Journal of Machine Learning Research 24 (2023) 1-33 Submitted 12/22; Revised 4/23; Published 4/23

Beyond the Golden Ratio for Variational Inequality Algorithms

Ahmet Alacaoglu[1] ALACAOGLU@WISC.EDU
Wisconsin Institute for Discovery
University of Wisconsin–Madison
Madison, WI, USA

Axel Böhm[1] AXEL.BOEHM@UNIVIE.AC.AT
Faculty of Mathematics
University of Vienna
Vienna, Austria

Yura Malitsky[1] YURII.MALITSKYI@UNIVIE.AC.AT
Faculty of Mathematics
University of Vienna
Vienna, Austria

Editor: Francesco Orabona

Abstract

We improve the understanding of the *golden ratio algorithm*, which solves monotone variational inequalities (VI) and convex-concave min-max problems via the distinctive feature of adapting the step sizes to the local Lipschitz constants. Adaptive step sizes not only eliminate the need to pick hyperparameters, but they also remove the necessity of global Lipschitz continuity and can increase from one iteration to the next.

We first establish the equivalence of this algorithm with popular VI methods such as reflected gradient, Popov or optimistic gradient descent-ascent (OGDA) in the unconstrained case with constant step sizes. We then move on to the constrained setting and introduce a new analysis that allows to use larger step sizes, to complete the bridge between the golden ratio algorithm and the existing algorithms in the literature. Doing so, we actually eliminate the link between the golden ratio $\frac{1+\sqrt{5}}{2}$ and the algorithm. Moreover, we improve the adaptive version of the algorithm, first by removing the maximum step size hyperparameter (an artifact from the analysis), and secondly, by adjusting it to nonmonotone problems with weak Minty solutions, with superior empirical performance.

Keywords: min-max, variational inequality, adaptive step size, nonmonotone

1. Introduction

With the increasing focus on min-max problems in applications, variational inequality (VI) algorithms have gained significant attention in machine learning. These algorithms focus on the following VI problem

$$\text{find } \mathbf{z}^* \in C \text{ such that } \langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0, \text{ for all } \mathbf{z} \in C, \tag{1}$$

1. Authors are ordered alphabetically.

for a monotone and Lipschitz operator $F\colon C \to \mathbb{R}^d$ and a convex closed set $C \subseteq \mathbb{R}^d$. The link between this template and convex-concave min-max problems such as

$$\min_{\mathbf{x}\in X}\max_{\mathbf{y}\in Y} f(\mathbf{x},\mathbf{y})$$

is well-known and follows by setting

$$\mathbf{z} = \begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}, \quad F(\mathbf{z}) = \begin{pmatrix}\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\\ -\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\end{pmatrix}, \quad C = X \times Y. \tag{2}$$

Many algorithms, dating back to 1970s, exist for solving (1), such as the extragradient method (Korpelevich, 1976), Popov's algorithm (Popov, 1980), forward-backward-forward (Tseng, 2000), reflected gradient (Malitsky, 2015; Malitsky and Tam, 2020), optimistic gradient descent-ascent (OGDA) (Daskalakis et al., 2018), dual extrapolation (Nesterov, 2007). The algorithm we focus in this paper is the golden ratio algorithm (GRAAL) due to (Malitsky, 2020), which iterates for $k \geq 0$ as

$$\begin{aligned}\bar{\mathbf{z}}^k &= \frac{\phi-1}{\phi}\mathbf{z}^k + \frac{1}{\phi}\bar{\mathbf{z}}^{k-1}\\ \mathbf{z}^{k+1} &= P_C(\bar{\mathbf{z}}^k - \alpha_k F(\mathbf{z}^k)).\end{aligned} \tag{aGRAAL}$$

with parameter $\phi > 1$ to be defined and a step size sequence α_k. The nonadaptive version of this algorithm uses $\alpha_k = \frac{\phi}{2L}$ where L is the global Lipschitz constant of F (Malitsky, 2020, Theorem 1).

However, what really separates (aGRAAL) from all the other VI algorithms listed above is the ability to provably use nonmonotone step sizes adapting to local Lipschitzness of F, with the adaptive rule

$$\alpha_k = \min\left(\gamma\alpha_{k-1}, \frac{\phi^2}{4\alpha_{k-2}}\frac{\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2}{\|F(\mathbf{z}^k) - F(\mathbf{z}^{k-1})\|^2}, \bar{\alpha}\right), \tag{3}$$

with $\gamma \leq \frac{1}{\phi} + \frac{1}{\phi^2} \in [1,2)$ and $\phi \in (1, \frac{1+\sqrt{5}}{2}]$. Higher ϕ not only gives a larger maximum step size as per (3), but also makes $\bar{\mathbf{z}}^k$ closer to the most recent iterate $\mathbf{z}^k$ instead of $\bar{\mathbf{z}}^{k-1}$. The first argument in (3) allows increasing step sizes between iterations and the second estimates a local Lipschitz constant of F. The last argument $\bar{\alpha}$ is required for theoretical reasons in (Malitsky, 2020) and generally picked as a large value in practice so that it will not be effective (see also Section 4).

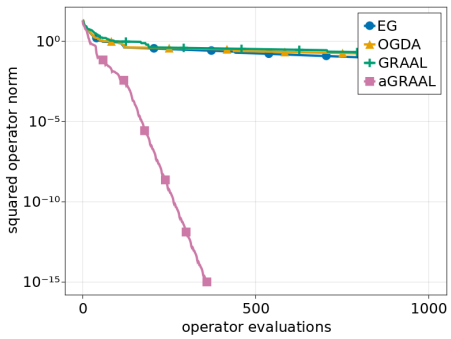

Figure 1: Policeman&Burglar matrix game example from (Nemirovski, 2013).

Compared to the constant step sizes such as $\frac{1}{L}$, the adaptive step sizes make GRAAL highly competitive in practice, in addition to relaxing the assumption of global Lipschitzness of F. Along with its unique empirical performance and generality, GRAAL keeps the good theoretical properties of nonadaptive algorithms: convergence of the sequence to a solution and $O(1/k)$ rate with monotonicity. A representative plot for the benefit of using step sizes as (3) is in Fig. 1.

1.1 Motivation and contributions.

This paper starts the formal study on understanding the peculiarity of GRAAL among the large sea of VI algorithms. Towards this goal, we contribute the following results for situating this algorithm in the literature and enhancing its theoretical understanding and practical merit:

1. In the unconstrained setting with a constant step size, GRAAL is equivalent to Popov's algorithm, OGDA and reflected gradient, when $\phi = 2$. This shows the convergence of GRAAL with $\phi = 2$ for free.

2. With constraints and a constant step size, the established connection is not sufficient and we introduce a novel analysis to prove the convergence of GRAAL with $\phi = 2$.

3. We improve the complexity of adaptive GRAAL from a quadratic dependence on the Lipschitz constant to a linear one, by eliminating the hyperparameter $\bar{\alpha}$.

4. We show that adaptive GRAAL has convergence guarantees for nonmonotone problems with weak Minty solutions and show that it reliably converges for some hard instances, even when no other method does so.

1.2 Notation and preliminaries.

We say that an operator F is monotone if $\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq 0$ for any $\mathbf{x}, \mathbf{y}$ and Lipschitz if $\|F(\mathbf{x}) - F(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$. We denote the projection as $P_C(\mathbf{z}) = \arg\min_{\mathbf{u}\in C} \|\mathbf{u} - \mathbf{z}\|^2$. We denote the golden ratio as $\varphi = \frac{1+\sqrt{5}}{2}$. The (restricted) *dual gap* function (see (Facchinei and Pang, 2003; Nesterov, 2007)) is a standard notion of suboptimality for VIs and is defined as

$$\mathrm{Gap}_S(\bar{\mathbf{z}}) = \max_{\mathbf{z}\in S} \langle F(\mathbf{z}), \bar{\mathbf{z}} - \mathbf{z}\rangle, \tag{4}$$

where S is a compact set. It is shown in (Nesterov, 2007, Lemma 1) that this is a valid suboptimality measure, since we will prove that the iterates of the algorithm remain in a compact set.

Assumption 1 *(i) The operator $F\colon C \to \mathbb{R}^d$ is monotone, (ii) F is L-Lipschitz, (iii) the set C is convex and closed, (iv) the set of solutions to* (1) *is nonempty.*

2. Connection of GRAAL and other VI algorithms

In this section, we assume that the operator F is monotone and L-Lipschitz. It is well-known that for such operators, resulting for example from min-max problems via (2), a naive forward evaluation

$$\mathbf{z}^{k+1} = P_C(\mathbf{z}^k - \alpha F(\mathbf{z}^k)) \tag{FB}$$

will not converge even for simple bilinear problems with any fixed step size — a property which makes even monotone VIs arguably more difficult to solve than the computation of stationary points in nonconvex minimization.

Having a nonconvergent scheme (FB), it is tempting to find a simple modification that will ensure convergence. There are two principal approaches to do so: one can try to change

either the argument $\mathbf{z}^k$ or the forward evaluation $F(\mathbf{z}^k)$ in (FB). Most algorithms opt for the latter option, whereas GRAAL opts for the former as we will see soon.

A strikingly simple, yet extremely powerful change is to replace $F(\mathbf{z}^k)$ by $F(\mathbf{z}^{k+1})$, which leads to

$$\mathbf{z}^{k+1} = P_C(\mathbf{z}^k - \alpha F(\mathbf{z}^{k+1})), \tag{PP}$$

which is known as the proximal point algorithm (Rockafellar, 1970). While this algorithm has great theoretical properties — for instance, it converges for any $\alpha > 0$ — it is an implicit method. Computing z^{k+1} is not a simple matter and usually each iteration of (PP) requires a call to another subsolver.

One of the earliest, and arguably most popular schemes, is the extragradient method (EG) which dates back to (Korpelevich, 1976). It relies on the change of the forward evaluation from $F(\mathbf{z}^k)$ to $F(P_C(\mathbf{z}^k - \alpha F(\mathbf{z}^k)))$. Alternatively we can write the method as

$$\bar{\mathbf{z}}^k = P_C(\mathbf{z}^k - \alpha F(\mathbf{z}^k)); \qquad \mathbf{z}^{k+1} = P_C(\mathbf{z}^k - \alpha F(\bar{\mathbf{z}}^k)). \tag{EG}$$

Interestingly, (Mokhtari et al., 2020; Nemirovski, 2004) showed that EG can be viewed as an approximation of (PP). Naturally, however, due to the explicit structure, the method requires an upper bound on the (constant) step size: $\alpha < \frac{1}{L}$. The point $\bar{\mathbf{z}}^k$ is commonly referred to as the *extrapolated point*.

By using an old evaluation of the operator to compute the extrapolated point instead of a new one, we could expect the quality of the iterates not to suffer much, leading to

$$\bar{\mathbf{z}}^k = P_C(\mathbf{z}^k - \alpha F(\bar{\mathbf{z}}^{k-1})); \qquad \mathbf{z}^{k+1} = P_C(\mathbf{z}^k - \alpha F(\bar{\mathbf{z}}^k)). \tag{Popov}$$

This strategy, proposed by (Popov, 1980), has similar guarantees as (EG) by requiring a single evaluation of the operator F. This comes at the cost of the need to reduce the step size, leading to the requirement $\alpha < \frac{1}{2L}$ (Hsieh et al., 2019).

About two decades after Popov (1980), Tseng (2000) proposed to modify EG to only require a single projection:

$$\bar{\mathbf{z}}^k = P_C(\mathbf{z}^k - \alpha F(\mathbf{z}^k)); \qquad \mathbf{z}^{k+1} = \bar{\mathbf{z}}^k - \alpha(F(\bar{\mathbf{z}}^k) - F(\mathbf{z}^k)), \tag{FBF}$$

leading to the forward-backward-forward method, with the same requirement for α as EG: $\alpha < \frac{1}{L}$.

As observed in (Böhm et al., 2022), applying Popov's idea to (FBF) and replacing $F(\mathbf{z}^k)$ by $F(\bar{\mathbf{z}}^{k-1})$, the obtained method can be conveniently written in one line

$$\bar{\mathbf{z}}^{k+1} = P_C(\bar{\mathbf{z}}^k - \alpha(2F(\bar{\mathbf{z}}^k) - F(\bar{\mathbf{z}}^{k-1}))), \tag{FoRB}$$

where the bound $\alpha < \frac{1}{2L}$ is imposed. This method was in greater generality (nonconstant step sizes) studied in (Malitsky and Tam, 2020) under the name forward-reflected-backward (FoRB), but is more widely known in the unconstrained setting as the optimistic-gradient-descent-ascent (OGDA) (Daskalakis et al., 2018). Again, (FoRB) can be seen as (FB) where $F(\mathbf{z}^k)$ is changed to $2F(\mathbf{z}^k) - F(\mathbf{z}^{k-1})$.

Another method proposed in (Malitsky, 2015) is projected reflected gradient method (PRG), which solves VI by requiring only one evaluation of F and iterates as

$$\bar{\mathbf{z}}^{k+1} = P_C(\bar{\mathbf{z}}^k - \alpha F(2\bar{\mathbf{z}}^k - \bar{\mathbf{z}}^{k-1})). \tag{PRG}$$

It is easy to see that the (PRG) method is equivalent to (FoRB) if the operator F is linear. The analysis of (PRG) in (Malitsky, 2015) requires $\alpha < \frac{\sqrt{2}-1}{L}$, which according to the above consideration is not tight when F is linear. Regarding the (FB) interpretation, we only need to change $F(\mathbf{z}^k)$ to $F(2\mathbf{z}^k - \mathbf{z}^{k-1})$.

A related method, relying on a very similar correction step as the one used in (FoRB), but applied after the projection, was proposed in (Csetnek et al., 2019) and is given by

$$\bar{\mathbf{z}}^{k+1} = P_C(\bar{\mathbf{z}}^k - \alpha F(\bar{\mathbf{z}}^k)) - \alpha(F(\bar{\mathbf{z}}^k) - F(\bar{\mathbf{z}}^{k-1})). \tag{shadow-DR}$$

In the unconstrained case this method is equivalent to (FoRB). The analysis in (Csetnek et al., 2019), however, requires the more restrictive $\alpha < \frac{1}{3L}$ and even provides a counterexample to show that this dependence is tight. It is not clear where this more conservative dependence on the Lipschitz constant comes from.

Last but not least, our main method of interest is the golden ratio algorithm (GRAAL), introduced in (Malitsky, 2020), given by

$$\bar{\mathbf{z}}^k = \frac{\phi - 1}{\phi}\mathbf{z}^k + \frac{1}{\phi}\bar{\mathbf{z}}^{k-1}; \qquad \mathbf{z}^{k+1} = P_C(\bar{\mathbf{z}}^k - \alpha F(\mathbf{z}^k)), \tag{GRAAL}$$

with the initialization $\bar{\mathbf{z}}^{-1} = \mathbf{z}^0$. Initially, the (nonadaptive) analysis in (Malitsky, 2020) requires choosing $\phi \le \varphi = \frac{1+\sqrt{5}}{2}$ (hence the name with golden ratio), together with a bound on the step size $\alpha \le \frac{\phi}{2L}$. In the following we will show that this bound can be relaxed, in which case we can recover some of the other methods mentioned in this section. Evidently, (GRAAL) can be seen as a modification of (FB) with the change $\mathbf{z}^k$ to $\bar{\mathbf{z}}^k$.

2.1 GRAAL is averaged PRG

From the first line of (GRAAL) we can rewrite

$$\mathbf{z}^k = \frac{\phi}{\phi - 1}\bar{\mathbf{z}}^k - \frac{1}{\phi - 1}\bar{\mathbf{z}}^{k-1} = \bar{\mathbf{z}}^k + \frac{1}{\phi - 1}\left(\bar{\mathbf{z}}^k - \bar{\mathbf{z}}^{k-1}\right). \tag{5}$$

Hence instead of viewing $\bar{\mathbf{z}}^k$ as a sequence of averaged iterates we can interpret $\mathbf{z}^k$ as a sequence of extrapolated iterates. Plugging (5) into the second line of (GRAAL) yields the identity claimed in the title, which gives for $\phi = 2$ a $\frac{1}{2}$-averaged version of (PRG):

$$\bar{\mathbf{z}}^{k+1} = \frac{1}{2}\bar{\mathbf{z}}^k + \frac{1}{2}P_C\left(\bar{\mathbf{z}}^k - \alpha F\big(2\bar{\mathbf{z}}^k - \bar{\mathbf{z}}^{k-1}\big)\right), \tag{6}$$

If the problem is unconstrained, GRAAL's $\bar{\mathbf{z}}^k$ sequence corresponds to the one generated by (PRG), but with scaled step size.

2.2 GRAAL is FoRB/OGDA/Popov in the unconstrained case

While (GRAAL) seems fundamentally different from methods such as (FoRB)/OGDA/(Popov) or (shadow-DR), which rely on previous evaluations of the operator F, they do turn out to be equivalent in the unconstrained setting, where $\bar{\mathbf{z}}^{k-1} = \mathbf{z}^k + \alpha F(\mathbf{z}^{k-1})$, via the simple identity:

$$\mathbf{z}^{k+1} = \bar{\mathbf{z}}^k - \alpha F(\mathbf{z}^k) = \frac{\phi - 1}{\phi}\mathbf{z}^k + \frac{1}{\phi}\bar{\mathbf{z}}^{k-1} - \alpha F(\mathbf{z}^k) = \mathbf{z}^k - \frac{\alpha}{\phi}\Big(\phi F(\mathbf{z}^k) - F(\mathbf{z}^{k-1})\Big).$$

To be precise, the equivalence to OGDA and others holds with $\phi = 2$ and for general ϕ, (GRAAL) is equivalent to the generalized version of OGDA, see (Ryu et al., 2019; Böhm, 2023), where 2 is replaced by ϕ and an appropriate scaling of the step size.

2.3 Summary

In the unconstrained setting, all the methods above collapse to two classes. On one hand, there is (EG)/(FBF), relying on two gradient evaluations per iteration and a step size constrained by $\frac{1}{L}$. On the other, there is the zoo of Popovesque methods: (GRAAL), (shadow-DR), (PRG), (FoRB), (Popov), and OGDA requiring only one call to F but paying the price of a smaller step size.

The established equivalences give us a proof of convergence for GRAAL with $\phi = 2$ in the *unconstrained* case since it reduces to known methods. However, these relationships are not useful to show the convergence of GRAAL with $\phi = 2$ in the *constrained* case which is much more common with min-max problems. It turns out that standard techniques are not sufficient for such a result, which motivates the next section, dedicated to proving this conclusion.

3. GRAAL with $\phi = 2$ for constrained problems

3.1 Dissection of GRAAL's analysis

We sketch the existing analysis of GRAAL in (Malitsky, 2020) and point out to the reason for the restrictive upper bound on ϕ. Then we see high level ideas on how to tighten this analysis en route to $\phi = 2$. For convenience, let us define

$$G(\mathbf{z}^k, \mathbf{z}) = 2\alpha\langle F(\mathbf{z}), \mathbf{z}^k - \mathbf{z}\rangle \quad \text{and} \quad \mathcal{E}_k(\mathbf{z}) := \frac{\phi}{\phi - 1}\|\bar{\mathbf{z}}^k - \mathbf{z}\|^2 + \frac{\phi}{2}\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2. \tag{7}$$

The former is important for showing the rate on the gap function since taking the maximum gives (4) for which we will show the $O(1/k)$ rate. The latter will serve as a Lyapunov (or energy) function.

The analysis of (Malitsky, 2020), given in Lemma 13 in the Appendix for convenience, results in the following key inequality for $k \geq 1$:

$$G(\mathbf{z}^k, \mathbf{z}) + \mathcal{E}_{k+1}(\mathbf{z}) \leq \mathcal{E}_k(\mathbf{z}) + \left(\phi - 1 - \frac{1}{\phi}\right)\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 - \frac{1}{\phi}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2. \tag{8}$$

Golden ratio appears to make $\phi - 1 - \frac{1}{\phi} = 0$. The analysis in (Malitsky, 2020) discards the *good term* $-\frac{1}{\phi}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2$ in the right-hand side to use a telescoping argument, common to methods described in Section 2.

We see that this good term is one index away from the main error term $\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2$. However, it is not one index forward, but instead backward, hence it is not immediate how to use it to relax the requirement of ϕ. This intuition can be formalized by summing the inequality and using the definition of $\mathcal{E}_{k+1}(\mathbf{z})$ which results in

$$\sum_{i=1}^{k} G(\mathbf{z}^i, \mathbf{z}) \leq \mathcal{E}_1(\mathbf{z}) - \frac{\phi}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \sum_{i=1}^{k}\left[\left(\phi - 1 - \frac{1}{\phi}\right)\|\mathbf{z}^{i+1} - \bar{\mathbf{z}}^i\|^2 - \frac{1}{\phi}\|\mathbf{z}^i - \bar{\mathbf{z}}^{i-1}\|^2\right].$$

We focus here on the main error term

$$-\frac{\phi}{2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2+\sum_{i=1}^{k}\left[\left(\phi-1-\frac{1}{\phi}\right)\|\mathbf{z}^{i+1}-\bar{\mathbf{z}}^i\|^2-\frac{1}{\phi}\|\mathbf{z}^i-\bar{\mathbf{z}}^{i-1}\|^2\right]. \tag{9}$$

If this unwieldy expression is bounded, the convergence rate of the gap follows immediately. It is easy to see that if $\phi = 2$, then (9) reduces to

$$-\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2+\frac{1}{2}\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^k\|^2-\frac{1}{2}\|\mathbf{z}^1-\bar{\mathbf{z}}^0\|^2, \tag{10}$$

which is not necessarily bounded due to the second term. In fact, a naive approach can be used for values slightly larger than the golden ratio. We know by Young's inequality that for any $\tau > 0$

$$-\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^k\|^2 \geq -(1+\tau)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2-(1+1/\tau)\|\bar{\mathbf{z}}^k-\mathbf{z}^k\|^2.$$

It is tedious but straightforward to properly adjust ϕ and τ so that the error term involving $\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^k\|^2$ is cancelled by using the negative terms in (10). However, this approach is definitely not tight due to spurious use of Young's inequality and only gives ϕ values up to 1.77, whereas we expect $\phi = 2$ to be tight, due to the connection with existing methods such as OGDA.

Another obvious remedy would be to simply *assume* that the term $\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^k\|^2$ is bounded, for example, by assuming the sequence $(\mathbf{z}^k)$ is bounded. However, this is not realistic since boundedness of C is a strong assumption, not holding for many min-max problems in practice. A prevalent example is constrained optimization where neither the primal nor the dual domain is bounded.

In the next section, we *prove* that the sequence $(\mathbf{z}^k)$ is bounded with $\phi = 2$, which will help us get convergence and rate results. Instead of the naive approach described above which tries to cancel the error in (10) by other terms and *loose* inequalities such as Young's, we will analyze the boundedness of the sequence by a novel induction argument on the *tight* inequality in (9).

3.2 GRAAL beyond the golden ratio

A standard way to prove boundedness of iterates is to identify a Lyapunov function (nonnegative and nonincreasing) including terms such as $\|\mathbf{z}^k-\mathbf{z}\|^2$. An example is $\mathcal{E}$ in (7). While this can be done with $\phi \leq \varphi$, it is unclear if it is possible with $\phi = 2$, due to the issue described in Section 3.1. To go around this difficulty, we have to use nonstandard tools and forgo the Lyapunov function argument and analyze boundedness of $(\mathbf{z}^k)$ directly via induction on the key inequality (8).

Theorem 1 *Let Assumption 1 hold and let $\phi = 2$, $\alpha \leq \frac{1}{L}$ in* (GRAAL)*. Then we have that $(\mathbf{z}^k)$ and $(\bar{\mathbf{z}}^k)$ are bounded sequences. In particular, we have*

$$\|\bar{\mathbf{z}}^k-\mathbf{z}^*\|^2 \leq 4\left(\|\mathbf{z}^1-\mathbf{z}^*\|^2+\|\mathbf{z}^0-\mathbf{z}^*\|^2\right) \leq 12\|\mathbf{z}^0-\mathbf{z}^*\|^2.$$

While we provide a formal proof for the case $\phi = 2$ here for simplicity, we also supply a computer aided proof via semi-definite programming, see Appendix A, which in addition covers the case for $\phi < 2$ and provides a better constant for $\phi = 2$.

Proof In (8), we set $\mathbf{z} = \mathbf{z}^*$ and use $G(\mathbf{z}^k, \mathbf{z}^*) \geq 0$ by (1). Next, we sum (8) with $\phi = 2$ which gives us

$$\begin{aligned} 2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 + \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 &\leq 2\|\bar{\mathbf{z}}^1 - \mathbf{z}^*\|^2 + \frac{1}{2}\|\mathbf{z}^1 - \mathbf{z}^0\|^2 + \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \\ &= \|\mathbf{z}^1 - \mathbf{z}^*\|^2 + \|\mathbf{z}^0 - \mathbf{z}^*\|^2 + \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2, \end{aligned} \tag{11}$$

where we used $\mathbf{z}^0 = \bar{\mathbf{z}}^0$. Define $R^2 = \|\mathbf{z}^1 - \mathbf{z}^*\|^2 + \|\mathbf{z}^0 - \mathbf{z}^*\|^2$. For the inductive step, assume that $\|\bar{\mathbf{z}}^i - \mathbf{z}^*\| \leq 2R$ for all $i \leq k$.

We will transform the error term $\frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2$ in (11) so that the other terms in the left-hand side of (11) can be used for (partial) cancellation. By the paralellogram law and definition $\bar{\mathbf{z}}^k = \frac{1}{2}\mathbf{z}^k + \frac{1}{2}\bar{\mathbf{z}}^{k-1}$, we have

$$\begin{aligned} \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 &= \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \|\mathbf{z}^k - \bar{\mathbf{z}}^k\|^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - 2\mathbf{z}^k + \bar{\mathbf{z}}^k\|^2 \\ &= \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \|\mathbf{z}^k - \bar{\mathbf{z}}^k\|^2 - 2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^k\|^2. \end{aligned}$$

Plugging in this equality to (11) and using the definition of $\bar{\mathbf{z}}^k$ gives us

$$2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 + 2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^k\|^2 \leq R^2 + \|\mathbf{z}^k - \bar{\mathbf{z}}^k\|^2 = R^2 + \frac{1}{4}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2. \tag{12}$$

The left-hand side is in a suitable form to apply another paralellogram law and obtain

$$2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 + 2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^k\|^2 = \|\mathbf{z}^k - \mathbf{z}^*\|^2 + 4\left\|\bar{\mathbf{z}}^{k+1} - \frac{\mathbf{z}^* + \mathbf{z}^k}{2}\right\|^2.$$

After combining this equality with (12), it follows that

$$\|\mathbf{z}^k - \mathbf{z}^*\|^2 + 4\left\|\bar{\mathbf{z}}^{k+1} - \frac{\mathbf{z}^* + \mathbf{z}^k}{2}\right\|^2 \leq R^2 + \frac{1}{4}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2. \tag{13}$$

We are now at the most critical point of the proof. For induction, we will combine the second term in the right-hand side of (13) with the terms in the left-hand side to obtain $\|\bar{\mathbf{z}}^{k-1} - \mathbf{z}^*\|^2$. Now continue with the following identity which can be verified by simple expansion of quadratics

$$\frac{1}{4}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2 - \|\mathbf{z}^k - \mathbf{z}^*\|^2 = -\frac{3}{4}\left\|\mathbf{z}^k - \frac{4}{3}\mathbf{z}^* + \frac{1}{3}\bar{\mathbf{z}}^{k-1}\right\|^2 + \frac{1}{3}\|\bar{\mathbf{z}}^{k-1} - \mathbf{z}^*\|^2.$$

Even though it is difficult to see the significance of this identity, a high level intuition is noticing that $\mathbf{z} \mapsto \frac{1}{4}\|\mathbf{z} - \bar{\mathbf{z}}^{k-1}\|^2 - \|\mathbf{z} - \mathbf{z}^*\|^2$ is maximized at $\mathbf{z} = \frac{4}{3}\mathbf{z}^* - \frac{1}{3}\bar{\mathbf{z}}^{k-1}$. We can then rewrite (13) as

$$3\left\|\frac{\mathbf{z}^k - \mathbf{z}^*}{2} + \frac{\bar{\mathbf{z}}^{k-1} - \mathbf{z}^*}{6}\right\|^2 + 4\left\|\bar{\mathbf{z}}^{k+1} - \frac{\mathbf{z}^* + \mathbf{z}^k}{2}\right\|^2 \leq R^2 + \frac{1}{3}\|\bar{\mathbf{z}}^{k-1} - \mathbf{z}^*\|^2. \tag{14}$$

Let $a = \mathbf{z}^k - \mathbf{z}^*$, $b = \bar{\mathbf{z}}^{k-1} - \mathbf{z}^*$. Then using this notation in (14), taking square root of both sides, using triangle inequality and the inductive assumption ($\|b\| \leq 2R$) implies that

$$\left\|\frac{1}{2}a + \frac{1}{6}b\right\|^2 \leq \frac{R^2}{3} + \frac{1}{9}\|b\|^2 \quad \Longrightarrow \quad \frac{1}{2}\|a\| \leq \sqrt{\frac{R^2}{3} + \frac{1}{9}\|b\|^2} + \frac{1}{6}\|b\| \leq \frac{\sqrt{7}+1}{3}R.$$

On the other hand, (14) along with inductive assumption gives us

$$\left\|\bar{\mathbf{z}}^{k+1} - \frac{\mathbf{z}^* + \mathbf{z}^k}{2}\right\|^2 \leq \frac{R^2}{4} + \frac{1}{12}\|\bar{\mathbf{z}}^{k-1} - \mathbf{z}^*\|^2 \quad \Longrightarrow \quad \left\|\bar{\mathbf{z}}^{k+1} - \frac{\mathbf{z}^* + \mathbf{z}^k}{2}\right\| \leq \sqrt{\frac{7R^2}{12}}.$$

We can now combine the last two estimations with triangle inequality to complete the induction

$$\begin{aligned}\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\| &\leq \left\|\bar{\mathbf{z}}^{k+1} - \frac{\mathbf{z}^* + \mathbf{z}^k}{2}\right\| + \left\|\frac{\mathbf{z}^* + \mathbf{z}^k}{2} - \mathbf{z}^*\right\| = \left\|\bar{\mathbf{z}}^{k+1} - \frac{\mathbf{z}^* + \mathbf{z}^k}{2}\right\| + \frac{1}{2}\|\mathbf{z}^k - \mathbf{z}^*\| \\ &\leq \sqrt{\frac{7}{12}} \cdot R + \frac{\sqrt{7}+1}{3} \cdot R < 2R. \end{aligned} \tag{15}$$

Hence, by induction we proved that $(\bar{\mathbf{z}}^k)$ is bounded. The definition $\mathbf{z}^k = 2\bar{\mathbf{z}}^k - \bar{\mathbf{z}}^{k-1}$ shows that $(\mathbf{z}^k)$ is also bounded. The final inequality uses $R^2 \leq 3\|\mathbf{z}^0 - \mathbf{z}^*\|^2$, which is shown in Lemma 14 in Appendix A. ■

Corollary 2 *Let Assumption 1 hold, $\phi = 2$, $\alpha = \frac{1-\varepsilon}{L}$ for any $\varepsilon \in (0,1)$ in* (GRAAL) *and $\mathbf{Z}^k = \frac{1}{k}\sum_{i=1}^k \mathbf{z}^i$. Then $(\mathbf{z}^k)$ converges to a solution of* (1) *and*

$$\mathrm{Gap}_S\left(\mathbf{Z}^k\right) \leq \frac{32L}{(1-\varepsilon)k}\|\mathbf{z}^0 - \mathbf{z}^*\|^2,$$

where $S = \{\mathbf{z} \in C\colon \|\mathbf{z} - \mathbf{z}^*\|^2 \leq 18\|\mathbf{z}^0 - \mathbf{z}^*\|^2\}$.

Proof We begin by proving the first part of the corollary. By the boundedness of the iterates, proved in Theorem 1, we have that $(\mathbf{z}^k)$ has a convergent subsequence, say $(\mathbf{z}^{k_i})$ and for some $\tilde{\mathbf{z}} \in C$, we have that $\mathbf{z}^{k_i} \to \tilde{\mathbf{z}}$. We next show that $\tilde{\mathbf{z}}$ is a solution of (1). For this, first notice that by summing the result of Lemma 13 with $\mathbf{z} = \mathbf{z}^*$, $\phi = 2$ and using the boundedness of $\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^k\|^2$ derived in Theorem 1, we have that

$$\sum_{k=0}^{\infty} \|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2 < +\infty.$$

Young's inequality gives

$$\begin{aligned}\|\mathbf{z}^k - \bar{\mathbf{z}}^k\|^2 &\leq 3\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2 + \frac{3}{2}\|\bar{\mathbf{z}}^k - \mathbf{z}^{k-1}\|^2 \\ &\leq \frac{15}{4}\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2 + \frac{3}{4}\|\bar{\mathbf{z}}^{k-1} - \mathbf{z}^{k-1}\|^2,\end{aligned}$$

where the second inequality is by the definition $\bar{\mathbf{z}}^k = \frac{1}{2}\mathbf{z}^k + \frac{1}{2}\bar{\mathbf{z}}^{k-1}$. Rearranging and summing this inequality gives

$$\sum_{k=0}^{\infty} \|\mathbf{z}^k - \bar{\mathbf{z}}^k\|^2 \leq \sum_{k=0}^{\infty} 15\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2 < +\infty.$$

where we used $\mathbf{z}^{-1} = \bar{\mathbf{z}}^{-1}$ for getting the first inequality.

As a result, we have that $\mathbf{z}^{k+1} - \mathbf{z}^k \to 0$ and $\bar{\mathbf{z}}^k - \mathbf{z}^k \to 0$ and hence $\bar{\mathbf{z}}^{k_i} \to \tilde{\mathbf{z}}$ and $\mathbf{z}^{k_i+1} \to \tilde{\mathbf{z}}$. We can take the limit in the prox-inequality

$$\langle \mathbf{z}^{k+1} - \bar{\mathbf{z}}^k + \alpha F(\mathbf{z}^k), \mathbf{z} - \mathbf{z}^{k+1} \rangle \geq 0$$

and use continuity of F to get that $\tilde{\mathbf{z}}$ is a solution. For simplicity, set $\tilde{\mathbf{z}} = \mathbf{z}^*$, where $\mathbf{z}^*$ is an arbitrary solution.

The result of Lemma 13 also gives the following inequality after using $\phi = 2$ and $\mathbf{z} = \mathbf{z}^*$:

$$2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 + \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \leq 2\|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 + \|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2 - \frac{1}{2}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2.$$

Since $\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2$ is bounded, the sequence on the left-hand side of the inequality is lower bounded and nonincreasing, hence convergent. In summary, we have that

$$\lim_{k\to\infty} 2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 + \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \text{ exists.}$$

As $\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \to 0$ and $\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 = 4\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^{k+1}\|^2 \to 0$, we further deduce that

$$\lim_{k\to\infty} \|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 \text{ exists.}$$

Since we previously showed that $\bar{\mathbf{z}}^{k_i} - \mathbf{z}^* \to 0$ for an arbitrary $\mathbf{z}^*$, we conclude that $\bar{\mathbf{z}}^k - \mathbf{z}^* \to 0$. It is easy to see that $\mathbf{z}^k - \mathbf{z}^* \to 0$ as well because $\bar{\mathbf{z}}^k - \mathbf{z}^k \to 0$.

Now we turn our attention to the convergence rate. We start by collecting some estimations from Theorem 1. In particular, the main conclusion of the theorem was that

$$\|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 \leq 4R^2 = 12\|\mathbf{z}^0 - \mathbf{z}^*\|^2. \tag{16}$$

Moreover, from (13), we have that

$$\begin{aligned}\|\mathbf{z}^k - \mathbf{z}^*\|^2 &\leq R^2 + \frac{1}{4}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2 \\ &\leq R^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{z}^*\|^2 + \frac{1}{2}\|\bar{\mathbf{z}}^{k-1} - \mathbf{z}^*\|^2.\end{aligned}$$

which together with (16) implies

$$\|\mathbf{z}^k - \mathbf{z}^*\|^2 \leq 6R^2 = 18\|\mathbf{z}^0 - \mathbf{z}^*\|^2. \tag{17}$$

This implies that the set S contains all iterates, which enables us to apply (Nesterov, 2007, Lemma 1) for the gap function.

We set $\phi = 2$ in Lemma 13, sum this inequality, divide by k, and take maximum over S to obtain

$$\mathrm{Gap}_S(\mathbf{Z}^k) \leq \frac{1}{2\alpha}\left(\max_{\mathbf{z}\in S} 2\|\bar{\mathbf{z}}^1 - \mathbf{z}\|^2 + \frac{1}{2}\|\mathbf{z}^1 - \mathbf{z}^0\|^2 + \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2\right),$$

where the $\frac{1}{2\alpha}$ factor on the right-hand side is because of $G(\mathbf{z}^k, \mathbf{z}) = 2\alpha\langle F(\mathbf{z}), \mathbf{z}^k - \mathbf{z}\rangle$ (see (7)).

By using $\bar{\mathbf{z}}^1 = \frac{1}{2}\mathbf{z}^1 + \frac{1}{2}\mathbf{z}^0$ due to $\bar{\mathbf{z}}^0 = \mathbf{z}^0$, we get

$$\mathrm{Gap}_S(\mathbf{Z}^k) \leq \frac{1}{2\alpha}\left(\max_{\mathbf{z}\in S}(\|\mathbf{z}^1 - \mathbf{z}\|^2 + \|\mathbf{z}^0 - \mathbf{z}\|^2) + \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2\right). \tag{18}$$

We will now use (16) and (17) to get

$$\frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \leq \|\mathbf{z}^{k+1} - \mathbf{z}^*\|^2 + \|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 \leq 30\|\mathbf{z}^0 - \mathbf{z}^*\|^2. \tag{19}$$

Moreover, we have for any $c > 0$ that

$$\begin{aligned}\max_{\mathbf{z}\in S}(\|\mathbf{z}^1 - \mathbf{z}\|^2 + \|\mathbf{z}^0 - \mathbf{z}\|^2) &\leq \max_{\mathbf{z}\in S} 2(1+c)\|\mathbf{z} - \mathbf{z}^*\|^2 + (1+1/c)(\|\mathbf{z}^1 - \mathbf{z}^*\|^2 + \|\mathbf{z}^0 - \mathbf{z}^*\|^2) \\ &\leq (36(1+c) + 3(1+1/c))\|\mathbf{z}^0 - \mathbf{z}^*\|^2, \end{aligned} \tag{20}$$

where we used Young's inequality for the first inquality and Lemma 14 in the second one. We apply (19) and (20) in (18) and pick c to minimize the corresponding term, to obtain the result. ■

Increasing ϕ from the golden ratio to 2 for the fixed step size regime not only emphasizes an interesting connection to OGDA, but also consistently improves empirical performance for monotone problems, see Fig. 2. Even though the adaptive version of GRAAL is typically superior in practice as per the experiments of Malitsky (2020), we want to point out that with constant step sizes, the need to pick $\phi \leq \varphi < 2$ caused GRAAL to perform worse than methods such as OGDA. The main merit of our result is showing that this is not the case.

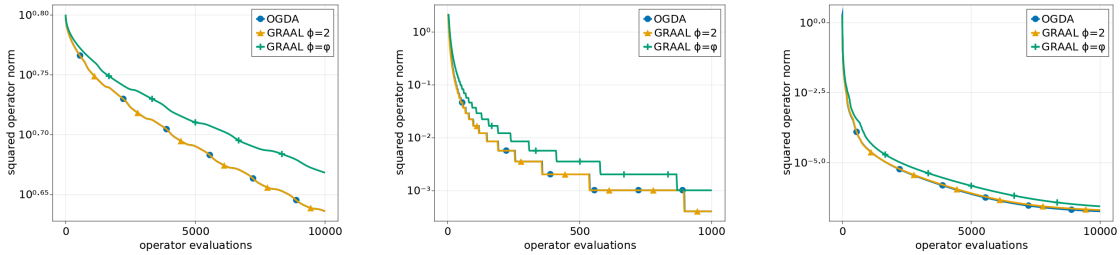

Figure 2: Left: The linearly constrained QP in (Yoon and Ryu, 2021). Middle: Test matrix given in (Nemirovski et al., 2009). Right: Randomly generated matrix game. Interestingly, even with constraints, GRAAL and OGDA perform almost identically.

4. Adaptive GRAAL: Removing hyperparameters and improving complexity

We first recall aGRAAL, proposed in (Malitsky, 2020):

$$\begin{aligned} \bar{\mathbf{z}}^k &= \frac{\phi-1}{\phi}\mathbf{z}^k + \frac{1}{\phi}\bar{\mathbf{z}}^{k-1} \\ \mathbf{z}^{k+1} &= P_C(\bar{\mathbf{z}}^k - \alpha_k F(\mathbf{z}^k)), \end{aligned} \tag{aGRAAL}$$

where α_k is picked as in (3) and $\phi \in \left(1, \frac{1+\sqrt{5}}{2}\right)$. As mentioned in (Malitsky, 2020), the hyperparameter $\bar{\alpha}$ in (3) is only for theoretical purposes and the suggested choice in practice is taking $\bar{\alpha}$ very large so that it will be ineffective in (3). However, there are two sides to this coin from a complexity point of view as we see now. Recall that (Malitsky, 2020) proved that the iterates remain in a bounded region and established the following worst case rate for the ergodic iterates

$$\mathrm{Gap}_S\left(\mathbf{Z}^k\right) \leq \frac{4L^2\bar{\alpha}}{k\phi^2} \underbrace{\max_{\mathbf{z}\in S}\left(\frac{\phi}{\phi-1}\|\mathbf{z}^1-\mathbf{z}\|^2 + \frac{\theta_0}{2}\|\mathbf{z}^0-\mathbf{z}\|^2\right)}_{=:D}, \tag{21}$$

where L denotes the (unknown) Lipschitz constant of F over this bounded region, S is any compact set containing this region, and $\theta_0 \geq 1$. On one hand, taking $\bar{\alpha}$ too large could make this rate vacuous. On the other, taking $\bar{\alpha}$ small may prevent taking large step sizes in (3). Another aspect of this bound is that the dependence on L is suboptimal, since most VI methods including *nonadaptive* GRAAL result in the rate $O(L/k)$. Obtaining linear dependence on L suggests taking $\bar{\alpha}$ as a large multiple of $\frac{1}{L}$, which not only requires the knowledge of L, but also could make the constant of the rate large as per the discussion above. This suboptimal worst-case complexity result may be seen as the cost of adaptivity. To avoid this conflict regarding the choice of $\bar{\alpha}$ in practice and the resulting complexity, we propose to remove $\bar{\alpha}$ in (3) and provide an analysis with the simpler step size rule

$$\alpha_k = \min\left(\gamma\alpha_{k-1}, \frac{\phi^2}{4\alpha_{k-2}}\frac{\|\mathbf{z}^k-\mathbf{z}^{k-1}\|^2}{\|F(\mathbf{z}^k)-F(\mathbf{z}^{k-1})\|^2}\right), \tag{22}$$

with $\gamma = \frac{1}{\phi} + \frac{1}{\phi^2}$ to complement the empirical success of the algorithm in the experiments of (Malitsky, 2020) with the choice (22). This rule not only removes the hyperparameter $\bar{\alpha}$ but as we show in the next theorem, it also results in the rate $O(L/k)$ which is only a small constant times worse than the rate of the nonadaptive method (see Remark 6 for a precise statement). As a result, this is a much smaller cost to pay for the worst-case complexity with adaptivity. With the proposed change, we obtain Alg. 1.

Algorithm 1 aGRAAL

Require: $\bar{\mathbf{z}}^0 = \mathbf{z}^0 \in C$, $\phi \in \left(1, \frac{1+\sqrt{5}}{2}\right)$, $\gamma \in \left(1, \frac{1}{\phi} + \frac{1}{\phi^2}\right]$, $\alpha_0 > 0$, $\theta_0 = \phi$.

$\mathbf{z}^1 = P_C(\mathbf{z}^0 - \alpha_0 F(\mathbf{z}^0))$ possibly by linesearch

for $k \geq 1$ **do**

$\alpha_k = \min\left(\gamma\alpha_{k-1}, \frac{\phi\theta_{k-1}}{4\alpha_{k-1}} \frac{\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2}{\|F(\mathbf{z}^k) - F(\mathbf{z}^{k-1})\|^2}\right)$

$\bar{\mathbf{z}}^k = \frac{\phi-1}{\phi}\mathbf{z}^k + \frac{1}{\phi}\bar{\mathbf{z}}^{k-1}$

$\mathbf{z}^{k+1} = P_C(\bar{\mathbf{z}}^k - \alpha_k F(\mathbf{z}^k))$

$\theta_k = \frac{\alpha_k}{\alpha_{k-1}}\phi$

We will start with the one iteration analysis from (Malitsky, 2020) which does not need any modification, so we state the result with a very brief proof to make the connection easy to follow. Note that the spurious $\bar{\alpha}$ term is not used in the analysis of (Malitsky, 2020, Theorem 2) in this result.

Lemma 3 *(Consequence of (Malitsky, 2020, Theorem 2)) Let Assumption 1 (i, iii, iv) hold and F be locally Lipschitz. Let $\phi \in \left(1, \frac{1+\sqrt{5}}{2}\right)$, $\gamma \leq \frac{1}{\phi} + \frac{1}{\phi^2}$ and $\mathbf{Z}^k = \frac{1}{\sum_{i=1}^k \alpha_i} \sum_{i=1}^k \alpha_i \mathbf{z}^i$. Then the iterates $(\mathbf{z}^k)$ of Alg. 1 are bounded and we have*

$$\mathrm{Gap}_S(\mathbf{Z}^k) \leq \frac{D}{2\sum_{i=1}^k \alpha_i} \quad \forall k \geq 1. \tag{23}$$

Proof In (Malitsky, 2020, Theorem 2) inequality (35) is only valid for $k \geq 2$. However due to our definition of $\mathbf{z}^1$ and since $\mathbf{z}^0 = \bar{\mathbf{z}}^0$, it already holds for $k \geq 1$. Then we can unroll the recursion another step until iteration $k = 1$ instead of $k = 2$. ■

4.1 Lower bounding the sum of step sizes

Looking at the bound in (23), we notice that in order to derive a complexity result, we require a lower bound for the sum of the step sizes. In the original proof in (Malitsky, 2020) such a bound is ensured by enforcing each α_i to be lower bounded by using $\bar{\alpha}$ to derive a $O(1/k)$ rate.

For aGRAAL the consecutive step sizes also depend on the initial α_0. Since the method is entirely adaptive and we do not assume any a priori knowledge about F, we want to make sure that α_0 is not too small and not too large. To this end, for the initialization, we recommend to use the linesearch procedure described below.

For brevity, we will write $L_k = \frac{\|F(\mathbf{z}^k) - F(\mathbf{z}^{k-1})\|}{\|\mathbf{z}^k - \mathbf{z}^{k-1}\|}$. Also, let L be the Lipschitz constant of F over the set $\overline{\mathrm{conv}}\{\mathbf{z}^0, \mathbf{z}^1, \dots\}$, which is bounded due to Lemma 3. It trivially holds that $L \geq L_k$ for all k. Without loss of generality, we assume that $L \geq 1$.

Let us choose α_0 via linesearch as follows: set α_0 to the largest number in $\{\gamma^{-i}\colon i = 0, 1, \dots\}$ (note here the use of γ given in Alg. 1) such that

$$\begin{aligned} &\mathbf{z}^1 = P_C(\mathbf{z}^0 - \alpha_0 F(\mathbf{z}^0)) \\ &\alpha_0 \|F(\mathbf{z}^1) - F(\mathbf{z}^0)\| \leq \frac{\phi}{2}\|\mathbf{z}^1 - \mathbf{z}^0\|. \end{aligned} \tag{24}$$

The second equation gives the upper bound for α_0: $\alpha_0 \leq \frac{\phi}{2L_1}$. Note that the linesearch always terminates because F is locally Lipschitz, and we can immediately obtain the following statements, which we will use later in Lemma 4. We have either $\alpha_0 = \gamma^{-0} = 1 > \frac{\phi}{2L} \geq \frac{\phi}{2\gamma L}$ or that $\gamma\alpha_0 = \gamma \cdot \gamma^{-i}$ violates the inequality above, that is

$$\gamma\alpha_0 \|F(\mathbf{z}^1) - F(\mathbf{z}^0)\| > \frac{\phi}{2}\|\mathbf{z}^1 - \mathbf{z}^0\|,$$

which also implies that $\alpha_0 \geq \frac{\phi}{2\gamma L_1} \geq \frac{\phi}{2\gamma L}$. Moreover, from the algorithm's update for α_1 we have either $\alpha_1 = \gamma\alpha_0 \geq \frac{\phi}{2L}$ or

$$\alpha_1 = \frac{\phi\theta_0}{4\alpha_0 L_1^2} = \frac{\phi^2}{4\alpha_0 L_1^2},$$

which implies that $\alpha_1 + \alpha_0 \geq 2\sqrt{\alpha_1\alpha_0} \geq \frac{\phi}{L_1}$. Using the upper bound for α_0, we deduce that $\alpha_1 \geq \frac{\phi}{2L_1} \geq \alpha_0$. To conclude, the suggested choice of α_0 in (24) ensures that

$$\alpha_1 \geq \frac{\phi}{2L} \quad \text{and} \quad \alpha_1 \geq \alpha_0.$$

The following lemma is our main technical contribution in Section 4, which will lead to Theorem 5 when combined with the previous lemma.

Lemma 4 *Let α_0 satisfy (24) and $c = \min_{m\geq 2} \frac{\phi}{m}\sqrt{\frac{\gamma^{m-1}-1}{\gamma-1}}$. Then we have*

$$\sum_{i=1}^{k} \alpha_i \geq \frac{(k-1)c}{L}.$$

As mentioned in the beginning of this section, we can no longer rely on a lower bound for every individual step size but want to use their structure to lower bound their sum directly. Specifically, whenever option two is active for α_i (meaning that the second component in (22) attains the minimum), then we can easily show that $\alpha_{i-2} + \alpha_i \geq \frac{\phi}{L}$. On the other hand, if option two is not present for a long time, then we will have a geometric progression $\alpha_i, \alpha_i\gamma, \alpha_i\gamma^2, \dots$, whose sum is also easy to bound. These are two main ingredients, however combining the two requires an intricate technical argument.

Proof We call $\alpha_k = \gamma\alpha_{k-1}$ and $\alpha_k = \frac{\phi^2}{4\alpha_{k-2}L_k^2}$ as the first and second options that the step size α_k can take respectively.

Claim 1. If α_k satisfies the second option, for $k \geq 2$, then $\alpha_{k-2} + \alpha_k \geq \frac{\phi}{L_k} \geq \frac{\phi}{L}$.

This follows directly from $a + b \geq 2\sqrt{ab}$.

Claim 2. $0 < c \leq \frac{\phi}{2}$.

The first inequality follows from $\gamma > 1$ and $\lim_{m\to\infty} \frac{\phi}{m}\sqrt{\frac{\gamma^{m-1}-1}{\gamma-1}} = \infty$. The second one follows from setting $m = 2$ in the definition of $c = \min_{m\geq 2} \frac{\phi}{m}\sqrt{\frac{\gamma^{m-1}-1}{\gamma-1}}$.

We wish to show that

$$\sum_{i=1}^{k} \alpha_i \geq \frac{c(k-1)}{L}.$$

If there does not exist an α_i smaller than $\frac{c}{L}$, then we are done. So assume that such an α_i exists.

For $s \geq 1$, let us call by a *tail* a maximal subsequence of consecutive elements $\alpha_{s+1}, \ldots, \alpha_{s+m}$ such that it starts from $\alpha_{s+1} \geq \frac{c}{L}$ and the rest of the elements are all smaller:

$$\alpha_j < \frac{c}{L} \quad \text{for all } j = s+2, \ldots, s+m \quad \text{and} \quad \alpha_{s+m+1} \geq \frac{c}{L}.$$

Notice that for every such a tail, $\alpha_{s+2} < \alpha_{s+1}$, which means that the second option for α_{s+2} is active. By Claim 1, this implies that $\alpha_s \geq \frac{c}{L}$. We call this element α_s as a *head*. Thus, for every tail $\alpha_{s+1}, \ldots, \alpha_{s+m}$ there is a preceding element α_s that is a head. This was the case when $s \geq 1$. If we have a tail with $s = 0$, we call this sequence $\alpha_1, \ldots, \alpha_m$ the *initial tail*. Note that the initial tail lacks a head (i.e. α_0 is not part of (23)), which is the reason why we will have to consider this case separately.

As a result, we can partition $(\alpha_i)_{i=1}^k$ into a non-overlapping sequence of *(i)* the initial tail (possibly empty), *(ii)* head-tail pairs, and *(iii)* elements larger than $\frac{c}{L}$. It is sufficient to show the bound for each part in our partition.

Head-tail sequence. For a head-tail sequence $\alpha_s, \alpha_{s+1}, \ldots, \alpha_{s+m}$, we will show that

$$\sum_{i=s}^{s+m} \alpha_i \geq \frac{c(m+1)}{L}.$$

As we have already mentioned, for α_{s+2} we have that $\alpha_{s+2} = \frac{\phi^2}{4\alpha_s L_{s+2}^2} \geq \frac{\phi^2}{4\alpha_s L^2}$. For elements $\alpha_{s+4}, \ldots, \alpha_{s+m}$ only the first option can occur, since otherwise there will be a contradiction with Claim 1. For α_{s+3}, however, both options are possible:

$$\alpha_{s+3} = \gamma\alpha_{s+2} \geq \frac{\gamma\phi^2}{4\alpha_s L^2} \quad \text{or} \quad \alpha_{s+3} = \frac{\phi^2}{4\alpha_{s+1} L_{s+3}^2} \geq \frac{\phi^2}{4\alpha_{s+1} L^2}.$$

If $\alpha_{s+3} = \gamma\alpha_{s+2} \geq \frac{\gamma\phi^2}{4\alpha_s L^2}$, then

$$\begin{aligned}
\sum_{i=s}^{s+m} \alpha_i &\geq \alpha_s + \alpha_{s+1} + \frac{\phi^2}{4\alpha_s L^2} + \frac{\gamma\phi^2}{4\alpha_s L^2} + \cdots + \frac{\gamma^{m-2}\phi^2}{4\alpha_s L^2} \\
&= \alpha_{s+1} + \alpha_s + \frac{\phi^2}{4\alpha_s L^2}(1 + \cdots + \gamma^{m-2}) \\
&= \alpha_{s+1} + \alpha_s + \frac{\phi^2}{4\alpha_s L^2}\frac{\gamma^{m-1}-1}{\gamma-1} \\
&\geq \alpha_{s+1} + \frac{\phi}{L}\sqrt{\frac{\gamma^{m-1}-1}{\gamma-1}} \\
&\geq \frac{c}{L} + \frac{mc}{L} = \frac{c(m+1)}{L}, \qquad (25)
\end{aligned}$$

where the last inequality follows from $\alpha_{s+1} \geq \frac{c}{L}$ and our choice of constant c.

If the second option is active for α_{s+3}, that is $\alpha_{s+3} \geq \frac{\phi^2}{4\alpha_{s+1}L^2}$, we have a similar estimation to (25):

$$\begin{aligned}
\sum_{i=s}^{s+m} \alpha_i &\geq \alpha_s + \alpha_{s+1} + \alpha_{s+2} + \frac{\phi^2}{4\alpha_{s+1} L^2} + \frac{\gamma\phi^2}{4\alpha_{s+1} L^2} + \cdots + \frac{\gamma^{m-3}\phi^2}{4\alpha_{s+1} L^2} \\
&= (\alpha_s + \alpha_{s+2}) + \alpha_{s+1} + \frac{\phi^2}{4\alpha_{s+1} L^2}(1 + \cdots + \gamma^{m-3}) \\
&= (\alpha_s + \alpha_{s+2}) + \alpha_{s+1} + \frac{\phi^2}{4\alpha_{s+1} L^2}\frac{\gamma^{m-2}-1}{\gamma-1} \\
&\geq (\alpha_s + \alpha_{s+2}) + \frac{\phi}{L}\sqrt{\frac{\gamma^{m-2}-1}{\gamma-1}} \\
&\geq \frac{2c}{L} + \frac{c(m-1)}{L} = \frac{c(m+1)}{L}, \qquad (26)
\end{aligned}$$

where the last inequality follows from $\alpha_s + \alpha_{s+2} \geq \frac{\phi}{L} \geq \frac{2c}{L}$ and the definition of c.

Hence, in both cases the desired bound holds.

Initial tail. For an empty initial tail there is nothing to prove. For a non-empty tail $\alpha_1, \ldots, \alpha_m$, we will show that

$$\sum_{i=1}^{m} \alpha_i \geq \frac{c(m-1)}{L}.$$

Since $\alpha_1 \geq \frac{\phi}{2L} \geq \frac{c}{L}$, the first option cannot be active for α_2. Hence, $\alpha_2 = \frac{\phi^2}{4\alpha_0 L_2^2} \geq \frac{\phi^2}{4\alpha_0 L^2}$. First of all, note that for $m = 1$ the desired inequality obviously holds: $\sum_{i=1}^{1} \alpha_i \geq 0$. For $m = 2$, we have by Claim 2

$$\sum_{i=1}^{2} \alpha_i \geq \alpha_0 + \alpha_2 = \alpha_0 + \frac{\phi^2}{4\alpha_0 L^2} \geq \frac{\phi}{L} \geq \frac{c}{L}.$$

Thus, for the rest of the proof we suppose that $m \geq 3$.

For steps $\alpha_4, \ldots, \alpha_m$ only the first option can be active (by Claim 1). For α_3 we have two cases:

1. The first option for α_3 is active, that is $\alpha_3 = \gamma\alpha_2$. Then similarly to (25) we have

$$\begin{aligned}\sum_{i=1}^{m} \alpha_i &\geq \alpha_1 + \alpha_2 + \gamma\alpha_2 + \cdots + \gamma^{m-2}\alpha_2 \\ &\geq \alpha_1 + \frac{\phi^2}{4\alpha_0 L^2}\frac{\gamma^{m-1}-1}{\gamma-1} \\ &\geq \alpha_0 + \frac{\phi^2}{4\alpha_0 L^2}\frac{\gamma^{m-1}-1}{\gamma-1} \\ &\geq \frac{\phi}{L}\sqrt{\frac{\gamma^{m-1}-1}{\gamma-1}} \\ &\geq \frac{cm}{L}. \end{aligned} \tag{27}$$

2. The second option for α_3 is active, that is $\alpha_3 = \frac{\phi^2}{4\alpha_1 L_3^2} \geq \frac{\phi^2}{4\alpha_1 L^2}$. Then similarly to (26) we have

$$\begin{aligned}\sum_{i=1}^{m} \alpha_i &= \alpha_1 + \alpha_2 + \alpha_3 + \gamma\alpha_3 + \cdots + \gamma^{m-3}\alpha_3 \\ &= \alpha_1 + \alpha_2 + \alpha_3(1 + \cdots + \gamma^{m-3}) \\ &\geq \alpha_1 + \alpha_2 + \frac{\phi^2}{4\alpha_1 L^2}\frac{\gamma^{m-2}-1}{\gamma-1} \\ &\geq \alpha_2 + \frac{\phi}{L}\sqrt{\frac{\gamma^{m-2}-1}{\gamma-1}} \\ &\geq \frac{c(m-1)}{L}, \end{aligned} \tag{28}$$

   where we used that $\alpha_2 \geq 0$.

Let us summarize what we have proved:

- the sequence $\alpha_1, \ldots, \alpha_k$ can be divided into an initial part $\alpha_1, \ldots \alpha_m$, some non-overlapping head-tails, and the remaining elements;
- for every head-tail $\alpha_s, \ldots, \alpha_{s+m}$ we showed that $\sum_{i=0}^{m} \alpha_{s+i} \geq \frac{c(m+1)}{L}$;
- for the initial tail we showed that $\sum_{i=1}^{m} \alpha_i \geq \frac{c(m-1)}{L}$;
- the remaining elements in $\alpha_1, \ldots, \alpha_k$ are always greater or equal than $\frac{c}{L}$.

Hence, for each subsequence of length l, the sum of its elements is at least $\frac{c(l-1)}{L}$. This allows us to conclude that

$$\sum_{i=1}^{k} \alpha_i \geq \frac{c(k-1)}{L}.$$

■

Theorem 5 *Let Assumption 1 (i, iii, iv) hold and F be locally Lipschitz. Let $\phi \in \left(1, \frac{1+\sqrt{5}}{2}\right)$, $\gamma = \frac{1}{\phi} + \frac{1}{\phi^2}$, $c = \min_{m \geq 2} \frac{\phi}{m} \sqrt{\frac{\gamma^{m-1}-1}{\gamma-1}} > 0$, and $\mathbf{Z}^k = \frac{1}{\sum_{i=1}^k \alpha_i} \sum_{i=1}^k \alpha_i \mathbf{z}^i$. Then Alg. 1 with α_0 as in (24) has the rate*

$$\mathrm{Gap}_S\left(\mathbf{Z}^k\right) \leq \frac{LD}{2c(k-1)}.$$

Proof The result follows by using the definition of $\mathbf{Z}^k$ on the result of Lemma 3 and the lower bound of $\sum_{i=1}^k \alpha_k \geq \frac{(k-1)c}{L}$ derived in Lemma 4. ■

The last thing left to understand is the value of c. The next remark shows that it is large enough for a meaningful choice of parameters.

Remark 6 *Setting $\phi = 1.5$ as in the experiments of (Malitsky, 2020) direct calculation gives $c \geq 0.5$.*

By proving a novel lower bound on the sum of the step sizes, we are able to prove a $\mathcal{O}(\frac{1}{k})$ convergence rate without requiring an artificial upper bound $\bar{\alpha}$ on each individual step. Not only does this simplify the aGRAAL method, but also removes the spurious L^2 dependence from (Malitsky, 2020), showing that there is basically no extra cost of adaptivity.

5. Nonmonotone problems with weak Minty solutions

In Section 3, for monotone problems, we observed empirically that increasing ϕ leads to a certain speed-up, as in Fig. 2. In this section, we turn to a special class of *nonmonotone* problems and show that in some cases *smaller* ϕ can be also favorable.

In particular, we consider the class of unconstrained problems $F(\mathbf{z}^*) = 0$ exhibiting a *weak Minty* solution, which is given by a point $\mathbf{z}^*$ such that

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

This notion was recently introduced in (Diakonikolas et al., 2021) in the context of von Neumann ratio games — a nonconvex-nonconcave min-max problem — and further investigated in (Lee and Kim, 2021b; Pethick et al., 2022; Böhm, 2023). While it is difficult to verify the existence of such solutions in practice, the template simultaneously captures different generalizations of monotonicity like the existence of a Minty solution (Malitsky, 2020; Mertikopoulos et al., 2019) and negative comonotonicity (Lee and Kim, 2021a; Bauschke et al., 2020) in order to study nonconvex-nonconcave min-max problems. See (Lee and Kim,

2021a) for the implications between these different monotonicity concepts and (Gorbunov et al., 2022) for a comparison in terms of last iterate convergence. To our knowledge, we give the first adaptive algorithm for this setting that does not require a linesearch at every iteration and only relies on *local* Lipschitz continuity.

5.1 Constant step size

We first show the convergence of unconstrained GRAAL with a constant step size. Note that the discussion below will therefore also hold for the generalized version of OGDA, as observed in Section 2.2.

Theorem 7 *Let F be L-Lipschitz and fulfill Assumption* (WM)*, with $\rho < \frac{1}{L}$. Let $\phi > 1$ be such that $\delta := \frac{2-\phi}{\phi L} - \rho > 0$. Let $(\mathbf{z}^k)$ and $(\bar{\mathbf{z}}^k)$ be the iterates generated* (GRAAL) *with $\alpha = \frac{2-\phi}{L}$. Then*

$$\min_{i=1,\dots,k} \|F(\mathbf{z}^k)\|^2 \leq \frac{L\phi}{k(2-\phi)(\phi-1)\delta}\left(\|\bar{\mathbf{z}}^1 - \mathbf{z}^*\|^2 + \frac{2}{\phi}\|\mathbf{z}^1 - \mathbf{z}^0\|^2\right).$$

In the limiting case when ϕ is close to 1 we can allow for $\rho < \frac{1}{L}$ in Theorem 7, which is precisely the best possible dependence for the (adaptive) EG+ method proven in (Pethick et al., 2022). For a more moderate choice, for example $\phi = \varphi = \frac{\sqrt{5}+1}{2}$, the bound on ρ tightens to $\rho < \frac{1}{2L}$. See also Fig. 3 for empirical evidence of the need to reduce ϕ for problems with weak Minty solution.

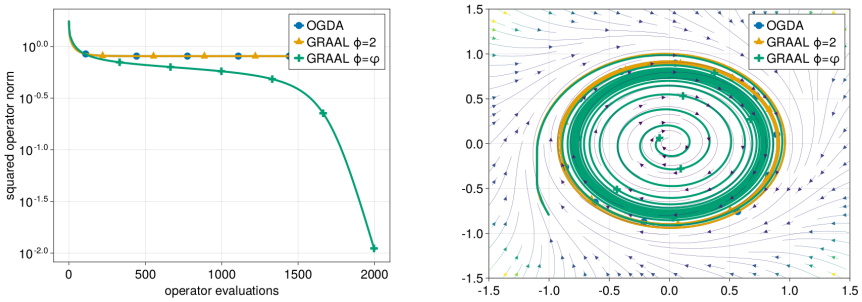

Figure 3: A special parametrization of the Polar Game example from (Pethick et al., 2022) (see Section 5.3.2 for details), showing the need to reduce ϕ for nonmonotone problems with weak Minty solutions.

Theorem 7 has the drawback of requiring knowledge of the weak Minty parameter ρ to set ϕ such that $\delta > 0$. In (Pethick et al., 2022) a similar problem was partially circumvented by an elaborate way to choose the corresponding parameter adaptively. Whether something similar can be done here is an open question. In practice, one can use a simple approach: run GRAAL with certain value of ϕ; if it does not converge, decrease and try again.

5.2 Adaptive step size

We continue with the guarantees of unconstrained aGRAAL under Assumption (WM).

Theorem 8 *Let F be locally Lipschitz and fulfill Assumption* (WM)*, and let $(\mathbf{z}^k)$ be the sequence generated by* (aGRAAL)*. As long as $\delta := \inf_k \alpha_k\Big(1 + \frac{1}{\phi} - \gamma\phi\Big) - \rho > 0$ we have*

$$\min_{i=1,\dots,k+1} \|F(\mathbf{z}^i)\|^2 \leq \frac{D}{\delta \sum_{i=1}^{k+1} \alpha_i} \leq \frac{cLD}{k\delta},$$

where c denotes the constant given in Lemma 4.

Remark 9 *In particular, picking $\phi = 1.1$ and $\gamma = 1.1$ gives the condition $\rho < 0.69\alpha_k$. As ϕ and γ get closer to 1, we can allow for $\rho < \alpha_k$, which would precisely be the dependence between ρ and the step size proven in (Pethick et al., 2022) for CurvatureEG+, a version of EG with linesearch.*

Corollary 10 *In the setting of Theorem 8, if the iterates happen to remain bounded, we only require*

$$\inf_k \frac{\alpha_{k+1}}{\phi} + \alpha_k\Big(1 + \frac{1}{\phi} - \gamma\phi\Big) - \rho > 0,$$

to deduce a similar convergence statement, but with modified constant.

Remark 11 *The condition in Corollary 10 between the step sizes and the weak Minty parameter ρ reduces, in the limiting case $\phi, \gamma \to 1$, to $\alpha_{k+1} + \alpha_k > \rho$. This corresponds to doubling of the range of ρ presented in Remark 9.*

Remark 12 (Lower bounding α_k.) *In this section we relaxed the condition between the Lipschitz constant L and parameter ρ in* (WM) *to a condition between α_k and ρ. We do so in the hope that aGRAAL is able to take larger steps, since they are based on the local Lipschitz constants. Unfortunately, we cannot give a good lower bound on any individual α_k in general. We see from the experiments that the step size sometimes does become lower than the one chosen by linesearch (see Fig. 4). This seems to, however, not harm the performance of the algorithm.*

5.3 Experiments

As discussed in the main section of the paper, in order to solve weak Minty problems, for all known methods two things are important: *(i)* the method needs to be modified in a way that could be interpreted as making it more "conservative"; *(ii)* step sizes should be large.

In the case of the golden ratio algorithm, the former means averaging with more of the (old) $\bar{\mathbf{z}}^{k-1}$ iterate, i.e. decreasing ϕ, see Fig. 4. For EG this means reducing the step size in the update of the $\mathbf{z}^k$ sequence, as proposed in (Diakonikolas et al., 2021; Pethick et al., 2022). For the constant step size methods in question it seems like this is all we can do. However, by choosing the step sizes adaptively, be it via linesearch as for the CurvatureEG+ method from (Pethick et al., 2022), or by (3) in the case of (aGRAAL), we can hope to take steps larger than the global Lipschitz constant would allow.

5.3.1 Forsaken

In Fig. 4 the following problem formulation is used, which originated in (Hsieh et al., 2021, Example 5.2) as a particularly difficult instance of min-max under the name of "Forsaken":

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

where $f(z) = \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$. This problem exhibits a solution at $(x^*, y^*) \approx (0.08, 0.4)$, but also *two* limit cycles not containing any critical point of the objective function.

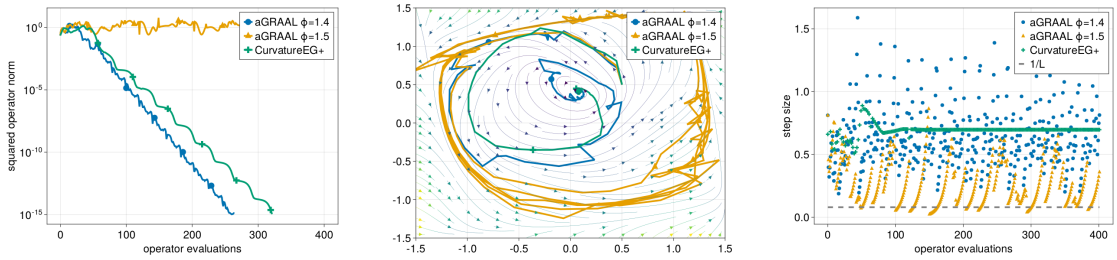

Figure 4: The "Forsaken" example of (Hsieh et al., 2021, Example 5.2), which is further explored in (Pethick et al., 2022). If ϕ is too large in aGRAAL, the method gets stuck in the limit cycle. Only for small ϕ the method converges, and does so rapidly.

5.3.2 Polar Game

Fig. 3 and 5 display results on the so-called Polar Game introduced in (Pethick et al., 2022, Example 3) using the following parametrization for $a > 0$

$$F(x, y) = (\psi(x, y) - y, \psi(y, x) + x),$$

where $\psi(x, y) = a\frac{1}{4}x(-1 + x^2 + y^2)(-1 + 4x^2 + 4y^2)$. The problem exhibits a limit cycle attracting solutions away from the solution in the center.

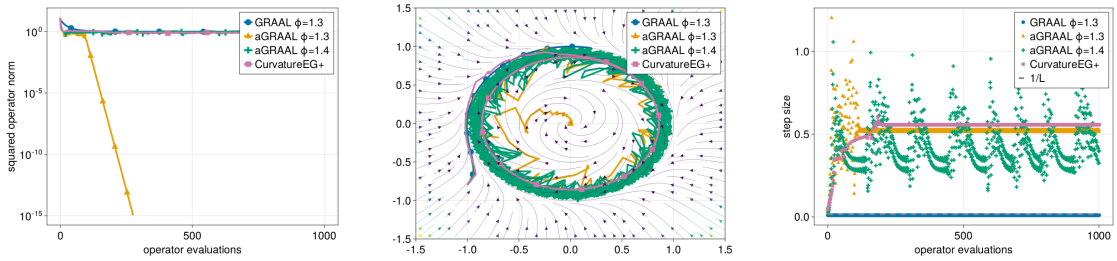

Figure 5: A different parametrization of the Polar Game, $a = 3$, showing the importance of adaptive step sizes. We observe that (aGRAAL) is the only method able to converge.

Fig. 5 shows that reducing ϕ alone will not be enough for problems where $\rho > \frac{1}{L}$ is not satisfied (the blue line does not converge). One has to choose larger step sizes.

5.3.3 A lower bound (Pethick et al., 2022, Example 5)

In Pethick et al. (2022) it was shown that EG+ (with arbitrary small second step size) may diverge if $\rho > \frac{1}{L}$ via the following unconstrained min-max problems

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} axy + \frac{b}{2}(x^2 - y^2),$$

for $a > 0$ and $b < 0$. The associated operator F is simply given by

$$F(x, y) = (ay + bx, by - ax)$$

and is Lipschitz continuous with constant $L = \sqrt{a^2 + b^2}$ and $(0, 0)$ is a weak Minty solution with constant $\rho = -\frac{2b}{a^2+b^2}$.

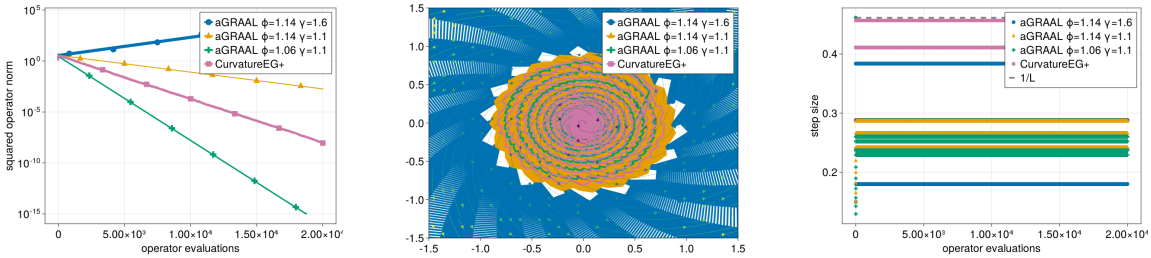

Figure 6: A special parametrization ($a^2 = 3.7, b = -1$) of (Pethick et al., 2022, Example 5), illustrating the fact that sometimes γ does indeed need to be chosen smaller as predicted by theory.

In Fig. 6 we see that, as predicted by Theorem 8 and Corollary 10, not only reducing ϕ but also decreasing γ can prevent divergence. At first glance, this is somewhat surprising as larger γ allows for a faster increase of the step size from one iteration to the next, and we have already seen that larger step sizes are important for convergence for this class of problems. We suspect this to be rooted in the special step size rule of aGRAAL, see (3), where one large step can lead to decrease in the step size of the next iteration.

Acknowledgments

The work of A. Böhm was funded by the Austrian Science Fund (FWF): W1260-N35. The work of A. Alacaoglu was funded by NSF awards 2023239 and 2224213; and DOE ASCR Subcontract 8F-30039 from Argonne National Laboratory. The work of Yura Malitsky was supported by the Wallenberg AI, Autonomous Systems and Software Program funded by the Knut and Alice Wallenberg Foundation, No. 305286.

Appendix A. Details on Section 3

For convenience, we give a full description of the algorithm

Algorithm 2 GRAAL

Require: $\bar{\mathbf{z}}^{-1} = \mathbf{z}^0 \in C$, $\phi \in (1, 2]$.

1: **for** $k \geq 0$ **do**
2: $\quad \bar{\mathbf{z}}^k = \frac{\phi-1}{\phi}\mathbf{z}^k + \frac{1}{\phi}\bar{\mathbf{z}}^{k-1}$
3: $\quad \mathbf{z}^{k+1} = P_C(\bar{\mathbf{z}}^k - \alpha F(\mathbf{z}^k))$

A.1 Proof of Corollary 2

We start with the following lemma which essentially summarizes the existing result of (Malitsky, 2020) for one iteration of the algorithm. Letting $\phi = 2$ and $\varepsilon = 0$ in this lemma gives (8) which is the starting point of Theorem 1.

Lemma 13 (based on Theorem 2 in (Malitsky, 2020)) *Let Assumption 1 hold and* $\alpha = \frac{\phi-2\varepsilon}{2L}$ *for any* $\varepsilon \in [0, \phi/2)$ *in* (GRAAL). *Then for any* $\mathbf{z} \in C$

$$G(\mathbf{z}^k, \mathbf{z}) + \frac{\phi}{\phi-1}\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}\|^2 + \frac{\phi}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \left(\phi - 1 - \frac{1}{\phi}\right)\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \\ \leq \frac{\phi}{\phi-1}\|\bar{\mathbf{z}}^k - \mathbf{z}\|^2 + \frac{(\phi-2\varepsilon)^2}{2\phi}\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2 - \frac{1}{\phi}\|\mathbf{z}^k - \bar{\mathbf{z}}^{k-1}\|^2.$$

Proof [of Lemma 13] By applying prox-inequality to the definition of $\mathbf{z}^{k+1}$, we have for any $\mathbf{z} \in C$ that

$$\langle \mathbf{z}^{k+1} - \bar{\mathbf{z}}^k + \alpha F(\mathbf{z}^k), \mathbf{z} - \mathbf{z}^{k+1}\rangle \geq 0. \tag{29}$$

Similarly, with the definition of $\mathbf{z}^k$, we have

$$\langle \mathbf{z}^k - \bar{\mathbf{z}}^{k-1} + \alpha F(\mathbf{z}^{k-1}), \mathbf{z}^{k+1} - \mathbf{z}^k\rangle \geq 0. \tag{30}$$

Since $\mathbf{z}^k - \bar{\mathbf{z}}^{k-1} = \frac{1+\phi}{\phi}(\mathbf{z}^k - \bar{\mathbf{z}}^k) = \phi(\mathbf{z}^k - \bar{\mathbf{z}}^k)$, we can rewrite (30) as

$$\langle \phi(\mathbf{z}^k - \bar{\mathbf{z}}^k) + \alpha F(\mathbf{z}^{k-1}), \mathbf{z}^{k+1} - \mathbf{z}^k\rangle \geq 0. \tag{31}$$

Summing up (29) and (31) yields

$$\langle \mathbf{z}^{k+1} - \bar{\mathbf{z}}^k, \mathbf{z} - \mathbf{z}^{k+1}\rangle + \phi\langle \mathbf{z}^k - \bar{\mathbf{z}}^k, \mathbf{z}^{k+1} - \mathbf{z}^k\rangle + \alpha\langle F(\mathbf{z}^k), \mathbf{z} - \mathbf{z}^k\rangle \\ + \alpha\langle F(\mathbf{z}^k) - F(\mathbf{z}^{k-1}), \mathbf{z}^k - \mathbf{z}^{k+1}\rangle \geq 0. \tag{32}$$

Expressing the first two terms in (32) by norms and using the definition of $G(\mathbf{z}^k, \mathbf{z})$ from (7), we deduce

$$G(\mathbf{z}^k, \mathbf{z}) + \|\mathbf{z}^{k+1} - \mathbf{z}\|^2 \leq \|\bar{\mathbf{z}}^k - \mathbf{z}\|^2 - \|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 + 2\alpha\langle F(\mathbf{z}^k) - F(\mathbf{z}^{k-1}), \mathbf{z}^k - \mathbf{z}^{k+1}\rangle \\ + \phi(\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 - \|\mathbf{z}^k - \bar{\mathbf{z}}^k\|^2). \tag{33}$$

The definition of $\bar{\mathbf{z}}^k$ gives

$$\begin{aligned}\|\mathbf{z}^{k+1}-\mathbf{z}\|^2 &= \frac{\phi}{\phi-1}\|\bar{\mathbf{z}}^{k+1}-\mathbf{z}\|^2 - \frac{1}{\phi-1}\|\bar{\mathbf{z}}^{k}-\mathbf{z}\|^2 + \frac{\phi}{(\phi-1)^2}\|\bar{\mathbf{z}}^{k+1}-\bar{\mathbf{z}}^{k}\|^2 \\ &= \frac{\phi}{\phi-1}\|\bar{\mathbf{z}}^{k+1}-\mathbf{z}\|^2 - \frac{1}{\phi-1}\|\bar{\mathbf{z}}^{k}-\mathbf{z}\|^2 + \frac{1}{\phi}\|\mathbf{z}^{k+1}-\bar{\mathbf{z}}^{k}\|^2, \end{aligned} \tag{34}$$

where the second equality used $\bar{\mathbf{z}}^{k+1}-\bar{\mathbf{z}}^k = \frac{\phi-1}{\phi}(\mathbf{z}^{k+1}-\bar{\mathbf{z}}^k)$.

Additionally, from $\alpha = \frac{\phi-2\varepsilon}{2L}$ and Young's inequality, we have

$$\begin{aligned}2\alpha\langle F(\mathbf{z}^k)-F(\mathbf{z}^{k-1}), \mathbf{z}^k-\mathbf{z}^{k+1}\rangle &\le (\phi-2\varepsilon)\|F(\mathbf{z}^k)-F(\mathbf{z}^{k-1})\|\|\mathbf{z}^{k+1}-\mathbf{z}^k\| \\ &\le \frac{\phi}{2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2 + \frac{(\phi-2\varepsilon)^2}{2\phi}\|\mathbf{z}^k-\mathbf{z}^{k-1}\|^2. \end{aligned} \tag{35}$$

Using (34), (35), and $\phi^2\|\mathbf{z}^k-\bar{\mathbf{z}}^k\|^2 = \|\mathbf{z}^k-\bar{\mathbf{z}}^{k-1}\|^2$ on (33) gives the desired conclusion. ■

Lemma 14 (Upper bound of R^2 used in Theorem 1) *Let Assumption 1 hold. Then for the first iteration of Alg. 2, we have that*

$$R^2 = \|\mathbf{z}^1-\mathbf{z}^*\|^2 + \|\mathbf{z}^0-\mathbf{z}^*\|^2 \le 3\|\mathbf{z}^0-\mathbf{z}^*\|^2.$$

Proof Recall that $\mathbf{z}^1 = P_C(\mathbf{z}^0-\alpha F(\mathbf{z}^0))$ and $\mathbf{z}^* = P_C(\mathbf{z}^*-\alpha F(\mathbf{z}^*))$. The former is because $\bar{\mathbf{z}}^0 = \mathbf{z}^0$ by the initialization in Alg. 2. The latter follows by the definition of $\mathbf{z}^*$ as a solution of (1). By firm-nonexpansiveness of P_C, we have

$$\|\mathbf{z}^1-\mathbf{z}^*\|^2 \le \|(\mathbf{z}^0-\alpha F(\mathbf{z}^0))-(\mathbf{z}^*-\alpha F(\mathbf{z}^*))\|^2 - \|\mathbf{z}^0-\alpha F(\mathbf{z}^0)-\mathbf{z}^1+\alpha F(\mathbf{z}^*)\|^2.$$

Expanding the terms in the right-hand side we have that

$$\begin{aligned}\|\mathbf{z}^1-\mathbf{z}^*\|^2 \le \|\mathbf{z}^0-\mathbf{z}^*\|^2 - \|\mathbf{z}^1-\mathbf{z}^0\|^2 + 2\alpha\langle F(\mathbf{z}^0)-F(\mathbf{z}^*), \mathbf{z}^0-\mathbf{z}^1\rangle \\ - 2\alpha\langle F(\mathbf{z}^0)-F(\mathbf{z}^*), \mathbf{z}^0-\mathbf{z}^*\rangle. \end{aligned} \tag{36}$$

By monotonicity, we have $\langle F(\mathbf{z}^0)-F(\mathbf{z}^*), \mathbf{z}^0-\mathbf{z}^*\rangle \ge 0$. By using $\alpha \le \frac{1}{L}$, we have

$$2\alpha\langle F(\mathbf{z}^0)-F(\mathbf{z}^*), \mathbf{z}^0-\mathbf{z}^1\rangle \le 2\alpha L\|\mathbf{z}^0-\mathbf{z}^*\|\|\mathbf{z}^0-\mathbf{z}^1\| \le \|\mathbf{z}^0-\mathbf{z}^*\|^2 + \|\mathbf{z}^0-\mathbf{z}^1\|^2.$$

Combining the last two estimates for the inner products in (36), we have

$$\|\mathbf{z}^1-\mathbf{z}^*\|^2 \le 2\|\mathbf{z}^0-\mathbf{z}^*\|^2,$$

which gives the result. ■

A.2 Alternative proof of Theorem 1 via SDP

The proof of boundedness of $(\mathbf{z}^k)$ presented in the main part was not very standard. In this section, we provide an alternative proof of that fact which is based on a semidefinite program combined with induction. This approach is easily generalizable to the case of $\phi < 2$ even though we give it for $\phi = 2$ for simplicity.

As before, we assume that $\|\bar{\mathbf{z}}^i - \mathbf{z}^*\|^2 \le 2R^2$ for all $i = 1, \dots, k$ and we must show that $\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 \le 2R^2$. Our main tool will be inequality (11) which holds for all $k \ge 1$ and which we recall here

$$2\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 + \|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 \le \|\mathbf{z}^1 - \mathbf{z}^*\|^2 + \|\mathbf{z}^0 - \mathbf{z}^*\|^2 + \frac{1}{2}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \tag{37}$$

Without loss of generality, we assume that $\mathbf{z}^* = 0$ and that $R^2 = \|\mathbf{z}^1 - \mathbf{z}^*\|^2 + \|\mathbf{z}^0 - \mathbf{z}^*\|^2 \le 1$. The first assumption is valid because we can always redefine sequences as $\mathbf{z}^k := \mathbf{z}^k - \mathbf{z}^*$ and $\bar{\mathbf{z}}^k := \bar{\mathbf{z}}^k - \mathbf{z}^*$, while the second — because we can rescale the norm $\|\cdot\|$ by any positive number.

Using these two assumptions and that $\mathbf{z}^k = 2\bar{\mathbf{z}}^k - \bar{\mathbf{z}}^{k-1}$ we can rewrite (37) as

$$2\|\bar{\mathbf{z}}^{k+1}\|^2 + \|2\bar{\mathbf{z}}^{k+1} - \bar{\mathbf{z}}^k - 2\bar{\mathbf{z}}^k + \bar{\mathbf{z}}^{k-1}\|^2 \le 1 + \frac{1}{2}\|2\bar{\mathbf{z}}^{k+1} - \bar{\mathbf{z}}^k - \bar{\mathbf{z}}^k\|^2,$$

or

$$2\|\bar{\mathbf{z}}^{k+1}\|^2 + \|2\bar{\mathbf{z}}^{k+1} - 3\bar{\mathbf{z}}^k + \bar{\mathbf{z}}^{k-1}\|^2 \le 1 + 2\|\bar{\mathbf{z}}^{k+1} - \bar{\mathbf{z}}^k\|^2. \tag{38}$$

The latter inequality is quadratic in $\bar{\mathbf{z}}^{k+1}, \bar{\mathbf{z}}^k, \bar{\mathbf{z}}^{k-1}$. Hence, after expanding the norms, we can rewrite it in a matrix notation as

$$\operatorname{tr}(\mathbf{G} \cdot \mathbf{M}) \le 1,$$

where $\mathbf{G}$ is a Gram matrix

$$\mathbf{G} = \begin{bmatrix} \langle \bar{\mathbf{z}}^{k+1}, \bar{\mathbf{z}}^{k+1} \rangle & \langle \bar{\mathbf{z}}^{k+1}, \bar{\mathbf{z}}^{k} \rangle & \langle \bar{\mathbf{z}}^{k+1}, \bar{\mathbf{z}}^{k-1} \rangle \\ \langle \bar{\mathbf{z}}^{k}, \bar{\mathbf{z}}^{k+1} \rangle & \langle \bar{\mathbf{z}}^{k}, \bar{\mathbf{z}}^{k} \rangle & \langle \bar{\mathbf{z}}^{k}, \bar{\mathbf{z}}^{k-1} \rangle \\ \langle \bar{\mathbf{z}}^{k-1}, \bar{\mathbf{z}}^{k+1} \rangle & \langle \bar{\mathbf{z}}^{k-1}, \bar{\mathbf{z}}^{k} \rangle & \langle \bar{\mathbf{z}}^{k-1}, \bar{\mathbf{z}}^{k-1} \rangle \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 4 & -4 & 2 \\ -4 & 7 & -3 \\ 2 & -3 & 1 \end{bmatrix}.$$

By induction assumption we have that $\mathbf{G}_{22} \le 2$ and $\mathbf{G}_{33} \le 2$ and we must show that $\mathbf{G}_{11} \le 2$. Consider the following semidefinite program

$$\begin{aligned} \max_{\mathbf{G}} \ & \mathbf{G}_{11} \quad \text{subject to} \\ & \mathbf{G} \succcurlyeq 0 \\ & \operatorname{tr}(\mathbf{G} \cdot \mathbf{M}) \le 1 \\ & \mathbf{G}_{22} \le 2 \\ & \mathbf{G}_{33} \le 2 \end{aligned}$$

If we can show that its optimal value is less than 2, then we are done: it will automatically imply that $\|\bar{\mathbf{z}}^{k+1}\|^2 \le 2$. Now, by solving it, we obtain $\mathbf{G}_{11} \approx 1.49259$.

Therefore, we have proved that for all k, $\|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 \le 2R^2$.

Remark 15 *The semidefinite program actually allows us to show a slightly tighter bound:* $\|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 \le 1.2R^2$*, but we kept the constant* 2 *for consistency with the previous approach.*

Appendix B. Details for Section 4

B.1 Implementation of the linesearch

In (24) we outline a very particular linesearch which decreases α_0 by γ in every step. Since the canonical choice of ϕ proposed in (Malitsky, 2020) yields the small $\gamma = 1.1$ this could result in many backtracking iterations if our initial guess is bad. For practical implementation, therefore, it is better to use first a coarse reduction of α_0, say with a factor of 10 and only at the end to switch to a fine factor γ.

Appendix C. Weak Minty proofs

C.1 Constant step size

By following the line of reasoning as in the monotone case one can only derive $\rho < \frac{1}{2L}$. We do not present the proof here but it is only a simple modification similar to the one we show later for the adaptive step size. In order to derive the $\rho < \frac{1}{L}$ bound (matching EG) one has to take a completely different approach. The high-level reason is that the monotone analysis requires $\alpha \leq \frac{\phi}{2L}$. This is counter intuitive as a smaller ϕ makes (GRAAL) more conservative since the iterates are more anchored to $\bar{\mathbf{z}}^k$, and at the same time requires a smaller step size. A more conservative averaging of the iterates should allow for a more aggressive step size, which is precisely the behavior we observe here. For convenience let $\mathbf{g}^k = F(\mathbf{z}^k)$.

Lemma 16 *Let F be L-Lipschitz and fulfill Assumption* (WM). *Let $(\mathbf{z}^k)$ and $(\bar{\mathbf{z}}^k)$ be the iterates generated* (GRAAL). *Then*

$$\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 \leq \|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 + \frac{\phi-1}{\phi}\alpha\rho\|\mathbf{g}^k\|^2 + \frac{4}{\phi(\phi-1)}\alpha\langle \mathbf{g}^k - \mathbf{g}^{k-1}, \mathbf{z}^k - \mathbf{z}^{k+1}\rangle \\ - \frac{3-\phi}{\phi-1}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 - \frac{\alpha^2(\phi+1)}{\phi^2(\phi-1)}\|\mathbf{g}^k - \mathbf{g}^{k-1}\|^2. \quad (39)$$

Proof First we observe simply from the definition of the iterates

$$\begin{aligned}
\|\bar{\mathbf{z}}^{k+1} - \mathbf{z}^*\|^2 &= \left\|\frac{\phi-1}{\phi}\mathbf{z}^{k+1} + \frac{1}{\phi}\bar{\mathbf{z}}^k - \mathbf{z}^*\right\|^2 \\
&= \left\|\frac{\phi-1}{\phi}(\bar{\mathbf{z}}^k - \alpha\mathbf{g}^k) + \frac{1}{\phi}\bar{\mathbf{z}}^k - \mathbf{z}^*\right\|^2 \\
&= \left\|\bar{\mathbf{z}}^k - \frac{\phi-1}{\phi}\alpha\mathbf{g}^k - \mathbf{z}^*\right\|^2 \\
&= \|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 - 2\frac{\phi-1}{\phi}\alpha\langle \mathbf{g}^k, \bar{\mathbf{z}}^k - \mathbf{z}^*\rangle + \left\|\frac{\phi-1}{\phi}\alpha\mathbf{g}^k\right\|^2 \quad (40) \\
&\leq \|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 + \frac{\phi-1}{\phi}\alpha\rho\|\mathbf{g}^k\|^2 - 2\frac{\phi-1}{\phi}\alpha\langle \mathbf{g}^k, \bar{\mathbf{z}}^k - \mathbf{z}^k\rangle + \left\|\frac{\phi-1}{\phi}\alpha\mathbf{g}^k\right\|^2 \\
&= \|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 + \frac{\phi-1}{\phi}\alpha\rho\|\mathbf{g}^k\|^2 - 2\frac{\phi-1}{\phi}\alpha\langle \mathbf{g}^k, \frac{1}{\phi}\alpha\mathbf{g}^{k-1}\rangle + \left\|\frac{\phi-1}{\phi}\alpha\mathbf{g}^k\right\|^2 \\
&= \|\bar{\mathbf{z}}^k - \mathbf{z}^*\|^2 + \frac{\phi-1}{\phi}\alpha\rho\|\mathbf{g}^k\|^2 - \frac{\phi-1}{\phi}\alpha^2\left\langle \mathbf{g}^k, \frac{2}{\phi}\mathbf{g}^{k-1} - \frac{\phi-1}{\phi}\mathbf{g}^k\right\rangle,
\end{aligned}$$

where we used (WM) to deduce the inequality. Now we want to prove the following equality

$$-\frac{\phi-1}{\phi}\alpha^2\left\langle \mathbf{g}^k, \frac{2}{\phi}\mathbf{g}^{k-1}-\frac{\phi-1}{\phi}\mathbf{g}^k\right\rangle = \frac{4}{\phi(\phi-1)}\alpha\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \mathbf{z}^k-\mathbf{z}^{k+1}\rangle \\ -\frac{3-\phi}{\phi-1}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2-\frac{\alpha^2(\phi+1)}{\phi^2(\phi-1)}\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2. \quad (41)$$

We can verify this by expanding all iterates in terms of operators. Observe first that

$$\mathbf{z}^{k+1}-\mathbf{z}^k=\bar{\mathbf{z}}^k-\alpha\mathbf{g}^k-\mathbf{z}^k=\frac{1}{\phi}(\bar{\mathbf{z}}^{k-1}-\mathbf{z}^k)-\alpha\mathbf{g}^k=\frac{1}{\phi}\alpha\mathbf{g}^{k-1}-\alpha\mathbf{g}^k=\alpha\frac{1}{\phi}(\mathbf{g}^{k-1}-\mathbf{g}^k)-\alpha\frac{\phi-1}{\phi}\mathbf{g}^k. \quad (42)$$

Now we use this to deduce

$$\begin{aligned}-\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2 &= -\alpha^2\left\|\frac{1}{\phi}(\mathbf{g}^{k-1}-\mathbf{g}^k)-\frac{\phi-1}{\phi}\mathbf{g}^k\right\|^2 \\ &= -\frac{\alpha^2}{\phi^2}\|\mathbf{g}^{k-1}-\mathbf{g}^k\|^2+\frac{\phi-1}{\phi^2}\alpha^2\langle \mathbf{g}^k, \mathbf{g}^{k-1}-\mathbf{g}^k\rangle-\alpha^2\left(\frac{\phi-1}{\phi}\right)^2\|\mathbf{g}^k\|^2 \\ &= -\frac{\alpha^2}{\phi^2}\|\mathbf{g}^{k-1}-\mathbf{g}^k\|^2+\frac{\phi-1}{\phi^2}\alpha^2\langle \mathbf{g}^k, 2\mathbf{g}^{k-1}-2\mathbf{g}^k-(\phi-1)\mathbf{g}^k\rangle \\ &= -\frac{\alpha^2}{\phi^2}\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2+\frac{\phi-1}{\phi^2}\alpha^2\left\langle \mathbf{g}^k, 2\mathbf{g}^{k-1}-(\phi+1)\mathbf{g}^k\right\rangle.\end{aligned}$$

Therefore, by multiplying both sides by $\frac{3-\phi}{\phi-1}$ we get

$$-\frac{3-\phi}{\phi-1}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2=-\frac{3-\phi}{\phi-1}\frac{\alpha^2}{\phi^2}\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2+\frac{3-\phi}{\phi^2}\alpha^2\left\langle \mathbf{g}^k, 2\mathbf{g}^{k-1}-(\phi+1)\mathbf{g}^k\right\rangle. \quad (43)$$

Again, by going to $\mathbf{g}^k$ everywhere we have

$$\begin{aligned}\frac{\alpha}{\phi}\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \mathbf{z}^k-\mathbf{z}^{k+1}\rangle &= \frac{\alpha^2}{\phi}\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \mathbf{g}^k-\frac{1}{\phi}\mathbf{g}^{k-1}\rangle \\ &= \frac{\alpha^2}{\phi}\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \frac{1}{\phi}(\mathbf{g}^k-\mathbf{g}^{k-1})+(1-\frac{1}{\phi})\mathbf{g}^k\rangle \\ &= \frac{\alpha^2}{\phi^2}\|\mathbf{g}^{k-1}-\mathbf{g}^k\|^2+\frac{\phi-1}{\phi^2}\alpha^2\langle \mathbf{g}^k, \mathbf{g}^k-\mathbf{g}^{k-1}\rangle,\end{aligned}$$

and therefore by multiplying both sides by $\frac{4}{\phi-1}$

$$\frac{4}{\phi-1}\frac{\alpha}{\phi}\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \mathbf{z}^k-\mathbf{z}^{k+1}\rangle=\frac{4}{\phi-1}\frac{\alpha^2}{\phi^2}\|\mathbf{g}^{k-1}-\mathbf{g}^k\|^2+\frac{4}{\phi^2}\alpha^2\langle \mathbf{g}^k, \mathbf{g}^k-\mathbf{g}^{k-1}\rangle. \quad (44)$$

By combining (44) and (43) we deduce

$$-\frac{\phi-1}{\phi}\alpha^2\left\langle \mathbf{g}^k, \frac{2}{\phi}\mathbf{g}^{k-1}-\frac{\phi-1}{\phi}\mathbf{g}^k\right\rangle = \frac{4}{\phi(\phi-1)}\alpha\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \mathbf{z}^k-\mathbf{z}^{k+1}\rangle \\ -\frac{3-\phi}{\phi-1}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2+\left(\frac{3-\phi}{\phi-1}\frac{\alpha^2}{\phi^2}-\frac{4}{\phi-1}\frac{\alpha^2}{\phi^2}\right)\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2.$$

We can deduce (41) by simplifying the above equation. The statement of the lemma is obtained by combining (41) and (40). ■

Proof [of Theorem 7] We deduce from (42)

$$\frac{\phi-1}{\phi^2}\alpha^2\|\mathbf{g}_k\|^2 = \frac{\alpha^2}{\phi^2(\phi-1)}\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2+\frac{2\alpha}{\phi(\phi-1)}\langle \mathbf{g}^{k-1}-\mathbf{g}^k, \mathbf{z}^k-\mathbf{z}^{k+1}\rangle+\frac{1}{\phi-1}\|\mathbf{z}^k-\mathbf{z}^{k+1}\|^2. \tag{45}$$

Adding (45) to (39) gives

$$\|\bar{\mathbf{z}}^{k+1}-\mathbf{z}^*\|^2+\frac{\phi-1}{\phi}\alpha(\frac{\alpha}{\phi}-\rho)\|\mathbf{g}^k\|^2 \le \|\bar{\mathbf{z}}^k-\mathbf{z}^*\|^2+\frac{2}{\phi(\phi-1)}\alpha\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \mathbf{z}^k-\mathbf{z}^{k+1}\rangle \\ -\frac{2-\phi}{\phi-1}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2-\frac{\alpha^2\phi}{\phi^2(\phi-1)}\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2 \tag{46}$$

By Young's inequality we deduce

$$\frac{2}{\phi(\phi-1)}\alpha\langle \mathbf{g}^k-\mathbf{g}^{k-1}, \mathbf{z}^k-\mathbf{z}^{k+1}\rangle \le \frac{\alpha}{\phi(\phi-1)L}\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2+\frac{\alpha L}{\phi(\phi-1)}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2. \tag{47}$$

Combining (47) and (46) yields

$$\|\bar{\mathbf{z}}^{k+1}-\mathbf{z}^*\|^2+\frac{\phi-1}{\phi}\alpha\left(\frac{\alpha}{\phi}-\rho\right)\|\mathbf{g}^k\|^2 \le \|\bar{\mathbf{z}}^k-\mathbf{z}^*\|^2 \\ -\left(\frac{2-\phi}{\phi-1}-\frac{\alpha L}{\phi(\phi-1)}\right)\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2+\left(\frac{\alpha}{L\phi(\phi-1)}-\frac{\alpha^2\phi}{\phi^2(\phi-1)}\right)\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2.$$

Therefore, in order to telescope we require

$$2\phi-\phi^2-\alpha L \ge \alpha L-\alpha^2L^2 \tag{48}$$

and $2\phi-\phi^2 \ge \alpha L$ for the last term to be nonnegative. The condition (48) can be simplified to

$$\frac{2-\phi}{L} \ge \alpha,$$

and the nonnegativity condition becomes redundant. We observe that, if $\rho < \frac{1}{L}$, then we can pick ρ close enough to one such that $\frac{2-\phi}{\phi L} > \rho$. The final bound we obtain by

$$\left(\frac{\alpha}{L\phi(\phi-1)}-\frac{\alpha^2\phi}{\phi^2(\phi-1)}\right)\|\mathbf{g}^k-\mathbf{g}^{k-1}\|^2 \le \left(\frac{\alpha L}{\phi(\phi-1)}-\frac{\alpha^2L^2}{\phi(\phi-1)}\right)\|\mathbf{z}^k-\mathbf{z}^{k-1}\|^2$$

where

$$\frac{\alpha L}{\phi(\phi-1)}-\frac{\alpha^2L^2}{\phi(\phi-1)} = \frac{\phi-2-(\phi-2)^2}{\phi(\phi-1)} = \frac{3\phi-\phi^2-2}{\phi(\phi-1)} \overset{\phi\ge 1}{\le} \frac{2\phi-2}{\phi(\phi-1)} = \frac{2}{\phi}. \tag{49}$$

■

C.2 Adaptive step size

The analysis of (aGRAAL) relies on a similar energy function as before. So with slight abuse of notation we (re-)define for the purpose of this section

$$\mathcal{E}_{k+1}(\mathbf{z}) := \frac{\phi}{\phi-1}\|\bar{\mathbf{z}}^{k+1}-\mathbf{z}\|^2 + \frac{\theta_k}{2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2.$$

Proof [of Theorem 8] The first few steps of our analysis follow the one presented in (Malitsky, 2020) so we do not reproduce them here. We continue from (34) in (Malitsky, 2020), which reads

$$\mathcal{E}_{k+1}(\mathbf{z}) \leq \mathcal{E}_k(\mathbf{z}) + \Big(\theta_k - 1 - \frac{1}{\phi}\Big)\alpha_k^2\|F(\mathbf{z}^k)\|^2 - \theta_k\|\mathbf{z}^k-\bar{\mathbf{z}}^k\|^2 - 2\alpha_k\langle F(\mathbf{z}^k), \mathbf{z}^k-\mathbf{z}\rangle.$$

By choosing $\mathbf{z}=\mathbf{z}^*$ and using the weak Minty property (WM) this reduces to

$$\mathcal{E}_{k+1}(\mathbf{z}^*) \leq \mathcal{E}_k(\mathbf{z}^*) + \alpha_k\left(\Big(\theta_k - 1 - \frac{1}{\phi}\Big)\alpha_k + \rho\right)\|F(\mathbf{z}^k)\|^2 - \theta_k\|\mathbf{z}^k-\bar{\mathbf{z}}^k\|^2. \tag{50}$$

Note that

$$\theta_k\|\mathbf{z}^k-\bar{\mathbf{z}}^k\|^2 = \frac{\theta_k}{\phi^2}\|\mathbf{z}^k-\bar{\mathbf{z}}^{k-1}\|^2 = \frac{\theta_k\alpha_{k-1}^2}{\phi^2}\|F(\mathbf{z}^{k-1})\|^2 = \frac{\alpha_{k-1}\alpha_k}{\phi}\|F(\mathbf{z}^{k-1})\|^2. \tag{51}$$

Plugging (51) into (50) yields

$$\mathcal{E}_{k+1}(\mathbf{z}^*) + \alpha_k\left(\frac{\alpha_{k-1}}{\phi}\|F(\mathbf{z}^{k-1})\|^2 + \left(\alpha_k\Big(1+\frac{1}{\phi}-\theta_k\Big)-\rho\right)\|F(\mathbf{z}^k)\|^2\right) \leq \mathcal{E}_k(\mathbf{z}^*). \tag{52}$$

If for all k

$$\alpha_k\left(\alpha_k\Big(1+\frac{1}{\phi}-\theta_k\Big)-\rho\right) > 0,$$

which can be ensured, by the fact that $\theta_k \leq \gamma\phi$, if

$$\alpha_k\Big(1+\frac{1}{\phi}-\gamma\phi\Big)-\rho > 0, \tag{53}$$

holds, then after telescoping, where we unroll the recursion until $k=1$ as argued in the proof of Lemma 3, to obtain

$$\sum_{i=1}^{k}\alpha_i\left((1+\frac{1}{\phi}-\gamma\phi)\alpha_i-\rho\right)\|F(\mathbf{z}^i)\|^2 + \mathcal{E}_{k+1}(\mathbf{z}^*) \leq \mathcal{E}_1(\mathbf{z}^*).$$

■

Remark 17 *Note that in order to guarantee just* $\liminf_k \|F(\mathbf{z}^k)\| = 0$ *it is sufficient to ask for*

$$\sum_{i=1}^{k}\alpha_i\left((1+\frac{1}{\phi}-\gamma\phi)\alpha_i-\rho\right) \to \infty,$$

meaning that we can allow for arbitrarily many step size to not fulfill the condition $\delta > 0$*, but for the sequence on average.*

In the presence of bounded iterates we were able to relax the conditions of Theorem 8. Let M denote the diameter of the ball containing the iterates.

Proof [of Corollary 10] After telescoping (52) we get

$$\mathcal{E}_{k+1}(\mathbf{z}^*) + \sum_{i=1}^{k} \alpha_i \left(\frac{\alpha_{i+1}}{\phi} + \alpha_i\Big(1 + \frac{1}{\phi} - \theta_i\Big) - \rho \right) \|F(\mathbf{z}^i)\|^2 \leq \mathcal{E}_1(\mathbf{z}^*) + \frac{\alpha_{k+1}\alpha_k}{\phi}\|F(\mathbf{z}^k)\|^2. \tag{54}$$

Let us first observe that the nonnegativity of the factor in front of $\|F(\mathbf{z}^i)\|^2$ can be guaranteed via

$$\frac{\alpha_{k+1}}{\phi} + \alpha_k\Big(1 + \frac{1}{\phi} - \theta_k\Big) \geq \frac{\alpha_{k+1}}{\phi} + \alpha_k\Big(1 + \frac{1}{\phi} - \gamma\phi\Big),$$

since $\theta_k \leq \gamma\phi$. Also, the last term on the right hand side of (54) can be bounded via

$$\alpha_{k+1}\alpha_k\|F(\mathbf{z}^k)\|^2 \leq \frac{\alpha_{k+1}}{\alpha_k}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \leq \gamma\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2.$$

Thus we obtain

$$\sum_{i=1}^{k} \alpha_i \left(\frac{\alpha_{i+1}}{\phi} + \alpha_i\Big(1 + \frac{1}{\phi} - \gamma\phi\Big) - \rho \right) \|F(\mathbf{z}_i)\|^2 \leq \mathcal{E}_1(\mathbf{z}^*) + \frac{\gamma}{\phi}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2,$$

where the last term on the right remains bounded due to the assumed boundedness of the iterates. To deduce the precise constant we observe

$$\frac{\phi}{\phi - 1}\|\bar{\mathbf{z}}^1 - \mathbf{z}\|^2 + \frac{\theta_0}{2}\|\mathbf{z}^1 - \mathbf{z}^0\|^2 + \frac{\gamma}{\phi}\|\mathbf{z}^{k+1} - \bar{\mathbf{z}}^k\|^2 \leq 2M^2 + 2M^2 + 2M^2.$$

Thus,

$$\min_{i=1,\dots,k} \|F(\mathbf{z}_i)\|^2 \leq \frac{6}{\delta \sum_{i=1}^{k} \alpha_i} M^2.$$

■

Appendix D. Details on experiments

D.1 Monotone problems

For the left plot in Fig. 2 we use the Lagrangian formulation of a linearly constrained quadratic program

$$L(x, y) = \frac{1}{2}x^T H x - h^T x - \langle Ax - b, y\rangle,$$

where $x, y, h, b \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}$ with $d = 100$. We use the parametrization proposed in (Ouyang and Xu, 2021) and further studied in (Yoon and Ryu, 2021) which provides a particularly difficult instance. The middle and right plot in Fig. 2 are special instances of bilinear matrix games of the form

$$\min_{x\in\Delta^d} \max_{y\in\Delta^d} x^T A y,$$

where Δ^d denotes the d-dimensional unit simplex $\{x \in \mathbb{R}^d : x \geq 0, \sum_{i=1}^{d} x_i = 1\}$. For our experiments we used $d = 50$.

D.2 Weak minty experiments

D.2.1 Algorithms

For (aGRAAL) ϕ is usually given in the legend, except for Fig. 1 where we used the default $\phi = 1.5$. If γ is not given in the legend we use the theoretical upper bound $1/\phi + 1/\phi^2$.

For CurvatureEG+ we use a backtracking linesearch initialized with $\nu\|JF(\mathbf{z}^k)\|^{-1}$, where we use $\nu = 0.99$ and JF denotes the Jacobian of F. We ignore the extra cost of this initialization but do count the extra gradient evaluations from the backtracking, where in every step the step size is decreased by $\tau = 0.9$.

D.2.2 Polar Game

The unique solution for this problem is in the origin. The Lipschitz constant L and the weak Minty parameter ρ we approximate numerically, via a grid search on the interval $[-1.1, 1.1] \times [-1.1, 1.1]$. In Fig. 3 the value $a = 1/3$ is used whereas in Fig. 5 is given by $a = 3.7$. For $a = 1/3$ we get $L \approx 6.94$ and $\rho \approx 0.09$, whereas for $a = 3$ we compute $L \approx 61.4$ and $\rho \approx 0.72$. In the latter case we observe from the result of the linesearch, see Fig. 5, that these global estimates are quite pessimistic but even locally the necessary condition $\rho < \alpha_k$ is not satisfied for CurvatureEG+, which is why we observe its divergence.

References

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

Axel Böhm. Solving nonconvex-nonconcave min-max problems exhibiting weak minty solutions. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

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. *SIAM Journal on Mathematics of Data Science*, 4(2):750–771, 2022.

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

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 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*, pages 2746–2754. PMLR, 2021.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Eduard Gorbunov, Adrien Taylor, Samuel Horváth, and Gauthier Gidel. Convergence of proximal point and extragradient-based methods beyond monotonicity: the case of negative comonotonicity. *arXiv preprint arXiv:2210.13831*, 2022.

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*, pages 4337–4348. PMLR, 2021.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Galina M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekon. Mat. Metody*, 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, editors, *Advances in Neural Information Processing Systems*, 2021a.

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.

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.

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*, pages 1497–1507. PMLR, 2020.

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

Arkadi Nemirovski. Mini-course on convex programming algorithms. Lecture notes, 2013.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

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

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.

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.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

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

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*, pages 12098–12109. PMLR, 2021.