Solving Nonconvex-Nonconcave Min-Max Problems exhibiting Weak Minty Solutions

Anonymous authors
Paper under double-blind review

Abstract

We investigate a structured class of nonconvex-nonconcave min-max problems exhibiting so-called *weak Minty* solutions, a notion which was only recently introduced, but is able to simultaneously capture different generalizations of monotonicity. We prove novel convergence results for a generalized version of the optimistic gradient method (OGDA) in this setting, matching the $1/k$ rate for the best iterate in terms of the squared operator norm recently shown for the extragradient method (EG). In addition we propose an adaptive step size version of EG, which does not require knowledge of the problem parameters. paragraph.

1 Introduction

The recent success of machine learning models which can be described by min-max optimization, such as generative adversarial networks Goodfellow et al. (2014), adversarial learning Madry et al. (2018), adversarial example games Bose et al. (2020) or actor-critic methods Pfau & Vinyals (2016), has sparked interest in such saddle point problems. While methods have been identified, which (mostly) work in practice, the setting in which the objective function is nonconvex in the minimization and nonconcave in the maximization component remains theoretically poorly understood and even shows intractability results Daskalakis et al. (2021); Lee & Kim (2021a). Recently, Daskalakis et al. (2020) studied a class of nonconvex-nonconcave min-max problems and observed that the extragradient method (EG) showed good converge behavior in the experiments. Surprisingly, the problems did not seem to exhibit any of the known tame properties such as monotonicity, or Minty solutions. Later, Diakonikolas et al. (2021) found the appropriate notion (see Assumption 1), which is weaker than the existence of a Minty solution (a assumption extensively used in the literature Malitsky (2020); Mertikopoulos et al. (2019); Dang & Lan (2015); Liu et al. (2020; 2021)) and also generalizes the concept of negative comonotonicity Bauschke et al. (2020); Combettes & Pennanen (2004); Lee & Kim (2021a). Due to these unifying and generalizing properties the notion of weak Minty solutions was promptly studied in Pethick et al. (2022); Lee & Kim (2021b).

Assumption 1 (Weak Minty solution)**.** *Given an operator $F : \mathbb{R}^d \to \mathbb{R}^d$ there exists a point $u^* \in \mathbb{R}^d$ and a parameter $\rho > 0$ such that*

$$\langle F(u), u - u^* \rangle \geq -\frac{\rho}{2}\|F(u)\|^2 \quad \forall u \in \mathbb{R}^d. \tag{1}$$

Additionally, Diakonikolas et al. (2021) proved that a generalization of the EG Korpelevich (1976) is able to solve problems which exhibit such solutions with a complexity of $\mathcal{O}(\varepsilon^{-1})$ for the squared operator norm. This modification which they title EG+, is based on an aggressive extrapolation step combined with a conservative update step. Such a step size policy has already been explored in the context of a stochastic version of EG in Hsieh et al. (2020).

In a similar spirit we investigate a variant of the optimistic gradient descent ascent (OGDA) Daskalakis et al. (2018); Popov (1980)/Forward-Reflected-Backward (FoRB) Malitsky & Tam (2020) method. We pose the question, and give an affirmative answer to:

Can OGDA match the convergence guarantees of EG in the presence of weak Minty solutions?

In particular, we show that the following modification of the OGDA method, given for step size $a > 0$ and parameter $0 < \gamma \leq 1$ by

$$u_{k+1} = u_k - a\Big((1+\gamma)F(u_k) - F(u_{k-1})\Big), \quad \forall k \geq 0,$$

is able to match the bounds of EG+ Diakonikolas et al. (2021); Pethick et al. (2022):

$$u_k = \bar{u}_k - aF(\bar{u}_k), \quad \bar{u}_{k+1} = \bar{u}_k - \gamma a F(u_k), \quad \forall k \geq 0,$$

by only requiring one gradient oracle call per iteration. In Figure 1 we see that beyond the theoretical guarantees OGDA+ can even provide convergence where EG+ does not.

Note that OGDA is most commonly written in the form where $\gamma = 1$, see Daskalakis et al. (2018); Malitsky & Tam (2020); Böhm et al. (2020), with the exception of two recent works which have investigated a more general coefficient see Ryu et al. (2019); Mokhtari et al. (2020). While the previous references target the monotone setting the true importance of γ only shows up in the presence of weak Minty solutions as in this case we *require* it to be larger than 1 to guarantee convergence — a phenomenon not present for monotone problems.

Connection to min-max. When considering a general (smooth) min-max problem

$$\min_x \max_y \; f(x,y)$$

the operator F mentioned in Assumption 1 arises naturally as $F(u) := [\nabla_x f(x,y), -\nabla_y f(x,y)]$ with $u = (x,y)$, c.f. Gidel et al. (2019). However, by studying saddle point problems from this more general perspective of variational inequalities (VIs), see (SVI), via the operator F we can simultaneously capture more settings such as certain equilibrium problems, see Facchinei & Pang (2007).

About the weak Minty parameter ρ. The parameter ρ in the definition of weak Minty solutions (1) plays a crucial role in the analysis and the experiments. In particular it is necessary that the step size is larger than a multiple of ρ. At the same time, as typical, the step size is constrained from above by the reciprocal of the Lipschitz constant of F. For example, since the authors of Diakonikolas et al. (2021) require the step size to be less than $\frac{1}{L}$, their convergence statement only holds if $\rho < \frac{1}{4L}$ for the choice $\gamma = \frac{1}{2}$. This was later improved in Pethick et al. (2022) to $\frac{1}{2L}$ for the particular choice of $\gamma = \frac{1}{2}$ and to $\frac{1}{L}$ for γ even smaller. As in the monotone setting, OGDA however, requires a smaller step size than EG. Nevertheless, through a different analysis we are able to match the most general condition on the weak Minty parameter $\rho < \frac{1}{L}$ for appropriate γ and a.

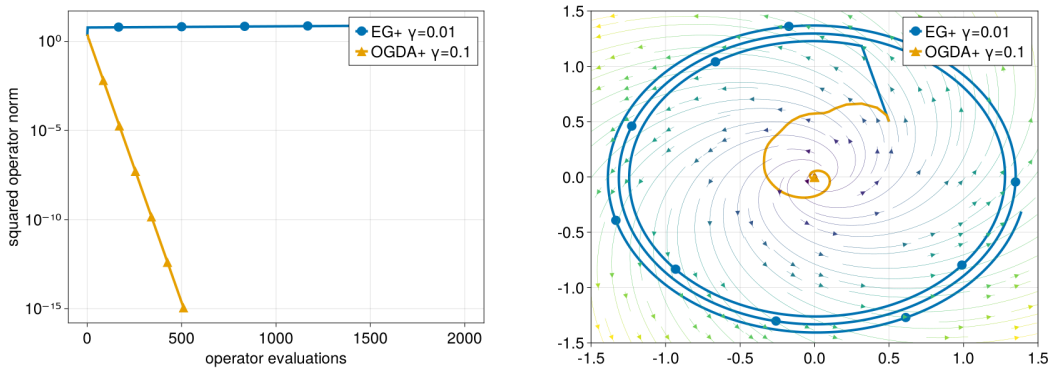
Contribution.

1. Building on the recently introduced notion of weak solutions to the Minty variational inequality, see Diakonikolas et al. (2021), we prove a novel convergence rate of $\mathcal{O}(1/k)$ in terms of the squared operator norm for a modification of OGDA, which matches the one of EG.

2. Under the stronger assumption that the operator is even monotone we recover the results of Ryu et al. (2019); Chavdarova et al. (2021), but with significantly improved (i.e. larger) step size.

3. We prove a complexity bound of $\mathcal{O}(\varepsilon^{-2})$ for a stochastic version of our newly proposed OGDA+ method.

4. Additionally, we propose an adaptive step size version of EG+, which is able to obtain the same convergence guarantees without any knowledge of the Lipschitz constant of the operator F, and therefore possibly even take larger steps in regions of low curvature and allow for convergence where a fixed step size policy does not.



(a) Performance of methods.

(b) Trajectories of iterates.

Figure 1: Problem (29), with $\rho = \frac{1}{L}$ meaning that convergence is not covered by theory. Since the Lipschitz constant can be computed analytically we choose the step size accordingly. Only OGDA+ is able to converge.

1.1 Related literature

Since there is an extensive literature on convergence rates in terms of a gap function or distance to a solution for monotone problems as well as generalizations such as nonconvex-concave Boţ & Böhm (2020); Lin et al. (2020), convex-nonconcave Xu et al. (2020) or under the Polyak-Łojasiewicz assumption, see Yang et al. (2020), we will only focus on the *nonconvex-nonconcave* setting and rates in terms of the *operator norm* for the monotone setting.

Operator norm for monotone problems. While in the convex-concave (or monotone) setting notions like *function values*, *primal-dual gap* or *distance to the solution* are frequently used, the operator norm as measure of optimality has received less attention in the literature. Nevertheless, Ryu et al. (2019) showed that OGDA exhibits a $\mathcal{O}(1/k)$ rate and that simultaneous gradient descent ascent with anchoring can almost match this rate (i.e. $\mathcal{O}(1/k^{1-\varepsilon})$ for $\varepsilon \in (0,1)$). Later Yoon & Ryu (2021) showed that EG with anchoring even achieves a rate of $\mathcal{O}(1/k^2)$ together with a matching lower bound.

Weak minty. In Diakonikolas et al. (2021) it was noticed that a particular parametrization of the von Neumann ratio game exhibits a new type of solution, which they titled weak Minty, without possessing any of the known properties such as (negative) comonotonicity or Minty solutions. They showed convergence in the presence of such solutions for EG if the extrapolation step size is twice as large as the update step. Later Pethick et al. (2022) showed that the condition on the weak Minty parameter can be relaxed by reducing the length of the update step even further and they do so in an adaptive way. In order to not require any other hyperparameters they also propose a backtracking line search, which might come at the cost of additional gradient computations or the use of second order information (in contrast to the adaptive step size we propose in Algorithm 2). In Lee & Kim (2021b) a different approach is taken by restricting the attention to the min-max setting and using multiple ascent steps per descent step, obtaining the same $\mathcal{O}(1/k)$ rate as EG.

Minty solutions. Many works have shown different approaches for when the problem at hand exhibits a Minty solution, see (MVI). The authors of Liu et al. (2021) showed that weakly monotone VIs can be solved by successively adding a quadratic proximity term and repeatedly optimizing the resulting strongly monotone VI with any convergent method. In Mertikopoulos et al. (2019) the convergence of the OGDA method was proven, but without any rate. In Malitsky (2020) it was noted that the convergence proof for the golden ratio algorithm (GRAAL) works without any modification. See also Dang & Lan (2015) for a non-euclidean version of EG and Liu et al. (2020) for adaptive methods. While the assumption that a Minty solution exists is a generalization of the monotone setting it is difficult to find nonmonotone problems that do possess such solutions. In our setting, see Assumption 1, the Minty inequality (MVI) is allowed to be violated at every point by a factor proportional to the squared operator norm.

Negative comonotonicity. While previously studied under the name of *cohypomonotonicity* Combettes & Pennanen (2004) the notion of negative comonotonicity was recently explored in Bauschke et al. (2020). It provides a generalization of monotonicity, but in a direction different from the notion of Minty solutions and only a few works have analyzed methods in this setting. The authors of Lee & Kim (2021a) analyzed an anchored version of EG and showed an improved convergence rate of $\mathcal{O}(1/k^2)$ (in terms of the squared operator norm). It is an open question whether such an acceleration is possible in the more general setting of weak Minty solutions (any Stampacchia solution to the VI given by negatively comonotone operator is a weak Minty solution).

Interaction dominance. The authors of Grimmer et al. (2020) investigate the notion of *α-interaction dominance* for nonconvex-nonconcave min-max problems and showed that the proximal-point method converges sublinearly if this condition holds in y and linearly if it holds in both components. Furthermore Lee & Kim (2021a) showed that if a problem is interaction dominant in both components, then it is also negatively comonotone.

Optimism. The beneficial effects of introducing the simple modification commonly known as optimism have recently sparked the interest of the machine learning community Daskalakis et al. (2018); Daskalakis & Panageas (2018); Liang & Stokes (2019); Gidel et al. (2019). Its name originates from online optimization Rakhlin & Sridharan (2013a;b). The idea dates back even further Popov (1980) and has been studied in the mathematical programming community as well Malitsky (2015); Malitsky & Tam (2020); Csetnek et al. (2019).

2 Preliminaries

2.1 Notions of solution

We summarize the most commonly used notions of solutions appearing in the context of variational inequalities (VIs) and beyond. Note that these are commonly defined in terms of a constraint set $C \subset \mathbb{R}^d$. A *Stampacchia*[1] Kinderlehrer & Stampacchia (2000) solution of the VI given by $F : \mathbb{R}^d \to \mathbb{R}^d$ is defined as a point u^* such that

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

In particular we only consider in this manuscript the unconstrained case $C = \mathbb{R}^d$ in which case the above condition reduces to $F(u^*) = 0$. Very much related with a long tradition is the following. A *Minty*[1] solution is given by a point $u^* \in C$ such that

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

If F is continuous, a Minty solution of the VI is always a Stampacchia solution. The reverse is in general not true but holds for example if the operator F is monotone. In particular, there exist nonmonotone problems with Stampacchia solutions but without any Minty solutions.

2.2 Notions of monotonicity

The aim of this section is to recall some elementary and some more recent notions of monotonicity and the connection between those. We call an operator F **monotone** if

$$\langle F(u) - F(v), u - v \rangle \geq 0.$$

Such operators arise naturally as the gradients of convex functions, solution operators of convex-concave min-max problems or from equilibrium problems.

Two notions frequently used in the literature that fall in this class are **strongly monotone** operators fulfilling

$$\langle F(u) - F(v), u - v \rangle \geq \mu \|u - v\|^2.$$

[1]Sometimes Stampacchia and Minty solutions are referred to as *strong* and *weak* solutions respectively, see Nesterov (2007), but we will refrain from this nomenclature, as it is confusing in the context of *weak Minty* solutions.

These form a very well-behaved subclass and usually lead to linear convergence of a variety of methods. They appear as (sub)-gradients of strongly convex functions or strongly-convex-strongly-concave min-max problems. A second subclass of monotone operators are so-called γ-**cocoercive** operator fulfilling

$$\langle F(u) - F(v), u - v\rangle \geq \beta\|F(u) - F(v)\|^2. \tag{2}$$

They appear as gradients of smooth convex functions. In this case the above inequality holds with β equal to the reciprocal of the gradients Lipschitz constant. At first sight it is not clear why lower bounding the term on the left hand side by either the difference of arguments or difference of operators would give nicer behavior. But in fact, strong monotonicity is typically the much "stronger" property. But there is also a certain duality between the two: If F is either strongly monotone or cocoercive then F^{-1} fulfills the other property (or a set-valued generalization thereof). We also want to point that if F is the gradient of a convex function f then F^{-1} is nothing other than the gradient of the convex conjugate f^*.

Leaving the monotone world. Both subclasses of monotonicity introduced above can be used to venture into the non-monotone world. Since general non-monotone operators might be very ill-behaved it makes sense to find settings that extend the monotone one, but still remain tractable. First and foremost, the by now well-studied setting of ν-**weak monotonicity**

$$\langle F(u) - F(v), u - v\rangle \geq -\nu\|u - v\|^2.$$

Such operators arise as the (sub)-gradients of the well-studied class (see Davis & Drusvyatskiy (2018); Drusvyatskiy & Paquette (2019)) of weakly convex functions — a rather generic class of functions as it includes all functions *without upward cusps*. In particular every smooth function (Lipschitz gradient) turns out to fulfill this property. On the other hand, extending the notion of cocoercivity to allow for negative coefficients, referred to as **cohypomonotonicity**, has received much less attention Bauschke et al. (2020); Combettes & Pennanen (2004) and is given by

$$\langle F(u) - F(v), u - v\rangle \geq -\gamma\|F(u) - F(v)\|^2.$$

Behavior w.r.t. the solution. While the above properties are standard assumption in the literature, it is usually sufficient to ask for the corresponding condition to hold when one of the arguments is a (Stampacchia) solution. This means instead of monotonicity it is enough to obtain standard convergence results, see Gorbunov et al. (2021), to ask for the operator F to be **star-monotone** Pennanen (2001), i.e.

$$\langle F(u), u - u^*\rangle \geq 0$$

or **star-cocoercive** Gorbunov et al. (2021)

$$\langle F(u), u - u^*\rangle \geq \gamma\|F(u)\|^2.$$

In this spirit, we can provide a new interpretation to the assumption of the existence of a weak Minty solution as asking for the operator F to be **negatively star-cocoercive** (with respect to at least one solution). Furthermore, we want to point out that while the above star notions are sometimes required to hold for all solutions u^*, in the following we only require it to hold for a single solution.

3 OGDA for problems with weak Minty solutions

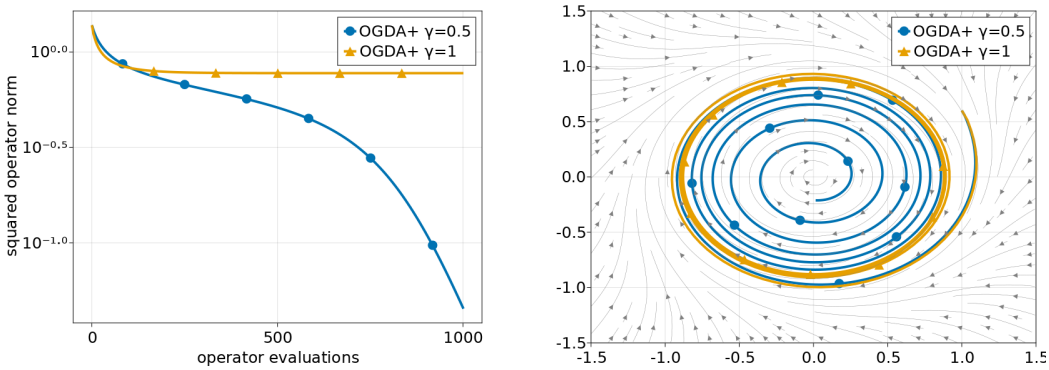
The generalized version of OGDA, whose name we equip in the spirit of Diakonikolas et al. (2021); Pethick et al. (2022), with a "+" to highlight the presence of the additional parameter γ, is given by:

Algorithm 1 OGDA+

Require: Starting point $u_0 = u_{-1} \in \mathbb{R}^d$, step size $a > 0$ and parameter $0 < \gamma \leq 1$.

for $k = 0, 1, \ldots$ **do**

$u_{k+1} = u_k - a\Big((1+\gamma)\,F(u_k) - F(u_{k-1})\Big)$



(a) Performance of methods.

(b) Sign of $\langle F(u), u - u^* \rangle$.

Figure 2: **Polar game** from Pethick et al. (2022) with $\frac{1}{8L} > \rho > \frac{1}{10L}$. Shows the need to reduce γ in OGDA+.

Theorem 3.1. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be L-Lipschitz continuous satisfying Assumption 1 with $\frac{1}{L} > \rho$ and $(u_k)_{k\geq 0}$ be the iterates generated by Algorithm 1 with step size a satisfying $a > \rho$ and*

$$aL \leq \frac{1-\gamma}{1+\gamma}. \tag{3}$$

Then, for all $k \geq 0$

$$\frac{1}{k}\sum_{i=0}^{k-1} \|F(u_i)\|^2 \leq \frac{1}{ka\gamma(a-\rho)}(\|u_0 + aF(u_0) - u^*\|^2).$$

In particular as long as $\rho < \frac{1}{L}$ we can find a γ small enough such that the above bound holds.

The first observation is that we would like to choose a as large as possible as this allows us to treat the largest class of problems with $\rho < a$. In order to be able to choose the step size a large we have to decrease γ as evident from (12). This, however degrades the speed of the algorithm as it makes the effective update steps smaller — the same effect can be observed Pethick et al. (2022) for EG+ and is therefore to be expected.

Furthermore, we want to point out that the condition $\rho < \frac{1}{L}$, is precisely the best possible bound for EG+ in Pethick et al. (2022).

Remark 3.1. As typical for nonconvex optimization Nesterov (2013) we do not get any bounds on the last iterate but rather on the average operator norms along the trajectory, which can, however, easily be translated into one of the two other popular types of statements. On the one hand this average is equal to the expectation of a uniformly at **random** chosen **iterate**, i.e.

$$\mathbb{E}\left[\|F(\hat{u}_k)\|^2\right] = \frac{1}{k}\sum_{i=0}^{k-1} \|F(u_i)\|^2,$$

where $\hat{u}_k \sim \mathcal{U}(u_0, \ldots, u_{k-1})$. On the other hand, it can be lower bounded by the operator norm of the **best iterate**

$$\min_{i=0,\ldots,k-1} \|F(u_i)\|^2 \leq \frac{1}{k}\sum_{i=0}^{k-1} \|F(u_i)\|^2.$$

Both of these forms are frequently used in the literature Daskalakis et al. (2020); Gorbunov et al. (2021) and we just recall them here for the convenience of the reader.

3.1 Improved bounds under monotonicity

While the above theorem also holds if the operator F is monotone, we can modify the proof slightly to obtain a better dependence on the parameters:

Theorem 3.2. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be monotone and L-Lipschitz. If*

$$aL < \frac{2-\gamma}{2+\gamma} \tag{4}$$

then, the iterates generated by OGDA+ converge to a solution. In particular, we can choose $\gamma = 1$ and $a < \frac{1}{3L}$.

3.2 OGDA+ stochastic

If the exact operator F is replaced by an unbiased estimator $\tilde{F}$ such that $\mathbb{E}\left[\tilde{F}(u_k) \,\middle|\, u_{k-1}\right] = F(u_k)$, which has bounded variance $\mathbb{E}\left[\|\tilde{F}(u_k) - F(u_k)\|^2\right] \leq \sigma^2$, we show that we can still guarantee convergence.

Theorem 3.3. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be L-Lipschitz satisfying Assumption 1 with $\frac{3}{4L} > \rho$ and let $(u_k)_{k\geq 0}$ be the sequence of iterates generated by stochastic OGDA+, with a and γ satisfying*

$$aL \leq \frac{7 - \sqrt{1+48\gamma^2}}{8(1+\gamma)}, \tag{5}$$

then, to visit an ε stationary point we require

$$\mathcal{O}\left(\frac{1}{\varepsilon}\frac{1}{a\gamma(a-\rho)}\|u_1 + ag_0 - u^*\|^2 \max\left\{1, \frac{16\sigma^2}{\varepsilon}\right\}\right)$$

calls to the stochastic oracle, with large batch sizes of order $\mathcal{O}(\varepsilon^{-2})$.

We recall that in the monotone case the analysis for $\gamma = 1$ (i.e. classical OGDA) allows for the larger step size $\frac{1}{2L}$. It can be suspected that the above analysis is therefor not tight. However it still provides significant improvement over the bounds proved in Ryu et al. (2019); Chavdarova et al. (2021).

4 EG+ with adaptive step sizes

In this section we present Algorithm 2 that is able to solve the previously mentioned problems without any knowledge of the Lipschitz constant L, as it is typically difficult to compute in practice. Additionally, it is well known that rough estimates will lead to small step sizes and slow convergence behavior. However, in the presence of weak Minty solutions there is additional interest in choosing large step sizes. We observed in Theorem 3.1 and related works such as Diakonikolas et al. (2021) and Pethick et al. (2022) the fact that a crucial ingredient in the analysis is that the step size is chosen larger than a multiple of the weak Minty parameter ρ to guarantee convergence at all. For these reasons we want to outline a method using adaptive step sizes estimating the local Lipschitz constant.
Since the analysis of OGDA+ is already quite involved in the constant step size regime we choose to equip EG+ with an *adaptive step size* which estimates the inverse of the (local) Lipschitz constant, see (6). Due the fact that the literature on adaptive methods, especially in the context of VIs is so vast we do not aim to give a comprehensive review but highlight only few with especially interesting properties. In particular we do not want to touch on methods with linesearch procedure which typically result in multiple gradient computations per iteration, such as Tseng (1991); Malitsky & Tam (2020).

We use a simple and therefore widely used adaptive step size choices which naively estimates the local Lipschitz constant and forces a monotone decreasing behavior. Such step sizes have been used extensively for monotone VIs, see Yang & Liu (2018); Boţ et al. (2020), and similarly in the context of the mirror-prox method which corresponds to EG in the setting of (non-euclidean) Bregman distances, see Antonakopoulos et al. (2019).

A version of EG with a different adaptive step size choice has been investigated Antonakopoulos et al. (2021); Bach & Levy (2019) with the unique feature that it is able to achieve the optimal rates for both smooth and nonsmooth problems without modification. However, these rates are only for monotone VIs and are in terms of the gap function.

One of the drawbacks of adaptive methods resides in the fact that the step sizes are typically required to be nonincreasing which results in poor behavior if a high curvature area was visited by the iterates before reaching a low curvature region. To the best of our knowledge the only method which is allowed to use nonmonotone step size to treat VIs, and does not use a possibly costly linesearch, is the golden ratio algorithm Malitsky (2020). It comes with the additional benefit of not requiring a global bound on the Lipschitz constant of F at all. While it is known that this method converges under the stronger assumption of existence of Minty solutions, a quantitative convergence result is still open.

Algorithm 2 EG+ with adaptive step size

Require: Starting points $u_0, \bar{u}_0 \in \mathbb{R}^d$, initial step size a_0 and parameter $\tau \in (0,1)$ and $0 < \gamma \leq 1$.

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

2: Find the step size:

$$a_k = \min\left\{a_{k-1}, \frac{\tau\|u_{k-1} - \bar{u}_{k-1}\|}{\|F(u_{k-1}) - F(\bar{u}_{k-1})\|}\right\}. \tag{6}$$

3: Compute next iterate:

$$\begin{aligned} u_k &= \bar{u}_k - a_k F(\bar{u}_k) \\ \bar{u}_{k+1} &= \bar{u}_k - a_k \gamma F(u_k). \end{aligned}$$

Clearly, a_k is monotonically decreasing by construction. Moreover, it is bounded away from zero by the simple observation that $a_k \geq \min\{a_0, \tau/L\} > 0$. The sequence therefore converges to a positive number which we denote by $a_\infty := \lim_k a_k$.

Theorem 4.1. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be L-Lipschitz that satisfy Assumption 1 with $a_\infty > 2\rho$ and let $(u_k)_{k\geq 0}$ be the iterates generated by Algorithm 2 with $\gamma = \frac{1}{2}$ and $\tau \in (0,1)$. Then, there exists a $k_0 \in \mathbb{N}$, such that*

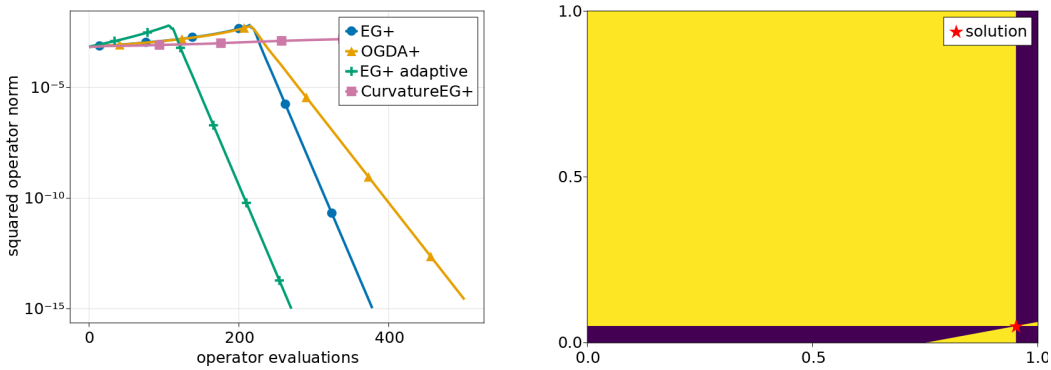
$$\frac{1}{k-k_0}\sum_{i=k_0}^{k}\|F(u_k)\|^2 \leq \frac{1}{k-k_0}\frac{L}{\tau(\frac{a_\infty}{2}-\rho)}\|\bar{u}_{k_0} - u^*\|^2.$$

Algorithm 2 presented above provides several benefits, but also some drawbacks. The main advantage resides in the fact that the Lipschitz constant of the operator F does not need to be known. Moreover, the step size choice presented in (6) might allow us to take steps much larger than what would be suggested by a global Lipschitz constant if the iterates never — or only during later iterations — visits the region of high curvature (large local L). In such cases these larger step sizes come with the additional advantage that they allow us to solve a richer class of problems as we are able to relax the condition $\rho < \frac{1}{4L}$ in the case of EG+ and Algorithm 1 to $\rho < a_\infty/2$ where $a_\infty = \lim_k a_k \geq \tau/L$.

On the other hand, we face the problem that the bounds in Theorem 4.1 only hold after an unknown number of initial iterations when $a_k/a_{k+1} \leq \frac{1}{\tau}$ is finally satisfied. In theory this might take long if the curvature around the solution is much higher than in the starting area as this will force the need to decrease the step size very late into the solution process resulting in the quotient a_k/a_{k+1} being too large. This drawback could be mitigated by choosing τ smaller. However, this will result in poor performance due to small step sizes.

5 Numerical experiments

In the following we compare EG+ method from Diakonikolas et al. (2021) with the two methods we propose OGDA+ and EG+ with adaptive step size, see Algorithm 1 and Algorithm 2 respectively. Last but not least we also include the CurvatureEG+ method from Pethick et al. (2022) which is a modification of EG+ and adaptively chooses the ratio of extrapolation and update step. In addition a backtracking linesearch is performed with an initial guess made by second order information, whose extra cost we ignore in the experiments.



(a) Performance of methods.

(b) Sign of $\langle F(u), u - u^* \rangle$.

Figure 3: **Ratio game**. A particularly difficult parametrization of (7) suggested in Daskalakis et al. (2020). Figure 3b shows the sign of $\langle F(u), u - u^* \rangle$ parametrized as $u := (x, 1-x, y, 1-y)$, with yellow representing a negative sign and purple a positive one. Step Sizes for EG+ and OGDA+ are tuned with the latter method using half the step size of the former.

5.1 Von Neumann's ratio game

We consider von Neumann's *ratio game* von Neumann (1945) recently explored in Daskalakis et al. (2020); Diakonikolas et al. (2021). It is given by

$$\min_{x \in \Delta^m} \max_{y \in \Delta^n} V(x, y) = \frac{\langle x, Ry \rangle}{\langle x, Sy \rangle}, \tag{7}$$

where $R \in \mathbb{R}^{m \times n}$ and $S \in \mathbb{R}_+^{m \times n}$ with $\langle x, Sy \rangle > 0$ for all $x \in \Delta^m, y \in \Delta^n$, with $\Delta^d := \{z \in \mathbb{R}^d : z_i \geq 0, \sum_{i=1}^d z_i = 1\}$ denoting the unit simplex. Expression (7) can be interpreted as the value $V(\pi_x, \pi_y)$ for a stochastic game with a single state and mixed strategies.

In Figure 3b we see an illustration of a particularly difficult instance of (7) discussed in Daskalakis et al. (2020), highlighting the fact that the Stampacchia solution is not a Minty solution, even when restricted to arbitrarily close ball around it (yellow area touching the solution). Interestingly we still observe good convergence behavior, although an estimated ρ is more then ten times larger than the estimated Lipschitz constant. Furthermore, Figure 3a shows that a hand tuned step size (in the case of EG+ and OGDA+) can lead to good convergence behavior, although it is chosen much larger than $1/L$. For the update step CurvatureEG+ ends up choosing a much smaller fraction of the extrapolation step than the other methods, resulting in slower convergence behavior.

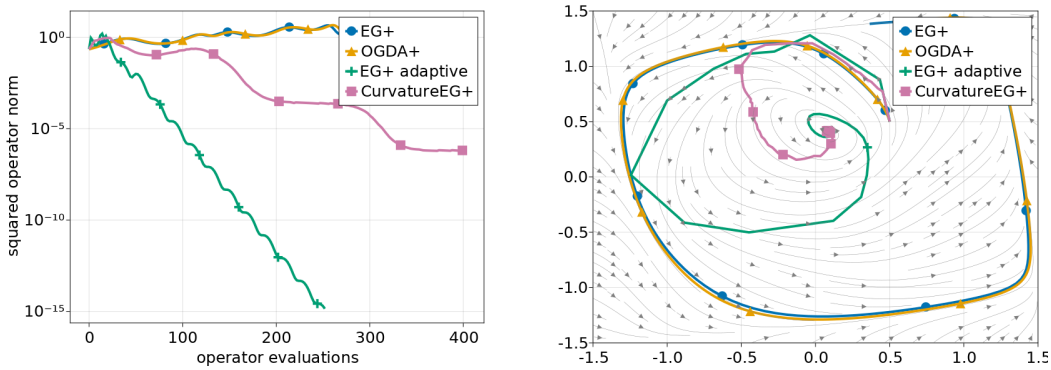
5.2 Forsaken

A particularly difficult min-max toy example with "Forsaken" solution was proposed in Example 5.2 of Hsieh et al. (2021), and is given by

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} x(y - 0.45) + \varphi(x) - \varphi(y), \tag{8}$$

where $\varphi(z) = \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$. This problem exhibits a Stampacchia solution at $(x^*, y^*) \approx (0.08, 0.4)$, but also *two* limit cycles not containing any critical point of the objective function. In addition, Hsieh et al. (2021) also observed that the limit cycle closer to the solution repels possible trajectories of iterates, thus "shielding" the solution. Later, Pethick et al. (2022) noticed that, restricted to the box $\|(x, y)\|_\infty < \frac{3}{2}$ the above mentioned solution is weak Minty with $\rho \approx 0.477761$, which is much larger than $\frac{1}{L} \approx 0.08$. In line with these observations we can see in Figure 4 that none of the fixed step size methods with step size bounded by $\frac{1}{L}$ converge. In light of this observation Pethick et al. (2022) proposed a backtracking linesearch which potentially allows for larger steps than predicted by the global Lipschitz constant. Similarly, our proposed adaptive step size version of EG+, see Algorithm 2, is also able to break through the repelling limit cycle

and converge the solution. On top of this, it does so at a faster rate and without the need of additional computations in the backtracking procedure.



(a) Performance of methods.

(b) Trajectories of iterates.

Figure 4: **Forsaken.** An illustration of different methods on problem (8), originally suggested in Hsieh et al. (2021). Step Sizes for EG+ and OGDA+ are chosen according to theory, $\frac{1}{L}$ and $\frac{1}{2L}$ respectively. Only Algorithm 2 and CurvatureEG+ are able to choose a step size large enough to withstand the repellent limit cycle.

6 Conclusion

Many interesting questions remain in the realm of min-max problems — especially when leaving the convex-concave setting. Very recently Gorbunov et al. (2021) showed that the $\mathcal{O}(1/k)$ bound on the squared operator norm for EG even holds for the **last iterate**, and not just the best one (at least in the monotone setting). Deriving a similar statement in the presence of weak Minty solutions outside of the monotone world seems too much to ask for as the proof in Gorbunov et al. (2021) relies on the fact that the operator norm monotonically decreases, which is not the case here (which can be seen from the experiments).

Overall, our analysis and experiments seem to provide evidence that there is little advantage of using OGDA+ over EG+ for most problems as the lower iteration cost is offset by the smaller step size, except in the case of problem (29) displayed in Figure 1.

On the other hand we observe that previous paradigm in pure minimization of "smaller step size ensures convergence" but "larger step size gets there faster", where the latter is typically constrained by the reciprocal of the gradients Lipschitz constant, does not seem to hold true for min-max problems anymore. The analysis of different methods in the presence of weak Minty solutions shows that convergence can be lost if the step size is too small and sometimes needs to be larger than $\frac{1}{L}$, which one can typically only hope for in adaptive methods. Our EG+ method with adaptive step size achieves this even without the additional cost of a backtracking linesearch as used for the CurvatureEG+ method of Pethick et al. (2022).

References

Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox method for variational inequalities with singular operators. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/c2890d44d06bafb6c7b4aa194857ccbc-Paper.pdf`.

Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. Adaptive extra-gradient methods for min-max optimization and games. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=R0a0kFI3dJx`.

Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, pp. 164–194. PMLR, 2019.

Heinz H Bauschke, Walaa M Moursi, and Xianfu Wang. Generalized monotone operators and their averaged resolvents. *Mathematical Programming*, pp. 1–20, 2020.

Axel Böhm, Michael Sedlmayer, Ernö Robert Csetnek, and Radu Ioan Boţ. Two steps at a time – taking GAN training in stride with Tseng's method. *arXiv:2006.09033*, 2020.

Avishek Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and William L. Hamilton. Adversarial example games. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/65586803f1435736f42a541d3a924595-Abstract.html`.

Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv:2007.13605*, 2020.

Radu Ioan Boţ, Michael Sedlmayer, and Phan Tu Vuong. A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANs. *arXiv:2003.07886*, 2020.

Tatjana Chavdarova, Michael I Jordan, and Manolis Zampetakis. Last-iterate convergence of saddle point optimizers via high-resolution differential equations. *arXiv preprint arXiv:2112.13826*, 2021.

Patrick L Combettes and Teemu Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

Ernö Robert Csetnek, Yura Malitsky, and Matthew K Tam. Shadow Douglas–Rachford splitting for monotone inclusions. *Applied Mathematics & Optimization*, 80(3):665–678, 2019.

Cong D Dang and Guanghui Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Computational Optimization and applications*, 60(2):277–310, 2015.

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

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=SJJySbbAZ`.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5527–5540. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/3b2acfe2e38102074656ed938abf4ac3-Paper.pdf`.

Constantinos Daskalakis, Stratis Skoulakis, and Manolis Zampetakis. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1466–1478, 2021.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

Jelena Diakonikolas, Constantinos Daskalakis, and Michael Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1laEnA5Ym`.

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

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: $O(1/k)$ last-iterate convergence for monotone variational inequalities and connections with cocoercivity. *arXiv preprint arXiv:2110.04261*, 2021.

Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of the proximal point method for nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*, 2020. URL `https://arxiv.org/pdf/2006.08667.pdf`.

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pp. 4337–4348. PMLR, 2021.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16223–16234. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/ba9a56ce0a9bfa26e8ed9e10b2cc8f46-Paper.pdf`.

David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12: 747–756, 1976.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021a. URL `https://openreview.net/forum?id=AYAgKFl78z`.

Sucheol Lee and Donghwan Kim. Semi-anchored multi-step gradient descent ascent method for structured nonconvex-nonconcave composite minimax problems. *arXiv preprint arXiv:2105.15042*, 2021b.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 907–915. PMLR, 2019.

Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJxIm0VtwH`.

Mingrui Liu, Hassan Rafique, Qihang Lin, and Tianbao Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *The Journal of Machine Learning Research*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

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

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1): 383–410, 2020.

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

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg8jjC9KQ`.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

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

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Teemu Pennanen. On the range of monotone composite mappings. *Journal of Nonlinear and Convex Analysis*, 2(2), 2001.

Thomas Pethick, Puya Latafat, Panos Patrinos, Olivier Fercoq, and Volkan Cevher. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/pdf?id=2_vhkAMARk`.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, pp. 993–1019, 2013a.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pp. 3066–3074, 2013b.

Ernest K Ryu, Kun Yuan, and Wotao Yin. Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems and GANs. *arXiv preprint arXiv:1905.10899*, 2019.

Paul Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138, 1991.

John von Neumann. A model of general economic equilibrium. *The Review of Economic Studies*, 13(1):1–9, 1945.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv:2006.02032*, 2020.

Jun Yang and Hongwei Liu. A modified projected gradient method for monotone variational inequalities. *Journal of Optimization Theory and Applications*, 179(1):197–211, 2018.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

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

A Omitted proofs

A.1 OGDA+

For convenience we will sometimes use the notation g_k for $F(u_k)$ for all $k \geq -1$.

Lemma A.1. *Let (u_k) be the sequence of iterates generated by Algorithm 1, then*

$$\begin{aligned}\|u_{k+1} + ag_k - u^*\|^2 \leq \|u_k + ag_{k-1} - u^*\|^2 + a\gamma\rho\|g_k\|^2 + 4a\gamma^{-1}\langle g_k - g_{k-1}, u_k - u_{k+1}\rangle \\ - (2\gamma^{-1} - 1)\|u_{k+1} - u_k\|^2 - a^2(1 + 2\gamma^{-1})\|g_k - g_{k-1}\|^2.\end{aligned} \tag{9}$$

Proof. From the update of the method we deduce for all $k \geq 0$

$$\begin{aligned}\|u_{k+1} + ag_k - u^*\|^2 &= \|u_k - a\gamma g_k + ag_{k-1} - u^*\|^2 \\ &= \|u_k + ag_{k-1} - u^*\|^2 - 2\langle u_k - u^*, a\gamma g_k\rangle - \langle 2ag_{k-1} - a\gamma g_k, a\gamma g_k\rangle \\ &\leq \|u_k + ag_{k-1} - u^*\|^2 + a\gamma\rho\|g_k\|^2 - \langle 2ag_{k-1} - a\gamma g_k, a\gamma g_k\rangle,\end{aligned} \tag{10}$$

where we used the weak Minty assumption to deduce the last inequality. It remains to derive the following equality

$$\begin{aligned}-\langle 2ag_{k-1} - a\gamma g_k, a\gamma g_k\rangle = 4a\gamma^{-1}\langle g_k - g_{k-1}, u_k - u_{k+1}\rangle - (2\gamma^{-1} - 1)\|u_{k+1} - u_k\|^2 \\ -a^2(1 + 2\gamma^{-1})\|g_k - g_{k-1}\|^2.\end{aligned} \tag{11}$$

This can be seen by expressing every difference of iterates in terms of gradients according to Algorithm 1, giving

$$\begin{aligned}4a\gamma^{-1}\langle g_k - g_{k-1}, u_k - u_{k+1}\rangle &= 4a\gamma^{-1}\langle g_k - g_{k-1}, a(g_k - g_{k-1}) + a\gamma g_k\rangle \\ &= 4a^2\gamma^{-1}\|g_k - g_{k-1}\|^2 + 4a^2\langle g_k, g_k - g_{k-1}\rangle\end{aligned}$$

and

$$\begin{aligned}&- (2\gamma^{-1} - 1)\|u_{k+1} - u_k\|^2 \\ &\quad = (1 - 2\gamma^{-1})\|a(g_k - g_{k-1}) + a\gamma g_k\|^2 \\ &\quad = (a^2 - 2a^2\gamma^{-1})\|g_k - g_{k-1}\|^2 + (2\gamma a^2 - 4a^2)\langle g_k, g_k - g_{k-1}\rangle + (a^2\gamma^2 - 2a^2\gamma)\|g_k\|^2 \\ &\quad = (a^2 - 2a^2\gamma^{-1})\|g_k - g_{k-1}\|^2 + \langle g_k, (a^2\gamma^2 - 4a^2)g_k - (2\gamma a^2 - 4a^2)g_{k-1}\rangle.\end{aligned}$$

Adding the previous two equalities to $-a^2(1 + 2\gamma^{-1})\|g_k - g_{k-1}\|^2$ proves (11). Combining (10) and (11) proves the desired statement. □

We are actually going to show a slightly more general version of Theorem 3.1, which introduces an additional parameter λ. Note that for $\lambda = \gamma^{-1}$ we recover the original statement of Theorem 3.1. This allows us to cover the analysis of the monotone case in one proof. In particular $\lambda = 0$ yields the statement of Theorem 3.2.

Theorem A.1. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be L-Lipschitz continuous satisfying Assumption 1 with $a\lambda\gamma > \rho$ for some $0 \leq \lambda \leq 2\gamma^{-1}$ and $(u_k)_{k\geq 0}$ the sequence of iterates generated by Algorithm 1 with*

$$aL \leq \frac{2 - \lambda\gamma - \gamma}{2 - \lambda\gamma + \gamma}. \tag{12}$$

Then, for all $k \geq 0$

$$\frac{1}{k}\sum_{i=0}^{k-1}\|F(u_i)\|^2 \leq \frac{1}{ka\gamma(a\lambda\gamma-\rho)}\|u_0 + aF(u_0) - u^*\|^2.$$

In particular as long as $\rho < \frac{1}{L}$ we can find a γ small enough such that the above bound holds.

Proof. Using the definition of u_{k+1} we get

$$a^2\lambda\gamma^2\|g_k\|^2 = \lambda\|u_{k+1}-u_k\|^2 + \lambda\|a(g_k - g_{k-1})\|^2 + 2\lambda a\langle u_{k+1}-u_k, g_k - g_{k-1}\rangle. \tag{13}$$

Combining (13) and Lemma A.1 we deduce

$$\begin{aligned}\|u_{k+1} + ag_k - u^*\|^2 + a\gamma(a\lambda\gamma-\rho)\|g_k\|^2 \leq \|u_k + ag_{k-1} - u^*\|^2 - a^2(1+2\gamma^{-1}-\lambda)\|g_k - g_{k-1}\|^2 \\ - (2\gamma^{-1}-1-\lambda)\|u_{k+1}-u_k\|^2 + (4a\gamma^{-1}-2a\lambda)\langle u_k - u_{k+1}, g_k - g_{k-1}\rangle.\end{aligned} \tag{14}$$

Now, we get via Young's inequality

$$(4\gamma^{-1}a - 2a\lambda)\langle u_k - u_{k+1}, g_k - g_{k-1}\rangle \leq La(2\gamma^{-1}-\lambda)\|u_k - u_{k+1}\|^2 + L^{-1}a(2\gamma^{-1}-\lambda)\|g_k - g_{k-1}\|^2.$$

Combining the previous two inequalities

$$\begin{aligned}\|u_{k+1} + ag_k - u^*\|^2 + a\gamma(a\gamma\lambda-\rho)\|g_k\|^2 \leq \|u_k + ag_{k-1} - u^*\|^2 \\ + (aL^{-1}(2\gamma^{-1}-\lambda) - a^2(1+2\gamma^{-1}-\lambda))\|g_k - g_{k-1}\|^2 - (2\gamma^{-1}-1-\lambda - La(2\gamma^{-1}-\lambda))\|u_{k+1}-u_k\|^2\end{aligned} \tag{15}$$

We now use the Lipschitz continuity[2] of F to deduce that

$$\begin{aligned}\|u_{k+1} + ag_k - u^*\|^2 + a\gamma(a\gamma\lambda-\rho)\|g_k\|^2 \leq \|u_k + ag_{k-1} - u^*\|^2 \\ + (aL(2\gamma^{-1}-\lambda) - a^2L^2(1+2\gamma^{-1}-\lambda))\|u_k - u_{k-1}\|^2 - (2\gamma^{-1}-1-\lambda - La(2\gamma^{-1}-\lambda))\|u_{k+1}-u_k\|^2.\end{aligned}$$

The fact that the terms can be telescoped is, with $\alpha = La$ equivalent to

$$2\gamma^{-1} - 1 - \lambda - (2\alpha\gamma^{-1} - \alpha\lambda) \geq -\alpha^2(1+2\gamma^{-1}) + \alpha^2\lambda + (2\alpha\gamma^{-1} - \alpha\lambda)$$

which can be simplified to

$$\alpha^2(2+\gamma-\lambda\gamma) - \alpha(4-2\lambda\gamma) + 2 - \gamma - \lambda\gamma \geq 0$$

By solving for α we get the condition

$$\alpha \leq \frac{2-\lambda\gamma-\gamma}{2-\lambda\gamma+\gamma} \tag{16}$$

where from the condition $2\gamma^{-1} - 1 - \lambda - (2\alpha\gamma^{-1} - \alpha\lambda) \geq 0$ we deduce $\alpha \leq \frac{2-\lambda\gamma-\gamma}{2-\lambda\gamma}$, which is redundant in light of (16). The statement follows since we chose $u_0 = u_{-1}$. □

A.2 Improved bounds under monotonicity

Proof of Theorem 3.2. From Theorem A.1 we deduce via $\lambda = 0$

$$\|u_k + aF(u_{k-1}) - u^*\|^2 + C\sum_{i=0}^{k-1}\|u_{i+1} - u_i\|^2 \leq \|u_0 + aF(u_0) - u^*\|^2$$

with $C = \big(\alpha^2(2+\gamma) - 4\alpha + 2 - \gamma\big)$. Therefore, and since $aL < \frac{2-\gamma}{2+\gamma}$ implies that $C > 0$, the iterates remain bounded and $\lim_{k\to\infty}\|u_{k+1} - u_k\|^2 = 0$. Because

$$u_{k+1} - u_k + a^2(F(u_k) - F(u_{k-1})) = \gamma a^2 F(u_k)$$

every cluster point of (u_k) is a solution. By the lemma of Opial the entire sequence of iterates converges thus to a solution. □

[2]Strictly speaking we would have to assume that the term before $\|g_k - g_{k-1}\|^2$ is positive, but if it is not we can just discard it and be done with the proof.

A.3 OGDA+ stochastic

For the stochastic analysis we use for convenience the notation $\Delta_{k+1} = \mathbb{E}\big[\|u_{k+1} - u_k\|^2\big]$.

Lemma A.2. *Let (u_k) be the sequence of iterates generated by stochastic OGDA+, then*

$$L_{k+1} + a(a-\rho)\mathbb{E}\big[\|g_k\|^2\big] \leq L_k + 16\sigma_k + \tilde{C}_1\Delta_k - \tilde{C}_2\Delta_{k+1}.$$

where $\tilde{C}_1 := \frac{4}{3}\max\{0, aL\gamma^{-1} - a^2L^2(1+\gamma^{-1})\}$ and $\tilde{C}_2 := \gamma^{-1} - 1 - aL\gamma^{-1}$.

Proof of Lemma A.2. From the update of the method we deduce for all $k \geq 0$

$$\begin{aligned}\mathbb{E}\big[\|u_{k+1} + a\tilde{g}_k - u^*\|^2\big] &= \|u_k - a\gamma\tilde{g}_k + a\tilde{g}_{k-1} - u^*\|^2 \\ &= \mathbb{E}\big[\|u_k + a\tilde{g}_{k-1} - u^*\|^2 - \langle 2a\tilde{g}_{k-1} - a\tilde{g}_k, a\gamma\tilde{g}_k\rangle\big] - 2a\mathbb{E}[\langle u_k - u^*, \gamma\tilde{g}_k\rangle] \\ &\leq \mathbb{E}\big[\|u_k + a\tilde{g}_{k-1} - u^*\|^2 + a\gamma\rho\|g_k\|^2 - \langle 2a\tilde{g}_{k-1} - a\gamma\tilde{g}_k, a\gamma\tilde{g}_k\rangle\big],\end{aligned} \tag{17}$$

where we used

$$2\mathbb{E}[\langle u_k - u^*, \tilde{g}_k\rangle] = 2\mathbb{E}[\mathbb{E}[\langle u_k - u^*, \tilde{g}_k\rangle \,|\, u_k]] = 2\mathbb{E}[\langle u_k - u^*, g_k\rangle] \geq -\rho\mathbb{E}\big[\|g_k\|^2\big]$$

to deduce the last inequality. It remains to note that

$$\begin{aligned}-\langle 2a\tilde{g}_{k-1} - a\gamma\tilde{g}_k, a\gamma\tilde{g}_k\rangle = 4a\gamma^{-1}\langle \tilde{g}_k - \tilde{g}_{k-1}, u_k - u_{k+1}\rangle - (2\gamma^{-1} - 1)\|u_{k+1} - u_k\|^2 \\ -a^2(1+2\gamma^{-1})\|\tilde{g}_k - \tilde{g}_{k-1}\|^2,\end{aligned} \tag{18}$$

which follows immediately the way we deduced (11). Now using the definition of u_{k+1} we get

$$\begin{aligned}a^2\gamma\mathbb{E}\big[\|g_k\|^2\big] = a^2\gamma\mathbb{E}\big[\|\mathbb{E}[\tilde{g}_k \,|\, u_k]\|^2\big] \leq a^2\gamma\mathbb{E}\big[\|\tilde{g}_k\|^2\big] = \mathbb{E}\big[\gamma^{-1}\|u_{k+1} - u_k\|^2\big] \\ +\mathbb{E}\big[\gamma^{-1}\|a(\tilde{g}_k - \tilde{g}_{k-1})\|^2 + 2a\gamma^{-1}\langle u_{k+1} - u_k, \tilde{g}_k - \tilde{g}_{k-1}\rangle\big].\end{aligned} \tag{19}$$

Combining (17) to (19) we deduce

$$\begin{aligned}\mathbb{E}\big[\|u_{k+1} + a\tilde{g}_k - u^*\|^2\big] + a\gamma(a-\rho)\mathbb{E}\big[\|g_k\|^2\big] \leq \|u_k + a\tilde{g}_{k-1} - u^*\|^2 - a^2(1+\gamma^{-1})\|\tilde{g}_k - \tilde{g}_{k-1}\|^2 \\ - (\gamma^{-1} - 1)\|u_{k+1} - u_k\|^2 + 2a\gamma^{-1}\langle u_k - u_{k+1}, \tilde{g}_k - \tilde{g}_{k-1}\rangle\end{aligned}$$

We get via Young's inequality

$$2a\gamma^{-1}\langle u_k - u_{k+1}, \tilde{g}_k - \tilde{g}_{k-1}\rangle \leq a\gamma^{-1}L\|u_k - u_{k+1}\|^2 + L^{-1}a\gamma^1\|\tilde{g}_k - \tilde{g}_{k-1}\|^2.$$

Combining the previous two inequalities

$$\begin{aligned}\mathbb{E}\big[\|u_{k+1} + a\tilde{g}_k - u^*\|^2\big] + a\gamma(a-\rho)\mathbb{E}\big[\|g_k\|^2\big] \leq \mathbb{E}\big[\|u_k + a\tilde{g}_{k-1} - u^*\|^2\big] \\ + \mathbb{E}\big[(L^{-1}a\gamma^{-1} - a^2(1+\gamma^{-1}))\|\tilde{g}_k - \tilde{g}_{k-1}\|^2 - (\gamma^{-1} - 1 - La\gamma^{-1})\|u_{k+1} - u_k\|^2\big]\end{aligned} \tag{20}$$

Next, we need to estimate the difference of the gradient estimators via the difference of the true

$$\begin{aligned}\|\tilde{g}_k - \tilde{g}_{k-1}\|^2 &\leq \left(1 + \frac{1}{3}\right)\|g_k - g_{k-1}\|^2 + 4\|g_k - \tilde{g}_k + \tilde{g}_{k-1} - g_{k-1}\|^2 \\ &\leq \frac{4}{3}L^2\|u_k - u_{k-1}\|^2 + 8\|g_k - \tilde{g}_k\|^2 + 8\|\tilde{g}_{k-1} - g_{k-1}\|^2,\end{aligned}$$

where we used the Lipschitz continuity of the operator. Therefore, by taking the expectation, we obtain

$$\mathbb{E}\big[\|\tilde{g}_k - \tilde{g}_{k-1}\|^2\big] \leq \frac{4}{3}L^2\mathbb{E}\big[\|u_k - u_{k-1}\|^2\big] + 8(\sigma_{k-1}^2 + \sigma_k^2). \tag{21}$$

Plugging (21) into (20) we deduce

$$\begin{aligned}\mathbb{E}\big[\|u_{k+1} + ag_k - u^*\|^2\big] + a\gamma(a-\rho)\mathbb{E}\big[\|g_k\|^2\big] \leq \mathbb{E}\big[\|u_k + ag_{k-1} - u^*\|^2\big] + 16(L^{-1}a\gamma^{-1} - a^2(1+\gamma^{-1}))\sigma_{k-1}^2 \\ + \max\left\{0, \frac{4}{3}(La\gamma^{-1} - a^2L^2(1+\gamma^{-1}))\right\}\Delta_k - (\gamma^{-1} - 1 - aL\gamma^{-1})\Delta_{k+1}.\end{aligned}$$

□

Proof of Theorem 3.3. Simple algebra shows that (5) ensures that $\tilde{C}_2 \geq \tilde{C}_1$. So we can iteratively apply the statement of Lemma A.2 to deduce

$$a\gamma(a-\rho)\sum_{i=0}^{k-1}\|g_i\|^2 \leq \|u_1 + ag_0 - u^*\|^2 + \tilde{C}_1\|u_1 - u_0\|^2 + 16\sum_{i=0}^{k-1}\sigma_i^2.$$

By taking per iteration $B := \max\{1, \frac{16}{\sigma\varepsilon}\}$ independent samples per iteration, i.e. replacing $\tilde{F}(u_k)$ by

$$\frac{1}{B}\sum_{i=1}^{B}\tilde{F}(u_k), \tag{22}$$

we arrive at a total oracle call complexity as claimed. □

A.4 EG+ with adaptive step size

Lemma A.3. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be L-Lipschitz and satisfy Assumption 1. Then, for the iterates generated by Algorithm 2, it holds that*

$$\begin{aligned}\gamma^{-1}\|\bar{u}_{k+1} - u^*\|^2 \leq \gamma^{-1}\|\bar{u}_k - u^*\|^2 - a_k(a_k\gamma(1-\gamma)-\rho)\|F(u_k)\|^2 \\ - \Big(1 - \frac{\tau a_k}{a_{k+1}}\Big)(\|u_k - \bar{u}_k\|^2 + \|u_k - \bar{u}_{k+1}\|^2).\end{aligned} \tag{23}$$

Proof of Lemma A.3. We start by splitting the following term into three

$$\langle a_kF(u_k), u_k - u^*\rangle = \langle a_kF(u_k), \bar{u}_{k+1} - u^*\rangle + \langle a_kF(\bar{u}_k), u_k - \bar{u}_{k+1}\rangle + a_k\langle F(u_k) - F(\bar{u}_k), u_k - \bar{u}_{k+1}\rangle. \tag{24}$$

By expressing $F(u_k)$ via the definition of $\bar{u}_{k+1}$ and the three point identity we obtain

$$\begin{aligned}\langle a_kF(u_k), \bar{u}_{k+1} - u^*\rangle &= \gamma^{-1}\langle \bar{u}_k - \bar{u}_{k+1}, \bar{u}_{k+1} - u^*\rangle \\ &= \frac{\gamma^{-1}}{2}(\|\bar{u}_k - u^*\|^2 - \|\bar{u}_{k+1} - \bar{u}_k\|^2 - \|\bar{u}_{k+1} - u^*\|^2).\end{aligned} \tag{25}$$

Similarly, by expressing $F(\bar{u}_k)$ via the definition of u_k we deduce

$$\begin{aligned}\langle a_kF(\bar{u}_k), u_k - \bar{u}_{k+1}\rangle &= \langle \bar{u}_k - u_k, u_k - \bar{u}_{k+1}\rangle \\ &= \frac{1}{2}\left(\|\bar{u}_{k+1} - \bar{u}_k\|^2 - \|u_k - \bar{u}_k\|^2 - \|u_k - \bar{u}_{k+1}\|^2\right).\end{aligned} \tag{26}$$

Lastly, via the Cauchy-Schwarz inequality

$$\begin{aligned}a_k\langle F(u_k) - F(\bar{u}_k), u_k - \bar{u}_{k+1}\rangle &\leq a_k\|F(u_k) - F(\bar{u}_k)\|\|u_k - \bar{u}_{k+1}\| \\ &\overset{(6)}{\leq} \frac{a_k\tau}{a_{k+1}}\|u_k - \bar{u}_k\|\|u_k - \bar{u}_{k+1}\| \\ &\leq \frac{a_k\tau}{2a_{k+1}}(\|u_k - \bar{u}_k\|^2 + \|u_k - \bar{u}_{k+1}\|^2).\end{aligned} \tag{27}$$

Combining (24) to (27) and Assumption 1, together with the observation that $\gamma a_kF(u_k) = \bar{u}_{k+1} - \bar{u}_k$, gives

$$\begin{aligned}\gamma^{-1}\|\bar{u}_{k+1} - u^*\|^2 + a_k(a_k\gamma(1-\gamma)-\rho)\|F(u_k)\|^2 \leq \gamma^{-1}\|\bar{u}_k - u^*\|^2 \\ - \Big(1 - \frac{\tau a_k}{a_{k+1}}\Big)(\|u_k - \bar{u}_k\|^2 + \|u_k - \bar{u}_{k+1}\|^2).\end{aligned}$$

We see that the best possible dependence between step size and ρ is achieved for $\gamma = \frac{1}{2}$. □

Proof of Theorem 4.1. As a_k converges to a_∞, the quotient a_k/a_{k+1} converges to 1. In particular, there exists an index $k_0 \in \mathbb{N}$ such that $a_k/a_{k+1} \leq \frac{1}{\tau}$ for all $k \geq k_0$, because $\tau < 1$. We can therefore drop the last term in (23) and sum up to obtain

$$\|\bar{u}_{k+1} - u^*\|^2 + \sum_{i=k_0}^{k} a_i \left(\frac{a_i}{2} - \rho\right) \|F(u_k)\|^2 \leq \|\bar{u}_{k_0} - u^*\|^2.$$

The desired statement follows by observing that $a_i \geq a_\infty \geq \tau/L$. □

Note that the above proof for the adaptive version of EG+ provides an improvement in the dependence between ρ and L over the analysis Diakonikolas et al. (2021) even in the constant step size regime.

B Additional statements and proofs

For the sake of completeness we provide a proof of the elementary fact that Minty solutions are a stronger requirement than Stampachia solutions.

Lemma B.1. *If F is continuous then every Minty solution is also a Stampacchia solution.*

Proof. Let w^* be a solution to the Minty VI and $z = \alpha w^* + (1-\alpha)u$ for an arbitrary $u \in \mathbb{R}^m$ and $\alpha \in (0,1)$, then

$$\langle F(\alpha w^* + (1-\alpha)u), (1-\alpha)(u - w^*)\rangle \geq 0.$$

This implies that

$$(1-\alpha)\langle F(\alpha w^* + (1-\alpha)u), (u - w^*)\rangle \geq 0.$$

By dividing by $(1-\alpha)$ and then taking the limit $\alpha \to 1$ we obtain that w^* is a solution of the Stampacchia formulation. □

C Numerics

For all experiments, if not specified otherwise, we used for OGDA+ and the adaptive version of EG+ the parameter $\gamma = \frac{1}{2}$. For the step size choice of Algorithm 2 we use $\tau = 0.99$. For the CurvatureEG+ method of Pethick et al. (2022) we use $\delta = -0.49$.

C.1 Ratio game

The data used to generate the instance displayed in Figure 3 was suggested in Daskalakis et al. (2020) and is given by

$$R = \begin{pmatrix} -0.6 & -0.3 \\ 0.6 & -0.3 \end{pmatrix}$$

and

$$S = \begin{pmatrix} 0.9 & 0.5 \\ 0.8 & 0.4 \end{pmatrix}.$$

This results in the following objective function for the min-max problem

$$V(x,y) = (-1.2xy + 0.9y - 0.3)/(0.4y + 0.1x + 0.4),$$

which gives rise to the optimality conditions

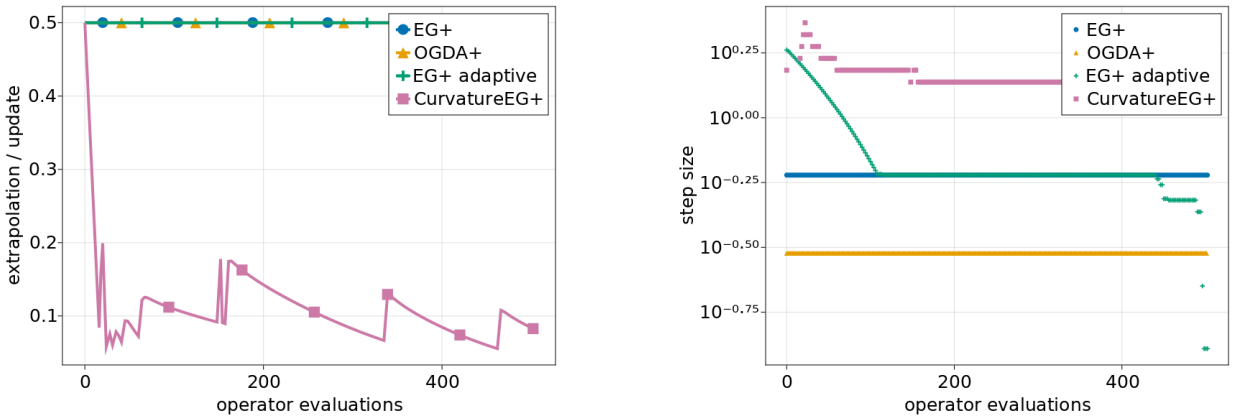
$$-0.12x^* - 0.39x + 0.48 = 0$$

and

$$-0.48y^2 - 0.57y + 0.03 = 0$$

with the solution $(x^*, y^*) = (0.951941, 0.050485)$. For the experiments we used an estimated Lipschitz constant $L = \frac{5}{3}$.

In Figure 5 we can see the reason for the slow convergence behavior of the CurvatureEG+ method observed in Figure 3. Not only is the step size computed by the backtracking procedure smaller than the one we use for adaptive EG+, see (6), also the second step (update step) uses an even smaller fraction of the already smaller extrapolation step size.



(a) Ratio of extrapolation to update step

(b) Extrapolation step size.

Figure 5: **Ratio game**. An illustration of the step sizes used by different methods, as well as the ratio of extrapolation to update step. CurvatureEG+ chooses its own ratio adaptively and does so for this example in a seemingly overly conservative way, resulting in slow convergence observed in Figure 3.

C.1.1 Polar Game

For Figure 2 we used the so-called Polar Game introduced in Pethick et al. (2022) which is given by

$$F(x, y) = (\psi(x, y) - y, \psi(y, x) + x), \tag{28}$$

where $\psi(x, y) = \frac{1}{16}ax(-1 + x^2 + y^2)(-9 + 16x^2 + 16y^2)$ and parameter $a > 0$. In Figure 2 we used $a = \frac{1}{3}$.

C.2 Lower bound example

The following min-max problem was introduced in Pethick et al. (2022) as a lower bound on the dependence between ρ and L for EG+:

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \xi xy + \frac{\zeta}{2}(x^2 - y^2). \tag{29}$$

In particular Theorem 3.4 from Pethick et al. (2022) states that EG+ (with $\gamma = \frac{1}{2}$) and constant step size $a = \frac{1}{L}$ converges if and only if $(0, 0)$ is a weak Minty solution with $\rho < \frac{1}{2L}$ where ρ and L can be computed explicitly in the above example and are given by

$$L = \sqrt{\xi^2 + \zeta^2} \quad \text{and} \quad \rho = 2\frac{\zeta^2}{\xi^2 + \zeta^2}. \tag{30}$$

Figure 1 is obtained by choosing ξ and ζ such that we get exactly $\rho = \frac{1}{2L}$ and the theory therefore predicting divergence of EG+, which is exactly what is empirically observed. Although, the general upper bound proved in Theorem 3.1 only states convergence in the case $\rho < \frac{1}{4L}$, we observe rapid convergence of OGDA+ for this example showcasing that it can drastically outperform EG+ in some scenarios.

For the experiments in Figure 1 we use $\xi = 1$ and $\zeta = -\sqrt{15}$, which gives $\rho = \frac{1}{2L}$ according to (30)