A Fast Optimistic Method for Monotone Variational Inequalities

Michael Sedlmayer [1] **Dang-Khoa Nguyen** [2] **Radu Ioan Boţ** [1 2]

Abstract

We study monotone variational inequalities that can arise as optimality conditions for constrained convex optimisation or convex-concave minimax problems and propose a novel algorithm that uses only one gradient/operator evaluation and one projection onto the constraint set per iteration. The algorithm, which we call *fOGDA-VI*, achieves a $o(1/k)$ rate of convergence in terms of the restricted gap function as well as the natural residual for the *last iterate*. Moreover, we provide a convergence guarantee for the sequence of iterates to a solution of the variational inequality. These are the best theoretical convergence results for numerical methods for (only) monotone variational inequalities reported in the literature. To empirically validate our algorithm we investigate a two-player matrix game with mixed strategies of the two players. Concluding, we show promising results regarding the application of fOGDA-VI to the training of generative adversarial nets.

1. Introduction

Variational inequalities are fundamental models in various fields such as optimisation, e.g., when determining primal-dual pairs of optimal solutions of constrained convex optimisation problems (Bauschke & Combettes, 2011), economics, game theory (Morgenstern & Von Neumann, 1953), or partial differential equations. Recently, they have attracted particularly significant attention in the area of machine learning due to the fundamental role they play, for instance, in multi agent reinforcement learning (Omidshafiei et al., 2017), robust adversarial learning (Madry et al., 2018) and the training of generative adversarial networks (GANs) (Goodfellow et al., 2014; Goodfellow, 2016).

[1]Research Network Data Science, University of Vienna, Vienna, Austria [2]Faculty of Mathematics, University of Vienna, Vienna, Austria. Correspondence to: Michael Sedlmayer <michael.sedlmayer@univie.ac.at>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1.1. Problem Setting

In the following we consider $\mathbb{R}^d$ with its standard inner product denoted by $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|$. Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a monotone operator, i.e.,

$$\langle F(w) - F(z), w - z \rangle \geq 0 \quad \forall w, z \in \mathbb{R}^d,$$

which is also L-Lipschitz continuous, i.e.,

$$\|F(w) - F(z)\| \leq L \|w - z\| \quad \forall w, z \in \mathbb{R}^d.$$

Furthermore, let C be a nonempty closed convex subset of $\mathbb{R}^d$. Then the (strong) classical *variational inequality* problem consists of finding $z^* \in C$ such that

$$\langle F(z^*), z - z^* \rangle \geq 0 \quad \forall z \in C. \tag{1}$$

For the following considerations we assume that the solution set of (1) is nonempty, i.e., $\Omega := \{z^* \in C \mid \langle F(z^*), z - z^* \rangle \geq 0 \quad \forall z \in C\} \neq \emptyset$.

Note that in the case of F being monotone and continuous, the above *strong* formulation is equivalent to the following problem,

$$\langle F(z), z - z^* \rangle \geq 0 \quad \forall z \in C, \tag{2}$$

which is known as the *weak* version of the variational inequality. Writing $N_C(z) := \{w \in \mathbb{R}^d \mid \langle v - z, w \rangle \leq 0 \quad \forall v \in C\}$, for $z \in C$, and $N_C(z) := \emptyset$, for $z \notin C$, to denote the normal cone of C, condition (1) is equivalent to the following *monotone inclusion*, where want to find $z^* \in \mathbb{R}^d$ such that

$$0 \in F(z^*) + N_C(z^*). \tag{3}$$

1.2. Contribution

We introduce an accelerated first order method for solving the constrained variational inequality problem (1) that uses a single operator evaluation and a single projection in each iteration. Our proposed algorithm, called fOGDA-VI, exhibits a $o(1/k)$ rate of convergence for the last iterate which is better than the $\mathcal{O}(1/k)$ results for other accelerated algorithms. Moreover, fOGDA-VI exhibits convergence of the generated sequence to a solution of the variational inequality under investigation, which is not necessarily the case for other accelerated methods (Cai et al., 2022a; Cai & Zheng, 2022) that have been proposed for (1).

1.3. Overview

This paper is structured as follows. In Section 2 we discuss suitable convergence measures and (accelerated) solution methods for monotone variational inequalities governed by a monotone and Lipschitz operator. The algorithm fOGDA-VI and the accompanying convergence results are presented in Section 3, which is followed by illustrations of the empirical performance of the proposed method when solving two-player matrix games and in the training of GANs in Section 4.

2. Solving Variational Inequalities

In this section we recall appropriate measures of convergence for solution methods for monotone variational inequalities and provide an overview on the most important solution methods from the literature, both nonaccelerated and accelerated ones, for solving (1).

2.1. Convergence Measures

We start with presenting three suitable measures that are commonly used to judge the quality of prospective solutions.

Restricted gap function For $z^* \in \Omega, z_0 \in \mathbb{R}^d$ and $\delta(z_0) := \|z^* - z_0\|$, the restricted *gap* function associated with the variational inequality (1) is defined as

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

It is also known as *merit* function (Nesterov, 2007) and it measures how much the statement of (2) is violated.

In the above definition, $\mathbb{B}(z; \delta) := \left\{ w \in \mathbb{R}^d \mid \|w - z\| \leq \delta \right\}$ denotes the closed ball centred at $z \in \mathbb{R}^d$ with radius $\delta > 0$. The restriction of the supremum to a bounded set, particularly the ball $\mathbb{B}(z^*; \delta(z_0))$ in our case, is essential to avoid an infinitely large gap when C is unbounded.

Tangent residual Another quantity that can be used to measure the quality of a solution candidate with respect to the variational inequality (1) is based on the observation that the latter it is equivalent to the monotone inclusion (3). The so-called *tangent residual* is given by

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

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

$$F(z^*) = 0, \tag{4}$$

to variational inequalities by measuring the distance from 0 to $F(z) + N_C(z)$. Note, if $z \notin C$ we have $N_C(z) = \emptyset$ and thus $r(z) = +\infty$.

Natural residual Another useful convergence measure is the *natural residual*, which in fact is upper bounded by the tangent residual, see Section B.1. For this we write $P_C(z)$ to denote the projection of $z \in \mathbb{R}^d$ onto the closed convex set C, which is uniquely defined and given by $P_C(z) = \arg\min_{w \in C} \|w - z\|$. Using the characterisation of the projection via the normal cone (see Proposition 6.46 in (Bauschke & Combettes, 2011)) we observe that

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

This motivates to look at

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

which is also known as *fixed point residual*. Note, in the unconstrained case (4) the two residuals coincide

$$r(z) = \texttt{Res}(z) = \|F(z)\| \quad \forall z \in \mathbb{R}^d.$$

2.2. Solution Methods

In this work we are interested exclusively in first order methods that are fully splitting, i.e., algorithms that only use direct evaluations of the operator F and projections onto C as main building blocks. As a general Lipschitz continuous operator is not necessarily cocoercive, the simplest first order method that is splitting – the Forward-Backward (FB) algorithm – can not be used to solve (1).

2.2.1. Nonaccelerated Solution Methods

Extragradient (EG) method Korpelevich (1976) and Antipin (1976) proposed to take a second forward evaluation of F in each iteration in order to solve (1). This results in the following scheme for $k \geq 0$

$$\text{EG:} \left\lfloor \begin{array}{l} w_k = P_C\left[z_k - \gamma F(z_k)\right] \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k)\right] \end{array} \right. \tag{5}$$

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

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

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

in both the unconstrained (Nemirovski, 2004; Nesterov, 2007; Mokhtari et al., 2020) and the constrained case (Hsieh et al., 2019), which further seems to be optimal (Ouyang & Xu, 2021). The *best iterate* convergence in terms of the tangent residual, however, is known to yield a rate of $\mathcal{O}(1/\sqrt{K})$ (Korpelevich, 1976; Facchinei & Pang, 2003), i.e.,

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

The more desirable *last iterate* convergence rate for EG was derived only recently in the unconstrained case (Gorbunov et al., 2022) which was then extended to the constrained case as well (Cai et al., 2022b). In fact,

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

as $k \to +\infty$. The result for the restricted gap function (Golowich et al., 2020b) as well as for the residuals is actually tight, meaning that the convergence rate for the averaged iterates is better than for the last iterate. Nevertheless, we emphasise that the latter one is more appealing and that the averaged iterates might still show acceptable behaviour while the actual trajectory of iterates cycles around the set of solutions (Mertikopoulos et al., 2018).

Popov's method In the saddle point setting Popov (1980) introduced the following algorithm which, when applied to (3), reads for $k \geq 1$

$$\text{Popov:} \left\lfloor \begin{array}{l} w_k = P_C\left[z_k - \gamma F(w_{k-1})\right] \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k)\right] \end{array} \right. \tag{6}$$

which converges to a solution of (1) for $0 < \gamma < 1/2L$. The update rule of (6) is very similar to (5) but requires F to be evaluated only once per iteration. Actually, Popov differs from EG only in the first block, where $F(z_k)$ is replaced by $F(w_{k-1})$. In the unconstrained case, (6) can be written in one line, yielding a method usually known as Optimistic Gradient Descent Ascent (OGDA), a name that was coined by works on GAN training (Daskalakis et al., 2018; Daskalakis & Panageas, 2018).

Given the close connection between EG and OGDA, it is not surprising that many convergence rate results hold in a similar way. In terms of the restricted gap OGDA converges like $\mathcal{O}(1/k)$ and $\mathcal{O}(1/\sqrt{k})$ for averaged iterates (Mokhtari et al., 2020) and last iterates (Golowich et al., 2020b), respectively, where the latter one is optimal. The convergence rate in terms of the residuals is $\mathcal{O}(1/\sqrt{k})$ and it can not be improved in general (Golowich et al., 2020a; Chavdarova et al., 2021a; Cai et al., 2022b), as seen for EG.

We have seen that both EG and Popov's method require two projections in each iteration. One might think that this is necessary to obtain convergent algorithms for the variational inequality (3) when the operator F is merely Lipschitz continuous and not cocoercive. This is not the case, however, and in the following we will look at algorithms that need only one evaluation of the projection operator per iteration.

Forward-Backward-Forward (FBF) method One of these single-call projection methods is the FBF method. It was proposed by Tseng (2000) and applied to (3) it iterates for $k \geq 0$

$$\text{FBF:} \left\lfloor \begin{array}{l} w_k = P_C\left[z_k - \gamma F(z_k)\right] \\ z_{k+1} = w_k - \gamma F(w_k) + \gamma F(z_k) \end{array} \right. \tag{7}$$

which converges to a solution of (1) for $0 < \gamma < 1/L$. Notice that in each iteration this iterative scheme performs two evaluations of F along the sequences $(z_k)_{k\geq 0}$ and $(w_k)_{k\geq 0}$, similar to EG – the first line of FBF and EG is even identical. In the second line, however, instead of performing another projection the forward step regarding the intermediate iterate w_k is corrected by the previous update $F(z_k)$. Moreover, for the unconstrained problem (4), i.e., in the absence of projections, FBF and EG are equivalent.

Forward-Reflected-Backward (FRB) method Another single-call projection method that even requires only one evaluation of F like the basic FB algorithm was proposed by Malitsky and Tam (2020). The FRB algorithm is given for $k \geq 1$ by

$$\text{FRB:} \left\lfloor \; z_{k+1} = P_C\left[z_k - 2\gamma F(z_k) + \gamma F(z_{k-1})\right] \right. \tag{8}$$

and converges to a solution of (1) for $0 < \gamma < 1/2L$. Note that FRB can be deducted from FBF by reusing $F(w_{k-1})$ instead of $F(z_k)$ in the first line of (7), similarly to how Popov's method can be obtained from EG. Hence (8) coincides with (6) and OGDA in the unconstrained case.

Projected Reflected Gradient (RG) method Before investigating FRB, Malitsky (2015) introduced another similar method where the order of the reflection and the forward step is reversed. In particular, a second forward step can be avoided by evaluating F at an appropriate linear combination of the iterates. This gives rise to the following method for $k \geq 1$

$$\text{RG:} \left\lfloor \begin{array}{l} w_k = 2z_k - z_{k-1} \\ z_{k+1} = P_C\left[z_k - \gamma F(w_k)\right] \end{array} \right. \tag{9}$$

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

Despite the similarities in the construction and iterate convergence, nonasymptotic convergence is less understood in the case of single-call projection methods. For example Banert and Boţ (2018) derived for Tseng's method an ergodic $\mathcal{O}(1/k)$ rate in terms of function values in the context of convex optimisation; see also (Böhm et al., 2022) for an ergodic $\mathcal{O}(1/\sqrt{k})$ convergence result in terms of the restricted gap function in the stochastic setting. For Malitsky's RG algorithm convergence in terms of the gap function and residuals like $\mathcal{O}(1/\sqrt{k})$ for the last iterate was established recently (Cai & Zheng, 2022).

2.2.2. Accelerated Solution Methods

Extra Anchored Gradient (EAG) algorithm An accelerated algorithm for solving the monotone equation (4) that

is based on EG, called Extra Anchored Gradient (EAG) algorithm, was proposed by Yoon and Ryu (2021). It is designed by using *anchoring*, a technique that can be traced back to Halpern's algorithm (1967). This iterative scheme exhibits a convergence rate of

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

These considerations have been followed by extension of EAG to the constrained setting (Cai et al., 2022a), where the authors consider for $k \geq 0$

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

with $0 < \gamma < 1/\sqrt{3}L$. One can notice that the algorithm uses two operator evaluations and two projection steps, like EG, and in the unconstrained case it coincides with its projection free counterpart (Yoon & Ryu, 2021), maintaining the $\mathcal{O}(1/k)$ convergence rate for gap function and residuals.

Accelerated Reflected Gradient (ARG) algorithm Similar to the nonaccelerated methods from the previous subsection, one can also reduce the number of necessary operator evaluations and projections to one each per iteration. This was done by investigating an accelerated version (Cai & Zheng, 2022) of the projected reflected gradient method (9) with convergence rate $\mathcal{O}(1/k)$. The method is given for $k \geq 1$ by

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

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

It is worth mentioning that for constrained EAG (Cai et al., 2022a) and ARG (Cai & Zheng, 2022) there are no guarantees for the iterates to converge to a solution and that, in spite of the fast theoretical convergence rate, the effect of the *anchor* z_0, to which the algorithm returns in every iteration, on the convergence speed is a slowing one, as we will see in the numerical experiments.

Further accelerated algorithms Further variants of anchoring based algorithms have been proposed by Tran-Dinh (2022) and together with Luo (Tran-Dinh & Luo, 2021), which all exhibit the same convergence rate in terms of the operator norm as EAG for (4). Monotone inclusions are also considered in these works, with a more general operator than the normal cone, but this requires either taking a backward step or additionally asking for cocoercivity of F. In the same spirit, an $o(1/k)$ rate of convergence together with convergence of iterates was shown for an accelerated version of the Krasnosel'skiĭ-Mann algorithm (Boţ & Nguyen, 2022).

Explicit Fast OGDA (fOGDA) algorithm Another approach which is different from the Halpern-type methods mentioned above was investigated in (Boţ et al., 2022). An appropriate (explicit) discretisation of a second-order dynamical system with vanishing damping term gives rise to an accelerated algorithm related to OGDA, called *fast OGDA* (fOGDA), which will constitute a starting point for our considerations in the following. The fast OGDA algorithm (Boţ et al., 2022) for solving the monotone equation (4) is given for $k \geq 1$ by

$$\text{fOGDA:} \left| \begin{array}{l} w_k = z_k + \frac{k}{k+\alpha}(z_k - z_{k-1}) - \gamma\frac{\alpha}{k+\alpha}F(w_{k-1}) \\ z_{k+1} = w_k - \gamma\frac{2k+\alpha}{k+\alpha}(F(w_k) - F(w_{k-1})) \end{array} \right.$$

and converges to a solution of (4) for $0 < \gamma < 1/4L$ and $\alpha > 2$. It was shown that fOGDA exhibits convergence rates like

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

as $k \to +\infty$.

3. Main Results

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

3.1. Extending fOGDA to the Constrained Case

It can be seen empirically, that incorporating one (or more) projections to the unconstrained fOGDA method in a naive way, similar to FRB or Popov, is not sufficient to obtain a solution method for (1). Instead, the idea is to introduce for every $k \geq 1$ an appropriate element ζ_k of the normal cone $N_C(z_k)$. This is done by replacing $F(w_{k-1})$ by the sum $F(w_{k-1}) + \zeta_k$. One might think that because of this the suitable choice would be to take $\zeta_k \in N_C(w_{k-1})$, but this is not the case which can be motivated as follows. For the unconstrained case, i.e., when solving the monotone equation (4), we want to find a sequence $(z_k)_{k\geq 1}$ yielding $\|F(z_k)\| \to 0$. However, in the constrained case, i.e., when tackling the monotone inclusion (3), we aim to establish sequences $(z_k)_{k\geq 1}$, $(\zeta_k)_{k\geq 1}$ with $\zeta_k \in N_C(z_k)$ such that

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

Algorithm 1 fOGDA-VI

Input: momentum parameter $\alpha > 2$; starting values z_0, $w_0 \in \mathbb{R}^d$, $z_1 \in C$, $\zeta_1 \in N_C(z_1)$; step size $0 < \gamma < 1/4L$; number of iterations $K > 1$.

for $k = 1$ **to** K **do**

Compute

$$w_k = z_k + \frac{k}{k+\alpha}(z_k - z_{k-1}) - \gamma\frac{\alpha}{k+\alpha}(F(w_{k-1}) + \zeta_k)$$
$$z_{k+1} = P_C\left[w_k - \gamma\left(1 + \frac{k}{k+\alpha}\right)(F(w_k) - F(w_{k-1}) - \zeta_k)\right]$$
$$\zeta_{k+1} = \frac{k+\alpha}{\gamma(2k+\alpha)}(w_k - z_{k+1}) - (F(w_k) - F(w_{k-1}) - \zeta_k)$$

end for

Then instead of fOGDA we have an algorithm which is given for every $k \geq 1$ by

$$\begin{aligned} w_k &= z_k + \tfrac{k}{k+\alpha}(z_k - z_{k-1}) - \gamma\tfrac{\alpha}{k+\alpha}(F(w_{k-1}) + \zeta_k), \\ z_{k+1} &= w_k - \gamma\tfrac{2k+\alpha}{k+\alpha}(F(w_k) - F(w_{k-1}) + \zeta_{k+1} - \zeta_k). \end{aligned} \tag{10}$$

At first glance this method seems to be implicit, as both z_{k+1} and ζ_{k+1} appear on the same line. However, the second line in (10) can be used to formulate a projection step (see Appendix B.1 for details). Even though from the perspective of (13) the appearance of the normal cone element is a natural consequence of the projection, finding its correct formulation is a highly non-trivial task. These considerations give rise to Algorithm 1, which we call *fOGDA-VI*.

The two main differences of our proposed method – introduction of a projection at a specific spot as well as explicit computation of a particular element in the normal cone – are deemed to be necessary. Changing (or even neglecting) either of them results in algorithms that fail to converge in general.

Remark 3.1. Initialisation of fOGDA-VI, however, is easy as for general $z_1 \in C$ it is sufficient to take $\zeta_1 = 0$. With arbitrary $\hat{z} \in \mathbb{R}^d$, one can also take $z_1 := P_C(\hat{z})$ and $\zeta_1 := \hat{z} - z_1 \in N_C(z_1)$

3.2. Convergence Statements

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

Theorem 3.2. *Let $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 1. Then the sequence $(z_k)_{k\geq 0}$ converges to a solution of (1).*

The central idea for the proof is to define an appropriate family of energy functions $(\mathcal{G}_{\lambda,k})_{k\geq 0}$, where $\lambda \geq 0$ is a parameter that depends on α, which dissipate over the course of the algorithm to obtain convergence or summability of various helpful quantities. Even though we are not able to enforce the family of discrete energies $(\mathcal{G}_{\lambda,k})_{k\geq 0}$ to have an actual nonincreasing property, we can at least show that for every $k \geq 0$

$$\mathcal{G}_{\lambda,k+1} \leq (1 + d_{\lambda,k})\mathcal{G}_{\lambda,k} - b_{\lambda,k}, \tag{11}$$

with some sequences $(b_{\lambda,k})_{k\geq 0}$ and $(d_{\lambda,k})_{k\geq 0}$. The aim is to control these two sequences in such a way that we can still derive some beneficial asymptotic results for $(\mathcal{G}_{\lambda,k})_{k\geq 0}$. As the additional terms are not necessarily nonnegative, a novel Lyapunov analysis is needed. For instance, we show that there exist $0 \leq \underline{\lambda}(\alpha) < \overline{\lambda}(\alpha) \leq (3\alpha-2)/4$ such that every $\underline{\lambda}(\alpha) < \lambda < \overline{\lambda}(\alpha)$ provides an energy function $(\mathcal{G}_{\lambda,k})_{k\geq 0}$ that is bounded from below and nonnegative sequences $(b_{\lambda,k})_{k\geq 0}$ and $(d_{\lambda,k})_{k\geq 0}$ with $\sum_{k\geq 0} d_{\lambda,k} < +\infty$ such that the inequality (11) holds for k large enough. This allows us to conclude that $\lim_{k\to+\infty} \mathcal{G}_{\lambda,k} \in \mathbb{R}$ exists, see Lemma A.2 for more details.

From this we can then verify that the first condition of Opial's lemma, see Lemma A.3, is fulfilled, while its second condition follows from the maximal monotonicity of $F + N_C$, see Proposition A.4.

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

Theorem 3.3. *Let $z^* \in \Omega$ be a solution of (1) and let $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 1. Then, as $k \to +\infty$, we have*

$$\texttt{Gap}(z_k) = o\left(\frac{1}{k}\right) \quad \text{and} \quad \texttt{Res}(z_k) = o\left(\frac{1}{k}\right).$$

Remark 3.4. The tangent residual exhibits the same last iterate convergence rate as the natural residual, i.e.,

$$r(z_k) = o\left(\frac{1}{k}\right).$$

In fact, we use the observation $\texttt{Res}(z) \leq \|F(z) + \zeta\|$, see (14), in the proof of Theorem 3.3 to obtain the convergence rate for the natural residual. As the restricted gap and the natural residual are mostly used in the literature (and are probably more intuitive) to quantify the convergence behaviour of numerical methods for variational inequalities, we opted to present Theorem 3.3 in the above manner.

4. Numerical Experiments

In this section we provide two numerical experiments to complement our theoretical results. For the first one we treat a two-player zero sum game, which amounts to solving a bilinear saddle point problem constrained by standard simplexes. The second one consists of application of fOGDA-VI to the training of GANs.

4.1. Two-player Zero Sum Game

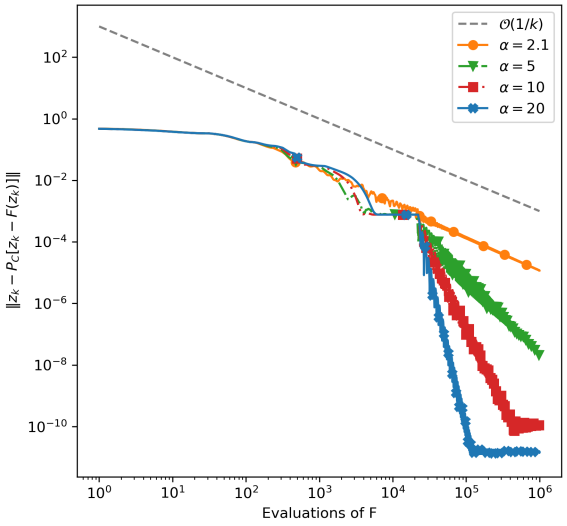


Figure 1. Comparison of different momentum parameters $\alpha > 2$ in Algorithm 1 in terms of the natural residual.

We aim to solve a two-player zero sum game with mixed strategies, which means that we need to solve the following bilinear saddle point problem,

$$\min_{x \in \Delta^m} \max_{y \in \Delta^n} \Phi(x, y) := x^T A y, \tag{12}$$

where $A \in \mathbb{R}^{m \times n}$ is a given pay-off matrix and $\Delta^d = \{v \in \mathbb{R}^d_+ \mid \sum_{i=1}^d v_i = 1\}$ denotes the d-dimensional standard simplex. Recall that a solution of (12) is given by a saddle point $(x^*, y^*) \in \Delta^m \times \Delta^n$ satisfying

$$\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*) \quad \forall (x, y) \in \Delta^m \times \Delta^n.$$

This leads to a monotone inclusion problem (3) with $C = \Delta^m \times \Delta^n$ and

$$F : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m \times \mathbb{R}^n, \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \nabla_x \Phi(x, y) \\ -\nabla_y \Phi(x, y) \end{pmatrix},$$

which gives

$$F(x, y) = \begin{pmatrix} Ay \\ -A^T x \end{pmatrix} = \begin{pmatrix} 0 & A \\ -A^T & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

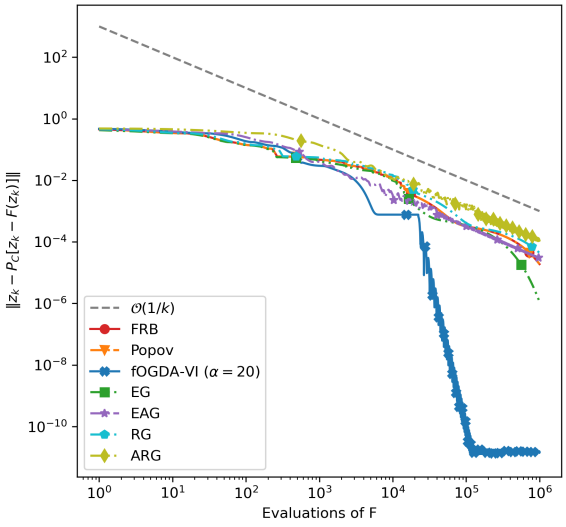


Figure 2. Comparison of different methods in terms of the natural residual.

Notice that F is Lipschitz continuous but not cocoercive, thus indeed the regular *Projected Gradient Descent Ascent* algorithm (PGDA), which is in this case nothing else than the FB algorithm, cannot be applied.

For our experiments we choose $d = n = 50$ and A to have entries drawn from the uniform distribution on the half-open interval $[0, 1)$. As the parameter $\alpha > 2$ can be chosen arbitrarily, we do a comparison of different values in terms of the natural residual in Figure 1 to gain more insight. Note that there is no upper bound for α that would be given on the basis of the theoretical considerations. As it turns out, from a certain point on after a period where all values of α perform similarly, bigger choices for α seem to give better results with faster convergence (even though the convergence rate of $o(1/k)$ is the same for all choices).

Table 1. Comparison of fOGDA-VI with LA-GDA in terms of Fréchet Inception Distance (FID; lower is better) and Inception Score (IS; higher is better). We report the best obtained scores, averaged over 5 runs with 500,000 iterations each. For all considered methods we evaluated the last (non averaged) iterates, the uniform average, the exponential moving average (EMA) and the EMA on the "slow weights" for the method incorporating "lookahead". Best scores for each metric are in boldface.

	FID				IS			
Method	non avg.	uniform avg.	EMA	EMA-slow	non avg.	uniform avg.	EMA	EMA-slow
fOGDA	18.49 ± 1.09	17.38 ± 1.69	18.51 ± 1.13	–	$7.82 \pm .07$	$8.7 \pm .15$	$8.1 \pm .15$	–
LA-GDA	$16.7 \pm .67$	$16.02 \pm .84$	$16.84 \pm .71$	**15.31** ± 1.27	$7.88 \pm .08$	**8.76** $\pm .19$	$8.29 \pm .07$	$8.59 \pm .1$

When going to "extremely big" choices of α we could not only observe further boost in convergence speed, but also increased oscillatory behaviour after a certain point. Whether fOGDA-VI for $\alpha \to +\infty$ amounts to a convergent method is not obvious; for the unconstrained fOGDA by (Boţ et al., 2022) one can see that this would lead to the unaccelerated OGDA method.

Concluding, we show a comparison of different methods that we have encountered in Sections 2.2.1 and 2.2.2 in Figure 2 where we report results on the natural residual. We see that fOGDA-VI clearly outperforms all other methods while only requiring one evaluation of F and one projection in each iteration.

4.2. GAN Training

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) form a powerful class of generative models that can produce for example unseen realistic images. Originally the problem was posed as a zero sum game between two adversarial players played by two neural networks, called *generator* and *discriminator*, that try to minimise and maximise the same loss function, respectively. The minimax structure of the underlying optimisation problem generally leads to cycling behaviour during the training process, making GANs notoriously hard to optimise (Mescheder et al., 2017; 2018). As it was shown empirically that principled methods that are used to solve variational inequalities (Gidel et al., 2019) and monotone inclusions (Böhm et al., 2022) can prove beneficial in the training process, we will apply our proposed algorithm fOGDA-VI to train ResNet architectures on the CIFAR-10 dataset.

For our experiments we use ResNet (He et al., 2016) architectures, see Appendix C.1, with the hinge version of the adversarial non-saturating loss (Miyato et al., 2018) trained on the CIFAR-10 (Krizhevsky, 2009) data set, which consists of 60,000 $(32\times32\times3)$-images in 10 classes, with 6,000 images per class. The metrics used to evaluate the generated images are the inception score (Salimans et al., 2016) (IS; higher is better) and the Fréchet inception distance (Heusel et al., 2017) (FID; lower is better), both computed on 50,000 samples in their original implementations.

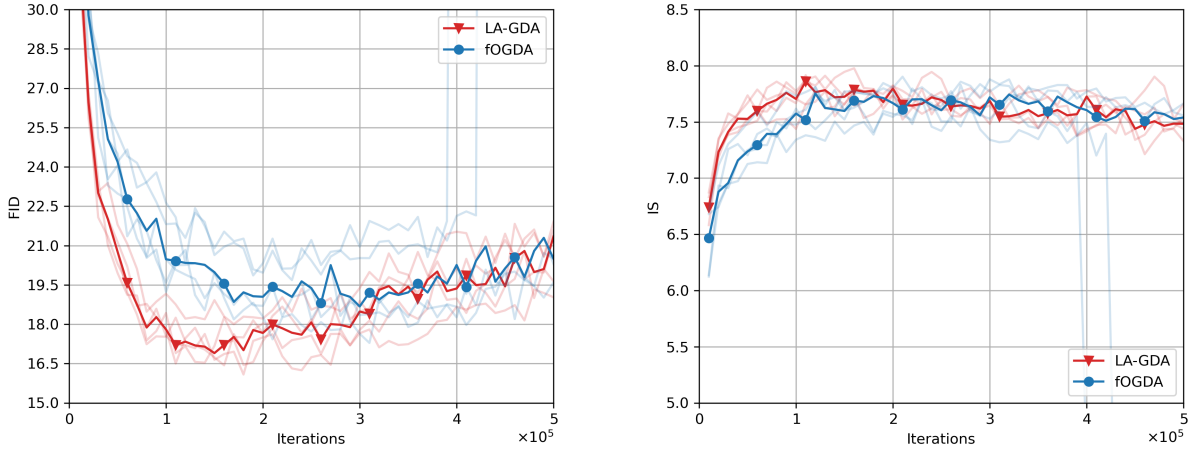
Furthermore, in our experiments instead of stochastic gradients we use the Adam optimiser (Kingma & Ba, 2014) with parameters $\beta_1 = 0$ and $\beta_2 = 0.9$ that were used in recent experiments (Chavdarova et al., 2021b) outperforming the class-dependent BigGAN (Brock et al., 2019) model on CIFAR-10. Additionally, we keep the batch size and the ratio of discriminator and generator updates the same as in (Chavdarova et al., 2021b).

Since we have mini batch updates for the GAN experiments instead of taking the full gradient we perform significantly more steps to incorporate the entire gradient information. Because of this the iterator k in Algorithm 1 might grow too large soon, so we conducted experiments to update the iterator k only every n-th step with different choices of n. Furthermore, we also did a hyperparameter search regarding the learning rate and the momentum parameter $\alpha > 2$ in Algorithm 1.

Table 2. Overall best obtained scores in terms of Fréchet Inception Distance (FID; lower is better) and Inception Score (IS; higher is better) for the last (non averaged) iterates. Best scores for each metric are in boldface.

Method	FID	IS
fOGDA-VI	15.69	8.91
LA-GDA	**14.09**	**9.06**

When performing the hyperparameter search for the momentum parameter α, we experienced that, just as in the theoretically justified setting of Section 4.1, bigger values seemed to perform better. Regarding the frequency of iterator updates we also observed better behaviour for bigger values of n in general. The parameters we used for the fOGDA-VI experiments were $\alpha = 100$ and $n = 1000$, and a learning rate of $\gamma = 0.0001$. We compared the results obtained by fOGDA-VI with the best method from (Chavdarova et al., 2021b), a variant of Gradient Descent Ascent incorporating averaging during the training which they call "lookahead", resulting in a method we denote by LA-GDA, for the convergence properties of which no theoretical evidence is available. For the LA-GDA experiments we kept all hyperparameters as reported by Chavdarova et al. (2021b). For both methods



(a) FID. (b) IS.

Figure 3. The median and the individual runs are illustrated with ticker solid lines and transparent lines, respectively. We report (a) FID and (b) IS for the last (non averaged) iterates.

we conducted 5 runs with 500,000 iterations each.

To further reduce the cycling characteristics of GAN training two commonly used techniques in practice are the (uniform) averaging and the exponential moving averaging (EMA) of the network weights. The beneficial effects of uniform and exponential averaging can also be observed in our experiments (see Table 1), however uniform averaging seems to have a stronger impact on the scores.

In Table 1 we report a comparison of fOGDA-VI with LA-GDA in terms of Fréchet Inception Distance (FID; lower is better) (Heusel et al., 2017) and Inception Score (IS; higher is better) (Salimans et al., 2016). We report the best obtained scores, averaged over 5 runs with 500,000 iterations each. For both methods we evaluated the last (non averaged) iterates, the uniform average, the exponential moving average (EMA) and the EMA on the "slow weights" for LA-GDA. One can see that while both considered methods give comparable results and show similar behaviour, the best score for both FID and IS is obtained by LA-GDA. An interesting observation is that the FID scores obtained by LA-GDA seem to be significantly worse than those reported in (Chavdarova et al., 2021b), while it exhibits higher reported IS values.

The scores reported in Table 2, where we list the overall best obtained values for both metrics, support the observations from Table 1. In general, the results for fOGDA-VI and LA-GDA are on a similar level with the latter giving the altogether best scores.

Figure 3 shows all five individual runs and the respective median for both methods in terms of (a) FID and (b) IS. It can be observed that LA-GDA achieves better results than fOGDA-VI during the first 200,000 iterations, while from then on the both methods achieve similar scores. As it appears, the medians of fOGDA-VI seem to stay more consistently on the level of the optimal scores while LA-GDA worsens again over time.

5. Conclusion

In this work we proposed a novel algorithm, called fOGDA-VI, to solve monotone variational inequalities that recovers the explicit fOGDA method from (Boţ et al., 2022) in the unconstrained case. We showed that fOGDA-VI exhibits a better rate of convergence than other accelerated methods, giving a rate of convergence like $o(1/k)$ in terms of the restricted gap function and the tangent and natural residuals, while still maintaining convergence of the iterates to a solution of the variational inequality under investigation. To validate our method in practice we treated a constrained bilinear minimax problem for which we obtained superior behaviour on this theoretically justified task. Moreover, application of fOGDA-VI to the training of GANs gives promising results even in practical settings that do not warrant the required assumptions.

Acknowledgements

The authors would like to thank the anonymous reviewers and the program chairs for their valuable suggestions and comments that have improved the quality of the paper.

Michael Sedlmayer would like to acknowledge support from the Austrian Research Promotion Agency (FFG), project "Smart operation of wind turbines under icing conditions (SOWINDIC)". Dang-Khoa Nguyen would like to acknowledge support from the Austrian Science Fund (FWF), project P 34922. Radu Ioan Boţ would like to acknowledge partial support from the Austrian Science Fund (FWF), projects W 1260 and P 34922.

References

Antipin, A. S. On a method for convex programs using a symmetrical modification of the lagrange function. *Ekonomika i Matematicheskie Metody*, 12(6):1164–1173, 1976.

Banert, S. and Boţ, R. I. A forward-backward-forward differential equation and its asymptotic properties. *Journal of Convex Analysis*, 25(2):371–388, 2018.

Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, New York, 2011.

Bauschke, H. H. and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces, Second Edition*. Springer, Cham, 2017.

Böhm, A., Sedlmayer, M., Csetnek, E. R., and Boţ, R. I. Two steps at a time—taking GAN training in stride with Tseng's method. *SIAM Journal on Mathematics of Data Science*, 4(2):750–771, 2022.

Boţ, R. I. and Nguyen, D.-K. Fast Krasnosel'skiĭ-Mann algorithm with a convergence rate of the fixed point iteration of $o(1/k)$. *arXiv preprint arXiv:2206.09462*, 2022.

Boţ, R. I., Csetnek, E. R., and Nguyen, D.-K. Fast OGDA in continuous and discrete time. *arXiv preprint arXiv:2203.10947*, 2022.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Cai, Y. and Zheng, W. Accelerated single-call methods for constrained min-max optimization. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

Cai, Y., Oikonomou, A., and Zheng, W. Accelerated algorithms for monotone inclusions and constrained nonconvex-nonconcave min-max optimization. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022a.

Cai, Y., Oikonomou, A., and Zheng, W. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022b.

Chavdarova, T., Jordan, M. I., and Zampetakis, M. Last-iterate convergence of saddle point optimizers via high-resolution differential equations. In *The 13th Annual Workshop on Optimization for Machine Learning (OPT2021)*, 2021a.

Chavdarova, T., Pagliardini, M., Stich, S. U., Fleuret, F., and Jaggi, M. Taming GANs with lookahead-minmax. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=ZW0yXJyNmoG.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pp. 9236–9246, 2018.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with optimism. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJJySbbAZ.

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

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on Generative Adversarial Networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1laEnA5Ym.

Golowich, N., Pattathil, S., and Daskalakis, C. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 33:20766–20778, 2020a.

Golowich, N., Pattathil, S., Daskalakis, C., and Ozdaglar, A. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *Conference on Learning Theory*, pp. 1758–1784. PMLR, 2020b.

Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv:1701.00160*, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Gorbunov, E., Loizou, N., and Gidel, G. Extragradient method: $\mathcal{O}(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022.

Halpern, B. Fixed points of nonexpanding maps. *Bulletin of the American Mathematical Society*, 73(6):957–961, 1967.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

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

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

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Malitsky, Y. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

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

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2703–2717. SIAM, 2018.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of GANs. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/4588e674d3f0faf985047d4c3f13ed0d-Paper.pdf.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, 2018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Mokhtari, A., Ozdaglar, A. E., and Pattathil, S. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020.

Morgenstern, O. and Von Neumann, J. *Theory of games and economic behavior*. Princeton university press, 1953.

Nemirovski, A. Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Omidshafiei, S., Pazis, J., Amato, C., How, J. P., and Vian, J. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pp. 2681–2690. PMLR, 2017.

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

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep

learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Popov, L. D. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Tran-Dinh, Q. The connection between Nesterov's accelerated methods and Halpern fixed-point iterations. *arXiv preprint arXiv:2203.04869*, 2022.

Tran-Dinh, Q. and Luo, Y. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150*, 2021.

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

Yoon, T. and Ryu, E. K. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *International Conference on Machine Learning*, pp. 12098–12109. PMLR, 2021.

A. Auxiliary Results

In the following we present auxiliary results that will be necessary for the convergence analysis in Appendix B.

We start this section with a result which characterises the convergence of sequences.

Lemma A.1 (see Lemma 32 in (Boţ et al., 2022)). *Let $a \geq 1$ and $(q_k)_{k\geq 1}$ be a bounded sequence in $\mathbb{R}^d$ such that*

$$\lim_{k\to+\infty}\left(q_{k+1}+\frac{k}{a}\left(q_{k+1}-q_k\right)\right)=p\in\mathbb{R}^d.$$

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

The following result is a particular instance of Lemma 5.31 in (Bauschke & Combettes, 2017).

Lemma A.2. *Let $(a_k)_{k\geq 1}$, $(b_k)_{k\geq 1}$, $(d_k)_{k\geq 1}$ and $(d_k)_{k\geq 1}$ be sequences of real numbers. Assume that $(a_k)_{k\geq 1}$ is bounded from below, $(b_k)_{k\geq 1}$ and $(d_k)_{k\geq 1}$ are nonnegative sequences such that $\sum_{k\geq 1} d_k < +\infty$. If*

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

then the following statements are true:

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

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

To show convergence of the sequence of iterates we will use the following result, which is a finite dimensional version of the so-called *Opial Lemma* (Opial, 1967).

Lemma A.3. *Let $S \subseteq \mathbb{R}^d$ be a nonempty set and $(z_k)_{k\geq 1} \subseteq \mathbb{R}^d$ a sequence such that the following two conditions hold:*

(i) for every $z^* \in S$, $\lim_{k\to+\infty} \|z_k - z^*\|$ exists;

(ii) every cluster point of $(z_k)_{k\geq 1}$ belongs to S.

Then $(z_k)_{k\geq 0}$ converges to an element in S.

The following result about maximal monotone operators will be crucial in the convergence analysis to verify item (ii) from Lemma A.3.

Proposition A.4 (see Proposition 20.33 in (Bauschke & Combettes, 2011)). *Let $A : \mathbb{R}^d \to 2^{\mathbb{R}^d}$ be maximal monotone, with $\operatorname{gra} A$ denoting the graph of A. Then $\operatorname{gra} A$ is closed, i.e., for every sequence $(z_k, v_k)_{k\geq 1}$ in $\operatorname{gra} A$ and every $(z, v) \in \mathbb{R}^d \times \mathbb{R}^d$, if $z_k \to z$ and $v_k \to v$, then $(z, v) \in \operatorname{gra} A$.*

Lemma A.5. *Let $a, b, c \in \mathbb{R}$ be such that $a \neq 0$ and $b^2 - ac \leq 0$. The following statements are true:*

(i) if $a > 0$, then

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

(ii) if $a < 0$, then

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

B. Convergence Analysis

In the following section we will establish the necessary convergence analysis to prove the main theoretical results of this work Theorem 3.2 and Theorem 3.3.

B.1. Notation and Preliminary Considerations

We start with proving the fact, that indeed

$$\texttt{Res}(z) \leq r(z) \quad \forall z \in \mathbb{R}^d.$$

To this end is sufficient to only consider the case when $z \in C$ as otherwise the inequality trivially holds. Let $z \in C$ and $\zeta \in N_C(z)$, then by the following equivalence (see Proposition 6.45 in (Bauschke & Combettes, 2011))

$$p = P_C(v) \quad \Leftrightarrow \quad \exists \zeta \in N_C(p) \text{ such that } v - p = \zeta, \tag{13}$$

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

$$\texttt{Res}(z) = \|z - P_C\left[z - F(z)\right]\| = \|P_C\left[z + \zeta\right] - P_C\left[z - F(z)\right]\| \leq \|F(z) + \zeta\|. \tag{14}$$

Since $\zeta \in N_C(z)$ was arbitrary, we conclude that $\texttt{Res}(z) \leq r(z)$.

The characterisation (13) can be also used to deduce a projection step from the second line of (10). For all $k \geq 1$ we have

$$z_{k+1} = w_k - \gamma\left(1 + \frac{k}{k+\alpha}\right)(F(w_k) - F(w_{k-1}) + \zeta_{k+1} - \zeta_k)$$

with $\zeta_k \in N_C(z_k)$, which using (13) is equivalent to

$$z_{k+1} = P_C\left[w_k - \gamma\left(1 + \frac{k}{k+\alpha}\right)(F(w_k) - F(w_{k-1}) - \zeta_k)\right],$$

using that for $\lambda > 0$

$$\zeta \in N_C(z) \quad \Leftrightarrow \quad \lambda\zeta \in N_C(z).$$

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

$$v_k := F(w_{k-1}) + \zeta_k, \tag{15}$$

which we plug into Algorithm 1 to obtain

$$w_k = z_k + \frac{k}{k+\alpha}\Big(z_k - z_{k-1}\Big) - \gamma\frac{\alpha}{k+\alpha}v_k, \tag{16a}$$

$$z_{k+1} = w_k - \gamma\left(1 + \frac{k}{k+\alpha}\right)\Big(v_{k+1} - v_k\Big). \tag{16b}$$

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

$$\gamma = \frac{1-\varepsilon}{4L}. \tag{17}$$

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

$$\begin{aligned}\|\zeta_{k+1} + F\left(z_{k+1}\right) - v_{k+1}\| &= \|F\left(z_{k+1}\right) - F\left(w_k\right)\| \leq L\left\|z_{k+1} - w_k\right\| \\ &\leq \gamma L\left\|v_{k+1} - v_k\right\| \leq \frac{1-\varepsilon}{4}\left\|v_{k+1} - v_k\right\| \leq \frac{1}{4}\left\|v_{k+1} - v_k\right\| \leq \left\|v_{k+1} - v_k\right\|.\end{aligned} \tag{18}$$

By summing up (16a) and (16b) we find for every $k \geq 1$

$$(k+\alpha)\left(z_{k+1} - z_k\right) - k\left(z_k - z_{k-1}\right) = -\alpha\gamma v_{k+1} - 2\gamma k\left(v_{k+1} - v_k\right). \tag{19}$$

Let $0 \leq \lambda \leq \alpha - 1$ and $z^* \in \Omega$. Following (Boţ et al., 2022) we denote for every $k \geq 1$

$$u_{\lambda,k} := 2\lambda (z_k - z^*) + 2k (z_k - z_{k-1}) + \frac{3\alpha - 2}{\alpha - 1}\gamma k v_k, \tag{20}$$

$$\begin{aligned}\mathcal{E}_{\lambda,k} := {} & \frac{1}{2}\|u_{\lambda,k}\|^2 + 2\lambda(\alpha - 1 - \lambda)\|z_k - z^*\|^2 + \frac{2(\alpha-2)}{\alpha-1}\lambda\gamma k \langle z_k - z^*, v_k\rangle \\ & + \frac{\alpha-2}{\alpha-1}\gamma^2 k\left(\frac{3\alpha-2}{2(\alpha-1)}k + \alpha\right)\|v_k\|^2,\end{aligned} \tag{21}$$

$$\begin{aligned}\mathcal{G}_{\lambda,k} := {} & \mathcal{E}_{\lambda,k} - \frac{2(\alpha-2)}{\alpha-1}\gamma k^2 \langle z_k - z_{k-1}, F(z_k) - F(w_{k-1})\rangle \\ & + \frac{\alpha-2}{\alpha-1}\gamma^2 k\sqrt{k}\left((1-\varepsilon)\sqrt{k} + \alpha\right)\|v_k - v_{k-1}\|^2.\end{aligned} \tag{22}$$

B.2. Properties of the Energy Functions

We collect the properties of the energy functions $(\mathcal{E}_{\lambda,k})_{k\geq 1}$ and $(\mathcal{G}_{\lambda,k})_{k\geq 1}$ in the following results. Please note that the statement of the first lemma can be deduced from eq. (81) in (Boţ et al., 2022), taking into account an appropriate formal correspondence between certain quantities. For better comprehensibility and in order to be able to refer to equations later on, we provide its proof nevertheless.

Lemma B.1. *Let $z^* \in \Omega$ and $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 1. For $0 \leq \lambda \leq \alpha - 1$, the following identity holds for every $k \geq 1$*

$$\begin{aligned}&\mathcal{E}_{\lambda,k+1} - \mathcal{E}_{\lambda,k} \\ &\quad = -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle \\ &\qquad + 2(\lambda + 1 - \alpha)(2k + \alpha + 1)\|z_{k+1} - z_k\|^2 \\ &\qquad + \frac{2}{\alpha-1}\Big(4(\alpha-1)(\lambda+1-\alpha) - \alpha(\alpha-2)\Big)\gamma k\langle z_{k+1} - z_k, v_{k+1}\rangle \\ &\qquad + \frac{2}{\alpha-1}\Big(2\alpha(\alpha-1)(\lambda+1-\alpha) + \alpha - 2(\alpha-1)^2\Big)\gamma\langle z_{k+1} - z_k, v_{k+1}\rangle \\ &\qquad - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1} - z_k, v_{k+1} - v_k\rangle - \frac{\alpha-2}{\alpha-1}\gamma^2 k(2k+\alpha)\|v_{k+1} - v_k\|^2 \\ &\qquad - \frac{\alpha-2}{\alpha-1}\gamma^2\left(2(3\alpha-2)k + 2\alpha^2 + \alpha - 2\right)\|v_{k+1}\|^2.\end{aligned} \tag{23}$$

Proof. Recall that by the definition in (20), we have for every $k \geq 1$

$$u_{\lambda,k} = 2\lambda(z_k - z^*) + 2k(z_k - z_{k-1}) + \frac{3\alpha-2}{\alpha-1}\gamma k v_k.$$

Similarly,

$$u_{\lambda,k+1} = 2\lambda(z_{k+1} - z^*) + 2(k+1)(z_{k+1} - z_k) + \frac{3\alpha-2}{\alpha-1}\gamma(k+1)v_{k+1}. \tag{24}$$

Thus, after subtraction we deduce from (19) that

$$\begin{aligned}&u_{\lambda,k+1} - u_{\lambda,k} \\ &\quad = 2(\lambda+1-\alpha)(z_{k+1} - z_k) + 2(k+\alpha)(z_{k+1} - z_k) - 2k(z_k - z_{k-1}) \\ &\qquad + \frac{3\alpha-2}{\alpha-1}\gamma v_{k+1} + \frac{3\alpha-2}{\alpha-1}\gamma k(v_{k+1} - v_k) \\ &\quad = 2(\lambda+1-\alpha)(z_{k+1} - z_k) + \frac{\alpha - 2(\alpha-1)^2}{\alpha-1}\gamma v_{k+1} + \frac{2-\alpha}{\alpha-1}\gamma k(v_{k+1} - v_k).\end{aligned} \tag{25}$$

For $k \geq 1$ we know that

$$\frac{1}{2}\left(\|u_{\lambda,k+1}\|^2 - \|u_{\lambda,k}\|^2\right) = \langle u_{\lambda,k+1}, u_{\lambda,k+1} - u_{\lambda,k}\rangle - \frac{1}{2}\|u_{\lambda,k+1} - u_{\lambda,k}\|^2, \tag{26}$$

and that for every $k \geq 0$

$$\begin{aligned} &2\lambda\left(\alpha-1-\lambda\right)\left(\left\|z_{k+1}-z^*\right\|^2-\left\|z_k-z^*\right\|^2\right) \\ &\quad=4\lambda\left(\alpha-1-\lambda\right)\left\langle z_{k+1}-z^*, z_{k+1}-z_k\right\rangle-2\lambda\left(\alpha-1-\lambda\right)\left\|z_{k+1}-z_k\right\|^2. \end{aligned} \tag{27}$$

We use the relations (24) and (25) to derive for every $k \geq 1$

$$\begin{aligned} &\left\langle u_{\lambda,k+1}, u_{\lambda,k+1}-u_{\lambda,k}\right\rangle \\ &\quad=4\lambda\left(\lambda+1-\alpha\right)\left\langle z_{k+1}-z^*, z_{k+1}-z_k\right\rangle \\ &\qquad+\frac{2}{\alpha-1}\left(\alpha-2\left(\alpha-1\right)^2\right)\lambda\gamma\left\langle z_{k+1}-z^*, v_{k+1}\right\rangle \\ &\qquad-\frac{2\left(\alpha-2\right)}{\alpha-1}\lambda\gamma k\left\langle z_{k+1}-z^*, v_{k+1}-v_k\right\rangle+4\left(\lambda+1-\alpha\right)\left(k+1\right)\left\|z_{k+1}-z_k\right\|^2 \\ &\qquad+\frac{2}{\alpha-1}\Big(\left(3\alpha-2\right)\left(\lambda+1-\alpha\right)+\alpha-2\left(\alpha-1\right)^2\Big)\gamma\left(k+1\right)\left\langle z_{k+1}-z_k, v_{k+1}\right\rangle \\ &\qquad-\frac{2\left(\alpha-2\right)}{\alpha-1}\gamma\left(k+1\right)k\left\langle z_{k+1}-z_k, v_{k+1}-v_k\right\rangle \\ &\qquad+\frac{1}{\left(\alpha-1\right)^2}\left(\alpha-2\left(\alpha-1\right)^2\right)\left(3\alpha-2\right)\gamma^2\left(k+1\right)\left\|v_{k+1}\right\|^2 \\ &\qquad-\frac{1}{\left(\alpha-1\right)^2}\left(\alpha-2\right)\left(3\alpha-2\right)\gamma^2\left(k+1\right)k\left\langle v_{k+1}, v_{k+1}-v_k\right\rangle, \end{aligned} \tag{28}$$

and

$$\begin{aligned} -\frac{1}{2}\left\|u_{\lambda,k+1}-u_{\lambda,k}\right\|^2 &= -2\left(\lambda+1-\alpha\right)^2\left\|z_{k+1}-z_k\right\|^2 \\ &\quad-\frac{2}{\alpha-1}\left(\alpha-2\left(\alpha-1\right)^2\right)\left(\lambda+1-\alpha\right)\gamma\left\langle z_{k+1}-z_k, v_{k+1}\right\rangle \\ &\quad-\frac{1}{2\left(\alpha-1\right)^2}\left(\alpha-2\left(\alpha-1\right)^2\right)^2\gamma^2\left\|v_{k+1}\right\|^2-\frac{\left(\alpha-2\right)^2}{2\left(\alpha-1\right)^2}\gamma^2k^2\left\|v_{k+1}-v_k\right\|^2 \\ &\quad+\frac{2(\alpha-2)}{\alpha-1}\left(\lambda+1-\alpha\right)\gamma k\left\langle z_{k+1}-z_k, v_{k+1}-v_k\right\rangle \\ &\quad+\frac{\alpha-2}{\left(\alpha-1\right)^2}\left(\alpha-2\left(\alpha-1\right)^2\right)\gamma^2k\left\langle v_{k+1}, v_{k+1}-v_k\right\rangle. \end{aligned} \tag{29}$$

After some algebra, we see that

$$\begin{aligned} &\Big(\left(3\alpha-2\right)\left(\lambda+1-\alpha\right)+\alpha-2\left(\alpha-1\right)^2\Big)\left(k+1\right)-\left(\lambda+1-\alpha\right)\left(\alpha-2\left(\alpha-1\right)^2\right) \\ &\quad=\Big(\left(3\alpha-2\right)\left(\lambda+1-\alpha\right)+\alpha-2\left(\alpha-1\right)^2\Big)k+2\alpha\left(\alpha-1\right)\left(\lambda+1-\alpha\right) \\ &\qquad+\alpha-2\left(\alpha-1\right)^2 \\ &\quad=\Big(\left(3\alpha-2\right)\left(\lambda+1-\alpha\right)+\alpha-2\left(\alpha-1\right)^2+\left(\alpha-2\right)\lambda\Big)k-\left(\alpha-2\right)\lambda k \\ &\qquad+2\alpha\left(\alpha-1\right)\left(\lambda+1-\alpha\right)+\alpha-2\left(\alpha-1\right)^2 \\ &\quad=\Big(4\left(\alpha-1\right)\left(\lambda+1-\alpha\right)-\alpha\left(\alpha-2\right)\Big)k-\left(\alpha-2\right)\lambda k \\ &\qquad+2\alpha\left(\alpha-1\right)\left(\lambda+1-\alpha\right)+\alpha-2\left(\alpha-1\right)^2. \end{aligned} \tag{30}$$

By plugging (28) and (29) into (26), and by taking into consideration the relation (30), we get for every $k \geq 1$

$$\begin{aligned}
\frac{1}{2}\left(\|u_{\lambda,k+1}\|^2 - \|u_{\lambda,k}\|^2\right) &= 4\lambda\left(\lambda+1-\alpha\right)\langle z_{k+1}-z^*, z_{k+1}-z_k\rangle \\
&+ \frac{2}{\alpha-1}\left(\alpha - 2\left(\alpha-1\right)^2\right)\lambda\gamma\langle z_{k+1}-z^*, v_{k+1}\rangle \\
&- \frac{2\left(\alpha-2\right)}{\alpha-1}\lambda\gamma k\langle z_{k+1}-z^*, v_{k+1}-v_k\rangle \\
&+ 2\left(\lambda+1-\alpha\right)\left(2k+\alpha+1-\lambda\right)\|z_{k+1}-z_k\|^2 \\
&+ \frac{2}{\alpha-1}\Big(4\left(\alpha-1\right)\left(\lambda+1-\alpha\right) - \alpha\left(\alpha-2\right) - \left(\alpha-2\right)\lambda\Big)\gamma k\langle z_{k+1}-z_k, v_{k+1}\rangle \\
&+ \frac{2}{\alpha-1}\Big(2\alpha\left(\alpha-1\right)\left(\lambda+1-\alpha\right) + \alpha - 2\left(\alpha-1\right)^2\Big)\gamma\langle z_{k+1}-z_k, v_{k+1}\rangle \\
&- \frac{2\left(\alpha-2\right)}{\alpha-1}\gamma k\left(k+\alpha-\lambda\right)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle - \frac{\left(\alpha-2\right)^2}{2\left(\alpha-1\right)^2}\gamma^2k^2\|v_{k+1}-v_k\|^2 \\
&+ \frac{1}{2\left(\alpha-1\right)^2}\left(\alpha - 2\left(\alpha-1\right)^2\right)\gamma^2\left(2\left(3\alpha-2\right)k + 2\alpha^2+\alpha-2\right)\|v_{k+1}\|^2 \\
&- \frac{\alpha-2}{\left(\alpha-1\right)^2}\gamma^2k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\langle v_{k+1}, v_{k+1}-v_k\rangle.
\end{aligned} \tag{31}$$

Furthermore, one can show that for every $k \geq 1$ we get

$$\begin{aligned}
&\left(k+1\right)\langle z_{k+1}-z^*, v_{k+1}\rangle - k\langle z_k - z^*, v_k\rangle \\
&\quad= \langle z_{k+1}-z^*, v_{k+1}\rangle + k\left(\langle z_{k+1}-z^*, v_{k+1}\rangle - \langle z_k-z^*, v_k\rangle\right) \\
&\quad= \langle z_{k+1}-z^*, v_{k+1}\rangle + k\langle z_{k+1}-z^*, v_{k+1}-v_k\rangle \\
&\qquad- k\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle + k\langle z_{k+1}-z_k, v_{k+1}\rangle
\end{aligned} \tag{32}$$

and

$$\begin{aligned}
&\left(k+1\right)\Big(\left(3\alpha-2\right)\left(k+1\right) + 2\alpha\left(\alpha-1\right)\Big)\|v_{k+1}\|^2 - k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\|v_k\|^2 \\
&\quad= \left(2\left(3\alpha-2\right)k + 2\alpha^2+\alpha-2\right)\|v_{k+1}\|^2 \\
&\qquad+ k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\left(\|v_{k+1}\|^2 - \|v_k\|^2\right) \\
&\quad= \left(2\left(3\alpha-2\right)k + 2\alpha^2+\alpha-2\right)\|v_{k+1}\|^2 \\
&\qquad+ 2k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\langle v_{k+1}, v_{k+1}-v_k\rangle \\
&\qquad- k\Big(\left(3\alpha-2\right)k + 2\alpha\left(\alpha-1\right)\Big)\|v_{k+1}-v_k\|^2.
\end{aligned} \tag{33}$$

In addition, direct computations show that

$$\alpha - 2\left(\alpha-1\right)^2 + \alpha - 2 = -2\left(\alpha-1\right)\left(\alpha-2\right)$$

and

$$-\left(\alpha-2\right)^2 - \left(\alpha-2\right)\left(3\alpha-2\right) = -4\left(\alpha-2\right)\left(\alpha-1\right).$$

Hence, multiplying (32) by $2\lambda\gamma(\alpha-2)/(\alpha-1) \geq 0$ and (33) by $\gamma^2(\alpha-2)/2(\alpha-1)^2 > 0$, followed by summing up the resulting identities with (27) and (31), yields (23) for every $k \geq 1$. □

Lemma B.2. *Let $z^* \in \Omega$ and $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 1. For $0 \leq \lambda \leq \alpha - 1$, the following statements are true:*

(i) for every $k \geq k_0 := \max\left\{2, \left\lceil \frac{1}{\alpha-2} \right\rceil\right\}$ *the following holds:*

$$\begin{aligned}
\mathcal{G}_{\lambda,k+1} - \mathcal{G}_{\lambda,k} \leq{}& \frac{(\alpha-1)(\alpha-2)}{\varepsilon(k+1)^2}\lambda^2 \|z_{k+1} - z^*\|^2 \\
&- 4(\alpha-2)\lambda\gamma \langle z_{k+1} - z^*, \zeta_{k+1} + F(z_{k+1})\rangle \\
&+ 4(\eta_0 k + \eta_1)\gamma \langle z_{k+1} - z_k, v_{k+1}\rangle + \left(\eta_2 k + \kappa_0\sqrt{k}\right)\|z_{k+1} - z_k\|^2 \\
&+ 4\left(\eta_3 k + \kappa_1\sqrt{k}\right)\gamma^2\|v_{k+1}\|^2 - \frac{\alpha-2}{\alpha-1}\mu_k\gamma^2\|v_{k+1} - v_k\|^2,
\end{aligned}$$

where

$$\begin{aligned}
\mu_k &:= (k+1)\left(\varepsilon(k+1) + \alpha^2\sqrt{k+1} + (\alpha-4)\right) - (\alpha-2), \\
\eta_0 &:= \frac{1}{2(\alpha-1)}\left(4(\alpha-1)(\lambda+1-\alpha) - \alpha(\alpha-2)\right) < 0, \\
\eta_1 &:= \frac{1}{2(\alpha-1)}\left(2\alpha(\alpha-1)(\lambda+1-\alpha) + \alpha - 2(\alpha-1)^2\right) < 0, \\
\eta_2 &:= 4(\lambda+1-\alpha) \leq 0, \\
\eta_3 &:= -\frac{1}{2(\alpha-1)}(\alpha-2)(3\alpha-2) < 0, \\
\kappa_0 &:= \frac{1}{\alpha-1}(\alpha-2)\sqrt{\alpha-2} > 0, \\
\kappa_1 &:= \frac{1}{4(\alpha-1)}(\alpha-2)\alpha > 0;
\end{aligned} \tag{34}$$

(ii) for every $k \geq 1$ *one has the following lower bound for the quantity* $\mathcal{G}_{\lambda,k}$

$$\begin{aligned}
\mathcal{G}_{\lambda,k} \geq{}& \frac{\alpha-2}{4(3\alpha-2)}\left\|4\lambda(z_k - z^*) + 2k(z_k - z_{k-1}) + \frac{2(3\alpha-2)}{\alpha-1}\gamma k v_k\right\|^2 \\
&+ \frac{(\alpha-2)^2}{4(3\alpha-2)(\alpha-1)}k^2\|z_k - z_{k-1}\|^2 + 2(\alpha-1)\lambda\left(1 - \frac{4\lambda}{3\alpha-2}\right)\|z_k - z^*\|^2.
\end{aligned} \tag{35}$$

Proof. (i) Let $k \geq 2$ be fixed. By the definition of $\mathcal{G}_{\lambda,k}$ in (22), we have for every $k \geq 2$

$$\begin{aligned}
&\mathcal{G}_{\lambda,k+1} - \mathcal{G}_{\lambda,k} \\
&\quad= \mathcal{E}_{\lambda,k+1} - \mathcal{E}_{\lambda,k} - \frac{2(\alpha-2)}{\alpha-1}\gamma\Big[(k+1)^2\langle z_{k+1} - z_k, F(z_{k+1}) - F(w_k)\rangle \\
&\qquad\qquad - k^2\langle z_k - z_{k-1}, F(z_k) - F(w_{k-1})\rangle\Big] \\
&\qquad + \frac{(\alpha-2)\alpha}{\alpha-1}\gamma^2\left[(k+1)\sqrt{k+1}\|v_{k+1} - v_k\|^2 - k\sqrt{k}\|v_k - v_{k-1}\|^2\right] \\
&\qquad + \frac{(\alpha-2)(1-\varepsilon)}{\alpha-1}\gamma^2\left[(k+1)^2\|v_{k+1} - v_k\|^2 - k^2\|v_k - v_{k-1}\|^2\right].
\end{aligned} \tag{36}$$

By using the definition of η_0, η_1, η_2 and η_3 in (34), for every $k \geq 1$ we deduce from (23) that

$$\begin{aligned}
&\mathcal{E}_{\lambda,k+1} - \mathcal{E}_{\lambda,k} \\
&\quad = -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle + \Big(\eta_2 k + 2(\lambda+1-\alpha)(\alpha+1)\Big)\|z_{k+1} - z_k\|^2 \\
&\qquad + 4(\eta_0 k + \eta_1)\gamma\langle z_{k+1} - z_k, v_{k+1}\rangle - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1} - z_k, v_{k+1} - v_k\rangle \\
&\qquad - \frac{\alpha-2}{\alpha-1}k(2k+\alpha)\gamma^2\|v_{k+1} - v_k\|^2 + \left(4\eta_3 k - \frac{\alpha-2}{\alpha-1}\left(2\alpha^2+\alpha-2\right)\right)\gamma^2\|v_{k+1}\|^2 \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1} - z_k, v_{k+1} - v_k\rangle \\
&\qquad + \Big(4(\eta_0 k + \eta_1)\gamma\langle z_{k+1} - z_k, v_{k+1}\rangle + \eta_2 k\|z_{k+1} - z_k\|^2 + 4\eta_3 k\gamma^2\|v_{k+1}\|^2\Big) \\
&\qquad - \frac{\alpha-2}{\alpha-1}k(2k+\alpha)\gamma^2\|v_{k+1} - v_k\|^2,
\end{aligned} \tag{37}$$

where the inequality comes from the fact that $0 \leq \lambda \leq \alpha - 1$ and $\alpha > 2$. Plugging (37) into (36) yields for every $k \geq 2$

$$\begin{aligned}
&\mathcal{G}_{\lambda,k+1} - \mathcal{G}_{\lambda,k} \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle - \frac{\alpha-2}{\alpha-1}\left(2k^2+\alpha k\right)\gamma^2\|v_{k+1} - v_k\|^2 \\
&\qquad - \frac{2(\alpha-2)}{\alpha-1}\gamma\Big[(k+1)^2\langle z_{k+1} - z_k, F(z_{k+1}) - F(w_k)\rangle \\
&\qquad\qquad - k^2\langle z_k - z_{k-1}, F(z_k) - F(w_{k-1})\rangle\Big] \\
&\qquad + \frac{(\alpha-2)\alpha}{\alpha-1}\gamma^2\Big[(k+1)\sqrt{k+1}\|v_{k+1} - v_k\|^2 - k\sqrt{k}\|v_k - v_{k-1}\|^2\Big] \\
&\qquad + \frac{(\alpha-2)(1-\varepsilon)}{\alpha-1}\gamma^2\Big[(k+1)^2\|v_{k+1} - v_k\|^2 - k^2\|v_k - v_{k-1}\|^2\Big] \\
&\qquad + \Big(4(\eta_0 k + \eta_1)\gamma\langle z_{k+1} - z_k, v_{k+1}\rangle + \eta_2 k\|z_{k+1} - z_k\|^2 + 4\eta_3 k\gamma^2\|v_{k+1}\|^2\Big) \\
&\qquad - \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1} - z_k, v_{k+1} - v_k\rangle.
\end{aligned} \tag{38}$$

Our next aim is to derive upper estimates for the first two terms on the right-hand side of (38), which will eventually simplify the subsequent three terms. From the Cauchy-Schwarz inequality and (18) we have for every $k \geq 1$

$$\begin{aligned}
&-4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, v_{k+1}\rangle = -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, \zeta_{k+1} + F(w_k)\rangle \\
&\quad = -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, \zeta_{k+1} + F(z_{k+1})\rangle \\
&\qquad + 4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, F(z_{k+1}) - F(w_k)\rangle \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, \zeta_{k+1} + F(z_{k+1})\rangle \\
&\qquad + 4(\alpha-2)\lambda\gamma\|z_{k+1} - z^*\|\,\|F(z_{k+1}) - F(w_k)\| \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, \zeta_{k+1} + F(z_{k+1})\rangle \\
&\qquad + 2(\alpha-2)\lambda\gamma\|z_{k+1} - z^*\|\,\|v_{k+1} - v_k\| \\
&\quad \leq -4(\alpha-2)\lambda\gamma\langle z_{k+1} - z^*, \zeta_{k+1} + F(z_{k+1})\rangle + \frac{(\alpha-1)(\alpha-2)}{\varepsilon(k+1)^2}\lambda^2\|z_{k+1} - z^*\|^2 \\
&\qquad + \frac{\alpha-2}{\alpha-1}\varepsilon\gamma^2(k+1)^2\|v_{k+1} - v_k\|^2.
\end{aligned} \tag{39}$$

For $\zeta_k \in N_C(z_k)$ and $\zeta_{k+1} \in N_C(z_{k+1})$, the monotonicity of N_C and F, together with the relation (19) and the fact that for every $k \geq 1$

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

yield for every $k \geq 1$

$$\begin{aligned}
&-\frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle \\
&\quad\leq \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\left\langle z_{k+1}-z_k, \Big(\zeta_{k+1}+F(z_{k+1})-v_{k+1}\Big)-\Big(\zeta_k+F(z_k)-v_k\Big)\right\rangle \\
&\quad= \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\left\langle z_{k+1}-z_k, \Big(F(z_{k+1})-F(w_k)\Big)-\Big(F(z_k)-F(w_{k-1})\Big)\right\rangle \\
&\quad= \frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad-\frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, F(z_k)-F(w_{k-1})\rangle \\
&\quad= \frac{2(\alpha-2)}{\alpha-1}\gamma (k+1)^2\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad-\frac{2(\alpha-2)}{\alpha-1}\gamma k^2\langle z_k-z_{k-1}, F(z_k)-F(w_{k-1})\rangle \\
&\qquad+\frac{2(\alpha-2)}{\alpha-1}\gamma\big((\alpha-2)k-1\big)\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad+\frac{2(\alpha-2)}{\alpha-1}\alpha\gamma^2 k\langle v_{k+1}, F(z_k)-F(w_{k-1})\rangle \\
&\qquad+\frac{4(\alpha-2)}{\alpha-1}\gamma^2 k^2\langle v_{k+1}-v_k, F(z_k)-F(w_{k-1})\rangle.
\end{aligned} \tag{40}$$

By Young's inequality together with (18) for every $k \geq \left\lceil \frac{1}{\alpha-2} \right\rceil$ we obtain

$$\begin{aligned}
&\frac{2(\alpha-2)}{\alpha-1}\gamma\big((\alpha-2)k-1\big)\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\Big(\sqrt{(\alpha-2)k-1}\,\|z_{k+1}-z_k\|^2 \\
&\qquad\qquad+\gamma^2\big((\alpha-2)k-1\big)\sqrt{(\alpha-2)k-1}\,\|F(z_{k+1})-F(w_k)\|^2\Big) \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\sqrt{(\alpha-2)k}\,\|z_{k+1}-z_k\|^2 \\
&\qquad+(\alpha-2)\sqrt{\alpha-1}\gamma^2(k+1)\sqrt{k+1}\,\|F(z_{k+1})-F(w_k)\|^2 \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\sqrt{(\alpha-2)k}\,\|z_{k+1}-z_k\|^2 \\
&\qquad+4(\alpha-2)\sqrt{\alpha-1}\gamma^4L^2(k+1)\sqrt{k+1}\,\|v_{k+1}-v_k\|^2 \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\sqrt{(\alpha-2)k}\,\|z_{k+1}-z_k\|^2+(\alpha-2)\alpha\gamma^2(k+1)\sqrt{k+1}\,\|v_{k+1}-v_k\|^2,
\end{aligned} \tag{41}$$

where in the second estimate we use the fact that $(\alpha-2)k-1 \leq (\alpha-1)(k+1)$, while in the last one we combine $\sqrt{\alpha-1} \leq \alpha$ and $\gamma L < 1/4 < 1$.

In addition, for every $k \geq 2$ we derive

$$\begin{aligned}
&\frac{2(\alpha-2)}{\alpha-1}\alpha\gamma^2 k\langle v_{k+1}, F(z_k)-F(w_{k-1})\rangle \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\,\|v_{k+1}\|^2+\frac{\alpha-2}{\alpha-1}\alpha\gamma^2 k\sqrt{k}\,\|F(z_k)-F(w_{k-1})\|^2 \\
&\quad\leq \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\,\|v_{k+1}\|^2+\frac{\alpha-2}{\alpha-1}\alpha\gamma^2 k\sqrt{k}\,\|v_k-v_{k-1}\|^2 \\
&\quad= \frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\,\|v_{k+1}\|^2+\frac{\alpha-2}{\alpha-1}\alpha\gamma^2(k+1)\sqrt{k+1}\,\|v_{k+1}-v_k\|^2 \\
&\qquad-\frac{\alpha-2}{\alpha-1}\alpha\gamma^2\Big[(k+1)\sqrt{k+1}\,\|v_{k+1}-v_k\|^2-k\sqrt{k}\,\|v_k-v_{k-1}\|^2\Big],
\end{aligned}$$

and, by using the Cauchy-Schwarz inequality and (18),

$$\begin{aligned}
&\frac{4(\alpha-2)}{\alpha-1}\gamma^2k^2\langle v_{k+1}-v_k, F(z_k)-F(w_{k-1})\rangle \\
&\quad\leq \frac{2(\alpha-2)}{\alpha-1}(1-\varepsilon)\gamma^2k^2\|v_{k+1}-v_k\|\|v_k-v_{k-1}\| \\
&\quad\leq \frac{\alpha-2}{\alpha-1}(1-\varepsilon)\gamma^2k^2\left(\|v_{k+1}-v_k\|^2+\|v_k-v_{k-1}\|^2\right) \\
&\quad= \frac{4(\alpha-2)}{\alpha-1}\gamma^3Lk^2\left(\|v_{k+1}-v_k\|^2+\|v_k-v_{k-1}\|^2\right) \\
&\quad\leq -\frac{4(\alpha-2)}{\alpha-1}\gamma^3L\left((k+1)^2\|v_{k+1}-v_k\|^2-k^2\|v_k-v_{k-1}\|^2\right) \\
&\qquad+\frac{8(\alpha-2)}{\alpha-1}\gamma^3L(k+1)^2\|v_{k+1}-v_k\|^2 \\
&\quad= -\frac{4(\alpha-2)}{\alpha-1}\gamma^3L\left((k+1)^2\|v_{k+1}-v_k\|^2-k^2\|v_k-v_{k-1}\|^2\right) \\
&\qquad+\frac{2(\alpha-2)}{\alpha-1}(1-\varepsilon)\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2,
\end{aligned} \tag{42}$$

where we want to recall that the first equality comes from (17). By plugging (41) and (42) into (40), then combining the result with (39), we get after rearranging the terms for every $k\geq k_0$

$$\begin{aligned}
&-4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, v_{k+1}\rangle-\frac{2(\alpha-2)}{\alpha-1}\gamma k(k+\alpha)\langle z_{k+1}-z_k, v_{k+1}-v_k\rangle \\
&\quad\leq \frac{2(\alpha-2)}{\alpha-1}\gamma\Big[(k+1)^2\langle z_{k+1}-z_k, F(z_{k+1})-F(w_k)\rangle \\
&\qquad\qquad -k^2\langle z_k-z_{k-1}, F(z_k)-F(w_{k-1})\rangle\Big] \\
&\qquad-\frac{\alpha-2}{\alpha-1}\alpha\gamma^2\left[(k+1)\sqrt{k+1}\|v_{k+1}-v_k\|^2-k\sqrt{k}\|v_k-v_{k-1}\|^2\right] \\
&\qquad-\frac{4(\alpha-2)}{\alpha-1}\gamma^3L\left[(k+1)^2\|v_{k+1}-v_k\|^2-k^2\|v_k-v_{k-1}\|^2\right] \\
&\qquad-4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle \\
&\qquad+\frac{\alpha-2}{\alpha-1}\left(\left(2k^2+\alpha k\right)-\mu_k\right)\gamma^2\|v_{k+1}-v_k\|^2 \\
&\qquad+\frac{1}{\alpha-1}(\alpha-2)\sqrt{(\alpha-2)k}\|z_{k+1}-z_k\|^2+\frac{\alpha-2}{\alpha-1}\alpha\gamma^2\sqrt{k}\|v_{k+1}\|^2 \\
&\qquad+\frac{1}{\varepsilon(k+1)^2}(\alpha-1)(\alpha-2)\lambda^2\|z_{k+1}-z^*\|^2,
\end{aligned} \tag{43}$$

where we set

$$\begin{aligned}
\mu_k &:= \left(2k^2+\alpha k\right)-\varepsilon(k+1)^2-(\alpha-1)\alpha(k+1)\sqrt{k+1}-\alpha(k+1)\sqrt{k+1} \\
&\quad-2(1-\varepsilon)(k+1)^2 \\
&= \varepsilon(k+1)^2+(\alpha-4)k-2-\alpha^2(k+1)\sqrt{k+1} \\
&= (k+1)\left(\varepsilon(k+1)+\alpha^2\sqrt{k+1}+\alpha-4\right)-(\alpha-2).
\end{aligned}$$

Finally, summing up (38) and (43), we obtain the desired estimate.

(ii) Observe that

$$\begin{aligned}
&\frac{2(\alpha-2)}{\alpha-1}\lambda\gamma k\langle z_k-z^*,v_k\rangle+\frac{(\alpha-2)(3\alpha-2)}{2(\alpha-1)^2}\gamma^2k^2\|v_k\|^2\\
&\quad=\frac{\alpha-2}{3\alpha-2}\left(\frac{2(3\alpha-2)}{\alpha-1}\lambda\gamma k\langle z_k-z^*,v_k\rangle+\frac{(3\alpha-2)^2}{2(\alpha-1)^2}\gamma^2k^2\|v_k\|^2\right)\\
&\quad=\frac{1}{3\alpha-2}(\alpha-2)\left(\frac{1}{2}\left\|2\lambda(z_k-z^*)+\frac{3\alpha-2}{\alpha-1}\gamma kv_k\right\|^2-2\lambda^2\|z_k-z^*\|^2\right).
\end{aligned}$$

By the definition of $u_{\lambda,k}$ in (20) and by using the identity

$$\|x\|^2+\|y\|^2=\frac{1}{2}\left(\|x+y\|^2+\|x-y\|^2\right)\quad\forall x,y\in\mathbb{R}^d,$$

we deduce that for every $k\geq 1$

$$\begin{aligned}
\mathcal{E}_{\lambda,k}&=\frac{1}{2}\|u_{\lambda,k}\|^2+2\lambda(\alpha-1-\lambda)\|z_k-z^*\|^2+\frac{2(\alpha-2)}{\alpha-1}\lambda\gamma k\langle z_k-z^*,v_k\rangle\\
&\quad+\frac{\alpha-2}{\alpha-1}\gamma^2k\left(\frac{1}{2(\alpha-1)}(3\alpha-2)k+\alpha\right)\|v_k\|^2\\
&=\frac{1}{2}\left\|2\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{3\alpha-2}{\alpha-1}\gamma kv_k\right\|^2\\
&\quad+2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2+\frac{\alpha-2}{\alpha-1}\alpha\gamma^2k\|v_k\|^2\\
&\quad+\frac{\alpha-2}{2(3\alpha-2)}\left\|2\lambda(z_k-z^*)+\frac{3\alpha-2}{\alpha-1}\gamma kv_k\right\|^2\\
&=\frac{\alpha}{3\alpha-2}\left\|2\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{3\alpha-2}{\alpha-1}\gamma kv_k\right\|^2\\
&\quad+2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2+\frac{\alpha-2}{\alpha-1}\alpha\gamma^2k\|v_k\|^2\\
&\quad+\frac{\alpha-2}{4(3\alpha-2)}\left\|4\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{2(3\alpha-2)}{\alpha-1}\gamma kv_k\right\|^2\\
&\quad+\frac{\alpha-2}{3\alpha-2}k^2\|z_k-z_{k-1}\|^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathcal{G}_{\lambda,k}&=\mathcal{E}_{\lambda,k}-\frac{2(\alpha-2)}{\alpha-1}\gamma k^2\langle z_k-z_{k-1},F(z_k)-F(w_{k-1})\rangle\\
&\quad+\frac{\alpha-2}{\alpha-1}\gamma^2k\sqrt{k}\left((1-\varepsilon)\sqrt{k}+\alpha\right)\|v_k-v_{k-1}\|^2\\
&\geq\frac{\alpha-2}{4(3\alpha-2)}\left\|4\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{2(3\alpha-2)}{\alpha-1}\gamma kv_k\right\|^2\\
&\quad+\frac{\alpha-2}{3\alpha-2}k^2\|z_k-z_{k-1}\|^2+2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2\\
&\quad-\frac{2(\alpha-2)}{\alpha-1}\gamma k^2\langle z_k-z_{k-1},F(z_k)-F(w_{k-1})\rangle\\
&\quad+\frac{\alpha-2}{\alpha-1}\gamma^2k\sqrt{k}\left(4\gamma L\sqrt{k}+\alpha\right)\|v_k-v_{k-1}\|^2.
\end{aligned}$$

Now we use relation (18) and apply Lemma A.5 with $(a,b,c) := \left(1/2, -2\gamma, 2\gamma/L\right)$ to verify that for every $k \geq 1$

$$\begin{aligned}
&\frac{1}{2}\left\|z_k - z_{k-1}\right\|^2 - 4\gamma\left\langle z_k - z_{k-1}, F\left(z_k\right) - F\left(w_{k-1}\right)\right\rangle + 8\gamma^3 L\left\|v_k - v_{k-1}\right\|^2 \\
&\quad \geq \frac{1}{2}\left\|z_k - z_{k-1}\right\|^2 - 4\gamma\left\langle z_k - z_{k-1}, F\left(z_k\right) - F\left(w_{k-1}\right)\right\rangle + \frac{2\gamma}{L}\left\|F\left(z_k\right) - F\left(w_{k-1}\right)\right\|^2 \\
&\quad \geq 0.
\end{aligned}$$

Combining the last two estimates, for every $k \geq 1$ one can easily conclude that

$$\begin{aligned}
\mathcal{G}_{\lambda,k} &\geq \frac{\alpha-2}{4\left(3\alpha-2\right)}\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha-2\right)}{\alpha-1}\gamma k v_k\right\|^2 \\
&\quad + \left(\alpha-2\right)\left(\frac{1}{3\alpha-2} - \frac{1}{4\left(\alpha-1\right)}\right)k^2\left\|z_k - z_{k-1}\right\|^2 \\
&\quad + 2\left(\alpha-1\right)\lambda\left(1 - \frac{4\lambda}{3\alpha-2}\right)\left\|z_k - z^*\right\|^2 \\
&= \frac{\alpha-2}{4\left(3\alpha-2\right)}\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha-2\right)}{\alpha-1}\gamma k v_k\right\|^2 \\
&\quad + \frac{\left(\alpha-2\right)^2}{4\left(3\alpha-2\right)\left(\alpha-1\right)}k^2\left\|z_k - z_{k-1}\right\|^2 + 2\left(\alpha-1\right)\lambda\left(1 - \frac{4\lambda}{3\alpha-2}\right)\left\|z_k - z^*\right\|^2,
\end{aligned}$$

which is the desired inequality. □

The following lemma will be helpful in the main proof.

Lemma B.3. *The following statements are true:*

(i) there exist two parameters

$$0 \leq \underline{\lambda}\left(\alpha\right) < \overline{\lambda}\left(\alpha\right) \leq \frac{3\alpha-2}{4} \tag{44}$$

such that for every λ satisfying $\underline{\lambda}\left(\alpha\right) < \lambda < \overline{\lambda}\left(\alpha\right)$ one can find an integer $k_\lambda \geq 1$ with the property that the following inequality holds for every $k \geq k_\lambda$

$$\begin{aligned}
R_k := &\sqrt{\frac{5\alpha-2}{2\left(3\alpha-2\right)}}\left(\eta_2 k + \kappa_0\sqrt{k}\right)\left\|z_{k+1} - z_k\right\|^2 + 4\gamma\left(\eta_0 k + \eta_1\right)\left\langle z_{k+1} - z_k, v_{k+1}\right\rangle \\
&+ 4\sqrt{\frac{5\alpha-2}{2\left(3\alpha-2\right)}}\gamma^2\left(\eta_3 k + \kappa_1\sqrt{k}\right)\left\|v_{k+1}\right\|^2 \leq 0;
\end{aligned} \tag{45}$$

(ii) there exists a positive integer k_ε such that for every $k \geq k_\varepsilon$ we have

$$\mu_k \geq \frac{\varepsilon}{2}\left(k+1\right)^2. \tag{46}$$

Proof. (i) For the quadratic expression in R_k we calculate

$$\begin{aligned}
\frac{\Delta_k'}{4\gamma^2} &:= \left(\eta_0 k + \eta_1\right)^2 - \frac{5\alpha-2}{2\left(3\alpha-2\right)}k\left(\eta_2\sqrt{k} + \kappa_0\right)\left(\eta_3\sqrt{k} + \kappa_1\right) \\
&= \left(\eta_0^2 - \frac{5\alpha-2}{2\left(3\alpha-2\right)}\eta_2\eta_3\right)k^2 - \frac{5\alpha-2}{2\left(3\alpha-2\right)}\left(\eta_2\kappa_1 + \kappa_0\eta_3\right)k\sqrt{k} \\
&\quad + \left(2\eta_0\eta_1 - \frac{5\alpha-2}{2\left(3\alpha-2\right)}\kappa_0\kappa_1\right)k + \eta_1^2.
\end{aligned}$$

Since $\left(\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3\right)k^2$ is the dominant term in the above polynomial, it suffices to guarantee that

$$\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3 < 0 \tag{47}$$

holds in order to ensure the existence of some integer $k_\lambda \geq 1$ such that $\Delta_k' \leq 0$ for every $k \geq k_\lambda$ and to obtain from here, due to Lemma A.5 (ii), that $R_k \leq 0$ for every $k \geq k_\lambda$.

It remains to show that there exists a choice of λ for which (47) is true. We set $\xi := \lambda + 1 - \alpha \leq 0$ and get

$$\begin{aligned}
\eta_0 &= \frac{1}{2(\alpha-1)}\left(4(\alpha-1)(\lambda+1-\alpha) - \alpha(\alpha-2)\right) \\
&= \frac{1}{2(\alpha-1)}\left(4(\alpha-1)\xi - \alpha(\alpha-2)\right), \\
\eta_2\eta_3 &= -\frac{2}{\alpha-1}(\alpha-2)(3\alpha-2)(\lambda+1-\alpha) = -\frac{2}{\alpha-1}(\alpha-2)(3\alpha-2)\xi.
\end{aligned}$$

This means that we have to guarantee that there exists a choice for ξ satisfying

$$\begin{aligned}
&\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3 \\
&= \frac{1}{4(\alpha-1)^2}\left(\left(4(\alpha-1)\xi - \alpha(\alpha-2)\right)^2 + 4(5\alpha-2)(\alpha-1)(\alpha-2)\xi\right) \\
&= \frac{1}{4(\alpha-1)^2}\left(16(\alpha-1)^2\xi^2 + 4(\alpha-1)(\alpha-2)(3\alpha-2)\xi + \alpha^2(\alpha-2)^2\right) < 0,
\end{aligned}$$

which is nothing else than

$$16(\alpha-1)^2\xi^2 + 4(\alpha-1)(\alpha-2)(3\alpha-2)\xi + \alpha^2(\alpha-2)^2 < 0. \tag{48}$$

A direct computation shows that

$$\Delta_\xi := 16(\alpha-1)^2(\alpha-2)^2\left((3\alpha-2)^2 - 4\alpha^2\right) = 16(\alpha-1)^2(\alpha-2)^3(5\alpha-2) > 0.$$

Hence, in order to get (48), we have to choose ξ between the two roots of the quadratic function arising in this formula, in other words

$$\begin{aligned}
\xi_1(\alpha) &:= \frac{1}{32(\alpha-1)^2}\left(-4(\alpha-1)(\alpha-2)(3\alpha-2) - \sqrt{\Delta_\xi}\right) \\
&= -\frac{1}{8(\alpha-1)}(\alpha-2)\left(3\alpha-2+\sqrt{(\alpha-2)(5\alpha-2)}\right) \\
&< \xi = \lambda + 1 - \alpha < \xi_2(\alpha) := \frac{1}{32(\alpha-1)^2}\left(-4(\alpha-1)(\alpha-2)(3\alpha-2) + \sqrt{\Delta_\xi}\right) \\
&= -\frac{1}{8(\alpha-1)}(\alpha-2)\left(3\alpha-2-\sqrt{(\alpha-2)(5\alpha-2)}\right).
\end{aligned}$$

Obviously $\xi_1(\alpha) < 0$ and from Vieta's formula $\xi_1(\alpha)\cdot\xi_2(\alpha) = \frac{\alpha^2(\alpha-2)^2}{16(\alpha-1)^2}$, it follows that we must have $\xi_2(\alpha) < 0$ as well.

Therefore, going back to λ, in order to be sure that $\eta_0^2 - \frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3 < 0$ this must be chosen such that

$$\alpha - 1 + \xi_1(\alpha) < \lambda < \alpha - 1 + \xi_2(\alpha).$$

Next we will show that

$$0 < \alpha - 1 - \frac{1}{8(\alpha-1)}(\alpha-2)(3\alpha-2) < \frac{3\alpha}{4} - \frac{1}{2}. \tag{49}$$

Indeed, the inequality on the left-hand side follows immediately, since

$$\begin{aligned}\alpha-1-\frac{1}{8(\alpha-1)}(\alpha-2)(3\alpha-2)&=\frac{1}{8(\alpha-1)}\left(5\alpha^2-8\alpha+4\right)\\&=\frac{1}{8(\alpha-1)}\left(\alpha^2+4(\alpha-1)^2\right)>0.\end{aligned}$$

Using this relation, one can notice that the inequality on the right hand side of (49) can be equivalently written as

$$5\alpha^2-8\alpha+4<2(\alpha-1)(3\alpha-2)\Leftrightarrow 0<\alpha^2-2\alpha=\alpha(\alpha-2),$$

which is true as $\alpha>2$.

From (49) we immediately deduce that

$$0<\alpha-1+\xi_2(\alpha)\quad\text{ and }\quad\alpha-1+\xi_1(\alpha)<\frac{3\alpha}{4}-\frac{1}{2}.$$

This allows us to choose $\underline{\lambda}<\overline{\lambda}$, where

$$\begin{aligned}\underline{\lambda}(\alpha)&:=\alpha-1+\xi_1(\alpha)\\&=\frac{1}{8(\alpha-1)}\alpha^2+\frac{1}{2}(\alpha-1)-\frac{1}{8(\alpha-1)}(\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)}\\\overline{\lambda}(\alpha)&:=\min\left\{\frac{3\alpha}{4}-\frac{1}{2},\alpha-1+\xi_2(\alpha)\right\}\\&=\min\left\{\frac{3\alpha}{4}-\frac{1}{2},\frac{1}{8(\alpha-1)}\alpha^2+\frac{1}{2}(\alpha-1)+\frac{1}{8(\alpha-1)}(\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)}\right\},\end{aligned}$$

since

$$\frac{1}{8(\alpha-1)}\alpha^2+\frac{1}{2}(\alpha-1)-\frac{1}{8(\alpha-1)}(\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)}>0.$$

Indeed, as $(\alpha-1)\sqrt{\alpha-1}>(\alpha-2)\sqrt{\alpha-2}$ and $4\sqrt{\alpha-1}>\sqrt{5\alpha-2}$ we can easily deduce that

$$\alpha^2+4(\alpha-1)^2>4(\alpha-1)^2>(\alpha-2)\sqrt{(\alpha-2)(5\alpha-2)}$$

and the claim follows.

In conclusion, choosing λ to satisfy $\underline{\lambda}(\alpha)<\lambda<\overline{\lambda}(\alpha)$, we have

$$\eta_0^2-\frac{5\alpha-2}{2(3\alpha-2)}\eta_2\eta_3<0$$

and therefore there exists some integer $k_\lambda\geq 1$ such that $R_k\leq 0$ for every k_λ.

(ii) For every $k\geq 1$ we have

$$\mu_k-\frac{\varepsilon}{2}(k+1)^2=\frac{\varepsilon}{2}(k+1)^2+\alpha^2(k+1)\sqrt{k+1}+(\alpha-4)(k+1)-(\alpha-2),$$

and the conclusion is obvious. □

The following proposition plays a key role in proving the convergence rates in Proposition B.5 which will be used to prove Theorem 3.2.

Proposition B.4. *Let $z^*\in\Omega$ and $(z_k)_{k\geq 0}$, $(w_k)_{k\geq 0}$, $(\zeta_k)_{k\geq 0}$ be the sequences generated by Algorithm 1 and let $(v_k)_{k\geq 0}$ be the sequence defined by (15). Then the following statements are true:*

(i) the following hold:

$$\sum_{k\geq 1}\langle z_k - z^*, F(z_k)+\zeta_k\rangle < +\infty, \tag{50a}$$

$$\sum_{k\geq 1} k\left\|v_{k+1}-v_k\right\|^2 < +\infty, \tag{50b}$$

$$\sum_{k\geq 1} k\left\|z_{k+1}-z_k\right\|^2 < +\infty, \tag{50c}$$

$$\sum_{k\geq 1} k\left\|F(w_k)+\zeta_{k+1}\right\|^2 < +\infty; \tag{50d}$$

(ii) the sequence $(z_k)_{k\geq 0}$ is bounded and the following hold as $k\to+\infty$:

$$\left\|z_k - z_{k-1}\right\| = \mathcal{O}\left(\frac{1}{k}\right), \quad \left\|\zeta_k + F\left(w_{k-1}\right)\right\| = \mathcal{O}\left(\frac{1}{k}\right), \quad \left\|\zeta_k + F\left(z_k\right)\right\| = \mathcal{O}\left(\frac{1}{k}\right),$$
$$\left\langle z_k - z^*, \zeta_k + F\left(z_k\right)\right\rangle = \mathcal{O}\left(\frac{1}{k}\right), \quad \left\langle z_k - z^*, F\left(z_k\right)\right\rangle = \mathcal{O}\left(\frac{1}{k}\right);$$

(iii) there exist $0 \leq \underline{\lambda}(\alpha) < \overline{\lambda}(\alpha) \leq {}^{(3\alpha-2)}/_4$ such that for every $\underline{\lambda}(\alpha) < \lambda < \overline{\lambda}(\alpha)$ the sequences $(\mathcal{E}_{\lambda,k})_{k\geq 1}$ and $(\mathcal{G}_{\lambda,k})_{k\geq 2}$ converge.

Proof. According to Lemma B.3 there exist $\underline{\lambda}(\alpha) < \overline{\lambda}(\alpha)$ such that (44) holds. We choose $\underline{\lambda}(\alpha) < \lambda < \overline{\lambda}(\alpha)$ and get, according to the same result, an integer $k_\lambda \geq 1$ such that for every $k \geq k_\lambda$ the inequality (45) holds. In addition, according to Lemma B.3(ii), we get a positive integer k_ε such that (46) holds for every $k \geq k_\varepsilon$.

This means that for every $k \geq k_1 := \max\{k_0, k_\lambda, k_\varepsilon\}$, where k_0 is the positive integer provided by Lemma B.2(i), we have

$$\begin{aligned}
&\mathcal{G}_{\lambda,k+1} - \mathcal{G}_{\lambda,k} \\
&\quad\leq \frac{(\alpha-1)(\alpha-2)\lambda^2}{\varepsilon(k+1)^2}\left\|z_{k+1}-z^*\right\|^2 - 4(\alpha-2)\lambda\gamma\left\langle z_{k+1}-z^*, \zeta_{k+1}+F\left(z_{k+1}\right)\right\rangle \\
&\quad- \frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\left\|v_{k+1}-v_k\right\|^2 + \left[\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k + \kappa_0\sqrt{k}\right]\left\|z_{k+1}-z_k\right\|^2 \\
&\quad+ \left[\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_3 k + \kappa_1\sqrt{k}\right]4\gamma^2\left\|v_{k+1}\right\|^2.
\end{aligned}$$

Since $\eta_2, \eta_3 < 0$ and $\kappa_0, \kappa_1 \geq 0$, we can find some $k_2 \geq k_1$ large enough such that for every $k \geq k_2$ we get

$$\begin{aligned}
\mathcal{G}_{\lambda,k+1} &\leq \mathcal{G}_{\lambda,k} + \frac{(\alpha-1)(\alpha-2)\lambda^2}{\varepsilon(k+1)^2}\left\|z_{k+1}-z^*\right\|^2 - 4(\alpha-2)\lambda\gamma\left\langle z_{k+1}-z^*, \zeta_{k+1}+F\left(z_{k+1}\right)\right\rangle \\
&\quad- \frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\left\|v_{k+1}-v_k\right\|^2 + \frac{1}{2}\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\left\|z_{k+1}-z_k\right\|^2 \\
&\quad+ \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)2\gamma^2\eta_3 k\left\|v_{k+1}\right\|^2.
\end{aligned} \tag{51}$$

In view of (35), we get that $\mathcal{G}_{\lambda,k} \geq 0$ for every $k \geq 1$ thus the sequence $(\mathcal{G}_{\lambda,k})_{k\geq 2}$ is bounded from below. Moreover, by setting

$$C_0 := \frac{1}{2\varepsilon}(\alpha-2)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)^{-1} > 0,$$

we assert that

$$\frac{(\alpha-1)(\alpha-2)\lambda^2}{\varepsilon(k+1)^2}\|z_{k+1}-z^*\|^2 = \frac{C_0}{(k+1)^2}\cdot 2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2$$
$$\leq \frac{C_0}{(k+1)^2}\mathcal{G}_{\lambda,k+1},$$

Under these premises, we deduce from (51) that for every $k \geq k_2$

$$\left(1-\frac{C_0}{(k+1)^2}\right)\mathcal{G}_{\lambda,k+1} \leq \mathcal{G}_{\lambda,k} - 4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle$$
$$-\frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2 \tag{52}$$
$$+\frac{1}{2}\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\|z_{k+1}-z_k\|^2 + \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)2\gamma^2\eta_3 k\|v_{k+1}\|^2.$$

Taking $k_3 := \max\left\{k_2, \left\lceil\sqrt{C_0}-1\right\rceil\right\}$ we conclude for every $k \geq k_3$ that

$$\left(1-\frac{C_0}{(k+1)^2}\right)^{-1} = \frac{(k+1)^2}{(k+1)^2-C_0} = 1+\frac{C_0}{(k+1)^2-C_0} > 1.$$

Hence, for every $k \geq k_3$, the inequality (52) leads to

$$\mathcal{G}_{\lambda,k+1} \leq \left(1+\frac{C_0}{(k+1)^2-C_0}\right)\mathcal{G}_{\lambda,k} - 4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle$$
$$-\frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2$$
$$+\frac{1}{2}\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\|z_{k+1}-z_k\|^2 + \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)2\gamma^2\eta_3 k\|v_{k+1}\|^2,$$

which is nothing else than the inequality (11) with

$$b_{\lambda,k} := 4(\alpha-2)\lambda\gamma\langle z_{k+1}-z^*, \zeta_{k+1}+F(z_{k+1})\rangle + \frac{\alpha-2}{2(\alpha-1)}\varepsilon\gamma^2(k+1)^2\|v_{k+1}-v_k\|^2$$
$$-\frac{1}{2}\left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)\eta_2 k\|z_{k+1}-z_k\|^2 - \left(1-\sqrt{\frac{5\alpha-2}{2(3\alpha-2)}}\right)2\gamma^2\eta_3 k\|v_{k+1}\|^2$$
$$\geq 0,$$
$$d_{\lambda,k} := \frac{C_0}{(k+1)^2-C_0} > 0.$$

Using Lemma A.2 we obtain (50) as well as convergence of the sequence $(\mathcal{G}_{\lambda,k})_{k\geq 1}$.

Since $(\mathcal{G}_{\lambda,k})_{k\geq 1}$ converges, it is also bounded from above, which, according to (35), implies that the following estimate holds for every $k \geq k_3$

$$\frac{\alpha-2}{3\alpha-2}\left\|4\lambda(z_k-z^*)+2k(z_k-z_{k-1})+\frac{2(3\alpha-2)}{\alpha-1}\gamma k v_k\right\|^2$$
$$+\frac{(\alpha-2)^2}{4(3\alpha-2)(\alpha-1)}k^2\|z_k-z_{k-1}\|^2 + 2(\alpha-1)\lambda\left(1-\frac{4\lambda}{3\alpha-2}\right)\|z_k-z^*\|^2$$
$$\leq \mathcal{G}_{\lambda,k} \leq \sup_{k\geq 1}\mathcal{G}_{\lambda,k} < +\infty.$$

From here we obtain the boundedness of the sequences

$$\left(4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha-2\right)}{\alpha-1}\gamma k v_k\right)_{k\geq 1},$$
$$\left(k\left(z_k - z_{k-1}\right)\right)_{k\geq 1} \quad \text{and} \quad \left(z_k\right)_{k\geq 0}.$$

In particular, for every $k \geq k_3$ we have

$$\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2(3\alpha-2)}{\alpha-1}\gamma k v_k\right\| \leq C_1 := \sqrt{\frac{3\alpha-2}{\alpha-2}\sup_{k\geq 1}\mathcal{G}_{\lambda,k}} < +\infty,$$
$$k\left\|z_k - z_{k-1}\right\| \leq C_2 := \frac{2}{\alpha-2}\sqrt{\left(3\alpha-2\right)\left(\alpha-1\right)\sup_{k\geq 1}\mathcal{G}_{\lambda,k}} < +\infty,$$
$$\left\|z_k - z^*\right\| \leq C_3 := \sqrt{\frac{1}{2\left(\alpha-1\right)\lambda}\left(1 - \frac{4\lambda}{3\alpha-2}\right)^{-1}\sup_{k\geq 1}\mathcal{G}_{\lambda,k}} < +\infty. \tag{53}$$

Using the triangle inequality, we deduce from here that for every $k \geq k_3$

$$\left\|v_k\right\| \leq \frac{\alpha-1}{2\left(3\alpha-2\right)\gamma k}\left\|4\lambda\left(z_k - z^*\right) + 2k\left(z_k - z_{k-1}\right) + \frac{2\left(3\alpha-2\right)}{\alpha-1}\gamma k v_k\right\|$$
$$+ \frac{\alpha-1}{\left(3\alpha-2\right)\gamma}\left\|z_k - z_{k-1}\right\| + \frac{2\left(\alpha-1\right)\lambda}{\left(3\alpha-2\right)\gamma k}\left\|z_k - z^*\right\| \leq \frac{C_4}{k},$$

where

$$C_4 := \frac{\alpha-1}{2\left(3\alpha-2\right)\gamma}\left(C_1 + 2C_2 + 4\overline{\lambda}\left(\alpha\right)C_3\right) > 0.$$

The statement (50b) yields

$$\lim_{k\to+\infty} k\left\|v_{k+1} - v_k\right\| = 0 \quad \Rightarrow \quad C_5 := \sup_{k\geq 1}\left\{k\left\|v_{k+1} - v_k\right\|\right\} < +\infty, \tag{54}$$

which, together with (18) implies that for every $k \geq k_3$

$$\begin{aligned}\left\|\zeta_{k+1} + F\left(z_{k+1}\right)\right\| &\leq \left\|\zeta_{k+1} + F\left(z_{k+1}\right) - v_{k+1}\right\| + \left\|v_{k+1}\right\| \\ &\leq \left\|v_{k+1} - v_k\right\| + \left\|v_{k+1}\right\| \leq \frac{C_6}{k},\end{aligned} \tag{55}$$

where

$$C_6 := C_4 + C_5 > 0.$$

The remaining assertion follows from the fact that $\zeta_k \in N_C\left(z_k\right)$ where $z_k \in C$ by definition, the Cauchy-Schwarz inequality and the boundedness of $\left(z_k\right)_{k\geq 0}$, namely, for every $k \geq k_3$ we deduce

$$0 \leq \left\langle z_k - z^*, F\left(z_k\right)\right\rangle \leq \left\langle z_k - z^*, \zeta_k + F\left(z_k\right)\right\rangle \leq \left\|z_k - z^*\right\|\left\|\zeta_k + F\left(z_k\right)\right\| \leq \frac{C_3 C_6}{k-1}.$$

To complete the proof, we are going to show that in fact

$$\lim_{k\to+\infty}\mathcal{E}_{\lambda,k} = \lim_{k\to+\infty}\mathcal{G}_{\lambda,k} \in \mathbb{R}.$$

Indeed, we already have seen that

$$\lim_{k\to+\infty}\left(k+1\right)\left\|v_{k+1} - v_k\right\| = \lim_{k\to+\infty}\left\|v_{k+1}\right\| = 0,$$

which, by the Cauchy-Schwarz inequality and (18) yields

$$0 \leq \lim_{k\to+\infty} k^2\left|\left\langle z_k - z_{k-1}, F\left(z_k\right) - F\left(w_{k-1}\right)\right\rangle\right| \leq C_2 \lim_{k\to+\infty} k\left\|F\left(z_k\right) - F\left(w_{k-1}\right)\right\|$$
$$\leq C_2 \lim_{k\to+\infty} k\left\|v_k - v_{k-1}\right\| = 0.$$

From here we obtain the desired statement. □

B.3. Proofs of the Main Results

Proof of Theorem 3.2. Let $\underline{\lambda}(\alpha) < \overline{\lambda}(\alpha)$ be the parameters provided by Lemma B.3 such that (44) holds and with the property that for every $\underline{\lambda}(\alpha) < \lambda < \overline{\lambda}(\alpha)$ there exists an integer $k_\lambda \geq 1$ such that for every $k \geq k_\lambda$ the inequality (45) holds.

For every $k \geq 1$ we set

$$p_k := \frac{1}{2}(\alpha - 1)\|z_k - z^*\|^2 + k\langle z_k - z^*, z_k - z_{k-1} + 2\gamma v_k\rangle, \tag{56}$$

$$q_k := \frac{1}{2}\|z_k - z^*\|^2 + 2\gamma \sum_{i=1}^{k} \langle z_i - z^*, v_i\rangle. \tag{57}$$

Then one can see that for every $k \geq 2$ we have

$$q_k - q_{k-1} = \langle z_k - z^*, z_k - z_{k-1}\rangle - \frac{1}{2}\|z_k - z_{k-1}\|^2 + 2\gamma\langle z_k - z^*, v_k\rangle,$$

and thus

$$(\alpha - 1)q_k + k(q_k - q_{k-1}) = p_k + 2(\alpha - 1)\gamma \sum_{i=1}^{k} \langle z_i - z^*, v_i\rangle - \frac{k}{2}\|z_k - z_{k-1}\|^2.$$

From (21) and (20), direct computation shows that for every $k \geq 1$

$$\begin{aligned}
\mathcal{E}_{\lambda,k} &= \frac{1}{2}\left\|2\lambda(z_k - z^*) + 2k(z_k - z_{k-1}) + \frac{3\alpha - 2}{\alpha - 1}\gamma k v_k\right\|^2 + 2\lambda(\alpha - 1 - \lambda)\|z_k - z^*\|^2 \\
&\quad + \frac{2(\alpha - 2)}{\alpha - 1}\lambda\gamma k\langle z_k - z^*, v_k\rangle + \frac{\alpha - 2}{\alpha - 1}\gamma^2 k\left(\frac{1}{2(\alpha - 1)}(3\alpha - 2)k + \alpha\right)\|v_k\|^2 \\
&= 2\lambda(\alpha - 1)\|z_k - z^*\|^2 + 4\lambda k\langle z_k - z^*, z_k - z_{k-1} + 2\gamma v_k\rangle + \frac{\alpha - 2}{\alpha - 1}\alpha\gamma^2 k\|v_k\|^2 \\
&\quad + \frac{k^2}{2}\left(\left\|2(z_k - z_{k-1}) + \frac{3\alpha - 2}{\alpha - 1}\gamma v_k\right\|^2 + \frac{(\alpha - 2)(3\alpha - 2)}{(\alpha - 1)^2}\gamma^2\|v_k\|^2\right).
\end{aligned} \tag{58}$$

Therefore, for every $\underline{\lambda}(\alpha) < \lambda_1 < \lambda_2 < \overline{\lambda}(\alpha)$ we can conclude

$$\begin{aligned}
\mathcal{E}_{\lambda_2,k} - \mathcal{E}_{\lambda_1,k} &= 4(\lambda_2 - \lambda_1)\left(\frac{1}{2}(\alpha - 1)\|z_k - z^*\|^2 + 2k\langle z_k - z^*, (z_k - z_{k-1}) + s v_k\rangle\right) \\
&= 4(\lambda_2 - \lambda_1)p_k.
\end{aligned}$$

Hence, according to the previous theorem, the limit $\lim_{k\to+\infty}(\mathcal{E}_{\lambda_2,k} - \mathcal{E}_{\lambda_1,k}) \in \mathbb{R}$ exists, which implies further that the limit

$$\lim_{k\to+\infty} p_k \in \mathbb{R} \text{ exists.} \tag{59}$$

Further, we observe that for every $k \geq 2$

$$\sum_{i=2}^{k} |\langle z_i - z^*, F(w_{i-1}) - F(z_i)\rangle| \leq \sum_{i=2}^{k} \|z_i - z^*\|\,\|F(w_{i-1}) - F(z_i)\| \tag{60a}$$

$$\leq \frac{1}{2}\sum_{i=2}^{k}\frac{1}{i^2}\|z_i - z^*\|^2 + \frac{1}{2}\sum_{i=2}^{k} i^2\|F(w_{i-1}) - F(z_i)\|^2$$

$$\leq \frac{1}{2}\sum_{i=2}^{+\infty}\frac{1}{i^2}\|z_i - z^*\|^2 + \frac{1}{2}\sum_{i=2}^{+\infty} i^2\|F(w_{i-1}) - F(z_i)\|^2 < +\infty, \tag{60b}$$

where (60a) comes from the Cauchy-Schwarz inequality, the first sum in (60b) is finite due to (53), while the second series is convergent because of (18) and (50b). This means the series $\sum_{k\geq 2}\langle z_k - z^*, F(w_{k-1}) - F(z_k)\rangle$ is absolutely convergent, thus convergent.

By taking into consideration (50a), it follows from here that the limit

$$\begin{aligned}&\lim_{k\to+\infty}\sum_{i=1}^{k}\langle z_i - z^*, v_i\rangle\\ &\quad= \lim_{k\to+\infty}\sum_{i=1}^{k}\langle z_i - z^*, \zeta_i + F(z_i)\rangle + \lim_{k\to+\infty}\sum_{i=1}^{k}\langle z_i - z^*, F(w_{i-1}) - F(z_i)\rangle \in \mathbb{R}\end{aligned}$$

exists. In addition, thanks to (50c), we have $\lim_{k\to+\infty} k\|z_{k+1} - z_k\|^2 = 0$, consequently,

$$\lim_{k\to+\infty}(\alpha - 1)q_k + k(q_k - q_{k-1}) \in \mathbb{R} \text{ exists.}$$

According to Proposition B.4, we have that $(q_k)_{k\geq 1}$ is bounded due to the boundedness of $(z_k)_{k\geq 0}$ and the fact that $\lim_{k\to+\infty}\sum_{i=1}^{k}\langle z_i - z^*, v_i\rangle \in \mathbb{R}$ exists. Therefore, we can apply Lemma A.1 to guarantee the existence of the limit $\lim_{k\to+\infty} q_k \in \mathbb{R}$. By the definition of q_k in (57) and the fact that the sequence $\left(\sum_{i=1}^{k-1}\langle z_i - z^*, v_i\rangle\right)_{k\geq 1}$ converges, we conclude that $\lim_{k\to+\infty}\|z_k - z^*\| \in \mathbb{R}$ exists. The hypothesis (i) in the Opial Lemma (see Lemma A.3) is fulfilled.

Let w be a cluster point of $(z_k)_{k\geq 0}$, which means that there exists a subsequence $\{z_{k_n}\}_{n\geq 0}$ such that

$$z_{k_n} \to w \text{ as } n \to +\infty.$$

It follows from Proposition B.4 that

$$F(z_{k_n}) + \zeta_{k_n} \to 0 \quad \text{as } n \to +\infty.$$

The maximal monotonicity of $F + N_C$ implies that $0 \in (N_C + F)(w)$, meaning that hypothesis (ii) of Lemma A.3 is also verified. The proof of the convergence of the iterates is therefore completed. □

Before finally proving Theorem 3.3 we show convergence rates of various helpful quantities.

Proposition B.5. *Let $z^* \in \Omega$ and $(z_k)_{k\geq 0}$ be the sequence generated by Algorithm 1. Then, as $k \to +\infty$, the following hold:*

$$\|z_k - z_{k-1}\| = o\left(\frac{1}{k}\right), \quad \langle F(z_k), z_k - z^*\rangle = o\left(\frac{1}{k}\right), \quad \langle \zeta_k + F(z_k), z_k - z^*\rangle = o\left(\frac{1}{k}\right)$$
$$\|\zeta_k + F(z_k)\| = o\left(\frac{1}{k}\right), \quad \|\zeta_k + F(w_{k-1})\| = o\left(\frac{1}{k}\right).$$

Proof. Let $\underline{\lambda}(\alpha) < \overline{\lambda}(\alpha)$ be the parameters provided by Lemma B.3 such that (44) holds and with the property that for every $\underline{\lambda}(\alpha) < \lambda < \overline{\lambda}(\alpha)$ there exists an integer $k_\lambda \geq 1$ such that for every $k \geq k_\lambda$ the inequality (45) holds. We fix $\underline{\lambda}(\alpha) < \lambda < \overline{\lambda}(\alpha)$ and recall that according to Proposition B.4(iii) the sequence $(\mathcal{E}_{\lambda,k})_{k\geq 1}$ converges.

We set for every $k \geq 1$

$$h_k := \frac{k^2}{2}\left(\left\|2(z_k - z_{k-1}) + \frac{3\alpha - 2}{\alpha - 1}\gamma v_k\right\|^2 + \frac{(\alpha - 2)(3\alpha - 2)}{(\alpha - 1)^2}\gamma^2\|v_k\|^2\right),$$

and notice that, in view of (58) and (56), we have

$$\mathcal{E}_{\lambda,k} = 4\lambda p_k + \frac{4(\alpha - 2)}{\alpha - 1}\alpha\gamma^2\gamma^2 k\|v_k\|^2 + h_k.$$

Proposition B.4 asserts that

$$\lim_{k\to\infty} k\|v_k\|^2 = 0,$$

which, together with $\lim_{k\to+\infty} \mathcal{E}_{\lambda,k} \in \mathbb{R}$ and $\lim_{k\to+\infty} p_k \in \mathbb{R}$ (see also (59)), yields the existence of

$$\lim_{k\to+\infty} h_k \in \mathbb{R}.$$

In addition, (50c) and (50d) in Proposition B.4 guarantee that

$$\sum_{k\geq 1} \frac{1}{k} h_k \leq 4 \sum_{k\geq 1} k \left\|z_k - z_{k-1}\right\|^2 + \frac{(3\alpha-2)(7\alpha-6)}{2(\alpha-1)^2} \gamma^2 \sum_{k\geq 1} k \left\|v_k\right\|^2 < +\infty.$$

Consequently, $\lim_{k\to+\infty} h_k = 0$, which yields

$$\lim_{k\to\infty} k \left\| 2(z_k - z_{k-1}) + \frac{3\alpha-2}{\alpha-1}\gamma v_k \right\| = \lim_{k\to\infty} k \left\|v_k\right\| = 0.$$

This immediately implies $\lim_{k\to+\infty} k \left\|z_k - z_{k-1}\right\| = 0$. The fact that

$$\lim_{k\to+\infty} k \left\|\zeta_k + F(z_k)\right\| = 0$$

follows from (18), (54) and (55), since

$$0 \leq \lim_{k\to+\infty} k \left\|\zeta_k + F(z_k)\right\| \leq \lim_{k\to+\infty} k \left\|v_k - v_{k-1}\right\| + \lim_{k\to+\infty} k \left\|v_k\right\| = 0.$$

Finally, using the Cauchy-Schwarz inequality and the fact that $(z_k)_{k\geq 0}$ is bounded, we obtain that $\lim_{k\to+\infty} k \langle z_k - z^*, F(z_k)\rangle = \lim_{k\to+\infty} k \langle z_k - z^*, \zeta_k + F(z_k)\rangle = 0$. □

Now we are able to prove the convergence rates in terms of the restricted gap and the natural gap.

Proof of Theorem 3.3. For every $k \geq 1$, using successively the monotonicity of F, the fact that $\zeta_k \in N_C(z_k)$, where $z_k \in C$ by its definition, and the Cauchy-Schwarz inequality, we deduce that for every $u \in C \cap \mathbb{B}(z^*; \delta(z_0))$

$$\begin{aligned}
\langle F(u), z_k - u\rangle &\leq \langle F(z_k), z_k - u\rangle \leq \langle \zeta_k + F(z_k), z_k - u\rangle \\
&= \langle \zeta_k + F(z_k), z_k - z^*\rangle + \langle \zeta_k + F(z_k), z^* - u\rangle \\
&\leq \langle \zeta_k + F(z_k), z_k - z^*\rangle + \left\|\zeta_k + F(z_k)\right\| \left\|z^* - u\right\| \\
&\leq \langle \zeta_k + F(z_k), z_k - z^*\rangle + \delta(z_0) \left\|\zeta_k + F(z_k)\right\|.
\end{aligned}$$

Therefore, it follows from Proposition B.5 that

$$\begin{aligned}
\texttt{Gap}(z_k) &= \max_{u\in C\cap\mathbb{B}(z^*;\delta(z_0))} \langle F(u), z_k - u\rangle \leq \langle \zeta_k + F(z_k), z_k - z^*\rangle + \delta(z_0) \left\|\zeta_k + F(z_k)\right\| \\
&= o\left(\frac{1}{k}\right) \quad \text{as } k \to +\infty.
\end{aligned}$$

Concluding, by (14) we obtain

$$\texttt{Res}(z_k) \leq \left\|\zeta_k + F(z_k)\right\| = o\left(\frac{1}{k}\right), \quad \text{as } k \to +\infty,$$

and the proof is complete. □

C. Implementation Details

In this section we report the details on the implementations for our GAN experiments.

C.1. Architecture

In Table 3 we describe the architectures that were used in the experiments on CIFAR-10. The models were selected replicating the set-up of (Miyato et al., 2018; Chavdarova et al., 2021b).

Table 3. ResNet architecture used for the CIFAR-10 experiments.

Generator (G)
Input: $z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
Linear $128 \to 4,096$
G-ResBlock
G-ResBlock
G-ResBlock
Batch Normalisation
ReLU
conv. (kernel: 3×3, $256 \to 3$, stride: 1, pad: 1)
$\tanh(\cdot)$
Discriminator (D)
Input: $x \in \mathbb{R}^{3\times32\times32}$
D-ResBlock
D-ResBlock
D-ResBlock
D-ResBlock
ReLU
Avg. Pool (kernel: 8×8) Linear $128 \to 1$
Spectral Normalisation

C.2. Hyperparameters

In Table 4 we list the hyperparameters that were used for fOGDA-VI to obtain the results on CIFAR-10. The hyperparameters for LA-GDA were the same as in (Chavdarova et al., 2021b).

Table 4. Hyperparameters used for the GAN experiments on CIFAR-10.

fOGDA-VI	
Batch size	= 128
Iterations	= 500, 000
Adam β_1	= 0.0
Adam β_2	= 0.9
Update ratio D/G	= 5
Learning rate for discriminator	$= 1 \times 10^{-4}$
Learning rate for generator	$= 1 \times 10^{-4}$
fOGDA α	= 100
fOGDA n	= 1000

C.3. PyTorch Code

In the following we report the code of the wrapper for the fOGDA-VI optimiser written using the PyTorch (Paszke et al., 2019) framework.

```
from torch.optim import Optimizer


class fOGDA(Optimizer):
    def __init__(self, optimizer, alpha=100, increment_iterator_every=1000):
        print(
            f"Using fOGDA (alpha={alpha};"
            f"increment iterator every {increment_iterator_every} step(s))."
        )
        self.optimizer = optimizer
        self.defaults = self.optimizer.defaults
        self.param_groups = self.optimizer.param_groups
        self.state = self.optimizer.state
```

```
        # fOGDA parameters
        self.alpha = alpha
        self.increment_iterator_every = increment_iterator_every
        self.iteration = 0
        self.k = 2
        self.params_copy = []
        self.old_params_copy = []
        self.updates = []
        self.old_updates = []
        self.old_difference_of_updates = []

    def step(self, closure=None):
        loss = None
        if closure is not None:
            loss = closure()

        no_old_params = len(self.old_params_copy) == 0
        no_old_updates = len(self.old_updates) == 0
        no_old_difference_of_updates = len(self.old_difference_of_updates) == 0

        # initialise (old) parameters
        if len(self.params_copy) > 0:
            raise RuntimeError("Something bad happend here...")
        for group in self.param_groups:
            for p in group["params"]:
                self.params_copy.append(p.data.clone())
                if no_old_params:
                    self.old_params_copy.append(p.data.clone())

        # reverse engineer update from optimizer step
        self.optimizer.step()
        i = -1
        if len(self.updates) > 0:
            raise RuntimeError("Something bad happend here...")
        for group in self.param_groups:
            for p in group["params"]:
                i += 1
                self.updates.append(self.params_copy[i] - p.data)

        # initialise old updates and difference of updates
        if (not no_old_updates and no_old_difference_of_updates) or (
            not no_old_difference_of_updates and no_old_updates
        ):
            raise RuntimeError("Something bad happend here...")
        if no_old_updates and no_old_difference_of_updates:
            for p in self.updates:
                self.old_updates.append(p.clone())
                self.old_difference_of_updates.append(torch.zeros_like(p))

        # compute fOGDA coefficients
        theta_p = self.alpha / (self.alpha + self.k + 1)
        theta = self.alpha / (self.alpha + self.k)
        theta_m = self.alpha / (self.alpha + self.k - 1)

        # compute new weights with fOGDA update
        i = -1
        for group in self.param_groups:
            for p in group["params"]:
                i += 1
                (
                    self.old_params_copy[i],
                    self.old_updates[i],
                    self.old_difference_of_updates[i],
                    p.data,
                ) = (
```

```
                    self.params_copy[i],
                    self.updates[i],
                    self.updates[i] - self.old_updates[i],
                    self.params_copy[i]
                    + (1 - theta_p)
                    * (self.params_copy[i] - self.old_params_copy[i])
                    - theta_p * self.updates[i]
                    - (2 - theta)
                    * (2 - theta_p)
                    * (self.updates[i] - self.old_updates[i])
                    + (2 - theta_m) * (1 - theta_p)
                    * self.old_difference_of_updates[i],
                )

        self.iteration += 1
        if self.iteration % self.increment_iterator_every == 0:
            # update iterator k
            self.k += 1

        # free parameters
        self.params_copy = []
        self.updates = []

        return loss
```