Addressing Min-Max Challenges in Nonconvex-Nonconcave Problems with Solutions Exhibiting Weak Minty Properties

Abstract

This research examines a specific category of structured nonconvex-nonconcave min-max problems that demonstrate a characteristic known as weak Minty solutions. This concept, which has only recently been defined, has already demonstrated its effectiveness by encompassing various generalizations of monotonicity at the same time. We establish new convergence findings for an enhanced variant of the optimistic gradient method (OGDA) within this framework, achieving a convergence rate of 1/k for the most effective iteration, measured by the squared operator norm, a result that aligns with the extragradient method (EG). Furthermore, we introduce a modified version of EG that incorporates an adaptive step size, eliminating the need for prior knowledge of the problem's specific parameters.

1 Introduction

The recent advancements in machine learning models, particularly those that can be formulated as min-max optimization problems, have generated significant interest in saddle point problems. Examples of these models include generative adversarial networks, adversarial learning frameworks, adversarial example games, and actor-critic methods. While practical methods have been developed that generally perform well, the theoretical understanding of scenarios where the objective function is nonconvex in the minimization component and nonconcave in the maximization component remains limited, with some research even suggesting intractability in certain cases.

A specific subset of nonconvex-nonconcave min-max problems was analyzed, and it was found that the extragradient method (EG) exhibited favorable convergence behavior in experimental settings. Surprisingly, these problems did not appear to possess any of the recognized favorable characteristics, such as monotonicity or Minty solutions. Subsequently, a suitable concept was identified (see Assumption 1), which is less restrictive than the presence of a Minty solution (a condition frequently employed in the existing literature) and also extends the idea of negative comonotonicity. Because of these properties that unify and generalize, the concept of weak Minty solutions was quickly investigated.

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

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

Moreover, it has been demonstrated that a modified version of EG is capable of addressing problems with such solutions, achieving a complexity of $O(\epsilon^{-1})$ for the squared operator norm. This adaptation, referred to as EG+, is based on a bold extrapolation step followed by a cautious update step. A similar step size approach has been previously examined in the context of a stochastic variant of EG.

In a similar vein, we explore a variation of the optimistic gradient descent ascent (OGDA), also known as Forward-Reflected-Backward (FoRB). We address the following question with an affirmative answer:

Can OGDA achieve convergence guarantees comparable to those of EG when dealing with weak Minty solutions?

Specifically, we demonstrate that a modified version of the OGDA method, defined for a step size $a > 0$ and a parameter $0 < \gamma \leq 1$ as follows:

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

can achieve the same convergence bounds as EG+ by requiring only a single gradient oracle call in each iteration.

It is worth noting that OGDA is most frequently expressed in a form where $\gamma = 1$. However, two recent studies have examined a more generalized coefficient. While these earlier studies focused on the monotone setting, the true significance of γ becomes

apparent only when dealing with weak Minty solutions. In this context, we find that γ must be greater than 1 to ensure convergence, a phenomenon that is not observed in monotone problems.

When examining a general smooth min-max problem:

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

the operator F mentioned in Assumption 1 naturally emerges as $F(u) := [\nabla_x f(x, y), -\nabla_y f(x, y)]$ with $u = (x, y)$. However, by examining saddle point problems from the broader viewpoint of variational inequalities (VIs) through the operator F, we can concurrently address more scenarios, such as certain equilibrium problems.

The parameter ρ in the definition of weak Minty solutions (1) is crucial for both the analysis and the experiments. Specifically, it is essential that the step size exceeds a value proportional to ρ. Simultaneously, as is typical, the step size is limited from above by the inverse of the Lipschitz constant of F. For instance, since some researchers require the step size to be less than $\frac{1}{4L}$, their convergence claim is valid only if $\rho < \frac{1}{4L}$. This condition was later improved to $\rho < \frac{1}{2L}$ for the choice $\gamma = 1$ and to $\rho < \frac{1}{L}$ for even smaller values of γ. As in the monotone setting, OGDA 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.

1.1 Contribution

Our contributions are summarized as follows:

1. We establish a new convergence rate of $O(1/k)$, measured by the squared operator norm, for a modified version of OGDA, which we call OGDA+. This rate matches that of EG and builds upon the recently introduced concept of weak solutions to the Minty variational inequality.
2. Even when a stronger condition is imposed, specifically that the operator is also monotone, we enhance the range of feasible step sizes for OGDA+ and obtain the most favorable result known for the standard method ($\gamma = 1$).
3. We demonstrate a complexity bound of $O(\epsilon^{-2})$ for a stochastic variant of the OGDA+ method.
4. We also introduce an adaptive step size version of EG+. This version achieves the same convergence guarantees without requiring any knowledge of the Lipschitz constant of the operator F. Consequently, it can potentially take larger steps in areas with low curvature, enabling convergence where a fixed step size strategy might fail.

1.2 Related literature

We will concentrate on the nonconvex-nonconcave setting, as there is a substantial body of work 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, convex-nonconcave, or under the Polyak-Łojasiewicz assumption.

Weak Minty. It was observed that a specific parameterization of the von Neumann ratio game exhibits a novel type of solution, termed "weak Minty," without having any of the previously known characteristics like (negative) comonotonicity or Minty solutions. Convergence in the presence of such solutions was demonstrated for EG, provided that the extrapolation step size is twice as large as the update step. Subsequently, it was shown that the condition on the weak Minty parameter can be relaxed by further reducing the length of the update step, and this is done adaptively. To avoid the need for additional hyperparameters, a backtracking line search is also proposed, which may incur extra gradient computations or require second-order information (in contrast to the adaptive step size we propose in Algorithm 3). A different approach is taken by focusing on the min-max setting and using multiple ascent steps per descent step, achieving the same $O(1/k)$ rate as EG.

Minty solutions. Numerous studies have presented various methods for scenarios where the problem at hand has a Minty solution. It was shown that weakly monotone VIs can be solved by iteratively adding a quadratic proximity term and repeatedly optimizing the resulting strongly monotone VI using any convergent method. The convergence of the OGDA method was proven, but without a specific rate. It was noted that the convergence proof for the golden ratio algorithm (GRAAL) is valid without any changes. While the assumption that a Minty solution exists is a generalization of the monotone setting, it is challenging to find non-monotone problems that possess such solutions. In our setting, as per Assumption 1, the Minty inequality (MVI) can be violated at any point by a factor proportional to the squared operator norm.

Negative comonotonicity. Although previously studied under the term "cohypomonotonicity," the concept of negative comonotonicity has recently been explored. It offers a generalization of monotonicity, but in a direction distinct from the concept of Minty solutions, and only a limited number of studies have examined methods in this context. An anchored version of EG was studied, and an improved convergence rate of $O(1/k^2)$ (in terms of the squared operator norm) was shown. Similarly, an accelerated version of the reflected gradient method was investigated. Whether such acceleration is possible in the more general setting of weak Minty solutions remains an open question (any Stampacchia solution to the VI given by a negatively comonotone operator is a weak Minty solution). Another intriguing observation was made, where for cohypomonotone problems, a monotonically decreasing gradient norm was demonstrated when using EG. However, we did not observe this in our experiments, emphasizing the need to differentiate this class from problems with weak Minty solutions.

Interaction dominance. The concept of α-interaction dominance for nonconvex-nonconcave min-max problems was investigated, and it was shown that the proximal-point method converges sublinearly if this condition is met in y and linearly if it is met in both components. Furthermore, it was demonstrated that if a problem is interaction dominant in both components, it is also negatively comonotone.

Optimism. The positive effects of introducing the simple modification commonly known as optimism have recently attracted the attention of the machine learning community. Its name comes from online optimization. The idea dates back even further and has also been studied in the mathematical programming community.

2 Preliminaries

2.1 Notions of solution

We outline the most frequently used solution concepts in the context of variational inequalities (VIs) and related areas. These concepts are typically defined with respect to a constraint set $C \subseteq \mathbb{R}^d$. A Stampacchia solution of the VI given by $F : \mathbb{R}^d \to \mathbb{R}^d$ is a point u^* such that:

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

In this work, we only consider the unconstrained case where $C = \mathbb{R}^d$, and the above condition simplifies to $F(u^*) = 0$. Closely related is the following concept: A Minty solution is a point $u^* \in C$ such that:

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

For a continuous operator F, a Minty solution of the VI is always a Stampacchia solution. The converse is generally not true but holds, for example, if the operator F is monotone. Specifically, there are nonmonotone problems with Stampacchia solutions but without any Minty solutions.

2.2 Notions of monotonicity

This section aims to revisit some fundamental and more contemporary concepts of monotonicity and the relationships between them. An operator F is considered monotone if:

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

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

Two frequently studied notions that fall into this category are strongly monotone operators, which satisfy:

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

and cocoercive operators, which fulfill:

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

Strongly monotone operators emerge as gradients of strongly convex functions or in strongly-convex-strongly-concave min-max problems. Cocoercive operators appear, for instance, as gradients of smooth convex functions, in which case (2) holds with β equal to the inverse of the gradient's Lipschitz constant.

Departing from monotonicity. Both of the aforementioned subclasses of monotonicity can serve as starting points for exploring the non-monotone domain. Given that general non-monotone operators may display erratic behavior, such as periodic cycles and spurious attractors, it is reasonable to seek settings that extend the monotone framework while remaining manageable. First and foremost is the extensively studied setting of ν-weak monotonicity:

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

Such operators arise as the gradients of the well-studied class of weakly convex functions, a rather general class of functions as it includes all functions without upward cusps. In particular, every smooth function with a 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 and is given by:

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

Clearly, if a Stampacchia solution exists for such an operator, then it also fulfills Assumption 1.

Behavior with respect to the solution. While the above properties are standard assumptions in the literature, it is usually sufficient to require the corresponding condition to hold when one of the arguments is a (Stampacchia) solution. This means that instead of monotonicity, it is enough to ask for the operator F to be star-monotone, i.e.,

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

or star-cocoercive,

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

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, which we denote with a "+" to emphasize 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 $0 < \gamma < 1$.
for $k = 0, 1, ...$ **do**
$u_{k+1} = u_k - a((1+\gamma)F(u_k) - F(u_{k-1}))$
end for

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 let $(u_k)_{k\geq 0}$ 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 $k \geq 0$,

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

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 $\rho < a$. To be able to choose a large step size a, we must decrease γ, as evident from (3). However, this degrades the algorithm's speed by making the update steps smaller. The same effect can be observed for EG+ and is therefore not surprising. One could derive an optimal γ (i.e., minimizing the right-hand side) from Theorem 3.1, but this results in a non-intuitive cubic dependence on ρ. In practice, the strategy of decreasing γ until convergence is achieved, but not further, yields reasonable results.

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

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 $aL = \frac{2-\gamma}{2+\gamma} - \epsilon$ for $\epsilon > 0$, then the iterates generated by OGDA+ fulfill

$$\min_{i=0,\dots,k-1} \|F(u_i)\|^2 \leq \frac{2}{ka^2\gamma^2\epsilon}\|u_0 + aF(u_0) - u^*\|^2.$$

In particular, we can choose $\gamma = 1$ and $a < \frac{1}{2L}$.

There are different works discussing the convergence of OGDA in terms of the iterates or a gap function with $a < \frac{1}{2L}$. However, we want to compare the above bound to more similar results on rates for the best iterate in terms of the operator norm. The same rate as ours for OGDA is shown, but requires the conservative step size bound $a \leq \frac{1}{16L}$. This was later improved to $a \leq \frac{1}{3L}$. All of these only deal with the case $\gamma = 1$. The only other reference that deals with a generalized (i.e., not necessarily $\gamma = 1$) version of OGDA is another work, where the resulting step size condition is $a \leq \frac{2-\gamma}{4L}$, which is strictly worse than ours for any γ. To summarize, not only do we show for the first time that the step size of a generalization of OGDA can go above $\frac{1}{2L}$, but we also provide the least restrictive bound for any value of γ.

3.2 OGDA+ stochastic

In this section, we discuss the setting where, instead of the exact operator F, we only have access to a collection of independent estimators $F(\cdot, \xi_i)$ at every iteration. We assume here that the estimator F is unbiased, i.e., $\mathbb{E}[F(u_k, \xi)|u_{k-1}] = F(u_k)$, and has bounded variance $\mathbb{E}[\|F(u_k, \xi) - F(u_k)\|^2] \leq \sigma^2$. We show that we can still guarantee convergence by using batch sizes B of order $O(\epsilon^{-1})$.

Algorithm 2 stochastic OGDA+
Require: Starting point $u_0 = u_{-1} \in \mathbb{R}^d$, step size $a > 0$, parameter $0 < \gamma \leq 1$ and batch size B.
for $k = 0, 1, ...$ **do**
Sample i.i.d. $(\xi_i)_{i=1}^B$ and compute estimator $\tilde{g}_k = \frac{1}{B}\sum_{i=1}^B F(u_k, \xi_i^k)$
$u_{k+1} = u_k - a((1+\gamma)\tilde{g}_k - \tilde{g}_{k-1})$
end for

Theorem 3.3. Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be L-Lipschitz satisfying Assumption 1 with $\frac{1}{L} > \rho$, and let $(u_k)_{k\geq 0}$ be the sequence of iterates generated by stochastic OGDA+, with a and γ satisfying $\rho < a < \frac{1-\gamma}{1+\gamma}\frac{1}{L}$. Then, to visit an ϵ-stationary point such that $\min_{i=0,\dots,k-1} \mathbb{E}[\|F(u_i)\|^2] < \epsilon$, we require

$$\frac{1}{ka\gamma(a-\rho)}\|u_0 + a\tilde{g}_0 - u^*\|^2 \max\left\{1, \frac{4\sigma^2}{aL\epsilon}\right\}$$

calls to the stochastic oracle $\tilde{F}$, with large batch sizes of order $O(\epsilon^{-1})$.

In practice, large batch sizes of order $O(\epsilon^{-1})$ are typically not desirable; instead, a small or decreasing step size is preferred. In the weak Minty setting, this causes additional trouble due to the necessity of large step sizes to guarantee convergence. Unfortunately, the current analysis does not allow for variable γ.

4 EG+ with adaptive step sizes

In this section, we present Algorithm 3, which 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 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, meaning that no step size needs to be supplied by the user and no line-search is carried out.

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 (4). Due to 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 a few with especially interesting properties. In particular, we do not want to touch on methods with a linesearch procedure, which typically result in multiple gradient computations per iteration.

We use a simple and therefore widely used step size choice that naively estimates the local Lipschitz constant and forces a monotone decreasing behavior. Such step sizes have been used extensively for monotone VIs and similarly in the context of the mirror-prox method, which corresponds to EG in the setting of (non-Euclidean) Bregman distances.

A version of EG with a different adaptive step size choice has been investigated, 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 that is allowed to use nonmonotone step sizes to treat VIs and does not use a possibly costly linesearch is the golden ratio algorithm. 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 the existence of Minty solutions, a quantitative convergence result is still open.

Algorithm 3 EG+ with adaptive step size

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

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

Find the step size:

$$a_k = \min\left\{a_{k-1}, \frac{\tau\|\bar{u}_k - \bar{u}_{k-1}\|}{\|F(\bar{u}_k) - F(\bar{u}_{k-1})\|}\right\} \tag{4}$$

Compute next iterate:

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

end for

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 satisfies Assumption 1, where u^* denotes any weak Minty solution, with $a_\infty > 2\rho$, and let $(u_k)_{k\geq 0}$ be the iterates generated by Algorithm 3 with $\gamma = \frac{1}{2}$ and $\tau \in (0,1)$. Then, there exists a $k_0 \in \mathbb{N}$ such that

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

Algorithm 3 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 (4) 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, visit 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+ 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 a long time 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. Even for monotone problems where this type of step size has been proposed, this problem could not be circumvented, and authors instead focused on the convergence of the iterates without any rate.

5 Numerical experiments

In the following, we compare the EG+ method with the two methods we propose: OGDA+ and EG+ with adaptive step size (see Algorithm 1 and Algorithm 3, respectively). Last but not least, we also include the CurvatureEG+ method, which is a modification of EG+ that adaptively chooses the ratio of extrapolation and update steps. In addition, a backtracking linesearch is performed with an initial guess made by second-order information, whose extra cost we ignore in the experiments.

5.1 Von Neumann's ratio game

We consider von Neumann's ratio game, which is given by:

$$\min_{x \in \Delta_m} \max_{y \in \Delta_n} V(x, y) = \frac{\langle x, Ry \rangle}{\langle x, Sy \rangle}, \tag{5}$$

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 $\Delta := \{z \in \mathbb{R}^d : z_i > 0, \sum_{i=1}^d z_i = 1\}$ denoting the unit simplex. Expression (5) can be interpreted as the value $V(x, y)$ for a stochastic game with a single state and mixed strategies.

We see an illustration of a particularly difficult instance of (5). Interestingly, we still observe good convergence behavior, although an estimated ρ is more than ten times larger than the estimated Lipschitz constant.

5.2 Forsaken

A particularly difficult min-max toy example with a "Forsaken" solution was proposed and is given by:

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} x(y - 0.45) + \phi(x) - \phi(y), \tag{6}$$

where $\phi(z) = \frac{1}{6}z^6 - \frac{2}{4}z^4 + \frac{1}{4}z^2 - \frac{1}{2}z$. 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, it was also observed that the limit cycle closer to the solution repels possible trajectories of iterates, thus "shielding" the solution. Later, it was noticed that, restricted to the box $\|(x, y)\|_\infty < 3$, the above-mentioned solution is weak Minty with $\rho \geq 2 \cdot 0.477761$, which is much larger than $\frac{1}{2L} \approx 0.08$. In line with these observations, we can see that none of the fixed step size methods with a step size bounded by $\frac{1}{L}$ converge. In light of this observation, a backtracking linesearch was proposed, which potentially allows for larger steps than predicted by the global Lipschitz constant. Similarly, our proposed adaptive step size version of EG+ (see Algorithm 3) is also able to break through the repelling limit cycle and converge to the solution. On top of this, it does so at a faster rate and without the need for additional computations in the backtracking procedure.

5.3 Lower bound example

The following min-max problem was introduced as a lower bound on the dependence between ρ and L for EG+:

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \mu x y + \frac{\zeta}{2}(x^2 - y^2). \tag{7}$$

In particular, it was stated that EG+ (with any γ) and constant step size $a = \frac{1}{L}$ converges for this problem if and only if $(0, 0)$ is a weak Minty solution with $\rho < \frac{1-\gamma}{L}$, where ρ and L can be computed explicitly in the above example and are given by:

$$L = \sqrt{\mu^2 + \zeta^2} \quad \text{and} \quad \rho = \frac{\mu^2 - \zeta^2}{2\mu}.$$

By choosing $\mu = 3$ and $\zeta = -1$, we get exactly $\rho = \frac{1}{L}$, therefore predicting divergence of EG+ for any γ, which is exactly what is empirically observed. Although the general upper bound proved in Theorem 3.1 only states convergence in the case $\rho < \frac{1}{L}$, we observe rapid convergence of OGDA+ for this example, showcasing that it can drastically outperform EG+ in some scenarios.

6 Conclusion

Many intriguing questions persist in the domain of min-max problems, particularly when departing from the convex-concave framework. Very recently, it was demonstrated that the $O(1/k)$ bounds on the squared operator norm for EG and OGDA for the last iterate (and not just the best one) are valid even in the negatively comonotone setting. Deriving a comparable statement in the presence of merely weak Minty solutions remains an open question.

In general, our analysis and experiments seem to suggest that there is minimal benefit in employing OGDA+ over EG+ for the majority of problems, as the reduced iteration cost is counterbalanced by the smaller step size. An exception is presented by problem (7), which is not covered by theory, and OGDA+ is the only method capable of converging.

Finally, we note that the 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 gradient's Lipschitz constant, does not appear to hold true for min-max problems anymore. The analysis of various methods in the presence of weak Minty solutions indicates that convergence can be lost if the step size is excessively 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 accomplishes this even without the added expense of a backtracking linesearch.article graphicx