universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation / Title of the Doctoral Thesis

"Convergence Rate Analysis of Optimisation and Minimax Algorithms for Machine Learning"

verfasst von / submitted by

Michael Sedlmayer, BSc MSc

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 796 605 405
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Mathematik
Betreut von / Supervisor:	Univ.-Prof. Dr. Radu Ioan Boţ
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Tara Lee Andrews

Acknowledgements

I would like to take the opportunity to thank all the people who contributed, both directly and indirectly, to this thesis. First and foremost I want to express my deep gratitude to my supervisor Professor Radu Ioan Boţ for his exceptional supervision and being a paragon of academic sincerity and perseverance, constantly striving for excellence. Next, I would like to thank Professor Tara Lee Andrews, who provided me with an informative insight into the rich field of Digital Humanities. I am also grateful to Professors Pontus Giselsson and Defeng Sun for agreeing to review this PhD thesis.

My research was financially supported by the University of Vienna, Research Network Data Science and the Austrian Research Promotion Agency (FFG), project "Smart operation of wind turbines under icing conditions (SOWINDIC)". Additionally, I want to thank the Vienna School of Mathematics and the Vienna Graduate School on Computational Optimization for all the generous resources and opportunities provided.

Furthermore, I am happy that I have been part of Professor Boţ's research group and gladly mention my colleagues Axel, Buris, David, Dennis, Khoa, Matúš, Mikhail, Robert, Sorin and Vuong. Thank you for a positive, friendly and productive atmosphere. I would like to extend my thankfulness also to my other colleagues from Oskar-Morgenstern-Platz and Kolingasse. Thank you for all the lunch breaks that so often transformed into deep conversations about the most diverse topics. I very much appreciate your professional and personal opinions that frequently give impulse for fresh thoughts.

Especially, I want to thank you, Axel, for starting as a student mentor, being an outstanding colleague and becoming a friend.

Last, but not least, I want to thank my family and friends for their constant support and all the precious diversions – without you only a fraction would have been possible. Tobias, thank you for making working from home much more enjoyable, and thank you, Victoria, for helping me swim.

Abstract

Optimisation and saddle point problems, where the minimisation is accompanied by an inner maximisation, are well-known to be successfully treated by methods from monotone inclusion and variational inequality theory. In this context we are particularly interested in first order methods that are full splitting. This means that only gradient information for smooth functions and proximal evaluations of simple convex nonsmooth functions are used. Furthermore we require that the algorithms do not include convoluted inner loops but evaluate the involved operators separately.

We do this on three different conceptual levels: (i) monotone inclusions that are most general and include (ii) variational inequalities that already contain more structure and include (iii) minimax/saddle point problems that arise from game theory or in the context of determining primal-dual pairs of optimal solutions of constrained convex optimisation problems. We propose full splitting, first order solution methods, establish asymptotic convergence of the algorithm and prove convergence rates in the case of variational inequalities and minimax problems.

To empirically validate all considered methods we provide simple, conceptual problems that showcase the convergence behaviour of the proposed methods. This is complemented by more complex experiments covering relevant real-world machine learning applications (for example in Digital Humanities), treating, among other things, Generative Adversarial Nets (GANs) and Multikernel Support Vector Machines. GANs have proven to be a powerful class of generative models that are notoriously hard to optimise by conventional training methods, but can be cast as a minimax problem and observably benefit from employing more principled algorithms established for this type of problem.

Zusammenfassung

Optimierungs- und Sattelpunktprobleme, bei denen die Minimierung von einer inneren Maximierung begleitet wird, werden bekannterweise erfolgreich durch Methoden aus der Theorie zu Monotonen Inklusionen und Variationsungleichungen gelöst. In diesem Kontext sind wir besonders an Verfahren erster Ordnung interessiert, die „full splitting“ sind. Das heißt, dass ausschließlich Gradienteninformationen von glatten Funktionen und proximale Auswertungen von einfachen konvexen nichtglatten Funktionen verwendet werden. Außerdem verlangen wir, dass die Algorithmen keine verschachtelten inneren Schleifen beinhalten, sondern alle involvierten Operatoren separat ausgewertet werden.

Damit beschäftigen wir uns auf drei unterschiedlichen konzeptionellen Ebenen: (i) Monotonen Inklusionen, die am allgemeinsten sind und (ii) Variationsungleichungen beinhalten, die bereits mehr Struktur aufweisen und (iii) Minimax-/Sattelpunktprobleme inkludieren, die sich in der Spieltheorie oder im Kontext der Bestimmung von primal-dualen Paaren von optimalen Lösungen zu eingeschränkten konvexen Optimierungsproblemen ergeben. Wir stellen „full splitting“ Lösungsverfahren erster Ordnung vor, zeigen asymptotische Konvergenz der Algorithmen und beweisen nicht-asymptotische Konvergenzraten im Fall von Variationsungleichungen und Minimax-Problemen.

Um alle betrachteten Methoden empirisch zu validieren, stellen wir einfache, konzeptuelle Probleme vor, die das Konvergenzverhalten der vorgestellten Verfahren anschaulich machen. Das wird von komplexeren Experimenten komplimentiert, die relevante reale Anwendungen (beispielsweise in den Digitalen Geisteswissenschaften – „Digital Humanities“) des Maschinellen Lernens umfassen, wo wir uns unter anderem mit „Generative Adversarial Nets“ (GANs) und „Multikernel Support Vector Machines“ beschäftigen. GANs haben sich als eine leistungsstarke Klasse generativer Modelle erwiesen, die mit herkömmlichen Trainingsmethoden bekanntermaßen schwer zu optimieren sind. Sie können jedoch als Minimax-Problem betrachtet werden und profitieren deutlich von der Anwendung fundierter Algorithmen, die für diese Art von Problemen entwickelt wurden.

Contents

Gib s' in ehrliche Händ'.

(D.R.)

1. Introduction

Optimisation and saddle point problems, where the minimisation is accompanied by an inner maximisation, are well-known to be successfully treated by methods from monotone inclusion and variational inequality theory. In this context we are particularly interested in first order methods that are full splitting. This means that only gradient information for smooth functions and proximal evaluations of simple convex nonsmooth functions are used. Furthermore we require that the algorithms do not include convoluted inner loops but evaluate the involved operators separately.

In the setting of minimising the sum of a smooth convex and a nonsmooth convex function, the problem translated to monotone inclusions corresponds to finding a zero of the sum of a monotone cocoercive and a maximal monotone operator. For this type of problem the most basic and most established solution method is the Forward-Backward algorithm [39,91,120], attracting constant interest for extensions and improvements [66, 131]. In particular we want to mention the idea of incorporating past iterates, for example by inertial effects [9,132]. If one of the two functions to be minimised is composed with a linear operator more principled methods are necessary. One family of algorithms consists of so-called *primal-dual* methods, which simultaneously solve the primal and the dual optimisation problem. Preferably, this is achieved by splitting schemes [25, 33, 44, 50, 75, 139]. One of the most prominent members of this class is given by the *alternating direction method of multipliers* (ADMM) [37, 48, 49, 61, 62, 88], where the Lagrangian function is augmented by an additional penalty term and partial updates on the dual variables are used.

The Lagrangian formulation of a convex constrained minimisation problem itself gives rise to another problem class of its own interest – so-called *minimax* or *saddle point* problems. Treating the equivalent inclusion problem by the FB method, we can observe that this approach does not converge in general. This behaviour stems from the fact that the involved single-valued operator is merely Lipschitz continuous and not necessarily cocoercive, which can be seen even in the simple case of treating a bilinear minimax objective. Well-known algorithms to solve this type of problem are Tseng's *Forward-Backward-Forward* (FBF) algorithm [137] and the *Extragradient* (EG) method [6, 85]. However, both EG and FBF require two forward evaluations per iteration to converge to a solution, leading to a considerable increase of computational cost. A possible remedy is to reuse previous evaluations of the involved operator. Doing so, one can reduce the number of necessary forward evaluations to only one – similar to FB. The algorithm that one receives by doing this for EG and FBF is a method established by *Popov* [123] and the *Forward-Reflected-Backward* (FRB) algorithm [100], respectively. In the unconstrained case both reduce to the same method which is usually known as *Optimistic Gradient*

Descent Ascent (OGDA), a name coined by the machine learning community [54, 55]. Other methods known to converge for Lipschitz continuous operators are the projected *Reflected Gradient* (RG) algorithm [98], which can be seen as a variant of FRB where the order of the reflection and the forward step is reversed, a shadow Douglas-Rachford algorithm [52], which is related to FRB as well but applies the correction term regarding the forward evaluation after the backward step, and the *Golden Ratio Algorithm* [99].

In order to increase the quality of the convergence behaviour of the gradient method, Polyak introduced the heavy ball method in his pioneering work [122]. The idea of adding inertial effects was employed and refined in various settings – most prominently in the context of solving smooth convex minimisation problems by Nesterov [113], but also for monotone inclusions [2] and nonsmooth convex minimisation problems [3]. Furthermore, it has been shown that applying inertial effects in the case of nonconvex minimisation problems allows the detection of different critical points [30, 122]. When applied to convex minimisation problems, the acceleration techniques introduced by Nesterov [113] or those defined via asymptotically similar constructions [12,14,43] lead to an improved convergence rate for the sequence of objective function values. When applied to monotone inclusions, in absence of function values, the same lead to improved convergence rates for the sequences of discrete velocities and Yosida regularisations of the governing operator [9, 10, 11]. If the monotone inclusion is governed by a variational inequality or saddle point problem, the most common acceleration technique is related to Halpern's algorithm [74]. This *anchoring* technique has been applied for example to extend EG [40,140] or RG [42], and used to propose further anchoring based algorithms [35, 135, 136]. Recently, some connections between this approach and Nesterov's acceleration [113, 115] have been found [119, 135].

Saddle point problems arise traditionally in game theory [110] or for example when determining primal-dual pairs of optimal solutions of convex optimisation problems [17]. However, in recent years they have witnessed increased interest due to many relevant and challenging applications in the field of machine learning, with the most prominent ones being multi agent reinforcement learning [116], robust adversarial learning [96] and the training of Generative Adversarial Networks (GANs) [70]. Even though the problems in reality are often not of this form, in the classical setting the minimax objective comprises a smooth convex-concave coupling function with Lipschitz continuous gradient and a (potentially nonsmooth) regulariser in each variable, leading to a convex-concave objective in total. Typically, nonsmoothness is only introduced via regularisers, as usually the coupling function is accessed through gradient evaluations. Recently there is another development, although with the saddle point problem *not* being convex-concave, where the assumption on differentiability of the coupling function in both components is weakened to only one component [23,92,125]. Problems with nonsmooth but convex-concave coupling functions, however, have not been studied so far.

In this thesis we will study the extension of existing methods or problems, provide rigorous convergence analyses for the proposed algorithms and prove convergence rates whenever the type of problem admits it.

1.1. Overview

In **Chapter 2** we will introduce notation and relevant preliminary and auxiliary results that are used in the subsequent chapters. We will recall basic elements from monotone operator theory and convex analysis, and state convergence theorems of Opial type in the discrete and continuous setting which will be particularly important to establish the convergence of the sequence of generated iterates to a solution.

This is followed by an extensive convergence analysis of (variants of) Tseng's Forward-Backward-Forward (FBF) method in **Chapter 3**, which is based on the articles [20,36]; see also [19, Chapter 5]. We will start with considerations about approaching the set of zeros of the sum of a maximal monotone operator and a single-valued monotone and Lipschitz continuous operator. On the one hand this will be done from a continuous-time perspective, where we formulate a second order dynamical system that approaches the solution set of the monotone inclusion problem to be solved. On the other hand we investigate a relaxed inertial FBF splitting algorithm – called *RIFBF* – that follows by explicit time discretisation. This is succeeded by convergence rate analysis of the "plain" FBF method applied to general variational inequalities and convex-concave minimax problems, both in the deterministic and the stochastic setting. We derive convergence rates of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$ in terms of the (restricted) gap function for the ergodic iterates in the deterministic and stochastic case, respectively. The convergence behaviour of the proposed methods is illustrated by applications to a bilinear saddle point problem. We close this chapter with empirical improvements in the training of Wasserstein GANs on the CIFAR-10 dataset.

Chapter 4 is dedicated to the study of a fast optimistic method for solving variational inequalities that can be obtained from constrained minimax problems, which uses only one gradient/operator evaluation and one projection onto the constraint set per iteration. Our proposed algorithm achieves a convergence rate of $o(1/k)$ in terms of the restricted gap function as well as the natural residual for the *last iterate*. To complete the theoretical results we provide a convergence guarantee for the (last) iterate to a solution of the variational inequality. To validate our method, which we call *fOGDA-VI*, we investigate a two-player matrix game with mixed strategies of the two players. Concluding, we apply fOGDA-VI to the training of generative adversarial nets.

In **Chapter 5** we aim to solve a convex-concave saddle point problem, where the convex-concave coupling function is smooth in one variable and nonsmooth in the other. We propose and investigate a novel algorithm under the name of *OGAProx*, consisting of an optimistic gradient ascent step in the smooth variable and a proximal step in the nonsmooth component of the coupling function. We consider the situations convex-concave, convex-strongly concave and strongly convex-strongly concave related to the saddle point problem under investigation. Regarding iterates we obtain (weak) convergence, a convergence rate of order $\mathcal{O}(1/K)$ and linear convergence like $\mathcal{O}(\theta^K)$ with $\theta < 1$, respectively. In terms of function values we obtain ergodic convergence rates of order $\mathcal{O}(1/K)$, $\mathcal{O}(1/K^2)$ and $\mathcal{O}(\theta^K)$ with $\theta < 1$, respectively. We validate our theoretical considerations on a

nonsmooth-linear saddle point problem, the training of multi kernel support vector machines and a classification problem incorporating minimax group fairness. This chapter is based on the article [32].

2. Preliminaries

For a real Hilbert space $\mathcal{H}$ we denote its inner product by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ with induced norm $\| \cdot \|_{\mathcal{H}}$, for which $\|x\|_{\mathcal{H}}^2 = \langle x, x \rangle_{\mathcal{H}}$ for all $x \in \mathcal{H}$. As in the following it will always be clear from the context which Hilbert space is meant, we will not mention the space explicitly when writing the inner product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$. Furthermore, we denote *strong* convergence or *convergence in norm* by the symbol $\to$, while we write $\rightharpoonup$ for *weak* convergence. The identity operator will be denoted by Id and we define $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x \geq 0\}$ and $\mathbb{R}_{++} := \{x \in \mathbb{R} \mid x > 0\}$.

For the remainder of this chapter let $\mathcal{H}$ and $\mathcal{X}$ be real Hilbert spaces.

2.1. Monotone operator theory

In the following chapters we will work with set-valued operators. An operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ maps elements of $\mathcal{H}$ to subsets of $\mathcal{H}$, i.e., for all $z \in \mathcal{H}$ we have $Az \subseteq \mathcal{H}$. We will denote its graph by

$$\operatorname{gra} A := \{(z, v) \in \mathcal{H} \times \mathcal{H} \mid v \in Az\}.$$

The *inverse* operator $A^{-1} : \mathcal{H} \to 2^{\mathcal{H}}$ is given by

$$A^{-1}v := \{z \in \mathcal{H} \mid v \in Az\},$$

and we write $\operatorname{zer}(A) := \{z \in \mathcal{H} \mid 0 \in Az\}$ for the set of zeros of the operator A. Moreover, we define the multiplication of an operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ with a scalar $\lambda \in \mathbb{R}$ and the sum of two operators $A, B : \mathcal{H} \to 2^{\mathcal{H}}$ by

$$\begin{aligned} (\lambda A)z &:= \{\lambda v \mid v \in Az\}, \\ (A + B)\, z &:= \{v + u \mid v \in Az,\, u \in Bz\}. \end{aligned}$$

Definition 2.1.1. A set-valued operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ is *monotone* if

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

It is maximal monotone if there exists no monotone operator $\tilde{A} : \mathcal{H} \to 2^{\mathcal{H}}$ such that $\operatorname{gra} A \subsetneq \operatorname{gra} \tilde{A}$.

The following definition gives a generalisation of monotone operators.

Definition 2.1.2. A set-valued operator $A : \mathcal{H} \to 2^{\mathcal{H}}$ is pseudo-monotone (on $\mathcal{H}$) if for all $(z, v),\, (w, u) \in \operatorname{gra} A$ the following implication holds true,

$$\langle v, w - z \rangle \geq 0 \;\Rightarrow\; \langle u, w - z \rangle \geq 0. \tag{2.1}$$

Let $C \subseteq \mathcal{H}$ be nonempty closed convex, then A is pseudo-monotone *on* C if (2.1) holds for all (z, v), $(w, u) \in \{(x, y) \in \operatorname{gra} A \mid x \in C\}$.

It is clear that every monotone operator is pseudo-monotone. However, an important difference between monotone and pseudo-monotone operators is that the sum of two of the latter are not necessarily pseudo-monotone any more.

The following result about maximal monotone operators will be crucial in the convergence analysis of the algorithms to be investigated.

Proposition 2.1.3 (see [17, Proposition 20.33]). *Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be maximal monotone. Then the following hold:*

(i) $\operatorname{gra} A$ *is sequentially closed in the strong-weak topology of $\mathcal{H} \times \mathcal{H}$, i.e., for every sequence $(z_k, v_k)_{k\geq 0}$ in $\operatorname{gra} A$ and every $(z, v) \in \mathcal{H} \times \mathcal{H}$, if $z_k \to z$ and $v_k \rightharpoonup v$, then $(z, v) \in \operatorname{gra} A$.*

(ii) $\operatorname{gra} A$ *is sequentially closed in the weak-strong topology of $\mathcal{H} \times \mathcal{H}$, i.e., for every sequence $(z_k, v_k)_{k\geq 0}$ in $\operatorname{gra} A$ and every $(z, v) \in \mathcal{H} \times \mathcal{H}$, if $z_k \rightharpoonup z$ and $v_k \to v$, then $(z, v) \in \operatorname{gra} A$.*

Definition 2.1.4. Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be a set-valued operator. The *resolvent* of A is

$$J_A := (\mathrm{Id} + A)^{-1}.$$

Definition 2.1.5. Let $B : \mathcal{H} \to \mathcal{H}$ be a single-valued operator. It is said to be

(i) *Lipschitz continuous* with constant $L \geq 0$, if

$$\|Bw - Bz\| \leq L \|w - z\| \quad \forall w, z \in \mathcal{H};$$

(ii) *nonexpansive*, if

$$\|Bw - Bz\| \leq \|w - z\| \quad \forall w, z \in \mathcal{H};$$

(iii) *cocoercive* with constant $\beta > 0$, if

$$\beta \|Bw - Bz\|^2 \leq \langle Bw - Bz, w - z\rangle \quad \forall w, z \in \mathcal{H};$$

(iv) *firmly nonexpansive*, if

$$\|Bw - Bz\|^2 \leq \langle Bw - Bz, w - z\rangle \quad \forall w, z \in \mathcal{H};$$

Remark 2.1.6. (i) Nonexpansiveness is the same as 1-Lipschitz continuity and firm nonexpansiveness is the same as 1-cocoercivity.

(ii) Let $B : \mathcal{H} \to \mathcal{H}$ be β-cocoercive. Then B is $1/\beta$-Lipschitz continuous by the Cauchy-Schwarz inequality and monotone due to the norm being nonnegative. The converse, however, is not true in general.

For the proof of the following statement please refer to [17, Proposition 23.10].

Proposition 2.1.7 (Minty Lemma). *Let $A : \mathcal{H} \to 2^{\mathcal{H}}$ be maximal monotone and $\gamma \in \mathbb{R}_{++}$. Then $J_{\gamma A} : \mathcal{H} \to \mathcal{H}$ is firmly nonexpansive and maximal monotone.*

2.2. Convex analysis

In this section we will discuss properties of convex functions taking values in the extended real line $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. Addition and multiplication are extended in a natural way from $\mathbb{R}$ to $\overline{\mathbb{R}}$, where we impose the following conventions:

$$\begin{aligned}(+\infty) + (-\infty) = (-\infty) + (+\infty) &= +\infty,\\ 0(+\infty) = (+\infty)0 &= +\infty,\\ 0(-\infty) = (-\infty)0 &= 0.\end{aligned}$$

Definition 2.2.1. A set $C \subseteq \mathcal{X}$ is said to be *convex*, if $\lambda x + (1-\lambda)y \in C$ for all x, $y \in C$ and all $\lambda \in [0,1]$.

Definition 2.2.2. Let $f : \mathcal{X} \to \overline{\mathbb{R}}$ be an extended-real valued function. Its *domain* is given by $\operatorname{dom} f = \{x \in \mathcal{X} \mid f(x) < +\infty\}$.

The function f is said to be

(i) *proper*, if $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in \mathcal{X}$;

(ii) *convex*, if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$ for all x, $y \in \mathcal{X}$ and all $\lambda \in [0,1]$;

(iii) *μ-strongly convex* ($\mu > 0$), if $f - \mu/2 \,\|\cdot\|^2$ is convex;

(iv) *(strongly) concave*, if $-f$ is (strongly) convex;

(v) *lower semicontinuous*, if $\liminf_{y \to x} \geq f(x)$ for all $x \in \mathcal{X}$;

(vi) *upper semicontinuous*, if $-f$ is lower semicontinuous.

By induction we can prove Jensen's inequality.

Lemma 2.2.3. *Let $f : \mathcal{X} \to \overline{\mathbb{R}}$ be convex. Let $K \geq 1$, $x_1, ..., x_K \in \mathcal{X}$ and $\lambda_1, ..., \lambda_K \in \mathbb{R}_+$ with $\sum_{k=1}^K \lambda_k = 1$. Then*

$$f\left(\sum_{k=1}^K \lambda_k x_k\right) \leq \sum_{k=1}^K \lambda_k f(x_k).$$

Definition 2.2.4. Let $f : \mathcal{X} \to \overline{\mathbb{R}}$ be a proper function. The *convex subdifferential* is defined as

$$\partial f(x) = \begin{cases} \{u \in \mathcal{X} \mid f(x) + \langle u, y - x\rangle \leq f(y) \quad \forall y \in \mathcal{X}\} & \text{if } f(x) \in \mathbb{R} \\ \emptyset & \text{otherwise.} \end{cases}$$

With the notion of convex subdifferential, we can extend the necessary and sufficient first order optimality condition.

Lemma 2.2.5 (Fermat's rule)**.** *Let $f : \mathcal{X} \to \overline{\mathbb{R}}$ be a proper function. Then*

$$x^* \in \arg\min_{x \in \mathcal{X}} f(x) \Leftrightarrow 0 \in \partial f(x^*).$$

The proof of the following proposition can be derived from [17, Proposition 16.8], the definition of the convex subdifferential and [17, Corollary 16.38 (iii)].

Proposition 2.2.6. *(i) Let $\mathcal{X}_1, ..., \mathcal{X}_m$, $m \geq 1$, be real Hilbert spaces and let $f_i : \mathcal{X}_i \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous for $1 \leq i \leq m$. Define $f : \mathcal{X}_1 \times ... \times \mathcal{X}_m \to \overline{\mathbb{R}}$ via $f(x_1, ..., x_m) = \sum_{i=1}^m f_i(x_i)$. Then for all $(x_1, ..., x_m) \in \mathcal{X}_1 \times ... \times \mathcal{X}_m$ we have*

$$\partial f(x_1, ..., x_m) = \partial f(x_1) \times ... \times \partial f(x_m).$$

(ii) Let $f, g : \mathcal{X} \to \overline{\mathbb{R}}$ be proper. Then $\partial f(x) + \partial g(x) \subseteq \partial(f + g)(x)$ for all $x \in \mathcal{X}$.

(iii) Let $f, g : \mathcal{X} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous with $\operatorname{dom} f = \mathcal{X}$. Then $\partial f(x) + \partial g(x) = \partial(f + g)(x)$ for all $x \in \mathcal{X}$.

The following prominent example connects some of the concepts we have encountered so far.

Example 2.2.7. *Let $C \subseteq \mathcal{X}$ be nonempty, closed and convex. Then the* indicator function *is defined by*

$$\delta_C := \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

Then δ_C is proper, convex and lower semicontinuous and

$$\partial \delta_C(x) = N_C(x) := \begin{cases} \{u \in \mathcal{X} \mid \langle u, y - x \rangle \leq 0 \quad \forall y \in \mathcal{X}\} & \text{if } x \in C \\ \emptyset & \text{otherwise,} \end{cases}$$

where $N_C : \mathcal{X} \to 2^{\mathcal{X}}$ denotes the normal cone *of C, which is a maximal monotone operator if C is nonempty, closed and convex.*

Definition 2.2.8. Let $f : \mathcal{X} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. Then we define the *proximal mapping* of f by

$$\operatorname{prox}_f : \mathcal{X} \to \mathcal{X}, \quad \operatorname{prox}_f(x) = \arg\min_{y \in \mathcal{X}} \left\{ f(y) + \frac{1}{2} \|y - x\|^2 \right\},$$

where $\operatorname{prox}_f(x)$ is well-defined as it is the unique solution of the above strongly convex optimisation problem.

The proof for the following statements can be found in [130] and [17, Proposition 16.34].

Proposition 2.2.9. *Let $f : \mathcal{X} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous. Then the following hold:*

(i) the subdifferential $\partial f : \mathcal{X} \to 2^{\mathcal{X}}$ is maximal monotone;

(ii) the resolvent of the subdifferential is given by $J_{\partial f} = \operatorname{prox}_f$;

(iii) for $x \in \mathcal{X}$ we have $p = \operatorname{prox}_f(x)$ if and only if $x - p \in \partial f(p)$, i.e., for all $y \in \mathcal{X}$ we have

$$f(y) \geq f(p) + \langle x - p, y - p \rangle .$$

Continuing Example 2.2.7 we see, that the proximal operator can be understood as a generalisation of the projection onto a set.

Example 2.2.10. *Let $C \subseteq \mathcal{X}$ be nonempty, closed and convex. Then from Proposition 2.2.9 (ii) we obtain for all $x \in \mathcal{X}$*

$$J_{N_C}(x) = \operatorname{prox}_{\partial \delta_C}(x) = \operatorname*{arg\,min}_{y \in \mathcal{X}} \left\{ \frac{1}{2} \|y - x\|^2 \right\} = P_C.$$

In particular, from Proposition 2.2.9 (iii) we deduct the well-known characterisation of the projection,

$$p = P_C(x) \Leftrightarrow [p \in C \text{ and } \langle y - p, x - p \rangle \leq 0 \quad \forall y \in C] .$$

In Remark 2.1.6 (ii) we mentioned that cocoercivity implies Lipschitz continuity, whereas the converse is not true in general. For the gradients of convex and differentiable functions, however, the two properties are equivalent. The proof for the following statement can be found for examle in [17, Corollary 18.16].

Proposition 2.2.11 (Baillon-Haddad)**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be convex and Fréchet differentiable and let $\beta \in \mathbb{R}_{++}$. Then ∇f is β-Lipschitz continuous if and only if ∇f is $1/\beta$-cocoercive.*

2.3. Convergence theorems of Opial type

To show weak convergence of sequences in Hilbert spaces we will use the following so-called *Opial Lemma*.

Lemma 2.3.1 (discrete Opial Lemma [117])**.** *Let $C \subseteq \mathcal{X}$ be a nonempty set and $(x_k)_{k \geq 0}$ a sequence in $\mathcal{X}$ such that the following two conditions hold:*

(a) for every $x^* \in C$, $\lim_{k \to +\infty} \|x_k - x^*\|$ exists;

(b) every weak sequential cluster point of $(x_k)_{k \geq 0}$ belongs to C.

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

The continuous version of the Opial Lemma reads as follows.

Lemma 2.3.2 (continuous Opial Lemma [27, Lemma 7])**.** *Let $C \subseteq \mathcal{X}$ be a nonempty set and $x : [0, +\infty) \to \mathcal{X}$ a given map. Assume that*

(a) for every $x^* \in C$, $\lim_{k\to+\infty} \|x(t) - x^*\|$ *exists;*

(b) every weak sequential cluster point of the map x *belongs to* C.

Then $x_\infty \in C$ *such that* $x(t)$ *converges weakly to* x_∞ *as* $t \to +\infty$.

2.4. Stochastics

We want to recall basic notions from measure and probability theory which can be found in any introductory book on this topic. To this end we partly follow [19, Section 2.4].

Definition 2.4.1. Let Ω be a set and let $\mathcal{A}$ be a σ-algebra on Ω. Then the tuple $(\Omega, \mathcal{A})$ is called a *measurable space*.

Definition 2.4.2. Let $(\Omega, \mathcal{A})$ be a measurable space. If it is additionally equipped with a probability measure $\mathbb{P}$, then the triple $(\Omega, \mathcal{A}, \mathbb{P})$ is called a *probability space*.

We call a measurable mapping from a probability space to a measurable space *random variable*. Usually when talking about random variables we will omit the spaces and sometimes even the probability measure, i.e., when talking about the expectation $\mathbb{E}[X]$ of the random variable X.

Definition 2.4.3 (Expectation)**.** For a random variable $X : \Omega :\to \mathbb{R}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ the *expected value* is defined as $\mathbb{E}[X] := \int_\Omega X(\omega)\, \mathrm{d}\mathbb{P}(\omega)$.

Definition 2.4.4. For a random variable X on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with finite expectation, its *conditional expectation* with respect to a sub σ-algebra $\mathcal{S}$ of $\mathcal{A}$ denoted by $\mathbb{E}[X \,|\, \mathcal{S}]$, is the $\mathcal{S}$-measurable random variable fulfilling

$$\int_A X \,\mathrm{d}\mathbb{P} = \int_A \mathbb{E}[X \,|\, \mathcal{S}] \,\mathrm{d}\mathbb{P} \quad \forall A \in \mathcal{S}.$$

Despite the abstract nature of this definition it is a standard problem in measure theory to show existence and uniqueness of such a random variable (in an almost sure sense). We will use the notation $\mathbb{E}[X \,|\, Y]$ for two random variables X and Y, where the conditional expectation of X is meant to be taken with respect to the σ-algebra generated by Y.

The following three substantial lemmas can be found for example in [134].

Lemma 2.4.5. *Let the assumptions of Definition 2.4.4 hold.*

(i) If X *is measurable with respect to* $\mathcal{S}$*, then* $\mathbb{E}[X \,|\, \mathcal{S}] = X$.

(ii) If X *is independent of* $\mathcal{S}$*, then* $\mathbb{E}[X \,|\, \mathcal{S}] = \mathbb{E}[X]$.

The two properties above are combined into the following statement.

Lemma 2.4.6 (Independence lemma). *Let f be a function of two arguments such that*

$$g(x) := \mathbb{E}[f(x, Y)].$$

If X and Y are two independent random variables, then

$$g(X) = \mathbb{E}[f(X, Y) \,|\, X].$$

In the stochastic setting of Section 3.4 we will deal with the case where the objective function $F : \mathcal{H} \to \mathbb{R}$ for all $z \in \mathcal{H}$ is given as $\mathbb{E}_\xi[F(z;\xi)]$ for a random variable ξ (with a slight abuse of notation). We write $\mathbb{E}_\xi$ to emphasize that ξ is stochastic and not z, but will drop the subscript later on. Typically one assumes the gradients of $F(z;\xi)$ with respect to z to be unbiased estimates for the gradient of F, i.e., that for all $z \in \mathcal{H}$ we have

$$\mathbb{E}_\xi[\nabla F(z;\xi)] = \nabla F(z).$$

When analysing the proposed algorithms in Section 3.4, due to use of stochastic gradients, the iterates $(z_k)_{k\geq 0}$ will turn to be stochastic themselves. Thus, by Lemma 2.4.6 we get

$$\mathbb{E}[\nabla F(z_k;\xi) \,|\, z_k] = \nabla F(z_k),$$

as long as ξ is independent of the iterate z_k.

Lemma 2.4.7 (Tower property). *Let $\mathcal{A}$ and $\mathcal{B}$ be two sigma algebras such that $\mathcal{A} \subseteq \mathcal{B}$. Then we have*

$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{B}] \,|\, \mathcal{A}] = \mathbb{E}[X \,|\, \mathcal{A}].$$

In particular the above lemma implies the *law of total expectation*, stating that for any sigma algebra $\mathcal{A}$ we have

$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{A}]] = \mathbb{E}[X].$$

2.5. Miscellaneous

In this section we gather various helpful auxiliary results. We start with statements regarding the existence of sequences, followed by a result related to quasi-Fejér monotone sequences

Lemma 2.5.1 (see [3]). *Let $(\varphi_k)_{k\geq 0}$, $(\alpha_k)_{k\geq 1}$, and $(\psi_k)_{k\geq 1}$ be sequences of nonnegative real numbers satisfying*

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

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 2.5.2 (see [31, Lemma 21]). *Let $a \geq 1$ and $(q_k)_{k\geq 1}$ be a bounded sequence in $\mathcal{H}$ such that*

$$\lim_{k\to+\infty} \left(q_{k+1} + \frac{k}{a}\left(q_{k+1} - q_k\right)\right) = p \in \mathcal{H}.$$

Then $\lim_{k\to+\infty} q_k = p$.

The following result is a particular instance of [18, Lemma 5.31].

Lemma 2.5.3. *Let $(a_k)_{k\geq 1}$, $(b_k)_{k\geq 1}$, $(d_k)_{k\geq 1}$ and $(d_k)_{k\geq 1}$ be sequences of real numbers. Assume that $(a_k)_{k\geq 1}$ is bounded from below, $(b_k)_{k\geq 1}$ and $(d_k)_{k\geq 1}$ are nonnegative sequences such that $\sum_{k\geq 1} d_k < +\infty$. If*

$$a_{k+1} \leq (1 + d_k)\, a_k - b_k \quad \forall k \geq 1, \tag{2.2}$$

then the following statements are true:

(i) the sequence $(b_k)_{k\geq 1}$ is summable, i.e., $\sum_{k\geq 1} b_k < +\infty$;

(ii) the sequence $(a_k)_{k\geq 1}$ is convergent.

The following two lemmas collect some elementary results that are used several times in the following, which will close this section.

Lemma 2.5.4. *Let $a, b, c, d \in \mathcal{X}$. Then the following hold:*

(i) "four point identity":

$$\langle a - b, c - d\rangle = \frac{1}{2}\left(\|a - d\|^2 - \|b - d\|^2 - \|a - c\|^2 \|b - c\|^2\right);$$

(ii) "Young's inequality":

$$\langle a, b\rangle \leq \frac{1}{2}\left(\|a\|^2 + \|b\|^2\right);$$

(iii) $\|a + b + c\|^2 \leq 3\left(\|a\|^2 + \|b\|^2 + \|c\|^2\right)$.

Lemma 2.5.5. *Let $a, b, c \in \mathbb{R}$ be such that $a \neq 0$ and $b^2 - ac \leq 0$. The following statements are true:*

(i) if $a > 0$, then

$$a\,\|x\|^2 + 2b\,\langle x, y\rangle + c\,\|y\|^2 \geq 0 \quad \forall x, y \in \mathcal{H};$$

(ii) if $a < 0$, then

$$a\,\|x\|^2 + 2b\,\langle x, y\rangle + c\,\|y\|^2 \leq 0 \quad \forall x, y \in \mathcal{H}.$$

3. FBF algorithm: extensions and convergence rates

In this chapter 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 endeavour aims to extend Tseng's forward-backward-forward (FBF) method by both using inertial effects and 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 discretisation, a convergence analysis in the general monotone case as well as when applied to the solving of pseudo-monotone variational inequalities. This is followed by convergence analysis of the "plain" FBF method applied to minimax problems and general variational inequalities. The derived error bounds are in terms of the (restricted) gap function for the ergodic iterates, both for the deterministic and the stochastic setting. Furthermore, we propose a seemingly new scheme which reuses previous gradients to mitigate the additional computational cost. In doing so we rediscover a known method, related to *Optimistic Gradient Descent Ascent (OGDA)*. For the deterministic and the stochastic problem we show a convergence rate of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$, respectively. We illustrate the proposed methods and their convergence behaviour by applications to a bilinear saddle point problem, in the context of which we also emphasise the interplay between the inertial and the relaxation parameters for RIFBF and showcase the obtained convergence rates. Our theoretical results are complemented with empirical improvements in the training of Wasserstein GANs on the CIFAR-10 dataset.

This chapter is based on the articles [20,36]; see also [19, Chapter 5].

3.1. Introduction

3.1.1. Motivation

One of the main motivations 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 constrained convex optimisation problems means actually solving convex-concave minimax problems [17], nevertheless, minimax problems are of own interest. Minimax problems arise traditionally in game theory [110] and, more recently, in robust adversarial learning [96] and the training of Generative Adversarial Networks

(GANs) [69, 70], as we will see in Section 3.5.

The general convex-concave minimax, or saddle point, problem is of the form

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Psi(x,y) := f(x) + \Phi(x,y) - g(y), \tag{MM}$$

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

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

The regularisers f and g are assumed to be proper, convex and lower semicontinuous. A solution of (MM) is given by a so-called saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$, fulfilling

$$\Psi(x^*, y) \leq \Psi(x^*, y^*) \leq \Psi(x, y^*) \quad \forall x \in \mathcal{X},\, \forall y \in \mathcal{Y},$$

or, equivalently, the system of optimality conditions

$$\begin{aligned} 0 &\in \partial[\Psi(\,\cdot\,, y^*)](x^*) = \partial f(x^*) + \nabla_x \Phi(x^*, y^*), \\ 0 &\in \partial[-\Psi(x^*, \,\cdot\,)](y^*) = \partial g(y^*) - \nabla_y \Phi(x^*, y^*), \end{aligned} \tag{3.1}$$

where $\partial f : \mathcal{X} \to 2^{\mathcal{X}}$ and $\partial g : \mathcal{Y} \to 2^{\mathcal{Y}}$ denote the convex subdifferential of the functions f and g, respectively. Defining

$$r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}, \quad (x,y) \mapsto f(x) + g(y),$$

and

$$F : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}, \quad (x,y) \mapsto \begin{pmatrix} \nabla_x \Phi(x,y) \\ -\nabla_y \Phi(x,y) \end{pmatrix},$$

we see that (3.1) is equivalent to

$$0 \in \partial r(x^*, y^*) + F(x^*, y^*).$$

Note that the assumptions stated above imply that r is proper, convex and lower semicontinuous, while F is monotone and L-Lipschitz continuous.

This motivates us also to investigate the following more general problem. Let $\mathcal{H}$ be a real Hilbert space, let $r : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex and lower semicontinuous function, and let $F : \mathcal{H} \to \mathcal{H}$ be a monotone and L-Lipschitz continuous operator ($L > 0$). The goal is to find $z^* \in \mathcal{H}$ such that

$$0 \in \partial r(z^*) + F(z^*). \tag{MI-VI}$$

Using the definition of the convex subdifferential, we see that (MI-VI) is equivalent to

$$r(z^*) + \langle F(z^*), z^* - z\rangle \leq r(z) \quad \forall z \in \mathcal{H}. \tag{VI-s}$$

This type of problem is well-established and is called *Variational Inequality (VI)*. The formulation (VI-s) is known as *strong* form of the VI, while its associated *weak* formulation reads as follows: find $z^* \in \mathcal{H}$ such that

$$r(z^*) + \langle F(z), z^* - z\rangle \leq r(z) \quad \forall z \in \mathcal{H}. \tag{VI-w}$$

Obviously, the notions of strong and weak formulation are strongly related and in the setting stated above even equivalent.

Lemma 3.1.1. *Let $r : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper, convex and lower semicontinuous function, and let $F : \mathcal{H} \to \mathcal{H}$ be a monotone operator. Then the following holds:*

(i) If $z^* \in \mathcal{H}$ is a solution of (VI-s)*, then it is also a solution of* (VI-w)*;*

(ii) If in addition F is continuous, then $z^* \in \mathcal{H}$ is a solution of (VI-s) *if and only if it is a solution of* (VI-w)*.*

Proof. (i) Let $z^* \in \mathcal{H}$ be a solution for the strong formulation (VI-s). Then by monotonicity of F we have that for all $z \in \mathcal{H}$

$$r(z) \geq r(z^*) + \langle F(z^*), z^* - z\rangle \geq r(z^*) + \langle F(z), z^* - z\rangle,$$

which is nothing else than z^* being a solution for the weak formulation (VI-w).

(ii) Let $z^* \in \mathcal{H}$ be a solution for the weak formulation (VI-w) and $z = \beta z^* + (1-\beta)w$ for an arbitrary but fixed $w \in \mathcal{H}$ and $\beta \in (0,1)$. Then

$$\langle F(\beta z^* + (1-\beta)w), (1-\beta)(w - z^*)\rangle + r(\beta z^* + (1-\beta)w) - r(z^*) \geq 0,$$

and convexity of r implies

$$(1-\beta)\langle F(\beta z^* + (1-\beta)w), w - z^*\rangle + (1-\beta)(r(w) - r(z^*)) \geq 0.$$

Dividing the above inequality by $(1-\beta)$ and taking the limit $\beta \to 1$ then gives

$$\langle F(z^*), w - z^*\rangle + r(w) - r(z^*) \geq 0.$$

As $w \in \mathcal{H}$ was arbitrary, we get that z^* is a solution for the strong formulation (VI-s).

□

Problem (VI-s) and (VI-w) is known as *general* strong and weak VI, respectively. For a specific choice of the function r we obtain a prestigious problem instance of particular interest.

Example 3.1.2. *Let $C \subseteq \mathcal{H}$ be a nonempty closed convex set. Then for the choice $r = \delta_C$ in* (VI-s) *and* (VI-w) *we obtain the* classical *strong and weak VI, respectively, where the task is to find $z^* \in C$ such that*

$$\langle F(z^*), z - z^* \rangle \geq 0 \quad \forall z \in C, \tag{3.2}$$

and

$$\langle F(z), z - z^* \rangle \geq 0 \quad \forall z \in C,$$

respectively.

A further generalisation of (MI-VI), and thus variational inequalities, is the so-called *Monotone Inclusion (MI)*, where we want to find $z^* \in \mathcal{H}$ such that

$$0 \in Az^* + Bz^*, \tag{MI}$$

with $A : \mathcal{H} \to 2^{\mathcal{H}}$ being a set-valued maximal monotone operator and $B : \mathcal{H} \to \mathcal{H}$ being a single-valued monotone and L-Lipschitz continuous operator ($L > 0$). Indeed this is a generalisation, as the subdifferential of a proper, convex and lower semicontinuous function is well-known to be maximal monotone.

Monotone inclusions governed by maximal monotone operators are often seen as generalisations of systems of optimality conditions for convex optimisation problems, however, (MI) shows that they can also be of own interest. In order to solve (MI) one cannot rely on solution methods developed for optimisation 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 chapter and in this way we will provide further evidence for the importance of monotone operators beyond the convex optimisation setting.

Remark 3.1.3. Formulation (MM) is indeed sufficiently general. For example if we look at the following saddle point problem where the regularisers are composed with linear operators,

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Psi(x, y) := f(Ax) + \Phi(x, y) - g(By), \tag{3.3}$$

where $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{H}$, $\mathcal{G}$ are real Hilbert spaces, $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, $g : \mathcal{G} \to \mathbb{R} \cup \{+\infty\}$ are proper, convex, lower semicontinuous functions, $A : \mathcal{X} \to \mathcal{H}$, $B : \mathcal{Y} \to \mathcal{G}$ are linear continuous mappings and the coupling function $\Phi : \mathcal{H} \times \mathcal{G} \to \mathbb{R}$ is convex-concave and Fréchet differentiable with Lipschitz continuous gradient. Then finding a saddle point of (3.3) amounts to solving

$$\begin{aligned} 0 &\in \partial (f \circ A)(x^*) + \nabla_x [\Phi(\cdot, y^*)](x^*), \\ 0 &\in \partial (g \circ B)(y^*) - \nabla_y [\Phi(x^*, \cdot)](y^*). \end{aligned}$$

Under certain regularity conditions (see [17, Section 16.4]) this is equivalent to

$$\begin{aligned} 0 &\in A^* \partial f(Ax^*) + \nabla_x \Phi(x^*, y^*), \\ 0 &\in B^* \partial g(By^*) - \nabla_y \Phi(x^*, y^*). \end{aligned}$$

Employing [17, Theorem 16.23] we see that the above holds if and only if there exist $(u^*, v^*) \in \mathcal{H} \times \mathcal{G}$ such that

$$\begin{aligned} 0 &\in \partial f^*(u^*) - Ax^* \\ 0 &= A^*u^* + \nabla_x \Phi(x^*, y^*) \\ 0 &\in \partial g^*(v^*) - By^* \\ 0 &= B^*v^* - \nabla_y \Phi(x^*, y^*), \end{aligned}$$

where f^* and g^* denote the Fenchel conjugate of f and g, respectively. Using the definition of the subdifferential and the convex-concave structure of the coupling function Φ we infer that this is fulfilled if and only if for all $(x, y, u, v) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \times \mathcal{G}$

$$\begin{aligned} &-f^*(u) + \langle u, Ax^* \rangle + \Phi(x^*, y) + g^*(v^*) - \langle v^*, By \rangle \\ &\quad \leq -f^*(u^*) + \langle u^*, Ax^* \rangle + \Phi(x^*, y^*) + g^*(v^*) - \langle v^*, By^* \rangle \\ &\quad \leq -f^*(u^*) + \langle u^*, Ax \rangle + \Phi(x, y^*) + g^*(v) - \langle v, By^* \rangle . \end{aligned}$$

Writing

$$\tilde{\Psi}((x, v), (y, u)) := -f^*(u) + \langle u, Ax \rangle + \Phi(x, y) + g^*(v) - \langle v, By \rangle ,$$

we see that the above chain of inequalities holds if and only if for all $(x, u) \in \mathcal{X} \times \mathcal{H}$ and all $(y, v) \in \mathcal{Y} \times \mathcal{G}$

$$\tilde{\Psi}((x^*, v^*), (y, u)) \leq \tilde{\Psi}((x^*, v^*), (y^*, u^*)) \leq \tilde{\Psi}((x, v), (y^*, u^*)),$$

which is the characterisation of a solution for the saddle point problem

$$\min_{(x,v) \in \mathcal{X} \times \mathcal{G}} \max_{(y,u) \in \mathcal{Y} \times \mathcal{H}} \tilde{f}(x, v) + \tilde{\Phi}((x, v), (y, u)) - \tilde{g}(y, u),$$

where

$$\tilde{f} : (x, v) \mapsto g^*(v) \quad \text{and} \quad \tilde{g} : (y, u) \mapsto f^*(u)$$

are proper, convex and lower semicontinuous functions, and

$$\tilde{\Phi} : ((x, v), (y, u)) \mapsto \langle u, Ax \rangle + \Phi(x, y) - \langle v, By \rangle$$

is convex-concave and differentiable with Lipschitz continuous gradient – matching the setting of (MM).

3.1.2. Related literature

Solution methods for solving (MI), when the operator B is cocoercive, have been developed intensively in the last decades [1, 9, 17, 27, 28]. The simplest method for solving (MI), when B is cocoercive, is the forward-backward method, which for a given starting point $z_0 \in \mathcal{H}$ iterates for all $k \geq 0$

$$z_{k+1} = J_{\gamma A} \left(\mathrm{Id} - \gamma B \right) z_k, \tag{3.4}$$

where γ is a positive step size chosen in $(0, 2/L)$ and J_A denotes the resolvent of A. 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 (see Remark 2.1.6 and Proposition 2.2.11). The iterative scheme (3.4) results from time-discretisation of the following first order dynamical system with time step size chosen equal to 1,

$$\begin{cases} \dot{z}(t) + z(t) = J_{\gamma A}\left(\mathrm{Id} - \gamma B\right) z(t), \\ z(0) = z_0. \end{cases}$$

For more about dynamical systems of implicit type associated to monotone inclusions and convex optimisation problems, see for example [1, 7, 21, 28].

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

$$\begin{cases} \ddot{z}(t) + \sigma(t)\dot{z}(t) + \tau(t)\left[z(t) - J_{\gamma A}\left(\mathrm{Id} - \gamma B\right) z(t)\right] = 0, \\ z(0) = z_0, \quad \dot{z}(0) = \dot{z}_0, \end{cases}$$

where $\sigma, \tau : [0, +\infty) \to [0, +\infty)$, was proposed and studied in [27] (see also [2, 7, 28]). Explicit time discretisation 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 relaxed inertial algorithms for monotone inclusions and convex optimisation problems in a series of papers, as they combine the advantages of both inertial effects and relaxation techniques. More precisely, they addressed the relaxed inertial proximal method [10, 11] and the relaxed inertial forward-backward method [9]. A relaxed inertial Douglas-Rachford algorithm for monotone inclusions has been proposed by [29]. [81] investigated the influence inertial effects and relaxation techniques have on the numerical performances of optimisation algorithms. The interplay between relaxation and inertial parameters for relative-error inexact under-relaxed algorithms has been addressed by [4,5].

Relaxation techniques are essential ingredients in the formulation of algorithms for monotone inclusions, as they provide more flexibility to the iterative schemes [17, 59]. 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 [122], who introduced the heavy ball method in order to speed up the convergence behaviour of the gradient algorithm and allow the detection of different critical points. This idea was employed and refined in various settings – most prominently in the context of solving smooth convex minimisation problems by Nesterov [113], but also for monotone inclusions [2] and nonsmooth convex minimisation problems [3]. When applied to convex minimisation problems, the acceleration techniques introduced by Nesterov [113] 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 [9, 10, 11] for the relaxed inertial proximal and forward-backward methods, to improved convergence rates for the sequences of discrete velocities and Yosida regularisations of the governing operator.

In the following we will focus on solving the monotone inclusion (MI) 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 discretisation of a second order dynamical system approaching the solution set of (MI). The forward-backward-forward (FBF) method was proposed by Tseng [137], iterating for all $k \geq 0$

$$\text{FBF:} \left| \begin{array}{l} w_k = J_{\gamma_k A}(\mathrm{Id} - \gamma_k B)z_k, \\ z_{k+1} = y_k - \gamma_k(By_k - Bz_k), \end{array} \right.$$

where $z_0 \in H$ is the starting point. The sequence $(z_k)_{k\geq 0}$ converges weakly to a solution of (MI) if the sequence of step sizes $(\gamma_k)_{k\geq 0}$ is chosen in the interval $(0, 1/L)$, where $L > 0$ is the Lipschitz constant of B. An inertial version of FBF, however in the absence of relaxation variables, was proposed by [26]. The convergence of the sequence of generated iterates has been proved in a very restrictive 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 [2,3]. 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 behaviour.

Recently, a forward-backward algorithm for solving (MI), when B is monotone and Lipschitz continuous, was proposed by [100]. This method requires in every iteration only one forward step instead of two, however, the sequence of step sizes has to be chosen constant in the interval $(0, 1/2L)$, which slows the algorithm down in comparison to FBF. A popular algorithm used to solve the variational inequality (MI-VI), when B is monotone and Lipschitz continuous, is Korpelevich's extragradient method [85]. Note that this algorithm can also be applied to the more general monotone inclusion problem (MI) with the step sizes to be chosen in the interval $(0, 1/L)$, however, this method requires two backward steps as well as two forward steps.

3.1.3. Outline

First, we will approach the solution set of (MI) from a continuous perspective by means of the trajectories generated by a second order dynamical system of FBF type in Section 3.2. 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 discretisation of the dynamical system gives rise to an algorithm of forward-backward-forward type with inertial and relaxation parameters which we call RIFBF.

In Section 3.3 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 [100] 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 step size 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 variational inequality problem (3.2) in the case when the operator involved is pseudo-monotone but not necessarily monotone. Pseudo-monotone operators for example appear in the consumer theory of mathematical economics [73] and as gradients of pseudo-convex functions [51], such as ratios of convex and concave functions in fractional programming [22]. To conclude this section we treat a bilinear saddle point problem which can be understood as a two-player zero-sum constrained game. In this context, we emphasise the interplay between the inertial and the relaxation parameters.

In Section 3.4 we deal with statements about convergence rates for the FBF method. We also propose a variant of FBF reusing previous gradients to reduce the computational cost per iteration, which turns out to be a known method, related to [100]. By developing a unifying scheme that captures both FBF and this method, we reveal a hitherto unknown connection. Using this approach we prove novel convergence rates in terms of the minimax gap for both methods in the context of saddle point problems. In the deterministic and stochastic setting we obtain rates of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$, respectively. Again, we treat bilinear saddle point problems to illustrate the obtained convergence rates.

Concluding, in Section 3.5 we employ variants of FBF for training Generative Adversarial Networks (GANs) which constitute a powerful 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 optimise. Using principled methods with convergence guarantees, even though only in the monotone setting, we show empirical improvements in the training of Wasserstein GANs on the CIFAR-10 dataset.

3.2. A second order dynamical system of FBF type

In this and the following section we aim to solve the monotone inclusion problem (MI). Recall, we are interested in finding $z^* \in \mathcal{H}$, with $\mathcal{H}$ being a real Hilbert space, such that

$$0 \in Az^* + Bz^*, \tag{MI}$$

where $A : \mathcal{H} \to 2^{\mathcal{H}}$ is a maximally monotone set-valued operator, and $B : \mathcal{H} \to \mathcal{H}$ is a monotone and Lipschitz continuous operator with Lipschitz constant $L > 0$, such that $\operatorname{zer}(A+B) := \{z \in \mathcal{H} : 0 \in Az + Bz\} \neq \emptyset$.

In this section we will for this reason focus on the study of the following dynamical system in connection with the monotone inclusion problem (MI),

$$\begin{cases} w(t) = J_{\gamma A}(\mathrm{Id} - \gamma B)z(t), \\ \ddot{z}(t) + \sigma(t)\dot{z}(t) + \tau(t)\left[z(t) - w(t) - \gamma\left(Bz(t) - Bw(t)\right)\right] = 0, \\ z(0) = z_0, \quad \dot{z}(0) = \dot{z}_0, \end{cases} \tag{3.5}$$

where $\sigma, \tau : [0, +\infty) \to [0, +\infty)$ are Lebesgue measurable functions, $0 < \gamma < 1/L$ and $z_0, \dot{z}_0 \in \mathcal{H}$.

We define $M : \mathcal{H} \to \mathcal{H}$ by

$$Mz = z - J_{\gamma A}(\mathrm{Id} - \gamma B)z - \gamma \left[Bz - B \circ J_{\gamma A}(\mathrm{Id} - \gamma B)z\right]. \tag{3.6}$$

Then (3.5) can be equivalently written as

$$\begin{cases} \ddot{z}(t) + \sigma(t)\dot{z}(t) + \tau(t)Mz(t) = 0, \\ z(0) = z_0, \quad \dot{z}(0) = \dot{z}_0. \end{cases} \tag{3.7}$$

The following result collects some properties of M.

Proposition 3.2.1. *Let M be defined as in (3.6). Then the following statements are true:*

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

(ii) M is Lipschitz continuous;

(iii) for all $z^* \in \mathrm{zer}(M)$ and all $z \in \mathcal{H}$ the following holds:

$$\langle Mz, z - z^* \rangle \geq \frac{1 - \gamma L}{(1 + \gamma L)^2} \|Mz\|^2.$$

Proof. (i) For $z \in \mathcal{H}$ we set $w := J_{\gamma A}(\mathrm{Id} - \gamma B)z$, thus $Mz = z - w - \gamma(Bz - Bw)$. Using the Lipschitz continuity of B we have

$$(1 - \gamma L)\|z - w\| \leq \|Mz\| = \|z - w - \gamma(Bz - Bw)\| \leq (1 + \gamma L)\|z - w\|. \tag{3.8}$$

Therefore, $z \in \mathrm{zer}(M)$ if and only if $z = w = J_{\gamma A}(\mathrm{Id} - \gamma B)z$, which is further equivalent to $z \in \mathrm{zer}(A + B)$.

(ii) Let $z, z' \in \mathcal{H}$ and $w := J_{\gamma A}(\mathrm{Id} - \gamma B)z$, and $w' := J_{\gamma A}(\mathrm{Id} - \gamma B)z'$. The Lipschitz continuity of B yields

$$\begin{aligned} \|Mz - Mz'\| &= \|z - w - \gamma(Bz - Bw) - z' + w' + \gamma(Bz' - Bw')\| \\ &\leq (1 + \gamma L)\left(\|z - z'\| + \|w - w'\|\right). \end{aligned}$$

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

$$\begin{aligned} \|w - w'\| &= \|J_{\gamma A}(\mathrm{Id} - \gamma B)z - J_{\gamma A}(\mathrm{Id} - \gamma B)z'\| \\ &\leq \|(\mathrm{Id} - \gamma B)z - (\mathrm{Id} - \gamma B)z'\| \leq (1 + \gamma L)\|z - z'\|. \end{aligned}$$

Therefore,

$$\|Mz - Mz'\| \leq (1 + \gamma L)(2 + \gamma L)\|z - z'\|,$$

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

(iii) Let $z^* \in \mathcal{H}$ be such that $0 \in (A+B)z^*$ and $z \in \mathcal{H}$. We denote $w := J_{\gamma A}(\mathrm{Id} - \gamma B)z$ and can write $(\mathrm{Id} - \gamma B)z \in (\mathrm{Id} + \gamma A)w$ or, equivalently,

$$\frac{1}{\gamma}(z-w) - (Bz - Bw) \in (A+B)w. \tag{3.9}$$

Using the monotonicity of $A+B$ we obtain

$$\left\langle \frac{1}{\gamma}(z-w) - (Bz-Bw), w - z^* \right\rangle \geq 0,$$

which is equivalent to

$$\langle z - w - \gamma(Bz - Bw), z - z^* \rangle \geq \langle z - w - \gamma(Bz - Bw), z - w \rangle.$$

This means that

$$\begin{aligned}\langle Mz, z - z^* \rangle &\geq \langle z - w - \gamma(Bz - Bw), z - w \rangle = \|z-w\|^2 - \gamma \langle Bz - Bw, z - w \rangle \\ &\geq \|z-w\|^2 - \gamma \|Bz - Bw\| \|z - w\| \geq (1 - \gamma L)\|z-w\|^2 \\ &\geq \frac{1-\gamma L}{(1+\gamma L)^2}\|Mz\|^2,\end{aligned}$$

where the last inequality follows from (3.8).

□

The following definition makes explicit which kind of solutions of the dynamical system (3.5) we are looking for. We recall that a function $z : [0,b] \to \mathcal{H}$ (where $b > 0$) is said to be absolutely continuous if there exists an integrable function $w : [0,b] \to \mathcal{H}$ such that

$$z(t) = z(0) + \int_0^t w(s)\,\mathrm{d}s \quad \forall t \in [0,b].$$

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

Definition 3.2.2. We say that $z : [0,+\infty) \to \mathcal{H}$ is a strong global solution of (3.5) if the following properties are satisfied:

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

(ii) $\ddot{z}(t) + \sigma(t)\dot{z}(t) + \tau(t)Mz(t) = 0$ for almost every $t \in [0,+\infty)$;

(iii) $z(0) = z_0$ and $\dot{z}(0) = \dot{z}_0$.

Since $M : \mathcal{H} \to \mathcal{H}$ is Lipschitz continuous, the existence and uniqueness of the trajectory of (3.5) follows from the Cauchy-Lipschitz Theorem for absolutely continuous trajectories.

Theorem 3.2.3. *(see [27, Theorem 4]) Let $\sigma, \tau : [0, +\infty) \to [0, +\infty)$ be Lebesgue measurable functions such that $\sigma, \tau \in L^1_{loc}([0, +\infty))$ (that is $\sigma, \tau \in L^1([0, b])$ for all $0 < b < +\infty$). Then for each $z_0, \dot{z}_0 \in \mathcal{H}$ there exists a unique strong global solution of the dynamical system* (3.5).

We will prove the convergence of the trajectories of (3.5) in a setting which requires the damping function σ and the relaxation function τ to fulfil the assumptions below. We refer to [27] for examples of functions which fulfil this assumption and want also to emphasise that when the two functions are constant we recover the conditions from [13].

Assumption 3.2.4. *$\sigma, \tau : [0, +\infty) \to [0, +\infty)$ are locally absolutely continuous and there exists $\nu > 0$ such that for almost every $t \in [0, +\infty)$ the inequalities*

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

hold true.

The result which states the convergence of the trajectories is adapted from the results by [27, Theorem 8]. However, it cannot be obtained as a direct consequence of it, since the operator M is not cocoercive as it is required there. Nevertheless, as seen in Proposition 3.2.1 (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 [27].

Theorem 3.2.5. *Let $\sigma, \tau : [0, +\infty) \to [0, +\infty)$ be functions satisfying Assumption 3.2.4 and $z_0, \dot{z}_0 \in \mathcal{H}$. Let $z : [0, +\infty) \to \mathcal{H}$ be the unique strong global solution of* (3.5). *Then the following statements are true:*

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

(ii) $\lim_{t\to+\infty} \dot{z}(t) = \lim_{t\to+\infty} \ddot{z}(t) = \lim_{t\to+\infty} Mz(t) = \lim_{t\to+\infty} [z(t) - w(t)] = 0$;

(iii) $z(t)$ converges weakly to an element in $\operatorname{zer}(A + B)$ as $t \to +\infty$.

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

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

Taking into account (3.7) we obtain for almost every $t \in [0, +\infty)$ that

$$\ddot{h}(t) + \sigma(t)\dot{h}(t) + \tau(t)\langle z(t) - z^*, Mz(t)\rangle = \|\dot{z}(t)\|^2,$$

which, together with Proposition 3.2.1 (iii), implies

$$\ddot{h}(t) + \sigma(t)\dot{h}(t) + \frac{1-\gamma L}{(1+\gamma L)^2}\tau(t)\|Mz(t)\|^2 \leq \|\dot{z}(t)\|^2 .$$

From this point, we can proceed as in the proof of [27, Theorem 8] – we will show now that from the last inequality we obtain the statements in (i) and (ii) and that the limit $\lim_{t\to+\infty} \|z(t) - z^*\| \in \mathbb{R}$ exists, which is the first assumption in the continuous version of the Opial Lemma, see Lemma 2.3.2.

Let us define $\kappa := \frac{1-\gamma L}{(1+\gamma L)^2} > 0$. Taking again into account (3.7) one obtains for every $t \in [0, +\infty)$

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

or, equivalently,

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

Combining this inequality with

$$\frac{\sigma(t)}{\tau(t)} \frac{d}{dt}\left(\|\dot{z}(t)\|^2\right) = \frac{d}{dt}\left(\frac{\sigma(t)}{\tau(t)} \|\dot{z}(t)\|^2\right) - \frac{\dot{\sigma}(t)\tau(t) - \sigma(t)\dot{\tau}(t)}{\tau^2(t)} \|\dot{z}(t)\|^2$$

and

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

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

$$\begin{aligned}\ddot{h}(t) + \frac{d}{dt}(\sigma h)(t) + \kappa \frac{d}{dt}\left(\frac{\sigma(t)}{\tau(t)} \|\dot{z}(t)\|^2\right) \\ + \left(\frac{\kappa\sigma^2(t)}{\tau(t)} + \kappa\frac{-\dot{\sigma}(t)\tau(t) + \sigma(t)\dot{\tau}(t)}{\tau^2(t)} - 1\right) \|\dot{z}(t)\|^2 + \frac{\kappa}{\tau(t)} \|\ddot{z}(t)\|^2 \leq 0.\end{aligned}$$

According to Assumption 3.2.4 for almost every $t \in [0, +\infty)$ the following inequality holds,

$$\ddot{h}(t) + \frac{d}{dt}(\sigma h)(t) + \kappa \frac{d}{dt}\left(\frac{\sigma(t)}{\tau(t)} \|\dot{z}(t)\|^2\right) + \nu \|\dot{z}(t)\|^2 + \frac{\kappa}{\bar{\tau}} \|\ddot{z}(t)\|^2 \leq 0, \tag{3.10}$$

where $\bar{\tau}$ denotes a positive upper bound for the function τ. From here we obtain that the function $t \mapsto \dot{h}(t) + \sigma(t)h(t) + \kappa\frac{\sigma(t)}{\tau(t)} \|\dot{z}(t)\|^2$, which is locally absolutely continuous, is monotonically decreasing. Hence there exists a real number N_1 such that

$$\dot{h}(t) + \sigma(t)h(t) + \kappa\frac{\sigma(t)}{\tau(t)} \|\dot{z}(t)\|^2 \leq N_1 \quad \forall t \in [0, +\infty), \tag{3.11}$$

which yields

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

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

$$h(T) \leq h(0)\exp(-\underline{\sigma}T) + \frac{N_1}{\underline{\sigma}} \left(1 - \exp(-\underline{\sigma}T)\right),$$

thus

$$h \text{ is bounded}$$

and, consequently,

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

On the other hand, from (3.11), it follows that for every $t \in [0, +\infty)$

$$\dot{h}(t) + \frac{\kappa\underline{\sigma}}{\bar{\tau}} \|\dot{z}(t)\|^2 \leq N_1,$$

hence

$$\langle z(t) - z^*, \dot{z}(t)\rangle + \frac{\kappa\underline{\sigma}}{\bar{\tau}} \|\dot{z}(t)\|^2 \leq N_1.$$

This yields, since z is bounded, that

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

and further that

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

Integrating (3.10) we obtain that there exists a real number $N_2 \in \mathbb{R}$ such that for every $t \in [0, +\infty)$

$$\dot{h}(t) + \sigma(t)h(t) + \kappa\frac{\sigma(t)}{\tau(t)} \|\dot{z}(t)\|^2 + \nu \int_0^t \|\dot{z}(s)\|^2 \,\mathrm{d}s + \frac{\kappa}{\bar{\tau}} \int_0^t \|\ddot{z}(s)\|^2 \,\mathrm{d}s \leq N_2,$$

which allows us to conclude that $\dot{z}, \ddot{z} \in L^2([0, +\infty); H)$. Finally, from (3.7) and Assumption 3.2.4 we deduce $Mz \in L^2([0, +\infty); H)$ and the proof of statement (i) is complete.

In order to prove (ii), we notice that for every $t \in [0, +\infty)$ the following is true,

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

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

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

$$\begin{aligned}\frac{d}{dt}\left(\frac{1}{2}\|M(z(t))\|^2\right) &= \left\langle M(z(t)), \frac{d}{dt}(Mz(t))\right\rangle \\ &\leq \frac{1}{2}\|M(z(t))\|^2 + \frac{(1+\gamma L)^2(2+\gamma L)^2}{2}\|\dot{z}(t)\|^2.\end{aligned}$$

By using (i) again, we derive $\lim_{t\to+\infty} M(z(t)) = 0$, while $\lim_{t\to+\infty} \ddot{z}(t) = 0$ follows from (3.7) and Assumption 3.2.4.

Finally, we have seen that the function $t \mapsto \dot{h}(t) + \sigma(t)h(t) + \kappa\frac{\sigma(t)}{\tau(t)}\|\dot{z}(t)\|^2$ is monotonically decreasing, thus from (i), (ii) and Assumption 3.2.4 we deduce that there exists

$$\lim_{t\to+\infty} \sigma(t)h(t) \in \mathbb{R}.$$

By taking also into account that $\lim_{t\to+\infty}\sigma(t) \in (0,+\infty)$ exists, we finally obtain that there exists

$$\lim_{t\to+\infty}\|z(t)-z^*\|^2 \in \mathbb{R}.$$

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 z is a zero of M, one cannot use the arguments from [27, 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 $\bar{z}$ be a weak sequential cluster point of z, which means that there exists a sequence $t_k \to +\infty$ such that $(z(t_k))_{k\geq 0}$ converges weakly to $\bar{z}$ as $k\to+\infty$. Since, according to (ii), $\lim_{t\to+\infty} Mz(t) = \lim_{t\to+\infty}[z(t)-w(t)] = 0$, we conclude that $(w(t_k))_{k\geq 0}$ also converges weakly to $\bar{z}$. According to (3.9) we have

$$\frac{1}{\gamma}(z(t_k)-w(t_k)) - (Bz(t_k)-Bw(t_k)) \in (A+B)w(t_k) \quad \forall k\geq 0. \tag{3.12}$$

Since B is Lipschitz continuous and $\lim_{k\to+\infty}\|z(t_k)-w(t_k)\| = 0$, the left hand side of (3.12) 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 $\mathcal{H}\times\mathcal{H}$. Therefore, taking the limit as $t_k\to+\infty$ in (3.12) we obtain $\bar{z}\in \operatorname{zer}(A+B)$.

Thus, the continuous Opial Lemma implies that $z(t)$ converges weakly to an element in $\operatorname{zer}(A+B)$ as $t\to+\infty$. □

Remark 3.2.6 (explicit discretisation). Explicit time discretisation of (3.5) with step size $s_k>0$, relaxation variable $\tau_k>0$, damping variable $\sigma_k>0$, and initial points z_0 and z_1 yields for all $k\geq 1$ the following iterative scheme:

$$\frac{1}{s_k^2}(z_{k+1}-2z_k+z_{k-1}) + \frac{\sigma_k}{s_k}(z_k-z_{k-1}) + \tau_k M v_k = 0, \tag{3.13}$$

where v_k is an extrapolation of z_k and z_{k-1} that will be chosen later. The Lipschitz continuity of M provides a certain flexibility to this choice. We can write (3.13) equivalently as

$$(\forall k\geq 1)\quad z_{k+1} = z_k + (1-\sigma_k s_k)(z_k-z_{k-1}) - s_k^2\tau_k M v_k.$$

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

$$\left\lfloor \begin{array}{l} v_k = z_k + \alpha_k(z_k-z_{k-1}) \\ w_k = J_{\gamma A}(\mathrm{Id}-\gamma B)v_k \\ z_{k+1} = (1-\rho_k)v_k + \rho_k\left(w_k - \gamma(Bw_k - Bv_k)\right), \end{array}\right.$$

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

3.3. Asymptotic convergence of a relaxed inertial FBF algorithm

In this section we investigate the convergence of the relaxed inertial algorithm derived in the previous section via explicit time discretisation of the second order dynamical system (3.5). Recall, we are still interested in solving the monotone inclusion (MI), where we want to find $z^* \in \mathcal{H}$ such that

$$0 \in Az^* + Bz^*,$$

with $A : \mathcal{H} \to 2^{\mathcal{H}}$ being a set-valued maximal monotone operator and $B : \mathcal{H} \to \mathcal{H}$ being a single-valued monotone and Lipschitz continuous operator with Lipschitz constant $L > 0$.

To this end we will address the following algorithm iterating for all $k \geq 1$

$$\text{RIFBF:} \left| \begin{array}{l} v_k = z_k + \alpha_k(z_k - z_{k-1}) \\ w_k = J_{\gamma_k A}(\mathrm{Id} - \gamma_k B)v_k \\ z_{k+1} = (1-\rho_k)v_k + \rho_k\left(w_k - \gamma_k(Bw_k - Bv_k)\right), \end{array} \right.$$

where z_0, $z_1 \in \mathcal{H}$ are starting points, $(\gamma_k)_{k\geq 1}$, $(\rho_k)_{k\geq 1} \subseteq \mathbb{R}_{++}$ are sequences of positive numbers, and $(\alpha_k)_{k\geq 1} \subseteq \mathbb{R}_+$ 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

$$\text{IFBF:} \left| \begin{array}{ll} v_k = z_k + \alpha_k(z_k - z_{k-1}) & \\ w_k = J_{\gamma_k A}(\mathrm{Id} - \gamma_k B)v_k & \forall k \geq 1 \\ z_{k+1} = w_k - \gamma_k(Bw_k - Bv_k) & \end{array} \right.$$

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

$$\text{RFBF:} \left| \begin{array}{ll} w_k = J_{\gamma_k A}(\mathrm{Id} - \gamma_k B)z_k & \\ z_{k+1} = (1-\rho_k)z_k + \rho_k\left(w_k - \gamma_k(Bw_k - Bz_k)\right) & \forall k \geq 1 \end{array} \right.$$

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

$$\text{FBF:} \left| \begin{array}{ll} w_k = J_{\gamma_k A}(\mathrm{Id} - \gamma_k B)z_k & \\ z_{k+1} = w_k - \gamma_k(Bw_k - Bz_k). & \forall k \geq 1 \end{array} \right.$$

Step size rules: Depending on the availability of the Lipschitz constant L of B , we have two different options for the choice of the sequence of step sizes $(\gamma_k)_{k\geq 1}$:

- **constant step size**: $\gamma_k := \gamma \in (0, 1/L)$ for all $k \geq 1$;
- **adaptive step size**: let $\mu \in (0,1)$ and $\gamma_1 > 0$. The step sizes for $k \geq 1$ are adaptively updated as follows

$$\gamma_{k+1} := \begin{cases} \min\left\{\gamma_k, \frac{\mu\|w_k - v_k\|}{\|Bw_k - Bv_k\|}\right\}, & \text{if } Bw_k - Bv_k \neq 0, \\ \gamma_k, & \text{otherwise.} \end{cases} \tag{3.14}$$

If the Lipschitz constant L of B is known in advance, then a constant step size can be chosen. Otherwise, the adaptive step size rule (3.14) is highly recommended. In the following, we provide the convergence analysis for the adaptive step size rule, as the constant step size rule can be obtained as a particular of it by setting $\gamma_1 := \gamma$ and $\mu := \gamma L$.

Proposition 3.3.1. *Let $\mu \in (0,1)$ and $\gamma_1 > 0$. The sequence $(\gamma_k)_{k\geq 1}$ generated by (3.14) is nonincreasing and*

$$\lim_{k\to+\infty} \gamma_k = \gamma \geq \min\left\{\gamma_1, \frac{\mu}{L}\right\}.$$

In addition,

$$\|Bw_k - Bv_k\| \leq \frac{\mu}{\gamma_{k+1}} \|w_k - v_k\| \quad \forall k \geq 1. \tag{3.15}$$

Proof. It is obvious from (3.14) that $\gamma_{k+1} \leq \gamma_k$ for all $k \geq 1$. Since B is Lipschitz continuous with Lipschitz constant L we obtain

$$\frac{\mu \|w_k - v_k\|}{\|Bw_k - Bv_k\|} \geq \frac{\mu}{L}, \text{ if } Bw_k - Bv_k \neq 0,$$

which together with (3.14) yields

$$\gamma_{k+1} \geq \min\left\{\gamma_1, \frac{\mu}{L}\right\} \quad \forall k \geq 1.$$

Thus, there exists

$$\gamma := \lim_{k\to+\infty} \gamma_k \geq \min\left\{\gamma_1, \frac{\mu}{L}\right\}.$$

The inequality (3.15) follows directly from (3.14). □

Proposition 3.3.2. *Let $u_k := w_k - \gamma_k(Bw_k - Bv_k)$ and $\theta_k := \frac{\gamma_k}{\gamma_{k+1}}$ for all $k \geq 1$. Then for all $z^* \in \operatorname{zer}(A+B)$ the following holds:*

$$\|u_k - z^*\|^2 \leq \|v_k - z^*\|^2 - \left(1 - \mu^2\theta_k^2\right)\|w_k - v_k\|^2 \quad \forall k \geq 1. \tag{3.16}$$

Proof. Let $k \geq 1$. Using (3.15) we have that

$$\begin{aligned}
\|v_k - z^*\|^2 &= \|v_k - w_k + w_k - u_k + u_k - z^*\|^2 \\
&= \|v_k - w_k\|^2 + \|w_k - u_k\|^2 + \|u_k - z^*\|^2 \\
&\quad + 2\langle v_k - w_k, w_k - z^*\rangle + 2\langle w_k - u_k, u_k - z^*\rangle \\
&= \|v_k - w_k\|^2 - \|w_k - u_k\|^2 + \|u_k - z^*\|^2 + 2\langle v_k - u_k, w_k - z^*\rangle \\
&= \|v_k - w_k\|^2 - \gamma_k^2\|Bw_k - Bv_k\|^2 + \|u_k - z^*\|^2 + 2\langle v_k - u_k, w_k - z^*\rangle \\
&\geq \|v_k - w_k\|^2 - \frac{\mu^2\gamma_k^2}{\gamma_{k+1}^2}\|w_k - v_k\|^2 + \|u_k - z^*\|^2 + 2\langle v_k - u_k, w_k - z^*\rangle.
\end{aligned} \tag{3.17}$$

On the other hand, since

$$(\mathrm{Id} + \gamma_k A)w_k \ni (\mathrm{Id} - \gamma_k B)v_k,$$

we have

$$v_k \in w_k + \gamma_k A w_k + \gamma_k B v_k = w_k - \gamma_k (B w_k - B v_k) + \gamma_k (A+B) w_k = u_k + \gamma_k (A+B) w_k.$$

Therefore,

$$\frac{1}{\gamma_k}(v_k - u_k) \in (A+B) w_k,$$

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

$$\langle v_k - u_k, w_k - z^* \rangle \geq 0.$$

Hence,

$$\|v_k - z^*\|^2 \geq \|v_k - w_k\|^2 - \frac{\mu^2 \gamma_k^2}{\gamma_{k+1}^2} \|w_k - v_k\|^2 + \|u_k - z^*\|^2,$$

which is (3.16). □

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

Proposition 3.3.3. *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*

$$0 < \rho_k \leq \frac{2}{1+\mu\theta_k} \quad \forall k \geq k_\rho, \tag{3.18}$$

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

$$H_k := \|z_k - z^*\|^2 - \alpha_k \|z_{k-1} - z^*\|^2 + 2\alpha_k \left(\alpha_k + \frac{1-\alpha_k}{\rho_k(1+\mu\theta_k)}\right) \|z_k - z_{k-1}\|^2. \tag{3.19}$$

Then there exists $k_0 \geq k_\rho$ such that for all $k \geq k_0$ the inequality

$$H_{k+1} - H_k \leq -\delta_k \|z_{k+1} - z_k\|^2 \tag{3.20}$$

holds, 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 3.3.2, with $u_k = w_k - \gamma_k(Bw_k - Bv_k)$, we have for all $k \geq 1$

$$\begin{aligned}
\|z_{k+1} - z^*\|^2 &= \|(1-\rho_k)v_k + \rho_k u_k - z^*\|^2 \\
&= \|(1-\rho_k)(v_k - z^*) + \rho_k(u_k - z^*)\|^2 \\
&= (1-\rho_k)\|v_k - z^*\|^2 + \rho_k \|u_k - z^*\|^2 - \rho_k(1-\rho_k)\|u_k - v_k\|^2 \\
&\leq (1-\rho_k)\|v_k - z^*\|^2 + \rho_k \|v_k - z^*\|^2 \\
&\quad - \rho_k\left(1 - \mu^2\theta_k^2\right)\|w_k - v_k\|^2 - \frac{1-\rho_k}{\rho_k}\|z_{k+1} - v_k\|^2 \\
&= \|v_k - z^*\|^2 - \rho_k\left(1 - \mu^2\theta_k^2\right)\|w_k - v_k\|^2 - \frac{1-\rho_k}{\rho_k}\|z_{k+1} - v_k\|^2.
\end{aligned} \tag{3.21}$$

Using (3.15) we obtain for all $k \geq 1$

$$\frac{1}{\rho_k}\|z_{k+1}-v_k\| = \|u_k-v_k\| \leq \|u_k-w_k\| + \|w_k-v_k\| = \gamma_k\|Bw_k-Bv_k\| + \|w_k-v_k\|$$
$$\leq \left(1+\frac{\mu\gamma_k}{\gamma_{k+1}}\right)\|w_k-v_k\| = (1+\mu\theta_k)\|w_k-v_k\|.$$

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

$$\frac{1-\mu\theta_k}{\rho_k(1+\mu\theta_k)}\|z_{k+1}-v_k\|^2 \leq \rho_k\left(1-\mu^2\theta_k^2\right)\|w_k-v_k\|^2.$$

We obtain from (3.21) that for all $k \geq k_1$

$$\begin{aligned}\|z_{k+1}-z^*\|^2 &\leq \|v_k-z^*\|^2 - \left(\frac{1-\mu\theta_k}{\rho_k(1+\mu\theta_k)}+\frac{1-\rho_k}{\rho_k}\right)\|z_{k+1}-v_k\|^2 \\ &= \|v_k-z^*\|^2 - \left(\frac{2}{\rho_k(1+\mu\theta_k)}-1\right)\|z_{k+1}-v_k\|^2.\end{aligned} \tag{3.22}$$

Now we will estimate the right-hand side of (3.22). For all $k \geq 1$ we have

$$\begin{aligned}\|v_k-z^*\|^2 &= \|z_k+\alpha_k(z_k-z_{k-1})-z^*\|^2 \\ &= \|(1+\alpha_k)(z_k-z^*)-\alpha_k(z_{k-1}-z^*)\|^2 \\ &= (1+\alpha_k)\|z_k-z^*\|^2 - \alpha_k\|z_{k-1}-z^*\|^2 + \alpha_k(1+\alpha_k)\|z_k-z_{k-1}\|^2,\end{aligned} \tag{3.23}$$

and

$$\begin{aligned}\|z_{k+1}-v_k\|^2 &= \|(z_{k+1}-z_k)-\alpha_k(z_k-z_{k-1})\|^2 \\ &= \|z_{k+1}-z_k\|^2 + \alpha_k^2\|z_k-z_{k-1}\|^2 - 2\alpha_k\langle z_{k+1}-z_k, z_k-z_{k-1}\rangle \\ &\geq (1-\alpha_k)\|z_{k+1}-z_k\|^2 + \left(\alpha_k^2-\alpha_k\right)\|z_k-z_{k-1}\|^2.\end{aligned} \tag{3.24}$$

Combining (3.22) with (3.23) and (3.24), together with (3.18) and using that $(\alpha_k)_{k\geq 1}$ is nondecreasing, we obtain for all $k \geq k_0 := \max\{k_1, k_\rho\}$

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

which is nothing else than (3.20). □

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

$$0 < \rho < \frac{2}{(1+\mu)} \frac{(1-\alpha)^2}{(2\alpha^2 - \alpha + 1)}. \tag{3.25}$$

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 (3.18) is fulfilled naturally with the choice (3.25).

Remark 3.3.4 (inertia versus relaxation). Inequality (3.25) represents the necessary trade-off between inertia and relaxation (see Figure 3.1 for two particular choices of μ). The expression is similar to the one derived by [10, Remark 2.13], the exception being an additional factor incorporating the step size parameter μ. This means that for given $0 \leq \alpha < 1$ the upper bound for the relaxation parameter is

$$\bar{\rho}(\alpha, \mu) = \frac{2}{(1+\mu)} \frac{(1-\alpha)^2}{(2\alpha^2 - \alpha + 1)}.$$

We further see that $\alpha \mapsto \bar{\rho}(\alpha, \mu)$ 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 $\rho_{\max}(\mu) := \bar{\rho}(0, \mu) = \frac{2}{1+\mu}$. On the other hand, when $\alpha \nearrow 1$, then $\bar{\rho}(\alpha, \mu) \searrow 0$. 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 $\mu \nearrow 1$.

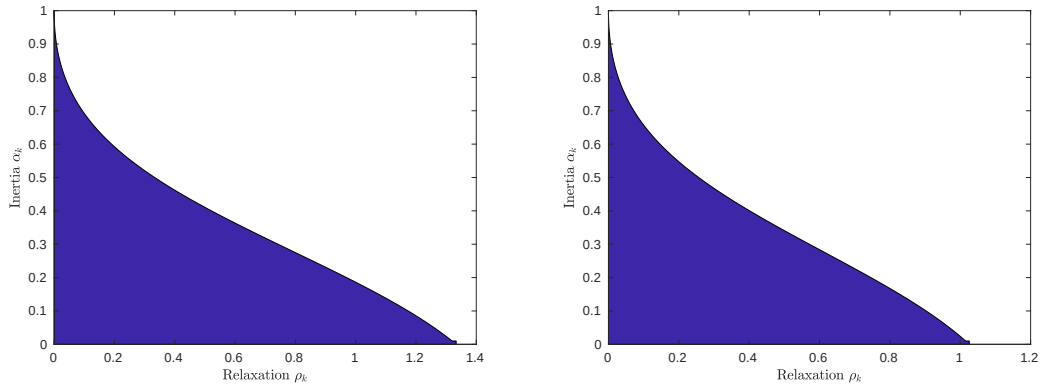


Figure 3.1.: Trade-off between inertia and relaxation for $\mu = 0.5$ (left) and $\mu = 0.95$ (right).

Proposition 3.3.5. *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 (3.18) holds, with the property that $\liminf_{k\to+\infty} \delta_k > 0$. For $z^* \in \operatorname{zer}(A+B)$ we define the sequence $(H_k)_{k\geq 1}$ as in (3.19). Then the following statements are true:*

(i) the sequence $(z_k)_{k\geq 0}$ is bounded;

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

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

Proof. (i) From (3.20) 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

$$H_{k_0} \geq H_k \geq \|z_k - z^*\|^2 - \alpha_k \|z_{k-1} - z^*\|^2 \geq \|z_k - z^*\|^2 - \alpha \|z_{k-1} - z^*\|^2$$

and, from here,

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

which implies that $(z_k)_{k\geq 0}$ 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 3.3.3.

□

We are now in the position to prove the main result of this section.

Theorem 3.3.6. *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 (3.18) holds, with the property that $\liminf_{k\to+\infty} \delta_k > 0$. Then the sequence $(z_k)_{k\geq 0}$ converges weakly to some element in $\operatorname{zer}(A+B)$ 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 2.3.1 for $C := \operatorname{zer}(A+B)$ are satisfied.

Let $z^* \in \operatorname{zer}(A+B)$. Indeed, it follows from (3.22) and (3.23) that for k large enough

$$\|z_{k+1} - z^*\|^2 \leq (1+\alpha_k)\|z_k - z^*\|^2 - \alpha_k\|z_{k-1} - z^*\|^2 + \alpha_k(1+\alpha_k)\|z_k - z_{k-1}\|^2.$$

Therefore, according to Lemma 2.5.1 and Proposition 3.3.5 (iii), $\lim_{k\to+\infty}\|z_k - z^*\|$ exists.

Let $\bar{z}$ be a weak limit point of $(z_k)_{k\geq 0}$ and a subsequence $(z_{k_l})_{l\geq 0}$ which converges weakly to $\bar{z}$ as $l \to +\infty$. From Proposition 3.3.5 (iii) we have

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

which, as $\liminf_{k\to+\infty}\delta_k > 0$, yields $\lim_{k\to+\infty}\|z_{k+1} - z_k\| = 0$. Recall that for $k \geq 1$ we have $u_k = w_k - \gamma_k(Bw_k - Bv_k)$. Since

$$\|u_k - v_k\| = \frac{1}{\rho_k}\|z_{k+1} - v_k\| = \frac{1}{\rho_k}\|z_{k+1} - z_k + \alpha_k(z_k - z_{k-1})\| \quad \forall k \geq 1,$$

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

$$\begin{aligned}\|u_k - v_k\| &= \|w_k - v_k - \gamma_k(Bw_k - Bv_k)\| \geq \|w_k - v_k\| - \gamma_k\|Bw_k - Bv_k\| \\ &\geq (1 - \mu\theta_k)\|w_k - v_k\|.\end{aligned}$$

Since $\lim_{k\to+\infty}(1-\mu\theta_k) = 1 - \mu > 0$, we can conclude that $\|w_k - v_k\| \to 0$ as $k \to +\infty$. Furthermore, for all $k \geq 1$ we have $v_k = z_k + \alpha_k(z_k - z_{k-1})$ and since $\|z_{k+1} - z_k\| \to 0$ as $k \to +\infty$ we deduce that $(v_{k_l})_{l\geq 0}$ converges weakly to $\bar{z}$ as $l \to +\infty$. Combining this with the fact that $\|w_k - v_k\| \to 0$ as $k \to +\infty$ then shows that $(w_{k_l})_{l\geq 0}$ and $(v_{k_l})_{l\geq 0}$ converges weakly to $\bar{z}$ as $l \to +\infty$, too. The definition of $(w_{k_l})_{l\geq 0}$ gives

$$\frac{1}{\gamma_{k_l}}(v_{k_l} - w_{k_l}) + Bw_{k_l} - Bv_{k_l} \in (A+B)w_{k_l} \quad \forall l \geq 0.$$

Using that $\left(\frac{1}{\gamma_{k_l}}(v_{k_l} - w_{k_l}) + Bw_{k_l} - Bv_{k_l}\right)_{l\geq 0}$ 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 $\mathcal{H} \times \mathcal{H}$, we obtain $0 \in (A+B)\bar{z}$, thus $\bar{z} \in \operatorname{zer}(A+B)$. □

Remark 3.3.7. In the particular case of the variational inequality (3.2), which corresponds to the case when A is the normal cone of a nonempty closed convex subset C of $\mathcal{H}$, we want to find $z^* \in C$ such that

$$\langle Bz^*, z - z^*\rangle \geq 0 \quad \forall z \in C.$$

Taking into account that $J_{\gamma N_C} = P_C$ is for all $\gamma > 0$ the projection operator onto C, the relaxed inertial FBF algorithm reads

$$\text{RIFBF-VI:} \left| \begin{array}{l} v_k = z_k + \alpha_k(z_k - z_{k-1}) \\ w_k = P_C(\mathrm{Id} - \gamma_k B)v_k \\ z_{k+1} = (1-\rho_k)v_k + \rho_k\left(w_k - \gamma_k(Bw_k - Bv_k)\right), \end{array} \right.$$

where z_0, $z_1 \in \mathcal{H}$ are starting points, $(\gamma_k)_{k\geq 1}$ and $(\rho_k)_{k\geq 1}$ are sequences of positive numbers, and $(\alpha_k)_{k\geq 1}$ is a sequence of nonnegative numbers. Then the algorithm converges weakly to a solution of (3.2) when B is a monotone and Lipschitz continuous operator in the hypotheses of Theorem 3.3.6.

As it has been shown in [24], it also converges when B is pseudo-monotone on $\mathcal{H}$, Lipschitz continuous, and sequentially weak-to-weak continuous, and also when $\mathcal{H}$ is finite dimensional, and B is pseudo-monotone on C and Lipschitz continuous.

Denoting $u_k := w_k - \gamma_k(Bw_k - Bv_k)$ and $\theta_k := \frac{\gamma_k}{\gamma_{k+1}}$ for all $k \geq 1$, then for all $z^* \in \mathrm{zer}(N_C + B)$ we get

$$\|u_k - z^*\|^2 \leq \|v_k - z^*\|^2 - \left(1 - \mu^2\theta_k^2\right)\|w_k - v_k\|^2 \quad \forall k \geq 1$$

which is nothing else than relation (3.16).

Indeed, since $w_k \in C$, we have $\langle Bz^*, w_k - z^*\rangle \geq 0$, and, further, the pseudo-monotonicity of B gives $\langle Bw_k, w_k - z^*\rangle \geq 0$ for all $k \geq 1$. On the other hand, since $w_k = P_C(\mathrm{Id} - \gamma_k B)v_k$, we have $\langle z^* - w_k, w_k - v_k + \gamma_k Bv_k\rangle \geq 0$ for all $k \geq 1$. The two inequalities yield

$$\langle w_k - z^*, v_k - u_k\rangle \geq 0 \quad \forall k \geq 1,$$

which, combined with (3.17), lead as in the proof of Proposition 3.3.2 to the conclusion.

Now, since relation (3.16) holds, the statements in Proposition 3.3.3 and Proposition 3.3.5 remain true and, as seen in the proof of Theorem 3.3.6, they guarantee that the limit $\lim_{k\to+\infty}\|z_k - z^*\| \in \mathbb{R}$ exists, and that

$$\lim_{k\to+\infty}\|w_k - v_k\| = \lim_{k\to+\infty}\|Bw_k - Bv_k\| = 0.$$

Having that, the weak converge of $(z_k)_{k\geq 0}$ to a solution of (3.2) follows, by arguing as in the proof by [24, Theorem 3.1], when B is pseudo-monotone on $\mathcal{H}$, Lipschitz-continuous and sequentially weak-to-weak continuous, and as in the proof by [24, Theorem 3.2], when $\mathcal{H}$ is finite dimensional, and B is pseudo-monotone on C and Lipschitz continuous.

3.3.1. Numerical experiments

In this subsection, we treat a bilinear saddle point problem, which is a usual deterministic example where all the assumptions are fulfilled to guarantee convergence. We illustrate the interplay between inertia and relaxation, for which we look at different quantities, e.g., the minimax gap, the fixed point residual or the discrete velocity of the iterates.

Bilinear saddle point problem

We want to solve the following bilinear saddle-point problem

$$\min_{x\in X}\max_{y\in Y}\Phi(x,y) := x^TAy + a^Tx + b^Ty, \tag{3.26}$$

with $A \in \mathbb{R}^{d\times n}$, $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^n$ and where $X \subseteq \mathbb{R}^d$ and $Y \subseteq \mathbb{R}^n$ are nonempty, closed and convex sets, in the sense that we want to find $x^* \in X$ and $y^* \in Y$ such that

$$\Phi(x^*,y) \leq \Phi(x^*,y^*) \leq \Phi(x,y^*)$$

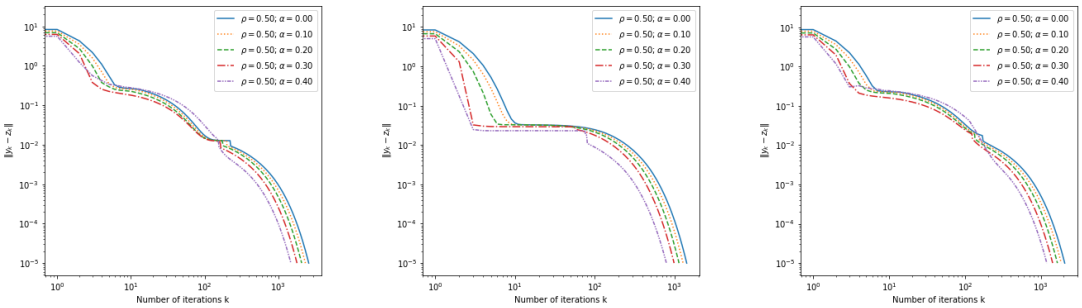
for all $x \in X$ and all $y \in Y$. The monotone inclusion to be solved in this case is of the form

$$0 \in N_{X\times Y}(x,y) + F(x,y),$$

where $N_{X\times Y}$ is the maximal monotone normal cone operator to $X \times Y$ and

$$F(x,y) = M\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ -b \end{pmatrix}, \text{ with } M = \begin{pmatrix} 0 & A \\ -A^T & 0 \end{pmatrix},$$

is monotone and Lipschitz continuous with Lipschitz constant $L = \|M\|_2$. Notice that F is not cocoercive.



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

Figure 3.2.: Behaviour of the fixed point residual $\|w_k - v_k\|$ for the constrained bilinear saddle-point problem with data drawn from different distributions ($\mu = 0.5$).

For our experiments we choose $d = n = 500$, 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 step size $\gamma_k = \gamma = \frac{\mu}{L}$, where $0 < \mu < 1$, constant inertial parameter $\alpha_k = \alpha$ and constant relaxation parameter $\rho_k = \rho$ for all $k \geq 1$. We set $z_k := (x_k, y_k)$ to fit the framework of our algorithm. The starting point z_0 is initialised randomly (with entries drawn from the uniform distribution on $[0, 1]$ for all three settings) and we set $z_1 = z_0$. We did not observe different behaviour

for various random trials, so each time we provide the results for only one run in the following.

In Figure 3.2 we can see the development of $\|w_k - v_k\|$ for problem (3.26) with data drawn from the different distributions for $\mu = 0.5$, $\rho = 0.5$ and various values of α. The quantity $\|w_k - v_k\|$ is the fixed point residual of the operator $J_{\gamma A}(\mathrm{Id} - \gamma B)$, for which according to the convergence analysis we have that $\|w_k - v_k\| \to 0$ as $k \to +\infty$. As solutions of the problem (3.26) are not available explicitly, we look at this meaningful surrogate instead. Also, if we have $w_k = v_k$ for some $k \geq 1$ then $z_{k+1} = w_k = v_k$ is a solution of (3.26). We see that the behaviour 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 any more 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	1.10	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 3.1.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a uniform distribution ($\mu = 0.5$)

In Table 3.1 we can see the necessary number of iterations for the algorithm to achieve $\|w_k - v_k\| \leq 10^{-5}$ 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 A.1.

As mentioned in Remark 3.3.4, 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.

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

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 3.1 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 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 3.2 ($\mu = 0.9$) and Table 3.3 ($\mu = 0.1$), we can observe a very interesting behaviour. Even though smaller μ allows for larger values of the relaxation parameter (see Remark 3.3.4), larger μ leads to better

results in general. This behaviour presumably is due to the role μ plays in the definition of the step size.

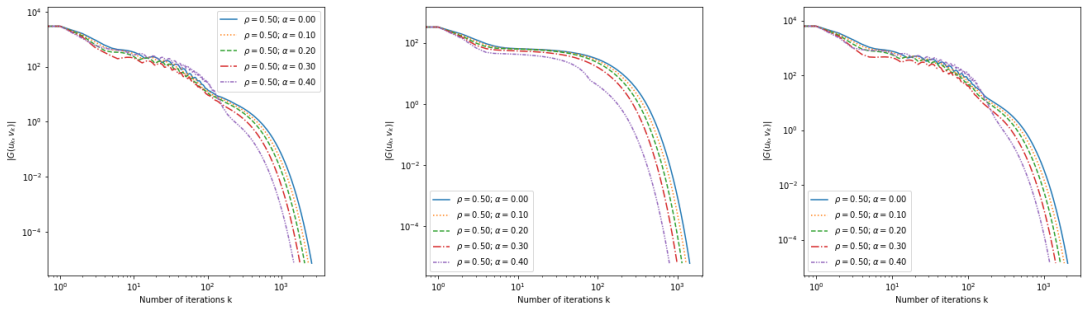
To get further insight into the convergence behaviour, we also look at the following gap function,

$$G(s,t) := \inf_{x \in X} \Phi(x,t) - \sup_{y \in Y} \Phi(s,y).$$

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

$$G(x_k, y_k) = -\|Ay_k + a\|_2 + b^T y_k - \left\|A^T x_k + b\right\|_2 - a^T x_k \quad \forall k \geq 0.$$

In Figure 3.3 we see the development of the absolute value of the gap $|G(x_k, y_k)|$ for problem (3.26) 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 behaviour 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 behaviour of the gap is not consistently better any more when α gets close to the limiting case. In the meantime the gap for maximal α is worst compared to all other parameter combinations, although 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.



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

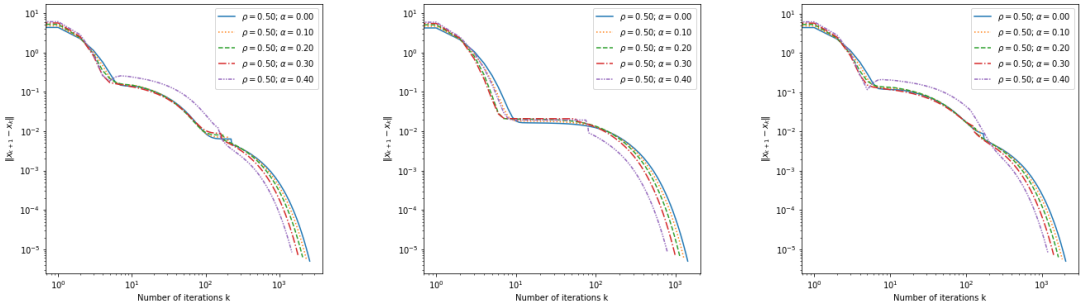
Figure 3.3.: Behaviour of the gap function $|G(x_k, y_k)|$ for the constrained bilinear saddle-point problem with data drawn from different distributions ($\mu = 0.5$).

Finally, in Figure 3.4 we look at the discrete velocity of the iterates $\|z_{k+1} - z_k\|$ 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

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	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ 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	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ $\geq$ 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	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ $\geq$ 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	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ $\geq$ 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	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	8,402	6,996	5,992	5,227	4,395	4,215	3,837	3,428	3,338	3,371	-	-	-	-
0.20	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	7,998	6,658	5,685	4,906	4,261	4,048	3,656	3,415	3,574	-	-	-	-	-
0.24	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	$\nearrow$ 9,501	7,592	6,309	5,263	4,497	4,225	3,841	3,626	-	-	-	-	-	-	-
0.28	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	8,994	7,178	5,872	4,861	4,297	4,065	3,835	-	-	-	-	-	-	-	-
0.32	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	$\nearrow$ 10,000	8,481	6,666	5,364	4,612	4,269	4,105	-	-	-	-	-	-	-	-	-
0.36	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	7,906	6,098	5,053	4,548	4,450	-	-	-	-	-	-	-	-	-	-
0.40	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	9,961	7,205	5,649	5,008	4,924	-	-	-	-	-	-	-	-	-	-	-
0.44	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	9,109	6,572	5,520	5,490	-	-	-	-	-	-	-	-	-	-	-	-
0.48	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	8,174	6,356	6,115	-	-	-	-	-	-	-	-	-	-	-	-	-
0.52	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	7,633	6,785	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.56	$\nearrow$ 10,000	$\nearrow$ $\geq$ 10,000	$\nearrow$ 10,000	7,642	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.60	$\nearrow$ 10,000	$\nearrow$ 10,000	9,747	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	$\nearrow$ 10,000	$\nearrow$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	$\nearrow$ 10,000	$\nearrow$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	$\nearrow$ 10,000	$\nearrow$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	$\nearrow$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	$\nearrow$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	$\nearrow$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	$\nearrow$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 3.3.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a uniform distribution ($\mu = 0.1$)

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 3.4.: Behaviour of the discrete velocity of the iterates $\|z_{k+1} - z_k\|$ for the constrained bilinear saddle-point problem with data drawn from different distributions ($\mu = 0.5$).

3.4. Convergence rates of Tseng's FBF algorithm

In the following section we aim to establish non-asymptotic convergence rates for the original version of the FBF algorithm [137], both in the deterministic and the stochastic setting. Due to the lack of a more specific structure the typical measure of optimality for the general monotone inclusion (MI) is given by the distance of the current iterate to a solution. Usually, without stronger assumptions on the problem, one can only show asymptotic convergence for this entity (see Section 3.3). In the case of monotone inclusions derived from variational inequalities (MI-VI) or minimax problems (MM), however, it is possible to show non-asymptotic convergence rates even in the general setting.

For this reason we will investigate the monotone inclusion corresponding to the general variational inequality (MI-VI) in the following. We want to remind that (MI-VI) consists of finding $z^* \in \mathcal{H}$ such that

$$0 \in \partial r(z^*) + F(z^*), \tag{MI-VI}$$

where we assume that $F : \mathcal{H} \to \mathcal{H}$ is a monotone and L-Lipschitz continuous operator ($L > 0$) and $r : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex and lower semicontinuous function. In particular, minimax problems (MM) constitute an especially interesting instance of (MI-VI). Recall, the general convex-concave minimax problem is given by

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Psi(x, y) := f(x) + \Phi(x, y) - g(y), \tag{MM}$$

where $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex-concave and differentiable with L-Lipschitz continuous gradient ($L > 0$), and $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ are proper, convex and lower semicontinuous. A solution of (MM) is given by a saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$, fulfilling

$$\Psi(x^*, y) \leq \Psi(x^*, y^*) \leq \Psi(x, y^*) \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}. \tag{SP}$$

The connection between the two problems can easily be seen when casting the optimality conditions for (SP) in the form of (MI-VI) via

$$r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}, \quad (x, y) \mapsto f(x) + g(y),$$

and

$$F : \mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}, \quad (x, y) \mapsto \begin{pmatrix} \nabla_x \Phi(x, y) \\ -\nabla_y \Phi(x, y) \end{pmatrix}. \tag{3.27}$$

3.4.1. Solution methods for monotone inclusion (MI-VI)

The connection between monotone inclusions and saddle point problems is of course not new. The application of the *Extragradient* method (EG) [85] to minimax problems has been studied in the seminal paper [112] under the name of *Mirror Prox* and a convergence rate of $\mathcal{O}(1/k)$ in terms of the function values has been proven. Even a stochastic version of the Mirror Prox algorithm has been studied in [82] with a convergence rate of $\mathcal{O}(1/\sqrt{k})$. Applied to problem (MI-VI), it iterates for all $k \geq 0$

$$\text{EG:} \left\lfloor \begin{array}{l} w_k = \operatorname{prox}_{\gamma_k r}(z_k - \gamma_k F(z_k)) \\ z_{k+1} = \operatorname{prox}_{\gamma_k r}(z_k - \gamma_k F(w_k)). \end{array} \right.$$

The *Forward-Backward-Forward* (FBF) or Tseng's method, introduced in [137] where convergence of the iterates was established, has not been studied rigorously for minimax problems in terms of function values yet, despite promising applications in [36] and *its advantage of it only requiring one projection*, whereas EG needs two. For all $k \geq 0$ it is given by

$$\text{FBF:} \left\lfloor \begin{array}{l} w_k = \operatorname{prox}_{\gamma_k r}(z_k - \gamma_k F(z_k)) \\ z_{k+1} = w_k - \gamma_k (F(w_k) - F(z_k)). \end{array} \right. \tag{3.28}$$

Both, EG and FBF, have the "disadvantage" of needing two gradient evaluations per iteration. A possible remedy — suggested in [64] for EG under the name of *extrapolation from the past* — is to reuse previous gradients. In a similar fashion we consider for all $k \geq 1$

$$\text{FBFp:} \left\lfloor \begin{array}{l} w_k = \operatorname{prox}_{\gamma_k r}(z_k - \gamma_k F(w_{k-1})) \\ z_{k+1} = w_k - \gamma_k (F(w_k) - F(w_{k-1})), \end{array} \right. \tag{3.29}$$

where we replaced $F(z_k)$ by $F(w_{k-1})$ twice in (3.28). As a matter of fact, the above method can be written exclusively in terms of $(w_k)_{k \geq 0}$ by incrementing the index k in the first line of (3.29) and then substituting z_{k+1} by the second line. This results in

$$w_{k+1} = \operatorname{prox}_{\gamma_k r}(w_k - \gamma_{k+1} F(w_k) - \gamma_k [F(w_k) - F(w_{k-1})]) \quad \forall k \geq 1. \tag{3.30}$$

This way we rediscover a known method which was studied in [100], where convergence of the iterates was established for general monotone inclusions under the name of *forward-reflected-backward*. It reduces to *optimistic mirror descent* [126,127] in the unconstrained case with constant step size $\gamma_k = \gamma$, giving

$$w_{k+1} = w_k - \gamma(2F(w_k) - F(w_{k-1})) \quad \forall k \geq 1, \tag{3.31}$$

which was introduced in [123] for saddle point problems and which has been proposed for the training of GANs under the name of *Optimistic Gradient Descent Ascent (OGDA)*, see [54, 55, 89]. Only recently a method very related to (3.30) was proposed in [52] and is characterized by applying the correction term after the projection,

$$w_{k+1} = \text{prox}_{\gamma r}(w_k - \gamma F(w_k)) - \gamma(F(w_k) - F(w_{k-1})) \quad \forall k \geq 1. \tag{3.32}$$

Evidently, in the unconstrained case and for constant step size the methods (3.29), (3.30), (3.31) and (3.32) are all equivalent.

All of the above methods and extensions rely solely on the monotone operator formulation of the saddle point problem where the two components x and y play a symmetric role. Taking the special minimax structure into consideration, [75] showed convergence of a method that uses an optimistic step (3.31) in one component and a regular gradient step in the other, thus requiring less storing of past gradients in comparison to (3.30).

On the downside, however, by reducing the number of required gradient evaluations per iteration, the largest possible step size is reduced from $1/L$ (see [85] or Theorem 3.4.5) to $1/2L$ (see [64, 98, 100] or Theorem 3.4.5). To summarise, the number of required gradient evaluations is halved, but so is the step size, resulting in no clear net gain.

3.4.2. Measure of optimality

As we have already mentioned, in order to obtain convergence rates we will look at a more problem specific quantity than the distance to the solution set to measure the quality of a point with respect to (MI-VI). If F arises from a saddle point problem (MM), meaning that F has the form (3.27), we want to make use of the saddle function Ψ, similarly to the objective function in a pure optimisation setting. We do this via the *minimax gap*, which for a point $w = (u, v) \in \mathcal{X} \times \mathcal{Y}$ measures the violation of w not fulfilling (SP) and is given by

$$G^{MM}(w) = \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Psi(u, y) - \Psi(x, v) = \sup_{y\in\mathcal{Y}} \Psi(u, y) - \inf_{x\in\mathcal{X}} \Psi(x, v). \tag{3.33}$$

This minimax gap can be interpreted from a game theoretic point of view as the sum of the maximal payoffs achievable by the two players by playing their respective best responses, given the current strategy of the opponent.

In the more general monotone inclusion setting (MI-VI), where no function values are available, an appropriate generalisation of (3.33) is given for any $w \in \mathcal{H}$ by

$$G^{VI}(w) = \sup_{z\in\mathcal{H}} \langle F(z), w - z\rangle + r(w) - r(z). \tag{3.34}$$

It stems from the field of variational inequalities, where such a function is also known as *merit function* [114]. With the notion of VIs in mind (see Section 3.1.1), the above defined gap (3.34) becomes natural as it measures how much the statement of (VI-w) is violated.

If $r = \delta_\Omega$ is the indicator of a compact and convex set Ω it is clear that the supremum is only taken over $z \in \Omega$ and will thus be finite for $w \in \Omega$. Since the problem (MI-VI) is in general unconstrained and the supremum can be infinite, we consider, as done for example in [114], the *restricted* gap instead, where the above supremum is taken over an auxiliary compact set $B \subseteq \mathcal{H}$ and $B \subseteq \mathcal{X} \times \mathcal{Y}$, respectively, instead of the entire space. Thus we define the *restricted minimax gap* (identifying $w = (u, v)$)

$$G_B^{MM}(w) := \sup_{(x,y)\in B} \Psi(u, y) - \Psi(x, v), \tag{3.35}$$

and the *restricted merit function*

$$G_B^{VI}(w) := \sup_{z\in B} \langle F(z), w - z\rangle + r(w) - r(z). \tag{3.36}$$

In order to capture both instances at the same time, we define the following unifying gap

$$G_B(w) := \begin{cases} G_B^{MM}(w) & \text{if } F \text{ and } r \text{ come from (MM)} \\ G_B^{VI}(w) & \text{otherwise.} \end{cases} \tag{3.37}$$

Note that the restricted gap is in general only a reasonable measure of optimality for elements of B. It is nonnegative on B and zero for points of B which solve (MI-VI). Additionally, we want to be able to conclude that if a point has zero gap it solves (MI-VI). This is for example the case if it is in the interior of B, which can always be ensured if B is chosen large enough.

The relevance of the above two quantities G_B^{MM} and G_B^{VI} will be made clear by the following statements.

Theorem 3.4.1. *Let $B \subseteq \mathcal{X} \times \mathcal{Y}$. A point (x^*, y^*) in the interior of B solves the saddle point problem* (MM) *if and only if its restricted minimax gap* (3.35) *is zero, i.e., $G_B^{MM}(x^*, y^*) = 0$. For all other elements of B the gap is nonnegative.*

Proof. Let (x^*, y^*) be in the interior of B. If $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ is a saddle point, it clearly fulfils that $\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Psi(x^*, y) - \Psi(x, y^*) = 0$. On the other hand let $G_B^{MM}(x^*, y^*) = 0$. For an arbitrary point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we can choose $\beta \in (0, 1)$ large enough such that $(u, v) := \beta(x^*, y^*) + (1 - \beta)(x, y) \in B$. Therefore

$$\Psi(x^*, v) - \Psi(u, y^*) = \Psi(x^*, \beta y^* + (1 - \beta)y) - \Psi(\beta x^* + (1 - \beta)x, y^*) \leq 0.$$

Using the convex-concave structure of Ψ we deduce that

$$\beta\Psi(x^*, y^*) + (1 - \beta)\Psi(x^*, y) - \beta\Psi(x^*, y^*) - (1 - \beta)\Psi(x, y^*) \leq 0,$$

which implies that $\Psi(x^*, y) \leq \Psi(x, y^*)$. Since $(x, y) \in \mathcal{X} \times \mathcal{Y}$ was chosen arbitrary, (x^*, y^*) is a saddle point. □

Similarly, an analogous statement can be shown for the restricted merit function G_B^{VI}.

Theorem 3.4.2. *Let $B \subseteq \mathcal{H}$. A point z^* in the interior of B solves the monotone inclusion* (MI-VI) *if and only if its restricted merit function* (3.36) *is zero, i.e., $G_B^{VI}(z^*) = 0$. For all other elements of B the gap is nonnegative.*

For the proof, recall that under the given assumptions the monotone inclusion (MI-VI) is equivalent to the weak formulation of the variational inequality (VI-w) (see Section 3.1.1). That means that $z^* \in \mathcal{H}$ solves (MI-VI) if and only if

$$\langle F(z), z^* - z\rangle + r(z^*) - r(z) \leq 0 \quad \forall z \in \mathcal{H}. \tag{VI-w}$$

Proof. First, observe that for $w \in B$ we have

$$0 = \langle F(w), w - w\rangle + r(w) - r(w) \leq \sup_{z\in B} \langle F(z), w - z\rangle + r(w) - r(z) = G_B^{VI}(w).$$

Now, let z^* be in the interior of B. If z^* is a solution to the weak VI (VI-w), then clearly

$$\langle F(z), z^* - z\rangle + r(z^*) - r(z) \leq 0 \quad \forall z \in B,$$

and thus $G_B^{VI}(z^*) = 0$.

On the other hand let $G_B^{VI}(z^*) = 0$. Since z^* is in the interior of B, for an arbitrary $z \in \mathcal{H}$ we can choose $\beta \in (0,1)$ large enough such that $w := \beta z^* + (1-\beta)z \in B$. Therefore

$$\begin{aligned} 0 &\geq \langle F(w), z^* - w\rangle + r(z^*) - r(w) \\ &= \langle F(\beta z^* + (1-\beta)z), z^* - \beta z^* - (1-\beta)z\rangle + r(z^*) - r(\beta z^* + (1-\beta)z) \end{aligned}$$

This implies that

$$(1-\beta)\,\langle F(\beta z^* + (1-\beta)z), z - z^*\rangle + (1-\beta)(r(z) - r(z^*)) \geq 0.$$

Dividing by $(1-\beta)$ and then taking the limit $\beta \to 1$ gives

$$\langle F(z^*), z - z^*\rangle + r(z) - r(z^*) \geq 0.$$

Since $z \in \mathcal{H}$ was chosen arbitrary, z^* solves the strong form of the VI (VI-s) and thus (MI-VI). □

3.4.3. Convergence statements

We now present a novel unifying scheme for solving problem (MI-VI), which generalises FBF (3.28) and in addition recovers the method motivated in (3.29) as FBFp. Let us point out again that the latter algorithm was already introduced in [100] and corresponds to OGDA [54, 55, 123, 126] if F stems from the minimax setting (3.27).

Algorithm 3.4.3 (generalised FBF)**.** For a starting point $z_0 \in \mathcal{H}$ and step sizes $\gamma_k > 0$ we consider for all $k \geq 0$

$$\left\lfloor \begin{array}{l} w_k = \operatorname{prox}_{\gamma_k r}(z_k - \gamma_k F(\diamondsuit_k)) \\ z_{k+1} = w_k - \gamma_k(F(w_k) - F(\diamondsuit_k)). \end{array} \right.$$

For $\diamondsuit_k = z_k$ this reduces to the well known FBF method, whereas $\diamondsuit_k = w_{k-1}$, with the additional initial condition $w_{-1} = z_0$, reuses previous gradients (FBFp).

Consider the scenario where the operator F is given only in expectation, i.e., instead of solving problem (MI-VI) we aim to find $z^* \in \mathcal{H}$ such that

$$0 \in \partial r(z^*) + \mathbb{E}_{\xi \sim Q}[F(z^*; \xi)]$$

This means that instead of F itself only a stochastic estimator $F(\,\cdot\,; \xi)$ is accessible. In this case we adapt Algorithm 3.4.3 in the following way.

Algorithm 3.4.4 (generalised stochastic FBF)**.** For a starting point $z_0 \in \mathcal{H}$ and step sizes $\gamma_k > 0$ we consider for all $k \geq 0$

$$\left\lfloor \begin{array}{l} \xi_k \sim Q \quad (\text{optionally } \eta_k \sim Q) \\ w_k = \operatorname{prox}_{\gamma_k r}(z_k - \gamma_k F(\diamondsuit_k; \triangle_k)) \\ z_{k+1} = w_k - \gamma_k(F(w_k; \xi_k) - F(\diamondsuit_k; \triangle_k)). \end{array} \right.$$

For $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$ this results in a stochastic version of FBF, whereas $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$, with the additional initial condition $w_{-1} = z_0$ and $\xi_{-1} = \eta_0$, reuses previous gradients (stochastic FBFp).

Even though both deterministic methods encompassed by the unifying scheme in Algorithm 3.4.3 have been studied before, the stated convergence results in the following are new. Note that while the rate for FBF is completely new our result for FBFp provides only a generalisation of the known rate for OGDA, see [109]. Similarly, the stochastic version of FBF has been considered recently in [34] and rates in terms of the fixed point residual $\|z_k - w_k\|$ have been obtained, but no statements for function values have been derived. However, we want to point out that the stochastic version of FBFp has not been considered prior to this work.

Let in the following $B \subseteq \mathcal{H}$ (identify $\mathcal{H} \equiv \mathcal{X} \times \mathcal{Y}$ in the case of (MM)) be the compact set of the restricted (unifying) gap function (3.37) with $D := \sup_{w,z \in B} \|z - w\|$ denoting its diameter. For convenience in the estimation we assume that for each of the discussed methods we have the starting point $z_0 \in B$. Recall that $L > 0$ denotes the Lipschitz constant of the operator F. Furthermore, the averaged iterates $\bar{w}_K$ for $K \geq 1$ are given by

$$\bar{w}_K := \frac{1}{\sum_{k=0}^{K-1} \gamma_k} \sum_{k=0}^{K-1} \gamma_k w_k.$$

While our convergence results are stated in terms of the averaged iterates, a technique which can prove beneficial in practice [47, 64], having guarantees on the last iterate would be more in the spirit of nonconvex methods. Such results have been obtained in [54] and [45], but are for bilinear and strongly-convex-strongly-concave problems, respectively, which are known to allow for the derivation of better convergence rates. On the other hand the works [68,71,93] have been able to guarantee rates of convergence for the last iterate for convex-concave problems, but all of them only in the unconstrained setting. Furthermore [68] uses additional smoothness assumptions, [93] employs a second-order method and [71] uses the norm of the gradient as the measure of optimality.

First we will formulate the convergence statements for the four different methods, which is followed by the respective convergence analysis that uses to a certain extent unified intermediate results.

Let us start with the convergence statements for the deterministic method from Algorithm 3.4.3.

Theorem 3.4.5 (deterministic). *Let $(w_k)_{k\geq 0}$ be generated by Algorithm 3.4.3. If*

(i) FBF, i.e., $\diamondsuit_k = z_k$, with step size $0 < \gamma_k \leq \gamma \leq 1/L$, or

(ii) FBFp, i.e., $\diamondsuit_k = w_{k-1}$, with step size $0 < \gamma_k \leq \gamma \leq 1/2L$

is chosen, then for all $K \geq 1$ we have

$$G_B(\bar{w}_K) \leq \frac{D^2}{2\sum_{k=0}^{K-1}\gamma_k}.$$

If we take a constant step size $\gamma_k = \gamma$ for all $k \geq 0$ according to Theorem 3.4.5, this choice amounts to a convergence rate of

$$G_B(\bar{w}_K) \leq \frac{D^2}{2\gamma K}.$$

In order to derive similar convergence statements for the stochastic method from Algorithm 3.4.4, we need to assume (standard) properties of the gradient estimator $F(\cdot\,;\xi)$.

Assumption 3.4.6. *Unbiasedness:* $\mathbb{E}_\xi[F(z;\xi)] = F(z) \quad \forall z \in \mathcal{H}$.

Assumption 3.4.7. *Bounded variance:* $\mathbb{E}_\xi\left[\|F(z;\xi) - F(z)\|^2\right] \leq \sigma^2 \quad \forall z \in \mathcal{H}$.

We want to point out that the latter assumption is weaker than the commonly used bounded (sub)gradient hypothesis ($\mathbb{E}_\xi[\|F(z;\xi)\|^2] \leq \sigma^2$), which is known to conflict with properties like strong monotonicity, see [94].

In particular, as we will see in the proofs, we actually only need the above assumption to hold for all iterates w_k. Such an hypothesis is in practice difficult to check, but could be exploited in special cases where additional properties of the variance and boundedness of the iterates are known a priori.

Assumption 3.4.8. *The samples ξ_k are independent of the iterates w_k, for all $k \geq 0$.*

Equipped with these assumptions we are able to prove the following statement.

Theorem 3.4.9 (stochastic)**.** *Let Assumptions 3.4.6–3.4.8 hold and let $(w_k)_{k\geq 0}$ be generated by Algorithm 3.4.4.*

(i) If stochastic FBF is chosen, i.e., $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$, with step size $0 < \gamma_k \leq \gamma < \frac{1}{L}$, then for all $K \geq 1$ we have

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 4(1-\gamma^2L^2)^{-1}\sigma^2 \sum_{k=0}^{K-1} \gamma_k^2}{\sum_{k=0}^{K-1} \gamma_k}.$$

(ii) If stochastic FBFp is chosen, i.e., $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$, with step size $0 < \gamma_k \leq \gamma < \frac{1}{2\sqrt{2}L}$, then for all $K \geq 1$ we have

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 6\left(1 + \frac{2\gamma^2L^2}{1-8\gamma^2L^2}\right)\sigma^2 \sum_{k=0}^{K-1} \gamma_k^2}{\sum_{k=0}^{K-1} \gamma_k}.$$

Even though the step size in Theorem 3.4.9 can be chosen arbitrarily close to $1/L$ and $1/(2\sqrt{2}L)$ for stochastic FBF and stochastic FBFp, respectively, this does not mean it is advisable as doing so would lead to a deteriorating constant in the estimates. Bounding away the step sizes from the respective upper bound can be used to obtain a unified convergence statement, similar to Theorem 3.4.5. On the one hand, if we take $\gamma = 1/\sqrt{2}L$ we can conclude

$$\frac{4}{1-\gamma^2L^2} = 8.$$

On the other hand, for $\gamma = 1/3L$ we obtain

$$6\left(1 + \frac{2\gamma^2L^2}{1-8\gamma^2L^2}\right) = 18.$$

This means that with step sizes $0 < \gamma_k \leq 1/\sqrt{2}L$ and $0 < \gamma_k \leq 1/3L$ $(k \geq 0)$ in the stochastic FBF and FBFp method, respectively, for $K \geq 1$ we get the unified estimate

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2 + 18\sigma^2 \sum_{k=0}^{K-1} \gamma_k^2}{\sum_{k=0}^{K-1} \gamma_k}.$$

The above theorem exhibits a classical step size dependence [128], yielding convergence for sequences $(\gamma_k)_{k\geq 0}$ that are square summable $\sum_{k=0}^{\infty} \gamma_k^2 < +\infty$ but not summable $\sum_{k=0}^{\infty} \gamma_k = +\infty$. Additionally, if in the setting of Theorem 3.4.9 the step size is chosen to be $\gamma_k = \gamma/\sqrt{k+1}$, a convergence rate can be obtained and is given by

$$\mathbb{E}[G_B(\bar{w}_K)] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right). \tag{3.38}$$

If the step size does not go to zero, the gap can usually not be expected to vanish either due to the inherent stochastic noise. However, we can still show decrease in the gap up to a residual stemming from the variance. In particular, for a constant step size $\gamma_k = \gamma$ we have

$$\mathbb{E}[G_B(\bar{w}_K)] \leq \frac{D^2}{\gamma K} + 18\sigma^2\gamma. \tag{3.39}$$

Furthermore, if the number of iterations K is fixed beforehand, a conclusion similar to (3.38) can be obtained by choosing $\gamma = 1/\sqrt{K}$ in (3.39).

3.4.4. Proofs

Preparations

Before we tackle the actual proofs of Theorem 3.4.5 and Theorem 3.4.9, let us introduce some notation connected to the strong formulation of the variational inequality (VI-s) associated to the monotone inclusion (MI-VI). We define $h : \mathcal{H} \times \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ via

$$h(w, z) := \langle F(w), w - z\rangle + r(w) - r(z),$$

which will be used to bound the (restricted) unifying gap function (3.37), which we recall, is given by

$$G_B(w) = \begin{cases} G_B^{MM}(w) & \text{if } F \text{ and } r \text{ come from (MM)} \\ G_B^{VI}(w) & \text{otherwise.} \end{cases}$$

In a similar spirit we define

$$h^{MM}(w, z) := \Psi(u, y) - \Psi(x, v),$$

where we identify $w = (u, v)$ and $z = (x, y)$, and

$$h^{VI}(w, z) := \langle F(z), w - z\rangle + r(w) - r(z),$$

and see that

$$G_B^{MM}(w) = \sup_{z\in B} h^{MM}(w, z) \quad \text{and} \quad G_B^{VI}(w) = \sup_{z\in B} h^{VI}(w, z).$$

Lemma 3.4.10. *Let $(\gamma_k)_{k\geq 0} \subseteq \mathbb{R}_{++}$ and let $(w_k)_{k\geq 0} \subseteq \mathcal{H}$. Then for all $K \geq 1$ the following holds,*

$$\sup_{z\in B} \left\{ \frac{1}{\sum_{k=0}^{K-1} \gamma_k} \sum_{k=0}^{K-1} \gamma_k h(w_k, z) \right\} \geq G_B(\bar{w}_K).$$

Proof. We start with the case, when F is a general monotone operator. Using its monotonicity we deduce that for all w, $z \in \mathcal{H}$ we have

$$\langle F(w), w - z\rangle \geq \langle F(z), w - z\rangle .$$

Adding $r(w) - r(z)$ to the above inequality then gives

$$h(w,z) \geq h^{VI}(w,z). \tag{3.40}$$

On the other hand, if F and r are derived from a saddle point problem (MM), i.e., when

$$F(x,y) = \begin{pmatrix} \nabla_x \Phi(x,y) \\ -\nabla_y \Phi(x,y) \end{pmatrix} \quad \text{and} \quad r(x,y) = f(x) + g(y),$$

from the convex-concave structure of Φ we get for (u,v), $(x,y) \in \mathcal{X} \times \mathcal{Y}$ that

$$\Phi(u,y) - \langle \nabla_y \Phi(u,v), y - v \rangle \leq \Phi(u,v) \quad \text{and} \quad \Phi(u,v) + \langle \nabla_x \Phi(u,v), x - u \rangle \leq \Phi(x,v).$$

By summing up the two inequalities and writing $w = (u,v)$ and $z = (x,y)$ we obtain

$$\langle F(w), w - z \rangle = \left\langle \begin{array}{c} \nabla_x \Phi(u,v), u - x \\ -\nabla_y \Phi(u,v), v - y \end{array} \right\rangle \geq \Phi(u,y) - \Phi(x,v).$$

On a side note, again because of the convex-concave structure of Φ we further get

$$\langle F(w), w - z \rangle \geq \Phi(u,y) - \Phi(x,v) \geq \langle F(z), w - z \rangle .$$

Adding $r(w) - r(z)$ to the above inequality then gives

$$h(w,z) \geq h^{MM}(w,z). \tag{3.41}$$

From (3.40) and (3.41) and using Jensen's inequality, as $h^{VI}(\,\cdot\,, z)$ and $h^{MM}(\,\cdot\,, z)$ are convex functions, we deduce

$$\sum_{k=0}^{K-1} \gamma_k h(w_k, z) \geq \sum_{k=0}^{K-1} \gamma_k h^{VI}(w_k, z) \geq \left(\sum_{k=0}^{K-1} \gamma_k \right) h^{VI}(\bar{w}_K, z)$$

and

$$\sum_{k=0}^{K-1} \gamma_k h(w_k, z) \geq \sum_{k=0}^{K-1} \gamma_k h^{MM}(w_k, z) \geq \left(\sum_{k=0}^{K-1} \gamma_k \right) h^{MM}(\bar{w}_K, z),$$

respectively. Dividing by $\sum_{k=0}^{K-1} \gamma_k$ and then taking the supremum over $z \in B$ finally gives the desired result. □

For all $k \geq 0$ we will use the notations

$$Z_k := F(\diamondsuit_k; \triangle_k) - F(\diamondsuit_k) \quad \text{and} \quad W_k := F(w_k; \xi_k) - F(w_k). \tag{3.42}$$

to denote the error of the stochastic estimator. Furthermore, we will write $\mathbb{E}[\,\cdot\,|\,U]$ for the conditional expectation with respect to the random variable U.

We will also need the following lemma.

Lemma 3.4.11. *Let $(p_k)_{k\geq 0} \subseteq \mathcal{H}$ be a given sequence, let $z \in \mathcal{H}$ and let $(v_k)_{k\geq 0} \subseteq \mathcal{H}$ be defined recursively for all $k \geq 0$ via $v_{k+1} := v_k - p_k$ with given starting value $v_0 \in \mathcal{H}$, then*

$$\sum_{k=0}^{K-1} \langle p_k, v_k - z\rangle \leq \frac{1}{2}\|v_0 - z\|^2 + \frac{1}{2}\sum_{k=0}^{K-1}\|p_k\|^2 .$$

Proof. From Lemma 2.5.4 (i) we can conclude immediately that

$$\langle p_k, v_k - z\rangle = \langle v_k - v_{k+1}, v_k - z\rangle = \frac{1}{2}\left(\|v_k - z\|^2 - \|v_{k+1} - z\|^2 + \|v_{k+1} - v_k\|^2\right),$$

from which the statement of the lemma follows. □

We will continue with a unified decrease result which will be the starting point for all convergence proofs in the following.

Proposition 3.4.12. *Let $\beta > 0$, then for all $k \geq 0$ and $z \in \mathcal{H}$ we have*

$$\begin{aligned}\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}(1+\beta)L^2\gamma_k^2\|\diamondsuit_k - w_k\|^2 \\ + \gamma_k\langle W_k, z - w_k\rangle + (1+\beta^{-1})\gamma_k^2\left(\|W_k\|^2 + \|Z_k\|^2\right).\end{aligned} \tag{3.43}$$

Proof. Let $k \geq 0$ and $z \in \mathcal{H}$ be arbitrary. Using the decomposition (3.42) it follows that

$$\langle \gamma_k F(w_k; \xi_k), w_k - z\rangle = \gamma_k\langle W_k, w_k - z\rangle + \gamma_k\langle F(w_k), w_k - z\rangle . \tag{3.44}$$

Since $\mathrm{prox}_{\gamma_k r} = (\mathrm{Id} + \gamma_k \partial r)^{-1}$ we deduce that

$$\langle z - w_k, w_k - z_k + \gamma_k F(\diamondsuit_k; \triangle_k)\rangle \geq \gamma_k(r(w_k) - r(z)). \tag{3.45}$$

Adding (3.44) and (3.45) gives that

$$\langle \gamma_k(F(w_k;\xi_k) - F(\diamondsuit_k;\triangle_k)) + z_k - w_k, w_k - z\rangle \geq \gamma_k\langle W_k, w_k - z\rangle + \gamma_k h(w_k, z),$$

which, using the definition of z_{k+1}, is equivalent to

$$\langle z - w_k, z_{k+1} - z_k\rangle \geq \gamma_k\langle W_k, w_k - z\rangle + \gamma_k h(w_k, z). \tag{3.46}$$

We estimate the inner product on the left side of the inequality by using the four point identity to deduce

$$\begin{aligned}\langle z - w_k, z_{k+1} - z_k\rangle &= \langle z - z_k + z_k - w_k, z_{k+1} - z_k\rangle \\ &= \frac{1}{2}\left(\|z - z_k\|^2 - \|z_{k+1} - z\|^2 + \|z_{k+1} - w_k\|^2 - \|z_k - w_k\|^2\right).\end{aligned} \tag{3.47}$$

The first two summands are fine as they will cancel by telescoping arguments, so we are left with estimating $\|z_{k+1} - w_k\|^2$. By the definition of z_{k+1} we have that

$$\begin{aligned}\|z_{k+1} - w_k\|^2 &= \gamma_k^2 \|F(\diamondsuit_k; \triangle_k) - F(w_k; \xi_k)\|^2 \\ &= \gamma_k^2 \|F(\diamondsuit_k) - F(w_k) + Z_k - W_k\|^2 \\ &\leq (1+\beta)\gamma_k^2 \|F(\diamondsuit_k) - F(w_k)\|^2 + (1+\beta^{-1})\gamma_k^2 \|Z_k - W_k\|^2 \\ &\leq (1+\beta)\gamma_k^2 L^2 \|\diamondsuit_k - w_k\|^2 + 2(1+\beta^{-1})\gamma_k^2 \left(\|Z_k\|^2 + \|W_k\|^2\right),\end{aligned} \tag{3.48}$$

where we added and subtracted $F(\diamondsuit_k)$ and $F(w_k)$ and applied Young's inequality to deduce the result. Adding (3.48), (3.47) and (3.46) we conclude that

$$\begin{aligned}\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}(1+\beta)L^2\gamma_k^2 \|\diamondsuit_k - w_k\|^2 \\ + \gamma_k \langle W_k, z - w_k\rangle + (1+\beta^{-1})\gamma_k^2 \left(\|W_k\|^2 + \|Z_k\|^2\right).\end{aligned}$$

□

Deterministic methods

Now we can continue with the proofs for the deterministic methods from Section 3.4.3.

Proof of Theorem 3.4.5. For the deterministic methods we have $W_k = Z_k = 0$, thus we can use $\beta \to 0$ and (3.43) reduces to

$$\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}L^2\gamma_k^2 \|\diamondsuit_k - w_k\|^2. \tag{3.49}$$

(i) FBF: We start off by plugging $\diamondsuit_k = z_k$ into (3.49) to deduce that for all $k \geq 0$

$$\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}(1 - L^2\gamma_k^2)\|z_k - w_k\|^2.$$

From this it is clear that the step size is constrained by $0 < \gamma \leq 1/L$ as stated in the theorem. Let $K \geq 1$, then by summing up from $k = 0$ to $K-1$ and dividing by $\sum_{k=0}^{K-1} \gamma_k$ we obtain

$$\frac{1}{\sum_{k=0}^{K-1}\gamma_k}\sum_{k=0}^{K-1}\gamma_k h(w_k, z) \leq \frac{\|z_0 - z\|^2}{2\sum_{k=0}^{K-1}\gamma_k}.$$

The claimed statement is then derived by taking the supremum in z over B and applying Lemma 3.4.10.

(ii) FBFp: In this case we plug $\diamondsuit_k = w_{k-1}$ into (3.49) and for all $k \geq 0$ we obtain

$$\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 + \frac{1}{2}L^2\gamma_k^2\|w_{k-1} - w_k\|^2. \tag{3.50}$$

Now we need to bound the term $\|w_{k-1} - w_k\|^2$ by $\|z_k - w_k\|^2$. Since

$$2\|z_k - w_k\|^2 + 2\|z_k - w_{k-1}\|^2 \geq \|w_k - w_{k-1}\|^2, \tag{3.51}$$

we have for all $k \geq 1$

$$\begin{aligned}\|z_k - w_k\|^2 &\geq \frac{1}{2}\|w_{k-1} - w_k\|^2 - \|z_k - w_{k-1}\|^2 \\ &\geq \frac{1}{2}\|w_{k-1} - w_k\|^2 - L^2\gamma_{k-1}^2\|w_{k-1} - w_{k-2}\|^2.\end{aligned} \tag{3.52}$$

Whereas for $k = 0$, since $w_{-1} = z_0$, we have

$$\|z_0 - w_0\|^2 = \|w_{-1} - w_0\|^2. \tag{3.53}$$

Plugging (3.53) into (3.50) for $k = 0$ we get that

$$\gamma_0 h(w_0, z) + \frac{1}{2}\|z_1 - z\|^2 + \frac{1}{2}(1 - L^2\gamma_0^2)\|w_0 - w_{-1}\|^2 \leq \frac{1}{2}\|z_0 - z\|^2. \tag{3.54}$$

Plugging (3.52) into (3.50) we get that for all $k \geq 1$

$$\begin{aligned}&\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 + \frac{1}{2}\left(\frac{1}{2} - \gamma_k^2 L^2\right)\|w_k - w_{k-1}\|^2 \\ &\leq \frac{1}{2}\|z_k - z\|^2 + \frac{1}{2}\gamma_{k-1}^2 L^2\|w_{k-1} - w_{k-2}\|^2.\end{aligned} \tag{3.55}$$

From this we deduce that for all $k \geq 0$ we need to ensure that

$$0 < \gamma_k \leq \frac{1}{2L}$$

in order to be able to telescope. Now let $K \geq 1$ and sum up (3.55) from $k = 1$ to $K - 1$, which yields

$$\begin{aligned}&\sum_{k=1}^{K-1} \gamma_k h(w_k, z) + \frac{1}{2}\|z_k - z\|^2 + \frac{1}{2}\left(\frac{1}{2} - \gamma_{K-1}^2 L^2\right)\|w_{K-1} - w_{K-2}\|^2 \\ &\leq \frac{1}{2}\|z_1 - z\|^2 + \frac{1}{2}\gamma_0^2 L^2\|w_0 - w_{-1}\|^2.\end{aligned}$$

Adding (3.54) to the above equation, followed by dividing by $\sum_{k=0}^{K-1}\gamma_k$ gives

$$\frac{1}{\sum_{k=0}^{K-1}\gamma_k}\sum_{k=0}^{K-1}\gamma_k h(w_k, z) \leq \frac{\|z_0 - z\|^2}{2\sum_{k=0}^{K-1}\gamma_k},$$

where we used that $1 - \gamma_0^2 L^2 \geq \gamma_0^2 L^2$ to get rid of $\|w_0 - w_{-1}\|^2$. Again, the final statement follows by taking the supremum in z over B and applying Lemma 3.4.10.

□

Stochastic methods

We finish this subsection with the proofs for the stochastic methods from Section 3.4.3.

Proof of Theorem 3.4.9. (i) FBF: Plugging $\diamondsuit_k = z_k$ and $\triangle_k = \eta_k$ into (3.43) gives for all $k \geq 0$

$$\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 \leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\left(1 - (1+\beta)\gamma_k^2 L^2\right)\|z_k - w_k\|^2 + \gamma_k \langle W_k, z - v_k\rangle + \gamma_k \langle W_k, v_k - w_k\rangle + \left(1 + \beta^{-1}\right)\gamma_k^2 \left(\|W_k\|^2 + \|Z_k\|^2\right).$$

Let $K \geq 1$, then summing up this inequality from $k = 0$ to $K - 1$ gives

$$\sum_{k=0}^{K-1} \gamma_k h(w_k, z) \leq \frac{1}{2}\|z_0 - z\|^2 - \frac{1}{2}\sum_{k=0}^{K-1}\left(1 - (1+\beta)\gamma_k^2 L^2\right)\|z_k - w_k\|^2 + \sum_{k=0}^{K-1}\gamma_k \langle W_k, z - v_k\rangle + \sum_{k=0}^{K-1}\gamma_k \langle W_k, v_k - w_k\rangle + \left(1 + \beta^{-1}\right)\sum_{k=0}^{K-1}\gamma_k^2 \left(\|W_k\|^2 + \|Z_k\|^2\right).$$

Applying Lemma 3.4.11 with $p_k := -\gamma_k W_k$ and $v_{k+1} = v_k - p_k$ with $v_0 = z_0$, we deduce that

$$\sum_{k=0}^{K-1} \langle -\gamma_k W_k, v_k - z\rangle \leq \frac{1}{2}\|z_0 - z\|^2 + \frac{1}{2}\sum_{k=0}^{K-1}\gamma_k^2 \|W_k\|^2. \tag{3.56}$$

Let $0 < \gamma_k \leq \gamma < 1/L$ for all $k \geq 0$ and let $\beta > 0$ such that $1 + \beta = \left(\gamma^2 L^2\right)^{-1}$, then for all $k \geq 0$ we have

$$1 - (1+\beta)\gamma_k^2 L^2 \geq 0$$

and we obtain

$$1 + \beta^{-1} = \frac{1}{1 - \gamma^2 L^2}.$$

Together with (3.56) this gives

$$\sum_{k=0}^{K-1} \gamma_k h(w_k, z) \leq \|z_0 - z\|^2 + \sum_{k=0}^{K-1}\gamma_k \langle W_k, v_k - w_k\rangle + \frac{2}{1 - \gamma^2 L^2}\sum_{k=0}^{K-1}\gamma_k^2 \left(\|W_k\|^2 + \|Z_k\|^2\right).$$

Next, we take the supremum over $z \in B$ and the expectation to obtain

$$\mathbb{E}\left[\sup_{z \in B}\left\{\sum_{k=0}^{K-1}\gamma_k h(w_k, z)\right\}\right] \leq D^2 + 4\left(1 - \gamma^2 L^2\right)^{-1}\sigma^2 \sum_{k=0}^{K-1}\gamma_k^2,$$

where we used that

$$\begin{aligned}\mathbb{E}[\langle W_k, v_k - w_k\rangle] &= \mathbb{E}\Big[\mathbb{E}\big[\langle W_k, v_k - w_k\rangle \,\big|\, w_{[k]}, \xi_{[k-1]}\big]\Big] \\ &= \mathbb{E}\Big[\big\langle \mathbb{E}\big[W_k \,\big|\, w_{[k]}, \xi_{[k-1]}\big], v_k - w_k\big\rangle\Big] = 0,\end{aligned} \tag{3.57}$$

with $\xi_{[k-1]} = (\xi_0, ..., \xi_{k-1})$ and $w_{[k]} = (w_0, ..., w_k)$. The final statement follows by dividing by $\sum_{k=0}^{K-1} \gamma_k$ and applying Lemma 3.4.10.

(ii) FBFp: By using $\diamondsuit_k = w_{k-1}$ and $\triangle_k = \xi_{k-1}$ we deduce from (3.43) for all $k \geq 0$ that

$$\begin{aligned}\gamma_k h(w_k, z) + \frac{1}{2}\|z_{k+1} - z\|^2 &\leq \frac{1}{2}\|z_k - z\|^2 - \frac{1}{2}\|z_k - w_k\|^2 \\ &+ \frac{1}{2}(1+\beta)L^2\gamma_k^2\|w_{k-1} - w_k\|^2 + \gamma_k\langle W_k, z - v_k\rangle + \gamma_k\langle W_k, v_k - w_k\rangle \\ &+ (1+\beta^{-1})\gamma_k^2\left(\|W_k\|^2 + \|Z_k\|^2\right).\end{aligned} \tag{3.58}$$

Similar to the deterministic case, we want to bound the term $\|w_{k-1} - w_k\|^2$ by $\|z_k - w_k\|^2$. For $k = 0$ we have

$$-\frac{1}{2}\|z_0 - w_0\|^2 + \frac{1}{2}(1+\beta)L^2\gamma_0^2\|w_0 - w_{-1}\|^2 = -\frac{1}{2}\left(1 - (1+\beta)L^2\gamma_0^2\right)\|w_0 - w_{-1}\|^2,$$

while for $k \geq 1$, using (3.51), we get

$$\begin{aligned}\|z_k - w_k\|^2 &\geq \frac{1}{2}\|w_k - w_{k-1}\|^2 - \|z_k - w_{k-1}\|^2 \\ &\geq \frac{1}{2}\|w_k - w_{k-1}\|^2 - \gamma_{k-1}^2\|F(w_{k-1};\xi_{k-1}) - F(w_{k-2};\xi_{k-2})\|^2 \\ &\geq \frac{1}{2}\|w_k - w_{k-1}\|^2 \\ &\quad - 3\gamma_{k-1}^2\left(\|W_{k-1}\|^2 + \|W_{k-2}\|^2 + \|F(w_{k-1}) - F(w_{k-2})\|^2\right),\end{aligned}$$

where we obtained the last inequality by inserting $\pm F(w_{k-1})$ and $\pm F(w_{k-2})$ in the difference of the stochastic estimators and using Lemma 2.5.4 (iii). Summing

up (3.58) from $k = 0$ to $K - 1$ and using the two estimates above then gives

$$\begin{aligned}\sum_{k=0}^{K-1} \gamma_k h(w_k, z) \leq{}& \frac{1}{2}\|z_0 - z\|^2 + \sum_{k=0}^{K-1} \gamma_k \langle W_k, z - v_k\rangle + \sum_{k=0}^{K-1} \gamma_k \langle W_k, v_k - w_k\rangle \\ &+ (1+\beta^{-1}) \sum_{k=0}^{K-1} \gamma_k^2 \left(\|W_k\|^2 + \|Z_k\|^2\right) + \frac{3}{2}\sum_{k=1}^{K-1} \gamma_{k-1}^2 \left(\|W_{k-1}\|^2 + \|W_{k-2}\|^2\right) \\ &- \frac{1}{2}\left(1 - (1+\beta)L^2\gamma_0^2\right)\|w_0 - w_{-1}\|^2 \\ &- \frac{1}{2}\sum_{k=1}^{K-1}\left(\frac{1}{2} - (1+\beta)L^2\gamma_k^2\right)\|w_k - w_{k-1}\|^2 \\ &+ \frac{1}{2}\sum_{k=1}^{K-1} 3\gamma_{k-1}^2 \|F(w_{k-1}) - F(w_{k-2})\|^2 .\end{aligned}$$

Bounding the second term on the right hand side of the above inequality via Lemma 3.4.11 yields

$$\begin{aligned}\sum_{k=0}^{K-1} \gamma_k h(w_k, z) \leq{}& \|z_0 - z\|^2 + \frac{1}{2}\sum_{k=0}^{K-1} \gamma_k^2 \|W_k\|^2 + \sum_{k=0}^{K-1} \gamma_k \langle W_k, v_k - w_k\rangle \\ &+ (1+\beta^{-1}) \sum_{k=0}^{K-1} \gamma_k^2 \left(\|W_k\|^2 + \|Z_k\|^2\right) + \frac{3}{2}\sum_{k=1}^{K-1} \gamma_{k-1}^2 \left(\|W_{k-1}\|^2 + \|W_{k-2}\|^2\right) \\ &- \frac{1}{2}\left(1 - (1+\beta)L^2\gamma_0^2\right)\|w_0 - w_{-1}\|^2 \\ &- \frac{1}{2}\sum_{k=1}^{K-1}\left(\frac{1}{2} - (1+\beta)L^2\gamma_k^2\right)\|w_k - w_{k-1}\|^2 \\ &+ \frac{1}{2}\sum_{k=1}^{K-1} 3\gamma_{k-1}^2 \|F(w_{k-1}) - F(w_{k-2})\|^2 .\end{aligned}$$

After taking the supremum over $z \in B$ and expectation, facilitating (3.57) we derive

$$\begin{aligned}\mathbb{E}\left[\sup_{z\in B}\left\{\sum_{k=0}^{K-1} \gamma_k h(w_k, z)\right\}\right] \leq{}& D^2 + \left(\frac{1}{2} + 2(1+\beta^{-1}) + 3\right)\sigma^2 \sum_{k=0}^{K-1} \gamma_k^2 \\ &- \frac{1}{2}\left(1 - (1+\beta)L^2\gamma_0^2\right)\mathbb{E}\|w_0 - w_{-1}\|^2 \\ &- \frac{1}{2}\sum_{k=1}^{K-1}\left(\frac{1}{2} - (1+\beta)L^2\gamma_k^2\right)\mathbb{E}\|w_k - w_{k-1}\|^2 \\ &+ \frac{1}{2}\sum_{k=1}^{K-1} 3L^2\gamma_{k-1}^2 \mathbb{E}\|w_{k-1} - w_{k-2}\|^2 .\end{aligned}$$

Let $0 < \gamma_k \leq \gamma < 1/2\sqrt{2}L$ for all $k \geq 0$ and let $\beta > 0$ such that

$$\gamma = \frac{1}{\sqrt{2(4+\beta)}L}, \tag{3.59}$$

then for $k \geq$ we have

$$\frac{1}{2} - (1+\beta)L^2\gamma_k^2 \geq 3L^2\gamma_k^2 > 0. \tag{3.60}$$

Using telescoping arguments we obtain

$$\begin{aligned}\mathbb{E}\left[\sup_{z\in B}\left\{\sum_{k=0}^{K-1}\gamma_k h(w_k, z)\right\}\right] &\leq D^2 + 6(1+\beta^{-1})\sigma^2\sum_{k=0}^{K-1}\gamma_k^2 \\ &\quad - \frac{1}{2}\left(1 - (1+\beta)L^2\gamma_0^2 - 3L^2\gamma_0^2\right)\mathbb{E}\left\|w_0 - w_{-1}\right\|^2 \\ &\quad - \frac{1}{2}\left(\frac{1}{2} - (1+\beta)L^2\gamma_{K-1}^2\right)\mathbb{E}\left\|w_{K-1} - w_{K-2}\right\|^2 \\ &\leq D^2 + 6(1+\beta^{-1})\sigma^2\sum_{k=0}^{K-1}\gamma_k^2,\end{aligned} \tag{3.61}$$

where we used (3.60) in the last inequality twice to bound the remaining terms. Furthermore, from (3.59) we derive

$$\beta^{-1} = \frac{2\gamma^2L^2}{1 - 8\gamma^2L^2}.$$

Plugging this estimate into (3.61), dividing by $\sum_{k=0}^{K-1}\gamma_k$ and applying Lemma 3.4.10 deduces the final statement.

□

3.4.5. Numerical experiments

To finish this section we look at a bilinear minimax problem to showcase the convergence rates that we have derived for the deterministic methods, and address the limitations of *Gradient Descent Ascent* (GDA) which is still used as status quo for many machine learning applications.

Bilinear saddle point problem

Following [69, 105] and others we consider the one dimensional bilinear example to illustrate the cycling behaviour of (even bilinear) minimax problems. We augment this approach by adding a nonsmooth ℓ_1-regulariser for one player, with $\kappa > 0$, and constraints for the other, resulting in

$$\min_{x\in\mathbb{R}}\max_{y\in\mathbb{R}}\ \kappa\left|x\right| + xy - \delta_{[-1,1]}(y). \tag{3.62}$$

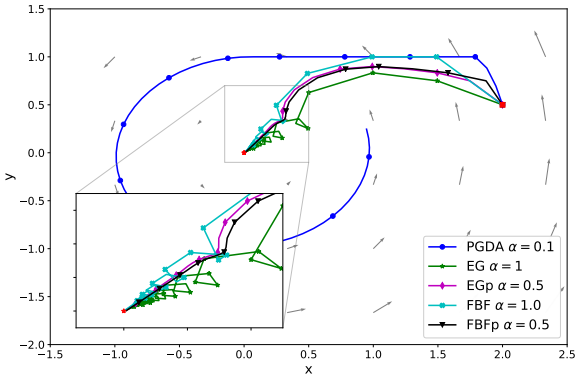


Figure 3.5.: Trajectories of the methods presented in Subsection 3.4.1 converging to a solution of problem (3.62).

For our experiments we chose $\kappa = 0.01$ and used $B = [-1, 1] \times [-1, 1]$ as the auxiliary set to compute the restricted gap function. We compared the methods presented in Subsection 3.4.1 with the proximal extension of GDA, called *Proximal Gradient Descent Ascent* (PGDA), which applied to problem (3.62) iterates for all $k \geq 0$

$$\text{PGDA:} \left\lfloor \begin{array}{l} x_{k+1} = \operatorname{prox}_{\kappa|\cdot|}(x_k - \gamma_k y_k) \\ y_{k+1} = P_{[-1,1]}(y_k + \gamma_k x_k)\,. \end{array} \right.$$

It is easy to see that the updates of PGDA are perpendicular to the iterates (up to the proximal evaluation), leading to oscillatory/diverging behaviour. By "EGp" we denote the method presented in [64] as *extrapolation from the past*.

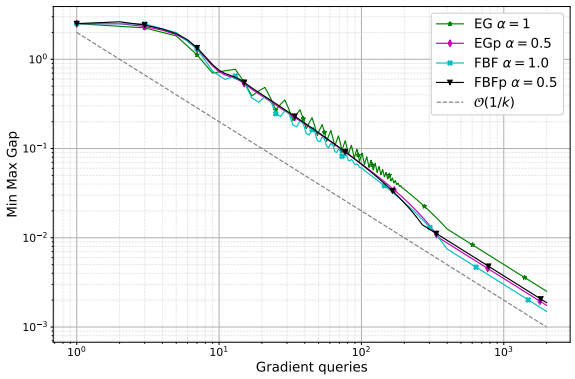


Figure 3.6.: Restricted gap function of the methods presented in Subsection 3.4.1 exhibiting a convergence rate of $\mathcal{O}(1/k)$.

The aforementioned issue of GDA and its proximal extension PGDA – cycling around the solution – is highlighted in Figure 3.5 where we can see that it fails to converge. The other methods, for which we display the averaged iterates, however do converge to a solution and show a decrease in the restricted gap according to Theorem 3.4.5 which can be seen in Figure 3.6. Interestingly, the two methods using extrapolation from the past seem to exhibit a more monotone decrease than FBF or EG.

3.5. Generative adversarial networks (GANs)

Generative Adversarial (Artificial Neural) Networks (GANs) [70] constitute a powerful 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, producing for example unseen realistic images. 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 and generated samples and tries to correctly 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 [70] was cast as a two-player zero sum game, i.e., a minimax problem of the form (MM),

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n}\ \mathbb{E}_{\rho\sim p}[\log\left(D_y(\rho)\right)]+\mathbb{E}_{\zeta\sim q}[\log\left(1-D_y(G_x(\zeta))\right)],$$

where x and y is the parametrisation of the generator and discriminator, respectively, D_y is the probability how certain the discriminator is that the input is real, G_x is the fake sample produced by the generator, and p and q is the real and some (latent) random distribution, respectively.

Due to problems with vanishing gradients in the training of such models, a successful alternative formulation called *Wasserstein GAN* (WGAN) [8] has been proposed. In this case the minimisation tries to reduce the Wasserstein distance between the true distribution p and the one learned by the generator q. Reformulating this distance via the Kantorovich-Rubinstein duality leads to an inner maximisation over 1-Lipschitz functions which are then approximated via neural networks, yielding the saddle point problem

$$\min_{x\in\mathbb{R}^d}\ \max_{\substack{y\in\mathbb{R}^n\\ \|y\|_{\mathrm{Lip}}\leq 1}}\ \mathbb{E}_{\rho\sim p}[D_y(\rho)]-\mathbb{E}_{\zeta\sim q}[D_y(G_x(\zeta))].$$

Conventionally, GANs are trained using variants of (stochastic) *Gradient Descent Ascent* (GDA) which are known to exhibit oscillatory behaviour [106] and thus fail to

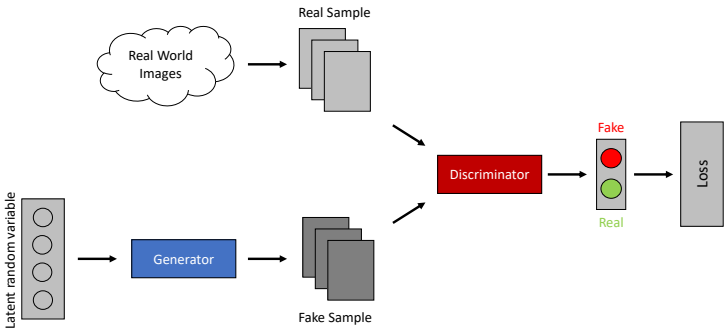


Figure 3.7.: Schematic illustration of Generative Adversarial Networks

converge even for simple bilinear saddle point problems, see [69]. We therefore propose the use of methods with provable convergence guarantees for (stochastic) convex-concave minimax problems, even though GANs are well known to not warrant these properties. Along similar considerations an adaptation of the *Extragradient method (EG)* [85] for the training of GANs was suggested in [64], whereas [54, 55, 89] studied *Optimistic Gradient Descent Ascent* (OGDA) based on *optimistic mirror descent* [126, 127]. We however investigate the *Forward-Backward-Forward (FBF)* method [137] from the previous sections, which, similar to EG, uses two gradient evaluations per update in order to circumvent the aforementioned issues but requires less projection/proximal steps per iteration.

Instead of trying to improve GAN performance with new architectures, loss functions, etc., we want to improve their training by contributing to the study of minimax problems. While the landscape of GAN training is far from matching the rigorous setting of monotonicity, the nonconvex-nonconcave setting remains either intractable in its full generality [56] or other simplifying assumption or other simplifying assumptions have to be made, for example the existence of Minty solutions [90, 103], which are again related to monotonicity.

We also want to point out that while extrapolation / optimistic steps are able to combat some of the oscillatory behaviour of minimax problems, another set of problems arises from the stochastic noise of the gradient evaluations which can be dealt with variance reduction techniques [45, 80] or reducing the step size (as we do). Other considerations have proposed to use the same sample in the for the two gradient evaluations leading to improved empirical performance [107]. Another approach was explored in [79] to use very different step sizes for the descent and ascent steps respectively (something that has been observed to be very beneficial in practice as well). Known methods were outfitted with *negative momentum* to improve stability in [65]. An additional technique proposed to deal with the oscillatory behaviour of minimax problems called *crossing-the-curl* was suggested in [63], whose authors used second order information to take a step perpendicular to the direction of rotation.

The aim of this section is to show how the use of methods with convergence guarantees, albeit only in the monotone setting, can yield better training performance for different architectures and objectives. In particular, we demonstrate that (inertial) FBF can perform at least as good as EG although requiring less evaluations of the regularisers. Note that in absence of any constraints or regularisation on the weights of at least one of the two networks, EG and FBF lead to equivalent updates of the iterates.

3.5.1. Wasserstein GANs trained on CIFAR-10

We now apply the proposed techniques from monotone inclusions to the training of Wasserstein GANs employing DCGAN [124] and ResNet [76] architectures. All models are trained on the CIFAR-10 dataset [86] which consists of 60,000 images in 10 different classes (with 50,000 training images and 10,000 test images) using an NVIDIA RTX 2080Ti GPU.

For the DCGAN experiments we work with the original WGAN formulation including weight clipping, since it includes regularisers innately (the indicator of a box for the weights of the discriminator). In addition we propose a modification of the WGAN formulation which replaces the box constraint on the discriminator's weights with an ℓ_1-regularisation, under the name of *WGAN-L1*. This results in a *soft thresholding* operation instead of the "harsh" clipping.

For the experiments on ResNet we use the WGAN-GP formulation [72] which penalises the norm of the gradient of the discriminator to enforce the Lipschitz constraint, together with spectral normalisation of the weight matrices [108] which has so far rather been considered as part of the architecture, but could at the same time (not rigorously) be seen as a regularisation step or as a projection onto the set of matrices with spectral norm less than 1.

	WGAN	
Method	IS	FID
AltAdam1	4.12±.06	56.44±.62
APD Adam	4.20±.04	53.97±.28
Extra Adam	4.07±.05	56.67±.61
IFBF Adam	**4.59**±.04	**45.25**±.60
FBF Adam	**4.54**±.04	**45.85**±.35
Opt. Adam	4.35±.06	50.41±.46

Table 3.4.: The best Inception Score (IS)[1] and Fréchet Inception Distance (FID) for the standard formulation *WGAN* with weight clipping.

Given the ubiquity and dominance of Adam [84] as an optimiser for many deep learning related training tasks, instead of using vanilla SGD we opt for Adam updates. This results in a method we call *FBF Adam*. Analogous approaches have been applied in [64]

and [54] resulting in *Extra Adam* and *Optimistic Adam*, respectively. We compare the aforementioned methods with the status-quo in GAN training, namely alternating one Adam step for each network: *AltAdam1*. For our experiments on the WGAN objective with weight clipping, we additionally considered two more methods – a variant of FBF with inertial effects, *IFBF Adam* ($\alpha = 0.05$), as well as a primal-dual algorithm for saddle point problems introduced by [75], *APD Adam*, which to the best of our knowledge has not been used for GAN training before. In general one would need stochastic versions of the used algorithms and which we do not provide in the case of RIFBF. A stochastic variant of the relaxed inertial FBF algorithm has been proposed and analysed in [53], where the convergence analysis of the stochastic numerical method massively relies on the one carried out for RIFBF.

Our hyperparameter search was limited to the step sizes for the newly introduced methods, while all other parameters were kept the same as in [64]. Note that we still report different values for the IS because we used the updated implementation [16]. It seems noteworthy that in the case of soft thresholding bigger step sizes performed better with the only exception of AltAdam1.

	WGAN-L1		WGAN-GP	
Method	IS	FID	IS	FID
AltAdam1	4.43±.03	50.86±2.17	6.01±.31	28.11±3.65
Extra Adam	4.67±.11	47.24±1.21	**6.58**±.08	21.40±.58
FBF Adam	**4.68**±.16	**46.60**±.76	6.57±.10	**21.22**±1.29
Opt. Adam	4.63±.13	47.98±1.49	6.42±.10	23.01±.95

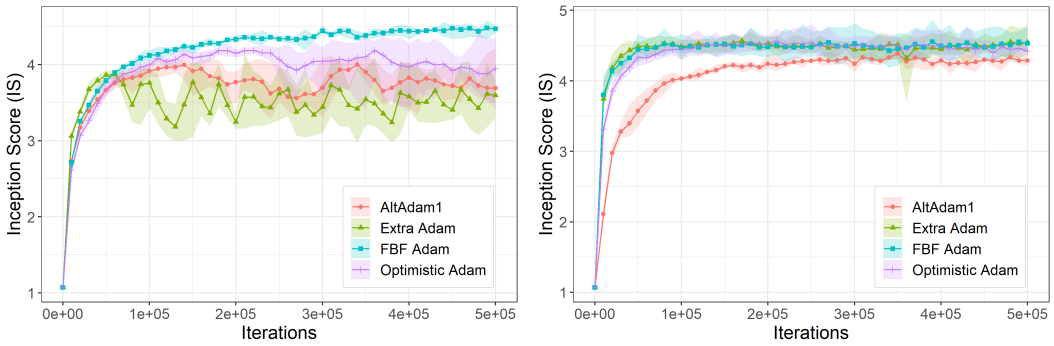
Table 3.5.: The best Inception Score (IS)[1] and Fréchet Inception Distance (FID). The column denoted by *WGAN-L1* and *WGAN-GP* refers to our regularised implementation using the 1-norm and the formulation with gradient penalty and spectral normalisation, respectively.

The two evaluation metrics used are the Inception Score (IS, higher is better) [133] and the Fréchet inception distance (FID, lower is better) [77], both computed on 50,000 samples. All reported values are averaged over 5 runs, where each consisted of 500,000 updates of the generator.

In Table 3.4 and Table 3.5 the best IS[1] and FID for each method are reported. FBF Adam performs at least as good as all considered competitors with respect to both evaluation metrics for WGAN-L1 and WGAN-GP, while in the case of the original WGAN objective with weight clipping FBF Adam and its inertial variant clearly outperform all other considered methods.

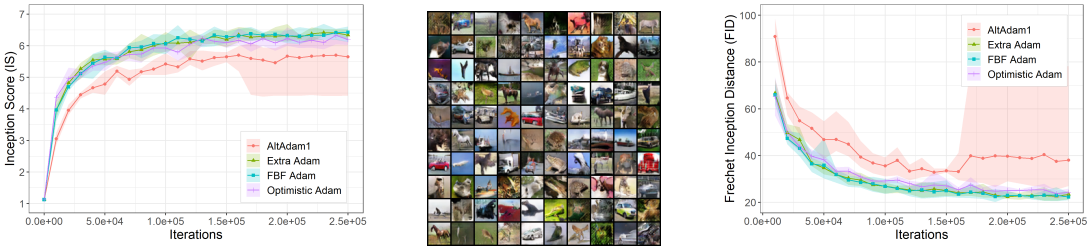
We see that even though we have only proved convergence for the monotone case in the deterministic setting, the variants of RIFBF even perform well in the training of GANs. As the theory suggests, making use of some inertial effects (regardless that Adam already

[1] In the case of the IS we use the updated and corrected implementation from [16], leading to lower reported values.



(a) IS for WGAN (weight clipping). (b) IS for WGAN-L1 (proximal operator).

Figure 3.8.: **DCGAN performance.** (a): Average and best/worst IS[1] on the WGAN objective with weight clipping. (b): Average and best/worst IS on the WGAN-L1 objective using the proximal operator (soft thresholding); The WGAN-L1 objective improves the IS in comparison to weight clipping and stabilises the behavior of all considered methods during the training procedure.



(a) Inception Score. (b) Generator sample. (c) FID.

Figure 3.9.: **ResNet performance**. Average and best/worst results regarding IS[1] and FID, see (a) and (b) respectively, using ResNet architecture on the WGAN-GP objective including spectral normalisation. (c): Samples from the generator trained with FBF Adam.

incorporates some momentum) seems to provide additional improvement of the numerical method in practice – at least for the (from an engineering point of view) modest original WGAN formulation.

One can also see that WGAN-L1 using the proximal operator improves the performance of all considered methods, decreasing the absolute and relative differences. Note that the results of the experiments with WGAN-L1 objective are comparable for the three methods with underlying convergence guarantees in the convex-concave case. Figure 3.8 shows the training progress regarding IS for each method and both problem formulations. The graphs suggest that making use of WGAN-L1 objective has a stabilising effect during training, leading to a smoother and more consistent learning curve

	WGAN-GP		
Method	non avg.	EMA	best
AltAdam1	28.11±3.65	—	—
Extra Adam	21.40±.58	18.29±.26	17.86
FBF Adam	**21.22**±1.29	**17.58**±.73	**16.65**
Opt. Adam	23.01±.95	21.00±1.33	19.28

Table 3.6.: Fréchet Inception Distance (FID) of regular iterates and averaged iterates using an exponential moving average (EMA), from training a ResNet with WGAN-GP formulation.

— a property that only FBF Adam seems to exhibit for weight clipping. Figure 3.9 as well as Table 3.5 show that for the WGAN-GP formulation FBF Adam maintains the improved performance of EG compared to GDA, while only requiring half the amount of spectral normalisations, resulting in time savings of up to 10% as reported in [108]. An even greater improvement by using FBF can be seen in Table 3.6 where we additionally consider the averaged iterates using an exponential moving average (EMA), which has already been observed to potentially have very positive effects [64].

In conclusion, the results suggest that employing methods that are designed to capture the nature of a problem, in this case a constrained/regularised minimax problem, is highly beneficial.

4. A fast optimistic method for variational inequalities

In this chapter we study numerical methods for solving variational inequalities that can be obtained from constrained minimax problems. We propose a novel algorithm that uses only one gradient/operator evaluation and one projection onto the constraint set per iteration, which in the unconstrained setting recovers the (explicit) "fast OGDA" (fOGDA) method [31] as a special case. It achieves a $o(1/k)$ rate of convergence in terms of the restricted gap function as well as the natural residual for the *last iterate*. To complete the theoretical results we provide a convergence guarantee for the (last) iterate to a solution of the variational inequality. To validate our method, which we call *fOGDA-VI*, we investigate a two-player matrix game with mixed strategies of the two players. Concluding, we show promising results regarding the application of fOGDA-VI to the training of generative adversarial nets.

4.1. Introduction

Let $\mathcal{H}$ be a real Hilbert space, let $F : \mathcal{H} \to \mathcal{H}$ be a monotone and L-Lipschitz continuous operator and let C be a nonempty closed convex subset of $\mathcal{H}$. Then the (strong) classical variational inequality problem (see Example 3.1.2) consists of finding $z^* \in \mathcal{H}$ such that

$$\langle F(z^*), z - z^* \rangle \geq 0 \quad \forall z \in C. \tag{4.1}$$

For the following considerations we assume that the solution set of (4.1) is nonempty, i.e., $\Omega := \{z \in C \mid \langle F(z), w - z \rangle \geq 0 \quad \forall w \in C\} \neq \emptyset$. Denoting the normal cone of C by N_C, condition (4.1) is equivalent to

$$0 \in F(z^*) + N_C(z^*). \tag{4.2}$$

Both (4.1) and its associated constrained saddle point problem are fundamental models in various fields such as optimisation, economics, game theory, and partial differential equations. Recently, they have attracted significant attention, in particular in the area of machine learning due to the fundamental role they play, for instance, in multi agent reinforcement learning [116], robust adversarial learning [96] and the training of generative adversarial networks (GANs) [70].

The aim of this work is to introduce an accelerated first order method for solving the constrained variational inequality problem (4.1). To this purpose, we first recall suitable measures that are commonly used to judge the quality of prospective solutions. We

will show that our proposed algorithm exhibits a better rate of convergence compared to other accelerated algorithms, while still maintaining the property that the generated sequence converges to a solution which is not necessarily the case for other accelerated methods [40, 42, 113].

4.1.1. Convergence measures

There are several ways to measure the convergence to a solution of (4.1) in the literature. For this let us first introduce some notation that is necessary in the following.

For $z \in \mathcal{H}$ and $\delta > 0$ we denote $\mathbb{B}(z;\delta) := \{w \in \mathcal{H} \mid \|w - z\| \leq \delta\}$. We write $P_C(z)$ to denote the projection of z onto a closed convex subset C of $\mathcal{H}$, which is uniquely defined.

Restricted gap function For z^*, $z_0 \in \Omega$ and $\delta(z_0) := \|z^* - z_0\|$, the *restricted gap function* [114], see also Section 3.4.2, associated with the variational inequality (4.1) is defined via

$$\texttt{Gap}(z) := \sup_{w \in C \cap \mathbb{B}(z^*;\delta(z_0))} \langle F(w), z - w \rangle \geq 0.$$

The restriction of the supremum to a bounded set, particularly the ball $\mathbb{B}(z^*;\delta(z_0))$ in our case, is essential to avoid an infinitely large gap when C is unbounded.

Tangent residual Another quantity that can be used to measure the quality of a candidate with respect to the variational inequality (4.1) is based on the observation that it is equivalent to the monotone inclusion (4.2). The so-called *tangent residual* is given by

$$r(z) := \inf_{\zeta \in N_C(z)} \|F(z) + \zeta\| .$$

In a straightforward way this quantity extends the usual measure $\|F(z)\|$ in the unconstrained setting of monotone equations, where the goal is to find $z^* \in \mathcal{H}$ such that

$$F(z^*) = 0, \tag{4.3}$$

to variational inequalities by measuring the distance of z being a solution and satisfying (4.2). Note, if $z \notin C$ we have $N_C(z) = \emptyset$ and thus $r(z) = +\infty$.

To also have a meaningful measure for elements not contained in the constraint set we can look at the following quantity.

Natural residual Another useful convergence measure is the *natural residual*, which in fact is upper bounded by the tangent residual. Using the characterisation of the projection via the normal cone (see [17, Proposition 6.45]) we observe that

$$0 \in F(z^*) + N_C(z^*) \quad \Leftrightarrow \quad z^* = P_C\left[z^* - F(z^*)\right].$$

This motivates to look at

$$\texttt{Res}(z) := \|z - P_C\left[z - F(z)\right]\|,$$

which is also known as *fixed point residual*. To see that indeed $\texttt{Res}(z) \leq r(z)$ for all $z \in \mathcal{H}$, it is sufficient to only consider the case when $z \in C$ as otherwise the inequality trivially holds. Let $z \in C$ and $\zeta \in N_C(z)$, then by the following equivalence (see [17, Proposition 6.45]),

$$p = P_C(v) \quad \Leftrightarrow \quad \exists \zeta \in N_C(p) \text{ such that } v - p = \zeta, \tag{4.4}$$

we obtain that $z = P_C[z + \zeta]$. Since the projection onto a nonempty closed convex set is nonexpansive, we then deduce that

$$\texttt{Res}(z) = \|z - P_C\left[z - F(z)\right]\| = \|P_C\left[z + \zeta\right] - P_C\left[z - F(z)\right]\| \leq \|F(z) + \zeta\|\,. \tag{4.5}$$

Since $\zeta \in N_C(z)$ was arbitrary, we conclude that $\texttt{Res}(z) \leq r(z)$. Note, in the unconstrained case (4.3) the two residuals coincide and are given by

$$r(z) = \texttt{Res}(z) = \|F(z)\|\,.$$

4.1.2. Related works

In this work we are interested exclusively in first order methods that are fully splitting, i.e., algorithms that only use projections onto C and direct evaluations of the operator F as main building blocks. As a general Lipschitz continuous operator is not necessarily cocoercive, the simplest first order method that is splitting – the Forward-Backward (FB) algorithm – can not be used to solve (4.1).

Extragradient (EG) method [6, 85] Korpelevich and Antipin proposed to take a second forward evaluation of F in each iteration in order to solve (4.1). This results in the following scheme

$$\text{EG:} \left\lfloor \begin{array}{l} w_k = P_C\left[z_k - \gamma F(z_k)\right], \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k)\right], \end{array} \right. \qquad \forall k \geq 0 \tag{4.6}$$

which converges to a solution for $0 < \gamma < 1/L$.

It is known that EG converges with a rate of $\mathcal{O}(1/K)$ in terms of the restricted gap function for the *averaged*, or *ergodic*, iterates

$$\bar{w}_K := \frac{1}{K} \sum_{k=1}^{K} w_k$$

in both the unconstrained [109, 112, 114] and the constrained case [78], which further seems to be optimal [118]. The *best iterate* convergence in terms of the tangent residual, however, is known to yield a rate of $\mathcal{O}(1/\sqrt{K})$ [60, 85], i.e.,

$$\min_{1 \leq k \leq K} r(z_k) = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \quad \text{as } k \to +\infty.$$

The more desirable *last iterate* convergence rate for EG was derived only recently in the unconstrained case [71] which was then extended to the constrained case as well [41]. In fact,

$$\texttt{Gap}(z_k) = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right) \quad \text{and} \quad \texttt{Res}(z_k) \leq r(z_k) = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right),$$

as $k \to +\infty$. The result for the restricted gap function [68] as well as for the residuals is actually tight, meaning that the convergence rate for the averaged iterates is better than for the last iterate. Nevertheless, we emphasize that the latter one is more appealing since the averaged iterates might still show acceptable behaviour while the actual trajectory of iterates cycles around the set of solutions periodically [104].

Popov's method [123] In the saddle point setting Popov [123] introduced the following algorithm which, when applied to (4.2), reads

$$\text{Popov:} \left\lfloor \begin{array}{l} w_k = P_C\left[z_k - \gamma F(w_{k-1})\right] \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k)\right] \end{array} \right. \quad \forall k \geq 0, \tag{4.7}$$

which converges to a solution for $0 < \gamma < 1/2L$. The update rule of (4.7) is very similar to (4.6) but requires F to be evaluated only once per iteration. Actually, Popov can be obtained from EG if we replace $F(z_k)$ by $F(w_{k-1})$ in the first line of (4.6). In the unconstrained case (4.7) can be written in one line, yielding a method usually known as Optimistic Gradient Descent Ascent (OGDA), a name that was coined by works on GAN training [54, 55].

Given the close connection between EG and OGDA, it is not surprising that many convergence rate results hold in a similar way. In terms of the restricted gap OGDA converges like $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$ for averaged iterates [109] and last iterates [68], respectively, where the latter one is optimal. The convergence rate in terms of the residuals is $\mathcal{O}(1/\sqrt{k})$ and it can not be improved in general [41, 46, 67], as seen for EG.

We have seen that both EG and Popov's method require two projections in each iteration. One might think that this is necessary to obtain convergent algorithms for the variational inequality (4.2) when the operator F is merely Lipschitz continuous and not cocoercive. This is not the case, however, and in the following we will look at algorithms that need only one evaluation of the projection operator per iteration.

Forward-Backward-Forward (FBF) method [137] One of these single-call projection methods is the FBF method. It was proposed by Tseng [137] and applied to (4.2) it iterates

$$\text{FBF:} \left\lfloor \begin{array}{l} w_k = P_C\left[z_k - \gamma F(z_k)\right] \\ z_{k+1} = w_k - \gamma\left(F(w_k) - F(z_k)\right) \end{array} \right. \quad \forall k \geq 0, \tag{4.8}$$

which converges to a solution for $0 < \gamma < 1/L$. Notice that in each iteration this iterative scheme performs two evaluations of F along the sequences $(z_k)_{k\geq 0}$ and $(w_k)_{k\geq 0}$, similar to EG – the first line of FBF and EG is even identical. In the second line, however,

instead of performing another projection the forward step regarding the intermediate iterate w_k is corrected by the previous update $F(z_k)$. Moreover, for the unconstrained problem (4.3), i.e., in the absence of projections, FBF and EG are equivalent.

Forward-Reflected-Backward (FRB) method [100] Another single-call projection method that even requires only one evaluation of F like the basic FB algorithm was proposed by Malitsky and Tam [100]. The FRB algorithm is given by

$$\text{FRB:} \left\lfloor \; z_{k+1} = P_C\left[z_k - 2\gamma F(z_k) + \gamma F(z_{k-1})\right] \qquad \forall k \geq 0, \right. \tag{4.9}$$

and converges to a solution for $0 < \gamma < 1/2L$. Note that FRB can be deducted from FBF by reusing $F(w_{k-1})$ instead of $F(z_k)$ in the first line of (4.8), similarly to how Popov's method can be obtained from EG. Hence (4.9) coincides with (4.7) and OGDA in the unconstrained case.

Projected Reflected Gradient (RG) method [98] Before investigating FRB, Malitsky introduced another similar method where the order of the reflection and the forward step is reversed [98]. In particular, a second forward step can be avoided by evaluating F at an appropriate linear combination of the iterates. This gives rise to the following method

$$\text{RG:} \left\lfloor \begin{array}{l} w_k = 2z_k - z_{k-1} \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k)\right] \end{array} \qquad \forall k \geq 0, \right. \tag{4.10}$$

which converges to a solution for $0 < \gamma < (\sqrt{2}-1)/L$.

Despite the similarities in the construction and iterate convergence, nonasymptotic convergence is less understood in the case of single-call projection methods. For example Banert and Boţ [15] derived for Tseng's method an ergodic $\mathcal{O}(1/k)$ rate in terms of function values in the context of convex optimisation; see also [20] for an ergodic $\mathcal{O}(1/\sqrt{k})$ convergence result in terms of the restricted gap function in the stochastic setting. For Malitsky's RG algorithm convergence in terms of the gap function and residuals like $\mathcal{O}(1/\sqrt{k})$ for the last iterate was established recently [42].

4.1.3. Accelerated methods

An accelerated algorithm for solving (4.3) that is based on EG, called Extra Anchored Gradient (EAG) algorithm, was proposed by Yoon and Ryu [140]. It is designed by using anchoring, a technique that can be traced back to Halpern's algorithm [74]. This iterative scheme exhibits a convergence rate of

$$\|F(z_k)\| = \mathcal{O}\left(\frac{1}{k}\right) \quad \text{as } k \to +\infty.$$

These considerations have been followed by extension of EAG to the constrained setting [40], where the authors consider

$$\text{EAG:} \left| \begin{array}{l} w_k = P_C\left[z_k - \gamma F(z_k) + \dfrac{1}{k+1}(z_0 - z_k)\right] \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k) + \dfrac{1}{k+1}(z_0 - z_k)\right] \end{array} \right. \qquad \forall k \geq 0,$$

with $0 < \gamma < 1/\sqrt{3}L$. One can notice that the algorithm uses two operator evaluations and two projection steps, like EG, and in the unconstrained case it coincides with its projection free counterpart [140], maintaining the $\mathcal{O}(1/k)$ convergence rate for gap function and residuals.

Similar to the nonaccelerated methods from the previous subsection, one can also reduce the number of necessary operator evaluations and projections to one each per iteration. This was done by investigating an accelerated version [42] of the projected reflected gradient method (4.10) with convergence rate $\mathcal{O}(1/k)$. The method is given by

$$\text{ARG:} \left| \begin{array}{l} w_k = 2z_k - z_{k-1} + \dfrac{1}{k+1}(z_0 - z_k) - \frac{1}{k}(z_0 - z_{k-1}) \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k) + \dfrac{1}{k+1}(z_0 - z_k)\right] \end{array} \right. \qquad \forall k \geq 0,$$

with $0 < \gamma \leq 1/12L$.

It is worth mentioning that for constrained EAG [40] and ARG [42] there are no guarantees for the iterates to converge to a solution.

Further variants of anchoring based algorithms have been proposed by Tran-Dinh [135] and together with Luo [136], which all exhibit the same convergence rate in terms of the operator norm $\left\|F(z^k)\right\|$ as EAG for (4.3). Monotone inclusions are also considered in these works, with a more general operator than N_C, but this requires either taking a backward step or additionally asking for cocoercivity of F. In the same spirit, an $o(1/k)$ rate of convergence together with convergence of iterates was shown for an accelerated version of the Krasnosel'skiĭ-Mann algorithm [35].

Tran-Dinh [135] as well as Park and Ryu [119] pointed out the existence of some connections between the anchoring approach and Nesterov's acceleration technique used for the minimisation of smooth and convex functions [113, 115].

A different approach, as it appears, from the Halpern-type methods mentioned above was investigated in [31]. An appropriate (explicit) discretisation of a second-order dynamical system gives rise to an accelerated algorithm related to OGDA, called *fast OGDA* (fOGDA), which will constitute a starting point for our considerations in the following. The fast OGDA algorithm [31] for solving the monotone equation (4.3) is given by

$$\text{fOGDA:} \left| \begin{array}{l} w_k = z_k + \dfrac{k}{k+\alpha}\Big(z_k - z_{k-1}\Big) - \gamma\dfrac{\alpha}{k+\alpha}F(w_{k-1}) \\ z_{k+1} = w_k - \gamma\left(1 + \dfrac{k}{k+\alpha}\right)\Big(F(w_k) - F(w_{k-1})\Big) \end{array} \right. \qquad \forall k \geq 1,$$

which converges for $0 < \gamma < 1/4L$ and $\alpha > 0$ large enough. It was shown that fOGDA exhibits convergence like

$$\texttt{Gap}(z_k) = o\left(\frac{1}{k}\right) \quad \text{and} \quad \|F(z_k)\| = o\left(\frac{1}{k}\right) \quad \text{as } k \to +\infty.$$

4.2. Main results

In this section, we will first motivate the necessary changes how to extend fOGDA to solve the variational inequality (4.1) and introduce our newly proposed method which we call *fOGDA-VI*. This is followed by formally stating the convergence results of fOGDA-VI – convergence of the iterates to a solution as well as convergence in terms of the restricted gap and the residuals like $o(1/k)$ for the *last* iterate.

4.2.1. Extending fOGDA to the constraint case

It can be seen empirically, that incorporating one (or more) projections into fOGDA in a naive way is not sufficient to obtain a solution method for (4.1). Instead, the idea is to introduce for all $k \geq 1$ an appropriate element ζ_k of the normal cone $N_C(z_k)$. This is done by replacing $F(w_{k-1})$ by the sum $F(w_{k-1}) + \zeta_k$. One might think that because of this the suitable choice would be to take $\zeta_k \in N_C(w_{k-1})$, but this is not the case which can be motivated as follows. For the unconstrained case, i.e., when solving the monotone equation (4.3), we want to find a sequence $(z_k)_{k\geq 1}$ yielding $\|F(z_k)\| \to 0$. However, in the constrained case, i.e., when tackling the monotone inclusion (4.2), we aim to establish sequences $(z_k)_{k\geq 1}$, $(\zeta_k)_{k\geq 1}$ with $\zeta_k \in N_C(z_k)$ such that

$$\|F(z_k) + \zeta_k\| \to 0.$$

Then instead of fOGDA we have an algorithm of the form

$$\begin{aligned} w_k &= z_k + \frac{k}{k+\alpha}\Big(z_k - z_{k-1}\Big) - \gamma\frac{\alpha}{k+\alpha}\Big(F(w_{k-1}) + \zeta_k\Big), \\ z_{k+1} &= w_k - \gamma\left(1 + \frac{k}{k+\alpha}\right)\Big(F(w_k) - F(w_{k-1}) + \zeta_{k+1} - \zeta_k\Big). \end{aligned} \tag{4.11}$$

At first glance this method seems to be implicit, as both z_{k+1} and ζ_{k+1} appear on the same line. The expression (4.11) can be used to formulate a projection step via (4.4), though.

These considerations give rise to the following algorithm.

Algorithm 4.2.1. Let $\alpha > 2$. For starting points z_0, $w_0 \in \mathcal{H}$, $z_1 \in C$, $\zeta_1 \in N_C(z_1)$ and

fixed step size $0 < \gamma < 1/4L$ we consider for all $k \geq 1$

$$\text{fOGDA-VI:} \left| \begin{array}{l} w_k = z_k + \dfrac{k}{k+\alpha}\Big(z_k - z_{k-1}\Big) - \gamma\dfrac{\alpha}{k+\alpha}\Big(F(w_{k-1}) + \zeta_k\Big), \\ z_{k+1} = P_C\left[w_k - \gamma\left(1 + \dfrac{k}{k+\alpha}\right)\Big(F(w_k) - F(w_{k-1}) - \zeta_k\Big)\right], \\ \zeta_{k+1} = \dfrac{k+\alpha}{\gamma(2k+\alpha)}\Big(w_k - z_{k+1}\Big) - \Big(F(w_k) - F(w_{k-1}) - \zeta_k\Big). \end{array} \right.$$

4.2.2. Convergence statements

The first main result concerns the convergence of the sequence of iterates to an element in Ω.

Theorem 4.2.2. *Let $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 4.2.1. Then the sequence $(z_k)_{k\geq 0}$ converges weakly to a solution of* (4.1).

The central idea for the proof is to define an appropriate nonincreasing energy function to obtain convergence or summability of various helpful quantities. Even though we are not able to enforce the discrete energy $(\mathcal{G}_{\lambda,k})_{k\geq 0}$, for suitable $\lambda \geq 0$, to have an actual nonincreasing property, we can at least show that for every $k \geq 0$

$$\mathcal{G}_{\lambda,k+1} \leq (1 + d_k)\mathcal{G}_{\lambda,k} - b_k, \tag{4.12}$$

with some nonnegative sequences $(b_k)_{k\geq 0}$, $(d_k)_{k\geq 0}$. The aim is to control these two sequences in such a way that we can still derive some beneficial asymptotic results of $(\mathcal{G}_{\lambda,k})_{k\geq 0}$. For instance, we show that there exist two constants $0 \leq \underline{\lambda}(\alpha) < \bar{\lambda}(\alpha)$ such that the inequality (4.12) holds with $\sum_{k\geq 0} d_k < +\infty$ for every $\lambda \in (\underline{\lambda}(\alpha), \bar{\lambda}(\alpha))$. This allows us to conclude that $\lim_{k\to+\infty} \mathcal{G}_{\lambda,k} \in \mathbb{R}$ exists, see Lemma 2.5.3 for more details.

From this we can then verify that the first condition of Opial's lemma, see Lemma 2.3.1, is fulfilled, while its second condition follows from the maximal monotonicity of $F + N_C$, see Proposition 2.1.3.

The asymptotic convergence of the iterates is complemented by statements about convergence rates in terms of the restricted gap as well as the natural residual for the last iterate.

Theorem 4.2.3. *Let $z^* \in \Omega$ be a solution of* (4.1) *and let $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 4.2.1. Then we have*

$$\texttt{Gap}(z_k) = o\left(\frac{1}{k}\right) \quad \textit{and} \quad \texttt{Res}(z_k) = o\left(\frac{1}{k}\right) \quad \textit{as } k \to +\infty.$$

4.3. Convergence analysis

In the following section we will establish the necessary convergence analysis to prove Theorem 4.2.2 and Theorem 4.2.3.

4.3.1. Notation and preliminary considerations

In the following for all $k \geq 1$ we will use the notation

$$v_k := F(w_{k-1}) + \zeta_k, \tag{4.13}$$

which we plug into Algorithm 4.2.1 to obtain

$$w_k = z_k + \frac{k}{k+\alpha}\Big(z_k - z_{k-1}\Big) - \gamma\frac{\alpha}{k+\alpha}v_k, \tag{4.14a}$$

$$z_{k+1} = w_k - \gamma\left(1 + \frac{k}{k+\alpha}\right)\Big(v_{k+1} - v_k\Big). \tag{4.14b}$$

Since $0 < \gamma < 1/4L$ there exists $0 < \varepsilon < 1$ such that

$$\gamma = \frac{1-\varepsilon}{4L}. \tag{4.15}$$

Hence, the definition of v_k together with the Lipschitz continuity of F give us

$$\begin{aligned}\|\zeta_{k+1} + F(z_{k+1}) - v_{k+1}\| &= \|F(z_{k+1}) - F(w_k)\| \leq L\|z_{k+1} - w_k\| \\ &\leq \gamma L\|v_{k+1} - v_k\| \leq \frac{1-\varepsilon}{4}\|v_{k+1} - v_k\| \leq \frac{1}{4}\|v_{k+1} - v_k\| \leq \|v_{k+1} - v_k\|.\end{aligned} \tag{4.16}$$

By summing up (4.14a) and (4.14b) we find

$$(k+\alpha)(z_{k+1} - z_k) - k(z_k - z_{k-1}) = -\alpha\gamma v_{k+1} - 2\gamma k(v_{k+1} - v_k). \tag{4.17}$$

Let $0 \leq \lambda \leq \alpha - 1$ and $z^* \in \Omega$. Following [31] we denote for every $k \geq 1$

$$u_{\lambda,k} := 2\lambda(z_k - z^*) + 2k(z_k - z_{k-1}) + \frac{3\alpha - 2}{\alpha - 1}\gamma k v_k, \tag{4.18}$$

$$\begin{aligned}\mathcal{E}_{\lambda,k} := \frac{1}{2}\|u_{\lambda,k}\|^2 &+ 2\lambda(\alpha - 1 - \lambda)\|z_k - z^*\|^2 + \frac{2(\alpha-2)}{\alpha-1}\lambda\gamma k\langle z_k - z^*, v_k\rangle \\ &+ \frac{\alpha-2}{\alpha-1}\gamma^2 k\left(\frac{3\alpha-2}{2(\alpha-1)}k + \alpha\right)\|v_k\|^2,\end{aligned} \tag{4.19}$$

$$\begin{aligned}\mathcal{G}_{\lambda,k} := \mathcal{E}_{\lambda,k} &- \frac{2(\alpha-2)}{\alpha-1}\gamma k^2\langle z_k - z_{k-1}, F(z_k) - F(w_{k-1})\rangle \\ &+ \frac{\alpha-2}{\alpha-1}\gamma^2 k\sqrt{k}\Big((1-\varepsilon)\sqrt{k} + \alpha\Big)\|v_k - v_{k-1}\|^2.\end{aligned} \tag{4.20}$$

4.3.2. Properties of the energy function

We collect the properties of the energy functions $(\mathcal{E}_{\lambda,k})_{k\geq 1}$ and $(\mathcal{G}_{\lambda,k})_{k\geq 1}$ in the following lemmas.

Lemma 4.3.1. *Let $z^* \in \Omega$ and $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 4.2.1. For $0 \leq \lambda \leq \alpha - 1$, the following identity holds for every $k \geq 1$*

$$\begin{aligned}
&\mathcal{E}_{\lambda,k+1} - \mathcal{E}_{\lambda,k} \\
&\quad = -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle \\
&\qquad + 2(\lambda+1-\alpha)(2k+\alpha+1)\|z_{k+1}-z_k\|^2 \\
&\qquad + \frac{2}{\alpha-1}\Big(4(\alpha-1)(\lambda+1-\alpha) - \alpha(\alpha-2)\Big)\gamma k\langle z_{k+1}-z_k, v_{k+1}\rangle \\
&\qquad + \frac{2}{\alpha-1}\Big(2\alpha(\alpha-1)(\lambda+1-\alpha) + \alpha - 2(\alpha-1)^2\Big)\gamma\langle z_{k+1}-z_k, v_{k+1}\rangle \\
&\qquad - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle - \frac{\alpha-2}{\alpha-1}\gamma^2 k(2k+\alpha)\|v_{k+1}-v_k\|^2 \\
&\qquad - \frac{\alpha-2}{\alpha-1}\gamma^2\left(2(3\alpha-2)k + 2\alpha^2+\alpha-2\right)\|v_{k+1}\|^2.
\end{aligned} \tag{4.21}$$

Proof. Recall that by the definition in (4.18), we have for every $k \geq 1$

$$u_{\lambda,k} = 2\lambda(z_k - z^*) + 2k(z_k - z_{k-1}) + \frac{3\alpha-2}{\alpha-1}\gamma k v_k$$

Similarly,

$$u_{\lambda,k+1} = 2\lambda(z_{k+1} - z^*) + 2(k+1)(z_{k+1} - z_k) + \frac{3\alpha-2}{\alpha-1}\gamma(k+1)v_{k+1}. \tag{4.22}$$

Thus, after substraction we deduce from (4.17) that

$$\begin{aligned}
&u_{\lambda,k+1} - u_{\lambda,k} \\
&\quad = 2(\lambda+1-\alpha)(z_{k+1}-z_k) + 2(k+\alpha)(z_{k+1}-z_k) - 2k(z_k - z_{k-1}) \\
&\qquad + \frac{3\alpha-2}{\alpha-1}\gamma v_{k+1} + \frac{3\alpha-2}{\alpha-1}\gamma k(v_{k+1}-v_k) \\
&\quad = 2(\lambda+1-\alpha)(z_{k+1}-z_k) + \frac{\alpha-2(\alpha-1)^2}{\alpha-1}\gamma v_{k+1} + \frac{2-\alpha}{\alpha-1}\gamma k(v_{k+1}-v_k).
\end{aligned} \tag{4.23}$$

For $k \geq 1$ we know that

$$\frac{1}{2}\left(\|u_{\lambda,k+1}\|^2 - \|u_{\lambda,k}\|^2\right) = \langle u_{\lambda,k+1}, u_{\lambda,k+1} - u_{\lambda,k}\rangle - \frac{1}{2}\|u_{\lambda,k+1} - u_{\lambda,k}\|^2, \tag{4.24}$$

and that for all $k \geq 0$

$$\begin{aligned}
&2\lambda(\alpha-1-\lambda)\left(\|z_{k+1}-z^*\|^2 - \|z_k - z^*\|^2\right) \\
&\quad = 4\lambda(\alpha-1-\lambda)\langle z_{k+1}-z^*, z_{k+1}-z_k\rangle - 2\lambda(\alpha-1-\lambda)\|z_{k+1}-z_k\|^2.
\end{aligned} \tag{4.25}$$

We use the relations (4.22) and (4.23) to derive for every $k \geq 1$

$$\begin{aligned}
&\langle u_{\lambda,k+1}, u_{\lambda,k+1} - u_{\lambda,k} \rangle \\
&\quad = 4\lambda(\lambda + 1 - \alpha) \langle z_{k+1} - z^*, z_{k+1} - z_k \rangle \\
&\qquad + \frac{2}{\alpha - 1} \Big(\alpha - 2(\alpha - 1)^2 \Big) \lambda\gamma \langle z_{k+1} - z^*, v_{k+1} \rangle \\
&\qquad - \frac{2(\alpha - 2)}{\alpha - 1} \lambda\gamma k \langle z_{k+1} - z^*, v_{k+1} - v_k \rangle + 4(\lambda + 1 - \alpha)(k+1) \|z_{k+1} - z_k\|^2 \\
&\qquad + \frac{2}{\alpha - 1} \Big((3\alpha - 2)(\lambda + 1 - \alpha) + \alpha - 2(\alpha - 1)^2 \Big) \gamma (k+1) \langle z_{k+1} - z_k, v_{k+1} \rangle \\
&\qquad - \frac{2(\alpha - 2)}{\alpha - 1} \gamma (k+1) k \langle z_{k+1} - z_k, v_{k+1} - v_k \rangle \\
&\qquad + \frac{1}{(\alpha - 1)^2} \Big(\alpha - 2(\alpha - 1)^2 \Big) (3\alpha - 2) \gamma^2 (k+1) \|v_{k+1}\|^2 \\
&\qquad - \frac{1}{(\alpha - 1)^2} (\alpha - 2)(3\alpha - 2) \gamma^2 (k+1) k \langle v_{k+1}, v_{k+1} - v_k \rangle ,
\end{aligned} \tag{4.26}$$

and

$$\begin{aligned}
-\frac{1}{2} \|u_{\lambda,k+1} - u_{\lambda,k}\|^2 &= -2(\lambda + 1 - \alpha)^2 \|z_{k+1} - z_k\|^2 \\
&\quad - \frac{2}{\alpha - 1} \Big(\alpha - 2(\alpha - 1)^2 \Big) (\lambda + 1 - \alpha) \gamma \langle z_{k+1} - z_k, v_{k+1} \rangle \\
&\quad - \frac{1}{2(\alpha - 1)^2} \Big(\alpha - 2(\alpha - 1)^2 \Big)^2 \gamma^2 \|v_{k+1}\|^2 - \frac{(\alpha - 2)^2}{2(\alpha - 1)^2} \gamma^2 k^2 \|v_{k+1} - v_k\|^2 \\
&\quad + \frac{2(\alpha - 2)}{\alpha - 1} (\lambda + 1 - \alpha) \gamma k \langle z_{k+1} - z_k, v_{k+1} - v_k \rangle \\
&\quad + \frac{\alpha - 2}{(\alpha - 1)^2} \Big(\alpha - 2(\alpha - 1)^2 \Big) \gamma^2 k \langle v_{k+1}, v_{k+1} - v_k \rangle .
\end{aligned} \tag{4.27}$$

After some algebra, we see that

$$\begin{aligned}
&\Big((3\alpha - 2)(\lambda + 1 - \alpha) + \alpha - 2(\alpha - 1)^2 \Big) (k+1) - (\lambda + 1 - \alpha) \Big(\alpha - 2(\alpha - 1)^2 \Big) \\
&\quad = \Big((3\alpha - 2)(\lambda + 1 - \alpha) + \alpha - 2(\alpha - 1)^2 \Big) k + 2\alpha(\alpha - 1)(\lambda + 1 - \alpha) \\
&\qquad + \alpha - 2(\alpha - 1)^2 \\
&\quad = \Big((3\alpha - 2)(\lambda + 1 - \alpha) + \alpha - 2(\alpha - 1)^2 + (\alpha - 2)\lambda \Big) k - (\alpha - 2)\lambda k \\
&\qquad + 2\alpha(\alpha - 1)(\lambda + 1 - \alpha) + \alpha - 2(\alpha - 1)^2 \\
&\quad = \Big(4(\alpha - 1)(\lambda + 1 - \alpha) - \alpha(\alpha - 2) \Big) k - (\alpha - 2)\lambda k \\
&\qquad + 2\alpha(\alpha - 1)(\lambda + 1 - \alpha) + \alpha - 2(\alpha - 1)^2 .
\end{aligned} \tag{4.28}$$

By plugging (4.26) and (4.27) into (4.24), and by taking into consideration the relation (4.28), we get for every $k \geq 1$

$$\begin{aligned}
\frac{1}{2}\left(\|u_{\lambda,k+1}\|^2 - \|u_{\lambda,k}\|^2\right) &= 4\lambda\left(\lambda+1-\alpha\right)\langle z_{k+1}-z^*, z_{k+1}-z_k\rangle \\
&+ \frac{2}{\alpha-1}\left(\alpha - 2\left(\alpha-1\right)^2\right)\lambda\gamma\langle z_{k+1}-z^*, v_{k+1}\rangle \\
&- \frac{2\left(\alpha-2\right)}{\alpha-1}\lambda\gamma k\langle z_{k+1}-z^*, v_{k+1}-v_k\rangle \\
&+ 2\left(\lambda+1-\alpha\right)\left(2k+\alpha+1-\lambda\right)\|z_{k+1}-z_k\|^2 \\
&+ \frac{2}{\alpha-1}\Big(4\left(\alpha-1\right)\left(\lambda+1-\alpha\right) - \alpha\left(\alpha-2\right) - \left(\alpha-2\right)\lambda\Big)\gamma k\langle z_{k+1}-z_k, v_{k+1}\rangle \\
&+ \frac{2}{\alpha-1}\Big(2\alpha\left(\alpha-1\right)\left(\lambda+1-\alpha\right) + \alpha - 2\left(\alpha-1\right)^2\Big)\gamma\langle z_{k+1}-z_k, v_{k+1}\rangle \\
&- \frac{2\left(\alpha-2\right)}{\alpha-1}\gamma k\left(k+\alpha-\lambda\right)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle - \frac{\left(\alpha-2\right)^2}{2\left(\alpha-1\right)^2}\gamma^2k^2\|v_{k+1}-v_k\|^2 \\
&+ \frac{1}{2\left(\alpha-1\right)^2}\left(\alpha-2\left(\alpha-1\right)^2\right)\gamma^2\left(2\left(3\alpha-2\right)k + 2\alpha^2+\alpha-2\right)\|v_{k+1}\|^2 \\
&- \frac{\alpha-2}{\left(\alpha-1\right)^2}\gamma^2k\left(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\right)\langle v_{k+1}, v_{k+1}-v_k\rangle .
\end{aligned} \tag{4.29}$$

Furthermore, one can show that for every $k \geq 1$ we get

$$\begin{aligned}
&\left(k+1\right)\langle z_{k+1}-z^*, v_{k+1}\rangle - k\langle z_k - z^*, v_k\rangle \\
&\quad = \langle z_{k+1}-z^*, v_{k+1}\rangle + k\left(\langle z_{k+1}-z^*, v_{k+1}\rangle - \langle z_k-z^*, v_k\rangle\right) \\
&\quad = \langle z_{k+1}-z^*, v_{k+1}\rangle + k\langle z_{k+1}-z^*, v_{k+1}-v_k\rangle \\
&\qquad - k\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle + k\langle z_{k+1}-z_k, v_{k+1}\rangle .
\end{aligned} \tag{4.30}$$

and

$$\begin{aligned}
&\left(k+1\right)\Big(\left(3\alpha-2\right)\left(k+1\right) + 2\alpha\left(\alpha-1\right)\Big)\|v_{k+1}\|^2 - k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\|v_k\|^2 \\
&\quad = \left(2\left(3\alpha-2\right)k + 2\alpha^2+\alpha-2\right)\|v_{k+1}\|^2 \\
&\qquad + k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\left(\|v_{k+1}\|^2 - \|v_k\|^2\right) \\
&\quad = \left(2\left(3\alpha-2\right)k + 2\alpha^2+\alpha-2\right)\|v_{k+1}\|^2 \\
&\qquad + 2k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\langle v_{k+1}, v_{k+1}-v_k\rangle \\
&\qquad - k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\|v_{k+1}-v_k\|^2 .
\end{aligned} \tag{4.31}$$

In addition, direct computations show that

$$\alpha - 2\left(\alpha-1\right)^2 + \alpha - 2 = -2\left(\alpha-1\right)\left(\alpha-2\right)$$

and

$$-(\alpha-2)^2-(\alpha-2)(3\alpha-2)=-4(\alpha-2)(\alpha-1).$$

Hence, multiplying (4.30) by $2\lambda\gamma(\alpha-2)/(\alpha-1)\geq 0$ and (4.31) by $\gamma^2(\alpha-2)/2(\alpha-1)^2>0$, followed by summing up the resulting identities with (4.25) and (4.29), yields (4.21) for every $k\geq 1$. □

Lemma 4.3.2. *Let $z^*\in\Omega$ and $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 4.2.1. For $0\leq\lambda\leq\alpha-1$, the following statements are true:*

(i) for every $k\geq k_0:=\max\left\{2,\left\lceil\frac{1}{\alpha-2}\right\rceil\right\}$ the following holds:

$$\begin{aligned}\mathcal{G}_{\lambda,k+1}-\mathcal{G}_{\lambda,k}\leq{}&\frac{(\alpha-1)(\alpha-2)}{\varepsilon(k+1)^2}\lambda^2\|z_{k+1}-z^*\|^2\\&-4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*,\zeta_{k+1}+F(z_{k+1})\rangle\\&+4(\eta_0k+\eta_1)\gamma\langle z_{k+1}-z_k,v_{k+1}\rangle+\left(\eta_2k+\kappa_0\sqrt{k}\right)\|z_{k+1}-z_k\|^2\\&+4\left(\eta_3k+\kappa_1\sqrt{k}\right)\gamma^2\|v_{k+1}\|^2-\frac{\alpha-2}{\alpha-1}\mu_k\gamma^2\|v_{k+1}-v_k\|^2,\end{aligned}$$

where

$$\begin{aligned}\mu_k&:=(k+1)\left(\varepsilon(k+1)+\alpha^2\sqrt{k+1}+(\alpha-4)\right)-(\alpha-2),\\\eta_0&:=\frac{1}{2(\alpha-1)}\left(4(\alpha-1)(\lambda+1-\alpha)-\alpha(\alpha-2)\right)<0,\\\eta_1&:=\frac{1}{2(\alpha-1)}\left(2\alpha(\alpha-1)(\lambda+1-\alpha)+\alpha-2(\alpha-1)^2\right)<0,\\\eta_2&:=4(\lambda+1-\alpha)\leq 0,\\\eta_3&:=-\frac{1}{2(\alpha-1)}(\alpha-2)(3\alpha-2)<0,\\\kappa_0&:=\frac{1}{\alpha-1}(\alpha-2)\sqrt{\alpha-2}>0,\\\kappa_1&:=\frac{1}{4(\alpha-1)}(\alpha-2)\alpha>0;\end{aligned}\tag{4.32}$$

(ii) for every $k\geq 1$ the quantity $\mathcal{G}_{\lambda,k}$ fulfils

$$\begin{aligned}\mathcal{G}_{\lambda,k}\geq{}&\frac{\alpha-2}{4(3\alpha-2)}\left\|4\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{2(3\alpha-2)}{\alpha-1}\gamma kv_k\right\|^2\\&+\frac{(\alpha-2)^2}{4(3\alpha-2)(\alpha-1)}k^2\|z_k-z_{k-1}\|^2+2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2.\end{aligned}\tag{4.33}$$

Proof. (i) Let $k \geq 2$ be fixed. By the definition of $\mathcal{G}_{\lambda,k}$ in (4.20), we have for every $k \geq 2$

$$\begin{aligned}
&\mathcal{G}_{\lambda,k+1} - \mathcal{G}_{\lambda,k} \\
&\quad = \mathcal{E}_{\lambda,k+1} - \mathcal{E}_{\lambda,k} - \frac{2(\alpha-2)}{\alpha-1}\gamma \Big[(k+1)^2 \langle z_{k+1} - z_k, F(z_{k+1}) - F(w_k)\rangle \\
&\qquad\qquad - k^2 \langle z_k - z_{k-1}, F(z_k) - F(w_{k-1})\rangle\Big] \\
&\quad + \frac{(\alpha-2)\alpha}{\alpha-1}\gamma^2 \Big[(k+1)\sqrt{k+1}\,\|v_{k+1} - v_k\|^2 - k\sqrt{k}\,\|v_k - v_{k-1}\|^2\Big] \\
&\quad + \frac{(\alpha-2)(1-\varepsilon)}{\alpha-1}\gamma^2 \Big[(k+1)^2\|v_{k+1} - v_k\|^2 - k^2\|v_k - v_{k-1}\|^2\Big].
\end{aligned} \tag{4.34}$$

By using the definition of η_0, η_1, η_2 and η_3 in (4.32), for every $k \geq 1$ we deduce from (4.21) that

$$\begin{aligned}
&\mathcal{E}_{\lambda,k+1} - \mathcal{E}_{\lambda,k} \\
&\quad = -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle + \Big(\eta_2 k + 2(\lambda+1-\alpha)(\alpha+1)\Big)\|z_{k+1} - z_k\|^2 \\
&\qquad + 4(\eta_0 k + \eta_1)\gamma\langle z_{k+1} - z_k, v_{k+1}\rangle - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1} - z_k, v_{k+1} - v_k\rangle \\
&\qquad - \frac{\alpha-2}{\alpha-1}k(2k+\alpha)\gamma^2\|v_{k+1} - v_k\|^2 + \left(4\eta_3 k - \frac{\alpha-2}{\alpha-1}\left(2\alpha^2 + \alpha - 2\right)\right)\gamma^2\|v_{k+1}\|^2 \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1} - z_k, v_{k+1} - v_k\rangle \\
&\qquad + \Big(4(\eta_0 k + \eta_1)\gamma\langle z_{k+1} - z_k, v_{k+1}\rangle + \eta_2 k\|z_{k+1} - z_k\|^2 + 4\eta_3 k\gamma^2\|v_{k+1}\|^2\Big) \\
&\qquad - \frac{\alpha-2}{\alpha-1}k(2k+\alpha)\gamma^2\|v_{k+1} - v_k\|^2,
\end{aligned} \tag{4.35}$$

where the inequality comes from the fact that $0 \leq \lambda \leq \alpha - 1$ and $\alpha > 2$. Plugging (4.35)

into (4.34) yields for every $k \geq 2$

$$\begin{aligned}
&\mathcal{G}_{\lambda,k+1} - \mathcal{G}_{\lambda,k} \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle - \frac{\alpha-2}{\alpha-1}\left(2k^2+\alpha k\right)\gamma^2\|v_{k+1}-v_k\|^2 \\
&\quad\quad - \frac{2(\alpha-2)}{\alpha-1}\gamma\Big[(k+1)^2\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\quad\quad\quad - k^2\langle z_k - z_{k-1}, F(z_k)-F(w_{k-1})\rangle\Big] \\
&\quad\quad + \frac{(\alpha-2)\alpha}{\alpha-1}\gamma^2\left[(k+1)\sqrt{k+1}\|v_{k+1}-v_k\|^2 - k\sqrt{k}\|v_k-v_{k-1}\|^2\right] \\
&\quad\quad + \frac{(\alpha-2)(1-\varepsilon)}{\alpha-1}\gamma^2\left[(k+1)^2\|v_{k+1}-v_k\|^2 - k^2\|v_k-v_{k-1}\|^2\right] \\
&\quad\quad + \Big(4(\eta_0 k+\eta_1)\gamma\langle z_{k+1}-z_k, v_{k+1}\rangle + \eta_2 k\|z_{k+1}-z_k\|^2 + 4\eta_3 k\gamma^2\|v_{k+1}\|^2\Big) \\
&\quad\quad - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle.
\end{aligned} \tag{4.36}$$

Our next aim is to derive upper estimates for the first two terms on the right-hand side of (4.36), which will eventually simplify the subsequent three terms. From the Cauchy-Schwarz inequality and (4.16) we have for all $k \geq 1$

$$\begin{aligned}
&-4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, v_{k+1}\rangle = -4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(w_k)\rangle \\
&\quad = -4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle \\
&\quad\quad + 4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, F(z_{k+1})-F(w_k)\rangle \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle \\
&\quad\quad + 4(\alpha-2)\lambda\gamma\|z_{k+1}-z^*\|\,\|F(z_{k+1})-F(w_k)\| \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle \\
&\quad\quad + 2(\alpha-2)\lambda\gamma\|z_{k+1}-z^*\|\,\|v_{k+1}-v_k\| \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle + \frac{(\alpha-1)(\alpha-2)}{\varepsilon(k+1)^2}\lambda^2\|z_{k+1}-z^*\|^2 \\
&\quad\quad + \frac{\alpha-2}{\alpha-1}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2.
\end{aligned} \tag{4.37}$$

For $\zeta_k \in N_C(z_k)$ and $\zeta_{k+1} \in N_C(z_{k+1})$, the monotonicity of N_C and F, together with the relation (4.17) and the fact that for all $k \geq 1$

$$\zeta_k + F(z_k) - v_k = F(z_k) - F(w_{k-1}),$$

yield for every $k \geq 1$

$$
\begin{aligned}
&-\frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle \\
&\quad\leq \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\left\langle z_{k+1}-z_k, \left(\zeta_{k+1}+F(z_{k+1})-v_{k+1}\right)-\left(\zeta_k+F(z_k)-v_k\right)\right\rangle \\
&\quad= \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\left\langle z_{k+1}-z_k, \left(F(z_{k+1})-F(w_k)\right)-\left(F(z_k)-F(w_{k-1})\right)\right\rangle \\
&\quad= \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad-\frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, F(z_k)-F(w_{k-1})\rangle \\
&\quad= \frac{2(\alpha-2)}{\alpha-1}\gamma (k+1)^2\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad-\frac{2(\alpha-2)}{\alpha-1}\gamma k^2\langle z_k-z_{k-1}, F(z_k)-F(w_{k-1})\rangle \\
&\qquad+\frac{2(\alpha-2)}{\alpha-1}\gamma((\alpha-2)k-1)\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad+\frac{2(\alpha-2)}{\alpha-1}\alpha\gamma^2 k\langle v_{k+1}, F(z_k)-F(w_{k-1})\rangle \\
&\qquad+\frac{4(\alpha-2)}{\alpha-1}\gamma^2 k^2\langle v_{k+1}-v_k, F(z_k)-F(w_{k-1})\rangle .
\end{aligned}
\tag{4.38}
$$

By Young's inequality together with (4.16) for every $k \geq \left\lceil \frac{1}{\alpha-2} \right\rceil$ we obtain

$$
\begin{aligned}
&\frac{2(\alpha-2)}{\alpha-1}\gamma((\alpha-2)k-1)\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\Big(\sqrt{(\alpha-2)k-1}\,\|z_{k+1}-z_k\|^2 \\
&\qquad\qquad +\gamma^2((\alpha-2)k-1)\sqrt{(\alpha-2)k-1}\,\|F(z_{k+1})-F(w_k)\|^2\Big) \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\sqrt{(\alpha-2)k}\,\|z_{k+1}-z_k\|^2 \\
&\qquad+(\alpha-2)\sqrt{\alpha-1}\gamma^2(k+1)\sqrt{k+1}\,\|F(z_{k+1})-F(w_k)\|^2 \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\sqrt{(\alpha-2)k}\,\|z_{k+1}-z_k\|^2 \\
&\qquad+4(\alpha-2)\sqrt{\alpha-1}\gamma^4L^2(k+1)\sqrt{k+1}\,\|v_{k+1}-v_k\|^2 \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\sqrt{(\alpha-2)k}\,\|z_{k+1}-z_k\|^2+(\alpha-2)\alpha\gamma^2(k+1)\sqrt{k+1}\,\|v_{k+1}-v_k\|^2 ,
\end{aligned}
\tag{4.39}
$$

where in the second estimate we use the fact that $(\alpha-2)k-1 \leq (\alpha-1)(k+1)$, while in the last one we combine $\sqrt{\alpha-1} \leq \alpha$ and $\gamma L < 1/4 < 1$.

In addition, for every $k \geq 2$ we derive

$$\begin{aligned}
&\frac{2(\alpha-2)}{\alpha-1}\alpha\gamma^2 k \langle v_{k+1}, F(z_k) - F(w_{k-1})\rangle \\
&\quad \leq \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\,\|v_{k+1}\|^2 + \frac{\alpha-2}{\alpha-1}\alpha\gamma^2 k\sqrt{k}\,\|F(z_k) - F(w_{k-1})\|^2 \\
&\quad \leq \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\,\|v_{k+1}\|^2 + \frac{\alpha-2}{\alpha-1}\alpha\gamma^2 k\sqrt{k}\,\|v_k - v_{k-1}\|^2 \\
&\quad = \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\,\|v_{k+1}\|^2 + \frac{\alpha-2}{\alpha-1}\alpha\gamma^2 (k+1)\sqrt{k+1}\,\|v_{k+1} - v_k\|^2 \\
&\qquad - \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\left[(k+1)\sqrt{k+1}\,\|v_{k+1} - v_k\|^2 - k\sqrt{k}\,\|v_k - v_{k-1}\|^2\right],
\end{aligned}$$

and, by using the Cauchy-Schwarz inequality and (4.16),

$$\begin{aligned}
&\frac{4(\alpha-2)}{\alpha-1}\gamma^2 k^2 \langle v_{k+1} - v_k, F(z_k) - F(w_{k-1})\rangle \\
&\quad \leq \frac{2(\alpha-2)}{\alpha-1}(1-\varepsilon)\gamma^2 k^2\,\|v_{k+1} - v_k\|\,\|v_k - v_{k-1}\| \\
&\quad \leq \frac{\alpha-2}{\alpha-1}(1-\varepsilon)\gamma^2 k^2\left(\|v_{k+1} - v_k\|^2 + \|v_k - v_{k-1}\|^2\right) \\
&\quad = \frac{4(\alpha-2)}{\alpha-1}\gamma^3 L k^2\left(\|v_{k+1} - v_k\|^2 + \|v_k - v_{k-1}\|^2\right) \\
&\quad \leq -\frac{4(\alpha-2)}{\alpha-1}\gamma^3 L\left((k+1)^2\,\|v_{k+1} - v_k\|^2 - k^2\,\|v_k - v_{k-1}\|^2\right) \\
&\qquad + \frac{8(\alpha-2)}{\alpha-1}\gamma^3 L(k+1)^2\,\|v_{k+1} - v_k\|^2 \\
&\quad = -\frac{4(\alpha-2)}{\alpha-1}\gamma^3 L\left((k+1)^2\,\|v_{k+1} - v_k\|^2 - k^2\,\|v_k - v_{k-1}\|^2\right) \\
&\qquad + \frac{2(\alpha-2)}{\alpha-1}(1-\varepsilon)\gamma^2(k+1)^2\,\|v_{k+1} - v_k\|^2,
\end{aligned} \tag{4.40}$$

where we want to recall that the first equality comes from (4.15). By plugging (4.39) and (4.40) into (4.38), then combining the result with (4.37), we get after rearranging

the terms for every $k \geq k_0$

$$\begin{aligned}
&-4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, v_{k+1}\rangle - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle \\
&\quad\leq \frac{2(\alpha-2)}{\alpha-1}\gamma\Big[(k+1)^2\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad\qquad - k^2\langle z_k - z_{k-1}, F(z_k)-F(w_{k-1})\rangle\Big] \\
&\quad - \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\Big[(k+1)\sqrt{k+1}\,\|v_{k+1}-v_k\|^2 - k\sqrt{k}\,\|v_k-v_{k-1}\|^2\Big] \\
&\quad - \frac{4(\alpha-2)}{\alpha-1}\gamma^3 L\Big[(k+1)^2\|v_{k+1}-v_k\|^2 - k^2\|v_k-v_{k-1}\|^2\Big] \\
&\quad - 4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle \\
&\quad + \frac{\alpha-2}{\alpha-1}\left(\left(2k^2+\alpha k\right)-\mu_k\right)\gamma^2\|v_{k+1}-v_k\|^2 \\
&\quad + \frac{1}{\alpha-1}(\alpha-2)\sqrt{(\alpha-2)k}\,\|z_{k+1}-z_k\|^2 + \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\,\|v_{k+1}\|^2 \\
&\quad + \frac{1}{\varepsilon(k+1)^2}(\alpha-1)(\alpha-2)\lambda^2\|z_{k+1}-z^*\|^2,
\end{aligned} \tag{4.41}$$

where we set

$$\begin{aligned}
\mu_k &:= \left(2k^2+\alpha k\right) - \varepsilon(k+1)^2 - (\alpha-1)\alpha(k+1)\sqrt{k+1} - \alpha(k+1)\sqrt{k+1} \\
&\quad - 2(1-\varepsilon)(k+1)^2 \\
&= \varepsilon(k+1)^2 + (\alpha-4)k - 2 - \alpha^2(k+1)\sqrt{k+1} \\
&= (k+1)\left(\varepsilon(k+1) + \alpha^2\sqrt{k+1} + \alpha - 4\right) - (\alpha-2).
\end{aligned}$$

Finally, summing up (4.36) and (4.41), we obtain the desired estimate.

(ii) Observe that

$$\begin{aligned}
&\frac{2(\alpha-2)}{\alpha-1}\lambda\gamma k\langle z_k - z^*, v_k\rangle + \frac{(\alpha-2)(3\alpha-2)}{2(\alpha-1)^2}\gamma^2k^2\|v_k\|^2 \\
&\quad = \frac{\alpha-2}{3\alpha-2}\left(\frac{2(3\alpha-2)}{\alpha-1}\lambda\gamma k\langle z_k-z^*, v_k\rangle + \frac{(3\alpha-2)^2}{2(\alpha-1)^2}\gamma^2k^2\|v_k\|^2\right) \\
&\quad = \frac{1}{3\alpha-2}(\alpha-2)\left(\frac{1}{2}\left\|2\lambda(z_k-z^*) + \frac{3\alpha-2}{\alpha-1}\gamma k v_k\right\|^2 - 2\lambda^2\|z_k-z^*\|^2\right)
\end{aligned}$$

By the definition of $u_{\lambda,k}$ in (4.18) and by using the identity

$$\|x\|^2 + \|y\|^2 = \frac{1}{2}\left(\|x+y\|^2 + \|x-y\|^2\right) \quad \forall x, y \in \mathcal{H},$$

we deduce that for every $k \geq 1$

$$\begin{aligned}
\mathcal{E}_{\lambda,k} &= \frac{1}{2}\left\|u_{\lambda,k}\right\|^2 + 2\lambda\left(\alpha-1-\lambda\right)\left\|z_k - z^*\right\|^2 + \frac{2\left(\alpha-2\right)}{\alpha-1}\lambda\gamma k\left\langle z_k - z^*, v_k\right\rangle \\
&\quad + \frac{\alpha-2}{\alpha-1}\gamma^2 k\left(\frac{1}{2\left(\alpha-1\right)}\left(3\alpha-2\right)k + \alpha\right)\left\|v_k\right\|^2 \\
&= \frac{1}{2}\left\|2\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{3\alpha-2}{\alpha-1}\gamma k v_k\right\|^2 \\
&\quad + 2\left(\alpha-1\right)\lambda\left(1 - \frac{4\lambda}{3\alpha-2}\right)\left\|z_k - z^*\right\|^2 + \frac{\alpha-2}{\alpha-1}\alpha\gamma^2 k\left\|v_k\right\|^2 \\
&\quad + \frac{\alpha-2}{2\left(3\alpha-2\right)}\left\|2\lambda\left(z_k - z^*\right) + \frac{3\alpha-2}{\alpha-1}\gamma k v_k\right\|^2 \\
&= \frac{\alpha}{3\alpha-2}\left\|2\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{3\alpha-2}{\alpha-1}\gamma k v_k\right\|^2 \\
&\quad + 2\left(\alpha-1\right)\lambda\left(1 - \frac{4\lambda}{3\alpha-2}\right)\left\|z_k - z^*\right\|^2 + \frac{\alpha-2}{\alpha-1}\alpha\gamma^2 k\left\|v_k\right\|^2 \\
&\quad + \frac{\alpha-2}{4\left(3\alpha-2\right)}\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha-2\right)}{\alpha-1}\gamma k v_k\right\|^2 \\
&\quad + \frac{\alpha-2}{3\alpha-2}k^2\left\|z_k - z_{k-1}\right\|^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathcal{G}_{\lambda,k} &= \mathcal{E}_{\lambda,k} - \frac{2\left(\alpha-2\right)}{\alpha-1}\gamma k^2\left\langle z_k - z_{k-1}, F\left(z_k\right) - F\left(w_{k-1}\right)\right\rangle \\
&\quad + \frac{\alpha-2}{\alpha-1}\gamma^2 k\sqrt{k}\left(\left(1-\varepsilon\right)\sqrt{k} + \alpha\right)\left\|v_k - v_{k-1}\right\|^2 \\
&\geq \frac{\alpha-2}{4\left(3\alpha-2\right)}\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha-2\right)}{\alpha-1}\gamma k v_k\right\|^2 \\
&\quad + \frac{\alpha-2}{3\alpha-2}k^2\left\|z_k - z_{k-1}\right\|^2 + 2\left(\alpha-1\right)\lambda\left(1 - \frac{4\lambda}{3\alpha-2}\right)\left\|z_k - z^*\right\|^2 \\
&\quad - \frac{2\left(\alpha-2\right)}{\alpha-1}\gamma k^2\left\langle z_k - z_{k-1}, F\left(z_k\right) - F\left(w_{k-1}\right)\right\rangle \\
&\quad + \frac{\alpha-2}{\alpha-1}\gamma^2 k\sqrt{k}\left(4\gamma L\sqrt{k} + \alpha\right)\left\|v_k - v_{k-1}\right\|^2.
\end{aligned}$$

Now we use relation (4.16) and apply Lemma 2.5.5 with $(a,b,c) := \left(1/2, -2\gamma, 2\gamma/L\right)$ to verify that for every $k \geq 1$

$$\begin{aligned}
&\frac{1}{2}\left\|z_k - z_{k-1}\right\|^2 - 4\gamma\left\langle z_k - z_{k-1}, F\left(z_k\right) - F\left(w_{k-1}\right)\right\rangle + 8\gamma^3 L\left\|v_k - v_{k-1}\right\|^2 \\
&\quad \geq \frac{1}{2}\left\|z_k - z_{k-1}\right\|^2 - 4\gamma\left\langle z_k - z_{k-1}, F\left(z_k\right) - F\left(w_{k-1}\right)\right\rangle + \frac{2\gamma}{L}\left\|F\left(z_k\right) - F\left(w_{k-1}\right)\right\|^2 \\
&\quad \geq 0.
\end{aligned}$$

Combining the last two estimates, for every $k \geq k_1$ one can easily conclude that

$$\begin{aligned}\mathcal{G}_{\lambda,k} &\geq \frac{\alpha-2}{4(3\alpha-2)}\left\|4\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{2(3\alpha-2)}{\alpha-1}\gamma k v_k\right\|^2 \\ &\quad+(\alpha-2)\left(\frac{1}{3\alpha-2}-\frac{1}{4(\alpha-1)}\right)k^2\|z_k-z_{k-1}\|^2 \\ &\quad+2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2 \\ &= \frac{\alpha-2}{4(3\alpha-2)}\left\|4\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{2(3\alpha-2)}{\alpha-1}\gamma k v_k\right\|^2 \\ &\quad+\frac{(\alpha-2)^2}{4(3\alpha-2)(\alpha-1)}k^2\|z_k-z_{k-1}\|^2+2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2,\end{aligned}$$

which is the desired inequality. □

The following lemma will be helpful in the main proof.

Lemma 4.3.3. *The following statements are true:*

(i) there exist two parameters

$$0 \leq \underline{\lambda}(\alpha) < \bar{\lambda}(\alpha) \leq \frac{3\alpha-2}{4}, \tag{4.42}$$

such that for every λ satisfying $\underline{\lambda}(\alpha) < \lambda < \bar{\lambda}(\alpha)$ one can find an integer $k_2(\lambda) \geq 1$ with the property that the following inequality holds for every $k \geq k_2(\lambda)$

$$\begin{aligned}R_k &:= \sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\Big(\eta_2 k+\kappa_0\sqrt{k}\Big)\|z_{k+1}-z_k\|^2+4\gamma\Big(\eta_0 k+\eta_1\Big)\langle z_{k+1}-z_k, v_{k+1}\rangle \\ &\quad+4\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\gamma^2\Big(\eta_3 k+\kappa_1\sqrt{k}\Big)\|v_{k+1}\|^2 \leq 0;\end{aligned} \tag{4.43}$$

(ii) there exists a positive integer k_3 such that for every $k \geq k_3$ we have

$$\mu_k \geq \frac{\varepsilon}{2}(k+1)^2. \tag{4.44}$$

Proof. (i) For the quadratic expression in R_k we calculate

$$\begin{aligned}\frac{\Delta_k'}{4\gamma^2} &:= (\eta_0 k+\eta_1)^2-\frac{5\alpha-2}{2(3\alpha-2)}k\Big(\eta_2\sqrt{k}+\kappa_0\Big)\Big(\eta_3\sqrt{k}+\kappa_1\Big) \\ &= \left(\eta_0^2-\frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3\right)k^2-\frac{5\alpha-2}{2(3\alpha-2)}(\eta_2\kappa_1+\kappa_0\eta_3)k\sqrt{k} \\ &\quad+\left(2\eta_0\eta_1-\frac{5\alpha-2}{2(3\alpha-2)}\kappa_0\kappa_1\right)k+\eta_1^2.\end{aligned}$$

Since $\left(\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3\right)k^2$ is the dominant term in the above polynomial, it suffices to guarantee that

$$\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3 < 0 \tag{4.45}$$

holds in order to ensure the existence of some integer $k_2(\lambda) \geq 1$ such that $\Delta_k' \leq 0$ for every $k \geq k_2(\lambda)$ and to obtain from here, due to Lemma 2.5.5 (ii), that $R_k \leq 0$ for every $k \geq k_2(\lambda)$.

It remains to show that there exists a choice of λ for which (4.45) is true. We set $\xi := \lambda + 1 - \alpha \leq 0$ and get

$$\begin{aligned}
\eta_0 &= \frac{1}{2(\alpha-1)}\left(4(\alpha-1)(\lambda+1-\alpha) - \alpha(\alpha-2)\right) \\
&= \frac{1}{2(\alpha-1)}\left(4(\alpha-1)\xi - \alpha(\alpha-2)\right), \\
\eta_2\eta_3 &= -\frac{2}{\alpha-1}(\alpha-2)(3\alpha-2)(\lambda+1-\alpha) = -\frac{2}{\alpha-1}(\alpha-2)(3\alpha-2)\xi.
\end{aligned}$$

This means that we have to guarantee that there exists a choice for ξ satisfying

$$\begin{aligned}
&\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3 \\
&= \frac{1}{4(\alpha-1)^2}\left(\left(4(\alpha-1)\xi - \alpha(\alpha-2)\right)^2 + 4(5\alpha-2)(\alpha-1)(\alpha-2)\xi\right) \\
&= \frac{1}{4(\alpha-1)^2}\left(16(\alpha-1)^2\xi^2 + 4(\alpha-1)(\alpha-2)(3\alpha-2)\xi + \alpha^2(\alpha-2)^2\right) < 0,
\end{aligned}$$

which is nothing else than

$$16(\alpha-1)^2\xi^2 + 4(\alpha-1)(\alpha-2)(3\alpha-2)\xi + \alpha^2(\alpha-2)^2 < 0. \tag{4.46}$$

A direct computation shows that

$$\Delta_\xi := 16(\alpha-1)^2(\alpha-2)^2\left((3\alpha-2)^2 - 4\alpha^2\right) = 16(\alpha-1)^2(\alpha-2)^3(5\alpha-2) > 0.$$

Hence, in order to get (4.46), we have to choose ξ between the two roots of the quadratic function arising in this formula, in other words

$$\begin{aligned}
\xi_1(\alpha) &:= \frac{1}{32(\alpha-1)^2}\left(-4(\alpha-1)(\alpha-2)(3\alpha-2) - \sqrt{\Delta_\xi}\right) \\
&= -\frac{1}{8(\alpha-1)}(\alpha-2)\left(3\alpha-2+\sqrt{(\alpha-2)(5\alpha-2)}\right) \\
&< \xi = \lambda + 1 - \alpha < \xi_2(\alpha) := \frac{1}{32(\alpha-1)^2}\left(-4(\alpha-1)(\alpha-2)(3\alpha-2) + \sqrt{\Delta_\xi}\right) \\
&= -\frac{1}{8(\alpha-1)}(\alpha-2)\left(3\alpha-2-\sqrt{(\alpha-2)(5\alpha-2)}\right).
\end{aligned}$$

Obviously $\xi_1(\alpha) < 0$ and from Vieta formula $\xi_1(\alpha) \cdot \xi_2(\alpha) = \frac{\alpha^2(\alpha-2)^2}{16(\alpha-1)^2}$, it follows that we must have $\xi_2(\alpha) < 0$ as well.

Therefore, going back to λ, in order to be sure that $\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3 < 0$ this must be chosen such that

$$\alpha - 1 + \xi_1(\alpha) < \lambda < \alpha - 1 + \xi_2(\alpha).$$

Next we will show that

$$0 < \alpha - 1 - \frac{1}{8(\alpha-1)}(\alpha-2)(3\alpha-2) < \frac{3\alpha}{4} - \frac{1}{2}. \tag{4.47}$$

Indeed, the inequality on the left-hand side follows immediately, since

$$\begin{aligned}\alpha - 1 - \frac{1}{8(\alpha-1)}(\alpha-2)(3\alpha-2) &= \frac{1}{8(\alpha-1)}\left(5\alpha^2 - 8\alpha + 4\right)\\ &= \frac{1}{8(\alpha-1)}\left(\alpha^2 + 4(\alpha-1)^2\right) > 0.\end{aligned}$$

Using this relation, one can notice that the inequality on the right hand side of (4.47) can be equivalently written as

$$5\alpha^2 - 8\alpha + 4 < 2(\alpha-1)(3\alpha-2) \Leftrightarrow 0 < \alpha^2 - 2\alpha = \alpha(\alpha-2),$$

which is true as $\alpha > 2$.

From (4.47) we immediately deduce that

$$0 < \alpha - 1 + \xi_2(\alpha) \quad \text{and} \quad \alpha - 1 + \xi_1(\alpha) < \frac{3\alpha}{4} - \frac{1}{2}.$$

This allows us to choose $\underline{\lambda} < \bar{\lambda}$, where

$$\begin{aligned}\underline{\lambda}(\alpha) &:= \alpha - 1 + \xi_1(\alpha)\\ &= \frac{1}{8(\alpha-1)}\alpha^2 + \frac{1}{2}(\alpha-1) - \frac{1}{8(\alpha-1)}(\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)}\\ \bar{\lambda}(\alpha) &:= \min\left\{\frac{3\alpha}{4} - \frac{1}{2}, \alpha - 1 + \xi_2(\alpha)\right\}\\ &= \min\left\{\frac{3\alpha}{4} - \frac{1}{2}, \frac{1}{8(\alpha-1)}\alpha^2 + \frac{1}{2}(\alpha-1) + \frac{1}{8(\alpha-1)}(\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)}\right\},\end{aligned}$$

since

$$\frac{1}{8(\alpha-1)}\alpha^2 + \frac{1}{2}(\alpha-1) - \frac{1}{8(\alpha-1)}(\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)} > 0.$$

Indeed, as $(\alpha-1)\sqrt{\alpha-1} > (\alpha-2)\sqrt{\alpha-2}$ and $4\sqrt{\alpha-1} > \sqrt{5\alpha-2}$ we can easily deduce that

$$\alpha^2 + 4(\alpha-1)^2 > 4(\alpha-1)^2 > (\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)}$$

and the claim follows.

In conclusion, choosing λ to satisfy $\underline{\lambda}(\alpha) < \lambda < \bar{\lambda}(\alpha)$, we have

$$\eta_0^2 - \frac{5\alpha - 2}{2(3\alpha - 2)}\eta_2\eta_3 < 0$$

and therefore there exists some integer $k_2(\lambda) \geq 1$ such that $R_k \leq 0$ for every $k \geq k_2(\lambda)$.

(ii) For every $k \geq 1$ we have

$$\mu_k - \frac{\varepsilon}{2}(k+1)^2 = \frac{\varepsilon}{2}(k+1)^2 + \alpha^2(k+1)\sqrt{k+1} + (\alpha - 4)(k+1) - (\alpha - 2),$$

and the conclusion is obvious. □

The following proposition plays a key role in proving the convergence rates in Proposition 4.3.5 which will be used to prove Theorem 4.2.2.

Proposition 4.3.4. *Let $z^* \in \Omega$ and $(z_k)_{k\geq 0}$, $(w_k)_{k\geq 0}$, $(\zeta_k)_{k\geq 0}$ be the sequences generated by Algorithm 4.2.1 and let $(v_k)_{k\geq 0}$ be the sequence defined by (4.13). Then the following statements are true:*

(i) the following hold:

$$\sum_{k\geq 1} \langle z_k - z^*, F(z_k) + \zeta_k \rangle < +\infty, \tag{4.48a}$$

$$\sum_{k\geq 1} k \|v_{k+1} - v_k\|^2 < +\infty, \tag{4.48b}$$

$$\sum_{k\geq 1} k \|z_{k+1} - z_k\|^2 < +\infty, \tag{4.48c}$$

$$\sum_{k\geq 1} k \|F(w_k) + \zeta_{k+1}\|^2 < +\infty; \tag{4.48d}$$

(ii) the sequence $(z_k)_{k\geq 0}$ is bounded and the following hold as $k \to +\infty$:

$$\|z_k - z_{k-1}\| = \mathcal{O}\left(\frac{1}{k}\right), \quad \|\zeta_k + F(w_{k-1})\| = \mathcal{O}\left(\frac{1}{k}\right), \quad \|\zeta_k + F(z_k)\| = \mathcal{O}\left(\frac{1}{k}\right),$$

$$\langle z_k - z^*, \zeta_k + F(z_k) \rangle = \mathcal{O}\left(\frac{1}{k}\right), \quad \langle z_k - z^*, F(z_k) \rangle = \mathcal{O}\left(\frac{1}{k}\right);$$

(iii) there exist $0 \leq \underline{\lambda}(\alpha) < \bar{\lambda}(\alpha) \leq {}^{(3\alpha-2)}/_4$ such that for every $\underline{\lambda}(\alpha) < \lambda < \bar{\lambda}(\alpha)$ the sequences $(\mathcal{E}_{\lambda,k})_{k\geq 1}$ and $(\mathcal{G}_{\lambda,k})_{k\geq 2}$ converge.

Proof. According to Lemma 4.3.3 there exist $\underline{\lambda}(\alpha) < \bar{\lambda}(\alpha)$ such that (4.42) holds. We choose $\underline{\lambda}(\alpha) < \lambda < \bar{\lambda}(\alpha)$ and get, according to the same result, an integer $k_2(\lambda) \geq 1$

such that for every $k \geq k_2(\lambda)$ the inequality (4.43) holds. In addition, according to Lemma 4.3.3(ii), we get a positive integer k_3 such that (4.44) holds for every $k \geq k_3$.

This means that for every $k \geq k_4(\lambda) := \max\{k_0, k_2(\lambda), k_3\}$, where k_0 is the positive integer provided by Lemma (4.3.2)(i), we have

$$\begin{aligned}
&\mathcal{G}_{\lambda,k+1} - \mathcal{G}_{\lambda,k} \\
&\quad \leq \frac{(\alpha-1)(\alpha-2)\lambda^2}{\varepsilon(k+1)^2}\|z_{k+1}-z^*\|^2 - 4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle \\
&\quad - \frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2 \\
&\quad + \left[\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k + \kappa_0\sqrt{k}\right]\|z_{k+1}-z_k\|^2 \\
&\quad + \left[\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_3 k + \kappa_1\sqrt{k}\right]4\gamma^2\|v_{k+1}\|^2.
\end{aligned}$$

Since $\eta_2, \eta_3 < 0$ and $\kappa_0, \kappa_1 \geq 0$, we can find some k_5 large enough such that for every $k \geq k_5$ we get

$$\begin{aligned}
\mathcal{G}_{\lambda,k+1} &\leq \mathcal{G}_{\lambda,k} + \frac{(\alpha-1)(\alpha-2)\lambda^2}{\varepsilon(k+1)^2}\|z_{k+1}-z^*\|^2 \\
&\quad - 4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle \\
&\quad - \frac{1}{2(\alpha-1)}(\alpha-2)\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2 \\
&\quad + \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\|z_{k+1}-z_k\|^2 + \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)4\gamma^2\eta_3 k\|v_{k+1}\|^2.
\end{aligned} \tag{4.49}$$

In view of (4.33), we get that $\mathcal{G}_{\lambda,k} \geq 0$ for every $k \geq 1$ thus the sequence $(\mathcal{G}_{\lambda,k})_{k\geq 2}$ is bounded from below. Moreover, by setting

$$C_0 := \frac{1}{2\varepsilon}(\alpha-2)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)^{-1} > 0,$$

we assert that

$$\begin{aligned}
\frac{(\alpha-1)(\alpha-2)\lambda^2}{\varepsilon(k+1)^2}\|z_{k+1}-z^*\|^2 &= \frac{C_0}{(k+1)^2}\cdot 2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2 \\
&\leq \frac{C_0}{(k+1)^2}\mathcal{G}_{\lambda,k+1},
\end{aligned}$$

Under these premises, we deduce from (4.49) that for every $k \geq k_5$

$$\begin{aligned}\left(1-\frac{C_0}{(k+1)^2}\right)\mathcal{G}_{\lambda,k+1} &\leq \mathcal{G}_{\lambda,k} - 4(\alpha-2)\lambda\gamma\left\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\right\rangle \\ &- \frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2 \\ &+ \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\|z_{k+1}-z_k\|^2 + \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)4\gamma^2\eta_3 k\|v_{k+1}\|^2.\end{aligned} \tag{4.50}$$

Taking $k_6 := \max\left\{k_5, \left\lceil\sqrt{C_0}-1\right\rceil\right\}$ we conclude for all $k \geq k_6$ that

$$\left(1-\frac{C_0}{(k+1)^2}\right)^{-1} = \frac{(k+1)^2}{(k+1)^2-C_0} = 1+\frac{C_0}{(k+1)^2-C_0} > 1.$$

Hence, for every $k \geq k_6$, the inequality (4.50) leads to

$$\begin{aligned}\mathcal{G}_{\lambda,k+1} &\leq \left(1+\frac{C_0}{(k+1)^2-C_0}\right)\mathcal{G}_{\lambda,k} - 4(\alpha-2)\lambda\gamma\left\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\right\rangle \\ &- \frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2 \\ &+ \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\|z_{k+1}-z_k\|^2 + \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)4\gamma^2\eta_3 k\|v_{k+1}\|^2,\end{aligned}$$

which is nothing else than the inequality (2.2) with

$$\begin{aligned}a_k &:= \mathcal{G}_{\lambda,k} \geq 0, \\ b_k &:= 4(\alpha-2)\lambda\gamma\left\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\right\rangle + \frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2 \\ &\quad - \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\|z_{k+1}-z_k\|^2 - \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)4\gamma^2\eta_3 k\|v_{k+1}\|^2 \\ &\geq 0, \\ d_k &:= \frac{C_0}{(k+1)^2-C_0} > 0.\end{aligned}$$

We obtain (4.48) as well as that the sequence $(\mathcal{G}_{\lambda,k})_{k\geq 1}$ converges due to Lemma 2.5.3.

Since $(\mathcal{G}_{\lambda,k})_{k\geq k_6}$ converges, it is bounded, which, according to (4.33), implies that the following estimate holds for every $k \geq k_6$

$$\begin{aligned}&\frac{\alpha-2}{3\alpha-2}\left\|4\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{2(3\alpha-2)}{\alpha-1}\gamma k v_k\right\|^2 \\ &\quad + \frac{(\alpha-2)^2}{4(3\alpha-2)(\alpha-1)}k^2\|z_k-z_{k-1}\|^2 + 2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2 \\ &\leq \mathcal{G}_{\lambda,k} \leq \sup_{k\geq 1}\mathcal{G}_{\lambda,k} < +\infty.\end{aligned}$$

From here we obtain the boundedness of the sequences

$$\left(4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha - 2\right)}{\alpha - 1}\gamma k v_k\right)_{k\geq 1},$$
$$\left(k\left(z_k - z_{k-1}\right)\right)_{k\geq 1} \quad \text{and} \quad \left(z_k\right)_{k\geq 0}.$$

In particular, for every $k \geq k_6$ we have

$$\begin{aligned}
&\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2(3\alpha - 2)}{\alpha - 1}\gamma k v_k\right\| \leq C_1 := \sqrt{\frac{3\alpha - 2}{\alpha - 2}\sup_{k\geq 1}\mathcal{G}_{\lambda,k}} < +\infty,\\
&k\left\|z_k - z_{k-1}\right\| \leq C_2 := \frac{2}{\alpha - 2}\sqrt{\left(3\alpha - 2\right)\left(\alpha - 1\right)\sup_{k\geq 1}\mathcal{G}_{\lambda,k}} < +\infty,\\
&\left\|z_k - z^*\right\| \leq C_3 := \sqrt{\frac{1}{2\left(\alpha - 1\right)\lambda}\left(1 - \frac{4\lambda}{3\alpha - 2}\right)^{-1}\sup_{k\geq 1}\mathcal{G}_{\lambda,k}} < +\infty.
\end{aligned} \tag{4.51}$$

Using the triangle inequality, we deduce from here that for every $k \geq k_6$

$$\begin{aligned}
\left\|v_k\right\| \leq{}& \frac{\alpha - 1}{2\left(3\alpha - 2\right)\gamma k}\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha - 2\right)}{\alpha - 1}\gamma k v_k\right\|\\
&+ \frac{\alpha - 1}{\left(3\alpha - 2\right)\gamma}\left\|z_k - z_{k-1}\right\| + \frac{2\left(\alpha - 1\right)\lambda}{\left(3\alpha - 2\right)\gamma k}\left\|z_k - z^*\right\| \leq \frac{C_4}{k},
\end{aligned}$$

where

$$C_4 := \frac{\alpha - 1}{2\left(3\alpha - 2\right)\gamma}\left(C_1 + 2C_2 + 4\bar{\lambda}\left(\alpha\right)C_3\right) > 0.$$

The statement (4.48b) yields

$$\lim_{k\to+\infty} k\left\|v_{k+1} - v_k\right\| = 0 \quad \Rightarrow \quad C_5 := \sup_{k\geq 1}\left\{k\left\|v_{k+1} - v_k\right\|\right\} < +\infty, \tag{4.52}$$

which, together with (4.16) implies that for every $k \geq k_6$

$$\begin{aligned}
\left\|\zeta_{k+1} + F\left(z_{k+1}\right)\right\| &\leq \left\|\zeta_{k+1} + F\left(z_{k+1}\right) - v_{k+1}\right\| + \left\|v_{k+1}\right\|\\
&\leq \left\|v_{k+1} - v_k\right\| + \left\|v_{k+1}\right\| \leq \frac{C_6}{k},
\end{aligned} \tag{4.53}$$

where

$$C_6 := C_4 + C_5 > 0.$$

The remaining assertion follows from the fact that $\zeta_k \in N_C\left(z_k\right)$ where $z_k \in C$ by definition, the Cauchy-Schwarz inequality and the boundedness of $\left(z_k\right)_{k\geq 0}$, namely, for every $k \geq k_6$ we deduce

$$0 \leq \left\langle z_k - z^*, F\left(z_k\right)\right\rangle \leq \left\langle z_k - z^*, \zeta_k + F\left(z_k\right)\right\rangle \leq \left\|z_k - z^*\right\|\left\|\zeta_k + F\left(z_k\right)\right\| \leq \frac{C_3 C_6}{k - 1}.$$

To complete the proof, we are going to show that in fact

$$\lim_{k\to+\infty} \mathcal{E}_{\lambda,k} = \lim_{k\to+\infty} \mathcal{G}_{\lambda,k} \in \mathbb{R}.$$

Indeed, we already have seen that

$$\lim_{k\to+\infty} (k+1)\,\|v_{k+1}-v_k\| = \lim_{k\to+\infty} \|v_{k+1}\| = 0,$$

which, by the Cauchy-Schwarz inequality and (4.16) yields

$$\begin{aligned} 0 \leq \lim_{k\to+\infty} k^2\,|\langle z_k - z_{k-1}, F(z_k) - F(w_{k-1})\rangle| &\leq C_2 \lim_{k\to+\infty} k\,\|F(z_k) - F(w_{k-1})\| \\ &\leq C_2 \lim_{k\to+\infty} k\,\|v_k - v_{k-1}\| = 0. \end{aligned}$$

From here we obtain the desired statement. □

4.3.3. Proofs of the main results

Proof of Theorem 4.2.2. Let $\underline{\lambda}(\alpha) < \bar{\lambda}(\alpha)$ be the parameters provided by Lemma 4.3.3 such that (4.42) holds and with the property that for every $\underline{\lambda}(\alpha) < \lambda < \bar{\lambda}(\alpha)$ there exists an integer $k_2(\lambda) \geq 1$ such that for every $k \geq k_2(\lambda)$ the inequality (4.43) holds.

For every $k \geq 1$ we set

$$p_k := \frac{1}{2}(\alpha-1)\,\|z_k - z^*\|^2 + k\,\langle z_k - z^*, z_k - z_{k-1} + 2\gamma v_k\rangle, \tag{4.54}$$

$$q_k := \frac{1}{2}\,\|z_k - z^*\|^2 + 2\gamma \sum_{i=1}^{k} \langle z_i - z^*, v_i\rangle. \tag{4.55}$$

Then one can see that for every $k \geq 2$ we have

$$q_k - q_{k-1} = \langle z_k - z^*, z_k - z_{k-1}\rangle - \frac{1}{2}\,\|z_k - z_{k-1}\|^2 + 2\gamma\,\langle z_k - z^*, v_k\rangle,$$

and thus

$$(\alpha-1)\,q_k + k\,(q_k - q_{k-1}) = p_k + 2(\alpha-1)\,\gamma \sum_{i=1}^{k} \langle z_i - z^*, v_i\rangle - \frac{k}{2}\,\|z_k - z_{k-1}\|^2.$$

From (4.19) and (4.18), direct computation shows that for every $k \geq 1$

$$\begin{aligned} \mathcal{E}_{\lambda,k} &= \frac{1}{2}\left\|2\lambda(z_k - z^*) + 2k(z_k - z_{k-1}) + \frac{3\alpha-2}{\alpha-1}\gamma k v_k\right\|^2 + 2\lambda(\alpha-1-\lambda)\,\|z_k - z^*\|^2 \\ &\quad + \frac{2(\alpha-2)}{\alpha-1}\lambda\gamma k\,\langle z_k - z^*, v_k\rangle + \frac{\alpha-2}{\alpha-1}\gamma^2 k\left(\frac{1}{2(\alpha-1)}(3\alpha-2)\,k + \alpha\right)\|v_k\|^2 \\ &= 2\lambda(\alpha-1)\,\|z_k - z^*\|^2 + 4\lambda k\,\langle z_k - z^*, z_k - z_{k-1} + 2\gamma v_k\rangle + \frac{\alpha-2}{\alpha-1}\alpha\gamma^2 k\,\|v_k\|^2 \\ &\quad + \frac{k^2}{2}\left(\left\|2(z_k - z_{k-1}) + \frac{3\alpha-2}{\alpha-1}\gamma v_k\right\|^2 + \frac{(\alpha-2)(3\alpha-2)}{(\alpha-1)^2}\gamma^2\,\|v_k\|^2\right). \end{aligned} \tag{4.56}$$

Therefore, for every $\underline{\lambda}(\alpha) < \lambda_1 < \lambda_2 < \bar{\lambda}(\alpha)$ we can conclude

$$\begin{aligned}\mathcal{E}_{\lambda_2,k} - \mathcal{E}_{\lambda_1,k} &= 4(\lambda_2 - \lambda_1)\left(\frac{1}{2}(\alpha-1)\|z_k - z^*\|^2 + 2k\langle z_k - z^*, (z_k - z_{k-1}) + sv_k\rangle\right) \\ &= 4(\lambda_2 - \lambda_1)\, p_k.\end{aligned}$$

Hence, according to the previous theorem, the limit $\lim_{k\to+\infty}(\mathcal{E}_{\lambda_2,k} - \mathcal{E}_{\lambda_1,k}) \in \mathbb{R}$ exists, which implies further that the limit

$$\lim_{k\to+\infty} p_k \in \mathbb{R} \text{ exists.} \tag{4.57}$$

Further, we observe that for every $k \geq 2$

$$\sum_{i=2}^{k} |\langle z_i - z^*, F(w_{i-1}) - F(z_i)\rangle| \leq \sum_{i=2}^{k} \|z_i - z^*\| \, \|F(w_{i-1}) - F(z_i)\| \tag{4.58a}$$

$$\leq 2\gamma C_4 C_5 \sum_{i=2}^{k} \frac{1}{i(i-1)} \tag{4.58b}$$

$$\leq 2\gamma C_4 C_5 \sum_{i=2}^{+\infty} \frac{1}{i(i-1)} < +\infty,$$

where (4.58a) comes from the Cauchy-Schwarz inequality, the estimate (4.58b) is a combination of (4.16), (4.51), and (4.52), and the constants C_3, C_4 are the ones defined in the proof of Theorem 4.3.4. This means the series $\sum_{k\geq 2}\langle z_k - z^*, F(w_{k-1}) - F(z_k)\rangle$ is absolutely convergent, thus convergent.

By taking into consideration (4.48a), it follows from here that the limit

$$\begin{aligned}&\lim_{k\to+\infty} \sum_{i=1}^{k} \langle z_i - z^*, v_i\rangle \\ &\quad = \lim_{k\to+\infty} \sum_{i=1}^{k} \langle z_i - z^*, \zeta_i + F(z_i)\rangle + \lim_{k\to+\infty} \sum_{i=1}^{k} \langle z_i - z^*, F(w_{i-1}) - F(z_i)\rangle \in \mathbb{R}\end{aligned}$$

exists. In addition, thanks to (4.48c), we have $\lim_{k\to+\infty} k\|z_{k+1} - z_k\|^2 = 0$, consequently,

$$\lim_{k\to+\infty} (\alpha-1)\, q_k + k(q_k - q_{k-1}) \in \mathbb{R} \text{ exists.}$$

According to Theorem 4.3.4, we have that $(q_k)_{k\geq 1}$ is bounded due to the boundedness of $(z_k)_{k\geq 0}$ and the fact that $\lim_{k\to+\infty}\sum_{i=1}^{k}\langle z_i - z^*, v_i\rangle \in \mathbb{R}$ exists. Therefore, we can apply Lemma 2.5.2 to guarantee the existence of the limit $\lim_{k\to+\infty} q_k \in \mathbb{R}$. By the definition of q_k in (4.55) and the fact that the sequence $\left(\sum_{i=1}^{k-1}\langle z_i - z^*, v_i\rangle\right)_{k\geq 1}$ converges, we conclude that $\lim_{k\to+\infty}\|z_k - z^*\| \in \mathbb{R}$ exists. The hypothesis (i) in the Opial Lemma (see Lemma 2.3.1) is fulfilled.

Let w be a weak sequential cluster point of $(z_k)_{k\geq 0}$, which means that there exists a subsequence $\{z_{k_n}\}_{n\geq 0}$ such that

$$z_{k_n} \rightharpoonup w \text{ as } n \to +\infty.$$

It follows from Theorem 4.3.4 that

$$F(z_{k_n}) + \zeta_{k_n} \to 0 \quad \text{as } n \to +\infty.$$

The maximal monotonicity of $F+N_C$ implies that $0 \in (N_C + F)(w)$ and therefore shows that the hypothesis (ii) of Lemma 2.3.1 is also verified. The proof of the convergence of the iterates is therefore completed. □

Before finally proving Theorem 4.3.5 we show convergence rates of various helpful quantities.

Proposition 4.3.5. *Let $z^* \in \Omega$ and $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 4.2.1. Then, as $k \to +\infty$, the following hold:*

$$\|z_k - z_{k-1}\| = o\left(\frac{1}{k}\right), \quad \langle F(z_k), z_k - z^*\rangle = o\left(\frac{1}{k}\right), \quad \langle \zeta_k + F(z_k), z_k - z^*\rangle = o\left(\frac{1}{k}\right)$$
$$\|\zeta_k + F(z_k)\| = o\left(\frac{1}{k}\right), \quad \|\zeta_k + F(w_{k-1})\| = o\left(\frac{1}{k}\right).$$

Proof. Let $\underline{\lambda}(\alpha) < \bar{\lambda}(\alpha)$ be the parameters provided by Lemma 4.3.3 such that (4.42) holds and with the property that for every $\underline{\lambda}(\alpha) < \lambda < \bar{\lambda}(\alpha)$ there exists an integer $k_2(\lambda) \geq 1$ such that for every $k \geq k_2(\lambda)$ the inequality (4.43) holds. We fix $\underline{\lambda}(\alpha) < \lambda < \bar{\lambda}(\alpha)$ and recall that according to Theorem 4.3.4(iii) the sequence $(\mathcal{E}_{\lambda,k})_{k\geq 1}$ converges.

We set for every $k \geq 1$

$$h_k := \frac{k^2}{2}\left(\left\|2(z_k - z_{k-1}) + \frac{3\alpha - 2}{\alpha - 1}\gamma v_k\right\|^2 + \frac{(\alpha - 2)(3\alpha - 2)}{(\alpha - 1)^2}\gamma^2 \|v_k\|^2\right),$$

and notice that, in view of (4.56) and (4.54), we have

$$\mathcal{E}_{\lambda,k} = 4\lambda p_k + \frac{4(\alpha - 2)}{\alpha - 1}\alpha\gamma^2\gamma^2 k \|v_k\|^2 + h_k.$$

Theorem 4.3.4 asserts that

$$\lim_{k\to\infty} k\|v_k\|^2 = 0,$$

which, together with $\lim_{k\to+\infty} \mathcal{E}_{\lambda,k} \in \mathbb{R}$ and $\lim_{k\to+\infty} p_k \in \mathbb{R}$ (see also (4.57)), yields the existence of

$$\lim_{k\to+\infty} h_k \in \mathbb{R}.$$

In addition, (4.48c) and (4.48d) in Theorem 4.3.4 guarantee that

$$\sum_{k\geq 1}\frac{1}{k}h_k \leq 4\sum_{k\geq 1} k\|z_k - z_{k-1}\|^2 + \frac{(3\alpha - 2)(7\alpha - 6)}{2(\alpha - 1)^2}\gamma^2 \sum_{k\geq 1} k\|v_k\|^2 < +\infty.$$

Consequently, $\lim_{k\to+\infty} h_k = 0$, which yields

$$\lim_{k\to\infty} k\left\|2\left(z_k - z_{k-1}\right) + \frac{3\alpha-2}{\alpha-1}\gamma v_k\right\| = \lim_{k\to\infty} k\left\|v_k\right\| = 0.$$

This immediately implies $\lim_{k\to+\infty} k\left\|z_k - z_{k-1}\right\| = 0$. The fact that

$$\lim_{k\to+\infty} k\left\|\zeta_k + F\left(z_k\right)\right\| = 0$$

follows from (4.16), (4.52) and (4.53), since

$$0 \leq \lim_{k\to+\infty} k\left\|\zeta_k + F\left(z_k\right)\right\| \leq \lim_{k\to+\infty} k\left\|v_k - v_{k-1}\right\| + \lim_{k\to+\infty} k\left\|v_k\right\| = 0.$$

Finally, using the Cauchy-Schwarz inequality and the fact that $(z_k)_{k\geq 0}$ is bounded, we obtain that $\lim_{k\to+\infty} k\left\langle z_k - z^*, F\left(z_k\right)\right\rangle = \lim_{k\to+\infty} k\left\langle z_k - z^*, \zeta_k + F\left(z_k\right)\right\rangle = 0$. □

Now we are able to prove the convergence rates in terms of the restricted gap and the natural gap.

Proof of Theorem 4.2.3. For every $k \geq 1$, using successively the monotonicity of F, the fact that $\zeta_k \in N_C(z_k)$, where $z_k \in C$ by its definition, and the Cauchy-Schwarz inequality, we deduce that for every $u \in C \cap \mathbb{B}(z^*; \delta(z_0))$

$$\begin{aligned}
\langle F(u), z_k - u\rangle &\leq \langle F(z_k), z_k - u\rangle \leq \langle \zeta_k + F(z_k), z_k - u\rangle \\
&= \langle \zeta_k + F(z_k), z_k - z^*\rangle + \langle \zeta_k + F(z_k), z^* - u\rangle \\
&\leq \langle \zeta_k + F(z_k), z_k - z^*\rangle + \left\|\zeta_k + F(z_k)\right\| \|z^* - u\| \\
&\leq \langle \zeta_k + F(z_k), z_k - z^*\rangle + \delta(z_0)\left\|\zeta_k + F(z_k)\right\|.
\end{aligned}$$

Therefore, it follows from Theorem 4.3.5 that

$$\begin{aligned}
\texttt{Gap}(z_k) &= \max_{u\in C\cap\mathbb{B}(z^*;\delta(z_0))} \langle F(u), z_k - u\rangle \leq \langle \zeta_k + F(z_k), z_k - z^*\rangle + \delta(z_0)\left\|\zeta_k + F(z_k)\right\| \\
&= o\left(\frac{1}{k}\right), \quad \text{as } k \to +\infty.
\end{aligned}$$

Concluding, by (4.5) we obtain

$$\texttt{Res}(z_k) \leq \left\|\zeta_k + F(z_k)\right\| = o\left(\frac{1}{k}\right), \quad \text{as } k \to +\infty,$$

and the proof is complete. □

4.4. Numerical experiments

In this section we provide two numerical experiments to complete our theoretical results. For the first one we want to solve a two-player zero-sum game, which amounts to solving a bilinear saddle point problem constrained by standard simplexes. The second one consists of application of fOGDA-VI to the training of GANs.

4.4.1. Two-player zero-sum game

We want to solve a two-player zero-sum game with mixed strategies, which means that we need to solve the following bilinear saddle point problem,

$$\min_{x\in\Delta^d}\max_{y\in\Delta^n} \Phi(x,y) := x^T A y, \tag{4.59}$$

where $A \in \mathbb{R}^{d\times n}$ is a given pay-off matrix and $\Delta^m = \{v \in \mathbb{R}^m_+ \mid \sum_{i=1}^m v_i = 1\}$ denotes the m-dimensional standard simplex. Recall that a solution of (4.59) is given by a saddle point $(x^*, y^*) \in \Delta^d \times \Delta^n$ satisfying

$$\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*) \quad \forall (x, y) \in \Delta^d \times \Delta^n.$$

This leads to a monotone inclusion problem (4.2) with $C = \Delta^d \times \Delta^n$ and

$$F : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^d \times \mathbb{R}^n, \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} 0 & A \\ -A^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} Ay \\ -A^T x \end{pmatrix}.$$

Notice that F is Lipschitz continuous but not cocoercive.

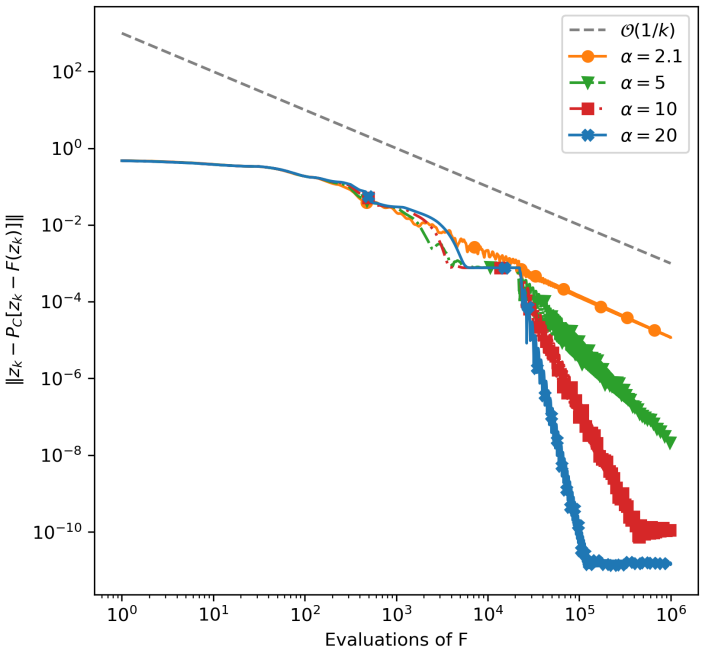


Figure 4.1.: Comparison of different momentum parameters $\alpha > 2$ in Algorithm 4.2.1 in terms of the natural residual

For our experiments we choose $d = n = 50$ and A to have entries drawn from the uniform distribution on the half-open interval $[0, 1)$. As the parameter $\alpha > 2$ can be chosen arbitrary, we do a comparison of different values in terms of the natural residual in Figure 4.1 to gain more insight. As it turns out, from a certain point on after a period where all values of α perform similarly, bigger choices for α seem to give better results with faster convergence.

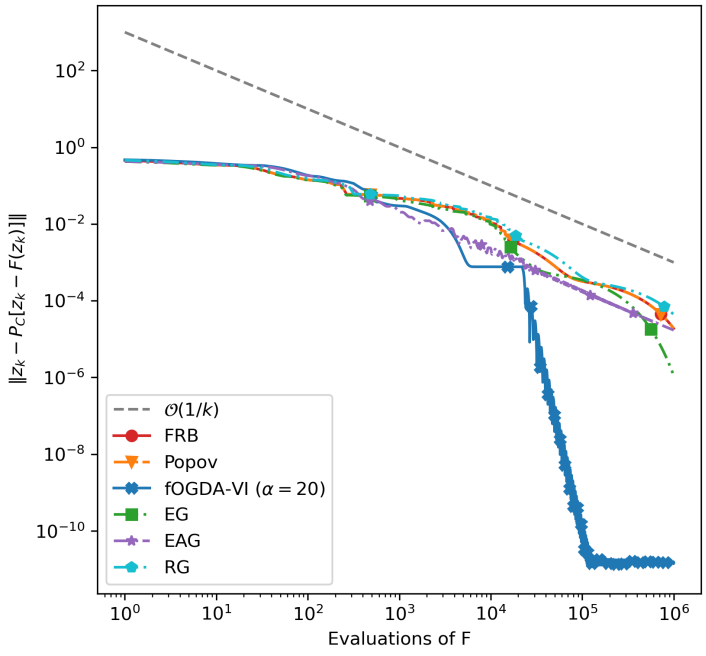


Figure 4.2.: Comparison of different methods in terms of the natural residual

Concluding, we show a comparison of different methods that we have encountered in Sections 4.1.2 and 4.1.3 in Figure 4.2 where we report results on the natural residual. We see that fOGDA-VI clearly outperforms all other methods while only requiring one evaluation of F and one projection in each iteration.

4.4.2. GAN training

Generative Adversarial Networks (GANs) [70] form a powerful class of generative models that can produce for example unseen realistic images. Originally the problem was posed as a zero-sum game between two adversarial players called *generator* and *discriminator* that try to minimise and maximise the same loss function, respectively. For more details about GANs please refer to Section 3.5.

As it was shown empirically that principled methods that are used to solve variational

inequalities [64] and monotone inclusions [20] can prove beneficial in the training process, we will apply our proposed algorithm fOGDA-VI to train ResNet architectures on the CIFAR-10 dataset.

For our experiments we use ResNet [76] architectures, see Appendix A.2.2, with the hinge version of the adversarial non-saturating loss [108] trained on the CIFAR-10 [86] data set, which consists of 60,000 $(32 \times 32 \times 3)$-images in 10 classes, with 6,000 images per class. The metrics used to evaluate the generated images are the inception score [133] (IS; higher is better) and the Fréchet inception distance [77] (FID; lower is better), both computed on 50,000 samples in their original implementations.

Furthermore, in our experiments instead of stochastic gradients we use the Adam optimizer [84] with parameters $\beta_1 = 0$ and $\beta_2 = 0.9$ that where used in recent experiments [47] outperforming the class-dependent BigGAN [38] model on CIFAR-10. Additionally, we keep the batch size and the ratio of discriminator and generator updates the same as in [47].

Since we have mini batch updates for the GAN experiments instead of taking the full gradient we perform significantly more steps to incorporate the entire gradient information. Because of this the iterator k in Algorithm 4.2.1 might grow too large soon, so we experiment to update the iterator k only every n steps with different choices of n. Furthermore, besides the inevitable search for the correct learning rate, we also need to find a good choice for the momentum parameter α.

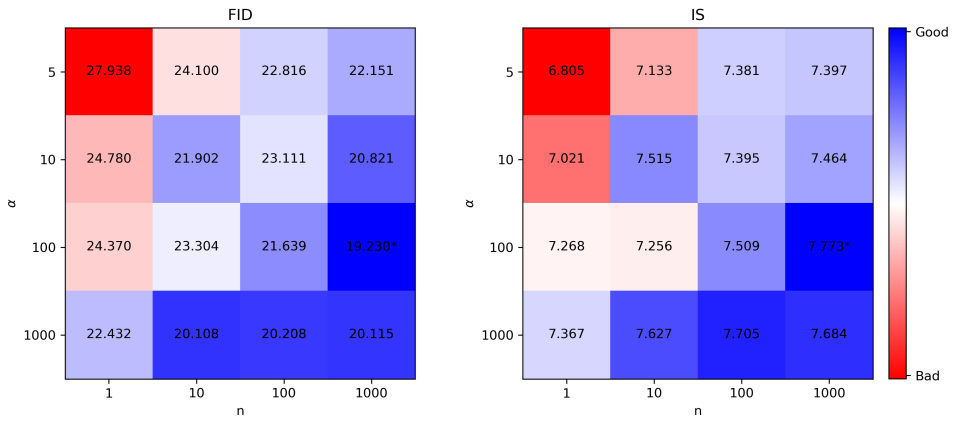
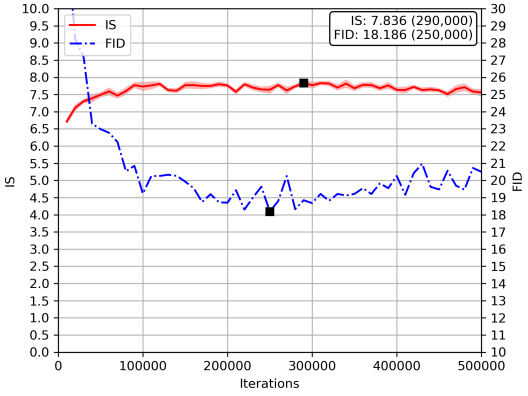


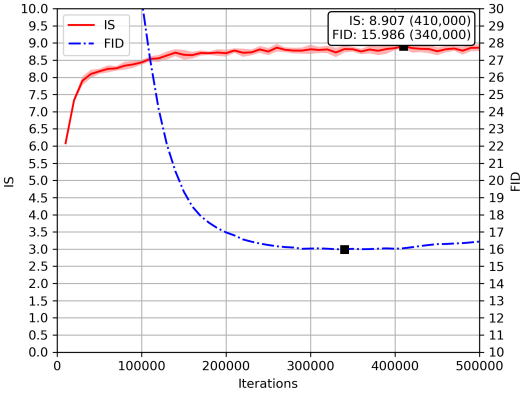
Figure 4.3.: Best FID and IS for different combinations of α and n achieved on one run with 100,000 iterations. The overall best result is indicated by an asterisk.

We performed a hyper parameter search for $\alpha = 5, 10, 100, 1000$ and $n = 1, 10, 100, 1000$, and two learning rates $\gamma = 0.0001, 0.0002$ with 100,000 iterations on one run each. As the smaller of the two learning rates achieved significantly better results in general, we checked for selected combinations of α and n with another learning rate ($\gamma = 0.00005$) which did not lead to further improvement. In Figure 4.3 we can see the best results

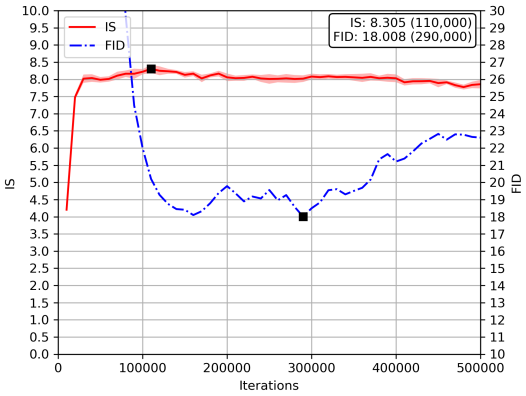
with respect to IS and FID for the learning rate $\gamma = 0.0001$. The best overall values are indicated by an asterisk and attained by $\alpha = 100$ and $n = 1000$. For this combination we then conducted a run with 500,000 iterations, which can be seen in Figure 4.4a.



(a) last



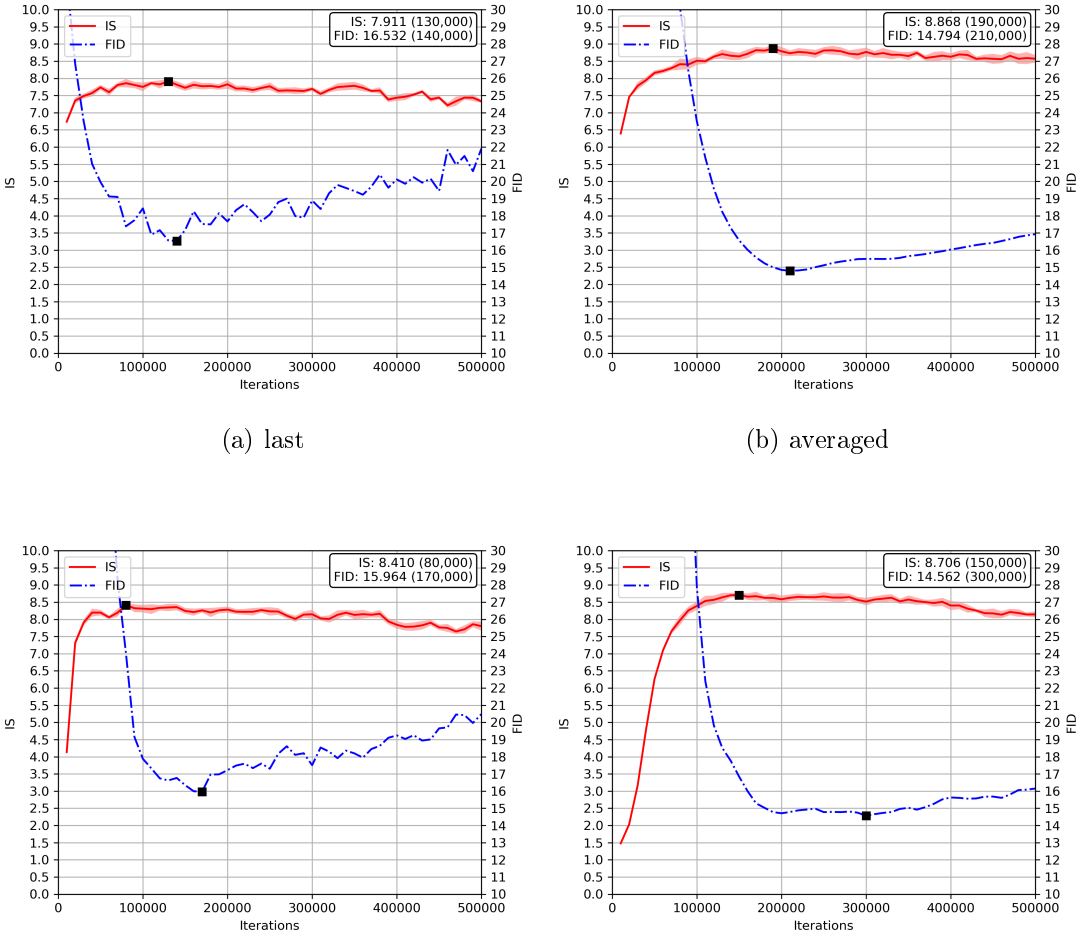
(b) averaged



(c) EMA

Figure 4.4.: Performance of training ResNets by fOGDA-VI with respect to FID and IS on one run with 500,000 iterations for (a) the last iterate, (b) the averaged iterates and (c) using EMA on the iterates.

To further reduce the cycling characteristics of GAN training two commonly used techniques in practice are the (uniform) averaging and the exponential moving averaging (EMA) of the network weights. In Figures 4.4b and 4.4c we can see that in our case these approaches are beneficial as well, with the largest improvement when using uniform averaging.



(a) last

(b) averaged

(c) EMA

(d) EMA slow

Figure 4.5.: Performance of training ResNets by LA-GDA with respect to FID and IS on one run with 500,000 iterations for (a) the last iterate, (b) the averaged iterates, (c) using EMA on the iterates and (d) using EMA on the "slow" iterates (see [47]).

We compare the results obtained by fOGDA-VI with the best method from [47], a variant of Gradient Descent Ascent incorporating averaging during the training which they call "lookahead", resulting in a method we denote by LA-GDA.

In Figure 4.5 we see that even though LA-GDA gives generally better results for the last iterate and when using EMA, the best result in terms of IS is achieved by fOGDA-VI for averaged iterates. The best FID value, however, is achieved by LA-GDA using EMA on the "slow weights". Note that using EMA on the "slow iterates" is only comparable for methods using lookahead.

5. An accelerated minimax algorithm for convex-concave saddle point problems with nonsmooth coupling function

In this chapter we aim to solve a convex-concave saddle point problem, where the convex-concave coupling function is smooth in one variable and nonsmooth in the other and *not* assumed to be linear in either. The problem is augmented by a nonsmooth regulariser in the smooth component. We propose and investigate a novel algorithm under the name of *OGAProx*, consisting of an optimistic gradient ascent step in the smooth variable coupled with a proximal step of the regulariser, and which is alternated with a proximal step in the nonsmooth component of the coupling function. We consider the situations convex-concave, convex-strongly concave and strongly convex-strongly concave related to the saddle point problem under investigation. Regarding iterates we obtain (weak) convergence, a convergence rate of order $\mathcal{O}(1/K)$ and linear convergence like $\mathcal{O}(\theta^K)$ with $\theta < 1$, respectively. In terms of function values we obtain ergodic convergence rates of order $\mathcal{O}(1/K)$, $\mathcal{O}(1/K^2)$ and $\mathcal{O}(\theta^K)$ with $\theta < 1$, respectively. We validate our theoretical considerations on a nonsmooth-linear saddle point problem, the training of multi kernel support vector machines and a classification problem incorporating minimax group fairness.

This chapter is based on the article [32].

5.1. Introduction

Saddle point – or minimax – problems arise traditionally in game theory [110] or for example in the context of determining primal-dual pairs of optimal solutions of convex optimisation problems [17]. However, in recent years they have witnessed increased interest due to many relevant and challenging applications in the field of machine learning, with the most prominent being the training of Generative Adversarial Networks (GANs) [70]. Even though the problems in reality are often not of this form, in the classical setting the minimax objective comprises a smooth convex-concave coupling function with Lipschitz continuous gradient and a (potentially nonsmooth) regulariser in each variable, leading to a convex-concave objective in total.

One well established method in practice due to its simplicity and computational efficiency is Gradient Descent Ascent (GDA), either in a simultaneous or in an alternating variant (for a recent comparison of the convergence behaviour of the two schemes we refer to [141]). However, naive application of GDA is known to lead to oscillatory behaviour or even divergence already in simple cases such as bilinear objectives. Most

algorithms with convergence guarantees in the general convex-concave setting make use of the formulation of the first order optimality conditions as monotone inclusion or variational inequality, treating both components in a symmetric fashion. For example we have the Extragradient method [85] whose application to minimax problems has been studied in [112] under the name of Mirror Prox, and the Forward-Backward-Forward method (FBF) [137] with application to saddle point problems in [20]. Both algorithms have even been successfully applied to the training of GANs (see [20, 64]), but, though being single-loop methods, suffer in practice from requiring two gradient evaluations per iteration. A possible way to avoid this is to reuse previous gradients. Doing this for FBF – as shown in [20] – recovers the Forward-Reflected Backward method [100] which was applied to saddle point problems under the name of Optimistic Mirror Descent and to GAN training under the name of Optimistic Gradient Descent Ascent [54, 55, 89].

The first method treating general coupling functions with an asymmetric scheme is the Accelerated Primal-Dual Algorithm (APD) by [75], involving an optimistic gradient ascent step in one component which is followed by a gradient descent step in the other one. In the special case of a bilinear coupling function APD recovers the Primal-Dual Hybrid Gradient Method (PDHG) [44]. In the case of the minimax objective being strongly convex-concave acceleration of PDHG is obtained in [44], which is also done for APD in [75], however only under the rather limiting assumption of linearity of the coupling function in one component.

In the following we introduce a novel algorithm *OGAProx* for solving a convex-concave saddle point problem, where the convex-concave coupling function is smooth in one variable and nonsmooth in the other, and it is augmented by a nonsmooth regulariser in the smooth component. OGAProx consists of an optimistic gradient ascent step in the smooth component of the coupling function combined with a proximal step of the regulariser, which is followed by a proximal step of the coupling function in the nonsmooth component. We will be also able to accelerate our method in the convex-strongly concave setting *without* linearity assumption on the coupling function. Furthermore, we prove linear convergence if the problem is strongly convex-strongly concave, yielding similar results as for PDHG [44] in the bilinear case.

So far in most works nonsmoothness is only introduced via regularisers, as the coupling function is typically accessed through gradient evaluations. Recently there is another development, although with the saddle point problem *not* being convex-concave, where the assumption on differentiability of the coupling function in both components is weakened to only one component [23]. As the evaluation of the proximal mapping does not require differentiability we will assume the coupling function to be smooth in only one component, too.

The remainder of the chapter is organised as follows. Next we will introduce the precise problem formulation and the setting we will work with, formulate the proposed algorithm OGAProx and state our contributions. Afterwards we will discuss the properties of our algorithm in the convex-concave and convex-strongly concave setting and state respective convergence results in Section 5.2. After that we will investigate the convergence of the method under the additional assumption of strong convexity-strong concavity in

Section 5.3. The chapter will be concluded by numerical experiments in Section 5.4, where we treat a simple nonsmooth-linear saddle point problem, the training of multi kernel support vector machines (with an application to a classification problem in the field of Digital Humanities) and a classification problem taking into account minimax group fairness.

5.1.1. Problem description

Consider the saddle point problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Psi(x,y) := \Phi(x,y) - g(y), \tag{5.1}$$

where $\mathcal{X}$, $\mathcal{Y}$ are real Hilbert spaces, $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is a coupling function with $\operatorname{dom}\Phi := \{(x,y) \in \mathcal{X} \times \mathcal{Y} \mid \Phi(x,y) < +\infty\} \neq \emptyset$ and $g : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ a regulariser. Throughout this chapter (unless otherwise specified) we will make the following assumptions:

- g is proper, lower semicontinuous and convex with modulus $\nu \geq 0$, i.e. $g - \frac{\nu}{2}\|\cdot\|^2$ is convex (notice that we also allow and consider the situation $\nu = 0$, in which case g is convex; otherwise g is strongly convex);
- for all $y \in \operatorname{dom} g$, $\Phi(\cdot, y) : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is proper, convex and lower semicontinuous;
- for all $x \in \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi) := \{u \in \mathcal{X} \mid \exists y \in \mathcal{Y} \text{ such that } (u,y) \in \operatorname{dom}\Phi\}$ we have that $\operatorname{dom}\Phi(x,\cdot) = \mathcal{Y}$ and $\Phi(x,\cdot) : \mathcal{Y} \to \mathbb{R}$ is concave and Fréchet differentiable. Moreover, $\operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi)$ is closed;
- there exist L_{yx}, $L_{yy} \geq 0$ such that for all (x,y), $(x',y') \in \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi) \times \operatorname{dom} g$ the following holds:

$$\left\|\nabla_y \Phi(x,y) - \nabla_y \Phi(x',y')\right\| \leq L_{yx}\left\|x - x'\right\| + L_{yy}\left\|y - y'\right\|. \tag{5.2}$$

By convention we set $+\infty - (+\infty) := +\infty$. Thus, the situation can be summarised by

$$\Psi(x,y) = \begin{cases} -\infty & \text{if } x \in \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi) \text{ and } y \notin \operatorname{dom} g, \\ \Phi(x,y) - g(y) & \text{if } x \in \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi) \text{ and } y \in \operatorname{dom} g, \\ +\infty & \text{if } x \notin \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi). \end{cases} \tag{5.3}$$

We are interested in finding a *saddle point* of (5.1), which is a point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ that fulfils the inequalities

$$\Psi(x^*, y) \leq \Psi(x^*, y^*) \leq \Psi(x, y^*) \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{5.4}$$

For the remainder we assume that such a saddle point exists.

The assumptions considered above ensure that for any saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ we have

$$x^* \in \mathrm{Pr}_{\mathcal{X}}(\mathrm{dom}\,\Phi), \quad y^* \in \mathrm{dom}\, g \quad \text{and} \quad \Psi(x^*, y^*) = \Phi(x^*, y^*) - g(y^*) \in \mathbb{R}.$$

Finding a saddle point of (5.1) amounts to solving the necessary and sufficient first order optimality conditions, given by the following coupled inclusion problems

$$0 \in \partial\left[\Phi(\,\cdot\,, y^*)\right](x^*) \quad \text{and} \quad 0 \in -\nabla_y \Phi(x^*, y^*) + \partial g(y^*).$$

Remark 5.1.1. In case Φ and g have full domain, Ψ is a convex-concave function with full domain and the set $\mathrm{Pr}_{\mathcal{X}}(\mathrm{dom}\,\Phi)$ is obviously closed. However, in order to allow more flexibility and to cover a wider range of problems (see also the last section with numerical experiments), our investigations are carried out in the more general setting given by the assumptions described above. Furthermore, these assumptions allow us to stay in the rigorous setting of the theory of convex-concave saddle functions as described by Rockafellar in [129] (see Definition 5.2.2 and Proposition 5.2.3 below).

Remark 5.1.2. Consider the nonsmooth convex optimisation with inequality constraints

$$\begin{aligned} \min_{x \in \mathcal{H}} \quad & f(x) \\ \text{s.t.} \quad & h_i(x) \leq 0 \quad i = 1, ..., m, \end{aligned} \tag{5.5}$$

where $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex and lower semicontinuous function and $h_i : \mathcal{H} \to \mathbb{R}$, $i = 1, ..., m$ are convex and continuous functions. The Lagrangian attached to (5.5) reads

$$L : \mathcal{H} \times \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}, \quad L(x, \lambda_1, ..., \lambda_m) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x).$$

Then the saddle point problem

$$\min_{x \in \mathcal{H}} \max_{(\lambda_1, ..., \lambda_m) \in \mathbb{R}^m_+} L(x, \lambda_1, ..., \lambda_m) = \min_{x \in \mathcal{H}} \max_{(\lambda_1, ..., \lambda_m) \in \mathbb{R}^m} L(x, \lambda_1, ..., \lambda_m) - \delta_{\mathbb{R}^m_+}(\lambda_1, ..., \lambda_m) \tag{5.6}$$

exhibits the structure of saddle point problem (5.1). It is known that if $(x^*, \lambda_1^*, ..., \lambda_m^*)$ is a saddle point of (5.6), then x^* is an optimal solution of the constrained convex optimisation problem (5.5) and $(\lambda_1^*, ..., \lambda_m^*)$ is an optimal solution of its Lagrange dual.

5.1.2. Algorithm

The algorithm we investigate performs an optimistic gradient ascent step of Φ followed by an evaluation of the proximal mapping of g in the variable y, while it carries out a purely proximal step of Φ in x. We will call this method *Optimistic Gradient Ascent – Proximal Point algorithm* (OGAProx) in the following. For all $k \geq 0$ we define

$$\text{OGAProx:} \left\lfloor \begin{aligned} y_{k+1} &= \mathrm{prox}_{\sigma_k g}\left(y_k + \sigma_k\left[(1+\theta_k)\nabla_y \Phi(x_k, y_k) - \theta_k \nabla_y \Phi(x_{k-1}, y_{k-1})\right]\right), \\ x_{k+1} &= \mathrm{prox}_{\tau_k \Phi(\,\cdot\,, y_{k+1})}(x_k), \end{aligned} \right. \tag{5.7}$$

with the conventions $x_{-1} := x_0$ and $y_{-1} := y_0$ for starting points $x_0 \in \mathrm{Pr}_{\mathcal{X}}(\mathrm{dom}\,\Phi)$ and $y_0 \in \mathrm{dom}\, g$. The particular choices of the sequences $(\sigma_k)_{k\geq 0}$, $(\tau_k)_{k\geq 0} \subseteq \mathbb{R}_{++}$ and $(\theta_k)_{k\geq 0} \subseteq (0,1]$ will be specified later.

5.1.3. Contribution

Let us summarize the main results:

1. We introduce a novel algorithm to solve saddle point problems with nonsmooth coupling functions, which in general is *not* assumed to be linear in either component.

2. We prove for the saddle function Ψ being
   a) convex-concave (see Theorem 5.2.7):
      - weak convergence of the generated sequence $(x_k, y_k)_{k\geq 0}$ to a saddle point (x^*, y^*) as $k \to +\infty$;
      - convergence of the minimax gap $\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)$ to zero like $\mathcal{O}(1/K)$ as $K \to +\infty$, where $(\bar{x}_K)_{K\geq 1}$ and $(\bar{y}_K)_{K\geq 1}$ are the *ergodic sequences* obtained by averaging $(x_k)_{k\geq 1}$ and $(y_k)_{k\geq 1}$, respectively;
   b) convex-strongly concave (see Theorem 5.2.10):
      - strong convergence of $(y_k)_{k\geq 0}$ to y^* like $\mathcal{O}(1/k)$ as $k \to +\infty$;
      - convergence of the minimax gap $\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)$ to zero like $\mathcal{O}(1/K^2)$ as $K \to +\infty$;
   c) strongly convex-strongly concave (see Theorem 5.3.3):
      - linear convergence of $(x_k, y_k)_{k\geq 0}$ to (x^*, y^*) like $\mathcal{O}(\theta^k)$, with $0 < \theta < 1$, as $k \to +\infty$;
      - linear convergence of the minimax gap $\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)$ to zero like $\mathcal{O}(\theta^K)$ as $K \to +\infty$.

5.2. Convex-(strongly) concave setting

First we will treat the case when the coupling function Φ is convex-concave and g is convex with modulus $\nu \geq 0$. In the case $\nu = 0$ this corresponds to $\Psi(x, y) = \Phi(x, y) - g(y)$ being convex-concave, while for $\nu > 0$ the saddle function $\Psi(x, y)$ is convex-strongly concave.

We will start with stating two assumptions on the step sizes of the algorithm which will be needed in the convergence analysis. These will be followed by a unified preparatory analysis for general $\nu \geq 0$ that will be the base to show convergence of the iterates as well as of the minimax gap. After that we will introduce a choice of parameters that satisfy the aforementioned assumptions. The section will be closed by convergence results for the convex-concave ($\nu = 0$) and the convex-strongly concave ($\nu > 0$) setting.

Assumption 5.2.1. *We assume that the step sizes τ_k, σ_k and the momentum parameter θ_k satisfy*

$$\tau_{k+1} \geq \frac{\tau_k}{\theta_{k+1}} \quad \text{and} \quad \sigma_{k+1} \geq \frac{\sigma_k}{\theta_{k+1}(1+\nu\sigma_k)} \quad \text{for all } k \geq 0. \tag{5.8}$$

Furthermore, we assume that there exist $\delta > 0$ and $(\alpha_k)_{k\geq 0} \subseteq \mathbb{R}_{++}$ such that

$$\frac{1-\delta}{\tau_k} \geq \frac{L_{yx}}{\alpha_{k+1}} \quad \text{and} \quad \frac{1-\delta}{\sigma_k} \geq L_{yx}\alpha_k\theta_k + L_{yy}(1+\theta_k) \quad \text{for all } k \geq 0, \tag{5.9}$$

where $\theta_0 := 1$.

5.2.1. Preliminary considerations

In this subsection we will make some preliminary considerations that will play an important role when proving the convergence properties of the numerical scheme given by (5.7). For all $k \geq 0$ we will use the notations

$$q_k := \nabla_y \Phi(x_k, y_k) - \nabla_y \Phi(x_{k-1}, y_{k-1}) \text{ and } s_k := \theta_k q_k + \nabla_y \Phi(x_k, y_k). \tag{5.10}$$

We take an arbitrary $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and let $k \geq 0$ be fixed. From the first line of (5.7) we derive

$$0 \in \partial g(y_{k+1}) + \frac{1}{\sigma_k}(y_{k+1} - y_k) - s_k, \tag{5.11}$$

and, as g is convex with modulus ν, this implies

$$\begin{aligned} g(y) &\geq g(y_{k+1}) + \langle s_k, y - y_{k+1}\rangle + \frac{1}{\sigma_k}\langle y_k - y_{k+1}, y - y_{k+1}\rangle + \frac{\nu}{2}\|y - y_{k+1}\|^2 \\ &= g(y_{k+1}) + \langle s_k, y - y_{k+1}\rangle + \frac{1}{2\sigma_k}\Big(\|y_k - y_{k+1}\|^2 + \|y - y_{k+1}\|^2 \\ &\quad - \|y - y_k\|^2\Big) + \frac{\nu}{2}\|y - y_{k+1}\|^2. \end{aligned} \tag{5.12}$$

From the second line of (5.7) we get

$$0 \in \partial[\Phi(\,\cdot\,, y_{k+1})](x_{k+1}) + \frac{1}{\tau_k}(x_{k+1} - x_k), \tag{5.13}$$

hence the convexity of $\Phi(\,\cdot\,, y)$ for $y \in \operatorname{dom} g$ yields

$$\begin{aligned} \Phi(x, y_{k+1}) &\geq \Phi(x_{k+1}, y_{k+1}) + \frac{1}{\tau_k}\langle x_k - x_{k+1}, x - x_{k+1}\rangle \\ &= \Phi(x_{k+1}, y_{k+1}) + \frac{1}{2\tau_k}\Big(\|x_k - x_{k+1}\|^2 + \|x - x_{k+1}\|^2 - \|x - x_k\|^2\Big). \end{aligned} \tag{5.14}$$

Combining (5.12) and (5.14) we obtain

$$\begin{aligned}\Psi(x_{k+1},y)-\Psi(x,y_{k+1}) &= \Phi(x_{k+1},y)-g(y)-\Phi(x,y_{k+1})+g(y_{k+1})\\ &\le \Phi(x_{k+1},y)-\Phi(x,y_{k+1})+\langle s_k,y_{k+1}-y\rangle-\frac{\nu}{2}\|y-y_{k+1}\|^2\\ &\quad+\frac{1}{2\sigma_k}\left(-\|y_k-y_{k+1}\|^2-\|y-y_{k+1}\|^2+\|y-y_k\|^2\right)\\ &\le \Phi(x_{k+1},y)-\Phi(x_{k+1},y_{k+1})+\langle s_k,y_{k+1}-y\rangle-\frac{\nu}{2}\|y-y_{k+1}\|^2\\ &\quad+\frac{1}{2\tau_k}\left(-\|x_k-x_{k+1}\|^2-\|x-x_{k+1}\|^2+\|x-x_k\|^2\right)\\ &\quad+\frac{1}{2\sigma_k}\left(-\|y_k-y_{k+1}\|^2-\|y-y_{k+1}\|^2+\|y-y_k\|^2\right),\end{aligned}$$

which, together with the concavity of Φ in the second variable and (5.10), gives

$$\begin{aligned}\Psi(x_{k+1},y)-\Psi(x,y_{k+1}) \le\ & \theta_k\langle q_k,y_{k+1}-y\rangle-\frac{\nu}{2}\|y-y_{k+1}\|^2\\ &-\langle\nabla_y\Phi(x_{k+1},y_{k+1}),y_{k+1}-y\rangle+\langle\nabla_y\Phi(x_k,y_k),y_{k+1}-y\rangle\\ &+\frac{1}{2\tau_k}\left(-\|x_k-x_{k+1}\|^2-\|x-x_{k+1}\|^2+\|x-x_k\|^2\right)\\ &+\frac{1}{2\sigma_k}\left(-\|y_k-y_{k+1}\|^2-\|y-y_{k+1}\|^2+\|y-y_k\|^2\right)\\ =\ & -\langle q_{k+1},y_{k+1}-y\rangle+\theta_k\langle q_k,y_k-y\rangle-\frac{\nu}{2}\|y-y_{k+1}\|^2\\ &+\frac{1}{2\tau_k}\left(-\|x_k-x_{k+1}\|^2-\|x-x_{k+1}\|^2+\|x-x_k\|^2\right)\\ &+\frac{1}{2\sigma_k}\left(-\|y_k-y_{k+1}\|^2-\|y-y_{k+1}\|^2+\|y-y_k\|^2\right)\\ &+\theta_k\langle q_k,y_{k+1}-y_k\rangle.\end{aligned}\tag{5.15}$$

By using (5.2) we can evaluate the last term in the above expression as follows

$$\begin{aligned}|\langle q_k,y-y_k\rangle| &\le \|q_k\|\,\|y-y_k\| \le (L_{yx}\|x_k-x_{k-1}\|+L_{yy}\|y_k-y_{k-1}\|)\|y-y_k\|\\ &\le \frac{L_{yx}}{2}\left(\alpha_k\|y-y_k\|^2+\frac{1}{\alpha_k}\|x_k-x_{k-1}\|^2\right)+\frac{L_{yy}}{2}\left(\|y-y_k\|^2+\|y_k-y_{k-1}\|^2\right),\end{aligned}\tag{5.16}$$

with $\alpha_k>0$ chosen such that (5.9) holds.

Writing (5.16) for $y:=y_{k+1}$ and combining the resulting inequality with (5.15) we derive

$$\Psi(x_{k+1},y)-\Psi(x,y_{k+1})\le a_k(x,y)-b_{k+1}(x,y)-c_k,\tag{5.17}$$

where

$$\begin{aligned}a_k(x,y) :=\ & \frac{1}{2\tau_k}\|x-x_k\|^2+\frac{1}{2\sigma_k}\|y-y_k\|^2+\theta_k\langle q_k,y_k-y\rangle+\theta_k\frac{L_{yx}}{2\alpha_k}\|x_k-x_{k-1}\|^2\\ &+\theta_k\frac{L_{yy}}{2}\|y_k-y_{k-1}\|^2,\end{aligned}$$

$$b_{k+1}(x,y) := \frac{1}{2\tau_k}\|x - x_{k+1}\|^2 + \frac{1}{2}\left(\frac{1}{\sigma_k} + \nu\right)\|y - y_{k+1}\|^2 + \langle q_{k+1}, y_{k+1} - y\rangle + \frac{L_{yx}}{2\alpha_{k+1}}\|x_{k+1} - x_k\|^2 + \frac{L_{yy}}{2}\|y_{k+1} - y_k\|^2,$$

and

$$c_k := \frac{1}{2}\left(\frac{1}{\tau_k} - \frac{L_{yx}}{\alpha_{k+1}}\right)\|x_{k+1} - x_k\|^2 + \frac{1}{2}\left(\frac{1}{\sigma_k} - L_{yy} - \theta_k(L_{yx}\alpha_k + L_{yy})\right)\|y_{k+1} - y_k\|^2.$$

Now, let us define for all $k \geq 0$

$$t_k := \frac{\theta_0}{\theta_0\theta_1\cdots\theta_k} \tag{5.18}$$

and notice that

$$\frac{t_k}{t_{k+1}} = \theta_{k+1}.$$

Relation (5.8) from Assumption 5.2.1 is equivalent to

$$\frac{t_k}{\tau_k} \geq \frac{t_{k+1}}{\tau_{k+1}} \quad \text{and} \quad t_k\left(\frac{1}{\sigma_k} + \nu\right) \geq \frac{t_{k+1}}{\sigma_{k+1}}, \tag{5.19}$$

which will be used in telescoping arguments in the following.

Let $K \geq 1$ and denote

$$T_K := \sum_{k=0}^{K-1} t_k, \quad \bar{x}_K := \frac{1}{T_K}\sum_{k=0}^{K-1} t_k x_{k+1}, \quad \bar{y}_K := \frac{1}{T_K}\sum_{k=0}^{K-1} t_k y_{k+1}. \tag{5.20}$$

Multiplying both sides of (5.17) by $t_k > 0$ as defined in (5.18), followed by summing up the inequalities for $k = 0, \ldots, K-1$ gives

$$\sum_{k=0}^{K-1} t_k\left(\Psi(x_{k+1}, y) - \Psi(x, y_{k+1})\right) \leq \sum_{k=0}^{K-1} t_k\left(a_k(x,y) - b_{k+1}(x,y) - c_k\right).$$

By Jensen's inequality, as $\Psi(\,\cdot\,, y) - \Psi(x, \,\cdot\,)$ is a convex function, we obtain

$$T_K\left(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)\right) \leq \sum_{k=0}^{K-1} t_k\left(\Psi(x_{k+1}, y) - \Psi(x, y_{k+1})\right),$$

and thus

$$T_K\left(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)\right) \leq \sum_{k=0}^{K-1} t_k\left(a_k(x,y) - b_{k+1}(x,y) - c_k\right). \tag{5.21}$$

Furthermore, using (5.19), we get for all $k \geq 0$

$$\begin{aligned} t_k b_{k+1}(x,y) &= \frac{t_k}{2\tau_k}\|x - x_{k+1}\|^2 + \frac{t_k}{2}\left(\frac{1}{\sigma_k} + \nu\right)\|y - y_{k+1}\|^2 + t_k\langle q_{k+1}, y_{k+1} - y\rangle \\ &\quad + t_k\frac{L_{yx}}{2\alpha_{k+1}}\|x_{k+1} - x_k\|^2 + t_k\frac{L_{yy}}{2}\|y_{k+1} - y_k\|^2 \\ &\geq \frac{t_{k+1}}{2\tau_{k+1}}\|x - x_{k+1}\|^2 + \frac{t_{k+1}}{2\sigma_{k+1}}\|y - y_{k+1}\|^2 + t_{k+1}\theta_{k+1}\langle q_{k+1}, y_{k+1} - y\rangle \\ &\quad + t_{k+1}\theta_{k+1}\frac{L_{yx}}{2\alpha_{k+1}}\|x_{k+1} - x_k\|^2 + t_{k+1}\theta_{k+1}\frac{L_{yy}}{2}\|y_{k+1} - y_k\|^2 \\ &= t_{k+1}a_{k+1}(x,y). \end{aligned}$$

Notice that by (5.9) in Assumption 5.2.1 there exists $\delta > 0$ such that for all $k \geq 0$

$$c_k \geq \delta\left(\frac{1}{2\tau_k}\|x_{k+1} - x_k\|^2 + \frac{1}{2\sigma_k}\|y_{k+1} - y_k\|^2\right) \geq 0. \tag{5.22}$$

For the following recall that $x_{-1} = x_0$ and $y_{-1} = y_0$, which implies $q_0 = 0$. By using the above two inequalities in (5.21) and writing (5.16) for $k = K$ we obtain

$$\begin{aligned} \Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K) &\leq \frac{1}{T_K}\sum_{k=0}^{K-1}\left(t_k a_k(x,y) - t_{k+1}a_{k+1}(x,y)\right) \\ &= \frac{1}{T_K}\left(t_0 a_0(x,y) - t_K a_K(x,y)\right) = \frac{1}{T_K}\left(\frac{t_0}{2\tau_0}\|x - x_0\|^2 + \frac{t_0}{2\sigma_0}\|y - y_0\|^2\right) \\ &\quad - \frac{t_K}{T_K}\left(\frac{1}{2\tau_K}\|x - x_K\|^2 + \frac{1}{2\sigma_K}\|y - y_K\|^2\right) \\ &\quad - \frac{t_K\theta_K}{T_K}\left(\langle q_K, y_K - y\rangle + \frac{L_{yx}}{2\alpha_K}\|x_K - x_{K-1}\|^2 + \frac{L_{yy}}{2}\|y_K - y_{K-1}\|^2\right) \\ &\leq \frac{1}{T_K}\left(\frac{t_0}{2\tau_0}\|x - x_0\|^2 + \frac{t_0}{2\sigma_0}\|y - y_0\|^2\right) \\ &\quad - \frac{t_K}{T_K}\left(\frac{1}{2\tau_K}\|x - x_K\|^2 + \frac{1}{2}\left(\frac{1}{\sigma_K} - \theta_K(L_{yx}\alpha_K + L_{yy})\right)\|y - y_K\|^2\right). \end{aligned} \tag{5.23}$$

By definition we have $t_0 = 1$ and by (5.9) that the last term of the above inequality is nonpositive, hence the following estimate for the minimax gap function evaluated at the ergodic sequences holds

$$\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K) \leq \frac{1}{T_K}\left(\frac{1}{2\tau_0}\|x - x_0\|^2 + \frac{1}{2\sigma_0}\|y - y_0\|^2\right) \ \forall K \geq 1. \tag{5.24}$$

With these considerations at hand – in specific we want to point out (5.17), (5.23) and (5.24) – we will be able to obtain convergence statements for the two settings $\nu = 0$ and $\nu > 0$.

In the following definition we adjust the term *proper* to the saddle point setting and refer to [129] for further considerations related to saddle functions.

Definition 5.2.2. A function $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ is called a *saddle function*, if $\Psi(\cdot\,, y)$ is convex for all $y \in \mathcal{Y}$ and $\Psi(x, \cdot\,)$ is concave for all $x \in \mathcal{Y}$. A saddle function Ψ is called *proper*, if there exists $(x', y') \in \mathcal{X} \times \mathcal{Y}$ such that $\Psi(x', y) < +\infty$ for all $y \in \mathcal{Y}$ and $-\infty < \Psi(x, y')$ for all $x \in \mathcal{X}$.

We conclude this preliminary subsection with a useful result regarding the minimax objective from (5.1).

Proposition 5.2.3. *The function $\Psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ defined via (5.3) is a proper saddle function (see Definition 5.2.2) such that $\Psi(\cdot\,, y)$ is lower semicontinuous for each $y \in \mathcal{Y}$ and $\Psi(x, \cdot\,)$ is upper semicontinuous for each $x \in \mathcal{X}$. Consequently, the operator*

$$(x, y) \mapsto \partial[\Psi(\cdot\,, y)](x) \times \partial[-\Psi(x, \cdot\,)](y)$$

is maximal monotone.

Proof. We choose $(x', y') \in \mathcal{X} \times \mathcal{Y}$ and distinguish four cases.

Firstly, we look at the case $y' \notin \operatorname{dom} g$. Then

$$\Psi(x, y') = \begin{cases} -\infty & \text{if } x \in \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi), \\ +\infty & \text{if } x \notin \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi), \end{cases}$$

thus $x \mapsto \Psi(x, y')$ is convex and lower semicontinuous, since $\operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi)$ is convex and closed. Secondly, if $y' \in \operatorname{dom} g$, then $g(y') \in \mathbb{R}$ and

$$\Psi(x, y') = \Phi(x, y') - g(y') \ \forall x \in \mathcal{X},$$

which means that $x \mapsto \Psi(x, y')$ is convex and lower semicontinuous. This proves that $\Psi(\cdot\,, y)$ is convex and lower semicontinuous for all $y \in \mathcal{Y}$.

On the other hand, if $x' \notin \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi)$, then

$$\Psi(x', y) = +\infty \ \forall y \in \mathcal{Y},$$

which means that $y \mapsto \Psi(x', y)$ is upper semicontinuous and concave. Finally, if $x' \in \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi)$, then

$$\Phi(x', y) \in \mathbb{R} \text{ and } -\Psi(x', y) = -\Phi(x', y) + g(y) \ \forall y \in \mathcal{Y}.$$

Hence $y \mapsto -\Psi(x', y)$ is proper, convex and lower semicontinuous, and so $y \mapsto \Psi(x', y)$ is concave and upper semicontinuous. This proves that $\Psi(x, \cdot\,)$ is concave and upper semicontinuous for all $x \in \mathcal{X}$.

Moreover, Ψ is a proper saddle function. By assumption we have $g(y) > -\infty$ for all $y \in \mathcal{Y}$ and there exists $x' \in \operatorname{Pr}_{\mathcal{X}}(\operatorname{dom}\Phi) \neq \emptyset$ such that $\operatorname{dom}\Phi(x', \cdot\,) = \mathcal{Y}$. Thus

$$\Psi(x', y) = \Phi(x', y) - g(y) < +\infty \ \forall y \in \mathcal{Y}.$$

Furthermore, by assumption there exist $y' \in \operatorname{dom} g \subseteq \mathcal{Y}$ such that $g(y') < +\infty$ and for all $x \in \mathcal{X}$ we have $\Phi(x, y') > -\infty$. Hence,

$$\Psi(x, y') = \Phi(x, y') - g(y') > -\infty \ \forall x \in \mathcal{X}.$$

The maximal monotonicity of $(x, y) \mapsto \partial[\Psi(\cdot\,, y)](x) \times \partial[-\Psi(x, \cdot\,)](y)$ follows from Corollary 1 and Theorem 3 in [129, pages 248-249]. □

5.2.2. Fulfilment of step size assumptions

In this subsection we will investigate a particular choice of parameters to fulfil Assumption 5.2.1 which is suitable for both cases of $\nu = 0$ and $\nu > 0$.

Proposition 5.2.4. *Let $\nu \geq 0$, $c_\alpha > L_{yx} \geq 0$, $\theta_0 = 1$ and τ_0, $\sigma_0 > 0$ such that*

$$(c_\alpha L_{yx}\tau_0 + 2L_{yy})\,\sigma_0 < 1.$$

We define

$$\theta_{k+1} := \frac{1}{\sqrt{1+\nu\sigma_k}}, \quad \tau_{k+1} := \frac{\tau_k}{\theta_{k+1}}, \quad \sigma_{k+1} := \theta_{k+1}\sigma_k \quad \text{for all } k \geq 0. \tag{5.25}$$

Then the sequence $(\tau_k)_{k\geq 0}$, $(\sigma_k)_{k\geq 0}$ and $(\theta_k)_{k\geq 0}$ fulfil (5.8) in Assumption 5.2.1 with equality and (5.9) for

$$\alpha_k := \begin{cases} c_\alpha\tau_0 & \text{if } k = 0, \\ c_\alpha\tau_{k-1} & \text{if } k \geq 1, \end{cases} \tag{5.26}$$

and

$$\delta := \min\left\{1 - \frac{L_{yx}}{c_\alpha},\, 1 - (c_\alpha L_{yx}\tau_0 + 2L_{yy})\,\sigma_0\right\} > 0. \tag{5.27}$$

Furthermore, for $(t_k)_{k\geq 0}$ defined as in (5.18) we have

$$t_k = \frac{\theta_0}{\theta_0\theta_1\cdots\theta_k} = \frac{\tau_k}{\tau_0} \quad \forall k \geq 0. \tag{5.28}$$

Proof. First, we show that the particular choice (5.25) fulfils (5.8) in Assumption 5.2.1 with equality. We see that for all $k \geq 0$

$$\tau_{k+1} = \frac{\tau_k}{\theta_{k+1}},$$

as well as

$$\sigma_{k+1} = \theta_{k+1}\sigma_k = \frac{\sigma_k}{\theta_{k+1}\frac{1}{\theta_{k+1}^2}} = \frac{\sigma_k}{\theta_{k+1}(1+\nu\sigma_k)},$$

follow straight forward by definition.

Next, we show that (5.9) in Assumption 5.2.1 holds for δ defined in (5.27) with the choices (5.25) and (5.26). The first inequality of (5.9) is equivalent to

$$1 - \delta \geq \frac{L_{yx}}{\alpha_{k+1}}\tau_k = \frac{L_{yx}}{c_\alpha} \quad \forall k \geq 0,$$

which clearly is fulfilled as

$$\delta \leq 1 - \frac{L_{yx}}{c_\alpha}.$$

On the other hand, the second inequality of (5.9) is equivalent to

$$1 - \delta \geq L_{yx}\alpha_k\theta_k\sigma_k + L_{yy}(1+\theta_k)\sigma_k \quad \forall k \geq 0.$$

By definition of the step size parameters (5.25) we have for all $k \geq 0$

$$\tau_{k+1}\sigma_{k+1} = \tau_0\sigma_0, \quad \theta_{k+1} \leq 1 = \theta_0, \quad \sigma_{k+1} \leq \sigma_0, \quad \theta_{k+1}\tau_{k+1} = \tau_k,$$

and thus

$$\begin{aligned} 1 - \delta &\geq L_{yx}\alpha_0\theta_0\sigma_0 + L_{yy}(1+\theta_0)\sigma_0 = c_\alpha L_{yx}\tau_0\sigma_0 + 2L_{yy}\sigma_0 \\ &\geq c_\alpha L_{yx}\theta_{k+1}^2\tau_{k+1}\sigma_{k+1} + L_{yy}(1+\theta_{k+1})\sigma_{k+1} \\ &= L_{yx}c_\alpha\tau_k\theta_{k+1}\sigma_{k+1} + L_{yy}(1+\theta_{k+1})\sigma_{k+1} \\ &= L_{yx}\alpha_{k+1}\theta_{k+1}\sigma_{k+1} + L_{yy}(1+\theta_{k+1})\sigma_{k+1}. \end{aligned}$$

This chain of inequalities holds since

$$\delta \leq 1 - (c_\alpha L_{yx}\tau_0 + 2L_{yy})\,\sigma_0.$$

Finally, using the definition of t_k and (5.25) we conclude that for all $k \geq 0$

$$t_k = \frac{\theta_0}{\theta_0\theta_1\cdots\theta_k} = \frac{\frac{\tau_0}{\tau_0}}{\frac{\tau_0}{\tau_0}\frac{\tau_0}{\tau_1}\cdots\frac{\tau_{k-1}}{\tau_k}} = \frac{\tau_k}{\tau_0}.$$

□

Remark 5.2.5. The choice $L_{yy} = 0$ in (5.2) which was considered in [75] in the convex-strongly concave setting corresponds to the case when the coupling function Φ is linear in y. We will prove convergence also for L_{yy} positive, which makes our algorithm applicable to a much wider range of problems, as we will see in the section with the numerical experiments.

When the coupling function $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is bilinear, that is $\Phi(x, y) = \langle y, Ax\rangle$ for some nonzero continuous linear operator $A : \mathcal{X} \to \mathcal{Y}$ then we are in the setting of [44]. In this situation one can choose $L_{yy} = 0$ and $L_{yx} = \|A\|$, and (5.27) yields

$$\delta = \min\left\{1 - \frac{\|A\|}{c_\alpha},\, 1 - c_\alpha\|A\|\tau_0\sigma_0\right\},$$

with $c_\alpha > \|A\|$. To guarantee $\delta > 0$ we fix $0 < \varepsilon < 1$ and set

$$c_\alpha = (1-\varepsilon)^{-1}\|A\|.$$

Hence, we need to satisfy

$$\tau_0\sigma_0\|A\|^2 < 1 - \varepsilon,$$

which heavily resembles the step size condition of [44, Algorithm 2]. Since

$$\operatorname{prox}_{\gamma\Phi(\cdot,y)}(x) = x - \gamma A^* y$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and all $\gamma > 0$, our OGAProx scheme becomes the primal-dual algorithm PDHG from [44].

5.2.3. Convergence results

In this subsection we combine the preliminary considerations with the choice of parameters (5.25) from Proposition 5.2.4.

We will start with the case $\nu = 0$ and constant step sizes, which gives weak convergence of the iterates to a saddle point (x^*, y^*) and convergence of the minimax gap evaluated at the ergodic iterates to zero like $\mathcal{O}(1/K)$. Afterwards we will consider the case $\nu > 0$, which leads to an accelerated version of the algorithm with improved convergence results. In this setting we obtain convergence of $(y_k)_{k\geq 0}$ to y^* like $\mathcal{O}(1/K)$ and convergence of the minimax gap evaluated at the ergodic iterates to zero like $\mathcal{O}(1/K^2)$.

Convex-concave setting

For the following we assume that the function g is convex with modulus $\nu = 0$, meaning it is merely convex. Using the results of the previous subsection we will show that with the choice (5.25) all the parameters are constant.

Proposition 5.2.6. *Let $c_\alpha > L_{yx} \geq 0$ and $\tau, \sigma > 0$ such that*

$$(c_\alpha L_{yx}\tau + 2L_{yy})\,\sigma < 1.$$

If $\nu = 0$, then the sequences $(\tau_k)_{k\geq 0}$, $(\sigma_k)_{k\geq 0}$ and $(\theta_k)_{k\geq 0}$ as defined in Proposition 5.2.4 are constant, in particular we have

$$\tau_k = \tau_0 := \tau, \quad \sigma_k = \sigma_0 := \sigma, \quad \theta_k = \theta_0 = 1 \quad \text{for all } k \geq 0. \tag{5.29}$$

Proof. As $\nu = 0$, (5.25) gives for all $k \geq 0$

$$\theta_{k+1} = \frac{1}{\sqrt{1+\nu\sigma_k}} = 1, \quad \tau_{k+1} = \frac{\tau_k}{\theta_{k+1}} = \tau_0, \quad \sigma_{k+1} = \theta_{k+1}\sigma_k = \sigma_0.$$

□

Next we will state and prove the convergence results in the convex-concave case.

Theorem 5.2.7. *Let $c_\alpha > L_{yx} \geq 0$ and $\tau, \sigma > 0$ such that*

$$(c_\alpha L_{yx}\tau + 2L_{yy})\,\sigma < 1.$$

Then the sequence $(x_k, y_k)_{k\geq 0}$ generated by OGAProx with the choice of constant parameters as in Proposition 5.2.6, namely,

$$\tau_k = \tau_0 := \tau, \quad \sigma_k = \sigma_0 := \sigma, \quad \theta_k = \theta_0 = 1 \quad \text{for all } k \geq 0,$$

converges weakly to a saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ of (5.1). Furthermore, let $K \geq 1$ and denote

$$\bar{x}_K = \frac{1}{K}\sum_{k=0}^{K-1} x_{k+1} \quad \text{and} \quad \bar{y}_K = \frac{1}{K}\sum_{k=0}^{K-1} y_{k+1}.$$

Then for all $K \geq 1$ and any saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ of (5.1) we have

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right).$$

Proof. First we will show weak convergence of the sequence of iterates $(x_k, y_k)_{k\geq 0}$ to some saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ of (5.1). For this we will use the Opial Lemma (see Lemma 2.3.1).

Let $k \geq 0$ and $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be an arbitrary but fixed saddle point. From (5.17) together with the choice (5.29) of constant parameters $\theta_k = 1$, $\tau_k = \tau$, $\sigma_k = \sigma$ and $\alpha_k = \alpha$ we obtain

$$\begin{aligned} 0 \leq \Psi(x_{k+1}, y^*) - \Psi(x^*, y_{k+1}) &\leq a_k(x^*, y^*) - b_{k+1}(x^*, y^*) - c_k \\ &= a_k(x^*, y^*) - a_{k+1}(x^*, y^*) - c_k, \end{aligned} \tag{5.30}$$

since

$$\begin{aligned} a_k(x^*, y^*) = \frac{1}{2\tau}\|x^* - x_k\|^2 + \frac{1}{2\sigma}\|y^* - y_k\|^2 + \langle q_k, y_k - y^*\rangle + \frac{L_{yx}}{2\alpha}\|x_k - x_{k-1}\|^2 \\ + \frac{L_{yy}}{2}\|y_k - y_{k-1}\|^2 = b_k(x^*, y^*), \end{aligned} \tag{5.31}$$

and

$$c_k = \frac{1}{2}\left(\frac{1}{\tau} - \frac{L_{yx}}{\alpha}\right)\|x_{k+1} - x_k\|^2 + \frac{1}{2}\left(\frac{1}{\sigma} - L_{yy} - (L_{yx}\alpha + L_{yy})\right)\|y_{k+1} - y_k\|^2.$$

We see that (5.31), writing (5.16) with $y = y^*$ and (5.8) in Assumption 5.2.1 yield

$$a_k(x^*, y^*) \geq \frac{1}{2\tau}\|x^* - x_k\|^2 + \frac{1}{2\sigma}\left(1 - \sigma(L_{yx}\alpha + L_{yy})\right)\|y^* - y_k\|^2 \geq 0. \tag{5.32}$$

Furthermore, from (5.30) and (5.22) we deduce

$$a_k(x^*, y^*) \geq a_{k+1}(x^*, y^*) + \delta\left(\frac{1}{2\tau}\|x_{k+1} - x_k\|^2 + \frac{1}{2\sigma}\|y_{k+1} - y_k\|^2\right).$$

Telescoping this inequality and taking into account (5.32) give

$$\lim_{k\to+\infty}(x_{k+1} - x_k) = \lim_{k\to+\infty}(y_{k+1} - y_k) = 0, \tag{5.33}$$

as well as the existence of the limit $\lim_{k\to+\infty} a_k(x^*, y^*) \in \mathbb{R}$.

From (5.32) we get that $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ are bounded sequences. Moreover, by using (5.2) and (5.33) in definition (5.10) we obtain that

$$(q_k)_{k\geq 0} \text{ converges strongly to } 0. \tag{5.34}$$

From the definition of $a_k(x^*, y^*)$ in (5.31), (5.33) and (5.34) we derive that

$$\exists \lim_{k\to+\infty}\left(\frac{1}{2\tau}\|x_k - x^*\|^2 + \frac{1}{2\sigma}\|y_k - y^*\|^2\right) \in \mathbb{R}.$$

Since this is true for an arbitrary saddle point $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$, we have that the first statement of the Opial Lemma holds.

Next we will show that all weak cluster points of $(x_k, y_k)_{k\geq 0}$ are in fact saddle points of (5.1). Assume that $(x_{k_n})_{n\geq 0}$ converges weakly to $x^* \in \mathcal{X}$ and $(y_{k_n})_{n\geq 0}$ converges weakly to $y^* \in \mathcal{Y}$ as $n \to +\infty$. From (5.13), (5.10) and (5.11) we have

$$\begin{aligned}\left(\frac{1}{\tau}(x_{k_n} - x_{k_n+1}), \frac{1}{\sigma}(y_{k_n} - y_{k_n+1}) + q_{k_n} - q_{k_n+1}\right) \\ \in \partial[\Phi(\,\cdot\,, y_{k_n+1})](x_{k_n+1}) \times (-\nabla_y \Phi(x_{k_n+1}, y_{k_n+1}) + \partial g(y_{k_n+1})) \\ = \partial[\Psi(\,\cdot\,, y_{k_n+1})](x_{k_n+1}) \times \partial[-\Psi(x_{k_n+1}, \,\cdot\,)](y_{k_n+1}),\end{aligned} \tag{5.35}$$

where we used that for all $k \geq 0$ we have $x_k \in \mathrm{Pr}_{\mathcal{X}}(\mathrm{dom}\,\Phi)$ and $y_k \in \mathrm{dom}\, g$. The sequence on the left hand side of the inclusion (5.35) converges strongly to $(0,0)$ as $n \to +\infty$ (according to (5.33) and (5.34)). Notice that the operator $(x, y) \mapsto \partial[\Psi(\,\cdot\,, y)](x) \times \partial[-\Psi(x, \,\cdot\,)](y)$ is maximal monotone (see Proposition 5.2.3), hence its graph is sequentially closed with respect to the strong $\times$ weak topology. From here we deduce

$$(0,0) \in \partial[\Psi(\,\cdot\,, y^*)](x^*) \times \partial[-\Psi(x^*, \,\cdot\,)](y^*),$$

from which we easily derive that (x^*, y^*) is a saddle point as it satisfies (5.4). This means that also the second statement of the Opial Lemma is fulfilled and we have weak convergence of $(x_k, y_k)_{k\geq 0}$ to a saddle point (x^*, y^*).

The remaining part is to show the convergence rate of the minimax gap of the ergodic sequences. Let $K \geq 1$ and $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be an arbitrary but fixed saddle point. Writing (5.24) for (x^*, y^*) yields

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{T_K}\left(\frac{1}{2\tau}\left\|x^* - x_0\right\|^2 + \frac{1}{2\sigma}\left\|y^* - y_0\right\|^2\right),$$

with

$$T_K = \sum_{k=0}^{K-1} t_k, \quad \bar{x}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k x_{k+1}, \quad \bar{y}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k y_{k+1}.$$

Using (5.28) to get $t_k = 1$ for all $k \geq 0$ in the above expressions gives

$$T_K = \sum_{k=0}^{K-1} t_k = K, \quad \bar{x}_K = \frac{1}{K}\sum_{k=0}^{K-1} x_{k+1}, \quad \bar{y}_K = \frac{1}{K}\sum_{k=0}^{K-1} y_{k+1}.$$

Finally we derive for all $K \geq 1$

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{K}\left(\frac{1}{2\tau}\left\|x^* - x_0\right\|^2 + \frac{1}{2\sigma}\left\|y^* - y_0\right\|^2\right).$$

□

Convex-strongly concave setting

For the remainder of this section we assume that the function g is convex with modulus $\nu > 0$, meaning it is ν-strongly convex. In this case the choice (5.25) leads to adaptive parameters and accelerated convergence.

Proposition 5.2.8. *Let $c_\alpha > L_{yx} \geq 0$, $\theta_0 = 1$ and $\tau_0, \sigma_0 > 0$ such that*

$$(c_\alpha L_{yx}\tau_0 + 2L_{yy})\,\sigma_0 < 1.$$

If $\nu > 0$ then $(\tau_k)_{k\geq 0}$, $(\sigma_k)_{k\geq 0}$ and $(\theta_k)_{k\geq 0}$ as defined in Proposition 5.2.4 are adaptive, in particular we have

$$\theta_{k+1} = \frac{1}{\sqrt{1+\nu\sigma_k}} < 1, \quad \tau_{k+1} = \frac{\tau_k}{\theta_{k+1}} > \tau_k, \quad \sigma_{k+1} = \theta_{k+1}\sigma_k < \sigma_k \quad \text{for all } k \geq 0. \tag{5.36}$$

Proof. The statements follow directly from Proposition 5.2.4 for $\nu > 0$. □

To obtain statements regarding the (accelerated) convergence rates in the convex-strongly concave setting, we look at the behaviour of the sequences of step size parameters $(\tau_k)_{k\geq 0}$ and $(\sigma_k)_{k\geq 0}$ for $k \to +\infty$.

Proposition 5.2.9. *Let $\theta_0 = 1$, $\tau_0 > 0$,*

$$0 < \sigma_0 \leq \frac{9+3\sqrt{13}}{2\nu},$$

and for all $k \geq 0$ denote

$$\gamma_k := \frac{\tau_k}{\sigma_k}.$$

Then with the choice of adaptive parameters (5.36) we have for all $k \geq 0$

$$\gamma_k \geq \frac{\nu^2\tau_0\sigma_0}{9}k^2 \quad \text{and} \quad \tau_k \geq \frac{\nu\tau_0\sigma_0}{3}k,$$

and for all $k \geq 1$

$$\sigma_k \leq \frac{3}{\nu}\frac{1}{k}.$$

Proof. By (5.25) we conclude that for all $k \geq 0$

$$\gamma_{k+1} = \gamma_k(1+\nu\sigma_k),$$

and further

$$\sigma_{k+1} = \sigma_k\sqrt{\frac{\gamma_k}{\gamma_{k+1}}},$$

which, applied recursively, gives

$$\sigma_k = \sigma_0\sqrt{\frac{\gamma_0}{\gamma_k}} = \sqrt{\tau_0\sigma_0}\frac{1}{\sqrt{\gamma_k}}.$$

We obtain

$$\gamma_{k+1} = \gamma_k(1 + \nu\sigma_k) = \gamma_k + \nu\sqrt{\tau_0\sigma_0}\sqrt{\gamma_k},$$

which we will use to show by induction that for all $k \geq 0$

$$\gamma_k \geq \frac{\nu^2\tau_0\sigma_0}{9}k^2. \tag{5.37}$$

For $k = 0$ the statement trivially holds, whereas for $k = 1$ we need to verify that

$$\gamma_1 = \gamma_0 + \nu\sqrt{\tau_0\sigma_0}\sqrt{\gamma_0} = \frac{\tau_0}{\sigma_0}(1 + \nu\sigma_0) \geq \frac{\nu^2\tau_0\sigma_0}{9},$$

which is equivalent to the following quadratic inequality

$$\sigma_0^2 - \frac{9}{\nu}\sigma_0 - \frac{9}{\nu^2} \leq 0,$$

and guaranteed to hold by our initial choice of $\sigma_0 > 0$. Now let $k \geq 1$ and assume that (5.37) holds. Then

$$\gamma_{k+1} = \gamma_k + \nu\sqrt{\tau_0\sigma_0}\sqrt{\gamma_k} \geq \frac{\nu^2\tau_0\sigma_0}{9}k^2 + \frac{\nu^2\tau_0\sigma_0}{3}k \geq \frac{\nu^2\tau_0\sigma_0}{9}(k+1)^2.$$

This shows the validity of (5.37) for all $k \geq 0$.

Now we can use inequality (5.37) to deduce the convergence behaviour of the sequences $(\tau_k)_{k\geq 0}$ and $(\sigma_k)_{k\geq 0}$ for $k \to +\infty$. We get for all $k \geq 0$

$$\tau_k = \sigma_k\gamma_k = \sqrt{\tau_0\sigma_0}\sqrt{\gamma_k} \geq \frac{\nu\tau_0\sigma_0}{3}k, \tag{5.38}$$

which, combined with

$$\tau_k\sigma_k = \frac{\tau_k^2}{\gamma_k} = \tau_0\sigma_0,$$

gives for all $k \geq 1$

$$\sigma_k \leq \frac{3}{\nu}\frac{1}{k}.$$

□

Now we are ready to prove the convergence results in the convex-strongly concave setting.

Theorem 5.2.10. *Let $c_\alpha > L_{yx} \geq 0$, $\theta_0 = 1$ and τ_0, $\sigma_0 > 0$ such that*

$$(c_\alpha L_{yx}\tau_0 + 2L_{yy})\,\sigma_0 < 1 \quad \text{and} \quad 0 < \sigma_0 \leq \frac{9 + 3\sqrt{13}}{2\nu}.$$

Let $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be a saddle point of (5.1). Then for $(x_k, y_k)_{k\geq 0}$ being the sequence generated by OGAProx with the choice of adaptive parameters

$$\theta_{k+1} = \frac{1}{\sqrt{1 + \nu\sigma_k}} < 1, \quad \tau_{k+1} = \frac{\tau_k}{\theta_{k+1}} > \tau_k, \quad \sigma_{k+1} = \theta_{k+1}\sigma_k < \sigma_k \quad \text{for all } k \geq 0,$$

we have for all $K \geq 1$

$$\|y^* - y_K\| \leq \frac{c_1}{K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right)^{\frac{1}{2}},$$

with $c_1 := \sqrt{\frac{18}{\nu^2\sigma_0\delta}}$, *where* $\delta > 0$ *is defined in* (5.27). *Furthermore, for* $K \geq 1$, *denote*

$$T_K = \sum_{k=0}^{K-1} t_k, \quad \bar{x}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k x_{k+1}, \quad \bar{y}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k y_{k+1},$$

where $t_k = \frac{\tau_k}{\tau_0}$ *for all* $k \geq 0$ *(see also* (5.28)*). Then for all* $K \geq 2$ *we get*

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{c_2}{K^2}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right),$$

with $c_2 := \frac{12}{\nu\sigma_0}$.

Proof. Let $K \geq 1$ and let $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be an arbitrary but fixed saddle point. First we will prove the convergence rate of the sequence of iterates $(y_k)_{k\geq 0}$. Plugging the particular choice of parameters (5.36) into (5.23) for (x^*, y^*), we obtain

$$\begin{aligned}
&\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2 \\
&\quad \geq \frac{1}{2\tau_0}\|x^* - x_K\|^2 + \frac{\tau_K}{\sigma_K}\left(1 - \sigma_K\theta_K(L_{yx}\alpha_K + L_{yy})\right)\frac{1}{2\tau_0}\|y^* - y_K\|^2 \\
&\quad \geq \gamma_K\frac{\delta}{2\tau_0}\|y^* - y_K\|^2,
\end{aligned}$$

where we use (5.9) in Assumption 5.2.1 for the last inequality. Combining this with (5.37) we derive

$$\|y^* - y_K\| \leq \frac{c_1}{K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right)^{\frac{1}{2}},$$

with $c_1 := \sqrt{\frac{18}{\nu^2\sigma_0\delta}}$.

Next we will show the convergence rate of the minimax gap at the ergodic sequences. Writing (5.24) for (x^*, y^*), we obtain

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{1}{T_K}\left(\frac{1}{2\tau_0}\|x^* - x_0\|^2 + \frac{1}{2\sigma_0}\|y^* - y_0\|^2\right). \tag{5.39}$$

Plugging the particular choice of $t_k = \frac{\tau_k}{\tau_0}$ for all $k \geq 0$ from (5.28) into the definition of T_K, together with (5.38) yields

$$T_K = \frac{1}{\tau_0}\sum_{k=0}^{K-1}\tau_k \geq \frac{\nu\sigma_0}{3}\sum_{k=0}^{K-1} k = \frac{\nu\sigma_0}{6}K(K-1).$$

Combining this inequality with (5.39), we obtain for all $K \geq 2$

$$0 \leq \Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K) \leq \frac{c_2}{K^2} \left(\frac{1}{2\tau_0} \|x^* - x_0\|^2 + \frac{1}{2\sigma_0} \|y^* - y_0\|^2 \right),$$

with $c_2 := \frac{12}{\nu\sigma_0}$, which concludes the proof. □

5.3. Strongly convex-strongly concave setting

For this section we assume that the function g is convex with modulus $\nu > 0$, meaning it is ν-strongly convex. In addition to the assumptions we had until now, for this section we also assume that for all $y \in \operatorname{dom} g$ the function $\Phi(\,\cdot\,, y) : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is μ-strongly convex with modulus $\mu > 0$. This means that the saddle function $(x, y) \mapsto \Psi(x, y)$ is strongly convex-strongly concave.

As in the previous section we will state two step size assumptions that will be needed for the convergence analysis. These again will be followed by preparatory observations and a result to guarantee the validity of the stated assumptions. The section will be closed with the formulation and proof of convergence results.

Assumption 5.3.1. *We assume that the step sizes τ_k, σ_k and the momentum parameter θ_k are constant*

$$\theta_k = \theta_0 =: \theta, \quad \tau_k = \tau_0 =: \tau, \quad \sigma_k = \sigma_0 =: \sigma \quad \forall k \geq 0,$$

and satisfy

$$1 + \mu\tau = \frac{1}{\theta}, \quad 1 + \nu\sigma = \frac{1}{\theta}, \tag{5.40}$$

with

$$0 < \theta < 1. \tag{5.41}$$

Furthermore, we assume that there exists $\alpha > 0$ such that

$$\frac{L_{yx}}{\alpha} \leq \frac{1}{\tau}, \quad L_{yy} \leq \frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{\sigma}, \tag{5.42}$$

with

$$1 - \theta\sigma(\alpha L_{yx} + L_{yy}) > 0. \tag{5.43}$$

5.3.1. Preliminary considerations

We take an arbitrary $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and let $k \geq 0$. Following similar considerations along (5.12)-(5.14), additionally taking into account the μ-strong convexity of $\Phi(\,\cdot\,, y)$ for

$y \in \operatorname{dom} g$, instead of (5.15) we derive

$$\begin{aligned}
\Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) \leq\ & \theta \langle q_k, y_k - y\rangle - \langle q_{k+1}, y_{k+1} - y\rangle \\
& - \frac{\mu}{2}\|x - x_{k+1}\|^2 - \frac{\nu}{2}\|y - y_{k+1}\|^2 + \theta \langle q_k, y_{k+1} - y_k\rangle \\
& + \frac{1}{2\tau}\left(-\|x_k - x_{k+1}\|^2 - \|x - x_{k+1}\|^2 + \|x - x_k\|^2\right) \\
& + \frac{1}{2\sigma}\left(-\|y_k - y_{k+1}\|^2 - \|y - y_{k+1}\|^2 + \|y - y_k\|^2\right) \\
\leq\ & \frac{1}{2\tau}\|x - x_k\|^2 + \frac{1}{2\sigma}\|y - y_k\|^2 + \theta \langle q_k, y_k - y\rangle \\
& - \frac{1 + \mu\tau}{2\tau}\|x - x_{k+1}\|^2 - \frac{1 + \nu\sigma}{2\sigma}\|y - y_{k+1}\|^2 - \langle q_{k+1}, y_{k+1} - y\rangle \\
& + \frac{\theta L_{yx}}{2\alpha}\|x_k - x_{k-1}\|^2 - \frac{1}{2\tau}\|x_{k+1} - x_k\|^2 \\
& + \frac{\theta L_{yy}}{2}\|y_k - y_{k-1}\|^2 - \frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{2\sigma}\|y_{k+1} - y_k\|^2 .
\end{aligned}$$

By (5.40) in Assumption 5.3.1 and for $\alpha > 0$ fulfilling (5.42)-(5.43), we obtain

$$\begin{aligned}
\Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) \leq\ & \frac{1}{2\tau}\|x - x_k\|^2 + \frac{1}{2\sigma}\|y - y_k\|^2 + \theta \langle q_k, y_k - y\rangle \\
& - \frac{1}{2\tau}\frac{1}{\theta}\|x - x_{k+1}\|^2 - \frac{1}{2\sigma}\frac{1}{\theta}\|y - y_{k+1}\|^2 - \langle q_{k+1}, y_{k+1} - y\rangle \\
& + \frac{\theta L_{yx}}{2\alpha}\|x_k - x_{k-1}\|^2 - \frac{1}{2\tau}\|x_{k+1} - x_k\|^2 \\
& + \frac{\theta L_{yy}}{2}\|y_k - y_{k-1}\|^2 - \frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{2\sigma}\|y_{k+1} - y_k\|^2 ,
\end{aligned}$$

which together with (5.42) and (5.43) gives

$$\begin{aligned}
\Psi(x_{k+1}, y) - \Psi(x, y_{k+1}) \leq\ & \frac{1}{2\tau}\left(\|x - x_k\|^2 - \frac{1}{\theta}\|x - x_{k+1}\|^2\right) \\
& + \frac{1}{2\sigma}\left(\|y - y_k\|^2 - \frac{1}{\theta}\|y - y_{k+1}\|^2\right) + \theta \langle q_k, y_k - y\rangle - \langle q_{k+1}, y_{k+1} - y\rangle \\
& + \frac{1}{2\tau}\left(\theta\|x_k - x_{k-1}\|^2 - \|x_{k+1} - x_k\|^2\right) + \frac{1}{2\tilde{\sigma}}\left(\theta\|y_k - y_{k-1}\|^2 - \|y_{k+1} - y_k\|^2\right),
\end{aligned} \tag{5.44}$$

where

$$\tilde{\sigma} := \frac{\sigma}{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}.$$

Let $K \geq 1$ and as in (5.20) denote

$$T_K = \sum_{k=0}^{K-1} t_k, \quad \bar{x}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k x_{k+1}, \quad \bar{y}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} t_k y_{k+1}.$$

with $t_k > 0$ defined as in (5.18), in other words

$$t_k = \theta^{-k} \quad \forall k \geq 0.$$

Multiplying both sides of (5.44) by $t_k > 0$ yields

$$\begin{aligned}
\frac{1}{\theta^k}\left(\Psi(x_{k+1}, y) - \Psi(x, y_{k+1})\right) \leq & \frac{1}{2\tau}\left(\frac{1}{\theta^k}\|x - x_k\|^2 - \frac{1}{\theta^{k+1}}\|x - x_{k+1}\|^2\right) \\
& + \frac{1}{2\sigma}\left(\frac{1}{\theta^k}\|y - y_k\|^2 - \frac{1}{\theta^{k+1}}\|y - y_{k+1}\|^2\right) \\
& + \frac{1}{\theta^{k-1}}\langle q_k, y_k - y\rangle - \frac{1}{\theta^k}\langle q_{k+1}, y_{k+1} - y\rangle \\
& + \frac{1}{2\tau}\left(\frac{1}{\theta^{k-1}}\|x_k - x_{k-1}\|^2 - \frac{1}{\theta^k}\|x_{k+1} - x_k\|^2\right) \\
& + \frac{1}{2\tilde{\sigma}}\left(\frac{1}{\theta^{k-1}}\|y_k - y_{k-1}\|^2 - \frac{1}{\theta^k}\|y_{k+1} - y_k\|^2\right).
\end{aligned}$$

Summing up the above inequality for $k = 0, \ldots, K-1$ and taking into account Jensen's inequality for the convex function $\Psi(\,\cdot\,, y) - \Psi(x, \,\cdot\,)$ give

$$\begin{aligned}
T_K\left(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)\right) &\leq \sum_{k=0}^{K-1}\frac{1}{\theta^k}\left(\Psi(x_{k+1}, y) - \Psi(x, y_{k+1})\right) \\
&\leq \frac{1}{2\tau}\left(\|x - x_0\|^2 - \frac{1}{\theta^K}\|x - x_K\|^2\right) + \frac{1}{2\sigma}\left(\|y - y_0\|^2 - \frac{1}{\theta^K}\|y - y_K\|^2\right) \\
&\quad - \frac{1}{\theta^{K-1}}\langle q_K, y_K - y\rangle - \frac{1}{\theta^{K-1}}\frac{1}{2\tau}\|x_K - x_{K-1}\|^2 - \frac{1}{\theta^{K-1}}\frac{1}{2\tilde{\sigma}}\|y_K - y_{K-1}\|^2 \\
&\leq \frac{1}{2\tau}\left(\|x - x_0\|^2 - \frac{1}{\theta^K}\|x - x_K\|^2\right) + \frac{1}{2\sigma}\left(\|y - y_0\|^2 - \frac{1}{\theta^K}\|y - y_K\|^2\right) \\
&\quad + \frac{1}{\theta^{K-1}}\frac{L_{yx}}{2}\left(\frac{1}{\alpha}\|x_K - x_{K-1}\|^2 + \alpha\|y_K - y\|^2\right) - \frac{1}{\theta^{K-1}}\frac{1}{2\tau}\|x_K - x_{K-1}\|^2 \\
&\quad + \frac{1}{\theta^{K-1}}\frac{L_{yy}}{2}\left(\|y_K - y_{K-1}\|^2 + \|y_K - y\|^2\right) - \frac{1}{\theta^{K-1}}\frac{1}{2\tilde{\sigma}}\|y_K - y_{K-1}\|^2 \\
&= \frac{1}{2\tau}\|x - x_0\|^2 + \frac{1}{2\sigma}\|y - y_0\|^2 \\
&\quad - \frac{1}{\theta^K}\frac{1}{2\tau}\|x - x_K\|^2 - \frac{1}{\theta^K}\frac{1 - \theta\sigma(\alpha L_{yx} + L_{yy})}{2\sigma}\|y - y_K\|^2 \\
&\quad - \frac{1}{2\theta^{K-1}}\left(\frac{1}{\tau} - \frac{L_{yx}}{\alpha}\right)\|x_K - x_{K-1}\|^2 - \frac{1}{2\theta^{K-1}}\left(\frac{1}{\tilde{\sigma}} - L_{yy}\right)\|y_K - y_{K-1}\|^2,
\end{aligned}$$

where in the second inequality we use (5.16). Omitting the last two terms which are non positive by (5.42), we obtain for all $K \geq 1$

$$\begin{aligned}
T_K\theta^K(\Psi(\bar{x}_K, y) - \Psi(x, \bar{y}_K)) + \frac{1}{2\tau}\|x - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y - y_K\|^2 \\
\leq \theta^K\left(\frac{1}{2\tau}\|x - x_0\|^2 + \frac{1}{2\sigma}\|y - y_0\|^2\right),
\end{aligned} \tag{5.45}$$

which we will use to obtain our convergence results in the following.

5.3.2. Fulfilment of step size assumptions

In this subsection we will investigate a particular choice of parameters τ, σ and θ such that Assumption 5.3.1 holds.

Proposition 5.3.2. *For $\alpha > 0$ define*

$$\tilde{\theta} := \max\left\{\frac{L_{yx}}{\alpha\mu + L_{yx}}, \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}}\right\}. \tag{5.46}$$

Let $\theta > 0$ such that

$$0 \leq \tilde{\theta} < \theta < 1, \tag{5.47}$$

and set

$$\tau = \frac{1}{\mu}\frac{1-\theta}{\theta} \quad \text{and} \quad \sigma = \frac{1}{\nu}\frac{1-\theta}{\theta}. \tag{5.48}$$

Then τ, σ and θ fulfil Assumption 5.3.1.

Proof. If $L_{yx} = L_{yy} = 0$, then the conclusion follows immediately. Assume that $L_{yx} + L_{yy} > 0$. It is easy to verify that definition (5.46) yields

$$0 < \tilde{\theta} < 1$$

and that (5.48) is equivalent to (5.40) where (5.41) is ensured by (5.47). Furthermore, plugging the specific form of the step sizes (5.48) into (5.42) we obtain for the first inequality of (5.42)

$$\frac{L_{yx}}{\alpha} \leq \frac{\mu\theta}{1-\theta},$$

which is equivalent to

$$\theta \geq \frac{L_{yx}}{\alpha\mu + L_{yx}}.$$

Note that by (5.47) we have

$$0 \leq \frac{L_{yx}}{\alpha\mu + L_{yx}} \leq \tilde{\theta} < \theta < 1.$$

Similarly, the second inequality of (5.42) is equivalent to the following quadratic inequality

$$\theta^2 - \frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}}\theta - \frac{L_{yy}}{\alpha L_{yx} + L_{yy}} \geq 0.$$

The non negative solution of the associated quadratic equation reads

$$\rho := \frac{1}{2}\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}} + \sqrt{\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}}\right)^2 + \frac{4L_{yy}}{\alpha L_{yx} + L_{yy}}}\right) \geq 0.$$

Since

$$0 \leq \rho < \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}} \leq \tilde{\theta} < \theta < 1,$$

the second inequality of (5.42) is also fulfilled. In order to see that

$$\rho < \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}},$$

we notice that this inequality is equivalent to

$$\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}}\right)^2 + \frac{4L_{yy}}{\alpha L_{yx} + L_{yy}} < \frac{(\nu^2 + 2\nu L_{yy} + (\alpha L_{yx} + 2L_{yy})^2)^2}{(\nu + \alpha L_{yx} + 2L_{yy})^2(\alpha L_{yx} + L_{yy})^2},$$

which holds if and only if

$$\begin{aligned}&((\alpha L_{yx} - \nu)^2 + 4L_{yy}(\alpha L_{yx} + L_{yy}))(\nu + \alpha L_{yx} + 2L_{yy})^2 \\ &< (\nu^2 + 2\nu L_{yy})^2 + 2(\nu^2 + 2\nu L_{yy})(\alpha L_{yx} + 2L_{yy})^2 + (\alpha L_{yx} + 2L_{yy})^4\end{aligned}$$

or, equivalently,

$$0 < 4\nu^2(\alpha L_{yx} + L_{yy})^2.$$

For the remaining condition (5.43) to hold we need to ensure

$$\theta > \frac{\alpha L_{yx} + L_{yy} - \nu}{\alpha L_{yx} + L_{yy}}.$$

For this we observe that

$$\begin{aligned}\rho &\geq \frac{1}{2}\left(\frac{\alpha L_{yx} - \nu}{\alpha L_{yx} + L_{yy}} + \sqrt{\left(\frac{\alpha L_{yx} + 2L_{yy} - \nu}{\alpha L_{yx} + L_{yy}}\right)^2}\right) \\ &\geq \frac{1}{2}\frac{\alpha L_{yx} - \nu + \alpha L_{yx} + 2L_{yy} - \nu}{\alpha L_{yx} + L_{yy}} = \frac{\alpha L_{yx} + L_{yy} - \nu}{\alpha L_{yx} + L_{yy}}.\end{aligned}$$

In conclusion, we obtain the following chain of inequalities

$$\frac{\alpha L_{yx} + L_{yy} - \nu}{\alpha L_{yx} + L_{yy}} \leq \rho < \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}} \leq \tilde{\theta} < \theta < 1,$$

which is satisfied by (5.47). □

5.3.3. Convergence results

Now we can combine the previous results and prove the convergence statements in the strongly convex-strongly concave setting.

Theorem 5.3.3. *Let $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be a saddle point of (5.1). Then for $(x_k, y_k)_{k\geq 0}$ being the sequence generated by OGAProx with the choice of parameters*

$$\tau = \frac{1}{\mu}\frac{1-\theta}{\theta}, \quad \sigma = \frac{1}{\nu}\frac{1-\theta}{\theta}, \quad 0 \leq \tilde{\theta} < \theta < 1,$$

with

$$\tilde{\theta} = \max\left\{\frac{L_{yx}}{\alpha\mu + L_{yx}}, \frac{\alpha L_{yx} + 2L_{yy}}{\nu + \alpha L_{yx} + 2L_{yy}}\right\},$$

for $\alpha > 0$, we denote for $K \geq 1$

$$T_K = \sum_{k=0}^{K-1} \theta^{-k}, \quad \bar{x}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} \theta^{-k} x_{k+1}, \quad \bar{y}_K = \frac{1}{T_K}\sum_{k=0}^{K-1} \theta^{-k} y_{k+1},$$

for which the following holds

$$\begin{aligned} 0 &\leq \theta(\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)) + \frac{1}{2\tau}\|x^* - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y^* - y_K\|^2 \\ &\leq \theta^K\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right), \end{aligned}$$

where $\tilde{\sigma} := \frac{\sigma}{1-\theta\sigma(\alpha L_{yx} + L_{yy})}$.

Proof. Let $K \geq 1$ and $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ be an arbitrary but fixed saddle point of (5.1). Writing (5.45) for (x^*, y^*) we get

$$\begin{aligned} 0 &\leq T_K\theta^K(\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)) + \frac{1}{2\tau}\|x^* - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y^* - y_K\|^2 \\ &\leq \theta^K\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right). \end{aligned}$$

Using

$$T_K = \sum_{k=0}^{K-1}\frac{1}{\theta^k} = \frac{1}{\theta^{K-1}}\frac{1-\theta^K}{1-\theta} \geq \frac{1}{\theta^{K-1}},$$

finally we obtain for all $K \geq 1$

$$\begin{aligned} 0 &\leq \theta(\Psi(\bar{x}_K, y^*) - \Psi(x^*, \bar{y}_K)) + \frac{1}{2\tau}\|x^* - x_K\|^2 + \frac{1}{2\tilde{\sigma}}\|y^* - y_K\|^2 \\ &\leq \theta^K\left(\frac{1}{2\tau}\|x^* - x_0\|^2 + \frac{1}{2\sigma}\|y^* - y_0\|^2\right), \end{aligned}$$

with $0 < \theta < 1$ as defined in (5.47). □

5.4. Numerical experiments

In this section we will treat three numerical applications of our method. The first one is of rather simple structure and has the purpose to highlight the convergence rates we obtained in the previous sections. The second one concerns multi kernel support vector machines to validate OGAProx on a more relevant application in practice, even though there are no theoretical guarantees for the "metric" reported there. The third numerical application addresses a classification problem incorporating minimax group fairness, which traces back to the solving of a minimax problem with nonsmooth coupling function.

5.4.1. Nonsmooth-linear problem

The first application we treat is to showcase the convergence rates we obtained in the previous sections and make a simple proof of concept. We look at the following nonsmooth-linear saddle point problem

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n}\Psi(x,y) := \langle [x]_+, Ay\rangle - \left(\delta_C(y) + \frac{\nu}{2}\|y\|^2\right), \tag{5.49}$$

with $\nu \geq 0$ and $A \in \mathbb{R}^{d\times n}$, $[\,\cdot\,]_+$ being the component-wise positive part,

$$[x]_+ = (\max\{0, x_i\})_{i=1}^d,$$

and C being the following convex polytope

$$C := \{y \in \mathbb{R}^n \mid Ay \geqq 0\}.$$

For $u = (u_i)_{i=1}^d$, $v = (v_i)_{i=1}^d \in \mathbb{R}^d$ the relation $u \geqq v$ denotes component-wise inequalities, namely,

$$u \geqq v \;\Leftrightarrow\; u_i \geq v_i \quad \text{for } 1 \leq i \leq d.$$

Then $g : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ with

$$g(y) := \delta_C(y) + \frac{\nu}{2}\|y\|^2$$

is proper, lower semicontinuous and convex with modulus $\nu \geq 0$ and $\operatorname{dom} g = C$. Moreover, $\Phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ with

$$\Phi(x,y) = \sum_{i=1}^d \max\{0, x_i\}(Ay)_i$$

has full domain, for all $x \in \mathbb{R}^d$ we have that $\Phi(x,\,\cdot\,)$ is linear and for all $y \in \operatorname{dom} g = C$ the function $\Phi(\,\cdot\,, y)$ is convex and continuous.

Furthermore, we obtain for all (x,y), $(x',y') \in \mathbb{R}^d \times \operatorname{dom} g$

$$\left\|\nabla_y\Phi(x,y) - \nabla_y\Phi(x',y')\right\| = \left\|A^T\left([x]_+ - [x']_+\right)\right\| \leq \|A\|\left\|x - x'\right\|,$$

hence (5.2) holds with $L_{yx} = \|A\|$ and $L_{yy} = 0$.

The algorithm (5.7) iterates for $k \geq 0$

$$\left\lfloor \begin{array}{l} v_k = y_k + \sigma_k \left[(1+\theta_k)\nabla_y \Phi(x_k, y_k) - \theta_k \nabla_y \Phi(x_{k-1}, y_{k-1})\right] \\ \quad = y_k + \sigma_k A^T((1+\theta_k)[x_k]_+ - \theta_k [x_{k-1}]_+), \\ y_{k+1} = \operatorname{prox}_{\sigma_k g}(v_k) = P_C\left(\frac{1}{1+\nu\sigma_k} v_k\right), \\ x_{k+1} = \operatorname{prox}_{\tau_k \Phi(\,\cdot\,, y_{k+1})}(x_k), \end{array} \right.$$

where the calculation of the orthogonal projection on the set C is a simple quadratic program and

$$\operatorname{prox}_{\tau\Phi(\,\cdot\,,y)}(x) = \left(\operatorname{prox}_{\tau(Ay)_i \max\{0,\,\cdot\,\}}(x_i)\right)_{i=1}^{d},$$

where, for $i = 1, ..., d$,

$$\operatorname{prox}_{\tau(Ay)_i \max\{0,\,\cdot\,\}}(x_i) = \begin{cases} x_i & \text{if } x_i \leq 0, \\ 0 & \text{if } 0 < x_i \leq \tau(Ay)_i, \\ x_i - \tau(Ay)_i & \text{if } x_i > \tau(Ay)_i. \end{cases}$$

By writing the first order optimality conditions and using Lagrange duality we obtain the following characterisation.

$$\begin{aligned} (x^*, y^*) \text{ is a saddle point of (5.49)} &\Leftrightarrow \begin{cases} 0 \in \partial\left(\langle [\,\cdot\,]_+, Ay^*\rangle - \delta_C(y^*) - \frac{\nu}{2}\|y^*\|^2\right)(x^*) \\ 0 \in \partial\left(-\left\langle A^T[x^*]_+, \,\cdot\,\right\rangle + \delta_C(\,\cdot\,) + \frac{\nu}{2}\|\,\cdot\,\|^2\right)(y^*) \end{cases} \\ &\Leftrightarrow \begin{cases} 0 \in \sum_{i=1}^d (Ay^*)_i \,\partial \max\{0, \,\cdot\,\}(x_i^*) \\ A^T[x^*]_+ - \nu y^* \in N_C(y^*) \end{cases} \\ &\Leftrightarrow \begin{cases} \forall i = 1, \ldots, d : ((Ay^*)_i > 0 \text{ and } x_i^* \leq 0) \text{ or} \\ \quad ((Ay^*)_i = 0 \text{ and } x_i^* \in \mathbb{R}) \\ \left\langle A^T[x^*]_+ - \nu y^*, y^*\right\rangle = 0 \\ \nu y^* - A^T[x^*]_+ \in A^T\left(\mathbb{R}^d_+\right) \end{cases} \\ &\Leftrightarrow \begin{cases} \forall i = 1, \ldots, d : ((Ay^*)_i > 0 \text{ and } x_i^* \leq 0) \text{ or} \\ \quad ((Ay^*)_i = 0 \text{ and } x_i^* \in \mathbb{R}) \\ \nu\|y^*\|^2 = \left\langle A^T[x^*]_+, y^*\right\rangle = \langle [x^*]_+, Ay^*\rangle = 0 \\ \nu y^* \in A^T\left([x^*]_+ + \mathbb{R}^d_+\right). \end{cases} \end{aligned}$$

This means, that for $\nu = 0$ we obtain

$$(x^*, y^*) \text{ is a saddle point of (5.49)} \Leftrightarrow \begin{cases} \forall i = 1, \ldots, d : ((Ay^*)_i > 0 \text{ and } x_i^* \leq 0) \text{ or} \\ \quad ((Ay^*)_i = 0 \text{ and } x_i^* \in \mathbb{R}) \\ 0 \in A^T\left([x^*]_+ + \mathbb{R}^d_+\right) \end{cases},$$

whereas for $\nu > 0$

$$(x^*, y^*) \text{ is a saddle point of (5.49)} \Leftrightarrow \begin{cases} y^* = 0 \\ 0 \in A^T \left([x^*]_+ + \mathbb{R}^d_+ \right) \end{cases}.$$

If $A \in \mathbb{R}^{d \times n}$ has full row rank the inclusion

$$0 \in A^T \left([x^*]_+ + \mathbb{R}^d_+ \right)$$

is equivalent to

$$x^* \leqq 0.$$

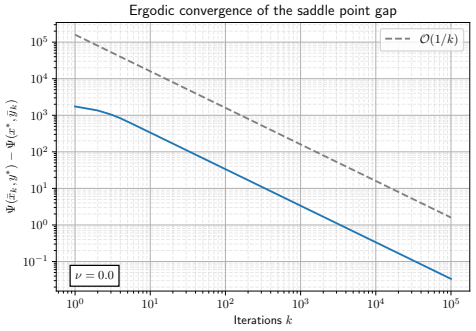


Figure 5.1.: Convergence of the minimax gap like $\mathcal{O}(1/K)$ for $\nu = 0$.

For the experiments we choose dimensions $d = 250$ and $n = 350$. For easier validation of the solution x^* we ensure that the matrix $A \in \mathbb{R}^{d \times n}$ with entries drawn from a uniform distribution on the interval $[-3, 3]$ has full row rank. The starting points $x_0 = x_{-1} \in \mathbb{R}^d$ and $y_0 \in \mathbb{R}^n$ have entries drawn from a uniform distribution on the interval $[-5, 5]$.

In the case $\nu = 0$, i.e., the regulariser g being merely convex, we proved weak asymptotic convergence of the iterates to some saddle point (x^*, y^*) and convergence of the minimax gap at the ergodic sequences to zero like $\mathcal{O}(1/K)$ for any saddle point. The latter is illustrated in Figure 5.1 for $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^n$ with $x^* \leqq 0$ and $y^* \in C$ with $y^* \neq 0$ for a single random initialisation.

Let $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^n$ be a saddle point. In the case $\nu > 0$, i.e., the regulariser g being ν-strongly convex, we proved strong non-asymptotic convergence of the sequence $(y_k)_{k \geq 0} \to y^*$ like $\mathcal{O}(1/K)$ and convergence of the minimax gap at the ergodic sequences to zero like $\mathcal{O}(1/K^2)$. The numerical behaviour of our method validating the theoretical claims for $\nu > 0$ is highlighted in Figure 5.2. The plots shown are for a single random initialisation and with the choice $\nu = \frac{3}{10}$.

5.4.2. Multi kernel support vector machine

The second application to test our method in practice is to learn a combined kernel matrix for a multi kernel *support vector machine* (SVM). We have a set of labelled training data

$$S_n = \{(a_1, b_1), \ldots, (a_n, b_n)\} \subseteq \mathbb{R}^m \times \{-1, 1\},$$

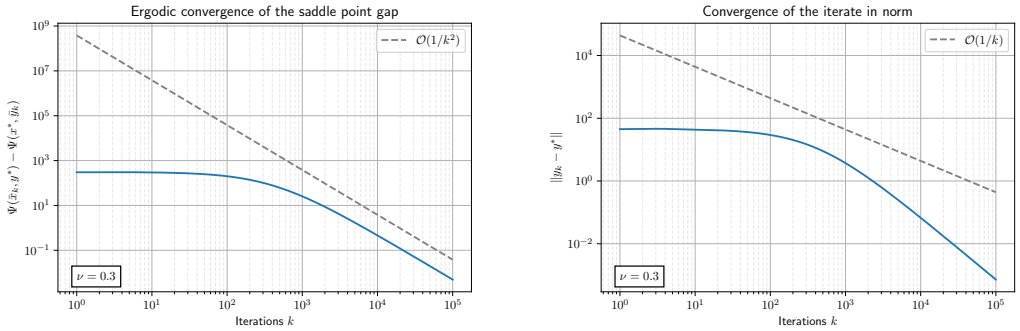


Figure 5.2.: Convergence of the minimax gap like $\mathcal{O}(\frac{1}{K^2})$ and of the sequence $(y_k)_{k\geq 0}$ in norm like $\mathcal{O}(\frac{1}{K})$ for $\nu > 0$.

where we call $b = (b_i)_{i=1}^n$, and a set of unlabelled test data

$$T_l = \{a_{n+1}, \ldots, a_{n+l}\} \subseteq \mathbb{R}^m.$$

We consider embeddings of the data according to a kernel function $\kappa : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ with the corresponding symmetric and positive semidefinite kernel matrix

$$\mathcal{K} = \begin{pmatrix} \mathcal{K}^{tr} & \mathcal{K}^{tr,t} \\ \mathcal{K}^{t,tr} & \mathcal{K}^{t} \end{pmatrix},$$

where $\mathcal{K}_{ij} = \kappa(a_i, a_j)$ for $i, j = 1, \ldots, n, n+1, \ldots, n+l$.

In the following e is a vector of appropriate size consisting of ones. According to [87] the problem of interest is

$$\min_{\substack{\mathcal{K}\in\mathbb{K} \\ \operatorname{trace}(\mathcal{K})=c}} \max_{\substack{0\leqq\alpha\leqq C \\ \langle\alpha,b\rangle=0}} \alpha^T e - \frac{1}{2}\alpha^T G(\mathcal{K}^{tr})\alpha - \frac{\nu}{2}\|\alpha\|_2^2, \tag{5.50}$$

where $\mathbb{K}$ is the model class of kernel matrices, $c \in (0, +\infty)$, $C \in (0, +\infty]$ and $\nu \in [0, +\infty)$ are model parameters and we define $G(\mathcal{K}^{tr}) := \operatorname{diag}(b)\mathcal{K}^{tr}\operatorname{diag}(b)$.

The set $\mathbb{K}$ is restricted to be the set of positive semidefinite matrices that can be written as a non negative linear combination of kernel matrices $\mathcal{K}_1, \ldots, \mathcal{K}_d$, i.e.,

$$\mathbb{K} = \left\{ \mathcal{K} \in S_+^m \;\middle|\; \mathcal{K} = \sum_{i=1}^d \eta_i \mathcal{K}_i,\ \eta_i \geq 0 \text{ for } i = 1, ..., d \right\}.$$

With this choice (5.50) becomes

$$\min_{\substack{\langle\eta,r\rangle=c \\ \eta\geqq 0}} \max_{\substack{0\leqq\alpha\leqq C \\ \langle\alpha,b\rangle=0}} \alpha^T e - \frac{1}{2}\sum_{i=1}^d \eta_i \alpha^T G(\mathcal{K}_i^{tr})\alpha - \frac{\nu}{2}\|\alpha\|^2, \tag{5.51}$$

where $\eta = (\eta_i)_{i=1}^d$ and $r = (r_i)_{i=1}^d$ with $r_i = \text{trace}(\mathcal{K}_i)$ for $i = 1, ..., d$. Assume $(\eta^*, \alpha^*) \in \mathbb{R}^d \times \mathbb{R}^n$ to be a saddle point of (5.51) and write

$$\mathcal{K}^* = \sum_{j=1}^{d} \eta_j^* \mathcal{K}_j.$$

Following the considerations of [75] we compute for $a_k \in T_l$ with $k \in \{n+1, \ldots, n+l\}$,

$$\mathcal{L}(a_k) = \text{sgn}\left(\sum_{i=1}^{n} b_i \alpha_i^* \mathcal{K}_{ik}^* + \gamma\right) = \text{sgn}\left(\sum_{i=1}^{n}\sum_{j=1}^{d} b_i \alpha_i^* \eta_j^* \left(\mathcal{K}_j\right)_{ik} + \gamma\right), \tag{5.52}$$

with

$$\gamma = b_{j_0}(1 - \nu\alpha_{j_0}^*) - \sum_{i=1}^{n} b_i \alpha_i^* \mathcal{K}_{ij_0}^* = b_{j_0}(1 - \nu\alpha_{j_0}^*) - \sum_{i=1}^{n}\sum_{j=1}^{d} b_i \alpha_i^* \eta_j^* \left(\mathcal{K}_j\right)_{ij_0},$$

for some $j_0 \in \{1, \ldots, n\}$ such that $0 < \alpha_{j_0}^* < C$.

After writing $x_i = \frac{r_i \eta_i}{c}$ for $i = 1, ..., d$ and augmenting the objective with an additional (strongly) convex penalisation term, we obtain

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^n} \delta_\Delta(x) + \frac{\mu}{2}\|x\|^2 - \frac{1}{2}\sum_{i=1}^{d} x_i y^T M_i y + y^T e - \left(\delta_Y(y) + \frac{\nu}{2}\|y\|^2\right), \tag{5.53}$$

where $\mu \geq 0$ and $M_i := \frac{c}{r_i} G(\mathcal{K}_i^{tr})$ for $i = 1, ..., d$,

$$\Delta := \{x \in \mathbb{R}^d \mid x \geqq 0,\ \langle x, e\rangle = 1\}$$

is the d-dimensional unit simplex and

$$Y := \{y \in \mathbb{R}^n \mid 0 \leqq y \leqq C,\ \langle y, b\rangle = 0\}$$

is the intersection of a box and a hyperplane.

In the notation of (5.1) we have $\Phi : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\Phi(x, y) = \delta_\Delta(x) + \frac{\mu}{2}\|x\|^2 - \frac{1}{2}\sum_{i=1}^{d} x_i y^T M_i y + y^T e,$$

and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ given by

$$g(y) = \delta_Y(y) + \frac{\nu}{2}\|y\|^2.$$

We see that Φ and g satisfy the assumptions considered for problem (5.1).

The algorithm (5.7) iterates as follows for $k \geq 0$

$$\left\lfloor \begin{array}{l} v_k = y_k + \sigma_k \left[(1+\theta_k)\nabla_y \Phi(x_k, y_k) - \theta_k \nabla_y \Phi(x_{k-1}, y_{k-1})\right], \\ y_{k+1} = \operatorname{prox}_{\sigma_k g}(v_k) = P_Y\left(\frac{1}{1+\nu\sigma_k} v_k\right), \\ x_{k+1} = \operatorname{prox}_{\tau_k \Phi(\,\cdot\,, y_{k+1})}(x_k) = P_\Delta\left(\frac{1}{1+\mu\tau_k}(x_k + \tau_k \xi^{y_{k+1}})\right), \end{array} \right.$$

where

$$\nabla_y \Phi(x,y) = -\left(\frac{1}{2}\sum_{i=1}^d x_i (M_i + M_i^T)\right) y + e = -\left(\sum_{i=1}^d x_i M_i\right) y + e \quad \text{for } (x,y) \in \Delta \times \mathbb{R}^n$$

and

$$\xi^y := \left(\frac{1}{2} y^T M_i y\right)_{i=1}^d.$$

To determine the correct step sizes and momentum parameter, we need to find Lipschitz constants for $\nabla_y \Phi$, i.e., L_{yx}, $L_{yy} \geq 0$ such that (5.2) holds. Recall, that we require for all (x,y), $(x',y') \in \Pr_{\mathcal{X}}(\operatorname{dom}\Phi) \times \operatorname{dom} g$

$$\left\|\nabla_y \Phi(x,y) - \nabla_y \Phi(x',y')\right\| \leq L_{yx}\left\|x - x'\right\| + L_{yy}\left\|y - y'\right\|,$$

with $\Pr_{\mathcal{X}}(\operatorname{dom}\Phi) = \Delta$ and $\operatorname{dom} g = Y$.

Let (x,y), $(x',y') \in \Delta \times Y$. Then

$$\begin{aligned} \left\|\nabla_y \Phi(x,y) - \nabla_y \Phi(x',y')\right\| &= \left\| -\sum_{i=1}^d x_i M_i y + e + \sum_{i=1}^d x_i' M_i y' - e \right\| \\ &= \left\| \sum_{i=1}^d x_i M_i y' - \sum_{i=1}^d x_i M_i y + \sum_{i=1}^d x_i' M_i y' - \sum_{i=1}^d x_i M_i y' \right\| \\ &\leq \left\| \sum_{i=1}^d x_i M_i (y - y') \right\| + \left\| \sum_{i=1}^d (x_i - x_i') M_i y' \right\| \\ &\leq \sum_{i=1}^d |x_i| \left\|M_i\right\| \left\|y - y'\right\| + \sum_{i=1}^d \left|x_i - x_i'\right| \left\|M_i\right\| \left\|y'\right\| \\ &\leq \left(\|x\|_1 \max_{1\leq i\leq d} \|M_i\|\right) \left\|y - y'\right\| + \left(\left\|y'\right\| \max_{1\leq i\leq d} \|M_i\|\right) \left\|x - x'\right\|_1 \\ &\leq \left(\|x\|_1 \max_{1\leq i\leq d} \|M_i\|\right) \left\|y - y'\right\| + \left(\left\|y'\right\| \sqrt{d} \max_{1\leq i\leq d} \|M_i\|\right) \left\|x - x'\right\|. \end{aligned}$$

As $x \in \Delta$, we have $\|x\|_1 = 1$ and since $y' \in Y$ we get $\|y'\| \leq C\sqrt{n}$. Thus we obtain

$$\left\|\nabla_y \Phi(x,y) - \nabla_y \Phi(x',y')\right\| \leq L_{yx}\left\|x - x'\right\| + L_{yy}\left\|y - y'\right\|,$$

with

$$L_{yx} = C\sqrt{dn} \max_{1\leq i\leq d} \|M_i\|\,, \quad L_{yy} = \max_{1\leq i\leq d} \|M_i\|\,.$$

For our experiments we use four different data sets from the "UCI Machine Learning Repository" [58]: the (original) Wisconsin *breast cancer* dataset [101] (699 total observations including 16 incomplete examples; 9 features), the Statlog *heart disease* data set (270 observations; 13 features), the *Ionosphere* data set (351 observations; 33 features) and the Connectionist Bench *Sonar* data set (208 observations; 60 features). All the data sets are normalised such that each feature column has zero mean and standard deviation equal to one.

Furthermore, we take $d = 3$ given kernel functions, namely a polynomial kernel function $k_1(a, a') = (1 + a^T a')^2$ of degree 2 for $\mathcal{K}_1$, a Gaussian kernel function $k_2(a, a') = \exp(-\frac{1}{2}(a - a')^T(a - a')/\frac{1}{10})$ for $\mathcal{K}_2$ and a linear kernel function $k_3(a, a') = a^T a'$ for $\mathcal{K}_3$. The resulting kernel matrices are normalised according to [87, Section 4.8], giving

$$r_i = \operatorname{trace}(\mathcal{K}_i) = n + l.$$

The model parameter $c > 0$ is chosen to be

$$c = \sum_{i=1}^{d} r_i = d(n + l),$$

and we set $C = 1$.

On this application we test the three proposed versions of OGAProx. We refer to the version of OGAProx with constant parameters from Section 5.2.3 as OGAProx-C1, to the one with adaptive parameters from Section 5.2.3 as OGAProx-A and to the one from Section 5.3.3 giving linear convergence with constant parameters as OGAProx-C2. The results are compared with those obtained by APD1 and APD2 from [75]. In their experiments on multi kernel SVMs they showed superiority of their method compared to *Mirror Prox* by [112] in terms of accuracy, runtime and relative error. They also argued that with APD they are able to obtain decent approximations of solutions of (5.51) by interior point methods such as MOSEK [111] taking about the same amount of runtime.

The main difference between APD and our method OGAProx is that for the first a gradient step in the first component is employed whereas for the latter a purely proximal step is used. To be able to employ APD2 with adaptive parameters for $\nu > 0$, the roles of x and y in (5.53) have to be switched, giving a different method than OGAProx-A. The runtime of both methods however is still very similar as both use the same number of gradient computations/storages and projections per iteration.

All algorithms are initialised with

$$x_0 = x_{-1} = \frac{1}{d}e, \quad y_0 = y_{-1} = 0.$$

Each data set is randomly partitioned into 80 % training and 20 % test set. The test set is used to judge the quality of the obtained model by predicting the labels via (5.52) and

computing the resulting test set accuracy (TSA). Note that the TSA is not guaranteed to converge or increase at all by our theoretical considerations, which only state convergence of the iterates and in terms of function values. The reported TSA values are the average over 10 random partitions. Due to occasionally occurring rather dramatic deflections of the TSA we actually compute 12 runs, but remove minimum and maximum values before calculating the mean.

1-norm soft margin classifier

For $\mu = \nu = 0$ the formulation (5.51) realises the so-called 1-norm soft margin classifier. In this case g is merely convex and we can only use the constant parameter choice from Section 5.2.3 with the name OGAProx-C1. We compare the results with those obtained by APD1 from [75].

		TSA at iteration k				
Method	Data set	$k = 250$	$k = 500$	$k = 1000$	$k = 1500$	$k = 2000$
OGAProx-C1	Breast cancer	97.15	97.37	97.08	93.94	**97.45**
	Heart disease	74.63	74.07	80.00	81.30	**82.78**
	Ionosphere	70.85	85.35	90.28	87.46	**93.24**
	Sonar	70.00	75.24	83.81	84.52	**85.95**
APD1	Breast cancer	97.23	97.37	**97.45**	94.01	**97.45**
	Heart disease	74.63	72.59	81.85	80.74	82.41
	Ionosphere	70.85	85.35	85.49	88.73	92.68
	Sonar	70.00	74.76	81.67	84.76	84.52

Table 5.1.: TSA of 1-norm soft margin classifier ($\mu = 0$, $\nu = 0$, $C = 1$) trained with OGAProx-C1 and APD1, averaged over 10 random partitions.

In the case of 1-norm soft margin classifier the results reported in Table 5.1 paint a clear picture. OGAProx outperforms APD on three out of four data sets and ties on one data set, achieving maximum TSA values of 97.45 %, 82.78 %, 93.24 % and 85.95 % on Breast cancer, Heart disease, Ionosphere and Sonar, respectively.

2-norm soft margin classifier

For $\mu = 0$ and $\nu > 0$ from (5.51) we obtain the so-called 2-norm soft margin classifier with $C = 1$. In this case g is ν-strongly convex and we can use both parameter choices from Section 5.2.3 and the one from Section 5.2.3 giving OGAProx-C1 and OGAProx-A, respectively. This time we compare the results with those obtained by APD1 as well as APD2 from [75].

We see in Table 5.2 that the situation for the 2-norm soft margin classifier is more diverse than previously with the 1-norm soft margin classifier. Comparing the two constant methods – OGAProx-C1 and APD1 – with each other, as well as the two adaptive

Method	Data set	TSA at iteration k				
		$k = 250$	$k = 500$	$k = 1000$	$k = 1500$	$k = 2000$
OGAProx-C1	Breast cancer	97.15	97.37	97.15	**97.45**	97.15
	Heart disease	75.19	75.00	77.78	83.52	83.52
	Ionosphere	70.99	85.35	89.86	87.89	91.27
	Sonar	70.71	77.86	81.90	85.71	**86.19**
APD1	Breast cancer	97.23	97.37	97.30	97.37	97.37
	Heart disease	75.37	67.78	80.74	82.22	**84.81**
	Ionosphere	71.27	85.35	88.87	89.72	**92.39**
	Sonar	70.48	76.43	83.33	84.76	85.71
OGAProx-A	Breast cancer	97.15	97.37	97.37	97.45	97.45
	Heart disease	76.11	73.70	83.70	81.30	**84.26**
	Ionosphere	70.85	85.21	86.34	90.42	**93.52**
	Sonar	70.48	76.90	83.33	82.62	84.76
APD2	Breast cancer	97.23	97.37	**97.59**	97.01	96.72
	Heart disease	76.11	71.30	81.48	78.70	83.15
	Ionosphere	71.13	85.35	84.79	84.93	90.42
	Sonar	70.24	75.95	84.05	84.52	**86.19**

Table 5.2.: TSA of 2-norm soft margin classifier ($\mu = 0$, $\nu = \frac{1}{2}$, $C = 1$) trained with OGAProx-C1, OGAProx-A, APD1 and APD2, averaged over 10 random partitions.

methods – OGAProx-A and APD2 – we see that in both cases two out of four times OGAProx is better than APD and vice versa. Notice that the two data sets with in general lower TSA, namely Heart disease and Sonar, seem to benefit from the regularising effect of $\nu > 0$, while those with already very good results on the other hand do not, compared to the results of the 1-norm soft margin classifier with $\nu = 0$. In addition note that the adaptive variant OGAProx-A improves on the result of OGAProx-C1 on three out of four data sets.

Regularised 2-norm soft margin classifier

For $\mu > 0$ and $\nu > 0$ from (5.51) we again obtain the so-called 2-norm soft margin classifier with $C = 1$, this time, however, in a regularised version. Now not only g is strongly convex, but also $\Phi(\cdot, y)$ and we can use all our parameter choices from Section 5.2.3, Section 5.2.3 and Section 5.3.3 yielding OGAProx-C1, OGAProx-A and OGAProx-C2, respectively. Once more we compare the results with those obtained by APD1 as well as APD2 from [75], pointing out that that OGAProx-C2 has no APD counterpart harnessing the additional strong convexity of the problem.

We see in Table 5.3 that for the regularised 2-norm soft margin classifier the situation

Method	Data set	TSA at iteration k				
		$k = 250$	$k = 500$	$k = 1000$	$k = 1500$	$k = 2000$
OGAProx-C1	Breast cancer	97.15	97.37	97.15	**97.52**	97.45
	Heart disease	75.19	73.52	77.22	83.15	83.70
	Ionosphere	70.99	85.35	87.89	91.41	**91.97**
	Sonar	70.48	78.81	83.33	84.76	**85.95**
APD1	Breast cancer	97.23	97.37	97.37	97.01	97.30
	Heart disease	75.19	68.89	75.56	79.81	**84.07**
	Ionosphere	71.27	85.35	86.06	89.15	91.69
	Sonar	70.71	76.43	83.10	85.48	85.48
OGAProx-A	Breast cancer	97.15	97.37	97.45	97.37	97.30
	Heart disease	76.11	70.93	82.78	80.74	83.52
	Ionosphere	70.85	85.21	85.92	89.86	**93.38**
	Sonar	70.24	76.43	82.86	86.19	86.19
APD2	Breast cancer	97.23	97.37	97.45	94.53	**97.52**
	Heart disease	76.11	71.67	80.00	79.26	**83.52**
	Ionosphere	71.13	85.35	86.90	92.39	91.13
	Sonar	70.24	75.00	82.62	84.52	**86.43**
OGAProx-C2	Breast cancer	97.15	97.45	**97.59**	97.15	96.57
	Heart disease	74.07	78.52	76.11	82.22	**83.70**
	Ionosphere	70.42	84.37	86.48	90.85	**92.25**
	Sonar	69.05	74.29	85.24	85.71	**86.19**

Table 5.3.: TSA of regularised 2-norm soft margin classifier ($\mu = 1$, $\nu = \frac{1}{2}$, $C = 1$) trained with OGAProx-C1, OGAProx-A, OGAProx-C2, APD1 and APD2, averaged over 10 random partitions.

is similar to the version without additional regulariser. This time for the constant methods, OGAProx-C1 and APD1, OGAProx is better than APD on three data sets while APD is better than OGAProx on only one. On the contrary, for the adaptive methods, OGAProx-A and APD2, it is the other way round. APD performs better than APD on three data sets while OGAProx is better than APD on only one. For the second version of OGAProx with constant parameter choice exhibiting linear convergence in both iterates and function values, there is no APD counterpart. When we compare the results for OGAProx-C2 to those of OGAProx-C1, then we see that the TSA values become better in general with improvements on three out of four data sets and one draw. On the Breast cancer data set OGAProx-C2 even delivers the maximum TSA over all considered methods.

Classification of ancient Armenian manuscripts

Another interesting application where we used SVMs for classification purposes was located in the field of *Digital Humanities*, a research field at the intersection of digital resources and computational methods, and the disciplines of the humanities. In our particular case we worked on a digital palaeography project for ancient Armenian manuscripts. *Palaeography* is the study of historic writing systems and historical manuscripts, including the analysis of historic handwriting or dating of manuscripts, and is concerned with the forms and processes of writing – not the textual content of documents.

The starting point of the project was an index of digitised manuscripts available online [138]. Some libraries provided direct access to the images to be downloaded and worked on, while others only granted online access to look at them on their homepage. Because of the nature of our experiments we had to restrict ourselves to the former. A huge difference between libraries was observable regarding properties concerning the quality of the images like their size or resolution, or whether they were available in greyscale or colour.

As rigorous web-scraping was out of scope, we worked with manuscripts from one single library (amounting to 49 manuscripts with a total of 24,307 pages) where both accessibility and quality were satisfactory. Palaeographers are interested in various attributes of a manuscript, especially in the precise date/year(s) when the manuscript was finished/written, the script used in the folio or the hand/scribe responsible for the production/text. To successfully apply machine learning to a task it is of utter importance to have an extensive and complete data set available for training and evaluation. In particular, when doing classification we are in need of a target variable, the *label*. The meta data of the 49 manuscripts considered, however, were not complete at all. Out of the 49 manuscripts we had 35 (with a total of 9,851 pages) with a given year but very unevenly distributed over several centuries, leading to an unbalanced data set not suitable for classification of the scripts. The hand or scribe of the manuscripts was only known in 19 cases with only single occurrences of each of the persons, rendering classification with respect to this target variable impossible. Regarding scripts we had 30 manuscripts (with a total of 18,401 pages) with data about script type. Out of the four Armenian scripts – *Erkata'gir*, *Sla'gir*, *Bolor'gir* and *Notr'gir* – only one, namely Erkata'gir, was missing in this subset of the data. This was the reason we decided to do classification regarding the script type of the manuscripts. Out of the 30 folios 23 consisted entirely of one script type each, which where used for the final analysis. To summarise, for Bolor'gir, Notr'gir and Sla'gir we had 7, 15, and 1 manuscript(s), respectively. A brief overview about the historical background of Armenian calligraphy can be found at [97], including sample images of the four Armenian script types.

To carry out the task of classification using images of the manuscripts, not looking into the textual content of the pages, we chose a similar approach to [83], where they worked on digital palaeography on medieval Latin manuscripts.

To this end we used a "bag of (visual) words" model, which is well-established in com-

puter vision and originated in automated text classification. The strategy for this is to extract local feature descriptors ("visual words") from the image, followed by encoding these local descriptors into a global descriptor which is then used for classification. For the first step, the identification of the local features, we used the *scale-invariant feature transform* (SIFT) [95]. This computer-vision algorithm is robust to image transformations like changes in scale, rotation, brightness or contrast. Because in alphabets letters might be derived by rotation of other letters, we disabled the rotation-invariance of the obtained SIFT key points.

The next step was to encode the local descriptors to form a global feature descriptor for each page. As a first simple approach we used the average of all local descriptors. In [83] more elaborate techniques were used involving clustering of the local descriptors, followed by employing so-called *i-vectors*, a tool from spoken language classification. As our results with simple averaging were already reasonably good, we did not try more sophisticated formulations.

The final step consisted of classifying the global descriptor into the correct script type class for which one can easily use (multikernel) SVMs. In [83] linear SVMs for binary one-vs-rest classification were used, i.e., for each script type a separate SVM was trained with all the instances of this script as positive example and all other as negative ones. In the test phase all SVMs were queried to rank the different predictions to determine the final class.

	Training Set				Test Set			
Script	Precision	Recall	F1	Support	Precision	Recall	F1	Support
Bolor'gir	0.83	1.00	0.91	10	0.73	0.91	0.81	170
Notr'gir	1.00	1.00	1.00	10	0.93	0.89	0.91	170
Sla'gir	1.00	0.80	0.89	10	0.94	0.75	0.83	170

Table 5.4.: Performance of DH1 in training and test phase with respect to precision, recall and f1-score.

	Training Set				Test Set			
Script	Precision	Recall	F1	Support	Precision	Recall	F1	Support
Bolor'gir	1.00	0.89	0.94	45	1.00	0.94	0.97	136
Notr'gir	0.94	0.98	0.96	45	0.96	0.99	0.97	136
Sla'gir	0.90	0.96	0.92	45	0.96	0.98	0.97	136

Table 5.5.: Performance of DH2 in training and test phase with respect to precision, recall and f1-score.

We considered three different scenarios, denoted by DH1, DH2 and DH3. For the training phase of DH1 we randomly chose 5 pages from two manuscripts written in Bolor'gir and Notr'gir each. Since only one manuscript written in Sla'gir (with 181

pages) was available for our experiments, hence we took 10 pages of this folio. For the test phase we took 170 out of the remaining 171 pages of the Sla'gir manuscript, while we took 34 pages of 5 manuscripts each for the other two script types. For the training of DH2 and DH3 we selected 15 pages of 3 manuscripts and 25 pages of 5 manuscripts each, respectively, and chose the test sets accordingly.

	Training Set				Test Set			
Script	Precision	Recall	F1	Support	Precision	Recall	F1	Support
Bolor'gir	0.99	0.98	0.99	125	1.00	0.93	0.96	56
Notr'gir	0.98	0.99	0.99	125	0.98	0.96	0.97	56
Sla'gir	0.98	0.98	0.98	125	0.92	1.00	0.96	56

Table 5.6.: Performance of DH3 in training and test phase with respect to precision, recall and f1-score.

In Tables 5.4-5.6 we see the performance of our model in the three different scenarios, both on the training and the test set. We observe that DH2 seems to be most robust, requiring relatively few training samples but still performing best on a relatively big test set which is the case for all three script types. This behaviour is confirmed in terms of the accuracy which can be found in Table 5.7.

	Training Set		Test Set	
	Accuracy	Support	Accuracy	Support
DH1	0.93	30	0.85	510
DH2	0.94	135	0.97	408
DH2	0.99	375	0.96	168

Table 5.7.: Accuracy of our model for the considered scenarios.

5.4.3. Classification incorporating minimax group fairness

We want to classify labelled data $(a_j, b_j)_{j=1}^n \subseteq \mathbb{R}^d \times \{\pm 1\}$, additionally taking into account so-called *minimax group fairness* [57, 102]. The data is divided into m groups $G_1, ..., G_m$, such that for $i \in [m] := \{1, ..., m\}$ we have $G_i = (a_{i_j}, b_{i_j})_{j=1}^{n_i} \subseteq (a_j, b_j)_{j=1}^n$ with $n_i := |G_i|$ and $i_j \in [n]$ for all $i \in [m]$ and all $j \in [n_i]$. *Fairness* is measured by worst-case outcomes across the considered groups. Hence we consider the following problem,

$$\min_{x \in \mathbb{R}^d} \max_{i \in [m]} f_i(x), \tag{5.54}$$

with

$$f_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} L(h_x(a_{i_j}), b_{i_j}),$$

where h_x is a function parametrised by x, mapping features to predicted labels, and L is a loss function measuring the error between the predicted and true labels.

It is easy to see that (5.54) is equivalent to

$$\min_{x\in\mathbb{R}^d}\max_{y\in\Delta_m}\sum_{i=1}^{m} y_i f_i(x),$$

where $\Delta_m := \{(v_1,...,v_m) \in \mathbb{R}^m \mid \sum_{i=1}^m v_i = 1,\ v_i \geq 0 \text{ for } i = 1,...,m\}$ denotes the probability simplex in $\mathbb{R}^m$. We will work with a linear (affine) predictor $h_x : \mathbb{R}^d \to \mathbb{R}$ given by

$$h_x(a) = a^T x,$$

with $x \in \mathbb{R}^d$ and $L : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ being the hinge loss, i.e.,

$$L(r,s) = \max\{0, 1 - sr\},$$

for r, $s \in \mathbb{R}$.

Combining all of the above we get

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^m} \Phi(x,y) - g(y), \tag{5.55}$$

with $\Phi : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ defined by

$$\Phi(x,y) = \sum_{i=1}^{m} y_i \frac{1}{n_i}\sum_{j=1}^{n_i} \max\{0, 1 - b_{i_j} a_{i_j}^T x\},$$

and $g : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ given by

$$g(y) = \delta_{\Delta_m}(y).$$

The function g is proper, lower semicontinuous and convex (with modulus $\nu = 0$). Furthermore, we observe that $\Phi(\cdot, y) : \mathbb{R}^d \to \mathbb{R}$ is proper, convex and lower semicontinuous for all $y \in \operatorname{dom} g = \Delta_m$ and for all $x \in \operatorname{Pr}_{\mathbb{R}^d}(\operatorname{dom}\Phi) = \mathbb{R}^d$ we have $\operatorname{dom}\Phi(x,\cdot) = \mathbb{R}^m$ and $\Phi(x,\cdot) : \mathbb{R}^m \to \mathbb{R}$ is concave and Fréchet differentiable. However, note that Φ is not differentiable in its first component.

Moreover the Lipschitz condition on the gradient is fulfilled as well. Indeed, for (x,y), $(x',y') \in \mathbb{R}^d \times \Delta_m$ we have

$$\left\|\nabla_y\Phi(x,y) - \nabla_y\Phi(x',y')\right\| \leq L_{yx}\left\|x - x'\right\| + L_{yy}\left\|y - y'\right\|,$$

with

$$L_{yx} = \sqrt{\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\left\|a_{i_j}\right\|^2} \quad \text{and} \quad L_{yy} = 0.$$

	Group S1		Group S2		Overall	
k	with fairness	without fairness	with fairness	without fairness	with fairness	without fairness
100	**95.78**	**95.78**	80.68	80.84	85.56	85.56
500	**95.78**	**95.78**	**81.15**	80.28	**85.93**	85.19
1000	**95.78**	**95.78**	**81.15**	80.28	**85.93**	85.19

Table 5.8.: TSA of the affine classifier after k iterations of OGAProx for the groups according to "sex", averaged over 5 random partitions.

Additionally, with $\tau > 0$ and $y \in \operatorname{dom} g$, we have for $x \in \mathbb{R}^d$

$$\operatorname{prox}_{\tau\Phi(\cdot,y)}(x) = \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \left\{ \tau \sum_{i=1}^{m} y_i \frac{1}{n_i} \sum_{j=1}^{n_i} \max\{0, 1 - b_{i_j} a_{i_j}^T u\} + \frac{1}{2} \|u - x\|^2 \right\}.$$

By introducing slack variables for the pointwise maximum, we see that the above minimisation problem is equivalent to the following quadratic program

$$\begin{aligned} \min_{\substack{u\in\mathbb{R}^d,\\ r_{ij}\in\mathbb{R},\\ i\in[m],\, j\in[n_i]}} \quad & \left\{ \tau \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_i \frac{1}{n_i} r_{ij} + \frac{1}{2} \|u - x\|^2 \right\}. \\ \text{s.t.} \quad & r_{ij} \geq 0 \quad \forall\, i \in [m],\, \forall\, j \in [n_i] \\ & r_{ij} + b_{i_j} a_{i_j}^T u \geq 1 \quad \forall\, i \in [m],\, \forall\, j \in [n_i] \end{aligned}$$

For our practical applications we consider the Statlog *heart disease* data set (270 observations; 13 features) from the "UCI Machine Learning Repository" [58] and consider two different groupings; one consisting of the sex of the patients, while the other one is regarding the patients' age. For "sex" we have two groups, that is female patients (Group S1) and male patients (Group S2), whereas for "age" we consider three groups, that is patients that are younger than 50 years old (Group A1), patients that are younger than 60 but at least 50 years old (Group A2), and patients that are 60 years of age or older (Group A3). The data set is randomly partitioned into 80 % training data and 20 % test data. The results in Table 5.8 and Table 5.9 are the values of the achieved test set accuracy (TSA) averaged over 5 random partitions. For each considered group we state the intragroup TSA together with the overall TSA for the entire test set.

Every time we report the results obtained by iterates of OGAProx governed by solving the minimax problem (5.55) taking into account the considered groups ("with fairness"), as well as the results obtained by not taking into account minimax group fairness ("without fairness"), i.e., solving the problem for a single extensive group $G_1 = (a_j, b_j)_{j=1}^n$ with $n_1 = n$, yielding the minimisation of the average loss over the whole population and leading to an "ordinary" minimisation problem.

	Group A1		Group A2		Group A3		Overall	
k	with fairness	without fairness	with fairness	without fairness	with fairness	without fairness	with fairness	without fairness
100	87.76	86.48	82.97	82.97	**86.93**	**86.93**	85.93	85.56
500	**88.71**	85.53	**83.84**	82.97	**86.93**	**86.93**	**86.67**	85.19
1000	**88.71**	85.53	**83.84**	82.97	**86.93**	**86.93**	**86.67**	85.19

Table 5.9.: TSA of the affine classifier after k iterations of OGAProx for the groups according to "age", averaged over 5 random partitions.

We see in Table 5.8 and Table 5.9 that taking into account the groups regarding "sex" and "age", respectively, is beneficial for training the affine classifier. In both cases "with fairness" achieves the highest TSA for each group and at the same time the highest overall TSA as well.

6. Conclusion

Starting from the monotone inclusion problem governed by a set-valued maximal monotone operator and a single-valued monotone and Lipschitz continuous operator we investigated a second order dynamical system of FBF type and a relaxed inertial FBF algorithm, called RIFBF, obtained as discretised counterpart of it, for which we were able to prove asymptotic convergence in both continuous and discrete case. These theoretical considerations were accompanied by applications to a bilinear saddle point problem, in the context of which we also emphasised the interplay between the inertial and the relaxation parameters.

After that we shifted our focus to general monotone variational inequalities and convex-concave saddle point problems. Our method of choice remained the FBF algorithm, but this time we were interested in results about convergence rates. We derived rates of $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$ in terms of the (restricted) gap function for the ergodic iterates in the deterministic and stochastic case, respectively. The obtained convergence rates were validated in practice on a simple bilinear minimax problem. Our theoretical results regarding (RI)FBF were complemented with empirical improvements in the training of Wasserstein GANs on the CIFAR-10 dataset.

This was followed by considerations regarding the classical (constrained) variational inequality that can be obtained from constrained saddle point problems. We investigated an accelerated method, which we call fOGDA-VI, with asymptotic convergence and convergence rates of $o(1/k)$ in terms of the restricted gap function as well as the natural residual for the last iterate. To validate the derived guarantees we treated a two-player matrix game with mixed strategies of the two players. Finally, we showed the beneficial effects of applying fOGDA-VI to the training of generative adversarial nets.

Concluding, we investigated a convex-concave saddle point problem, where the convex-concave coupling function is smooth in one variable and nonsmooth in the other. To solve this type of problem we proposed a novel algorithm under the name of OGAProx and considered the cases of the saddle function to be convex-concave, convex-strongly concave and strongly convex-strongly concave. Regarding iterates we established asymptotic convergence, a convergence rate of order $\mathcal{O}(1/k)$ and linear convergence like $\mathcal{O}(\theta^k)$, respectively, while in terms of function values we obtained ergodic convergence rates of order $\mathcal{O}(1/k)$, $\mathcal{O}(1/k^2)$ and $\mathcal{O}(\theta^k)$, respectively. We validated our theoretical considerations on a nonsmooth-linear saddle point problem, the training of multi kernel support vector machines (including an application to a classification problem in the field of Digital Humanities) and a classification problem incorporating minimax group fairness.

A. Appendix

A.1. Additional tables

In the following we report additional tables regarding the experiments from Section 3.3.1 for the standard normal and the 1-Poisson distribution.

A.1.1. 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 A.1.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ 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 A.2.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ 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	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 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	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 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	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 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	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 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	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 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	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 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	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	7,366	5,526	4,419	3,653	3,022	2,551	2,473	2,227	2,023	-	-	-	-	-	-	-
0.28	$\geq$ 10,000	$\geq$ 10,000	$\geq$ 10,000	6,979	5,235	4,165	3,396	2,703	2,448	2,344	2,111	-	-	-	-	-	-	-	-
0.32	$\geq$ 10,000	$\geq$ 10,000	9,885	6,592	4,929	3,847	3,014	2,622	2,490	2,214	-	-	-	-	-	-	-	-	-
0.36	$\geq$ 10,000	$\geq$ 10,000	9,304	6,201	4,568	3,439	2,894	2,639	2,345	-	-	-	-	-	-	-	-	-	-
0.40	$\geq$ 10,000	$\geq$ 10,000	8,723	5,775	4,128	3,148	2,920	2,513	-	-	-	-	-	-	-	-	-	-	-
0.44	$\geq$ 10,000	$\geq$ 10,000	8,139	5,246	3,667	3,219	2,737	-	-	-	-	-	-	-	-	-	-	-	-
0.48	$\geq$ 10,000	$\geq$ 10,000	7,507	4,563	3,724	3,049	-	-	-	-	-	-	-	-	-	-	-	-	-
0.52	$\geq$ 10,000	$\geq$ 10,000	6,708	4,416	3,518	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.56	$\geq$ 10,000	$\geq$ 10,000	5,748	4,298	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.60	$\geq$ 10,000	$\geq$ 10,000	5,719	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.64	$\geq$ 10,000	9,879	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.68	$\geq$ 10,000	8,680	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.72	$\geq$ 10,000	8,203	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.76	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.80	10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.84	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.88	$\geq$ 10,000	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table A.3.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from the standard normal distribution ($\mu = 0.1$)

A.1.2. 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 A.4.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ 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 A.5.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ 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	$\geq$ 10,000	$\geq$ 10,000	8,298	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
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 A.6.: Number of iterations necessary to achieve $\|w_k - v_k\| \leq 10^{-5}$ in the constrained bilinear saddle-point problem with data drawn from a Poisson distribution ($\mu = 0.1$)

A.2. Architectures

For our experiments we used the architectures proposed in [64].

A.2.1. DCGAN

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)
$\tanh(\cdot)$
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 A.7.: DCGAN architecture used for the CIFAR-10 experiments.

A.2.2. ResNet

Generator
Input: $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
Linear $128 \to 128 \times 4 \times 4$
ResBlock $128 \to 128$
ResBlock $128 \to 128$
ResBlock $128 \to 128$
ReLU
transposed conv. (kernel: 3×3, $128 \to 3$, stride: 1, pad: 1)
$\tanh(\cdot)$
Discriminator
Input: $x \in \mathbb{R}^{3\times32\times32}$
ResBlock $3 \to 128$
ResBlock $128 \to 128$
ResBlock $128 \to 128$
ResBlock $128 \to 128$
Linear $128 \to 1$

Table A.8.: ResNet architecture used for the CIFAR-10 experiments.

A.3. Hyperparameters

(DCGAN) WGAN Hyperparameters	
Batch size	$= 64$
Number of generator updates	$= 500,000$
Adam β_1	$= 0.5$
Adam β_2	$= 0.9$
Discriminator weight clipping	$= 0.01$
Learning rate for discriminator	$= 5 \times 10^{-4}$ (Extra Adam)
	$= 2 \times 10^{-4}$ (AltAdam1, FBF Adam, Optim. Adam)
Learning rate for generator	$= 5 \times 10^{-5}$ (Extra Adam)
	$= 2 \times 10^{-5}$ (AltAdam1, FBF Adam, Optim. Adam)

Table A.9.: Hyperparameters used for the WGAN formulation (with weight clipping).

For the WGAN formulation with *weight clipping*, see Table A.9, we used the extensively tuned hyperparameters from [64] for ExtraAdam, Adam1 and OptimisticAdam. Note

that our values of the Inception Score (IS) differ from the ones reported in [64] as we use the newer implementation of the IS proposed in [16]. For FBF Adam we tuned the step size and kept all other hyperparameters equal.

(DCGAN) WGAN-L1 Hyperparameters	
Batch size	$= 64$
Number of generator updates	$= 500,000$
Adam β_1	$= 0.5$
Adam β_2	$= 0.9$
Discriminator ℓ_1-regularisation	$= 1 \times 10^{-4}$
Learning rate for discriminator	$= 1 \times 10^{-3}$ (FBF Adam, Extra Adam)
	$= 5 \times 10^{-4}$ (Optim. Adam)
	$= 2 \times 10^{-4}$ (AltAdam1)
Learning rate for generator	$= 1 \times 10^{-4}$ (FBF Adam, Extra Adam)
	$= 5 \times 10^{-5}$ (Optim. Adam)
	$= 2 \times 10^{-5}$ (AltAdam1)

Table A.10.: Hyperparameters used for the WGAN-L1 formulation (with soft thresholding).

For our newly proposed WGAN-L1 formulation using 1-Norm regularisation, see Table A.10, we limited the hyperparameter search to the step sizes, with the values in Table A.9 as initial guesses. We chose the value performing the best in terms of IS and FID for a sample seed. All other parameters were kept the same as in [64].

(ResNet) WGAN-GP Hyperparameters	
Batch size	$= 64$
Number of generator updates	$= 250,000$
Adam β_1	$= 0.5$
Adam β_2	$= 0.9$
Gradient penalty	$= 10$
Power iterations for SN	$= 1$
Learning rate for discriminator	$= 5 \times 10^{-4}$ (FBF Adam, Extra Adam, Optim. Adam)
	$= 2 \times 10^{-5}$ (AltAdam1)
Learning rate for generator	$= 5 \times 10^{-4}$ (FBF Adam, Extra Adam, Optim. Adam)
	$= 2 \times 10^{-5}$ (AltAdam1)

Table A.11.: Hyperparameters used for the WGAN-GP formulation (with spectral normalization (SN)).

For the experiments based on the WGAN-GP formulation including spectral normal-

ization (SN) we limited the hyperparameter search to the step sizes, with the values recommended in [64] as initial guesses. We used a single power iteration for the SN as suggested in [108] and reduced the number of generator updates by a factor of two to ease the computational burden.

A.4. PyTorch codes

The codes below were written in the PyTorch [121] framework for the GAN experiments in Section 3.5 and Section 4.4.2.

A.4.1. FBF

```
from torch.optim import Optimizer


class FBF(Optimizer):
    """Base class for optimizers with Forward-Backward-Forward steps.

    Arguments:
    params (iterable): an iterable of :class:`torch.Tensor` s or
        :class:`dict` s. Specifies what Tensors should be optimized.
    defaults: (dict): a dict containing default values of optimization
        options (used when a parameter group doesn't specify them).
    inertia (float, optional): value for inertial FBF method
    """

    def __init__(self, params, defaults, inertia):
        super(FBF, self).__init__(params, defaults)
        self.updates_copy = []
        self.old_params_copy = []
        if inertia < 0.0:
            raise ValueError("Invalid inertia value: {}".format(inertia))
        self.inertia = inertia

    def update(self, p, group):
        raise NotImplementedError

    def extrapolation(self):
        """Performs the extrapolation step and saves a copy of the
        current update for the actual FBF step.
        """
        # Check if a copy of a previous update was already made.
        is_empty = len(self.updates_copy) == 0
        for group in self.param_groups:
            for p in group["params"]:
                u = self.update(p, group)
                # Save the current update for the FBF step. Several
                # auxiliary extrapolation steps can be made before the
                # actual step but only the update from the first
                # extrapolation is saved.
```

```
                if u is None:
                    if is_empty:
                        self.updates_copy.append(None)
                    continue
                else:
                    if is_empty:
                        self.updates_copy.append(u.data.clone())
                    # Do a gradient step
                    p.data.add_(u)

    def step(self, closure=None):
        """Performs the actual FBF step.

        Arguments:
            closure (callable, optional): A closure that reevaluates the
                model and returns the loss.
        """
        if len(self.updates_copy) == 0:
            raise RuntimeError("Need to call extrapolation before calling
                                                    step.")

        loss = None
        if closure is not None:
            loss = closure()

        have_inertia = self.inertia > 0.0
        no_old = len(self.old_params_copy) == 0
        i = -1
        for group in self.param_groups:
            for p in group["params"]:
                i += 1
                u = self.update(p, group)
                if u is None:
                    continue
                # Do another gradient step
                p.data.add_(u)
                v = self.updates_copy[i]
                if v is None:
                    continue
                # Update the parameters using the update saved during
                # extrapolation
                p.data.add_(-v)
                # In case of inertial FBF do convex combination with
                # previous iterate and store the current one
                if have_inertia:
                    if no_old:
                        # Only in very first iteration
                        self.old_params_copy.append(p.data.clone())
                    else:
                        p.data, self.old_params_copy[i] = (
                            (1 + self.inertia) * p.data
                            - self.inertia * self.old_params_copy[i],
                            p.data,
```

```
                        )
92
        # Free the old update
94      self.updates_copy = []
        return loss
```

A.4.2. fOGDA-VI

```
from torch.optim import Optimizer
2

4   class fOGDA(Optimizer):
        def __init__(self, optimizer, alpha=100, increment_iterator_every=
                                                1000):
6           print(
                f"Using fOGDA (alpha={alpha};"
8               f"increment iterator every {increment_iterator_every} step(s)
                                                    )."
            )
10          self.optimizer = optimizer
            self.defaults = self.optimizer.defaults
12          self.param_groups = self.optimizer.param_groups
            self.state = self.optimizer.state
14          # fOGDA parameters
            self.alpha = alpha
16          self.increment_iterator_every = increment_iterator_every
            self.iteration = 0
18          self.k = 2
            self.params_copy = []
20          self.old_params_copy = []
            self.updates = []
22          self.old_updates = []
            self.old_difference_of_updates = []
24
        def step(self, closure=None):
26          loss = None
            if closure is not None:
28              loss = closure()

30          no_old_params = len(self.old_params_copy) == 0
            no_old_updates = len(self.old_updates) == 0
32          no_old_difference_of_updates = len(self.old_difference_of_updates
                                                ) == 0

34          # initialise (old) parameters
            if len(self.params_copy) > 0:
36              raise RuntimeError("Something bad happend here...")
            for group in self.param_groups:
38              for p in group["params"]:
                    self.params_copy.append(p.data.clone())
40                  if no_old_params:
                        self.old_params_copy.append(p.data.clone())
```

```
42
        # reverse engineer update from optimizer step
44      self.optimizer.step()
        i = -1
46      if len(self.updates) > 0:
            raise RuntimeError("Something bad happend here...")
48      for group in self.param_groups:
            for p in group["params"]:
50              i += 1
                self.updates.append(self.params_copy[i] - p.data)
52
        # initialise old updates and difference of updates
54      if (not no_old_updates and no_old_difference_of_updates) or (
            not no_old_difference_of_updates and no_old_updates
56      ):
            raise RuntimeError("Something bad happend here...")
58      if no_old_updates and no_old_difference_of_updates:
            for p in self.updates:
60              self.old_updates.append(p.clone())
                self.old_difference_of_updates.append(torch.zeros_like(p)
                                                      )
62
        # compute fOGDA coefficients
64      theta_p = self.alpha / (self.alpha + self.k + 1)
        theta = self.alpha / (self.alpha + self.k)
66      theta_m = self.alpha / (self.alpha + self.k - 1)

68      # compute new weights with fOGDA update
        i = -1
70      for group in self.param_groups:
            for p in group["params"]:
72              i += 1
                (
74                  self.old_params_copy[i],
                    self.old_updates[i],
76                  self.old_difference_of_updates[i],
                    p.data,
78              ) = (
                    self.params_copy[i],
80                  self.updates[i],
                    self.updates[i] - self.old_updates[i],
82                  self.params_copy[i]
                    + (1 - theta_p)
84                  * (self.params_copy[i] - self.old_params_copy[i])
                    - theta_p * self.updates[i]
86                  - (2 - theta)
                    * (2 - theta_p)
88                  * (self.updates[i] - self.old_updates[i])
                    + (2 - theta_m) * (1 - theta_p)
90                  * self.old_difference_of_updates[i],
                )
92
        self.iteration += 1
```

```
94          if self.iteration % self.increment_iterator_every == 0:
                # update iterator k
96              self.k += 1

98          # free parameters
            self.params_copy = []
100         self.updates = []

102         return loss
```

Bibliography

[1] B. Abbas and H. 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.

[2] F. 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.

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

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

[5] M. M. Alves and R. 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.

[6] A. S. Antipin. On a method for convex programs using a symmetrical modification of the lagrange function. *Ekonomika i Matematicheskie Metody*, 12(6):1164–1173, 1976.

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

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[9] H. Attouch and A. Cabot. Convergence of a relaxed inertial forward–backward algorithm for structured monotone inclusions. *Applied Mathematics & Optimization*, 80(3):547–598, 2019.

[10] H. Attouch and A. Cabot. Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. *Mathematical Programming*, 184(1):243–287, 2020.

[11] H. Attouch and A. Cabot. Convergence rate of a relaxed inertial proximal algorithm for convex minimization. *Optimization*, 69(6):1281–1312, 2020.

[12] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, pages 1–43, 2020.

[13] H. Attouch and P.-E. 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.

[14] H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $o(1/k^2)$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[15] S. Banert and R. I. Bot. A forward-backward-forward differential equation and its asymptotic properties. *Journal of Convex Analysis*, 25(2):371–388, 2018.

[16] S. Barratt and R. Sharma. A note on the inception score. *arXiv:1801.01973*, 2018.

[17] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

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

[19] A. Böhm. *Quantitative convergence estimates of deterministic and stochastic methods for optimization and minimax problems*. Phd thesis, University of Vienna, Austria, 2020.

[20] A. Böhm, M. Sedlmayer, E. R. Csetnek, and R. I. Boţ. Two steps at a time—taking GAN training in stride with Tseng's method. *SIAM Journal on Mathematics of Data Science*, 4(2):750–771, 2022.

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

[22] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer, New York, 2006.

[23] R. I. Boţ and A. Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.

[24] R. I. Boţ, E. R. Csetnek, , and P. T. Vuong. The forward–backward–forward method from continuous and discrete perspective for pseudo-monotone variational inequalities in Hilbert spaces. *European Journal of Operational Research*, 287(1):49–60, 2020.

[25] R. I. Boţ and E. R. Csetnek. On the convergence rate of a forward-backward type primal-dual splitting algorithm for convex optimization problems. *Optimization*, 64(1):5–23, 2015.

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

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

[28] R. I. Boţ and E. 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.

[29] R. I. Boţ, E. R. Csetnek, and C. Hendrich. Inertial Douglas–Rachford splitting for monotone inclusion problems. *Applied Mathematics and Computation*, 256:472–487, 2015.

[30] R. I. Boţ, E. R. Csetnek, and S. C. László. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO Journal on Computational Optimization*, 4(1):3–25, 2016.

[31] R. I. Boţ, E. R. Csetnek, and D.-K. Nguyen. Fast OGDA in continuous and discrete time. *arXiv preprint arXiv:2203.10947*, 2022.

[32] R. I. Boţ, E. R. Csetnek, and M. Sedlmayer. An accelerated minimax algorithm for convex-concave saddle point problems with nonsmooth coupling function. *Computational Optimization and Applications*, pages 1–42, 2022.

[33] R. I. Boţ and C. Hendrich. Convergence analysis for a primal-dual monotone+ skew splitting algorithm with applications to total variation minimization. *Journal of Mathematical Imaging and Vision*, 2014.

[34] R. I. Boţ, P. Mertikopoulos, M. Staudigl, and P. T. Vuong. Minibatch forward-backward-forward methods for solving stochastic variational inequalities. *Stochastic Systems*, 11(2):112–139, 2021.

[35] R. I. Bot and D.-K. Nguyen. Fast Krasnosel'skiĭ-Mann algorithm with a convergence rate of the fixed point iteration of $o(1/k)$. *arXiv preprint arXiv:2206.09462*, 2022.

[36] R. I. Boţ, M. Sedlmayer, and P. T. Vuong. A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANs. *Journal of Machine learning research*, 2022. to appear.

[37] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[38] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[39] R. E. Bruck Jr. An iterative solution of a variational inequality for certain monotone operators in hilbert space. *Bulletin of the American Mathematical Society*, 81(5):890–892, 1975.

[40] Y. Cai, A. Oikonomou, and W. Zheng. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

[41] Y. Cai, A. Oikonomou, and W. Zheng. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022.

[42] Y. Cai and W. Zheng. Accelerated single-call methods for constrained min-max optimization. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

[43] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization theory and Applications*, 166(3):968–982, 2015.

[44] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 2011.

[45] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, pages 391–401, 2019.

[46] T. Chavdarova, M. I. Jordan, and M. Zampetakis. Last-iterate convergence of saddle point optimizers via high-resolution differential equations. In *The 13th Annual Workshop on Optimization for Machine Learning (OPT2021)*, 2021.

[47] T. Chavdarova, M. Pagliardini, S. U. Stich, F. Fleuret, and M. Jaggi. Taming GANs with lookahead-minmax. In *International Conference on Learning Representations*, 2021.

[48] L. Chen, D. Sun, and K.-C. Toh. An efficient inexact symmetric gauss–seidel based majorized admm for high-dimensional convex composite conic programming. *Mathematical Programming*, 161(1):237–270, 2017.

[49] L. Chen, D. Sun, and K.-C. Toh. A note on the convergence of ADMM for linearly constrained convex optimization problems. *Computational Optimization and Applications*, 66(2):327–343, 2017.

[50] L. Condat. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 2013.

[51] R. W. Cottle and J. A. Ferland. On pseudo-convex functions of nonnegative variables. *Mathematical Programming*, 1(1):95–101, 1971.

[52] E. R. Csetnek, Y. Malitsky, and M. K. Tam. Shadow Douglas–Rachford splitting for monotone inclusions. *Applied Mathematics & Optimization*, 80(3):665–678, 2019.

[53] S. Cui, U. Shanbhag, M. Staudigl, and P. T. Vuong. Stochastic relaxed inertial forward-backward-forward splitting for monotone inclusions in Hilbert spaces. *Computational Optimization and Applications*, 83(2):465–524, 2022.

[54] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

[55] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9236–9246, 2018.

[56] C. Daskalakis, S. Skoulakis, and M. Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478, 2021.

[57] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

[58] D. Dua and C. Graff. UCI machine learning repository, 2019.

[59] J. Eckstein and D. P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

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

[61] M. Fortin and R. Glowinski. Chapter III: On decomposition-coordination methods using an augmented lagrangian. In *Studies in Mathematics and Its Applications*, volume 15, pages 97–146. Elsevier, 1983.

[62] D. Gabay. Chapter IX: Applications of the method of multipliers to variational inequalities. In *Studies in Mathematics and Its Applications*, volume 15, pages 299–331. Elsevier, 1983.

[63] I. Gemp and S. Mahadevan. Global convergence to the equilibrium of gans using variational inequalities. *arXiv preprint arXiv:1808.01531*, 2018.

[64] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on Generative Adversarial Networks. In *International Conference on Learning Representations*, 2019.

[65] G. Gidel, R. A. Hemmat, M. Pezeshki, R. Le Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019.

[66] P. Giselsson. Nonlinear forward-backward splitting with projection correction. *SIAM Journal on Optimization*, 31(3):2199–2226, 2021.

[67] N. Golowich, S. Pattathil, and C. Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 33:20766–20778, 2020.

[68] N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pages 1758–1784. PMLR, 2020.

[69] I. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv:1701.00160*, 2016.

[70] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

[71] E. Gorbunov, N. Loizou, and G. Gidel. Extragradient method: $\mathcal{O}(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR, 2022.

[72] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[73] N. Hadjisavvas, S. Schaible, and N.-C. Wong. Pseudomonotone operators: a survey of the theory and its applications. *Journal of Optimization Theory and Applications*, 152(1):1–20, 2012.

[74] B. Halpern. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

[75] E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[76] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[77] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. 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.

[78] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[79] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *arXiv preprint arXiv:2003.10162*, 2020.

[80] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

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

[82] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[83] M. Kestemont, V. Christlein, and D. Stutzmann. Artificial paleography: computational approaches to identifying script types in medieval manuscripts. *Speculum*, 92(S1):S86–S109, 2017.

[84] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

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

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

[87] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5:27–72, 2004.

[88] M. Li, D. Sun, and K.-C. Toh. A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM Journal on Optimization*, 26(2):922–950, 2016.

[89] T. Liang and J. Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 907–915. PMLR, 2019.

[90] Q. Lin, M. Liu, H. Rafique, and T. Yang. Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities. *arXiv:1810.10207*, 2018.

[91] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[92] M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *J. Mach. Learn. Res.*, 22:169–1, 2021.

[93] N. Loizou, H. Berard, A. Jolicoeur-Martineau, P. Vincent, S. Lacoste-Julien, and I. Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR, 2020.

[94] N. Loizou, S. Vaswani, I. H. Laradji, and S. Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

[95] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[96] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[97] R. Malayan. The art of Armenian calligraphy. `https://armeniancalligraphy.com/history.html`. Accessed: 18 December 2022.

[98] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

[99] Y. Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, 2020.

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

[101] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1990.

[102] N. Martinez, M. Bertran, and G. Sapiro. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR, 2020.

[103] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019.

[104] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

[105] L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, 2018.

[106] L. Mescheder, S. Nowozin, and A. Geiger. The numerics of gans. *arXiv preprint arXiv:1705.10461*, 2017.

[107] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

[108] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[109] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.

[110] O. Morgenstern and J. Von Neumann. *Theory of games and economic behavior*. Princeton university press, 1953.

[111] MOSEK ApS. The MOSEK optimization toolbox for MATLAB manual. version 9.0, 2019.

[112] A. Nemirovski. Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[113] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Doklady Akademija Nauk USSR*, 269:543–547, 1983.

[114] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[115] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[116] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR, 2017.

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

[118] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

[119] J. Park and E. K. Ryu. Exact optimal accelerated complexity for fixed-point iterations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 17420–17457. PMLR, 2022.

[120] G. B. Passty. Ergodic convergence to a zero of the sum of monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 72(2):383–390, 1979.

[121] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019.

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

[123] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[124] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[125] H. Rafique, M. Liu, Q. Lin, and T. Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv:1810.02060*, 2018.

[126] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 993–1019, 2013.

[127] S. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[128] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[129] R. T. Rockafellar. Monotone operators associated with saddle-functions and minimax problems. *Nonlinear functional analysis*, 18(part 1):397–407, 1970.

[130] R. T. Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970.

[131] H. Sadeghi, S. Banert, and P. Giselsson. Forward-backward splitting with deviations for monotone inclusions. *arXiv preprint arXiv:2112.00776*, 2021.

[132] H. Sadeghi, S. Banert, and P. Giselsson. Incorporating history and deviations in forward–backward splitting. *arXiv preprint arXiv:2208.05498*, 2022.

[133] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[134] S. Shreve. *Stochastic calculus for finance I: the binomial asset pricing model*. Springer Science & Business Media, 2005.

[135] Q. Tran-Dinh. The connection between Nesterov's accelerated methods and Halpern fixed-point iterations. *arXiv preprint arXiv:2203.04869*, 2022.

[136] Q. Tran-Dinh and Y. Luo. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150*, 2021.

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

[138] C. Vidal-Gorène, A. Sargsyan, and E. Van Elverdinghe. Index of digitized Armenian manuscripts (v2.0.0). `https://www.armenian-manuscripts-index.com/`. Accessed: 18 December 2022.

[139] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 2013.

[140] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, pages 12098–12109. PMLR, 2021.

[141] G. Zhang, Y. Wang, L. Lessard, and R. B. Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 7659–7679. PMLR, 2022.