A Lyapunov analysis of Korpelevich's extragradient method with fast and flexible extensions

A Lyapunov analysis of Korpelevich's extragradient method with fast and flexible extensions

Manu Upadhyaya[⋆] Puya Latafat[†] Pontus Giselsson[⋆]

[⋆]Department of Automatic Control
Lund University, Lund, Sweden
{manu.upadhyaya, pontus.giselsson}@control.lth.se

[†]DYSCO (Dynamical Systems, Control, and Optimization)
IMT School for Advanced Studies Lucca, Lucca, Italy
puya.latafat@imtlucca.it

Abstract

We present a Lyapunov analysis of Korpelevich's extragradient method and establish an $\mathcal{O}(1/k)$ last-iterate convergence rate. Building on this, we propose flexible extensions that combine extragradient steps with user-specified directions, guided by a line-search procedure derived from the same Lyapunov analysis. These methods retain global convergence under practical assumptions and can achieve superlinear rates when directions are chosen appropriately. Numerical experiments highlight the simplicity and efficiency of this approach.

Keywords. Monotone inclusions, extragradient method, Lyapunov analysis, superlinear convergence

Mathematics subject classification 2020. 47J20, 47J25, 47J26, 47H05, 65K10, 65K15, 65J15

1 Introduction

In this work, we consider the inclusion problem

$$\text{find } z \in \mathcal{H} \ \text{ such that } \ 0 \in F(z) + \partial g(z), \tag{1.1}$$

where $F : \mathcal{H} \to \mathcal{H}$ is monotone and L_F-Lipschitz continuous for some $L_F \in \mathbb{R}_{++}$, $g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex, and lower semicontinuous function, and $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a real Hilbert space. Inclusion problems of the form (1.1) are known as *hemivariational inequalities* [25] or *(mixed) variational inequalities* [23, 29], and frequently arise in fundamental mathematical programming problems—either directly or through reformulation—including minimization, saddle-point, complementarity, Nash equilibrium, and fixed-point problems [6, 15]. The most common methods for solving (1.1) belong to the large class of extragradient-type methods [21, 33, 42]. For a recent review, see [40, 41]. Among these methods, the first and most widely recognized is Korpelevich's extragradient method [21]. Although originally proposed for the constrained case in which g is the indicator function of a nonempty, closed, and convex set, the method also applies to the more general setting in (1.1). Specifically, given an initial point $z^0 \in \mathcal{H}$ and a step-size parameter $\gamma \in \mathbb{R}_{++}$, its iterations are given by

$$\bar{z}^k = \operatorname{prox}_{\gamma g}\left(z^k - \gamma F(z^k)\right), \tag{1.2a}$$

$$z^{k+1} = \operatorname{prox}_{\gamma g}\left(z^k - \gamma F(\bar{z}^k)\right) \tag{1.2b}$$

for each $k \in \mathbb{N}_0$. A popular alternative is Tseng's forward-backward-forward method [42], given by

$$\bar{z}^k = \operatorname{prox}_{\gamma g}\left(z^k - \gamma F(z^k)\right), \tag{1.3a}$$

$$z^{k+1} = \bar{z}^k + \gamma(F(z^k) - F(\bar{z}^k)) \tag{1.3b}$$

for each $k \in \mathbb{N}_0$, that requires one less evaluation of the proximal operator $\operatorname{prox}_{\gamma g}$ per iteration. Classically, the convergence analyses of these methods rely on Fejér-type arguments [21, 42].

In this work, we propose an analysis centered around the Lyapunov function

$$\mathcal{V}(z^k, \bar{z}^k, z^{k+1}) = 2\gamma^{-1}\langle z^k - z^{k+1}, F(z^k) - F(\bar{z}^k)\rangle + \gamma^{-2}\|z^{k+1} - \bar{z}^k\|^2 + \gamma^{-2}\|z^k - z^{k+1}\|^2.$$

For the extragradient method, $\mathcal{V}_k$ serves as a nonnegative optimality measure for the inclusion problem (1.1) as shown in Proposition 2.1. The same property can also be easily verified for Tseng's method since the Lyapunov function reduces to $\mathcal{V}_k = \gamma^{-2}\|z^k - \bar{z}^k\|^2$ in this case. In the particular case when $g = 0$, both methods are identical, and the Lyapunov function reduces to $\mathcal{V}_k = \|F(z^k)\|^2$.

Besides being an optimality measure, we show in Theorem 2.2 that $\mathcal{V}_k$ satisfies a descent inequality for the extragradient method. Moreover, we show a Fejér-type inequality in which $\mathcal{V}_k$ appears as the residual term (see Theorem 2.4). By combining this result with the descent property for $\mathcal{V}_k$, we establish an $\mathcal{O}(1/k)$ last-iterate convergence rate for the extragradient method, as shown in Corollary 2.5. This improves the standard $\mathcal{O}(1/k)$ best-so-far convergence rate for the forward-backward residual that immediately follows from Korpelevich's classical analysis, i.e., $\min_{i \in [\![0,k]\!]} \|\bar{z}^i - z^i\|^2 = \mathcal{O}(1/k)$.

Interestingly, these results are particular to Korpelevich's extragradient method. We demonstrate through a simple counterexample in Example B.1 that the descent inequality fails for Tseng's method. Moreover, even for the extragradient method, it is crucial to leverage the specific structure of (1.1). Indeed, the claimed descent inequality fails, and even the convergence of the method does not hold if we replace ∂f with a maximally monotone operator $T : \mathcal{H} \to 2^{\mathcal{H}}$ and correspondingly the proximal operator $\operatorname{prox}_{\gamma g}$ in (1.2) with the resolvent $(\mathrm{Id} + \gamma T)^{-1}$. This broader setting is ruled out by a counterexample presented in Example B.2.

The second objective of this work is to develop flexible extragradient-type schemes that accommodate fast local directions while maintaining global convergence. In this regard, the seminal work [37] introduced an inexact regularized Newton method for solving monotone equations, i.e., when $g = 0$ in (1.1). Global convergence in [37] is achieved through a scheme that relies on the hyperplane projection idea from [36]. Specifically, an intricate line search is used to identify a hyperplane separating the current iterate from the solution set, and the method then projects onto this hyperplane to ensure global convergence. Consequently, even though a line search is performed, the convergence analysis still relies on Fejér-type techniques. In practice, however, the projection step can undermine the effectiveness of the Newtonian directions, resulting in slower convergence.

Another related work to ours is [39], which addresses the problem of finding fixed points of averaged operators. They propose a hybrid scheme accelerating many numerical algorithms under the Krasnosel'skiĭ–Mann framework. Similar to [37], their scheme incorporates a hyperplane projection step and achieves superlinear convergence under suitable assumptions. In addition, it allows for a general class of local directions, including quasi-Newton-type directions, providing greater flexibility in practice.

In contrast to the approaches mentioned earlier that use a line search to identify a separating hyperplane onto which the iterates are projected (see [37, 39]), our proposed schemes incorporate line search procedures grounded in our new Lyapunov analysis, directly aiming to reduce $\mathcal{V}_k$. We introduce three flexible algorithms tailored to specific instances of problem (1.1). `FLEX` (Algorithm 1) is introduced for finding zeros of F, `I-FLEX` (Algorithm 2) is applicable when F is injective, and `Prox-FLEX` (Algorithm 3) addresses problem (1.1) in its full generality. All three algorithms share the same guiding principle: at each iteration, one performs a convex combination of a standard extragradient step (1.2) and a step based on a user-specified direction. The specific weighting for this convex combination is determined by the line-search procedure, ensuring sufficient descent of the optimality measure $\mathcal{V}_k$. Similar to [39], our schemes accommodate a wide range of user-chosen directions, including quasi-Newton-type directions. A key feature enabling this approach is that $\mathcal{V}_k$ depends solely on values computed at each iteration and does not involve a solution to (1.1). This design ensures high flexibility while guaranteeing global (see Section 3) and superlinear convergence (see Section 5) when choosing suitable directions.

Our preliminary numerical experiments indicate that using quasi-Newton directions in our proposed algorithms yields favorable performance. In particular, limited-memory type-I and type-II Anderson acceleration exhibit promising results (see Section 6). In related work, [45] studies Anderson acceleration for finding fixed points of averaged operators, proposing a globalization strategy based on a stabilization and safeguarding mechanism—rather than a line search—that reverts to a nominal Krasnosel'skiĭ–Mann step whenever the Anderson acceleration step fails to sufficiently reduce the forward residual. More recently, [34] introduced an extragradient-based scheme with memory-one Anderson acceleration, which reduces overhead and allows for simple, explicit updates of the directions. Furthermore, [4] presents a quasi-Newton method tailored to minimax problems. Our theory offers a direct globalization strategy for such directions, applicable in the uniformly monotone and injective settings (see Theorem 3.4), or whenever the resulting directions are summable (see Theorem 3.2.(i)).

1.1 Organization

In Section 2, we formally introduce the new Lyapunov analysis for Korpelevich's extragradient method. Building on this framework, we present the three new algorithms in Section 3 and establish their global convergence under suitable assumptions. In Section 4, we provide detailed proofs of the results from the preceding section. Next, Section 5 focuses on superlinear convergence, including corresponding proofs for the proposed algorithms. Numerical experiments appear in Section 6, and we conclude in Section 7 with a summary of key findings and directions for future research. Finally, Appendix A offers background material on Korpelevich's extragradient method, and Appendix B presents the counterexamples mentioned earlier.

1.2 Notation and preliminaries

Let $\mathbb{N}_0$ denote set of nonnegative integers, $\mathbb{N}$ the set of postive integers, $\mathbb{Z}$ the set of integers, $[\![n,m]\!] = \{l \in \mathbb{Z} \mid n \leq l \leq m\}$ the set of integers inclusively between the integers n and m, $\mathbb{R}$

the set of real numbers, $\mathbb{R}_+$ the set of nonnegative real numbers, $\mathbb{R}_{++}$ the set of positive real numbers, $\mathbb{R}^n$ the set of all n-tuples of elements of $\mathbb{R}$, $\mathbb{R}^{m\times n}$ the set of real-valued matrices of size $m\times n$, if $M\in\mathbb{R}^{m\times n}$ then $[M]_{i,j}$ the i,j-th element of M, $\mathbb{S}^n$ the set of symmetric real-valued matrices of size $n\times n$, and $\mathbb{S}^n_+\subseteq\mathbb{S}^n$ the set of positive semidefinite real-valued matrices of size $n\times n$. Suppose that $1\le p<+\infty$, $K\subseteq\mathbb{N}_0$, and $\mathcal{U}\subseteq\mathcal{W}$, where $\mathcal{W}$ is a normed space. Then we define the space $\ell^p(K;\mathcal{U})=\{(u^k)_{k\in K}\in\mathcal{U}^K \mid \sum_{k\in K}\|u^k\|^p<+\infty\}$.

Throughout this paper, $(\mathcal{H},\langle\cdot,\cdot\rangle)$ will denote a real Hilbert space and $\|\cdot\|$ the canonical norm, which will be clear from the context. Let $F:\mathcal{H}\to\mathcal{H}$ be an operator. Suppose that $L_F\ge 0$. The operator F is said to be *L_F-Lipschitz continuous* if $\|F(x)-F(y)\|\le L_F\|x-y\|$ for each $x,y\in\mathcal{H}$. The operator F is said to be *monotone* if $0\le\langle F(x)-F(y),x-y\rangle$ for each $x,y\in\mathcal{H}$. Suppose that $\mu_F\ge 0$. The operator F is said to be *μ_F-strongly monotone* if $\mu_F\|x-y\|^2\le\langle F(x)-F(y),x-y\rangle$ for each $x,y\in\mathcal{H}$. Moreover, F is said to be *uniformly monotone with modulus* $\phi:\mathbb{R}_+\to\mathbb{R}_+$ if ϕ is increasing, vanishes only at 0, and $\phi(\|x-y\|)\le\langle F(x)-F(y),x-y\rangle$ for each $x,y\in\mathcal{H}$. For a general set-valued operator $T:\mathcal{H}\to 2^{\mathcal{H}}$, the set of *zeros* is denoted by $\operatorname{zer}(T)=\{x\in\mathcal{H}\mid 0\in T(x)\}$. The Cauchy–Schwarz inequality states that $|\langle x,y\rangle|\le\|x\|\|y\|$ for each $x,y\in\mathcal{H}$ and Young's inequality that $2\langle x,y\rangle\le\alpha\|x\|^2+\alpha^{-1}\|y\|^2$ for each $x,y\in\mathcal{H}$ and $\alpha>0$.

Given a function $g:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$, the *effective domain* of g is the set $\operatorname{dom} g=\{x\in\mathcal{H}\mid f(x)<+\infty\}$. The function g is said to be *proper* if $\operatorname{dom} g\neq\emptyset$. The *subdifferential* of a proper function g is the set-valued operator $\partial g:\mathcal{H}\to 2^{\mathcal{H}}$ defined as the mapping $x\mapsto\{u\in\mathcal{H}\mid\forall y\in\mathcal{H}, g(y)\ge g(x)+\langle u,y-x\rangle\}$. The function g is said to be *convex* if $g((1-\lambda)x+\lambda y)\le(1-\lambda)g(x)+\lambda g(y)$ for each $x,y\in\mathcal{H}$ and $0\le\lambda\le 1$. The function g is said to be *lower semicontinuous* if $\liminf_{y\to x}g(y)\ge g(x)$ for each $x\in\mathcal{H}$. If $C\subseteq\mathcal{H}$, the *indicator function of* C, denoted $\delta_C:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$, is defined as $\delta_C(x)=0$ if $x\in C$ and $\delta_C(x)=+\infty$ if $x\in\mathcal{H}\setminus C$.

Let $g:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ be proper, convex and lower semicontinuous, and let $\gamma>0$. Then the *proximal operator* $\operatorname{prox}_{\gamma g}:\mathcal{H}\to\mathcal{H}$ is defined as the single-valued operator given by

$$\operatorname{prox}_{\gamma g}(x)=\operatorname*{argmin}_{z\in\mathcal{H}}\left(g(z)+\tfrac{1}{2\gamma}\|x-z\|^2\right)$$

for each $x\in\mathcal{H}$ [6, Proposition 12.15]. If $x,p\in\mathcal{H}$, then $p=\operatorname{prox}_{\gamma g}(x)\Leftrightarrow\gamma^{-1}(x-p)\in\partial g(p)$ $\Leftrightarrow 0\le g(y)-g(p)-\langle\gamma^{-1}(x-p),y-p\rangle$ for each $y\in\mathcal{H}$ [6, Proposition 16.44, Proposition 16.6].

2 A new Lyapunov analysis

Classical convergence analyses of Korpelevich's extragradient method (1.2) typically rely on Fejér-type arguments, as discussed in Appendix A. In this section, we introduce a complementary Lyapunov inequality that not only leads to a last-iterate result but also forms the basis of the new algorithms presented in Section 3. Throughout this work, we investigate (1.1) under the following assumption.

Assumption I: *The following hold in problem* (1.1)*.*

(i) $F:\mathcal{H}\to\mathcal{H}$ *is monotone and L_F-Lipschitz continuous for some* $L_F\in\mathbb{R}_{++}$*.*

(ii) $g:\mathcal{H}\to\mathbb{R}\cup\{+\infty\}$ *is proper, convex and lower semicontinuous.*

Our analysis is centered around the Lyapunov function $\mathcal{V}:\mathcal{H}^3\to\mathbb{R}$ given by

$$\mathcal{V}(z,\bar z,z^+)=2\gamma^{-1}\langle z-z^+,F(z)-F(\bar z)\rangle+\gamma^{-2}\|z^+-\bar z\|^2+\gamma^{-2}\|z-z^+\|^2, \tag{2.1}$$

for each $(z,\bar{z},z^+)\in\mathcal{H}^3$. Proposition 2.1 establishes that $\mathcal{V}$ is generally a valid optimality measure for the inclusion problem in (1.1). For notational convenience, we define the algorithmic operators $T_1^\gamma, T_2^\gamma : \mathcal{H}\to\mathcal{H}$ by

$$T_1^\gamma = \operatorname{prox}_{\gamma g}\circ(\mathrm{Id}-\gamma F) \quad\text{and}\quad T_2^\gamma = \operatorname{prox}_{\gamma g}\circ(\mathrm{Id}-\gamma F\circ T_1^\gamma), \tag{2.2}$$

where $\gamma\in\mathbb{R}_{++}$ is the step-size parameter. With this notation, the iterates of (1.2) can be written compactly as $\bar{z}^k = T_1^\gamma(z^k)$ and $z^{k+1} = T_2^\gamma(z^k)$.

Proposition 2.1: *Suppose that Assumption I holds. Let $\gamma\in(0,1/L_F)$, $z\in\mathcal{H}$, $\bar{z}=T_1^\gamma(z)$ and $z^+=T_2^\gamma(z)$ where T_1^γ and T_2^γ are the algorithmic operators defined in (2.2), and $\mathcal{V}$ the Lyapunov function defined in (2.1). Then the following hold.*

(i) $\mathcal{V}(z,\bar{z},z^+)\geq(1-\gamma L_F)\gamma^{-2}(\|z^+-\bar{z}\|^2+\|z^+-z\|^2)\geq 0$.

(ii) $\mathcal{V}(z,\bar{z},z^+)=0$ *if and only if* $z=\bar{z}=z^+\in\operatorname{zer}(F+\partial g)$.

Proof.

2.1.(i): The inner product in the definition of $\mathcal{V}$ can be written as

$$\begin{aligned}\langle z-z^+,F(z)-F(\bar{z})\rangle &= \langle z-z^+,F(z)-F(z^+)\rangle+\langle z-z^+,F(z^+)-F(\bar{z})\rangle\\ &\geq \langle z-z^+,F(z^+)-F(\bar{z})\rangle\\ &\geq -\|z-z^+\|\|F(z^+)-F(\bar{z})\|\\ &\geq -L_F\|z-z^+\|\|z^+-\bar{z}\|\\ &\geq -\tfrac{L_F}{2}(\|z-z^+\|^2+\|z^+-\bar{z}\|^2),\end{aligned}$$

where monotonicity of F is used in the first inequality, the Cauchy–Schwarz inequality is used in the second inequality, Lipschitz continuity of F is used in the third inequality, and Young's inequality for products is used in the fourth inequality. The lower bound of $\mathcal{V}$ follows from using this inequality in (2.1) and the assumption $\gamma L_F\in(0,1)$.

2.1.(ii): Suppose that $\mathcal{V}(z,\bar{z},z^+)=0$. Then Proposition 2.1.(i) and Proposition A.2.(ii) imply that $z=\bar{z}=z^+\in\operatorname{zer}(F+\partial g)$. Conversely, suppose that $z=\bar{z}=z^+\in\operatorname{zer}(F+\partial g)$. Then it is clear from (2.1) that $\mathcal{V}(z,\bar{z},z^+)=0$. □

The following result shows that $\mathcal{V}$ is, in fact, a suitable Lyapunov function for the extragradient method (1.2), i.e., it fulfills a decent inequality. Moreover, the descent inequality neither contains a solution of (1.1) nor assumes the existence of a solution.

Theorem 2.2: *Suppose that Assumption I holds and the sequence $((z^k,\bar{z}^k))_{k\in\mathbb{N}_0}$ is generated by (1.2) with initial point $z^0\in\mathcal{H}$ and step-size parameter $\gamma\in\mathbb{R}_{++}$. Then*

$$\mathcal{V}_{k+1}\leq\mathcal{V}_k-(1-\gamma^2L_F^2)\gamma^{-2}\|z^{k+1}-\bar{z}^k\|^2 \tag{2.3}$$

for each $k\in\mathbb{N}_0$, where $\mathcal{V}_k=\mathcal{V}(z^k,\bar{z}^k,z^{k+1})$ for the Lyapunov function $\mathcal{V}$ defined in (2.1).

Proof. Note that the first and second proximal steps in (1.2) can equivalently be written via their subgradient characterization as

$$\gamma^{-1}(z^k-\bar{z}^k)-F(z^k)\in\partial g(\bar{z}^k) \quad\text{and}\quad \gamma^{-1}(z^k-z^{k+1})-F(\bar{z}^k)\in\partial g(z^{k+1}), \tag{2.4}$$

respectively. Using the subgradient inequality at the points z^{k+1}, z^{k+2} and $\bar{z}^{k+1}$, with the particular subgradients given in (2.4), it follows that

$$0 \geq g(z^{k+1}) - g(z) + \langle \gamma^{-1}(z^k - z^{k+1}) - F(\bar{z}^k), z - z^{k+1} \rangle, \tag{2.5a}$$

$$0 \geq g(z^{k+2}) - g(z) + \langle \gamma^{-1}(z^{k+1} - z^{k+2}) - F(\bar{z}^{k+1}), z - z^{k+2} \rangle, \tag{2.5b}$$

$$0 \geq g(\bar{z}^{k+1}) - g(z) + \langle \gamma^{-1}(z^{k+1} - \bar{z}^{k+1}) - F(z^{k+1}), z - \bar{z}^{k+1} \rangle. \tag{2.5c}$$

holds for any $z \in \mathcal{H}$, respectively. Picking $z = \bar{z}^{k+1}$ in (2.5a), $z = z^{k+1}$ in (2.5b), $z = z^{k+2}$ in (2.5c), summing the resulting inequalities, and multiplying by $2\gamma^{-1}$ gives

$$\begin{aligned} 0 \geq\ & 2\gamma^{-2}\langle z^k - z^{k+1}, \bar{z}^{k+1} - z^{k+1} \rangle - 2\gamma^{-1}\langle F(\bar{z}^k), \bar{z}^{k+1} - z^{k+1} \rangle \\ & + 2\gamma^{-2}\|z^{k+1} - z^{k+2}\|^2 - 2\gamma^{-1}\langle F(\bar{z}^{k+1}), z^{k+1} - z^{k+2} \rangle \\ & + 2\gamma^{-2}\langle z^{k+1} - \bar{z}^{k+1}, z^{k+2} - \bar{z}^{k+1} \rangle - 2\gamma^{-1}\langle F(z^{k+1}), z^{k+2} - \bar{z}^{k+1} \rangle. \end{aligned} \tag{2.6}$$

The first two inner products in (2.6) can be simplified as

$$\begin{aligned} A_k &= 2\gamma^{-2}\langle z^k - z^{k+1}, \bar{z}^{k+1} - z^{k+1} \rangle - 2\gamma^{-1}\langle F(\bar{z}^k), \bar{z}^{k+1} - z^{k+1} \rangle \\ &= \gamma^{-2}\|z^k - z^{k+1}\|^2 + \gamma^{-2}\|\bar{z}^{k+1} - z^{k+1}\|^2 - \gamma^{-2}\|z^k - \bar{z}^{k+1}\|^2 - 2\gamma^{-1}\langle F(\bar{z}^k), \bar{z}^{k+1} - z^{k+1} \rangle, \end{aligned}$$

where the identity $2\langle x, y \rangle = \|x\|^2 + \|y\|^2 - \|x - y\|^2$ for each $x, y \in \mathcal{H}$ is used in the second equality, while the remaining four terms in (2.6) can be simplified as

$$\begin{aligned} B_k &= 2\gamma^{-2}\|z^{k+1} - z^{k+2}\|^2 - 2\gamma^{-1}\langle F(\bar{z}^{k+1}), z^{k+1} - z^{k+2} \rangle \\ &\quad + 2\gamma^{-2}\langle z^{k+1} - \bar{z}^{k+1}, z^{k+2} - \bar{z}^{k+1} \rangle - 2\gamma^{-1}\langle F(z^{k+1}), z^{k+2} - \bar{z}^{k+1} \rangle \\ &= 2\gamma^{-1}\langle z^{k+1} - z^{k+2}, F(z^{k+1}) - F(\bar{z}^{k+1}) \rangle + \gamma^{-2}\|z^{k+2} - \bar{z}^{k+1}\|^2 + \gamma^{-2}\|z^{k+1} - z^{k+2}\|^2 \\ &\quad + \gamma^{-2}\|z^{k+1} - \bar{z}^{k+1}\|^2 - 2\gamma^{-1}\langle F(z^{k+1}), z^{k+1} - \bar{z}^{k+1} \rangle \\ &= \mathcal{V}_{k+1} + \gamma^{-2}\|z^{k+1} - \bar{z}^{k+1}\|^2 - 2\gamma^{-1}\langle F(z^{k+1}), z^{k+1} - \bar{z}^{k+1} \rangle, \end{aligned}$$

where the identity $2\langle x, y \rangle = \|x\|^2 + \|y\|^2 - \|x - y\|^2$ for each $x, y \in \mathcal{H}$ is used in the second equality, and the explicit expression for $\mathcal{V}_{k+1} = \mathcal{V}(z^{k+1}, \bar{z}^{k+1}, z^{k+2})$ is used in the third equality. Therefore, the inequality $A_k + B_k \leq 0$ in (2.6) can be rearranged as

$$\begin{aligned} \mathcal{V}_{k+1} \leq\ & \gamma^{-2}\|z^k - \bar{z}^{k+1}\|^2 - \gamma^{-2}\|z^k - z^{k+1}\|^2 - 2\gamma^{-2}\|\bar{z}^{k+1} - z^{k+1}\|^2 \\ & + 2\gamma^{-1}\underbrace{\langle F(\bar{z}^k) - F(z^{k+1}), \bar{z}^{k+1} - z^{k+1} \rangle}_{=C_k}. \end{aligned} \tag{2.7}$$

We can upper bound the term C_k as

$$\begin{aligned} C_k &= \langle F(z^k) - F(z^{k+1}), z^{k+1} - z^k \rangle + \langle F(\bar{z}^k) - F(z^k), z^{k+1} - z^k \rangle \\ &\quad + \langle F(\bar{z}^k) - F(z^{k+1}), \bar{z}^{k+1} + z^k - 2z^{k+1} \rangle \\ &\leq \langle F(\bar{z}^k) - F(z^k), z^{k+1} - z^k \rangle \\ &\quad + \tfrac{\gamma}{2}\|F(\bar{z}^k) - F(z^{k+1})\|^2 + \tfrac{1}{2\gamma}\|\bar{z}^{k+1} - 2z^{k+1} + z^k\|^2 \\ &\leq \langle F(\bar{z}^k) - F(z^k), z^{k+1} - z^k \rangle + \tfrac{\gamma L_F^2}{2}\|\bar{z}^k - z^{k+1}\|^2 \\ &\quad + \tfrac{1}{2\gamma}(2\|z^k - z^{k+1}\|^2 + 2\|\bar{z}^{k+1} - z^{k+1}\|^2 - \|z^k - \bar{z}^{k+1}\|^2) \end{aligned} \tag{2.8}$$

where monotonicity of F and Young's inequality is used in the first inequality, and Lipschitz continuity of F along with the identity $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$ for each $x, y \in \mathcal{H}$ is used in the last inequality.

Combining (2.7) and (2.8), and using $\mathcal{V}_k = 2\gamma^{-1}\langle z^k - z^{k+1}, F(z^k) - F(\bar{z}^k)\rangle + \gamma^{-2}\|z^{k+1} - \bar{z}^k\|^2 + \gamma^{-2}\|z^k - z^{k+1}\|^2$ gives

$$\begin{aligned}\mathcal{V}_{k+1} - \mathcal{V}_k &\le 2\gamma^{-1}\langle F(\bar{z}^k) - F(z^k), z^{k+1} - z^k\rangle + L_F^2\|\bar{z}^k - z^{k+1}\|^2 \\ &\quad + \gamma^{-2}\|z^k - z^{k+1}\|^2 - \mathcal{V}_k \\ &= -(1-\gamma^2 L_F^2)\gamma^{-2}\|\bar{z}^k - z^{k+1}\|^2,\end{aligned}$$

as claimed. ☐

Next, we present Corollary 2.3, which follows immediately from Theorem 2.2 by letting $g = 0$. Observe that the resulting inequality (2.10) strengthens [19, Lemma 3.2] and [10, Remark 2.1] by incorporating the additional residual term $(1-\gamma^2 L_F^2)\,\|F(z^k) - F(\bar{z}^k)\|^2$.

Corollary 2.3: *Suppose that Assumption I.(i) holds and the sequence $((z^k, \bar{z}^k))_{k\in\mathbb{N}_0}$ is generated by*

$$\begin{aligned}\bar{z}^k &= z^k - \gamma F(z^k), \\ z^{k+1} &= z^k - \gamma F(\bar{z}^k)\end{aligned}$$

for each $k \in \mathbb{N}_0$, with initial point $z^0 \in \mathcal{H}$ and step-size parameter $\gamma \in \mathbb{R}_{++}$. Then,

$$\|F(z^{k+1})\|^2 \le \|F(z^k)\|^2 - (1-\gamma^2 L_F^2)\|F(z^k) - F(\bar{z}^k)\|^2 \tag{2.10}$$

for each $k \in \mathbb{N}_0$. Moreover, if $\gamma \in (0, 1/L_F)$, then for any $k \in \mathbb{N}_0$ and $z^\star \in \operatorname{zer}(F)$ it holds that

$$\|F(z^k)\|^2 \le \frac{\|z^0 - z^\star\|^2}{\gamma^2(1-\gamma^2 L_F^2)(k+1)}. \tag{2.11}$$

Proof. Letting $g = 0$ in Theorem 2.2 gives (2.10). Using $g = 0$, (A.4) in Proposition A.3 gives

$$\|z^{i+1} - z^\star\|^2 \le \|z^i - z^\star\|^2 - \gamma^2(1-\gamma^2 L_F^2)\|F(z^i)\|^2. \tag{2.12}$$

for each $i \in \mathbb{N}_0$. Inductively summing (2.12) from $i = 0$ to $i = k$, rearranging, and dividing by $\gamma^2(1-\gamma^2 L_F^2)(k+1)$ gives that

$$\begin{aligned}\|F(z^k)\|^2 &\le \frac{1}{k+1}\sum_{i=0}^{k}\|F(z^i)\|^2 \\ &\le \frac{\sum_{i=0}^{k}\left(\|z^i - z^\star\|^2 - \|z^{i+1} - z^\star\|^2\right)}{\gamma^2(1-\gamma^2 L_F^2)(k+1)} \\ &\le \frac{\|z^0 - z^\star\|^2}{\gamma^2(1-\gamma^2 L_F^2)(k+1)}\end{aligned}$$

for each $k \in \mathbb{N}_0$, where (2.10) is used in the first inequality. ☐

The following result shows that $\mathcal{V}_k$, when scaled by a nonnegative constant, equals the residual of a Fejér-type inequality. As a direct consequence, this gives an $\mathcal{O}(1/k)$ last-iterate convergence result in terms of $\mathcal{V}_k$ as presented in Corollary 2.5.

Theorem 2.4: *Suppose that Assumption I holds, the sequence $((z^k, \bar{z}^k))_{k\in\mathbb{N}_0}$ is generated by (1.2) with initial point $z^0 \in \mathcal{H}$ and step-size parameter $\gamma \in (0, 1/L_F]$, and the sequence*

$(\mathcal{V}_k)_{k\in\mathbb{N}_0}$ *is given by* $\mathcal{V}_k = \mathcal{V}(z^k, \bar{z}^k, z^{k+1})$ *for each* $k \in \mathbb{N}_0$ *and the Lyapunov function* $\mathcal{V}$ *defined in* (2.1)*. Then, for any* $k \in \mathbb{N}_0$ *and* $z^\star \in \operatorname{zer}(F + \partial g)$ *it holds that*

$$\|z^{k+1} - z^\star\|^2 \leq \|z^k - z^\star\|^2 - \alpha(\gamma, L_F)\mathcal{V}_k, \tag{2.13}$$

where

$$\alpha(\gamma, L_F) = \tfrac{\gamma^2}{2}\left(\sqrt{5 - 4\gamma^2 L_F^2} - 1\right) \geq 0. \tag{2.14}$$

Proof. Note that (2.4) and $-F(z^\star) \in \partial g(z^\star)$ can equivalently be characterized by

$$0 \leq g(z) - g(\bar{z}^k) - \langle \gamma^{-1}(z^k - \bar{z}^k) - F(z^k), z - \bar{z}^k \rangle, \tag{2.15a}$$
$$0 \leq g(z) - g(z^{k+1}) - \langle \gamma^{-1}(z^k - z^{k+1}) - F(\bar{z}^k), z - z^{k+1} \rangle, \tag{2.15b}$$
$$0 \leq g(z) - g(z^\star) - \langle -F(z^\star), z - z^\star \rangle \tag{2.15c}$$

for each $z \in \mathcal{H}$. Picking $z = z^{k+1}$ in (2.15a), $z = z^\star$ in (2.15b), $z = \bar{z}^k$ in (2.15c), summing the resulting inequalities, and multiplying by 2γ gives

$$\begin{aligned}
0 \leq\ & -2\langle z^k - \bar{z}^k, z^{k+1} - \bar{z}^k \rangle + 2\gamma\langle F(z^k), z^{k+1} - \bar{z}^k \rangle \\
& -2\langle z^k - z^{k+1}, z^\star - z^{k+1} \rangle + 2\gamma\langle F(\bar{z}^k), z^\star - z^{k+1} \rangle \\
& -2\gamma\langle F(\bar{z}^k) - F(z^\star), \bar{z}^k - z^\star \rangle + 2\gamma\langle F(\bar{z}^k), \bar{z}^k - z^\star \rangle \\
\leq\ & \|z^k - z^{k+1}\|^2 - \|z^k - \bar{z}^k\|^2 - \|z^{k+1} - \bar{z}^k\|^2 + 2\gamma\langle F(z^k), z^{k+1} - \bar{z}^k \rangle \\
& \|z^k - z^\star\|^2 - \|z^k - z^{k+1}\|^2 - \|z^\star - z^{k+1}\|^2 + 2\gamma\langle F(\bar{z}^k), z^\star - z^{k+1} \rangle \\
& + 2\gamma\langle F(\bar{z}^k), \bar{z}^k - z^\star \rangle \\
=\ & \|z^k - z^\star\|^2 - \|z^\star - z^{k+1}\|^2 - \|z^k - \bar{z}^k\|^2 - \|z^{k+1} - \bar{z}^k\|^2 \\
& - 2\gamma\langle F(z^k) - F(\bar{z}^k), \bar{z}^k - z^{k+1} \rangle,
\end{aligned} \tag{2.16}$$

where the identity $-2\langle x, y\rangle = \|x - y\|^2 - \|x\|^2 - \|y\|^2$ for each $x, y \in \mathcal{H}$ and monotonicity of F is used in the second inequality. Picking $z = z^{k+1}$ in (2.15a), $z = \bar{z}^k$ in (2.15b), and summing the resulting inequalities gives

$$\begin{aligned}
0 &\leq g(z^{k+1}) - g(\bar{z}^k) - \langle \gamma^{-1}(z^k - \bar{z}^k) - F(z^k), z^{k+1} - \bar{z}^k \rangle \\
&\quad + g(\bar{z}^k) - g(z^{k+1}) - \langle \gamma^{-1}(z^k - z^{k+1}) - F(\bar{z}^k), \bar{z}^k - z^{k+1} \rangle \\
&= -\langle F(z^k) - F(\bar{z}^k), \bar{z}^k - z^{k+1} \rangle - \gamma^{-1}\|\bar{z}^k - z^{k+1}\|^2.
\end{aligned} \tag{2.17}$$

For notational simplicity, we let $\alpha = \alpha(\gamma, L_F)$ for $\alpha(\gamma, L_F)$ as in (2.14), where simple algebra shows that $\alpha \geq 0$ if and only if $\gamma L_F \leq 1$. Multiplying (2.17) with $2\alpha\gamma^{-1}$, and adding the result to (2.16) gives

$$\begin{aligned}
0 &\leq \|z^k - z^\star\|^2 - \|z^{k+1} - z^\star\|^2 - \|z^k - \bar{z}^k\|^2 - (1 + 2\alpha\gamma^{-2})\|z^{k+1} - \bar{z}^k\|^2 \\
&\quad - 2\gamma(1 + \alpha\gamma^{-2})\langle F(z^k) - F(\bar{z}^k), \bar{z}^k - z^{k+1} \rangle \\
&= \|z^k - z^\star\|^2 - \|z^{k+1} - z^\star\|^2 - \alpha\mathcal{V}_k + A_k,
\end{aligned}$$

where

$$\begin{aligned}
A_k =\ & -\|z^k - \bar{z}^k\|^2 - (1 + 2\alpha\gamma^{-2})\|z^{k+1} - \bar{z}^k\|^2 - 2\gamma(1 + \alpha\gamma^{-2})\langle F(z^k) - F(\bar{z}^k), \bar{z}^k - z^{k+1} \rangle \\
& + \alpha\mathcal{V}_k \\
=\ & -\|z^k - \bar{z}^k\|^2 - (1 + \alpha\gamma^{-2})\|z^{k+1} - \bar{z}^k\|^2 + \alpha\gamma^{-2}\|z^k - z^{k+1}\|^2 \\
& + 2\gamma\underbrace{\langle \alpha\gamma^{-2}(z^k - z^{k+1}) - (1 + \alpha\gamma^{-2})(\bar{z}^k - z^{k+1}), F(z^k) - F(\bar{z}^k) \rangle}_{=B_k},
\end{aligned} \tag{2.18}$$

where we substituted $\mathcal{V}_k$.

To complete the proof, it is enough to show that $A_k \le 0$. The last inner product in (2.18) can be upper bounded using Young inequality as

$$\begin{aligned} B_k &\le \tfrac{\gamma}{2}\|F(z^k)-F(\bar z^k)\|^2 + \tfrac{1}{2\gamma}\|\alpha\gamma^{-2}(z^k-z^{k+1})-(1+\alpha\gamma^{-2})(\bar z^k - z^{k+1})\|^2 \\ &\le \tfrac{\gamma L_F^2}{2}\|z^k-\bar z^k\|^2 \\ &\quad + \tfrac{1}{2\gamma}\big((1+\alpha\gamma^{-2})\|\bar z^k - z^{k+1}\|^2 + \alpha\gamma^{-2}(1+\alpha\gamma^{-2})\|z^k-\bar z^k\|^2 - \alpha\gamma^{-2}\|z^k-z^{k+1}\|^2\big), \end{aligned} \tag{2.19}$$

where Lipschitz continuity of F and the identity $\|\beta x-(1+\beta)y\|^2 = (1+\beta)\|y\|^2 + \beta(1+\beta)\|x-y\|^2 - \beta\|x\|^2$ for each $x,y\in\mathcal{H}$ and each $\beta\in\mathbb{R}$ [6, Corollary 2.15] are used in the second inequality. Substituting (2.19) in (2.18) gives

$$A_k \le -(1-\gamma^2L_F^2 - \alpha\gamma^{-2}(1+\alpha\gamma^{-2}))\|z^k-\bar z^k\|^2 = 0, \tag{2.20}$$

where the last equality follows from simple algebra after substituting $\alpha = \alpha(\gamma, L_F)$ as in (2.14). ☐

Corollary 2.5: *Suppose that Assumption I holds and the sequence $((z^k,\bar z^k))_{k\in\mathbb{N}_0}$ is generated by (1.2) with initial point $z^0\in\mathcal{H}$ and step-size parameter $\gamma\in(0,1/L_F)$. Then, for any $k\in\mathbb{N}_0$ and $z^\star\in\operatorname{zer}(F+\partial g)$ it holds that*

$$\mathcal{V}_k \le \frac{\|z^0-z^\star\|^2}{\alpha(\gamma,L_F)(k+1)}, \tag{2.21}$$

where $\alpha(\gamma,L_F)>0$ is defined in (2.14) and $\mathcal{V}_k = \mathcal{V}(z^k,\bar z^k,z^{k+1})$ for the Lyapunov function $\mathcal{V}$ defined in (2.1).

Proof. The last iterate convergence result in (2.21) follows by inductively summing (2.13), dividing by $\alpha(\gamma,L_F)(k+1)$, and using monotonicity of $\mathcal{V}_k$ as shown in (2.3). ☐

3 Algorithms for monotone inclusions

In light of the descent property established in Theorem 2.2, we propose line-search extensions of the extragradient method that combine the nominal steps of (1.2) with user-specified directions. This section focuses on identifying appropriate conditions to guarantee global convergence, with detailed proofs deferred to Section 4. We deliberately leave the choice of directions open at this stage and postpone the superlinear convergence analysis to Section 5. This abstraction offers flexibility in choosing methods—such as (inexact) (quasi-)Newton approaches, Anderson acceleration, or other suitable algorithms—for computing the directions.

In the first subsection, we consider the classical extragradient setting where $g=0$ and introduce a line search based on $\|F(z^k)\|^2$ and its descent inequality in (2.10). We then extend our approach in Section 3.2 to the more general setting of (1.1). Separating the analysis in this way reflects the stronger convergence results available when $g=0$, as well as the fact that, in this case, the line search is more computationally efficient.

3.1 Fast line-search extragradient

In this subsection, we focus on the case $g=0$ in (1.1), where the Lyapunov inequality (2.3) simplifies to (2.10). The first algorithm introduced here is `FLEX` (Algorithm 1), which can

Algorithm 1: FLEX (Fast Line-search EXtragradient)

Initialize: $z^0 \in \mathcal{H}$, $\gamma \in (0, 1/L_F)$, $(\rho, \sigma, \beta) \in (0,1)^3$, $M \in \mathbb{N}_0$

1: **for** $k = 0, 1, 2, \dots$ **do**

2: $\quad \bar{z}^k = z^k - \gamma F(z^k)$; $w^k = z^k - \gamma F(\bar{z}^k)$

3: $\quad$ Compute a direction $d^k \in \mathcal{H}$ at z^k

4: $\quad$ **if** $\|F(z^k + d^k)\| \leq \rho \|F(z^k)\|$ **then**

5: $\quad\quad z^{k+1} = z^k + d^k$

6: $\quad$ **else**

7: $\quad\quad$ Set $z^{k+1} = (1 - \tau_k) w^k + \tau_k (z^k + d^k)$ where τ_k is the largest number in $\{\beta^i \mid i \in [\![1, M]\!]\} \cup \{0\}$ such that

$$\|F(z^{k+1})\|^2 \leq \|F(z^k)\|^2 - \sigma(1 - \gamma^2 L_F^2) \|F(z^k) - F(\bar{z}^k)\|^2 \tag{3.1}$$

8: $\quad$ **end if**

9: **end for**

be viewed as a hybrid scheme in the same spirit as [20, Algorithm 5.16]. At each iteration, it computes a suitable direction d^k (see Section 5) and performs the updates $z^{k+1} = z^k + d^k$ whenever the contraction condition in Step 4 holds. Otherwise, it conducts a line search based on the descent inequality (2.10), serving as a *performance safeguard*.

Before we present the convergence results for FLEX, we offer some observations on the line-search procedure.

Remark 3.1: The line-search interpolation strategy in FLEX is designed to ensure global convergence while infusing local update directions in the algorithm. It differs from standard line-search procedures in some respects.

(i) After a finite number of backtracks, the method defaults to $\tau_k = 0$, at which point (3.1) is satisfied due to (2.10) in Corollary 2.3. Taking the nominal step after a finite number of trials is not just a practical consideration but is also theoretically grounded. Without additional assumptions, it is possible that $\|F(z^k) - F(\bar{z}^k)\| = 0$ even when no solution has been found, and (3.1) is not satisfied by any $\tau_k > 0$. Therefore, additional assumptions are required for such edge cases if an infinite backtracking strategy with known finite termination is to be employed. This is further explored in Section 3.1.1.

(ii) Enforcing a descent inequality as in (3.1) of Step 7 can be viewed as a performance safeguarding. As it is shown in Theorem 3.2.(i) below, the convergence of FLEX can be guaranteed provided that $(d^k)_{k \in \mathbb{N}_0} \in \ell^1(\mathbb{N}_0; \mathcal{H})$. Therefore, the descent inequality within the line-search procedure ensures that the directions contribute effectively to the convergence, preventing arbitrarily poor performance.

The next theorem establishes global convergence of FLEX under the assumption that the directions are summable, a setting that will be revisited in Section 5, see also Theorem 5.5.(iii). Alternatively, if F is uniformly monotone, the summability assumption is dropped, as shown in Theorem 3.2.(ii). Moreover, when F is μ_F-strongly monotone, as in Theorem 3.2.(iii), a linear convergence rate is achieved.

Theorem 3.2: *Suppose that Assumption I.(i) holds,* $\operatorname{zer}(F) \neq \emptyset$*, and the sequence* $(z^k)_{k \in \mathbb{N}_0}$ *is generated by FLEX (Algorithm 1). Then, the following hold.*

Algorithm 2: I-FLEX (Injective-FLEX)

Initialize: $z^0 \in \mathcal{H}$, $\gamma \in (0, 1/L_F)$, $(\sigma, \beta) \in (0,1)^2$

1: **for** $k = 0, 1, 2, \dots$ **do**

2: $\quad \bar{z}^k = z^k - \gamma F(z^k)$; $w^k = z^k - \gamma F(\bar{z}^k)$

3: $\quad$ Compute a direction $d^k \in \mathcal{H}$ at z^k

4: $\quad$ Set $z^{k+1} = (1-\tau_k)w^k + \tau_k(z^k + d^k)$ where τ_k is the largest number in $\{\beta^i \mid i \in \mathbb{N}_0\}$ such that

$$\|F(z^{k+1})\|^2 \le \|F(z^k)\|^2 - \sigma(1-\gamma^2 L_F^2)\|F(z^k) - F(\bar{z}^k)\|^2 \tag{3.3}$$

5: **end for**

(i) If $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0; \mathcal{H})$, then $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to some point in $\operatorname{zer}(F)$*.*

(ii) If F is uniformly monotone, then $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to some point in $\operatorname{zer}(F)$*.*

(iii) If there exists $0 < \mu_F \le L_F$ such that $\mu_F\|x-y\| \le \|F(x) - F(y)\|$ for each $x, y \in \mathcal{H}$, then $(z^k)_{k\in\mathbb{N}_0}$ converges strongly to some point in $\operatorname{zer}(F)$ *and*

$$\|F(z^{k+1})\|^2 \le \underbrace{\max\left(\rho^2, 1 - \sigma\gamma^2\mu_F^2(1-\gamma^2 L_F^2)\right)}_{\in(0,1)}\|F(z^k)\|^2 \tag{3.2}$$

for each $k \in \mathbb{N}_0$.

3.1.1 Variant of FLEX under injectivity

As highlighted in Remark 3.1(i), since $\|F(z^k) - F(\bar{z}^k)\| = 0$ can occur without reaching a solution, special considerations are necessary. In FLEX, this is addressed by employing an explicit finite termination in the line-search procedure and assuming that the directions d^k are summable. However, when the operator F is injective, it is possible to exploit the Lyapunov inequality in (2.10) directly to establish convergence results without additional assumptions. To this end, we introduce I-FLEX, which incorporates a more traditional line-search procedure similar to that used in [38, Algorithm PANOC]. However, the PANOC algorithm is developed for minimization problems and utilizes a fundamentally different Lyapunov function. Importantly, I-FLEX uses an infinite backtracking strategy with guaranteed finite termination since injectivity ensures that $\|F(z^k) - F(\bar{z}^k)\| = 0$ only when a solution has been found. Moreover, I-FLEX has two fewer parameters than FLEX, simplifying its implementation.

Proposition 3.3: *Suppose that Assumption I.(i) holds and F is injective. Then, independent of the choice of the direction d^k in Step 3 of I-FLEX (Algorithm 2), either there exists an iteration $k \in \mathbb{N}_0$ such that $z^k \in \operatorname{zer}(F)$ or the line search in Step 4 is well-defined for each iteration $k \in \mathbb{N}_0$.*

Proof. Follows from $\sigma \in (0,1)$, (2.10), continuity of F, and that $\|F(z^k) - F(\bar{z}^k)\| \ne 0$ if and only if $z^k \notin \operatorname{zer}(F)$. □

Theorem 3.4: *Suppose that Assumption I.(i) holds, $\operatorname{zer}(F) \ne \emptyset$, and the sequence $(z^k)_{k\in\mathbb{N}_0}$ is generated by I-FLEX (Algorithm 2).*

(i) If F is injective and weakly continuous, then each weak sequential cluster point of $(z^k)_{k\in\mathbb{N}_0}$ is in $\operatorname{zer}(F)$*.*

(ii) If F is injective and weakly continuous, and $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0; \mathcal{H})$, then $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to some point in $\operatorname{zer}(F)$*.*

Algorithm 3: Prox-FLEX (Proximal-FLEX)

Initialize: $z^0 \in \mathcal{H}$, $\gamma \in (0,1/L_F)$, $(\rho,\sigma,\beta) \in (0,1)^3$, $M \in \mathbb{N}_0$
Require: Lyapunov function $\mathcal{V}$ as in (2.1) and algorithmic operators T_1^γ, T_2^γ as in (2.2)

```
for k = 0,1,2,... do
    z̄^k = T_1^γ(z^k)                                  = prox_{γg}(z^k − γF(z^k))
    w^k = T_2^γ(z^k)                                  = prox_{γg}(z^k − γF(z̄^k))
    Compute a direction d^k ∈ H at z^k
    if V(z^k + d^k, T_1^γ(z^k + d^k), T_2^γ(z^k + d^k)) ≤ ρ^2 V(z^k, z̄^k, w^k) then
        z^{k+1} = z^k + d^k
    else
        Set z^{k+1} = (1 − τ_k)w^k + τ_k(z^k + d^k) where τ_k is the largest number in
        {β^i | i ∈ [[1, M]]} ∪ {0} such that
```

$$\mathcal{V}(z^{k+1}, T_1^\gamma(z^{k+1}), T_2^\gamma(z^{k+1})) \le \mathcal{V}(z^k, \bar{z}^k, w^k) - \sigma(1-\gamma^2 L_F^2)\gamma^{-2}\|w^k - \bar{z}^k\|^2 \quad (3.5)$$

```
    end if
end for
```

(iii) If F is uniformly monotone, then $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to some point in $\operatorname{zer}(F)$*.*

(iv) If there exists $0 < \mu_F \le L_F$ such that $\mu_F\|x-y\| \le \|F(x)-F(y)\|$ for each $x,y \in \mathcal{H}$, then $(z^k)_{k\in\mathbb{N}_0}$ converges strongly to some point in $\operatorname{zer}(F)$ *and*

$$\|F(z^{k+1})\|^2 \le \underbrace{(1-\sigma\gamma^2\mu_F^2(1-\gamma^2 L_F^2))}_{\in(0,1)}\|F(z^k)\|^2 \quad (3.4)$$

for each $k \in \mathbb{N}_0$.

3.2 Proximal fast line-search extragradient

A direct generalization of FLEX (Algorithm 1) in Section 3.1 is provided in Prox-FLEX (Algorithm 3) for the case when g in (1.1) is nonzero. Here, the Lyapunov inequality (2.3) from Theorem 2.2 is used to modify the standard extragradient method in (1.2); otherwise, the underlying approach remains the same. However, there is one important difference between FLEX and Prox-FLEX in terms of computations required per line-search trial. The condition (3.1) in FLEX requires only one additional F evaluation per trial while condition (3.5) in Prox-FLEX requires two additional F evaluations and two additional $\operatorname{prox}_{\gamma g}$ evaluations per trial. Next, we present a convergence result of Prox-FLEX.

Theorem 3.5: *Suppose that Assumption I holds,* $\operatorname{zer}(F+\partial g) \neq \emptyset$*, the sequence $(z^k)_{k\in\mathbb{N}_0}$ is generated by Prox-FLEX (Algorithm 3), and $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0;\mathcal{H})$. Then $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to some point in* $\operatorname{zer}(F+\partial g)$*.*

Remark 3.6: (i) If F is μ_F-strongly monotone, then the Lyapunov inequality (2.3) in Theorem 2.2 can be strengthened to include the additional term $-2\gamma^{-1}\mu_F\|z^{k+1} - z^k\|^2$ in the right-hand side; this follows from using strong monotonicity of F instead of monotonicity of F in the first inequality in (2.8). This observation suggests that the line-search condition (3.5) in Prox-FLEX can be replaced by

$$\begin{aligned}\mathcal{V}(z^{k+1}, \bar{z}^{k+1}, w^{k+1}) \le \mathcal{V}(z^k, \bar{z}^k, w^k) &- \sigma(1-\gamma^2 L_F^2)\gamma^{-2}\|w^k - \bar{z}^k\|^2 \\ &- 2\gamma^{-1}\sigma\mu_F\|w^k - z^k\|^2.\end{aligned} \quad (3.6)$$

Note that using Young's inequality, we get

$$\mathcal{V}(z^k,\bar{z}^k,w^k) \le (2\gamma^{-1}+\gamma^{-1}L_F+\gamma^{-2})\|w^k-z^k\|^2+(\gamma^{-1}L_F+\gamma^{-2})\|w^k-\bar{z}^k\|^2. \quad (3.7)$$

Combining (3.6), (3.7) and Step 5 in `Prox-FLEX` gives

$$\mathcal{V}(z^{k+1},\bar{z}^{k+1},w^{k+1}) \le \underbrace{\max\left(\rho^2, 1-\frac{\sigma\min\left((1-\gamma^2L_F^2),2\gamma\mu_F\right)}{2\gamma+\gamma L_F+1}\right)}_{\in(0,1)}\mathcal{V}(z^k,\bar{z}^k,w^k)$$

for each $k\in\mathbb{N}_0$, i.e. $(\mathcal{V}(z^k,\bar{z}^k,w^k))_{k\in\mathbb{N}_0}$ converges at least Q-linearly to zero. However, the resulting line-search condition is not always actionable since μ_F may not be known in many practical problems. Therefore, we have chosen not to consider the strongly monotone case further.

(ii) Similar to `I-FLEX`, `Prox-FLEX` can be modified to perform infinite backtracking on (3.5) with guaranteed finite termination, even without the strengthened line-search condition described above in Remark 3.6(i). This modification requires $\|w^k-\bar{z}^k\|$ to be an optimality measure, which holds when both F and $\operatorname{prox}_{\gamma g}$ are injective. However, since $\operatorname{prox}_{\gamma g}$ is rarely injective in practical applications, we omit this modification from our analysis.

4 Global convergence

This section provides detailed proofs of the results presented in Section 3. We start by providing two useful lemmas. The first lemma establishes that the iterates generated by `FLEX`, `I-FLEX`, and `Prox-FLEX` are quasi-Féjer monotone with respect to the solution set, which is an important tool in establishing global convergence. The second lemma contains some auxiliary results.

Lemma 4.1: *Suppose that Assumption I holds, $z^\star\in\operatorname{zer}(F+\partial g)$, $(d^k)_{k\in\mathbb{N}_0}\in\ell^1(\mathbb{N}_0;\mathcal{H})$, T_1^γ and T_2^γ are the algorithmic operators defined in (2.2), and $\mathcal{V}$ is the Lyapunov function given in (2.1). Let $(z^k)_{k\in\mathbb{N}_0}\in\mathcal{H}^{\mathbb{N}_0}$ such that $z^{k+1}=(1-\tau_k)w^k+\tau_k(z^k+d^k)$, where $\tau_k\in[0,1]$ and $w^k=T_2^\gamma(z^k)$ for each $k\in\mathbb{N}_0$. Then there exists a sequence $(\varepsilon_k)_{k\in\mathbb{N}_0}\in\ell^1(\mathbb{N}_0;\mathbb{R}_+)$ such that*

$$\|z^{k+1}-z^\star\|^2 \le \|z^k-z^\star\|^2+\varepsilon_k-(1-\tau_k)\alpha(\gamma,L_F)\mathcal{V}_k, \quad (4.1)$$

where $\alpha(\gamma,L_F)$ is defined in (2.14), $\mathcal{V}_k=\mathcal{V}(z^k,\bar{z}^k,w^k)$, and $\bar{z}^k=T_1^\gamma(z^k)$ for each $k\in\mathbb{N}_0$.

Proof. Note that Proposition 2.1.(i) gives that $\mathcal{V}_k\ge 0$ for each $k\in\mathbb{N}_0$. Using the identity

$$\|\tau x+(1-\tau)y\|^2=\tau\|x\|^2+(1-\tau)\|y\|^2-\tau(1-\tau)\|x-y\|^2$$

for each $x,y\in\mathcal{H}$ and $\tau\in\mathbb{R}$ [6, Corollary 2.15], and $z^{k+1}-z^\star=\tau_k(z^k+d^k-z^\star)+(1-\tau_k)(w^k-z^\star)$ for each $k\in\mathbb{N}_0$, we get that

$$\begin{aligned}
&\|z^{k+1}-z^\star\|^2\\
&\quad=\tau_k\|z^k+d^k-z^\star\|^2+(1-\tau_k)\|w^k-z^\star\|^2-\tau_k(1-\tau_k)\|z^k+d^k-w^k\|^2\\
&\quad\le\tau_k\|z^k+d^k-z^\star\|^2+(1-\tau_k)\|z^k-z^\star\|^2-(1-\tau_k)\alpha(\gamma,L_F)\mathcal{V}_k\\
&\quad\le\tau_k(\|z^k-z^\star\|+\|d^k\|)^2+(1-\tau_k)\|z^k-z^\star\|^2-(1-\tau_k)\alpha(\gamma,L_F)\mathcal{V}_k\\
&\quad\le\|z^k-z^\star\|^2+2\|z^k-z^\star\|\|d^k\|+\|d^k\|^2-(1-\tau_k)\alpha(\gamma,L_F)\mathcal{V}_k \qquad (4.2)
\end{aligned}$$

$$\leq \left(\|z^k - z^\star\| + \|d^k\|\right)^2 \tag{4.3}$$

for each $k \in \mathbb{N}_0$, where (2.13) and $\tau_k(1-\tau_k) \geq 0$ is used in the first inequality, the triangle inequality is used in the second inequality, $\tau_k \leq 1$ is used in the third inequality, and $(1-\tau_k)\alpha(\gamma, L_F) \geq 0$ is used in the last inequality. Taking the square root of (4.3) and inductively applying the resulting inequality gives that

$$\|z^k - z^\star\| \leq \|z^0 - z^\star\| + \textstyle\sum_{i=0}^{k-1}\|d^i\| \leq \underbrace{\|z^0 - z^\star\| + \textstyle\sum_{i=0}^{\infty}\|d^i\|}_{=E} < \infty \tag{4.4}$$

for each $k \in \mathbb{N}_0$, where the empty sum is interpreted as zero and E is finite since $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0; \mathcal{H})$. Therefore, (4.2) and (4.4) imply that

$$\|z^{k+1} - z^\star\|^2 \leq \|z^k - z^\star\|^2 + \underbrace{2E\|d^k\| + \|d^k\|^2}_{=\varepsilon_k} - (1-\tau_k)\alpha(\gamma, L_F)\mathcal{V}_k$$

for each $k \in \mathbb{N}_0$, where summability of $(\varepsilon_k)_{k\in\mathbb{N}_0}$ follows from $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0; \mathcal{H})$. □

Lemma 4.2: *Suppose that Assumption I holds, the sequences $(z^k)_{k\in\mathbb{N}_0}$, $(\bar{z}^k)_{k\in\mathbb{N}_0}$ and $(w^k)_{k\in\mathbb{N}_0}$ are generated by either `FLEX` (Algorithm 1), `I-FLEX` (Algorithm 2), or `Prox-FLEX` (Algorithm 3), and $\mathcal{V}$ is the Lyapunov function given in (2.1). Then, the following hold.*

- *(i) $(\mathcal{V}(z^k, \bar{z}^k, w^k))_{k\in\mathbb{N}_0}$ is convergent, which for `FLEX` and `I-FLEX` reduces to $(F(z^k))_{k\in\mathbb{N}_0}$ being convergent.*
- *(ii) $(\|w^k - \bar{z}^k\|^2)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0; \mathbb{R}_+)$, which for `FLEX` and `I-FLEX` can be written as $(\|F(z^k) - F(\bar{z}^k)\|^2)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0; \mathbb{R}_+)$.*
- *(iii) If F is uniformly monotone, then $(\|F(z^k)\|)_{k\in\mathbb{N}_0}$ converges to zero for `FLEX` and `I-FLEX`.*

Proof. First, we establish for `Prox-FLEX` that

$$\begin{aligned}\mathcal{V}(z^{k+1}, \bar{z}^{k+1}, w^{k+1}) &\leq \mathcal{V}(z^k, \bar{z}^k, w^k) \\ &\quad - \min((1-\gamma L_F)(1-\rho^2), \sigma(1-\gamma^2 L_F^2))\gamma^{-2}\|w^k - \bar{z}^k\|^2\end{aligned} \tag{4.5}$$

for each $k \in \mathbb{N}_0$. Note that Step 5 in `Prox-FLEX` implies that

$$\begin{aligned}\mathcal{V}(z^{k+1}, \bar{z}^{k+1}, w^{k+1}) &\leq \rho^2\mathcal{V}(z^k, \bar{z}^k, w^k) \\ &= \mathcal{V}(z^k, \bar{z}^k, w^k) - (1-\rho^2)\mathcal{V}(z^k, \bar{z}^k, w^k) \\ &\leq \mathcal{V}(z^k, \bar{z}^k, w^k) - (1-\gamma L_F)(1-\rho^2)\gamma^{-2}\|w^k - \bar{z}^k\|^2\end{aligned}$$

for each iteration k when the condition in Step 5 of `Prox-FLEX` is true, where Proposition 2.1.(i) is used in the last inequality. This combined with (3.5) in `Prox-FLEX` gives (4.5).

Second, since `Prox-FLEX` reduced to `FLEX` when $g = 0$, (4.5) implies that

$$\|F(z^{k+1})\|^2 \leq \|F(z^k)\|^2 - \min((1-\gamma L_F)(1-\rho^2), \sigma(1-\gamma^2 L_F^2))\|F(z^k) - F(\bar{z}^k)\|^2 \tag{4.6}$$

for each $k \in \mathbb{N}_0$, for `FLEX`.

4.2.(i): Follows from (4.6) for `FLEX`, (3.3) for `I-FLEX`, and (4.5) for `Prox-FLEX`, combined with the monotone convergence theorem.

4.2.(ii): Note that $(\|w^k - \bar{z}^k\|^2)_{k\in\mathbb{N}_0} = (\gamma^2\|F(z^k) - F(\bar{z}^k)\|^2)_{k\in\mathbb{N}_0}$ for `FLEX` and `I-FLEX`. The statement follows from (4.6) for `FLEX`, (3.3) for `I-FLEX`, and (4.5) for `Prox-FLEX`, combined with a telescoping summation argument.

4.2.(iii): Suppose that F is uniformly monotone, i.e., there exists an increasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, that vanishes only at 0, such that

$$\phi(\|x-y\|) \leq \langle x-y, F(x)-F(y)\rangle$$

for each $x, y \in \mathcal{H}$. Note that

$$\begin{aligned}\phi\left(\gamma\|F(z^k)\|\right) &= \phi\left(\|z^k - \bar{z}^k\|\right) \\ &\leq \langle z^k - \bar{z}^k, F(z^k) - F(\bar{z}^k)\rangle \\ &\leq \|z^k - \bar{z}^k\|\|F(z^k) - F(\bar{z}^k)\| \\ &= \gamma\|F(z^k)\|\|F(z^k) - F(\bar{z}^k)\| \xrightarrow[k\to\infty]{} 0,\end{aligned}$$

where $\bar{z}^k = z^k - \gamma F(z^k)$ is used in the first equality, the Cauchy–Schwarz inequality is used in the second inequality, $\bar{z}^k = z^k - \gamma F(z^k)$ is used in the last equality, and the convergence to zero in the last line follows from Lemma 4.2.(i) and Lemma 4.2.(ii). This proves the claim. □

4.1 Proofs regarding `FLEX`

Proof of Theorem 3.2.(i). This follows from Theorem 3.5, since `Prox-FLEX` (Algorithm 3) reduces to `FLEX` (Algorithm 1) when $g = 0$. □

Proof of Theorem 3.2.(ii). See Lemma 4.2.(iii). □

Proof of Theorem 3.2.(iii). Note that (3.1) gives that

$$\begin{aligned}\|F(z^{k+1})\|^2 \leq& \|F(z^k)\|^2 - \sigma(1-\gamma^2 L_F^2)\|F(z^k) - F(\bar{z}^k)\|^2 \\ \leq& \|F(z^k)\|^2 - \sigma\mu_F^2(1-\gamma^2 L_F^2)\|z^k - \bar{z}^k\| \\ =& (1-\sigma\gamma^2\mu_F^2(1-\gamma^2 L_F^2))\|F(z^k)\|^2 \end{aligned} \tag{4.7}$$

for each iteration k such that the condition in Step 4 in `FLEX` is false, where Step 2 in `FLEX` is used in the last equality. Combining (4.7) and Step 4 in `FLEX` gives (3.2). Moreover, $(\|F(z^k)\|)_{k\in\mathbb{N}_0}$ converges to zero since $\max(\rho^2, 1-\sigma\gamma^2\mu_F^2(1-\gamma^2 L_F^2)) \in (0,1)$. Since

$$\|z^k - z^\star\| \leq \mu_F^{-1}\|F(z^k) - F(z^\star)\| = \mu_F^{-1}\|F(z^k)\|$$

for each $k \in \mathbb{N}_0$, $(z^k)_{k\in\mathbb{N}_0}$ converges strongly to $z^\star \in \operatorname{zer}(F)$. □

4.2 Proofs regarding `I-FLEX`

Proof of Theorem 3.4.(i). Suppose that $(z^k)_{k\in K} \rightharpoonup z^\infty$. Weak continuity of F and $\bar{z}^k = z^k - \gamma F(z^k)$ give that $(\bar{z}^k)_{k\in K} \rightharpoonup \bar{z}^\infty = z^\infty - \gamma F(z^\infty)$. On the other hand, it follows from Lemma 4.2.(ii) and weak continuity of F that $F(z^\infty) = F(\bar{z}^\infty)$, which in view of the injectivity assumption of F, implies $z^\infty = \bar{z}^\infty = z^\infty - \gamma F(z^\infty)$. Therefore, $z^\infty \in \operatorname{zer}(F)$, as claimed. □

Proof of Theorem 3.4.(ii). Note that Theorem 3.4.(i) gives that each weak sequential cluster point of $(z^k)_{k\in\mathbb{N}_0}$ is in $\operatorname{zer}(F)$. Moreover, [6, Lemma 5.31] and (4.1) in Lemma 4.1 give that $(\|z^k - z^\star\|)_{k\in\mathbb{N}_0}$ convergence. Thus, [6, Lemma 2.47] gives that $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to some point in $\operatorname{zer}(F)$, as claimed. □

Proof of Theorem 3.4.(iii). See Lemma 4.2.(iii). □

Proof of Theorem 3.4.(iv). Note that (3.3) gives that

$$\begin{aligned}\|F(z^{k+1})\|^2 \leq& \|F(z^k)\|^2 - \sigma(1-\gamma^2 L_F^2)\|F(z^k) - F(\bar{z}^k)\|^2 \\ \leq& \|F(z^k)\|^2 - \sigma\mu_F^2(1-\gamma^2 L_F^2)\|z^k - \bar{z}^k\| \\ =& (1-\sigma\gamma^2\mu_F^2(1-\gamma^2 L_F^2))\|F(z^k)\|^2\end{aligned}$$

for each $k \in \mathbb{N}_0$, where Step 2 in I-FLEX is used in the last equality. Therefore, $(\|F(z^k)\|)_{k\in\mathbb{N}_0}$ converges to zero since $1-\sigma\gamma^2\mu_F^2(1-\gamma^2 L_F^2) \in (0,1)$. Since

$$\|z^k - z^\star\| \leq \mu_F^{-1}\|F(z^k) - F(z^\star)\| = \mu_F^{-1}\|F(z^k)\|$$

for each $k \in \mathbb{N}_0$, $(z^k)_{k\in\mathbb{N}_0}$ converges strongly to $z^\star \in \operatorname{zer}(F)$. □

4.3 Proofs regarding Prox-FLEX

Proof of Theorem 3.5. Set $\tau_k = 1$ for the iterations when the condition in Step 5 in Prox-FLEX is true and let $z^\star \in \operatorname{zer}(F+\partial g)$. Then (4.1) in Lemma 4.1 and [6, Lemma 5.31] imply that $(\|z^k - z^\star\|)_{k\in\mathbb{N}_0}$ converges. Thus, the proof is complete if we can show that weak sequential cluster points of $(z^k)_{k\in\mathbb{N}_0}$ belong to $\operatorname{zer}(F+\partial g)$, due to [6, Lemma 2.47].

For this, it suffices to show that $(\|\bar{z}^k - z^k\|)_{k\in\mathbb{N}_0}$ converges to zero. Indeed, suppose that $(z^k)_{k\in K} \rightharpoonup z^\infty$ for some $z^\infty \in \mathcal{H}$ and $(\|\bar{z}^k - z^k\|)_{k\in\mathbb{N}_0}$ converges to zero. Then $(\bar{z}^k)_{k\in K} \rightharpoonup z^\infty$. Moreover, the proximal evaluation in Step 2 in Prox-FLEX can equivalently be written as

$$\gamma^{-1}(z^k - \bar{z}^k) - F(z^k) + F(\bar{z}^k) \in (F+\partial g)(\bar{z}^k). \tag{4.8}$$

The left-hand side of (4.8) converges strongly to zero since F is continuous and $(\|z^k - \bar{z}^k\|)_{k\in\mathbb{N}_0}$ converges to zero. Moreover, the operator $F+\partial g$ is maximally monotone, since F is maximally monotone (by continuity and monotonicity [6, Corollary 20.28]), ∂g is maximally monotone [6, Theorem 20.48], and F has full domain [6, Corollary 25.5]. Thus, [6, Proposition 20.38] gives that $z^\infty \in \operatorname{zer}(F+\partial g)$, and by [6, Lemma 2.47] we conclude that $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to a point in $\operatorname{zer}(F+\partial g)$.

It remains to show that $(\|\bar{z}^k - z^k\|)_{k\in\mathbb{N}_0}$ converges to zero, which we do by showing that $(\mathcal{V}_k)_{k\in\mathbb{N}_0}$ converges to zero and applying Proposition 2.1.(i). Let $K_{<1} = \{k \in \mathbb{N}_0 \mid \tau_k < 1\}$. Suppose that $|K_{<1}| < \infty$. Then $\mathcal{V}_{k+1} \leq \rho^2\mathcal{V}_k$ for each $k \in \mathbb{N}_0$ such that $k > \max K_{<1}$, and $(\mathcal{V}_k)_{k\in\mathbb{N}_0}$ converges to zero since $\rho \in (0,1)$. On the contrary, suppose that $|K_{<1}| = \infty$. Let $\Gamma : K_{<1} \to K_{<1}$ such that $\Gamma(k) = \min\{i \in K_{<1} \mid k < i\}$ for each $k \in K_{<1}$. Let $k \in K_{<1}$, and notice that $\tau_k \leq \bar{\tau}$ for any such index, where $\bar{\tau} = \max\{\beta^i \mid i \in [\![1,M]\!]\} \cup \{0\} < 1$. Inductively summing (4.1) in Lemma 4.1 from k to $\Gamma(k)-1$ gives

$$\|z^{\Gamma(k)} - z^\star\|^2 \leq \|z^k - z^\star\|^2 - (1-\bar{\tau})\alpha(\gamma, L_F)\mathcal{V}_k + \sum_{i=k}^{\Gamma(k)-1} \varepsilon_i, \tag{4.9}$$

where we used the fact that $\tau_i = 1$ for any $i \in K_1$. Inductively summing over all $k \in K_{<1}$ in (4.9), rearranging, and dividing by $(1-\bar{\tau})\alpha(\gamma, L_F) > 0$ gives

$$\begin{aligned}\sum_{k\in K_{<1}} \mathcal{V}_k \leq& \frac{\sum_{k\in K_{<1}}(\|z^k - z^\star\|^2 - \|z^{\Gamma(k)} - z^\star\|^2 + \sum_{i=k}^{\Gamma(k)-1}\varepsilon_i)}{(1-\bar{\tau})\alpha(\gamma, L_F)} \\ \leq& \frac{\|z^{\min(K_{<1})} - z^\star\|^2 + \sum_{k=0}^{\infty}\varepsilon_k}{(1-\bar{\tau})\alpha(\gamma, L_F)} < \infty,\end{aligned} \tag{4.10}$$

where summability of $(\varepsilon_k)_{k\in\mathbb{N}_0}$ is used in the last inequality. Note that

$$\begin{aligned}\sum_{k=0}^{\infty}\mathcal{V}_k &= \sum_{k\in K_{<1}}\sum_{i=k}^{\Gamma(k)-1}\mathcal{V}_i\\ &\le \sum_{k\in K_{<1}}\sum_{i=k}^{\Gamma(k)-1}\rho^{2(i-k)}\mathcal{V}_k\\ &\le \frac{1}{1-\rho^2}\sum_{k\in K_{<1}}\mathcal{V}_k<\infty,\end{aligned}$$

where Step 5 in Prox-FLEX is used in the first inequality, the expression for the geometric series is used in the second inequality, and (4.10) is used in the last inequality. This completes the proof. □

5 Superlinear convergence

The convergence analyses presented so far have been blind to the choice of directions $(d^k)_{k\in\mathbb{N}_0}$; nevertheless, attaining a fast convergence rate relies on their precise choice. This section presents a minimal set of assumptions on the directions that ensure superlinear convergence. Our main focus will be on quasi-Newton-type directions that are computed as

$$d^k = -H_k R_\gamma(z^k), \quad \text{where} \quad R_\gamma = \tfrac{1}{\gamma}(\mathrm{Id} - \mathrm{prox}_{\gamma g}\circ(\mathrm{Id}-\gamma F)), \tag{5.1}$$

$\gamma\in\mathbb{R}_{++}$, $H_k:\mathcal{H}\to\mathcal{H}$ is a linear operator encapsulating information of the geometry of the residual mapping R_γ at z^k, and F and g satisfy Assumption I. The specific way H_k is computed determines the underlying quasi-Newton method (see Section 6 for details). Notice that the zeros of R_γ coincide with the set of solutions of (1.1). Moreover, when $g=0$, R_γ reduces to F, and the directions are given by $d^k=-H_kF(z^k)$. The following assumption on the directions $(d^k)_{k\in\mathbb{N}_0}$ can be seen as a boundedness assumption on the linear operators $(H_k)_{k\in\mathbb{N}_0}$. However, note that the assumption applies to directions beyond the ones given in (5.1).

Assumption II: *The sequence of directions $(d^k)_{k\in\mathbb{N}_0}$ used in FLEX, I-FLEX, or Prox-FLEX satisfies $\|d^k\|\le D\|R_\gamma(z^k)\|$ for each $k\in\mathbb{N}_0$ such that $k\ge K$, for some constants $D\ge 0$ and $K\in\mathbb{N}_0$, where R_γ denotes the residual operator as defined in (5.1) (the function g is set to zero in the particular cases of FLEX and I-FLEX).*

Assumption II is a natural assumption for directions defined in (5.1). For example, under suitable regularity conditions for regularized Newton directions—specifically when $g=0$—we demonstrate this in Proposition 5.1. Note that in Proposition 5.1, we assume that F is continuously Fréchet differentiable; however, this assumption is made solely for illustrative purposes and is not required elsewhere in the paper. Assumption II has also been utilized in the context of minimization and in finding zeros of nonexpansive maps, as seen in [1, Theorem 5.7.A3 and Theorem 5.8.A3] and [39, Assumption 2], respectively.

Proposition 5.1: *Let $F:\mathcal{H}\to\mathcal{H}$ be monotone and continuously Fréchet differentiable, and suppose that the Fréchet derivative $\mathsf{D}F$ at $z^\star\in\operatorname{zer}F$ is left invertible. Suppose that $(z^k)_{k\in\mathbb{N}_0},(d^k)_{k\in\mathbb{N}_0}\in\mathcal{H}^{\mathbb{N}_0}$ are such that*

$$r_k\ \mathrm{Id}+\mathsf{D}F(z^k))d^k=-F(z^k) \tag{5.2}$$

for some sequence $(r_k)_{k\in\mathbb{N}_0} \in \mathbb{R}_{++}^{\mathbb{N}_0}$*, and that* $(z^k)_{k\in\mathbb{N}_0}$ *converges strongly to* $z^\star$*. Then, Assumption II is satisfied with* $g = 0$*.*

Proof. Let $t > 0$ and note that

$$0 \leq \frac{\langle F(z+tv) - F(z), z+tv-z\rangle}{t^2} \xrightarrow[t\downarrow 0]{} \langle \mathsf{D}F(z)v, v\rangle \tag{5.3}$$

for any $v \in \mathcal{H}$ and for any $z \in \mathcal{H}$, by monotonicity of F. This implies that the bounded linear operator $r_k\ \mathrm{Id} + \mathsf{D}F(z^k)$ is r_k-strongly monotone, and therefore invertible for each $k \in \mathbb{N}_0$. This, in turn, ensures that the regularized Newton update (5.2) is well-defined, i.e., d^k is uniquely defined at each iteration.

Moreover, since $\mathsf{D}F(z^\star)$ is left invertible, there exists $c_1 > 0$ such that $\|\mathsf{D}F(z^\star)v\| \geq c_1\|v\|$ for any $v \in \mathcal{H}$ [5, Proposition 10.29]. This observation combined with $z^k \to z^\star$ and continuity of $\mathsf{D}F(\cdot)$ implies that there exists $c_2 > 0$ and $K \in \mathbb{N}_0$ such that $\|\mathsf{D}F(z^k)v\| \geq c_2\|v\|$ for any $v \in \mathcal{H}$ and for any $k \geq K$. Therefore,

$$\begin{aligned}
\|F(z^k)\|^2 &= \|r_k\ \mathrm{Id} + \mathsf{D}F(z^k)d^k\|^2 \\
&= \|r_k d^k\|^2 + 2r_k\langle \mathsf{D}F(z^k)d^k, d^k\rangle + \|\mathsf{D}F(z^k)d^k\|^2 \\
&\geq \|\mathsf{D}F(z^k)d^k\|^2 \\
&\geq c_2^2\|d^k\|^2
\end{aligned}$$

for each $k \geq K$, where the first inequality follows from (5.3) and $r_k > 0$. This establishes that Assumption II is satisfied with $D = 1/c_2$, when $g = 0$. □

Remark 5.2: In the case of `FLEX`, when the operator is strongly monotone, the sequence $(\|F(z^k)\|)_{k\in\mathbb{N}_0}$ converges Q-linearly to zero, as established in Theorem 3.2.(iii). This observation, combined with Assumption II is sufficient to conclude that $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0;\mathcal{H})$, thereby yielding global convergence as demonstrated in Theorem 3.2.(i). An analogous argument extends to `Prox-FLEX` after incorporating the strengthening discussed in Remark 3.6(i).

We proceed to quantify the quality of the directions used in the algorithms that guarantee fast convergence. The classical condition of [16, Chapter 7.5] for Newton-type methods identifies a sequence of directions $(d^k)_{k\in\mathbb{N}_0}$ relative to a sequence $(z^k)_{k\in\mathbb{N}_0}$ converging to $z^\star$ as superlinear if

$$\lim_{k\to\infty} \frac{\|z^k + d^k - z^\star\|}{\|z^k - z^\star\|} = 0. \tag{5.4}$$

This notion a priori assumes the convergence of the sequence $(z^k)_{k\in\mathbb{N}_0}$. Here, we use a slightly refined notion and define superlinear directions similar to [39, Definition VI.2].

Definition 5.3: *Suppose that* $\gamma \in \mathbb{R}_{++}$*,* $(z^k)_{k\in\mathbb{N}_0}, (d^k)_{k\in\mathbb{N}_0} \in \mathcal{H}^{\mathbb{N}_0}$*, Assumption I holds,* T_1^γ *and* T_2^γ *are the algorithmic operators defined in (2.2), and* $\mathcal{V}$ *is the Lyapunov function given in (2.1). Then we say that the sequence of directions* $(d^k)_{k\in\mathbb{N}_0}$ *is* superlinear relative to $(z^k)_{k\in\mathbb{N}_0}$ *if*

$$\lim_{k\to\infty} \frac{\mathcal{V}(z^k + d^k, T_1^\gamma(z^k + d^k), T_2^\gamma(z^k + d^k))}{\mathcal{V}(z^k, T_1^\gamma(z^k), T_2^\gamma(z^k))} = 0. \tag{5.5}$$

In the case of solving monotone equations where $g = 0$, addressed by `FLEX` and `I-FLEX`, (5.5) reduces to

$$\lim_{k\to\infty} \frac{\|F(z^k + d^k)\|}{\|F(z^k)\|} = 0. \tag{5.6}$$

As shown below in Theorem 5.5.(iii), this assumption in conjunction with Assumption II is sufficient to conclude $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0;\mathcal{H})$, establishing global sequential convergence by Theorem 3.5. See also Theorems 3.2.(i) and 3.4.(ii) for `FLEX` and `I-FLEX`, respectively.

Remark 5.4: The superlinear convergence results presented in Theorem 5.5 also hold under (5.4) of [16] since it implies the notion in Definition 5.3. Indeed, by Assumption II

$$\begin{aligned}\|d^k\|^2 &\le D^2\left(\|z^k - T_2^\gamma(z^k)\| + \|T_2^\gamma(z^k) - T_1^\gamma(z^k)\|\right)^2\\ &\le 2D^2\left(\|z^k - T_2^\gamma(z^k)\|^2 + \|T_2^\gamma(z^k) - T_1^\gamma(z^k)\|^2\right)\\ &\le \frac{2\gamma^2 D^2}{1-\gamma L_F}\mathcal{V}(z^k, T_1^\gamma(z^k), T_2^\gamma(z^k))\end{aligned}$$

for each $k \in \mathbb{N}_0$ such that $k \ge K$, where the triangle inequality is used in the first inequality and Proposition 2.1.(i) is used in the last inequality. Hence

$$\frac{\mathcal{V}(z^k+d^k, T_1^\gamma(z^k+d^k), T_2^\gamma(z^k+d^k))}{\mathcal{V}(z^k, T_1^\gamma(z^k), T_2^\gamma(z^k))} \le \frac{2\gamma^2 D^2}{(1-\gamma L_F)\alpha(\gamma, L_F)}\frac{\|z^k+d^k-z^\star\|^2}{\|d^k\|}, \tag{5.7}$$

for each $k \in \mathbb{N}_0$ such that $k \ge K$, where $\mathcal{V}(z^k+d^k, T_1^\gamma(z^k+d^k), T_2^\gamma(z^k+d^k)) \le \|z^k+d^k - z^\star\|^2/\alpha(\gamma, L_F)$ is used (see Theorem 2.4). Combining (5.7) with (5.4) and the fact that $\lim_{k\to\infty}\|z^k - z^\star\|/\|d^k\| = 1$ (see [16, Lemmma 7.5.7]) shows that the ratio on the left-hand-side of (5.7) vanishes. Therefore, (5.5) is a weaker condition than (5.4) under Assumption II. We also refer the reader to [16] for further details. See also [39, Theorem VI.7] for connection to other well-known conditions such as the Dennis-Moré assumption [12].

Theorem 5.5: *Suppose that Assumption I and Assumption II hold,* $\operatorname{zer}(F+\partial g) \neq \emptyset$, T_1^γ *and* T_2^γ *are the algorithmic operators defined in* (2.2), $\mathcal{V}$ *is the Lyapunov function given in* (2.1), *and* $(d^k)_{k\in\mathbb{N}_0}$ *is superlinear relative to the sequence* $(z^k)_{k\in\mathbb{N}_0}$ *generated by either* `FLEX` *(Algorithm 1),* `I-FLEX` *(Algorithm 2), or* `Prox-FLEX` *(Algorithm 3). Then, the following hold.*

(i) $z^{k+1} = z^k + d^k$ *for all* $k \in \mathbb{N}_0$ *sufficiently large.*

(ii) $(\mathcal{V}(z^k, T_1^\gamma(z^k), T_2^\gamma(z^k)))_{k\in\mathbb{N}_0}$ *converges to zero at least Q-superlinearly.*

(iii) $(d^k)_{k\in\mathbb{N}_0} \in \ell^1(\mathbb{N}_0;\mathcal{H})$ *with* $(d^k)_{k\in\mathbb{N}_0}$ *converging to zero at least R-superlinearly.*

(iv) If $\dim\mathcal{H} < \infty$, *then* $(z^k)_{k\in\mathbb{N}_0}$ *converges to some point* $z^\star \in \operatorname{zer}(F+\partial g)$ *at least R-superlinearly.*

Proof. The proof is presented for `Prox-FLEX`, with the necessary adjustments for `FLEX` and `I-FLEX` outlined at the end of the proof.

5.5.(i) Follows from (5.5) since Step 5 in `Prox-FLEX` is true for all $k \in \mathbb{N}_0$ sufficiently large.

5.5.(ii) Follows from 5.5.(i) and (5.5).

5.5.(iii) Note that Assumption II and Proposition 2.1.(i) give that

$$\|d^k\| \le \frac{D}{\gamma}\|z^k - T_1^\gamma(z^k)\| \le \frac{2D}{\sqrt{1-\gamma L_F}}\sqrt{\mathcal{V}(z^k, T_1^\gamma(z^k), T_2^\gamma(z^k))}$$

for each $k \in \mathbb{N}_0$ such that $k \ge K$. The claim now follows from 5.5.(ii).

5.5.(iv) Theorem 3.5 and 5.5.(iii) imply that the sequence $(z^k)_{k\in\mathbb{N}_0}$ converges to some point $z^\star \in \operatorname{zer}(F+\partial g)$. Since $z^{k+1}-z^k = d^k$ for all $k\in\mathbb{N}_0$ sufficiently large, 5.5.(iii) implies that $(z^{k+1}-z^k)_{k\in\mathbb{N}_0}$ converges to zero at least R-(super)linearly. In particular, there exists $\kappa\in\mathbb{R}_{++}$ and $(c_k)_{k\in\mathbb{N}_0}\in\mathbb{R}_{++}^{\mathbb{N}_0}$ such that $\lim_{k\to\infty} c_k = 0$ and

$$\|z^{k+1}-z^k\| \leq \kappa \prod_{i=1}^{k} c_i \tag{5.8}$$

for each $k\in\mathbb{N}_0$. Let $k,j\in\mathbb{N}_0$ such that $j>k$. The triangle inequality and (5.8) give that

$$\|z^k - z^\star\| \xleftarrow[j\to\infty]{} \|z^k - z^j\| \leq \sum_{\ell=k}^{j-1}\|z^\ell - z^{\ell+1}\| \leq \kappa\sum_{\ell=k}^{j-1}\prod_{i=1}^{\ell} c_i \xrightarrow[j\to\infty]{} \kappa\sum_{\ell=k}^{\infty}\prod_{i=1}^{\ell} c_i = \mu_k.$$

The sequence $(\mu_k)_{k\in\mathbb{N}_0}\in\mathbb{R}_{++}^{\mathbb{N}_0}$ converges to zero at least Q-superlinearly since

$$\frac{\mu_k}{\mu_{k-1}} = \frac{\sum_{\ell=k}^{\infty}\prod_{i=1}^{\ell} c_i}{\sum_{\ell=k-1}^{\infty}\prod_{i=1}^{\ell} c_i} = \frac{\left(\prod_{i=1}^{k-1} c_i\right)\left(\sum_{\ell=k}^{\infty}\prod_{i=k}^{\ell} c_i\right)}{\left(\prod_{i=1}^{k-1} c_i\right)\left(1+\sum_{\ell=k}^{\infty}\prod_{i=k}^{\ell} c_i\right)} = \frac{\left(\sum_{\ell=k}^{\infty}\prod_{i=k}^{\ell} c_i\right)}{\left(1+\sum_{\ell=k}^{\infty}\prod_{i=k}^{\ell} c_i\right)} \to 0$$

and $\lim_{k\to\infty}\sum_{\ell=k}^{\infty}\prod_{i=k}^{\ell} c_i = 0$. Thus, $(z^k)_{k\in\mathbb{N}_0}$ converges to $z^\star$ at least R-superlinearly, as claimed.

The assertions for `FLEX` follow directly, as setting $g=0$ reduces `Prox-FLEX` and its underlying assumptions to those of `FLEX`. For `I-FLEX`, the only distinction is that, for sufficiently large k, $\tau_k = 1$ is always accepted in Step 4 of `I-FLEX`, due to (5.6). All other arguments remain unchanged. □

6 Numerical experiments

In this section, we assess the performance of the proposed algorithms in Section 3 through a series of simulations on standard problems using both synthetic and real-world datasets. Code to replicate the experiments is made available online.[1] Table 1 contains a description of the algorithms used.

In the numerical experiments for `FLEX`, `I-FLEX`, and `Prox-FLEX`, we use directions $(d^k)_{k\in\mathbb{N}_0}$ based on quasi-Newton directions.

Anderson acceleration. The first set of quasi-Newton directions we use are the standard limited-memory type-I and type-II Anderson acceleration methods [3, 17]. These directions are computed via (5.1), i.e., $d^k = -H_k R_\gamma(z^k)$, where H_k differs between the type-I and type-II variants. Both methods employ a memory parameter $m\in\mathbb{N}$ and define $m_k = \min\{m,k\}$. They also maintain two buffer matrices:

$$Y_k = \begin{bmatrix} y^{k-m_k} & \cdots & y^{k-1}\end{bmatrix} \quad \text{and} \quad S_k = \begin{bmatrix} s^{k-m_k} & \cdots & s^{k-1}\end{bmatrix},$$

where $y^i = R_\gamma(z^{i+1}) - R_\gamma(z^i)$ and $s^i = z^{i+1}-z^i$. For type-I Anderson acceleration (denoted `AA-I`), we have

$$H_k = I + (S_k - Y_k)(S_k^\top Y_k)^{-1} S_k^\top,$$

whereas for type-II Anderson acceleration (denoted `AA-II`), we have

$$H_k = I + (S_k - Y_k)(Y_k^\top Y_k)^{-1} Y_k^\top.$$

Additional discussion can be found in [45].

[1] https://github.com/manuupadhyaya/flex

Table 1: Algorithms used in the numerical simulations (when applicable).

Method	Description
EG	Extragradient method (1.2) with $\gamma = 0.9/L_F$.
EAG-C	Extra anchored gradient with constant step size $\alpha = 1/(8L_F)$ [44, Section 2.1].
GRAAL	Golden ratio algorithm with $\phi = 2$ and $\alpha = 0.999/L_F$ [2, Algorithm 2], [23].
aGRAAL	Adaptive golden ratio algorithm with $\phi = (2+\sqrt{5})/2$, $\gamma = 1/\phi + 1/\phi^2$ and $\alpha_0 = 0.1$ [2, Algorithm 1], [23].
EG-AA	An extragradient-type method with type-II Anderson acceleration with memory $m = 1$ [34, Algorithm 1] using the parameter values described in [34, Section 4].
FISTA	Fast iterative shrinkage-thresholding algorithm with constant stepsize [7, Section 4].
FLEX	Algorithm 1 with $\gamma = 0.9/L_F$, $\beta = 0.3$, $\sigma = 0.1$, $\rho = 0.99$, and $M = 2$.
I-FLEX	Algorithm 2 with $\beta = 0.01$ and $\sigma = 0.1$.
Prox-FLEX	Algorithm 3 with $\gamma = 0.9/L_F$, $\beta = 0.3$, $\sigma = 0.1$, $\rho = 0.99$, and $M = 2$.

J-symmetric directions. We also incorporate directions derived from the J-symmetric quasi-Newton approach proposed in [4], which is developed for unconstrained minimax problems. This method exploits the so-called J-symmetric structure of the Hessian in such problems, allowing a rank-2 update of the (inverse) Hessian estimate that naturally generalizes the classic Powell's symmetric Broyden method from standard minimization to minimax optimization. The formula for updating H_k in (5.1) can be found in [4, Proposition 2.2]. We refer to this method as J-sym.

6.1 Quadratic minimax problem

Consider the quadratic convex-concave minimax problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \underset{y\in\mathbb{R}^n}{\text{maximize}}\ \mathcal{L}(x,y) \tag{6.1}$$

for the saddle function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathcal{L}(x,y) = \tfrac{1}{2}(x-x^\star)^\top A(x-x^\star) + (x-x^\star)^\top C(y-y^\star) - \tfrac{1}{2}(y-y^\star)^\top B(y-y^\star)$$

for each $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$, where $x^\star, y^\star \in \mathbb{R}^n$, $A, B \in \mathbb{S}^n_+$, and $C \in \mathbb{R}^{n\times n}$. A solution to the minimax problem (6.1) can be obtained by solving an associated saddle point problem, which in turn can be equivalently be written as (1.1) by letting $\mathcal{H} = \mathbb{R}^{2n}$ with the inner product set to the dot product, $g = 0$, and $F : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ as the monotone and L_F-Lipschitz continuous operator given by

$$F(z) = \begin{bmatrix} \nabla_x \mathcal{L}(x,y) \\ -\nabla_y \mathcal{L}(x,y) \end{bmatrix} = \begin{bmatrix} A(x-x^\star) + C(y-y^\star) \\ B(y-y^\star) - C^\top(x-x^\star) \end{bmatrix}$$

for each $z = (x,y) \in \mathbb{R}^n \times \mathbb{R}^n$, where[2]

$$L_F = \left\| \begin{bmatrix} A & C \\ -C^\top & B \end{bmatrix} \right\|.$$

[2]The matrix norm is taken as the spectral norm.

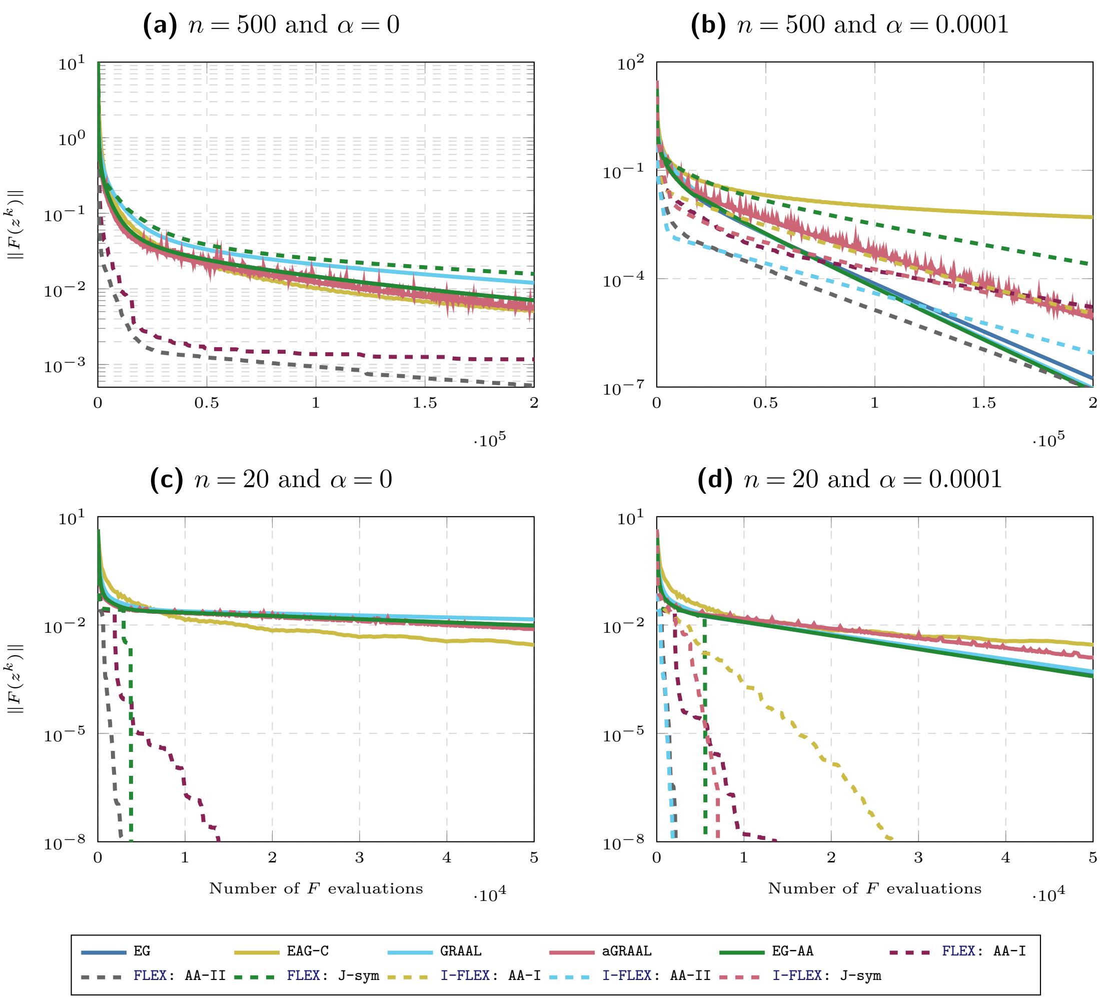

Figure 1: Convergence of algorithms on the quadratic minimax problem (6.1). Both `AA-I` and `AA-II` uses memory parameter $m = 20$.

We generate problem data as in [4, Section 5.1], which is outlined below. The results of the numerical experiments are presented in Figure 1. We see that `FLEX` and `I-FLEX` do very well for small problems, while larger ones are more challenging. Nevertheless, the use of `AA-II` directions in our algorithms systematically performs at the top.

Input: $\alpha \in \mathbb{R}_+$ and $n \in \mathbb{N}$
Output: $x^\star, y^\star \in \mathbb{R}^n$, $A, B \in \mathbb{S}^n_+$ and $C \in \mathbb{R}^{n\times n}$

Let $x^\star, y^\star \in \mathbb{R}^n$ such that $x_i^\star, y_i^\star \sim \mathcal{N}(0,1)$ for each $i \in [\![1,n]\!]$
Let $S \in \mathbb{R}^{n\times n}$ such that $[S]_{i,j} \sim \mathcal{N}(0, 1/\sqrt{n})$ for each $i,j \in [\![1,n]\!]$
$S \leftarrow (S + S^\top)/2$
$S \leftarrow S + (|\lambda_{\min}(S)| + 1)I$
$A \leftarrow \alpha S$
Repeat steps 2-4 with a different random seed and let $B \leftarrow \alpha S$
Repeat step 2 with a different random seed and let $C \leftarrow S$

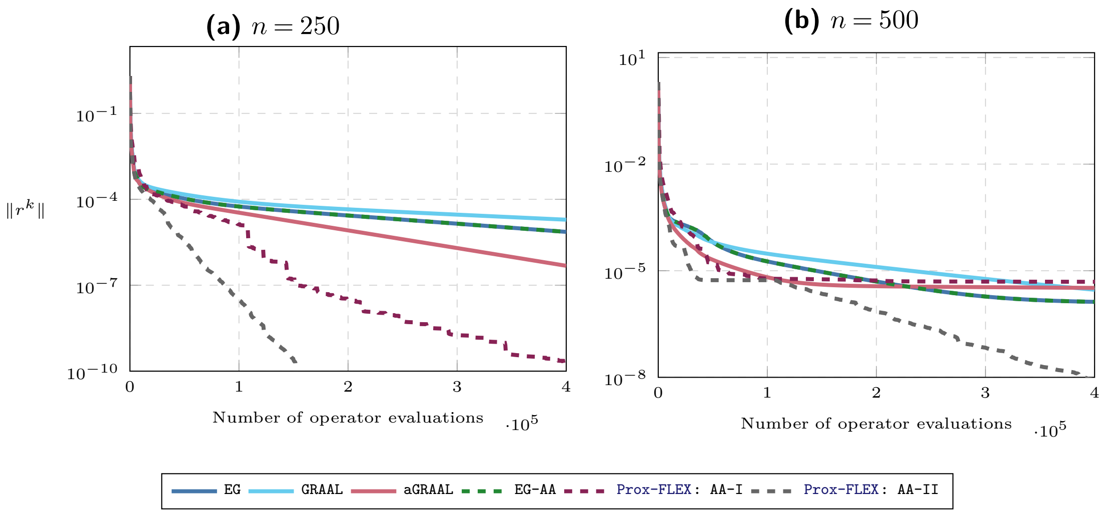

Figure 2: Convergence of algorithms on the bilinear zero-sum game with simplex constraints (6.2) where $r^k = R_{1/2L_F}(z^k)$ and R is the residual mapping in (5.1). Both `AA-I` and `AA-II` uses memory parameter $m = 10$ for Figure 2a and $m = 20$ for Figure 2b. The number of operator evaluations equals the number of F and $\operatorname{prox}_{\gamma g}$ evaluations.

6.2 Bilinear zero-sum game with simplex constraints

Consider the bilinear zero-sum game with simplex constraints given by

$$\underset{x\in\Delta^n}{\text{minimize}}\ \underset{y\in\Delta^n}{\text{maximize}}\ x^\top Ay \tag{6.2}$$

where $A \in \mathbb{R}^{n\times n}$ is the payoff matrix and $\Delta^n = \left\{w \in \mathbb{R}^n_+ \mid w^\top \mathbf{1} = 1\right\}$ is the probability simplex in $\mathbb{R}^n$, which is equivalent to finding a saddle point $(x^\star, y^\star) \in \Delta^n \times \Delta^n$ (which is guaranteed to exist), i.e.,

$$(x^\star)^\top Ay \leq (x^\star)^\top Ay^\star \leq x^\top Ay^\star$$

for each $(x,y) \in \Delta^n \times \Delta^n$. This, in turn, is equivalent to solving (1.1) by letting $\mathcal{H} = \mathbb{R}^{2n}$ with the inner product set to the dot product, $g = \delta_{\Delta^n\times\Delta^n}$, and $F : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ as the monotone and L_F-Lipschitz continuous operator given by

$$F(z) = \begin{bmatrix} Ay \\ -A^\top x \end{bmatrix}$$

for each $z = (x,y) \in \mathbb{R}^n \times \mathbb{R}^n$, where

$$L_F = \left\| \begin{bmatrix} 0 & A \\ -A^\top & 0 \end{bmatrix} \right\|.$$

We generate $A = S - S^\top$ for a random matrix $S \in \mathbb{R}^{n\times n}$ such that $[S]_{i,j} \sim \mathcal{N}(0,1)$ for each $i,j \in [\![1,n]\!]$, resulting in a skew-symmetric matrix A. The results of the numerical experiments are presented in Figure 2. We see that using `AA-II` directions in `Prox-FLEX` gives good performance.

6.3 Cournot–Nash equilibrium problem

Consider a noncooperative game with $n \in \mathbb{N}$ players, in which each player $i \in [\![1,n]\!]$ has to pick a strategy z_i that lies in $\mathcal{Z}_i$, a subset of a real Hilbert space $\mathcal{H}_i$, and has an associated loss function $\varphi_i : \mathcal{H} \to \mathbb{R}$, where $\mathcal{H} = \prod_{j=1}^n \mathcal{H}_j$. In this case, a pure strategy Nash equilibrium is a strategy profile $z = (z_1, \ldots, z_n) \in \mathcal{H}$ that solves the problem

$$\text{find } z \in \mathcal{H} \text{ such that } z_i \in \underset{x \in \mathcal{Z}_i}{\operatorname{Argmin}} \, \varphi_i(x; z_{\smallsetminus i}) \text{ for each } i \in [\![1, n]\!], \tag{6.3}$$

where we have used the notation $(x; z_{\smallsetminus i}) = (z_1, \ldots, z_{i-1}, x, z_{i+1}, \ldots, z_n)$ for each $x \in \mathcal{H}_i$ and $i \in [\![1, n]\!]$. In particular, assume that, for each $i \in [\![1, n]\!]$, the function $\varphi_i(\cdot\,; z_{\smallsetminus i}) : \mathcal{H}_i \to \mathbb{R}$ is convex for each $z \in \mathcal{H}$, the gradient $\nabla_{z_i} \varphi_i : \mathcal{H} \to \mathcal{H}_i$ exists and is Lipschitz continuous, and the set $\mathcal{Z}_i \subseteq \mathcal{H}_i$ is nonempty, closed and convex. Then (6.3) can equivalently be written as (1.1) by letting $F : \mathcal{H} \to \mathcal{H} : z \mapsto (\nabla_{z_i} \varphi_i(z))_{i=1}^n$ and $g = \delta_{\mathcal{Z}}$, where $\mathcal{Z} = \prod_{i=1}^n \mathcal{Z}_i$, and it is straightforward to verify that Assumption I holds.

Let us further specialize the model to the Cournot–Nash equilibrium problem for oligopolistic markets with concave-quadratic cost functions and a differentiated commodity, as presented in [8]. Such models are useful for policymakers and economists in analyzing market outcomes, assessing welfare effects, and evaluating the impact of various market interventions [9, 18, 26, 27, 30, 43]. In particular, in the model of [8], each producer $i \in [\![1, n]\!]$ chooses to produce and supply a quantity $z_i \in [0, T_i]$ of a differentiated commodity at a cost $c_i : \mathbb{R} \to \mathbb{R}$ such that

$$c_i(z_i) = a_i z_i^2 + b_i z_i,$$

for each $z_i \in \mathbb{R}$, where $T_i > 0$ denotes the maximum capacity of production, and $a_i < 0$ and $b_i > 0$ are numbers such that $b_i \geq -2T_i a_i$, ensuring that c_i is increasing on $[0, T_i]$. Moreover, each producer $i \in [\![1, n]\!]$ has a price per produced unit of the differentiated commodity[3], denoted by $p_i : \mathbb{R}^n \to \mathbb{R}$, that also depends on the other producers' supply, and is modeled by

$$p_i(z) = m_i - d_i \sum_{j=1}^n z_j$$

for each $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$, for some $m_i > b_i$ and $d_i > -a_i$, where the last two assumptions guarantee a positive profit in a monopolistic setting, i.e., when $n = 1$. Thus, given that the goal of each producer is to maximize profit, or equivalently minimize losses, an equilibrium state where no producer has any incentive to deviate unidirectionally from its production plan can be modeled by (6.3), with $\mathcal{Z}_i = [0, T_i]$, $\mathcal{H}_i = \mathbb{R}$, and

$$\varphi_i(z) = c_i(z_i) - z_i p(z)$$

for each $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ and $i \in [\![1, n]\!]$, which fulfill the assumptions in the first paragraph of this section. We identify F as

$$F(z) = \underbrace{\begin{bmatrix} 2(a_1 + d_1) & d_1 & d_1 & \cdots & d_1 \\ d_2 & 2(a_2 + d_2) & d_2 & \cdots & d_2 \\ d_3 & d_3 & 2(a_3 + d_3) & \cdots & d_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_n & d_n & d_n & \cdots & 2(a_n + d_n) \end{bmatrix}}_{=A} z + \begin{bmatrix} b_1 - m_1 \\ \vdots \\ b_n - m_n \end{bmatrix}$$

with Lipschitz constant $L_F = \|A\|$. We also note that [35] provides the existence of a solution in this case. We generate the data similar to the approach in [8, Section 4.1], as outlined below. The results of the numerical experiments are presented in Figure 3.

[3] Also known as the inverse demand function.

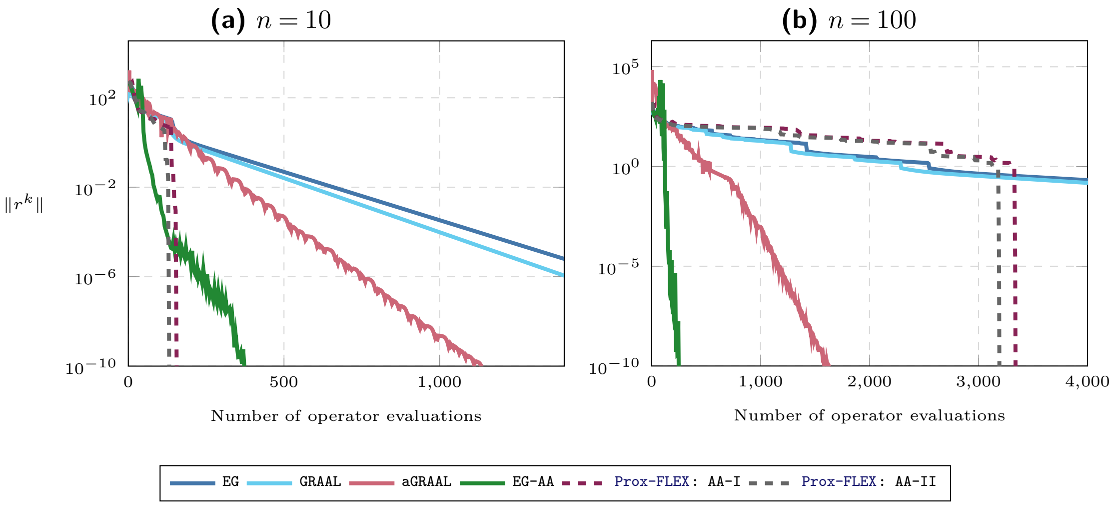

Figure 3: Convergence of algorithms on the Cournot–Nash equilibrium problem where $r^k = R_{1/2L_F}(z^k)$ and R is the residual mapping in (5.1). Both `AA-I` and `AA-II` uses memory parameter $m = 3$. The number of operator evaluations equals the number of F and $\mathrm{prox}_{\gamma g}$ evaluations.

Although $n = 100$ in Figure 3b is not representative of a real oligopolistic market, we include this larger problem size to evaluate the performance and scalability of the algorithms. We observe that `Prox-FLEX` has a superlinear drop-off in both cases and that `EG-AA` and `aGRAAL` scale well for this particular problem.

Input: $n \in \mathbb{N}$
Output: $((T_i, a_i, b_i, m_i, d_i))_{i=1}^n$

repeat
  For each $i \in [\![1, n]\!]$, sample m_i uniformly from $[150, 250]$
  For each $i \in [\![1, n]\!]$, sample b_i uniformly from $[30, 50]$
  For each $i \in [\![1, n]\!]$, sample T_i uniformly from $[3, 7]$
  For each $i \in [\![1, n]\!]$, sample d_i uniformly from $[5, 20]$
  Sort $(d_i)_{i=1}^n$ in increasing order
  For each $i \in [\![1, n]\!]$, sample u_i uniformly from $[-10, -5]$
  For each $i \in [\![1, n]\!]$, compute $a_i = d_i / u_i$
  Sort $(a_i)_{i=1}^n$ in decreasing order
  valid $\leftarrow$ `True`
  for $i \in [\![1, n]\!]$ **do**
    if $b_i < -2a_iT_i$ **or** $m_i \leq b_i$ **or** $d_i \leq -a_i$ **then**
      valid $\leftarrow$ `False`
      break
    end if
  end for
until valid is `True`

6.4 Sparse logistic regression

Consider the sparse logistic regression problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \log\left(1+\exp\left(-b_i\langle a_i,x\rangle\right)\right)+\lambda\|x\|_1 \tag{6.4}$$

where $(a_i,b_i)\in\mathbb{R}^n\times\{-1,1\}$ for each $i=1,\ldots,m$. The minimization problem (6.4) can equivalently be written as the inclusion problem (1.1) by letting $\mathcal{H}=\mathbb{R}^n$ with the inner product set to the dot product, $F:\mathbb{R}^n\to\mathbb{R}^n$ such that

$$F(x)=\sum_{i=1}^{m}-b_i a_i \sigma\left(-b_i\langle a_i,x\rangle\right)=K^\top\sigma(Kx) \tag{6.5}$$

where

$$\sigma:\mathbb{R}^m\to\mathbb{R}^m:\begin{bmatrix}u_1\\ \vdots\\ u_m\end{bmatrix}\mapsto\begin{bmatrix}\frac{\exp(u_1)}{1+\exp(u_1)}\\ \vdots\\ \frac{\exp(u_m)}{1+\exp(u_m)}\end{bmatrix},\quad K=\begin{bmatrix}-b_1a_1^\top\\ \vdots\\ -b_ma_m^\top\end{bmatrix}\in\mathbb{R}^{m\times n},$$

and $g=\lambda\|\cdot\|_1$. Moreover, note that Assumption I holds with $L_F=(1/4)\|K\|^2$. The results of the numerical experiments are presented in Figure 4. Although not designed specifically for minimization problems, we observe that `Prox-FLEX` with `AA-II` directions performs at the top in all but one problem.

7 Conclusions

This paper investigated algorithms for solving inclusion problems involving the sum of a monotone and Lipschitz continuous operator and the subdifferential of a proper, convex, and lower semicontinuous function. We proposed a new Lyapunov function for Korpelevich's extragradient method and established a last-iterate convergence result. Departing from the standard Fejér-type analysis, this Lyapunov-based optimality measure did not rely on a known solution to the inclusion problem. It underpinned three novel algorithms that extend the extragradient method. These algorithms balanced user-specified directions and standard extragradient steps, guided by carefully designed line search steps based on the new Lyapunov analysis. In addition to providing global convergence results under various assumptions, we showed that when the directions are superlinear, no backtracking is triggered, leading to superlinear convergence.

Future research directions include developing similar solution-independent Lyapunov functions for, e.g., the forward-reflected-backward method by Malitsky and Tam [24]. Another promising direction is to broaden the scope of the analysis beyond the monotone setting to include cohypomonotone operators [31], and the more general class of problems characterized by the weak Minty condition [13, 22, 32]. Additionally, further exploration is warranted to adapt the approach to the mirror prox framework [28].

A Background on Korpelevich's extragradient method

In the original paper [21], the extragradient method (1.2) was analyzed under the assumption that g is the indicator function of a nonempty, closed, and convex set, making the proximal operator reduce to the projection onto that set. However, as noted in [25], the extragradient method extends to the more general setting (1.1). The remainder of this section presents results in this more general context, with proofs included for completeness.

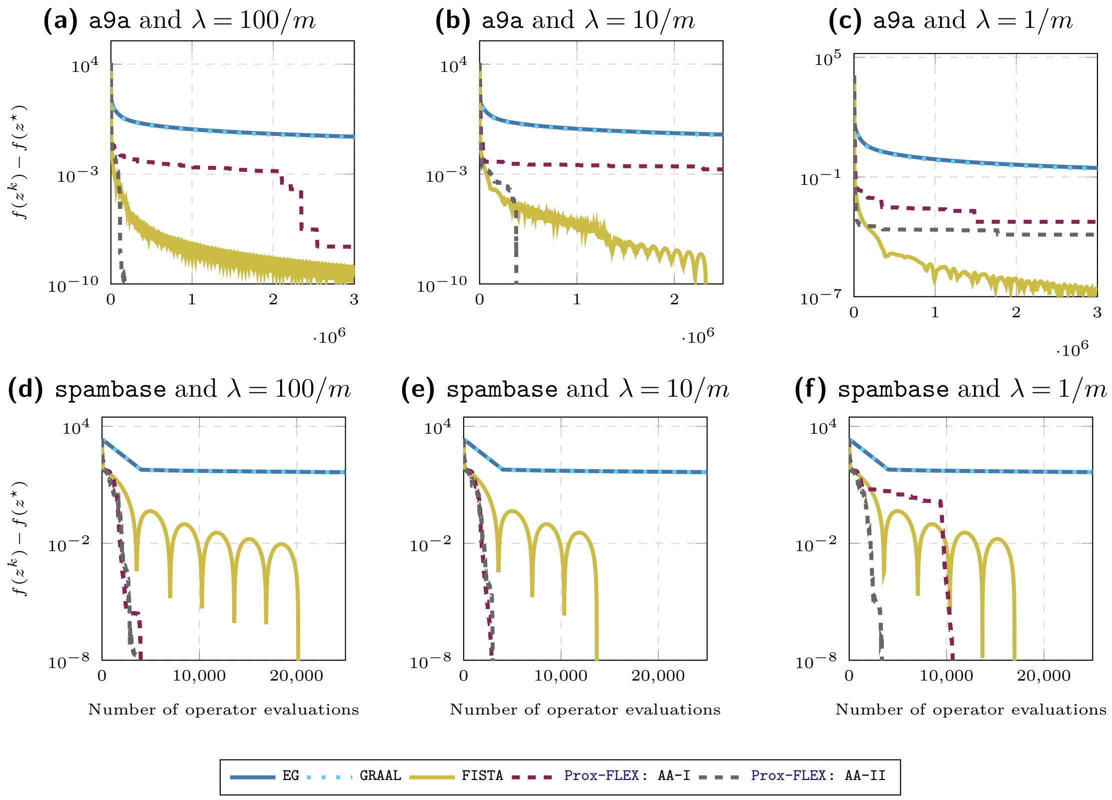

Figure 4: Convergence of algorithms on the sparse logistic regression problem (6.4), using the datasets `a9a` from [11] and `spambase` from [14]. Both `AA-I` and `AA-II` uses memory parameter $m = 10$ for Figures 4a to 4c and $m = 6$ for Figures 4d to 4f. The number of operator evaluations equals the number of F and $\mathrm{prox}_{\gamma g}$ evaluations.

Definition A.1: *Suppose that Assumption I holds and let $\gamma \in \mathbb{R}_{++}$. A point $z \in \mathcal{H}$ is said to be a fixed point of the extragradient method (1.2) if*

$$\bar{z} = \mathrm{prox}_{\gamma g}(z - \gamma F(z)), \tag{A.1a}$$

$$z = \mathrm{prox}_{\gamma g}(z - \gamma F(\bar{z})). \tag{A.1b}$$

Proposition A.2: *Suppose that Assumption I holds and let $\gamma \in \mathbb{R}_{++}$. Then, the following hold:*

(i) If $z \in \mathrm{zer}\,(F + \partial g)$, then z is a fixed point of the extragradient method, i.e., (A.1) holds, and $z = \bar{z}$.

(ii) If $\gamma \in (0, L_F)$, z is a fixed point of the extragradient method, and $\bar{z}$ is defined as in (A.1a), then $z = \bar{z} \in \mathrm{zer}\,(F + \partial g)$.

Proof. The proximal evaluations in (A.1a) and (A.1b) can equivalently be written via their subgradient characterization as

$$\gamma^{-1}(z - \bar{z}) - F(z) \in \partial g(\bar{z}), \tag{A.2a}$$

$$-F(\bar{z}) \in \partial g(z), \tag{A.2b}$$

respectively.

A.2.(i): Note that $z \in \operatorname{zer}(F+\partial g)$ and (A.1a) is equivalent to $-F(z) \in \partial g(z)$ and (A.2a), respectively. Using monotonicity of ∂g [6, Theorem 20.48], we get that

$$\begin{aligned} 0 &\leq \langle \gamma^{-1}(z-\bar{z}) - F(z) + F(z), \bar{z} - z\rangle \\ &= -\gamma^{-1}\|z-\bar{z}\|^2 \leq 0, \end{aligned}$$

since $\gamma \in \mathbb{R}_{++}$. We conclude that $z = \bar{z}$ and that (A.1) holds.

A.2.(ii): By using monotonicity of ∂g at the points $\bar{z}$ and z, and the corresponding subgradients in (A.2), we get that

$$\begin{aligned} 0 &\leq \langle \gamma^{-1}(z-\bar{z}) - F(z) + F(\bar{z}), \bar{z} - z\rangle \\ &= -\gamma^{-1}\|z-\bar{z}\|^2 + \langle F(\bar{z}) - F(z), \bar{z} - z\rangle \\ &\leq -\gamma^{-1}\|z-\bar{z}\|^2 + \|F(\bar{z}) - F(z)\|\|\bar{z} - z\| \\ &\leq (L_F - \gamma^{-1})\|z-\bar{z}\|^2, \end{aligned} \tag{A.3}$$

where the Cauchy–Schwarz inequality is used in the second inequality, and Lipschitz continuity of F in the third inequality. Since $L_F - \gamma^{-1} < 0$, we conclude from (A.3) that $z = \bar{z}$. That $z = \bar{z} \in \operatorname{zer}(F+\partial g)$ now follows from (A.2a) or (A.2b). □

Proposition A.3: *Suppose that Assumption I holds, the sequence* $((z^k, \bar{z}^k))_{k\in\mathbb{N}_0}$ *is generated by* (1.2) *with initial point* $z^0 \in \mathcal{H}$ *and step-size parameter* $\gamma \in \mathbb{R}_{++}$*, and* $z^\star \in \operatorname{zer}(F+\partial g)$*. Then*

$$\|z^{k+1} - z^\star\|^2 \leq \|z^k - z^\star\|^2 - (1-\gamma^2 L_F^2)\|\bar{z}^k - z^k\|^2 \tag{A.4}$$

for each $k \in \mathbb{N}_0$*. Moreover, if* $\gamma \in (0, 1/L_F)$*, then* $(z^k)_{k\in\mathbb{N}_0}$ *converges weakly to a point in* $\operatorname{zer}(F+\partial g)$*.*

Proof. Note that the first and second proximal evaluations in (1.2) are equivalent to

$$0 \leq g(z) - g(\bar{z}^k) - \langle \gamma^{-1}(z^k - \bar{z}^k) - F(z^k), z - \bar{z}^k\rangle \text{ for each } z \in \mathcal{H}, \tag{A.5}$$

and

$$0 \leq g(z) - g(z^{k+1}) - \langle \gamma^{-1}(z^k - z^{k+1}) - F(\bar{z}^k), z - z^{k+1}\rangle \text{ for each } z \in \mathcal{H}, \tag{A.6}$$

respectively, and the assumption $z^\star \in \operatorname{zer}(F+\partial g)$ is equivalent to

$$0 \leq g(z) - g(z^\star) - \langle -F(z^\star), z - z^\star\rangle \text{ for each } z \in \mathcal{H}. \tag{A.7}$$

Picking $z = z^{k+1}$ in (A.5), $z = z^\star$ in (A.6), $z = \bar{z}^k$ in (A.7), summing the resulting inequalities, and multiplying by 2γ gives

$$\begin{aligned} 0 &\leq 2\gamma g(z^{k+1}) - 2\gamma g(\bar{z}^k) - 2\langle z^k - \bar{z}^k - \gamma F(z^k), z^{k+1} - \bar{z}^k\rangle \\ &\quad + 2\gamma g(z^\star) - 2\gamma g(z^{k+1}) - 2\langle z^k - z^{k+1} - \gamma F(\bar{z}^k), z^\star - z^{k+1}\rangle \\ &\quad + 2\gamma g(\bar{z}^k) - 2\gamma g(z^\star) - 2\langle -\gamma F(z^\star), \bar{z}^k - z^\star\rangle \\ &= A_k + B_k, \end{aligned}$$

where

$$\begin{aligned} A_k &= -2\langle z^k - \bar{z}^k, z^{k+1} - \bar{z}^k\rangle - 2\langle z^k - z^{k+1}, z^\star - z^{k+1}\rangle \\ &= \|z^k - z^{k+1}\|^2 - \|z^k - \bar{z}^k\|^2 - \|z^{k+1} - \bar{z}^k\|^2 + \|z^k - z^\star\|^2 - \|z^k - z^{k+1}\|^2 \end{aligned}$$

$$
\begin{aligned}
&- \|z^\star - z^{k+1}\|^2 \\
&= \|z^k - z^\star\|^2 - \|z^\star - z^{k+1}\|^2 - \|z^k - \bar{z}^k\|^2 - \|z^{k+1} - \bar{z}^k\|^2,
\end{aligned}
$$

and

$$
\begin{aligned}
B_k &= 2\gamma\langle F(z^k), z^{k+1} - \bar{z}^k\rangle + 2\gamma\langle F(\bar{z}^k), z^\star - z^{k+1}\rangle - 2\gamma\langle -F(z^\star), \bar{z}^k - z^\star\rangle \\
&= 2\gamma\langle F(z^k), z^{k+1} - \bar{z}^k\rangle + 2\gamma\langle F(\bar{z}^k), z^\star - z^{k+1}\rangle + 2\gamma\langle F(\bar{z}^k), \bar{z}^k - z^\star\rangle \\
&\quad - 2\gamma\langle F(\bar{z}^k) - F(z^\star), \bar{z}^k - z^\star\rangle \\
&\leq 2\gamma\langle F(z^k), z^{k+1} - \bar{z}^k\rangle + 2\gamma\langle F(\bar{z}^k), z^\star - z^{k+1}\rangle + 2\gamma\langle F(\bar{z}^k), \bar{z}^k - z^\star\rangle \\
&= 2\gamma\langle F(z^k) - F(\bar{z}^k), z^{k+1} - \bar{z}^k\rangle \\
&\leq \gamma^2\|F(z^k) - F(\bar{z}^k)\|^2 + \|z^{k+1} - \bar{z}^k\|^2 \\
&\leq \gamma^2 L_F^2\|z^k - \bar{z}^k\| + \|z^{k+1} - \bar{z}^k\|^2,
\end{aligned}
$$

where monotonicity of F is used in the first inequality, Young's inequality is used in the second inequality, and Lipschitz continuity of F in the third inequality. We conclude that

$$
0 \leq A_k + B_k \leq \|z^k - z^\star\|^2 - \|z^\star - z^{k+1}\|^2 - (1 - \gamma^2 L_F^2)\|z^k - \bar{z}^k\|^2,
$$

which proves (A.4).

Next, note that (A.4) gives that $(\|z^k - z^\star\|)_{k\in\mathbb{N}_0}$ converges. Thus, $(z^k)_{k\in\mathbb{N}_0}$ is bounded and there exists a subsequence $(z^k)_{k\in K} \rightharpoonup z^\infty$ for some $z^\infty \in \mathcal{H}$ [6, Lemma 2.45]. Moreover, (A.4) and the requirement $\gamma \in (0, 1/L_F)$ give that $(\|\bar{z}^k - z^k\|^2)_{k\in\mathbb{N}_0}$ is summable, and therefore, $(\bar{z}^k)_{k\in K} \rightharpoonup z^\infty$. The first proximal evaluation in (1.2) can equivalently be written as

$$
\gamma^{-1}(z^k - \bar{z}^k) - F(z^k) + F(\bar{z}^k) \in (F + \partial g)(\bar{z}^k). \tag{A.8}
$$

The left-hand side of (A.8) converges strongly to zero since F is continuous and $(\|z^k - \bar{z}^k\|)_{k\in\mathbb{N}_0}$ converges to zero. Moreover, the operator $F + \partial g$ is maximally monotone, since F is maximally monotone (by continuity and monotonicity [6, Corollary 20.28]), ∂g is maximally monotone [6, Theorem 20.48], and F has full domain [6, Corollary 25.5]. Thus, [6, Proposition 20.38] gives that $z^\infty \in \operatorname{zer}(F + \partial g)$, and by [6, Lemma 2.47] we conclude that $(z^k)_{k\in\mathbb{N}_0}$ converges weakly to a point in $\operatorname{zer}(F + \partial g)$, as claimed. □

B Counterexamples

Example B.1: Let $\mathcal{V}_k = \mathcal{V}(z^k, \bar{z}^k, z^{k+1})$ for the Lyapunov function $\mathcal{V}$ defined in (2.1) and iterates $((z^k, \bar{z}^k))_{k\in\mathbb{N}_0}$ generated by Tseng's method (1.3). This example contains a particular instance of the inclusion problem (1.1), initial point $z^0 \in \mathcal{H}$, and step size $\gamma \in (0, 1/L_F)$ for which $\mathcal{V}_k$ increases between the first two consecutive iterations, thereby establishing that $\mathcal{V}_k$ has no (one-step) decent inequality in this case. In particular, consider $\mathcal{H} = \mathbb{R}^4$, $F : \mathbb{R}^4 \to \mathbb{R}^4$, and $g : \mathbb{R}^4 \to \mathbb{R} \cup \{+\infty\}$ such that

$$
F(z) = \begin{bmatrix} Ax \\ -A^\top y \end{bmatrix} \quad \text{and} \quad g(z) = \begin{cases} 0 & \text{if } x \in [-7, 6]^2 \text{ and } y \in [1, 8]^2, \\ +\infty & \text{otherwise} \end{cases}
$$

for each $z = (x, y) \in \mathbb{R}^2 \times \mathbb{R}^2$, respectively, where

$$
A = \begin{bmatrix} 7 & 6 \\ 1 & 0 \end{bmatrix}.
$$

It is straightforward to verify that Assumption I holds with[4]

$$L_F = \left\| \begin{bmatrix} 0 & A \\ -A^\top & 0 \end{bmatrix} \right\| \approx 9.25091.$$

By letting $z^0 = (-1,-7,-1,7)$ and $\gamma = 1/10$, Tseng's method gives that

$$\begin{aligned} \bar{z}^0 &= \left(-\frac{9}{2}, -\frac{69}{10}, 1, \frac{32}{5}\right), \\ z^1 &= \left(-\frac{277}{50}, -\frac{71}{10}, -\frac{36}{25}, \frac{43}{10}\right), \\ \bar{z}^1 &= \left(-7, -\frac{1739}{250}, 1, 1\right), \end{aligned}$$

and therefore

$$\mathcal{V}_0 = 1662 \quad \text{and} \quad \mathcal{V}_1 = \frac{1187246}{625} = 1899.5936,$$

establishing the claim. □

Example B.2: Consider the inclusion problem

$$\text{find } z \in \mathcal{H} \text{ such that } 0 \in F(z) + T(z)$$

where $F : \mathcal{H} \to \mathcal{H}$ satisfies Assumption I.(i) and $T : \mathcal{H} \to 2^{\mathcal{H}}$ is a maximally monotone operator. Let

$$\begin{aligned} \bar{z}^k &= (\mathrm{Id} + \gamma T)^{-1}(z^k - \gamma F(z^k)), \\ z^{k+1} &= (\mathrm{Id} + \gamma T)^{-1}(z^k - \gamma F(\bar{z}^k)), \end{aligned} \tag{B.1}$$

and $\mathcal{V}_k = \mathcal{V}(z^k, \bar{z}^k, z^{k+1})$ for the Lyapunov function $\mathcal{V}$ defined in (2.1). This example contains a particular problem instance for which $(z^k)_{k\in\mathbb{N}_0}$ diverges and $\mathcal{V}_k$ increases between the first two consecutive iterations. In particular, consider $\mathcal{H} = \mathbb{R}^2$, $z^0 = (10,10)$, $\gamma = 1/10$, and $F, T : \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$F(z) = \underbrace{\begin{bmatrix} 0 & 9 \\ -9 & 0 \end{bmatrix}}_{=A} \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{and} \quad T(z) = \underbrace{\begin{bmatrix} 0 & -4 \\ 4 & 0 \end{bmatrix}}_{=B} \begin{bmatrix} x \\ y \end{bmatrix}$$

for each $z = (x,y) \in \mathbb{R} \times \mathbb{R}$, where $L_F = 9$. Note that (B.1) reduces to

$$z^{k+1} = \underbrace{(I + \gamma B)^{-1} \Big(I - \gamma A \Big((I + \gamma B)^{-1} (I - \gamma A)\Big)\Big)}_{=C} z^k,$$

where $C \in \mathbb{R}^{2\times 2}$ has full rank and spectral radius ≈ 1.132596, which is greater than one. Therefore, we can conclude that $(z^k)_{k\in\mathbb{N}_0}$ diverges. Moreover, (B.1) gives that

$$\begin{aligned} \bar{z}^0 &= \left(\frac{215}{29}, \frac{465}{29}\right), \\ z^1 &= \left(\frac{3245}{1682}, \frac{26745}{1682}\right), \\ \bar{z}^1 &= \left(-\frac{447965}{97556}, \frac{1899785}{97556}\right), \end{aligned}$$

[4]The matrix norm is taken as the spectral norm.

$$z^2 = \left(-\frac{53118995}{5658248}, \frac{87834005}{5658248}\right),$$

and therefore

$$\mathcal{V}_0 = \frac{5875000}{841} \approx 6985.73127229489,$$
$$\mathcal{V}_1 = \frac{12676046875}{1414562} \approx 8961.11084208398,$$

establishing the second claim. □

Acknowledgments

M. Upadhyaya and P. Giselsson acknowledge support from the ELLIIT Strategic Research Area and the Wallenberg AI, Autonomous Systems, and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. Additionally, P. Giselsson acknowledges support from the Swedish Research Council. P. Latafat is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA - National Group for Mathematical Analysis, Probability and their Applications) of the Istituto Nazionale di Alta Matematica (INdAM - National Institute of Higher Mathematics). We thank Yurong Chen for highlighting some details in Section 6.3.

References

[1] M. Ahookhosh, A. Themelis, and P. Patrinos. "A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima". In: *SIAM Journal on Optimization* 31.1 (2021), pp. 653–685. DOI: 10.1137/19M1264783.

[2] A. Alacaoglu, A. Böhm, and Y. Malitsky. "Beyond the golden ratio for variational inequality algorithms". In: *Journal of Machine Learning Research* 24.172 (2023), pp. 1–33.

[3] D. G. Anderson. "Iterative procedures for nonlinear integral equations". In: *Journal of the Association for Computing Machinery (J. Assoc. Comput. Mach.)* 12.4 (1965), pp. 547–560. DOI: 10.1145/321296.321305.

[4] A. Asl, H. Lu, and J. Yang. "A J-symmetric quasi-Newton method for minimax problems." In: *Mathematical Programming* 204.1/2 (2024), pp. 207–254. DOI: 10.1007/s10107-023-01957-1.

[5] S. Axler. *Measure, integration & real analysis.* Graduate Texts in Mathematics: 282. Cham: Springer International Publishing, 2020. DOI: 10.1007/978-3-030-33143-6.

[6] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* CMS Books in Mathematics. Springer, 2017. DOI: 10.1007/978-3-319-48311-5.

[7] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM Journal on Imaging Sciences* 2.1 (2009), pp. 183–202. DOI: 10.1137/080716542.

[8] G. Bigi and M. Passacantando. "Differentiated oligopolistic markets with concave cost functions via Ky Fan inequalities." In: *Decisions in Economics and Finance* (2017), pp. 1–17. DOI: 10.1007/s10203-017-0187-7.

[9] G. Bonanno. "General equilibrium theory with imperfect competition". In: *Journal of Economic Surveys* 4.4 (1990), pp. 297–328. DOI: 10.1111/j.1467-6419.1990.tb00091.x.

[10] R. I. Boţ and E. Chenchene. *Extra-gradient method with flexible anchoring: strong convergence and fast residual decay.* 2024. arXiv: 2410.14369v1 [math.OC].

[11] C.-C. Chang and C.-J. Lin. “LIBSVM: A library for support vector machines”. In: *ACM Trans. Intell. Syst. Technol.* 2.3 (2011). DOI: 10.1145/1961189.1961199.

[12] J. E. Dennis Jr and J. J. Moré. “Quasi-Newton methods, motivation and theory”. In: *SIAM review* 19.1 (1977), pp. 46–89. DOI: 10.1137/1019005.

[13] J. Diakonikolas, C. Daskalakis, and M. Jordan. “Efficient methods for structured nonconvex-nonconcave min-max optimization”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2746–2754.

[14] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[15] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Vol. I.: Springer New York, NY, 2003. DOI: 10.1007/b97543.

[16] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Vol. II.: Springer New York, NY, 2003. DOI: 10.1007/b97544.

[17] H. Fang and Y. Saad. “Two classes of multisecant methods for nonlinear acceleration”. In: *Numerical Linear Algebra with Applications* 16.3 (2009), pp. 197–221. DOI: 10.1002/nla.622.

[18] D. Fudenberg and J. Tirole. *Dynamic models of oligopoly*. Fundamentals of pure and applied economics.: Harwood Academic Publishers, 1986.

[19] 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: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research.: PMLR, 2022, pp. 366–402.

[20] A. F. Izmailov and M. V. Solodov. *Newton-type methods for optimization and variational problems*. Vol. 1.: Springer Cham, 2014. DOI: 10.1007/978-3-319-04247-3.

[21] G. M. Korpelevich. “The extragradient method for finding saddle points and other problems”. In: *Matecon* 12 (1976), pp. 747–756.

[22] S. Lee and D. Kim. “Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems”. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34.: Curran Associates, Inc., 2021, pp. 22588–22600.

[23] Y. Malitsky. “Golden ratio algorithms for variational inequalities”. In: *Mathematical Programming* 184.1 (2020), pp. 383–410. DOI: 10.1007/s10107-019-01416-w.

[24] Y. Malitsky and M. K. Tam. “A forward-backward splitting method for monotone inclusions without cocoercivity”. In: *SIAM Journal on Optimization* 30.2 (2020), pp. 1451–1472. DOI: 10.1137/18M1207260.

[25] R. D. C. Monteiro and B. F. Svaiter. “Complexity of variants of Tseng’s modified F-B splitting and Korpelevich’s methods for hemivariational inequalities with applications to saddle-point and convex optimization problems”. In: *SIAM Journal on Optimization* 21.4 (2011), pp. 1688–1720. DOI: 10.1137/100801652.

[26] P. von Mouche and F. Quartieri, eds. *Equilibrium theory for Cournot oligopolies and related games*. 1st ed. Springer Series in Game Theory 978-3-319-29254-0.: Springer Cham, 2016. DOI: 10.1007/978-3-319-29254-0.

[27] L. D. Muu, V. H. Nguyen, and N. V. Quy. “On Nash–Cournot oligopolistic market equilibrium models with concave cost functions”. In: *Journal of Global Optimization* 41.3 (2008), pp. 351–364. DOI: 10.1007/s10898-007-9243-0.

[28] 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”. In: *SIAM Journal on Optimization* 15.1 (2004), pp. 229–251. DOI: 10.1137/S1052623403425629.

[29] M. A. Noor. “Mixed variational inequalities”. In: *Applied Mathematics Letters* 3.2 (1990), pp. 73–75. DOI: 10.1016/0893-9659(90)90018-7.

[30] K. Okuguchi and F. Szidarovszky. *The theory of oligopoly with multi-product firms*. 2nd ed. Originally published as volume 342 in the series: Lecture Notes in Economics.: Springer Berlin, Heidelberg, 1999. DOI: 10.1007/978-3-642-60169-9.

[31] T. Pennanen. “Local convergence of the proximal point algorithm and multiplier methods without monotonicity”. In: *Mathematics of Operations Research* 27.1 (2002), pp. 170–191. DOI: 10.1287/moor.27.1.170.331.

[32] T. Pethick, P. Latafat, P. Patrinos, O. Fercoq, and V. Cevher. “Escaping limit cycles: global convergence for constrained nonconvex-nonconcave minimax problems”. In: *International Conference on Learning Representations*. 2021.

[33] L. D. Popov. “A modification of the Arrow-Hurwicz method for search of saddle points”. In: *Mathematical Notes of the Academy of Sciences of the USSR* 28.5 (1980), pp. 845–848. DOI: 10.1007/BF01141092.

[34] X. Qu, W. Bian, and X. Chen. *An extra gradient Anderson-accelerated algorithm for pseudomonotone variational inequalities*. 2024. arXiv: 2408.06606 [math.OC].

[35] J. B. Rosen. “Existence and uniqueness of equilibrium points for concave N-person games”. In: *Econometrica* 33.3 (1965), pp. 520–534. DOI: 10.2307/1911749.

[36] M. V. Solodov and B. F. Svaiter. “A Hybrid Approximate Extragradient–Proximal Point Algorithm Using the Enlargement of a Maximal Monotone Operator”. In: *Set-Valued Analysis* 7.4 (1999), pp. 323–345. DOI: 10.1023/A:1008777829180.

[37] M. V. Solodov and B. F. Svaiter. “A globally convergent inexact Newton method for systems of monotone equations”. In: *Reformulation: Nonsmooth, piecewise smooth, semismooth and smoothing methods* (1999), pp. 355–369. DOI: 10.1007/978-1-4757-6388-1_18.

[38] L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos. “A simple and efficient algorithm for nonlinear model predictive control.” In: *IEEE 56th Annual Conference on Decision and Control (CDC)*. 2017, pp. 1939–1944. DOI: 10.1109/CDC.2017.8263933.

[39] A. Themelis and P. Patrinos. “SuperMann: a superlinearly convergent algorithm for finding fixed points of nonexpansive operators”. In: *IEEE Transactions on Automatic Control* 64.12 (2019), pp. 4875–4890. DOI: 10.1109/TAC.2019.2906393.

[40] Q. Tran-Dinh and N. Nguyen-Trung. *Accelerated extragradient-type methods – part 2: generalization and sublinear convergence rates under co-hypomonotonicity*. 2025. arXiv: 2501.04585 [math.OC].

[41] Q. Tran-Dinh and N. Nguyen-Trung. *Revisiting extragradient-type methods – part 1: generalizations and sublinear convergence rates*. 2024. arXiv: 2409.16859 [math.OC].

[42] P. Tseng. “A modified forward-backward splitting method for maximal monotone mappings”. In: *SIAM Journal on Control and Optimization* 38.2 (2000), pp. 431–446. DOI: 10.1137/S0363012998338806.

[43] X. Vives. “Cournot and the oligopoly problem”. In: *European Economic Review* 33.2 (1989), pp. 503–514. DOI: 10.1016/0014-2921(89)90129-3.

[44] 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: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research.: PMLR, 2021, pp. 12098–12109.

[45] J. Zhang, B. O’Donoghue, and S. Boyd. “Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations”. In: *SIAM Journal on Optimization* 30.4 (2020), pp. 3170–3197. DOI: 10.1137/18M1232772.