, the weak converge of $(x_k)_{k\geq 0}$ to a solution of (4) follows, by arguing as in the proof by (Boţ et al., 2020, Theorem 3.1), when B is pseudo-monotone on H, Lipschitz-continuous and sequentially weak-to-weak continuous, and as in the proof by (Boţ et al., 2020, Theorem 3.2), when H is finite dimensional, and B is pseudo-monotone on C and Lipschitz continuous.

4. Numerical Experiments

In this section, we provide two numerical experiments that complete our theoretical results. The first one, a bilinear saddle point problem, is a usual deterministic example where all the assumptions are fulfilled to guarantee convergence. For the second experiment we leave the save harbour of justified assumptions and frankly use RIFBF to treat a more complex (stochastic) problem and train a special kind of generative machine learning system that receives a lot of attention recently.

4.1 Bilinear Saddle Point Problem

We want to solve the following bilinear saddle-point problem

$$\min_{u\in U} \max_{v\in V} \Phi(u,v) := u^T A v + a^T u + b^T v, \tag{26}$$

with $A \in \mathbb{R}^{m\times n}$, $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$ and where $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ are nonempty, closed and convex sets, in the sense that we want to find $u^* \in U$ and $v^* \in V$ such that

$$\Phi(u^*, v) \leq \Phi(u^*, v^*) \leq \Phi(u, v^*)$$

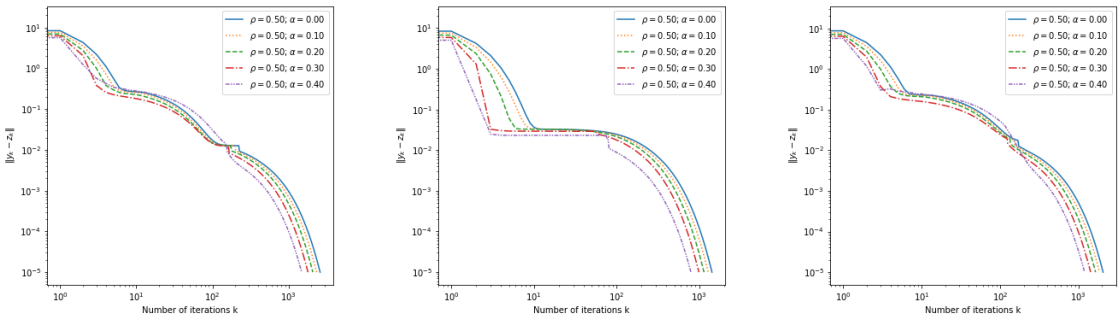
for all $u \in U$ and all $v \in V$. The monotone inclusion to be solved in this case is of the form

$$0 \in N_{U\times V}(u,v) + F(u,v),$$

where $N_{U\times V}$ is the maximal monotone normal cone operator to $U \times V$ and $F(u,v) = M\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} a \\ -b \end{pmatrix}$, with $M = \begin{pmatrix} 0 & A \\ -A^T & 0 \end{pmatrix}$, is monotone and Lipschitz continuous with Lipschitz constant $L = \|M\|_2$. Notice that F is not cocoercive.

For our experiments we choose $m = n = 500$, and A, a and b to have entries drawn from different random distributions; we look at the uniform distribution on the interval $[0, 1]$, the standard normal distribution and the 1-Poisson distribution. For the constraint sets U and V we take unit balls in the Euclidean norm. We use constant stepsize $\lambda_k = \lambda = \frac{\mu}{L}$, where $0 < \mu < 1$, constant inertial parameter $\alpha_k = \alpha$ and constant relaxation parameter $\rho_k = \rho$ for all $k \geq 1$. We set $x_k := (u_k, v_k)$ to fit the framework of our algorithm. The starting point x_0 is initialized randomly (with entries drawn from the uniform distribution on $[0, 1]$ for all three settings) and we set $x_1 = x_0$. We did not observe different behavior for various random trials, so each time we provide the results for only one run in the following.

In Figure 2 we can see the development of $\|y_k - z_k\|$ for problem (26) with data drawn from the different distributions for $\mu = 0.5$, $\rho = 0.5$ and various values of α. The quantity $\|y_k - z_k\|$ is the fixed point residual of the operator $J_{\lambda A}(I - \lambda B)$, for which according to the convergence analysis we have that $\|y_k - z_k\| \to 0$ as $k \to +\infty$. As solutions of the



(a) uniform (b) standard normal (c) Poisson

Figure 2: Behavior of the fixed point residual $\|y_k - z_k\|$ for the constrained bilinear saddle-point problem with data drawn from different distributions ($\mu = 0.5$).

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.32
0.00	$\geq$ 10,000	$\geq$ 10,000	6,493	4,327	3,245	2,596	2,166	1,860	1,629	1,447	1,234	1,108	1,017	941	929
0.04	$\geq$ 10,000	$\geq$ 10,000	6,233	4,154	3,115	2,493	2,080	1,787	1,562	1,375	1,171	1,065	977	-	-
0.08	$\geq$ 10,000	$\geq$ 10,000	5,973	3,981	2,985	2,389	1,995	1,713	1,496	1,277	1,121	1,024	937	-	-
0.12	$\geq$ 10,000	$\geq$ 10,000	5,713	3,807	2,856	2,286	1,910	1,635	1,427	1,194	1,077	984	-	-	-
0.16	$\geq$ 10,000	$\geq$ 10,000	5,453	3,634	2,726	2,184	1,824	1,551	1,336	1,140	1,038	-	-	-	-
0.20	$\geq$ 10,000	$\geq$ 10,000	5,192	3,461	2,597	2,082	1,735	1,461	1,231	1,099	-	-	-	-	-
0.24	$\geq$ 10,000	9,871	4,932	3,287	2,468	1,976	1,621	1,374	1,170	-	-	-	-	-	-
0.28	$\geq$ 10,000	9,350	4,671	3,114	2,339	1,858	1,502	1,282	-	-	-	-	-	-	-
0.32	$\geq$ 10,000	8,830	4,411	2,943	2,204	1,734	1,396	-	-	-	-	-	-	-	-
0.36	$\geq$ 10,000	8,309	4,150	2,770	2,035	1,589	1,327	-	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	7,787	3,891	2,571	1,866	1,486	-	-	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	7,266	3,633	2,351	1,730	-	-	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	6,744	3,340	2,149	-	-	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	6,223	3,012	2,000	-	-	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	5,707	2,709	-	-	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	5,130	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	4,480	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	3,968	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 1: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a uniform distribution ($\mu = 0.5$)

problem (26) are not available explicitly, we look at this meaningful surrogate instead. Also, if we have $y_k = z_k$ for some $k \geq 1$ then $x_{k+1} = y_k = z_k$ is a solution of (26).

We see that the behavior of the residual is similar for most combinations of parameters and all three settings. When the inertial parameter α is larger, the algorithm takes fewer iterations for the fixed point residual to reach the considered threshold of 10^{-5}. However, when α gets close to the limiting case the performance is not consistently better throughout the entire run anymore when the data is drawn from the uniform or the Poisson distribution.

Temporarily the residual is even worse than for smaller α, nevertheless the algorithm still terminates in fewer iterations in the end.

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.04
0.00	$\geq$ 10,000	7,850	3,927	2,620	1,967	1,574	1,309	1,110	948	837	734	705
0.04	$\geq$ 10,000	7,536	3,770	2,516	1,889	1,511	1,256	1,054	908	791	704	-
0.08	$\geq$ 10,000	7,222	3,613	2,411	1,810	1,447	1,201	999	865	751	-	-
0.12	$\geq$ 10,000	6,908	3,456	2,307	1,731	1,383	1,141	946	822	-	-	-
0.16	$\geq$ 10,000	6,595	3,300	2,202	1,652	1,317	1,071	893	774	-	-	-
0.20	$\geq$ 10,000	6,281	3,143	2,098	1,572	1,248	999	843	-	-	-	-
0.24	$\geq$ 10,000	5,967	2,987	1,993	1,491	1,170	934	-	-	-	-	-
0.28	$\geq$ 10,000	5,653	2,830	1,887	1,405	1,083	879	-	-	-	-	-
0.32	$\geq$ 10,000	5,340	2,674	1,780	1,302	1,000	-	-	-	-	-	-
0.36	$\geq$ 10,000	5,026	2,517	1,664	1,196	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	4,713	2,357	1,531	1,105	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	4,400	2,189	1,390	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	4,088	1,999	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	3,773	1,789	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	3,443	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	3,077	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	2,663	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.76	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.80	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.84	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-

Table 2: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a uniform distribution ($\mu = 0.9$)

In Table 1 we can see the necessary number of iterations for the algorithm to achieve $\|y_k - z_k\| \leq 10^{-5}$ for $\mu = 0.5$ and different choices of the parameters α and ρ. As the results are very similar for all three considered distributions, here we only report the case of data drawn from the uniform distribution on the interval $[0, 1]$. For the tables regarding the experiments for data drawn from the standard normal distribution and the 1-Poisson distribution please refer to Appendix B.

As mentioned in Remark 9, there is a trade-off between inertia and relaxation. The parameters α and ρ also need to fulfil the relations

$$0 \leq \alpha < 1 \quad \text{and} \quad 0 < \rho < \frac{2}{(1+\mu)} \frac{(1-\alpha)^2}{(2\alpha^2 - \alpha + 1)},$$

which is the reason why not every combination of α and ρ is valid.

We see that for a particular choice of the relaxation parameter the least number of iterations is achieved when the inertial parameter is as large as possible. If, on the other hand, we fix the inertial parameter, we observe that also larger values of ρ are better and lead to fewer iterations. To get a conclusion regarding the trade-off between the two parameters, Table 1 suggests that the influence of the relaxation parameter is stronger than that of the inertial parameter. Given that no numerical experiments are available for the

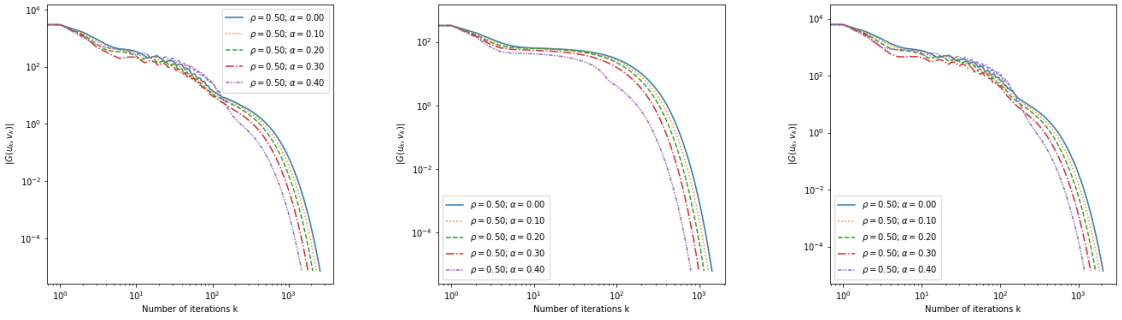
α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80
0.00	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,337	7,143	6,247	5,553	4,975	4,274	4,200	3,818	3,406	3,272	3,222	3,146	3,230
0.04	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,607	8,003	6,856	5,997	5,334	4,639	4,246	4,041	3,579	3,333	3,235	3,161	2,846	-
0.08	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,206	7,668	6,569	5,746	5,114	4,314	4,204	3,837	3,382	3,256	3,211	3,209	-	-
0.12	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,804	7,332	6,281	5,493	4,774	4,255	4,038	3,617	3,298	3,242	3,314	-	-	-
0.16	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,402	6,996	5,992	5,227	4,395	4,215	3,837	3,428	3,338	3,371	-	-	-	-
0.20	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	7,998	6,658	5,685	4,906	4,261	4,048	3,656	3,415	3,574	-	-	-	-	-
0.24	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,501	7,592	6,309	5,263	4,497	4,225	3,841	3,626	-	-	-	-	-	-	-
0.28	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,994	7,178	5,872	4,861	4,297	4,065	3,835	-	-	-	-	-	-	-	-
0.32	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,481	6,666	5,364	4,612	4,269	4,105	-	-	-	-	-	-	-	-	-
0.36	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	7,906	6,098	5,053	4,548	4,450	-	-	-	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,961	7,205	5,649	5,008	4,924	-	-	-	-	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,109	6,572	5,520	5,490	-	-	-	-	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,174	6,356	6,115	-	-	-	-	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	7,633	6,785	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	7,642	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	$\geq$ 10,000	9,747	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 3: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a uniform distribution ($\mu = 0.1$)

relaxed inertial proximal or forward-backward algorithms, we cannot say if this is a common phenomenon of relaxed inertial methods or one that is specific to RIFBF.

Even though the possible values for α get smaller if ρ goes to $\rho_{\max}(\mu) = \frac{2}{1+\mu}$, the number of iterations gets less for $\alpha > 0$. In particular, we want to point out that over-relaxation ($\rho > 1$) seems to be highly beneficial.

Comparing the results for different choices of μ in Table 2 ($\mu = 0.9$) and Table 3 ($\mu = 0.1$), we can observe a very interesting behavior. Even though smaller μ allows for larger values of the relaxation parameter (see Remark 9), larger μ leads to better results in general. This behavior presumably is due to the role μ plays in the definition of the stepsize.



(a) uniform (b) standard normal (c) Poisson

Figure 3: Behavior of the gap function $|G(u_k, v_k)|$ for the constrained bilinear saddle-point problem with data drawn from different distributions ($\mu = 0.5$).

To get further insight into the convergence behavior, we also look at the following gap function,

$$G(s,t) := \inf_{u \in U} \Phi(u,t) - \sup_{v \in V} \Phi(s,v).$$

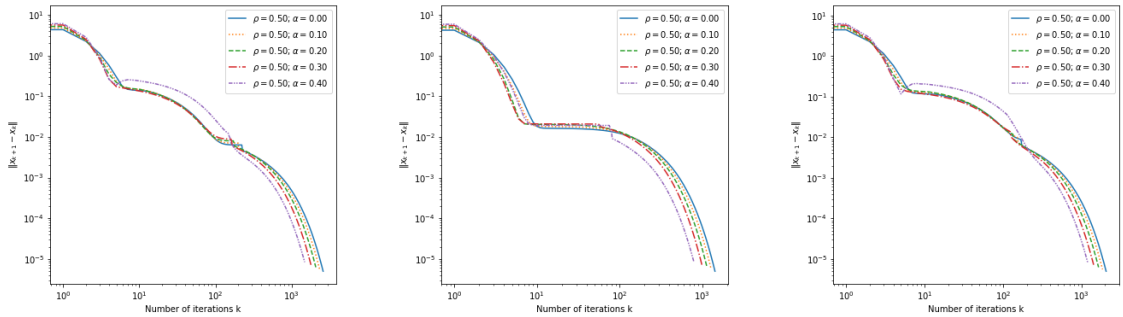
The quantity $(G(u_k, v_k))_{k \geq 0}$ should be a measure to judge the performance of the iterates $(u_k, v_k)_{k \geq 0}$, as for the optimum (u^*, v^*) we have $\Phi(u^*, v) \leq \Phi(u^*, v^*) \leq \Phi(u, v^*)$ for all $u \in U$ and all $v \in V$ and hence $G(u^*, v^*) = 0$.
Because of the particular choice of a bilinear objective and the constraint sets U and V the expressions can be actually computed in closed form and we get

$$G(u_k, v_k) = -\|Av_k + a\|_2 + b^T v_k - \|A^T u_k + b\|_2 - a^T u_k \quad \forall k \geq 0.$$

In Figure 3 we see the development of the absolute value of the gap $|G(u_k, v_k)|$ for problem (26) with data drawn from the different distributions for $\mu = 0.5$, $\rho = 0.5$ and various values of α. We see that, as in the case of the residual of the fixed point iteration, the behavior is similar for most combinations of parameters – the larger the inertial parameter α the better the results in general. However, in the case of the data being drawn from the uniform or the Poisson distribution, the behavior of the gap is not consistently better anymore when α gets close to the limiting case. In the meantime the gap for maximal α is

worst compared to all other parameter combinations, altough the algorithm still gives the best results in the end. As the theory suggests, the gap indeed decreases and tends to zero as the number of iterations grows larger.

Finally, in Figure 4 we look at the discrete velocity of the iterates $\|x_{k+1} - x_k\|$ which is square summable and thus vanishes for $k \to +\infty$. Again we look at the three settings where the data is drawn from a uniform, the standard normal and a Poisson distribution and plot the results for $\mu = 0.5$, $\rho = 0.5$ and various values of α. Once more larger inertial parameters give better results, with the biggest difference to the previous instances being an intermediate phase approximately up to iteration 100 where we do not have a clear picture yet and which occurs for all distributions.



(a) uniform (b) standard normal (c) Poisson

Figure 4: Behavior of the discrete velocity of the iterates $\|x_{k+1} - x_k\|$ for the constrained bilinear saddle-point problem with data drawn from different distributions ($\mu = 0.5$).

4.2 Generative Adversarial Networks (GANs)

Generative Adversarial (Artificial Neural) Networks (GANs) is a class of machine learning systems, where two "adversarial" networks compete in a (zero-sum) game against each other. Given a training set, this technique aims to learn to generate new data with the same statistics as the training set. The generative network, called generator, tries to mimic the original genuine data (distribution) and strives to produce samples that fool the other network. The opposing, discriminative network, called discriminator, evaluates both true samples as well as generated ones and tries to distinguish between them. The generator's objective is to increase the error rate of the discriminative network by producing novel samples that the discriminator thinks were not generated, while its opponent's goal is to successfully judge (with high certainty) whether the presented data is true or not. Typically, the generator learns to map from a latent space to a data distribution which is (hopefully) similar to the original distribution. However, one does not have access to the distribution itself, but only random samples drawn from it. The original formulation of GANs

by (Goodfellow et al., 2014) is as follows,

$$\min_u \max_v \Phi(u, v),$$

where $\Phi(u, v) = \mathbb{E}_{x\sim p}\left[\log\left(D_v(x)\right)\right] + \mathbb{E}_{x'\sim q_u}\left[\log\left(1 - D_v(x')\right)\right]$ is the value function, u and v is the parametrisation of the generator and discriminator, respectively, D_v is the probability how certain the discriminator is that the input is real, and p and q_u is the real and learned distribution, respectively.

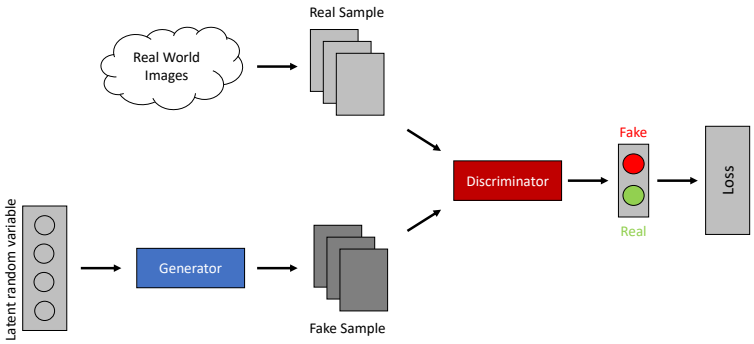


Figure 5: Schematic illustration of Generative Adversarial Networks.

Again, the problem

$$\min_{u\in U} \max_{v\in V} \Phi(u, v)$$

is understood in the sense that we want to find $u^* \in U$ and $v^* \in V$ such that

$$\Phi(u^*, v) \leq \Phi(u^*, v^*) \leq \Phi(u, v^*)$$

for all $u \in U$ and all $v \in V$. The corresponding inclusion problem then reads

$$0 \in N_{U\times V}(u, v) + F(u, v),$$

where $N_{U\times V}$ is the normal cone operator to $U \times V$ and $F(u, v) = \left(\nabla_u \Phi(u, v), -\nabla_v \Phi(u, v)\right)^T$. If U and V are nonempty, convex and closed sets, and $\Phi(u, v)$ is convex-concave, Fréchet differentiable and has a Lipschitz continuous gradient, then we have a variational inequality, the obtained theory holds and we can apply RIFBF-VI.

This motivates to use (variants of) FBF methods for the training of GANs, even though in practice the used value functions typically are not convex-concave and further, the gradient might not be Lipschitz continuous if it exists at all. Additionally, in general one needs stochastic versions of the used algorithms and which we do not provide in the case of RIFBF. A stochastic variant of the relaxed inertial forward-backward-forward algorithm RIFBF has been proposed and analyzed in (Cui et al., 2021) one and a half years after the first version of this article. The convergence analysis of the stochastic numerical method massively relies on the one carried out for RIFBF.

Recently, a first successful attempt of using methods coming from the field of variational inequalities for GAN training was done by (Gidel et al., 2019). In particular they applied the well established extra-gradient algorithm and some derived variations. In this spirit we frankly apply the FBF method and a variant with inertial effect ($\alpha = 0.05$), as well as a primal-dual algorithm for saddle point problems, introduced by (Hamedani and Aybat, 2021) (PDSP), which to the best of our knowledge has not been used for GAN training before, and compare the results to the best method (Extra Adam) from the work by (Gidel et al., 2019).

For our experiments we use the standard DCGAN architecture (Radford et al., 2016) where generator and discriminator consist (among other elements) of several convolutional and convolutional-transpose layers, respectively, that have shown to work very well with images. The opposing networks are trained on the CIFAR10 dataset (Krizhevsky, 2009) with the WGAN objective and weight clipping (Arjovsky et al., 2017), which uses the idea to minimize the Wasserstein distance between the true distribution and the one learned by the generator. Note that in absence of bound constraints on the weights of at least one of the two networks, the backward step in the FBF algorithm would be redundant, as we would project on the whole space and we obtain the unconstrained extra-gradient method.

Furthermore, in our experiments instead of stochastic gradients we use the Adam optimizer (Kingma and Ba, 2014) with the hyperparameters ($\beta_1 = 0.5$, $\beta_2 = 0.9$) that where used by (Gidel et al., 2019), as the best results there were achieved with this choice. Also we would like to mention that we only did a hyperparameter search for the stepsizes of the newly introduced methods, all other parameters we chose to be equal as in the aforementioned work.

Method	IS	FID
PDSP Adam	4.20 ± 0.04	53.97 ± 0.28
Extra Adam	4.07 ± 0.05	56.67 ± 0.61
FBF Adam	4.54 ± 0.04	45.85 ± 0.35
IFBF Adam ($\alpha = 0.05$)	$\mathit{4.59 \pm 0.04}$	$\mathit{45.25 \pm 0.60}$

Table 4: Best IS and FID scores (averaged over 5 runs) achieved on CIFAR10.

The model is evaluated using the inception score (IS) (reworked implementation by (Barratt and Sharma, 2018) that fixes some issues of the original one) as well as the Fréchet inception distance (FID) (Heusel et al., 2017), both computed on 50,000 samples. Experiments were run with 5 random seeds for 500,000 updates of the generator on a NVIDIA GeForce RTX 2080Ti GPU. Table 4 reports the best IS and FID achieved by each considered method. Note that the values of IS for Extra Adam differ from those stated by (Gidel et al., 2019), due to the usage of the corrected implementation of the score.

We see that even though we only have proved convergence for the monotone case in the deterministic setting, the variants of RIFBF perform well in the training of GANs. IFBF Adam outperforms all other considered methods, both for the IS and the FID. As the theory suggests, making use of some inertial effects (regardless that Adam already incorporates some momentum) seems to provide additional improvement of the numerical method in practice. The results suggest, that employing methods that are designed to

capture the nature of a problem, in this case a constrained minimax/saddle-point problem, is highly beneficial. In Figure 6 we provide samples of the generator trained with the different methods.

Acknowledgments

The authors would like to thank Julius Berner and Axel Böhm for fruitful discussions and their valuable suggestions and comments, and also to the anonymous reviewers for their recommendations which have improved the quality of the paper. Radu Boţ would like to acknowledge partial support from the Austrian Science Fund (FWF), projects I 2419-N32 and W 1260. Michael Sedlmayer would like to acknowledge support from the Austrian Research Promotion Agency (FFG), project "Smart operation of wind turbines under icing conditions (SOWINDIC)".

(a) PDSP Adam

(b) Extra Adam

(c) FBF Adam

(d) IFBF Adam

Figure 6: Comparison of the samples of a WGAN with weight clipping trained with the different methods.

Appendix A. Proof of Theorem 4

We will prove that from

$$\ddot{h}(t) + \gamma(t)\dot{h}(t) + \frac{1-\lambda L}{(1+\lambda L)^2}\tau(t)\|Mx(t)\|^2 \leq \|\dot{x}(t)\|^2 \quad \forall t \in [0,+\infty)$$

one obtains the statements (i) and (ii) in Theorem 4 as well as the existence of the limit $\lim_{t\to+\infty}\|x(t)-x^*\| \in \mathbb{R}$. We denote $\kappa := \frac{1-\lambda L}{(1+\lambda L)^2} > 0$.

Taking again into account (8) one obtains for every $t \in [0,+\infty)$

$$\ddot{h}(t) + \gamma(t)\dot{h}(t) + \frac{\kappa}{\tau(t)}\|\ddot{x}(t) + \gamma(t)\dot{x}(t)\|^2 \leq \|\dot{x}(t)\|^2$$

or, equivalently,

$$\ddot{h}(t) + \gamma(t)\dot{h}(t) + \frac{\kappa\gamma(t)}{\tau(t)}\frac{d}{dt}\big(\|\dot{x}(t)\|^2\big) + \left(\frac{\kappa\gamma^2(t)}{\tau(t)} - 1\right)||\dot{x}(t)||^2 + \frac{\kappa}{\tau(t)}||\ddot{x}(t)||^2 \leq 0.$$

Combining this inequality with

$$\frac{\gamma(t)}{\tau(t)}\frac{d}{dt}\big(\|\dot{x}(t)\|^2\big) = \frac{d}{dt}\left(\frac{\gamma(t)}{\tau(t)}\|\dot{x}(t)\|^2\right) - \frac{\dot{\gamma}(t)\tau(t) - \gamma(t)\dot{\tau}(t)}{\tau^2(t)}\|\dot{x}(t)\|^2$$

and

$$\gamma(t)\dot{h}(t) = \frac{d}{dt}(\gamma h)(t) - \dot{\gamma}(t)h(t) \geq \frac{d}{dt}(\gamma h)(t),$$

it yields for every $t \in [0,+\infty)$

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

According to Assumption 1 we have for almost every $t \in [0,+\infty)$ the inequality

$$\ddot{h}(t) + \frac{d}{dt}(\gamma h)(t) + \kappa\frac{d}{dt}\left(\frac{\gamma(t)}{\tau(t)}\|\dot{x}(t)\|^2\right) + \nu\|\dot{x}(t)\|^2 + \frac{\kappa}{\overline{\tau}}\|\ddot{x}(t)||^2 \leq 0, \tag{27}$$

where $\overline{\tau}$ denotes a positive upper bound for the function τ. From here we obtain that the function $t \mapsto \dot{h}(t) + \gamma(t)h(t) + \kappa\frac{\gamma(t)}{\tau(t)}\|\dot{x}(t)\|^2$, which is locally absolutely continuous, is monotonically decreasing. Hence there exists a real number M such that

$$\dot{h}(t) + \gamma(t)h(t) + \kappa\frac{\gamma(t)}{\tau(t)}\|\dot{x}(t)\|^2 \leq M \quad \forall t \in [0,+\infty), \tag{28}$$

which yields

$$\dot{h}(t) + \underline{\gamma}h(t) \leq M \quad \forall t \in [0,+\infty),$$

where $\underline{\gamma}$ denotes a positive lower bound for the function γ. By multiplying this inequality with $\exp(\underline{\gamma}t)$ and then by integrating from 0 to T, where $T > 0$, it yields

$$h(T) \leq h(0)\exp(-\underline{\gamma}T) + \frac{M}{\underline{\gamma}}(1 - \exp(-\underline{\gamma}T)),$$

thus

$$h \text{ is bounded}$$

and, consequently,

$$\text{the trajectory } x \text{ is bounded.}$$

On the other hand, from (28), it follows that for every $t \in [0, +\infty)$

$$\dot{h}(t) + \frac{\kappa\underline{\gamma}}{\overline{\tau}}\|\dot{x}(t)\|^2 \leq M,$$

hence

$$\langle x(t) - x^*, \dot{x}(t)\rangle + \frac{\kappa\underline{\gamma}}{\overline{\tau}}\|\dot{x}(t)\|^2 \leq M.$$

This yields, since x is bounded, that

$$\dot{x} \text{ is bounded,}$$

and further that

$$\dot{h} \text{ is bounded.}$$

Integrating (27) we obtain that there exists a real number $N \in \mathbb{R}$ such that for every $t \in [0, +\infty)$

$$\dot{h}(t) + \gamma(t)h(t) + \kappa\frac{\gamma(t)}{\tau(t)}\|\dot{x}(t)\|^2 + \nu\int_0^t ||\dot{x}(s)||^2 ds + \frac{\kappa}{\overline{\tau}}\int_0^t ||\ddot{x}(s)||^2 ds \leq N,$$

which allows us to conclude that $\dot{x}, \ddot{x} \in L^2([0, +\infty); H)$. Finally, from (8) and Assumption 1 we deduce $Mx \in L^2([0, +\infty); H)$ and the proof of statement (i) is complete.

In order to prove (ii), we notice that for every $t \in [0, +\infty)$ it holds

$$\frac{d}{dt}\left(\frac{1}{2}\|\dot{x}(t)\|^2\right) = \langle \dot{x}(t), \ddot{x}(t)\rangle \leq \frac{1}{2}\|\dot{x}(t)\|^2 + \frac{1}{2}\|\ddot{x}(t)\|^2,$$

which, according to (i), leads to $\lim_{t\to+\infty} \dot{x}(t) = 0$.

Further, for every $t \in [0, +\infty)$ we have

$$\frac{d}{dt}\left(\frac{1}{2}\|M(x(t))\|^2\right) = \left\langle M(x(t)), \frac{d}{dt}(Mx(t))\right\rangle \leq \frac{1}{2}\|M(x(t))\|^2 + \frac{(1+\lambda L)^2(2+\lambda L)^2}{2}\|\dot{x}(t)\|^2.$$

By using again (i), we obtain $\lim_{t\to+\infty} M(x(t)) = 0$, while $\lim_{t\to+\infty} \ddot{x}(t) = 0$ follows from from (8) and Assumption 1.

Finally, we have seen that the function $t \mapsto \dot{h}(t) + \gamma(t)h(t) + \kappa\frac{\gamma(t)}{\tau(t)}\|\dot{x}(t)\|^2$ is monotonically decreasing, thus from (i), (ii) and Assumption 1 we deduce that $\lim_{t\to+\infty} \gamma(t)h(t)$ exists and it is a real number. By taking also into account that $\lim_{t\to+\infty} \gamma(t) \in (0, +\infty)$ exists, we obtain that $\lim_{t\to+\infty} \|x(t) - x^*\| \in \mathbb{R}$ exists.

Appendix B. Additional Tables

B.1 Standard Normal Distribution

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.04
0.00	$\geq$ 10,000	4,217	2,110	1,407	1,056	845	705	604	529	471	424	400
0.04	$\geq$ 10,000	4,049	2,025	1,351	1,014	811	677	580	508	452	399	-
0.08	$\geq$ 10,000	3,880	1,941	1,295	972	778	648	556	487	433	-	-
0.12	$\geq$ 10,000	3,711	1,857	1,238	929	744	620	532	466	-	-	-
0.16	$\geq$ 10,000	3,543	1,772	1,182	887	710	592	508	441	-	-	-
0.20	$\geq$ 10,000	3,374	1,688	1,126	845	677	564	481	-	-	-	-
0.24	$\geq$ 10,000	3,206	1,604	1,070	803	643	533	-	-	-	-	-
0.28	$\geq$ 10,000	3,037	1,520	1,014	761	607	498	-	-	-	-	-
0.32	$\geq$ 10,000	2,868	1,435	958	717	564	-	-	-	-	-	-
0.36	$\geq$ 10,000	2,700	1,351	901	668	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	2,531	1,267	841	610	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	2,363	1,182	770	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	2,194	1,093	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	2,026	986	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	1,857	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	1,679	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	1,459	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.76	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.80	8,437	-	-	-	-	-	-	-	-	-	-	-
0.84	6,747	-	-	-	-	-	-	-	-	-	-	-

Table 5: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from the standard normal distribution ($\mu = 0.9$)

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.32
0.00	$\geq$ 10,000	7,127	3,565	2,378	1,785	1,429	1,191	1,022	894	795	716	614	600	537	522
0.04	$\geq$ 10,000	6,842	3,423	2,283	1,713	1,372	1,144	981	859	764	668	623	572	-	-
0.08	$\geq$ 10,000	6,557	3,281	2,188	1,642	1,315	1,096	940	823	731	615	604	542	-	-
0.12	$\geq$ 10,000	6,272	3,138	2,093	1,571	1,258	1,049	900	787	693	623	578	-	-	-
0.16	$\geq$ 10,000	5,988	2,996	1,999	1,500	1,201	1,001	858	744	626	607	-	-	-	-
0.20	$\geq$ 10,000	5,703	2,853	1,904	1,429	1,144	953	813	695	621	-	-	-	-	-
0.24	$\geq$ 10,000	5,418	2,711	1,809	1,358	1,087	900	751	642	-	-	-	-	-	-
0.28	$\geq$ 10,000	5,133	2,569	1,714	1,286	1,025	841	682	-	-	-	-	-	-	-
0.32	$\geq$ 10,000	4,848	2,427	1,619	1,212	952	758	-	-	-	-	-	-	-	-
0.36	$\geq$ 10,000	4,564	2,284	1,524	1,128	861	725	-	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	4,279	2,142	1,421	1,028	795	-	-	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	3,995	1,999	1,299	925	-	-	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	3,710	1,847	1,148	-	-	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	3,426	1,663	1,102	-	-	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	3,141	1,449	-	-	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	2,839	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	2,458	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	2,172	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	8,066	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 6: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from the standard normal distribution ($\mu = 0.5$)

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80
0.00	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	9,690	7,269	5,816	4,848	4,156	3,637	3,233	2,910	2,416	2,434	2,162	1,878	1,943	1,702	1,696	1,459
0.04	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	9,303	6,978	5,584	4,654	3,990	3,492	3,104	2,677	2,515	2,316	1,981	1,929	1,828	1,616	1,588	-
0.08	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	8,916	6,688	5,351	4,460	3,824	3,346	2,970	2,404	2,448	2,186	1,853	1,916	1,650	1,685	-	-
0.12	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	8,528	6,397	5,119	4,267	3,658	3,198	2,802	2,506	2,343	2,044	1,859	1,816	1,572	-	-	-
0.16	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	8,141	6,107	4,887	4,073	3,487	3,014	2,460	2,460	2,237	1,893	1,860	1,678	-	-	-	-
0.20	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	7,754	5,816	4,654	3,876	3,299	2,796	2,485	2,343	2,130	1,752	1,795	-	-	-	-	-
0.24	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	7,366	5,526	4,419	3,653	3,022	2,551	2,473	2,227	2,023	-	-	-	-	-	-	-
0.28	$\geqslant$ 10,000	$\geqslant$ 10,000	$\geqslant$ 10,000	6,979	5,235	4,165	3,396	2,703	2,448	2,344	2,111	-	-	-	-	-	-	-	-
0.32	$\geqslant$ 10,000	$\geqslant$ 10,000	9,885	6,592	4,929	3,847	3,014	2,622	2,490	2,214	-	-	-	-	-	-	-	-	-
0.36	$\geqslant$ 10,000	$\geqslant$ 10,000	9,304	6,201	4,568	3,439	2,894	2,639	2,345	-	-	-	-	-	-	-	-	-	-
0.40	$\geqslant$ 10,000	$\geqslant$ 10,000	8,723	5,775	4,128	3,148	2,920	2,513	-	-	-	-	-	-	-	-	-	-	-
0.44	$\geqslant$ 10,000	$\geqslant$ 10,000	8,139	5,246	3,667	3,219	2,737	-	-	-	-	-	-	-	-	-	-	-	-
0.48	$\geqslant$ 10,000	$\geqslant$ 10,000	7,507	4,563	3,724	3,049	-	-	-	-	-	-	-	-	-	-	-	-	-
0.52	$\geqslant$ 10,000	$\geqslant$ 10,000	6,708	4,416	3,518	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.56	$\geqslant$ 10,000	$\geqslant$ 10,000	5,748	4,298	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.60	$\geqslant$ 10,000	$\geqslant$ 10,000	5,719	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	10,000	9,879	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	10,000	8,680	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	10,000	8,203	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	$\geqslant$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 7: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from the standard normal distribution ($\mu = 0.1$)

B.2 1-Poisson Distribution

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.04
0.00	$\geq$ 10,000	6,183	3,093	2,063	1,549	1,241	1,036	888	769	675	589	565
0.04	$\geq$ 10,000	5,936	2,969	1,981	1,487	1,192	995	851	733	638	564	-
0.08	$\geq$ 10,000	5,689	2,846	1,899	1,426	1,143	953	812	698	603	-	-
0.12	$\geq$ 10,000	5,441	2,722	1,816	1,365	1,094	911	768	662	-	-	-
0.16	$\geq$ 10,000	5,194	2,598	1,734	1,304	1,044	866	723	623	-	-	-
0.20	$\geq$ 10,000	4,947	2,475	1,653	1,242	993	814	679	-	-	-	-
0.24	$\geq$ 10,000	4,700	2,352	1,571	1,180	941	756	-	-	-	-	-
0.28	$\geq$ 10,000	4,452	2,229	1,490	1,117	882	706	-	-	-	-	-
0.32	$\geq$ 10,000	4,205	2,106	1,407	1,046	810	-	-	-	-	-	-
0.36	$\geq$ 10,000	3,958	1,984	1,321	967	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	3,711	1,861	1,224	888	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	3,465	1,733	1,125	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	3,219	1,592	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	2,975	1,441	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	2,724	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	2,444	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	2,145	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.76	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.80	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-
0.84	9,852	-	-	-	-	-	-	-	-	-	-	-

Table 8: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a Poisson distribution ($\mu = 0.9$)

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.32
0.00	$\geq$ 10,000	$\geq$ 10,000	5,149	3,432	2,573	2,059	1,717	1,474	1,293	1,153	998	894	822	763	753
0.04	$\geq$ 10,000	9,888	4,943	3,294	2,470	1,977	1,648	1,416	1,242	1,103	944	861	791	-	-
0.08	$\geq$ 10,000	9,476	4,737	3,157	2,368	1,895	1,580	1,359	1,190	1,037	904	829	762	-	-
0.12	$\geq$ 10,000	9,064	4,530	3,019	2,264	1,812	1,513	1,300	1,138	964	871	799	-	-	-
0.16	$\geq$ 10,000	8,651	4,324	2,882	2,162	1,731	1,446	1,237	1,080	919	842	-	-	-	-
0.20	$\geq$ 10,000	8,239	4,118	2,744	2,059	1,650	1,379	1,164	996	888	-	-	-	-	-
0.24	$\geq$ 10,000	7,827	3,911	2,607	1,956	1,568	1,296	1,093	942	-	-	-	-	-	-
0.28	$\geq$ 10,000	7,414	3,705	2,469	1,855	1,477	1,202	1,031	-	-	-	-	-	-	-
0.32	$\geq$ 10,000	7,002	3,498	2,332	1,751	1,381	1,119	-	-	-	-	-	-	-	-
0.36	$\geq$ 10,000	6,589	3,292	2,197	1,624	1,273	1,067	-	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	6,176	3,085	2,049	1,490	1,196	-	-	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	5,762	2,881	1,874	1,386	-	-	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	5,349	2,658	1,714	-	-	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	4,935	2,404	1,611	-	-	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	4,524	2,169	-	-	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	4,085	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	3,576	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	3,198	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 9: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a Poisson distribution ($\mu = 0.5$)

α \ ρ	0.01	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80
0.00	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,060	6,714	5,752	5,031	4,471	4,015	3,532	3,438	3,129	2,893	2,827	2,767	2,728	2,761
0.04	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,675	7,737	6,445	5,521	4,829	4,294	3,773	3,504	3,303	2,964	2,890	2,790	2,784	2,646	-
0.08	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,271	7,414	6,175	5,290	4,627	4,119	3,556	3,448	3,152	2,893	2,838	2,786	2,672	-	-
0.12	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,867	7,090	5,905	5,058	4,424	3,874	3,523	3,316	3,026	2,920	2,829	2,863	-	-	-
0.16	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,462	6,766	5,634	4,826	4,210	3,619	3,471	3,198	2,968	2,946	2,895	-	-	-	-
0.20	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,057	6,441	5,362	4,581	3,960	3,563	3,354	3,142	3,022	3,032	-	-	-	-	-
0.24	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	7,652	6,114	5,080	4,263	3,677	3,522	3,288	3,192	-	-	-	-	-	-	-
0.28	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,670	7,244	5,781	4,750	3,957	3,650	3,469	3,366	-	-	-	-	-	-	-	-
0.32	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	9,128	6,831	5,383	4,373	3,831	3,695	3,588	-	-	-	-	-	-	-	-	-
0.36	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,582	6,374	4,951	4,210	3,982	3,863	-	-	-	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	8,023	5,842	4,628	4,352	4,233	-	-	-	-	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	7,357	5,382	4,824	4,681	-	-	-	-	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	6,655	5,474	5,215	-	-	-	-	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	$\geq$ 10,000	9,424	6,385	5,869	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	$\geq$ 10,000	8,431	6,725	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.60	10,000	10,000	8,298	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	10,000	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	10,000	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	10,000	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 10: Number of iterations necessary to achieve $\|y_k - z_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a Poisson distribution ($\mu = 0.1$)

Appendix C. DCGAN Architecture

Generator
Input: $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
Linear $128 \to 512 \times 4 \times 4$
Batch Normalization
ReLU
transposed conv. (kernel: 4×4, $512 \to 256$, stride: 2, pad: 1)
Batch Normalization
ReLU
transposed conv. (kernel: 4×4, $256 \to 128$, stride: 2, pad: 1)
Batch Normalization
ReLU
transposed conv. (kernel: 4×4, $128 \to 3$, stride: 2, pad: 1)
Tanh(.)
Discriminator
Input: $x \in \mathbb{R}^{3 \times 32 \times 32}$
conv. (kernel: 4×4, $1 \to 64$, stride: 2, pad: 1)
LeakyReLU (negative slope: 0.2)
conv. (kernel: 4×4, $64 \to 128$, stride: 2, pad: 1)
Batch Normalization
LeakyReLU (negative slope: 0.2)
conv. (kernel: 4×4, $128 \to 256$, stride: 2, pad: 1)
Batch Normalization
LeakyReLU (negative slope: 0.2)
Linear $128 \times 4 \times 4 \times 4 \to 1$

Table 11: DCGAN architecture for our experiments on CIFAR10.

References

Boushra Abbas and Hedy Attouch. Dynamical systems and forward-backward algorithms associated with the sum of a convex subdifferential and a monotone cocoercive operator. *Optimization*, 64(10):2223–2252, 2015.

Felipe Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000.

Felipe Alvarez and Hedy Attouch. An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. *Set-Valued Analysis*, 9(1-2):3–11, 2001.

Maicon M Alves and Raul T Marcavillaca. On inexact relative-error hybrid proximal extragradient, forward-backward and Tseng's modified forward-backward methods with inertial effects. *Set-Valued and Variational Analysis*, 28:301–325, 2020.

Maicon M Alves, Jonathan Eckstein, Marina Geremia, and Jefferson G Melo. Relative-error inertial-relaxed inexact versions of Douglas-Rachford and ADMM splitting algorithms. *Computational Optimization and Applications*, 75(2):389–422, 2020.

Anatoly S Antipin. Minimization of convex functions on convex sets by means of differential equations. *Differential Equations*, 30(9):1365–1375, 1994.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML 2017)*, volume 70, pages 214–223. PMLR, 2017.

Hedy Attouch and Alexandre Cabot. Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions. *Applied Mathematics & Optimization*, 80 (3):547–598, 2019a.

Hedy Attouch and Alexandre Cabot. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Mathematical Programming*, pages 1–45, 2019b.

Hedy Attouch and Alexandre Cabot. Convergence rate of a relaxed inertial proximal algorithm for convex minimization. *Optimization*, pages 1–32, 2019c.

Hedy Attouch and Paul-Emile Maingé. Asymptotic behavior of second-order dissipative evolution equations combining potential with non-potential effects. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(3):836–857, 2011.

Shane Barratt and Rishi Sharma. A note on the inception score. In *ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2017.

Jérôme Bolte. Continuous gradient projection method in Hilbert spaces. *Journal of Optimization Theory and its Applications*, 119(2):235–259, 2003.

Jonathan M Borwein and Adrian S Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, New York, 2006.

Radu I Boţ and Ernö R Csetnek. Second order forward-backward dynamical systems for monotone inclusion problems. *SIAM Journal on Control and Optimization*, 54(3):1423–1443, 2016a.

Radu I Boţ and Ernö R Csetnek. An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems. *Numerical Algorithms*, 71(3): 519–540, 2016b.

Radu I Boţ and Ernö R Csetnek. Convergence rates for forward-backward dynamical systems associated with strongly monotone inclusions. *Journal of Mathematical Analysis and Applications*, 457(2):1135–1152, 2018.

Radu I Boţ, Ernö R Csetnek, and Christopher Hendrich. Inertial Douglas-Rachford splitting for monotone inclusion problems. *Applied Mathematics and Computation*, 256(1):472–487, 2015.

Radu I Boţ, Ernö R Csetnek, and Phan T Vuong. The Forward-Backward-Forward method from discrete and continuous perspective for pseudo-monotone variational inequalities in Hilbert spaces. *European Journal of Operational Research*, 287:49–60, 2020.

Richard W Cottle and Jacques A Ferland. On pseudo-convex functions of nonnegative variables. *Mathematical Programming*, 1(1):95–101, 1971.

Shisheng Cui, Uday V Shanbhag, Mathias Staudigl, and Phan T Vuong. Stochastic Relaxed Inertial Forward-Backward-Forward splitting for monotone inclusions in Hilbert spaces. *arXiv:2107.10335*, 2021.

Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR 2019)*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

Nicolas Hadjisavvas, Siegfried Schaible, and Ngai-Ching Wong. Pseudomonotone operators: a survey of the theory and its applications. *Journal of Optimization Theory and Applications*, 152(1):1–20, 2012.

Erfan Y Hamedani and Necdet S Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637, 2017.

Franck Iutzeler and Julien M Hendrickx. A generic online acceleration scheme for optimization algorithms via relaxation and inertia. *Optimization Methods and Software*, 34(2): 383–405, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Master's thesis, University of Toronto, Canada, 2009.

Yura Malitsky and Matthew K Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Doklady Akademija Nauk USSR*, 269:543–547, 1983.

Zdzisław Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bulletin of the American Mathematical Society*, 73(4):591–597, 1967.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR 2016)*, 2016.

Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.