MOVING ANCHOR EXTRAGRADIENT METHODS FOR SMOOTH STRUCTURED MINIMAX PROBLEMS

JAMES K. ALCALA, YAT TIN CHOW, AND MAHESH SUNKULA

Abstract. This work introduces a moving anchor acceleration technique to extragradient algorithms for smooth structured minimax problems. The moving anchor is introduced as a generalization of the original algorithmic anchoring framework, i.e. the EAG method introduced in [32], in hope of further acceleration. We show that the optimal order of convergence in terms of worst-case complexity on the squared gradient, $O(1/k^2)$, is achieved by our new method (where k is the number of iterations). We have also extended our algorithm to a more general nonconvex-nonconcave class of saddle point problems using the framework of [14], which slightly generalizes [32]. We obtain similar order-optimal complexity results in this extended case. In both problem settings, numerical results illustrate the efficacy of our moving anchor algorithm variants, in particular by attaining the theoretical optimal convergence rate for first order methods, as well as suggesting a better optimized constant in the big O notation which surpasses the traditional fixed anchor methods in many cases. A proximal-point preconditioned version of our algorithms is also introduced and analyzed to match optimal theoretical convergence rates.

1. Introduction

Minimax, min-max, or saddle point problems of the form

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} L(x, y) \tag{1.1}$$

have received considerable attention from optimization researchers and, in particular, machine learning practitioners because of applications including but not limited to Game Theory, Online Learning, GANs [9], [3], adversarial learning [18], and reinforcement learning [7]. Measuring the duality gap $\sup_{y^* \in \mathbb{R}^m} L(x, y^*) - \inf_{x^* \in \mathbb{R}^n} L(x^*, y)$ on averaged (ergodic) iterates or last-iterates of algorithms is one natural way to measure the suboptimality of methods designed to solve (1.1). This is a clear analog to measuring suboptimality for algorithms for minimization problems. On the other hand, such a measurement is not as natural to consider when (1.1) is nonconvex-nonconcave, and as will be discussed, the convergence guarantees for this kind of measure may be limiting.

When problem (1.1) is differentiable, another meaningful measure of suboptimality is the squared gradient norm or Hamiltonian of L, $\text{Ham}_L(x, y) = \|\nabla L(x, y)\|^2$. (Sometimes this includes an extra factor of $\frac{1}{2}$, which is not included in this paper. No physical interpretation of this quantity is used here.) This suboptimality measure retains meaning for nonconvex-nonconcave problems and convergence rates on the squared gradient-norm have only recently attained order-optimal convergence rates in these problem settings. This is especially important, as many machine learning settings involve neural networks which result in problems that are inherently nonconvex-nonconcave - and as our results indicate, there may still be room for numerical improvements.

The EAG (extra-anchored gradient) class of algorithms, first introduced in [32], combines extragradient and the more recently developed anchoring methods in a single framework

James K. Alcala is partially supported by a UCR Dissertation Year Program Award.

Yat Tin Chow is partially supported by a Regents' Faculty Fellowship, University of California, Riverside.

to tackle smooth-structured convex-concave minimax problems. With the primary assumptions being $R-$smoothness and convexity-concavity of (1.1), EAG achieved $O(1/k^2) = \Omega(1/k^2)$ convergence rates on the squared gradient-norm; that is, the algorithm is order-optimal. This achievement has inspired a flurry of research activity in recent years [13], [28], [32]. To show optimality, the authors of [32] adapt arguments from [21], [22] to construct a worse-case analysis for a large class of algorithms that contain EAG.

As anchoring is relatively new compared to extragradient, much of the literature written as a direct consequence of these results emphasizes anchoring and other Halpern adjacent techniques [15], [30], [29]. However, the EAG class is not without limitations. The two sub-variants of EAG, EAG-V with varying step-size and EAG-C with constant step-size, have difficult convergence analyses and are both relegated to the convex-concave class of smooth functions. Addressing some of these issues, the authors of [14] introduced the Fast ExtraGradient Method, or FEG. This method generalizes the results of EAG and EG+ [6] to introduce the order-optimal pairing of the extragradient anchor to the setting of certain nonconvex-nonconcave problems (specifically, negative comonotone) and introduces an analysis dependent on terms that are less difficult to work with. Furthermore, their work improves upon the bounding constant attained in EAG in convex-concave problems while retaining optimal convergence rates for a broader class of problems that are of particular importance to machine learning practitioners, among many others.

In the spirit of these previous works, our contributions are as follows.

(1) We introduce a new technique, the 'moving anchor,' into the algorithmic settings of EAG-V and FEG under minimal assumptions. We demonstrate that in both settings, introducing the moving anchor retains order-optimal $O(1/k^2)$ convergence rates across a range of parameter choices that using the moving anchor gives one access to. One may recover the original fixed-anchor algorithms via parameter tuning, so our algorithms generalize much of the current anchoring literature.

(2) For both the EAG-V moving anchor and the FEG moving anchor, we run a variety of numerical examples by comparing multiple versions of our moving anchor algorithms with their fixed anchor counterparts. These numerical examples demonstrate the efficacy of our algorithm, as in all cases, a moving anchor algorithm variant in each example is the fastest algorithm by a constant or is comparable-to-better for all iterations. In addition, in many cases the fastest moving anchor algorithm appears to have a massive initial oscillation towards the fixed point that the fixed point algorithms seem to lack - this may be beneficial for reaching certain stopping criteria very quickly.

(3) We develop a theoretical version of the moving anchor algorithms (in both the convex-concave EAG-V and nonconvex-nonconcave FEG) with a proximal anchoring step with fruitful implications for future research.

2. Literature Review & Preliminaries

2.1. Halpern iteration and anchoring.

Introduced in 1967 and inspired by Browder's classical fixed point theorem, the Halpern iteration [11] is an algorithm built for approximating fixed point(s) of nonexpanding maps in a Hilbert space. Its convergence has been studied in [16], and it is extensively used in monotone inclusion-type problem settings [5], [30], [2]. A recent paper [29] draws an explicit connection between Halpern-inspired methods and Nesterov's AGM [23], linking two very active strains of acceleration literature.

Directly inspired by Halpern, algorithmic anchoring was recently introduced in the literature [27] and has since been utilized to establish optimal $O(1/k^2)$ convergence rates for smooth-structured convex-concave minimax problems [32]. Since then, these methods have been extended to the nonconvex-nonconcave, negative comonotone problem setting [14] and analogous settings for composite problems in a multi-step framework [15]. Interestingly, this latter framework introduces 'semi'-anchoring, where only one part of the descent-ascent step is anchored, and a unique anchor occurs at each step of the multi-step. To our knowledge, this is the first instance of an anchoring method that goes beyond a single fixed anchor. In [30], the authors develop an anchored Popov's scheme and a splitting version of the EAG developed in [32], with a similar analysis.

2.2. **Extragradient methods.** The extragradient method first appeared in [12] and has since been an important acceleration method extensively studied in the optimization literature [1], [31], [17], especially in the context of generative adversarial networks [9], [3] and adversarial training [18]. A classical result regarding these methods is that if $X \in \mathbb{R}^n, Y \in \mathbb{R}^m$ are compact domains, then for the duality gap $\max_{y* \in Y} L(x, y*) - \min_{x* \in X} L(x*, y)$, the ergodic iterate of extragradient-type methods [19], [24] have an $O(1/k)$ rate, which is order-optimal [25], [20]. Recently, it was shown that the last iterate convergence rate for extragradient also attains $O(1/k)$ convergence [10], with *only* monotonicity and Lipschitz assumptions. This closes the gap between the last-iterate and ergodic-iterate convergence rates for extragradient discussed in [8]. Another recent interesting result was attained in [6], where the authors developed the Extragradient+ method, a variant of extragradient extended to various nonconvex-nonconcave problem settings.

On the other hand, when the problem at hand has certain smoothness properties, the squared gradient norm $\|\nabla L\|^2$ for extragradient-type algorithms recently achieved order-optimal convergence of $O(1/k^2)$ [32], [14], thanks in part to the synthesis with anchoring. This breaks the bound of the SCLI class of algorithms discussed in [8], which contains the unmodified extragradient, because EAG is *not* SCLI, but specifically 2-CLI or in an extended class of 1-CLI algorithms. See Appendix D.2 of [32] for a best-iterate (NOT last iterate, at the time of writing this quantity doesn't seem to be known) convergence analysis of extragradient and Appendix E of [32] and [8] for more details on the relationships between these classes of algorithms. We conclude this discussion by remarking that for smooth problems, the bound on the squared gradient norm is meaningful in nonconvex-nonconcave problem settings, and as demonstrated in this and recent works, has room for numerical improvement.

2.3. **Preliminaries.** A saddle function $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is (non)convex-(non)concave if it is (non)convex in x for any fixed $y \in \mathbb{R}^m$ and (non)concave in y for any fixed $x \in \mathbb{R}^n$. A saddle point $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ is any point such that the inequality $L(\hat{x}, y) \leq L(\hat{x}, \hat{y}) \leq L(x, \hat{y})$ for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Solutions to (1.1) are defined as saddle points.

Throughout this paper, we assume the differentiability of L, and we are especially interested in the so-called *saddle operator* associated to L,

$$G_L(z) = \begin{bmatrix} \nabla_x L(x, y) \\ -\nabla y L(x, y) \end{bmatrix} \tag{2.1}$$

where the L subscript is omitted when the underlying saddle function is known. When our problem is convex-concave, the operator (2.1) is known to be monotone [26], meaning $\langle G_L(z_1) - G_L(z_2), z_1 - z_2 \rangle \geq 0 \ \forall z_1, z_2 \in \mathbb{R}^n \times \mathbb{R}^m$. We assume that this operator G_L is R-Lipschitz, or has certain stronger Lipschitz properties we detail later; this is sometimes referred to as L being R-smooth. With these properties in mind, one may introduce an assumption that generalizes monotonicity: let $\rho \in (-\frac{1}{2R}, +\infty)$. In this paper, we assume that when G_L is *not* monotone, it satisfies

$$\langle G_L(z_1) - G_L(z_2), z_1 - z_2 \rangle \geq \rho \|G_L(z_1) - G_L(z_2)\|^2 \ \forall z_1, z_2 \in \mathbb{R}^n \times \mathbb{R}^m.$$

When $\rho > 0$, this is called co-coercivity; when $\rho = 0$, this recovers monotonicity; when $\rho < 0$, this is called negative comonotonicity. This latter condition on (2.1) allows one to consider certain nonconvex-nonconcave problems L, and is also going to be a central focus of this work. Note, however, that these assumptions need not cover all smooth nonconvex-nonconcave problems of interest. Figure 1, Table 1, and Example 1 of [14] illustrate broader problem classes than negative comonotonicity that retain smoothness while being nonconvex-nonconcave.

Finally we state that although $\nabla L \neq G_L$, we have $\|\nabla L\| = \|G_L\|$, so we may use these expressions interchangeably.

3. Original Algorithm, EAG-V

The Extragradient Anchored Algorithm, or EAG with varying step size (EAG-V) has a simple statement and a relatively simple proof of convergence:

$$\begin{aligned}
z^{k+1/2} &= z^k + \beta_k(z^0 - z^k) - \alpha_k G(z^k) \\
z^{k+1} &= z^k + \beta_k(z^0 - z^k) - \alpha_k G(z^{k+1/2}) \\
\alpha_{k+1} &= \frac{\alpha_k}{1 - \alpha_k^2 R^2}\left(1 - \frac{(k+2)^2}{(k+1)(k+3)}\alpha_k^2 R^2\right) \\
&= \alpha_k \left(1 - \frac{1}{(k+1)(k+3)} \frac{\alpha_k^2 R^2}{1 - \alpha_k^2 R^2}\right)
\end{aligned}$$

with $\alpha_0 \in (0, 1/R)$, and R a predetermined constant. Here, G is the so-called saddle operator, $G := (\nabla_x L, -\nabla_y L)$ and L is a convex-concave saddle function in a minimax optimization problem. It is a nontrivial fact that G is monotone [32]. The structure of the α_k's and β_k's are detailed below alongside auxiliary sequences A_k and B_k. We state the convergence of this algorithm as a theorem and relay the details of its convergence via a specific Lyapunov functional as a lemma. For more details, including a version of EAG with a non-varying step size, see [32].

Theorem 3.1 (EAG-V convergence rate [32])**.** *Assume $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an R-smooth convex-concave function with a saddle point z^*. Assume further that $\alpha_0 \in (0, \frac{3}{4R})$ and define $\alpha_\infty = \lim_{k\to\infty} \alpha_k$. Then EAG-V converges, with rate*

$$||\nabla G(z^k)||^2 \leq \frac{4(1 + \alpha_0 \alpha_\infty R^2)}{\alpha_\infty^2} \frac{||z^0 - z^*||^2}{(k+1)(k+2)}$$

where $G = (\nabla L|_{x\in\mathbb{R}^n}, -\nabla L|_{-y\in-\mathbb{R}^m})$.

Since z^* is the saddle point, this theorem demonstrates $O(1/k^2)$ convergence of the algorithm. To derive this order of convergence, the following lemma is necessary.

Lemma 3.2 (EAG Lyapunov Functional [32])**.** *Let $\{\beta_k\}_{k\geq 0} \subseteq (0, 1)$ and $\alpha_0 \in (0, \frac{1}{R})$ be given. Consider the following sequences defined by the given recurrence relations for $k \geq 0$:*

$$\begin{aligned}
A_k &= \frac{\alpha_k}{2\beta_k} B_k \\
B_{k+1} &= \frac{B_k}{1 - \beta_k} \\
\alpha_{k+1} &= \frac{\alpha_k \beta_{k+1}(1 - \alpha_k^2 R^2 - \beta_k^2)}{\beta_k(1 - \beta_k)(1 - \alpha_k^2 R^2)} \qquad (3.1)
\end{aligned}$$

where $B_0 = 1$. Assume that $\alpha_k \in (0, \frac{1}{R})$ holds for all $k \geq 0$, and that L is $R-$smooth and convex-concave. Then the sequence $\{V_k\}_{k\geq}$ defined as

$$V_k := A_k\|G(z^k)\|^2 + B_k\langle G(z^k), z^k - z^0\rangle \tag{3.2}$$

is non-increasing.

Within (3.2), choosing $\beta_k = \frac{1}{k+2}$ yields $B_k = k+1$, $A_k = \frac{\alpha_k(k+2)(k+1)}{2}$, and the construction of α_{k+1} in (3.1).

4. EAG-V with moving anchor

In this section, we construct and analyze a new version of the EAG-V algorithm. Here, the anchoring point moves at each time step. We call this the moving anchor algorithm; it utilizes a similar extragradient step. Further down, we demonstrate comparable rates of convergence to the original EAG algorithm with varying step-size.

For the $k-th$ iterate of $z^0 \in \mathbb{R}^n \times \mathbb{R}^m$, the EAG-V with moving anchor is defined as

$$z^0 = \bar{z}^0$$

$$z^{k+1/2} = z^k + \frac{1}{k+2}(\bar{z}^k - z^k) - \alpha_k G(z^k) \tag{4.1}$$

$$z^{k+1} = z^k + \frac{1}{k+2}(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2}) \tag{4.2}$$

$$\bar{z}^{k+1} = \bar{z}^k + \gamma_{k+1}G(z^{k+1}) \tag{4.3}$$

The major structural difference here is the introduction of the regularly-updating $\bar{z}^k$, analogous to the role of z^0 in the EAG-V detailed in the previous section. (4.3) is the regular update for this anchor; it depends on the algorithm update (4.2) rather than exclusively on itself. All previously defined terms are the same as in the fixed anchor algorithm, now with

$$c_{k+1} \leq \frac{c_k}{1+\delta_k}, \tag{4.4}$$

$$\gamma_{k+1} \leq \frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}. \tag{4.5}$$

We choose δ_k so that $\sum_{k=0}^{\infty}\log(1+\delta_k) < \infty$. The c_k terms are part of the definition of the Lyapunov functional we use in our analysis; these come in handy when we use γ_k to absolve terms. Let $c_\infty := \lim_{k\to\infty} c_k = c_0\prod_{k=0}^{\infty}\frac{1}{1+\delta_k}$. As a general rule, one wishes to choose c_0 so that c_∞ satisfies some specified convergence constraint; these constraints will appear throughout the major convergence theorems in this section and the next section. While the choice of c_0 is therefore limited to according to certain problem/algorithm constraints, in general there seems to be much freedom in choosing c_0 and the sequence $\{\delta_k\}$. For the rest of this article, we take (4.4) and (4.5) to be given with equal signs instead of inequalities. Before we proceed with the analysis, we emphasize that the original EAG-V algorithm may be recovered simply by setting $\gamma_{k+1} := 0$ for all k.

Now, we give the definition of the Lyapunov functional and show that it is nonincreasing:

Lemma 4.1. *The Lyapunov functional*

$$V_k := A_k\|G(z^k)\|^2 + B_k\langle G(z^k), z^k - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2,$$

where all constants have been previously defined, is nonincreasing.

Proof. First we reorganize some of the algorithm statements and label them for use later.

$$z^k - z^{k+1} = \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}) \tag{4.6}$$
$$z^{k+1/2} - z^{k+1} = \alpha_k(G(z^{k+1/2}) - G(z^k)) \tag{4.7}$$
$$\bar{z}^k - z^{k+1} = (1-\beta_k)(\bar{z}^k - z^k) + \alpha_k G(z^{k+1/2}) \tag{4.8}$$
$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1} G(z^{k+1}) \tag{4.9}$$

(4.6) comes from rearranging (4.2), (4.7) comes from taking the difference between (4.1) and (4.2), (4.8) is $\bar{z}^k$ minus (4.2), and (4.9) is (4.3) rearranged. The overall goal of this proof is to show that the difference $V_k - V_{k+1}$ is nonnegative.

$$\begin{aligned}
&V_k - V_{k+1}\\
&\geq A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 \underbrace{+B_k\langle z^k - \bar{z}^k, G(z^k)\rangle}_{\text{I}}\\
&\quad \underbrace{-B_{k+1}\langle z^{k+1} - \bar{z}^{k+1}, G(z^{k+1})\rangle}_{\text{II}} + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2\\
&\quad \underbrace{-\frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^k) - G(z^{k+1})\rangle}_{\text{III}}
\end{aligned}$$

Notice that the last term above, III, is not part of the definition of V_k nor V_{k+1}. It has been introduced to aid in the proof and is nonnegative by the monotonicity of G. We would like to absolve any terms containing the $\bar{z}^k, \bar{z}^{k+1}$ terms. To accomplish this, our next goal is to focus on turning the labeled parts (I, II, III) of the above line into

$$\underbrace{\alpha_k B_{k+1}\langle G(z^{k+1/2}, G(z^{k+1})\rangle + \frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2 - \frac{\alpha_k B_k}{\beta k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle}_{\text{IV}}.$$

We now detail this process. The term I does not change. For II, on the other hand, we have

$$\begin{aligned}
&\underbrace{-B_{k+1}\langle z^{k+1} - \bar{z}^{k+1}, G(z^{k+1})\rangle}_{\text{II}}\\
&= B_{k+1}\langle \bar{z}^k - z^{k+1}, G(z^{k+1})\rangle - B_{k+1}\langle \bar{z}^k - \bar{z}^{k+1}, G(z^{k+1})\rangle && (4.10)\\
&= B_{k+1}\langle (1-\beta_k)(\bar{z}^k - z^k) + \alpha_k G(z^{k+1/2}), G(z^{k+1})\rangle - B_{k+1}\langle -\gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle && (4.11)
\end{aligned}$$

where the first equality comes from recognizing $z^{k+1} - \bar{z}^{k+1} = z^{k+1} - \bar{z}^k + \bar{z}^k - \bar{z}^{k+1}$ and the second comes from substituting in equality (4.8) and (4.9). For III,

$$\begin{aligned}
&\underbrace{-\frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^k) - G(z^{k+1})\rangle}_{\text{III}}\\
&= -\frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^k)\rangle + \frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^{k+1})\rangle && (4.12)\\
&= -\frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^k)\rangle + \frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^{k+1})\rangle,
\end{aligned}$$

where the last equality is a result of substituting in (4.6) in each of the first arguments of the two terms in (4.12). Now, we can begin simplify everything we've done to obtain IV.

$$\underbrace{\langle z^k - \bar{z}^k, G(z^k)\rangle}_{\text{I}} \tag{4.13}$$

$$\underbrace{\langle (1-\beta_k)(z^k - \bar{z}^k) - \alpha_k G(z^{k+1/2}) - \gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle}_{\text{II}} \tag{4.14}$$

$$\underbrace{-\frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^k)\rangle}_{\text{III}} \tag{4.15}$$

$$\underbrace{+\frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^{k+1})\rangle}_{\text{III}} \tag{4.16}$$

From here, we'll use two facts. First, $B_{k+1} = \frac{B_k}{1-\beta_k}$. This allows us to combine and cancel the very first component of (4.14) with the $\beta_k(z^k - \bar{z}^k)$ component of (4.16). Additionally, (4.13) cancels with the $\beta_k(z^k - \bar{z}^k)$ component of (4.15). This leaves us with

$$
\begin{aligned}
&= \underbrace{\alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle + B_{k+1}\langle \gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle}_{\text{II}} \\
&\quad \underbrace{-\frac{B_k\alpha_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k)\rangle + \frac{B_k\alpha_k}{\beta_k}\langle G(z^{k+1/2}), G(z^{k+1})\rangle}_{\text{III}} \\
&= \underbrace{\alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle + \frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2}_{\text{IV}} \\
&\quad \underbrace{-\frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle}_{\text{IV}},
\end{aligned}
$$

where the last equality is a result of applying the anchor update to get the norm squared term, and combining the latter two terms while leaving $G(z^{k+1/2})$ fixed.

Thus, we've shown

$$
\begin{aligned}
& A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 \\
& + B_k\langle z^k - \bar{z}^k, G(z^k)\rangle - B_{k+1}\langle z^{k+1} - \bar{z}^{k+1}, G(z^{k+1})\rangle \\
& + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 - \frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^k) - G(z^{k+1})\rangle \\
&= A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle && (4.17) \\
& - \frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle && (4.18) \\
& + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 + \frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2 && (4.19)
\end{aligned}
$$

Now, we continue on with our goal of absolving terms. From Cauchy, we have that

$$||z^* - \bar{z}^{k+1}\|^2 \le (1+\delta_k)\|z^* - \bar{z}^k\|^2 + (1 + \frac{1}{\delta_k})\|\bar{z}^k - \bar{z}^{k+1}\|^2 \tag{4.20}$$

and from the algorithm definition,

$$c_{k+1} = \frac{c_k}{1+\delta_k}, \quad \gamma_{k+1} = \frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}. \tag{4.21}$$

We apply (4.20) to (4.19) to obtain

$$\begin{aligned}
&\geq A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle \\
&-\frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle + c_k\|z^* - \bar{z}^k\|^2 \\
&-c_{k+1}\big((1+\delta_k)\|z^* - \bar{z}^k\|^2 + (1+\frac{1}{\delta_k})\|\bar{z}^k - \bar{z}^{k+1}\|^2\big) + \frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2
\end{aligned}$$

and now we apply (4.21):

$$\begin{aligned}
&\geq A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle \\
&-\frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle + c_k\|z^* - \bar{z}^k\|^2 \\
&-c_k\|z^* - \bar{z}^k\|^2 - \frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2 + \frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2 \\
&= A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle \\
&-\frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle + 0.
\end{aligned}$$

At this point, showing that the remaining terms are nonnegative is nontrivial, but directly follows the arguments made in the proof of Lemma 2 in [32]. Specifically, following (29) onwards in [32], one will find that

$$\begin{aligned}
&A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle \\
&-\frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle \\
&\geq 0,
\end{aligned}$$

which completes the proof. □

Now we have the primary result of this section.

Theorem 4.2. *The EAG-V algorithm with moving anchor, described above, together with the Lyapunov functional described in Lemma 4.1, has convergence rate*

$$\|G(z^k)\|^2 \leq \frac{4(\alpha_0 R^2 + c_0)\|z^0 - z^*\|^2}{\alpha_\infty (k+1)(k+2)}$$

as long as we assume $c_\infty \alpha_\infty \geq 1$.

Proof. For the most part, this argument parallels the analogous argument found in [32]. We use the Lyapunov functional to isolate and bound $\|G(z^k)\|^2$.

$$\begin{aligned}
V_k \leq V_0 &= \alpha_0\|G(z^0)\|^2 + c_0\|z^0 - z^*\|^2 \\
&\leq (\alpha_0 R^2 + c_0)\|z^0 - z^*\|^2
\end{aligned} \tag{4.22}$$

by $R-$smoothness. On the other hand,

$$\begin{aligned} V_k &= A_k\|G(z^k)\|^2 + B_k\langle G(z^k), z^k - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2 \\ &\geq A_k\|G(z^k)\|^2 + B_k\langle G(z^k), z^* - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2 \\ &\geq \frac{A_k}{2}\|G(z^k)\|^2 + (c_k - \frac{B_k^2}{2A_k})\|z^* - \bar{z}^k\|^2 \\ &= \frac{\alpha_k(k+1)(k+2)}{4}\|G(z^k)\|^2 + (c_k - \frac{k+1}{\alpha_k(k+2)})\|z^* - \bar{z}^k\|^2 \\ &\geq \frac{\alpha_\infty}{4}(k+1)(k+2)\|G(z^k)\|^2 + (c_\infty - \frac{1}{\alpha_\infty})\|z^* - \bar{z}^k\|^2 \\ &\geq \frac{\alpha_\infty}{4}(k+1)(k+2)\|G(z^k)\|^2 \end{aligned}$$

As long as $c_\infty \geq \frac{1}{\alpha_\infty}$, the second to last line above is positive, and we may focus on the inequality given to us by the last line above:

$$\frac{\alpha_\infty}{4}(k+1)(k+2)\|G(z^k)\|^2 \leq (\alpha_0 R^2 + c_0)\|z^0 - z^*\|^2.$$

Dividing both sides by the constant $\frac{\alpha_\infty}{4}(k+1)(k+2)$ gives the desired result. □

4.1. **Proof of convergence for $-\gamma_k$.** We next show that, for a slightly restricted choice of γ_k, our proof works for $-\gamma_k$ in place of γ_k. This is of interest as numerical results indicate that certain problem settings favor $-\gamma_k$ in terms of convergence speed by a constant, while $+\gamma_k$ seems to be favored in other settings.

Lemma 4.3. *In the setting of Lemma 4.1, replace γ_k with $-\gamma_k$ in the definition of the EAG-V algorithm with moving anchor, and suppose $\gamma_{k+1} = \min \frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}, \frac{e_{k+1}}{2B_{k+1}\|G(z^{k+1})\|^2}$, where $\sum e_k < \infty$. Then our Lyapunov functional is nonincreasing, and we attain the same order of convergence in our algorithm.*

Proof. First, note that the anchor update (4.3) has been modified to become

$$-\gamma_{k+1} \geq -\frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}, \tag{4.23}$$

resulting in the following modification to (4.9):

$$\bar{z}^k - \bar{z}^{k+1} = \gamma_{k+1}G(z^{k+1}). \tag{4.24}$$

We see the first adjustment in the previous lemma in the transition from line (4.10) to (4.11); note that we focus only on the terms dependent on (4.24):

$$\begin{aligned} &-B_{k+1}\langle \bar{z}^k - \bar{z}^{k+1}, G(z^{k+1})\rangle \\ &= -B_{k+1}\langle \gamma_{k+1}G(z^{k+1}), G(z^{k+1})\rangle \\ &= -B_{k+1}\langle (2\gamma_{k+1} - \gamma_{k+1})G(z^{k+1}), G(z^{k+1})\rangle \\ &= -B_{k+1}\langle 2\gamma_{k+1}G(z^{k+1}), G(z^{k+1})\rangle + B_{k+1}\langle \gamma_{k+1}G(z^{k+1}), G(z^{k+1})\rangle. \end{aligned} \tag{4.25}$$

The latter term in line (4.25) will go on and cancel in a quadratic form as in the proof of the original lemma. Continuing, one will be left over with the term $-B_{k+1}\langle 2\gamma_{k+1}G(z^{k+1}), G(z^{k+1})\rangle$.

At this point, if we proceed as in Lemma 4.1, we end up with the inequality

$$V_k - V_{k+1} \geq -2\gamma_{k+1}B_{k+1}\|G(z^{k+1})\|^2$$

or, after rearranging,

$$V_k - V_{k+1} + 2\gamma_{k+1}B_{k+1}\|G(z^{k+1})\|^2 \geq 0.$$

By construction, the left-hand side of the inequality should remain nonnegative. Now, because

$$\gamma_{k+1} \leq \frac{e_{k+1}}{2B_{k+1}\|G(z^{k+1})\|^2},$$

when we proceed as in the proof of Theorem 4.2 to show convergence, getting to the line (4.22), we get the inequality

$$\begin{aligned} V_k \leq & V_0 + \sum_{j=1}^{k-1} 2\gamma_j B_j \|G(z^j)\|^2 \\ \leq & V_0 + \sum_{j=1}^{k-1} e_j \\ \leq & V_0 + \sum_{j=1}^{\infty} e_j \\ = & CV_0, \end{aligned}$$

where C is a constant. This completes the proof that our algorithm has both a nonincreasing Lyapunov functional and the $O(1/k^2)$ convergence under the assumption of a (slightly restricted) negative γ_k term. □

It is worth noting that z^{k+1} is computed before γ_{k+1} within the algorithm, so the restriction in Lemma 4.3 and others like it may not be too restrictive to work with. Our toy numerical tests allowed us to simply put a negative sign in front of the γ_k terms to attain convergence matching the optimal rate, and which is in some cases markedly faster. Unfortunately, these results do not give much of an indication as to how exactly the tuning of γ_k benefits numerical convergence rates. We leave the theoretical exploration of this phenomena to future work.

5. Moving anchor in nonconvex/nonconcave minmax problems

In [14], the methods in [32] are expanded to a broader class of smooth structured nonconvex-nonconcave minimax problems at the same accelerated $O(1/k^2)$ convergence rate. This new algorithm is called the FEG, or Fast ExtraGradient method. We bring the idea of the moving anchor to this more general setting, and show that a moving anchor with more or less the same conditions in the convex-concave setting is also a feasible approach in this class of problems. Below we give the explicit definition of this FEG modified via a moving anchor, and state its convergence results via a nonincreasing Lyapunov functional and a theorem bounding the squared gradient norm.

The FEG with moving anchor, following [14], is given as

$$\begin{aligned} z^{k+1/2} &= z^k + \beta_k(\bar{z}^k - z^k) - (1-\beta_k)(\alpha_k + 2\rho_k)G(z^k) \\ z^{k+1} &= z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2}) - (1-\beta_k)2\rho_k G(z_k) \\ \bar{z}^{k+1} &= \bar{z}^k + \gamma_{k+1} G(z^{k+1}) \\ c_{k+1} &= \frac{c_k}{1+\delta_k} \\ \gamma_{k+1} &= \frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})} \end{aligned}$$

where $\{\delta_k\}$ is chosen so that $\sum_{i=0}^{\infty} \log(1+\delta_i) < \infty$, with $\{\gamma_k\}$, $\{c_k\}$, and c_∞ chosen in the same method given in the EAG-V with moving anchor, and, as before, $\bar{z}^0 = z^0$. Before we state the results, two remarks are needed:

Remark 5.1. For some $\rho \in \big(-\frac{1}{2R}, \infty\big), \langle G(z) - G(z'), z - z' \rangle \geq \rho \|G(z) - G(z')\|^2 \; \forall z, z' \in \mathbb{R}^m \times \mathbb{R}^n$. (Note z, z' are vectors, not matrices.) This is known as $\rho-$comonotonicity, and has three sub-conditions. For $\rho > 0$, we have cocoercivity; for $\rho = 0$, we have monotonicity; and with $\rho < 0$ we have (negative) comonotonicity. This condition will hold whenever any FEG variant is discussed throughout this work.

Remark 5.2. As in the EAG with moving anchor, one may recover the original fixed anchor FEG by setting $\gamma_k = 0$ for all k. This allows us to state our algorithm while also offering an easy reference point for the original fixed anchor version.

Lemma 5.3. *Suppose that the sequences $\{\alpha_k\}_{k\geq 0}$, $\{\beta_k\}_{k\geq 0}$, and $\{R_k\}_{k\geq 0} \subset (0,\infty)$, and $\{\rho_k\}_{k\geq 0} \subset \mathbb{R}$ satisfy $\alpha_0 \in (0,\infty), \alpha_k \in (0, \frac{1}{R_k}), \beta_0 = 1, \{\beta_k\}_{k\geq 1} \subseteq (0,1)$ for all k. Additionally, assume that the following bound, Lipschitz conditions, and comonotonicity conditions respectively hold for all $k \geq 0$:*

$$\frac{(1-\beta_{k+1})}{2\beta_{k+1}}(\alpha_{k+1} + 2\rho_{k+1}) - \rho_k \leq \frac{1}{2\beta_k}(\alpha_k + 2\rho_k) - \rho_k$$
$$\|G(z^1) - G(z^0)\| \leq R_0 \|z^1 - z^0\|$$
$$\|G(z^{k+1}) - G(z^{k+1/2})\| \leq R_k \|z^{k+1} - z^{k+1/2}\|$$
$$\langle G(z^{k+1}) - G(z^k), z^{k+1} - z^k \rangle \geq \rho_k \|G(z^{k+1}) - G(z^k)\|^2.$$

If also $A_0 = \dfrac{\alpha_0(L_0^2\alpha_0^2 - 1)}{2}, B_0 = 0, B_1 = 1$, and

$$A_k = \frac{B_k(1-\beta_k)}{2\beta_k}(\alpha_k + 2\rho_k) - B_k\rho_k, \; B_{k+1} = \frac{B_k}{1-\beta_k},$$

then the Lyapunov functional

$$V_k := A_k \|G(z^k)\|^2 - B_k \langle G(z^k), \bar{z}^k - z^k \rangle + c_k \|z^* - \bar{z}^k\|^2,$$

where z^* is a saddle point, is nonincreasing.

Proof. This proof proceeds similarly to that of the convex-concave, monotone case in the previous section. First, we write out some relations which will be used shortly:

$$z^{k+1} - z^k = \frac{\beta_k}{1-\beta_k}(\bar{z}^k - z^{k+1}) - \frac{\alpha_k}{1-\beta_k} G(z^{k+1/2}) - 2\rho_k G(z^k) \tag{5.1}$$
$$z^{k+1} - z^k = \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2}) - 2\rho_k(1-\beta_k)G(z^k) \tag{5.2}$$
$$z^{k+1} - z^{k+1/2} = \alpha_k((1-\beta_k)G(z^k) - G(z^{k+1/2})) \tag{5.3}$$
$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1} G(z^{k+1}) \tag{5.4}$$

As in the proof in the convex-concave case of EAG-V with moving anchor, we introduce a term to the difference of two arbitrary consecutive functionals in our sequence:

$$
\begin{aligned}
&V_k - V_{k+1} \\
\geq & A_k \|G(z^k)\|^2 - B_k \langle G(z^k), \bar{z}^k - z^k \rangle - A_{k+1} \|G(z^{k+1})\|^2 + B_{k+1} \langle G(z^{k+1}), \bar{z}^{k+1} - z^{k+1} \rangle \\
&+ c_k \|z^* - \bar{z}^k\|^2 - c_{k+1} \|z^* - \bar{z}^{k+1}\|^2 \\
&- \frac{B_k}{\beta_k} \big(\langle G(z^{k+1}) - G(z^k), z^{k+1} - z^k \rangle - \rho_k \|G(z^{k+1}) - G(z^k)\|^2 \big) \\
= & A_k \|G(z^k)\|^2 - B_k \langle G(z^k), \bar{z}^k - z^k \rangle - A_{k+1} \|G(z^{k+1})\|^2 + B_{k+1} \langle G(z^{k+1}), \bar{z}^{k+1} - z^{k+1} \rangle \qquad (5.5) \\
&+ c_k \|z^* - \bar{z}^k\|^2 - c_{k+1} \|z^* - \bar{z}^{k+1}\|^2 \\
&- \frac{B_k}{\beta_k} \langle G(z^{k+1}), z^{k+1} - z^k \rangle + \frac{B_k}{\beta_k} \langle G(z^k), z^{k+1} - z^k \rangle + \frac{B_k \rho_k}{\beta_k} \|G(z^{k+1}) - G(z^k)\|^2
\end{aligned}
$$

From here, we first simplify the introduced term further and then substitute (5.1) into the inner product which has a B_k out front, and then substitute (5.2) into the inner product with a B_{k+1} out front; each of these is in line (5.5). After some computation, this leads to

$$
\begin{aligned}
&V_k - V_{k+1} \\
\geq & \big(A_k - \frac{2B_k \rho_k (1-\beta_k)}{\beta_k}\big) \|G(z^k)\|^2 - A_{k+1} \|G(z^{k+1})\|^2 + \frac{\alpha_k B_k}{\beta_k (1-\beta_k)} \langle G(z^{k+1}), G(z^{k+1/2}) \rangle \\
&+ \frac{2\rho_k B_k}{\beta_k} \langle G(z^{k+1}), G(z^k) \rangle - \frac{\alpha_k B_k}{\beta_k} \langle G(z^k), G(z^{k+1/2}) \rangle + B_{k+1} \langle G(z^k), \bar{z}^{k+1} - \bar{z}^k \rangle \\
&+ \frac{B_k \rho_k}{\beta_k} \|G(z^{k+1}) - G(z^k)\|^2 + c_k \|z^* - \bar{z}^k\|^2 - c_{k+1} \|z^* - \bar{z}^{k+1}\|^2 \\
= & \big(A_k - \frac{B_k \rho_k (1-2\beta_k)}{\beta_k}\big) \|G(z^k)\|^2 - \big(A_k - \frac{B_k \rho_k}{\beta_k}\big) \|G(z^{k+1})\|^2 + \frac{\alpha_k B_k}{\beta_k (1-\beta_k)} \langle G(z^{k+1}), G(z^{k+1/2}) \rangle \qquad (5.6) \\
&- \frac{\alpha_k B_k}{\beta_k} \langle G(z^k), G(z^{k+1/2}) \rangle + B_{k+1} \langle G(z^{k+1}), \bar{z}^{k+1} - \bar{z}^k \rangle + c_k \|z^* - \bar{z}^k\|^2 - c_{k+1} \|z^* - \bar{z}^{k+1}\|^2. \qquad (5.7)
\end{aligned}
$$

Next, let's focus on the last three terms in (5.7): $B_{k+1} \langle G(z^{k+1}), \bar{z}^{k+1} - \bar{z}^k \rangle + c_k \|z^* - \bar{z}^k\|^2 - c_{k+1} \|z^* - \bar{z}^{k+1}\|^2$. By Cauchy-Schwartz,

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

Second, by construction

$$
B_{k+1} \langle G(z^{k+1}), \bar{z}^k - \bar{z}^{k+1} \rangle = \frac{B_{k+1}}{\gamma_{k+1}} \|\bar{z}^k - \bar{z}^{k+1}\|^2
$$

and

$$
c_{k+1} \leq \frac{c_k}{1+\delta_k}, \; \gamma_{k+1} \leq \frac{B_{k+1}}{c_{k+1}(1 + \frac{1}{\delta_k})}.
$$

Applying these facts to the three terms we're considering, we get that

$$
\begin{aligned}
&B_{k+1}\langle G(z^{k+1}), \bar{z}^{k+1} - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 \\
\geq &\frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^{k+1} - \bar{z}^k\|^2 + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\big((1+\delta_k)\|z^* - \bar{z}^k\|^2 + (1+\frac{1}{\delta_k})\|\bar{z}^k - \bar{z}^{k+1}\|^2\big) \\
\geq &\frac{B_{k+1}}{B_{k+1}}c_{k+1}(1+\frac{1}{\delta_k})\|\bar{z}^k - \bar{z}^{k+1}\|^2 + c_k\|z^* - \bar{z}^k\|^2 \\
&- c_{k+1}(1+\delta_k)\|z^* - \bar{z}^k\|^2 - c_{k+1}(1+\frac{1}{\delta_k})\|\bar{z}^k - \bar{z}^{k+1}\|^2 \\
\geq &c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}(1+\delta_k)\|z^* - \bar{z}^k\|^2 \geq c_k\|z^* - \bar{z}^k\|^2 - c_k\|z^* - \bar{z}^k\|^2 \geq 0.
\end{aligned}
$$

While this takes care of the latter three terms in lines (5.6) to (5.7), that everything else is nonnegative is a nontrivial argument. However, it directly follows the proof of Lemma 7.1 in [14], so as before we refer to their proof, and then our Lyapunov functional is also nonincreasing. □

Theorem 5.4 ($O(1/k^2)$ convergence rate for FEG with moving anchor)**.** *For the $R-$Lipschitz continuous and $\rho-$comonotone operator G where $\rho > -\frac{1}{2R}$, $z^* \in Z_*(G), Z_*(G) := \{z^* \in \mathbb{R}^d : G(z^*) = 0\}$, and $c_\infty - \frac{1}{\frac{1}{R}+2\rho} \geq 0$, the sequence $\{z^k\}_{k\geq 0}$ generated by FEG with moving anchor satisfies*

$$
\|G(z^k)\|^2 \leq \frac{4c_0\|z^0 - z^*\|^2}{k^2(\frac{1}{R}+2\rho)}
$$

for all $k \geq 1$.

Proof. Under the same assumptions as Lemma 5.3, we take $\alpha_k = 1/R, \beta_k = \frac{1}{k+1}, R_k = R$, which satisfy the conditions in the statement for all k greater than or equal to 0. These give us $B_k = k, A_k = \frac{k^2}{2}(\frac{1}{R}+2\rho) - k\rho$.

From here,

$$
c_0\|z^* - z^0\|^2 = V_0 \geq V_k = \left(\frac{k^2}{2}(\frac{1}{R}+2\rho) - k\rho\right)\|G(z^k)\|^2 - k\langle G(z^k), \bar{z}^k - z^k\rangle + c_k\|z^* - \bar{z}^k\|^2,
$$

so then

$$
\begin{aligned}
&\frac{k^2}{2}(\frac{1}{L}+2\rho)\|G(z^k)\|^2 + c_k\|z^* - \bar{z}^k\|^2 \\
\leq &k\langle G(z^k), \bar{z}^k - z^k\rangle + k\rho\|G(z^k)\|^2 + c_0\|z^* - z_0\|^2 \\
\leq &k\langle G(z^k), \bar{z}^k - z^*\rangle + c_0\|z^* - z_0\|^2 \text{ (by comonotonicity condition)} \\
\leq &k\|G(z^k)\|\|\bar{z}^k - z^*\| + c_0\|z^* - z_0\|^2 \\
\leq &\frac{k^2}{2\delta}\|G(z^k)\|^2 + \frac{\delta}{2}\|\bar{z}^k - z^*\|^2 + c_0\|z^* - z_0\|^2.
\end{aligned}
$$

From here, define $\frac{1}{\delta} = \frac{1}{2R} + \rho$. Then we have that

$$
\frac{k^2}{2}\left(\frac{1}{R} + 2\rho - \frac{1}{2R} - \rho\right)\|G(z^k)\|^2 + \left(c_\infty - \frac{1}{\frac{1}{R}+2\rho}\right)\|\bar{z}^k - z^*\|^2 \leq c_0\|z^* - z_0\|^2,
$$

and as long as the constant $c_\infty - \frac{1}{\frac{1}{R}+2\rho} \geq 0$, we obtain the desired result by dividing both sides of the inequality

$$
\frac{k^2}{2}\left(\frac{1}{2R} + \rho\right)\|G(z^k)\|^2 \leq c_0\|z^* - z_0\|^2
$$

by $\frac{k^2}{2}\left(\frac{1}{2R}+\rho\right)$. ☐

See [14]'s proof of Theorem 4.1 for the analogous result with a fixed anchor. Next, we show that having $-\gamma_{k+1}$ in place of γ_{k+1} may also, with some additional assumptions, provide a convergent algorithm.

Lemma 5.5. *In the setting of Lemma 5.3, replace γ_k with $-\gamma_k$ in the definition of the FEG algorithm with moving anchor, and suppose $\gamma_{k+1} = \min \frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}, \frac{e_{k+1}}{2B_{k+1}\|G(z^{k+1})\|^2}$, where $\sum e_k < \infty$. Then the Lyapunov functional described in Lemma 5.3 is nonincreasing, and we attain the same order of convergence for the FEG with moving anchor and $-\gamma_k$.*

Proof. The proof proceeds in exactly the same manner as that in Lemma 4.3. ☐

As in the EAG-V with moving anchor case, we suspect this restriction is not too major a restriction based off of numerical results, and that there is a 'better' way to show that the $-\gamma_k$ version of our algorithm converges.

6. Introducing a proximal term

6.1. **Modified EAG-V with moving anchor.** Throughout these notes, each of the algorithms thus far developed are explicit in nature. In this section we introduce and develop a version of these moving anchor algorithms that features a proximal term, and discuss this as a potential future avenue of exploration. We begin with developing the proximal version of the EAG-V with moving anchor.

Definition 6.1 (Modified EAG-V with moving anchor)**.** In the setting of EAG-V with moving anchor, consider equation (4.9) from the proof of Lemma 4.1:

$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1}G(z^{k+1})$$

and now let us consider the same equation with an additional term introduced:

$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1}G(z^{k+1}) - t_k(H(\bar{z}^k) - H(\bar{z}^{k+1})), \tag{6.1}$$

where H is a monotone operator and t_k is nonnegative. This only modifies the anchor update within the algorithm itself, and it does so in the following way:

$$\bar{z}^{k+1} = (I + t_k H)^{-1}(\bar{z}^k + \gamma_{k+1}G(z^{k+1}) + t_k H(\bar{z}^k)). \tag{6.2}$$

This is the modified EAG-V with moving anchor.

Lemma 6.2. *Under the same conditions as Lemma 4.1 and with H any monotone operator, t_k a nonnegative parameter, the Lyapunov functional for the modified EAG-V algorithm with moving anchor is nonincreasing. Specifically, replacing the previous $\bar{z}^{k+1}$ update in the unmodified EAG-V moving anchor algorithm with equation (6.2) still results in a nonincreasing Lyapunov functional.*

Proof. Within the proof of Lemma 4.1 recall the following line:

$$\begin{aligned}-B_{k+1}\langle z^{k+1} - \bar{z}^{k+1}, G(z^{k+1})\rangle \\ = B_{k+1}\langle \bar{z}^k - z^{k+1}, G(z^{k+1})\rangle - B_{k+1}\langle \bar{z}^k - \bar{z}^{k+1}, G(z^{k+1})\rangle.\end{aligned}$$

Within this proof that the functional is nonincreasing, the primary change is that we must use equation (6.1) for substituting $G(z^{k+1})$. This results in

$$\begin{aligned}
-B_{k+1}\langle \bar{z}^k - \bar{z}^{k+1}, G(z^{k+1})\rangle \\
&= -B_{k+1}\left\langle \bar{z}^k - \bar{z}^{k+1}, \frac{\bar{z}^k - \bar{z}^{k+1} + t_k(H(\bar{z}^k) - H(\bar{z}^{k+1}))}{-\gamma_{k+1}}\right\rangle \\
&= \frac{B_{k+1}}{\gamma_{k+1}}\left(\|\bar{z}^k - \bar{z}^{k+1}\|^2 + t_k\langle \bar{z}^k - \bar{z}^{k+1}, H(\bar{z}^k) - H(\bar{z}^{k+1}\rangle\right).
\end{aligned}$$

The term $\frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2$ will be utilized elsewhere (see Lemma 4.1) so we don't need to worry about it here, and the term $\frac{B_{k+1}}{\gamma_{k+1}} t_k\langle \bar{z}^k - \bar{z}^{k+1}, H(\bar{z}^k) - H(\bar{z}^{k+1}\rangle$ is nonnegative by monotonicity and the fact that t_k is also nonnegative. This completes the proof. □

Theorem 6.3. *The modified EAG-V algorithm with moving anchor has convergence rate* $O(1/k^2)$.

Remark 6.4. While H may be any monotone operator, in practice one may wish to take $H = G$.

6.2. Modified FEG with moving anchor.

Definition 6.5 (Proximal FEG with moving anchor)**.** In the setting of FEG with moving anchor, consider (5.4) from the proof of Lemma 5.3:

$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1} G(z^{k+1})$$

and now let's consider the same term with a proximal term introduced:

$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1} G(z^{k+1}) - t_k(H(\bar{z}^k) - H(\bar{z}^{k+1})),$$

where H is a monotone operator just as before. This modification affects the anchor update in the same way as in the previous case:

$$\bar{z}^{k+1} = (I + t_k H)^{-1}(\bar{z}^k + \gamma_{k+1} G(z^{k+1}) + t_k H(\bar{z}^k)) \tag{6.3}$$

Lemma 6.6. *Under the same conditions as Lemma 5.3 and with H any monotone operator, t_k nonnegative for all k, the Lyapunov functional for the modified FEG algorithm with moving anchor is nonincreasing. Specifically, replacing the previous $\bar{z}^{k+1}$ update in the unmodified FEG moving anchor algorithm with (6.3) still results in a nonincreasing Lyapunov functional.*

Proof. The proof proceeds in the same manner as in that of Lemma 6.2. The only minor difference is that in this case, we begin with $B_{k+1}\langle G(z^{k+1}), \bar{z}^{k+1} - \bar{z}^k\rangle$. We still obtain from this the terms

$$\frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2 + \frac{B_{k+1}}{\gamma_{k+1}} t_k\langle \bar{z}^k - \bar{z}^{k+1}, H(\bar{z}^k) - H(\bar{z}^{k+1}\rangle,$$

where the first term is utilized elsewhere in the larger proof of the functional being nonincreasing and the latter term is monotone, thus nonnegative. □

Theorem 6.7. *The modified FEG algorithm with moving anchor has convergence rate* $O(1/k^2)$.

7. Numerical experiments

In this section we detail several numerical experiments. First, we visualize two thousand iterations of EAG-V and FEG, each moving anchor versus the fixed anchor, on a toy 'almost bilinear' example. Next, we look at the log of the grad norm squared versus the log of iterations for the EAG examples. Note that this error graph is an example in the monotone convex-concave case. We then run a nonconvex-nonconcave negative comonotone example for FEG variants, where some interesting convergence behaviors among the moving anchor variants are exhibited. Finally, we study monotone FEG variants (moving and fixed anchor) on a nonlinear two player game. Throughout all of these examples, $c_1 = \pi^2/6$, $c_k = \frac{c_{k-1}}{1+\delta_{k-1}}(k = 2, 3, ...)$, and in all except for the last example, δ_k is chosen to be $\exp(k^2) - 1$.

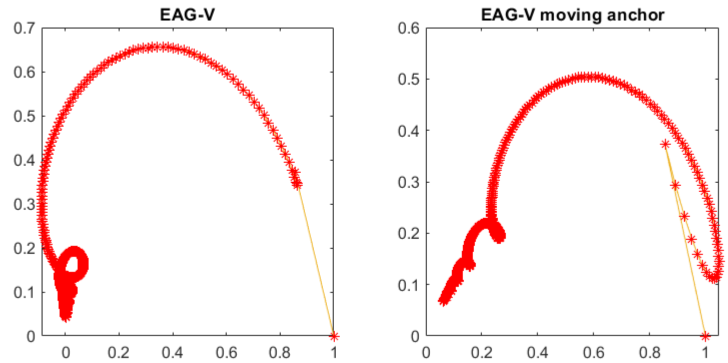

Figure 1. The first two thousand iterations of the EAG algorithm with varying step-size, or EAG-V, compared to the first two thousand iterations of the moving anchor EAG-V algorithm.

Figure 1 compares the iterations of EAG-V with a fixed anchor to the iterations for the moving anchor EAG-V. Figure 1, Figure 2, and Figure 3 all display iterations where the function used is the 'almost bilinear' function $f : \mathbb{R}^2 \to \mathbb{R}, f(x, y) = \epsilon\frac{\|x\|^2}{2} + \langle x, y\rangle - \epsilon\frac{\|y\|^2}{2}$. Here, ϵ is small, for these experiments set to 0.01, and the straightforward nature of the example allows for ease of visualizing the iterations as well as their differences when it comes to comparing convergence rates. In particular, the unique saddle-point is $(0, 0)$.

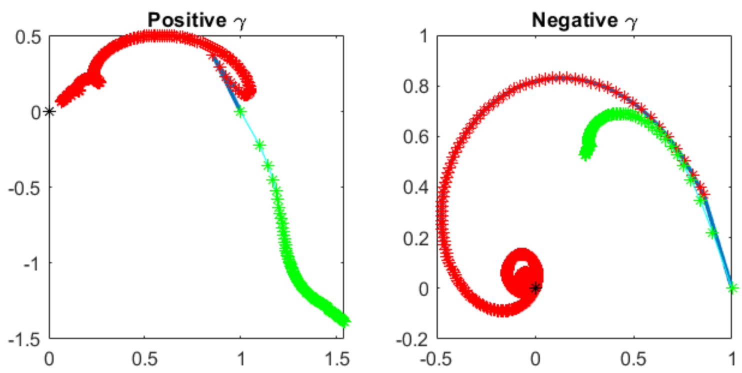

Figure 2. The two moving anchor EAG-V variants compared in red, along with their anchors in green.

Figure 2 compares, via the same function as Figure 1, the two moving anchor variants of EAG-V. When the γ_k parameter is positive, the anchor iterations moves away from the saddle and the algorithm updates very rapidly. When γ_k has only its sign changed to negative, the anchor (seen in green) seems to stay much closer to the iterations and the saddle-point. The iterations appear to converge at a markedly faster rate (by a constant) for this latter case over both the fixed anchor and the positive γ_k setting, an observation that is confirmed below.

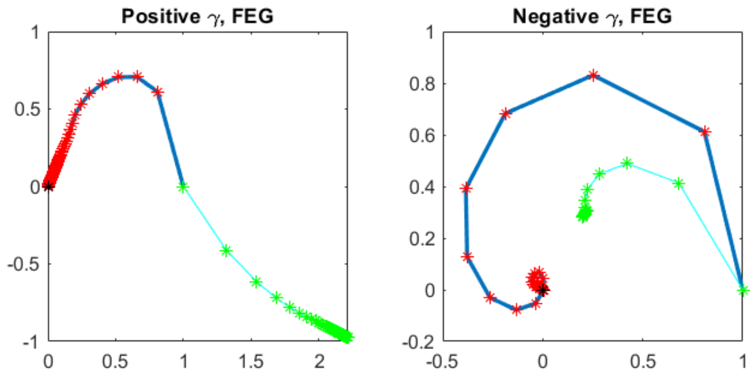

FIGURE 3. The two moving anchor FEG variants compared in red, along with their anchors in green.

Figure 3 compares the two moving anchor version of the FEG method, in the same manner as the comparison shown in Figure 2: red dots are the algorithm updates, green dots are the anchor updates, and the function is the 'almost bilinear' one previously described. Note that in both cases, the iterations seem to zone in on and converge to the saddle point in a much faster manner. In [14], the authors established that even on convex-concave problems, FEG performs at the same optimal order of convergence as EAG, but at a significantly faster rate. This behavior seems to have carried over to our algorithm where we introduce the moving anchor to these frameworks.

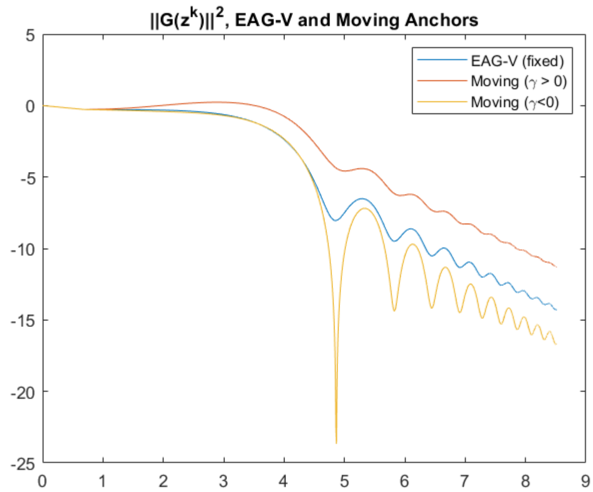

FIGURE 4. Comparison of the grad-norm squared of three EAG-V variants of interest on a toy 'almost bilinear' problem.

Figure 4 captures the behavior of $\|G(z^k)\|^2$ across all three convex-concave algorithms of interest: EAG-V, moving anchor EAG-V with positive γ_k, and moving anchor EAG-V with negative γ_k. Each algorithm attains the optimal order of convergence, while the negative γ_k algorithm is markedly faster than both algorithms by a constant. Identical behavior occurs under the same problem setting with the FEG and FEG with moving anchors (positive and negative γ_k), with the negative γ_k algorithm again being the fastest, so we do not include this figure here.

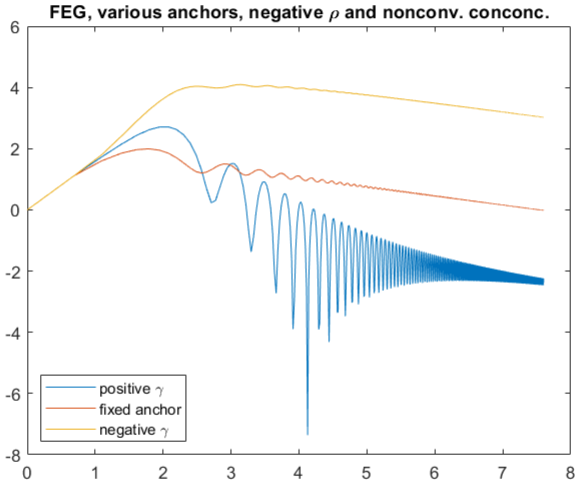

FIGURE 5. Comparison of the errors of three FEG variants in a nonconvex-nonconcave setting. Note the positive γ converges fastest by a constant.

Figure 5 captures the error of FEG across all three anchor variants in a numerical example that is explicitly comonotone and nonconvex-nonconcave:

$$L(x,y) = \frac{\rho R^2}{2}x^2 + R\sqrt{1-\rho^2 R^2}xy - \frac{\rho R^2}{2}y^2$$

with $L : \mathbb{R}^2 \to \mathbb{R}, R = 1, \rho = -1/3$ $1-$smooth and $-1/3$-negative comonotone. Interestingly, this is the only numerical example where the moving anchor with positive γ_k - of any variant - was the fastest of all three algorithms. The intuition is that the positive γ_k functions as negative γ_k in the monotone, convex-concave problem settings, pulling the iterations closer to the saddle. More examples in this vein may verify that this behavior with positive γ_k occurs only in the negative comonotone problem setting. More theoretical work to verify this observed numerical behavior will be one of our goals in a future work.

The final figure, Figure 6, compares three different monotone FEG variants on a particular nonlinear game that was studied extensively in [4]:

$$\min_{x\in\Delta^n}\max_{y\in\Delta^m} \frac{1}{2}\langle Qx, x\rangle + \langle Kx, y\rangle$$

where $Q = A^T A$ is positive semidefinite for $A \in \mathbb{R}^{k\times n}$ which has entries generated independently from the standard normal distribution, $K \in \mathbb{R}^{m\times n}$ with entries generated uniformly and independently from the interval $[-1, 1]$, and Δ^n, Δ^m are the $n-$ and $m-$simplices, respectively:

$$\Delta^n := \Big\{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\Big\},\ \Delta^m := \Big\{y \in \mathbb{R}^m_+ : \sum_{j=1}^m y_j = 1\Big\}.$$

One may interpret this as a two person game where player one has n strategies to choose from, choosing strategy i with probability x_i $(i = 1, ..., n)$ to attempt to minimize a loss,

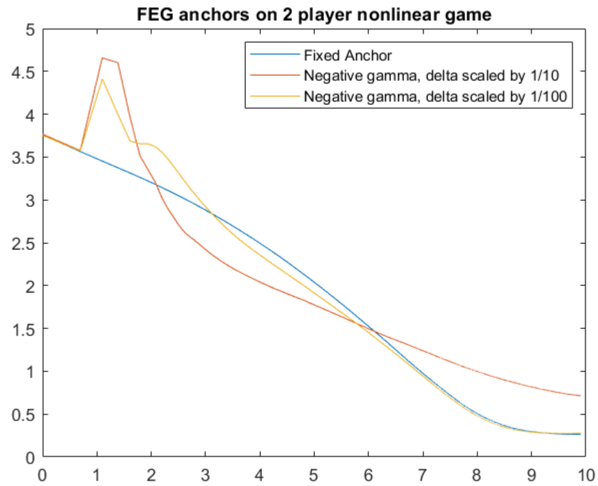

FIGURE 6. log of iterations versus log of the gradient norm-squared for monotone (that is, $\rho = 0$) FEG variants studied on a particular nonlinear game.

while the second player attempt to maximize their gain among m strategies with strategy j chosen with probability y_j $(j = 1, ..., m)$. The payoff is a quadratic function that depends on the strategy of both players. For this example, we used FEG fixed and moving anchor variants in the monotone (that is, $\rho = 0$) setting of the algorithm. We compare $20,000$ iterations of the log of the grad norm squared of the fixed anchor versus the same for a negative γ_k variant where the parameter δ is scaled by $1/10$ and a negative γ_k variant where δ is scaled by $1/100$. This is the first numerical example where we tune δ, a parameter used in controlling the step size. We chose $m = 500$, $k = 1000$, and $n = 2500$.

We remark that, initially, it seems the $\frac{1}{10}\delta$ variant is fastest by a constant, and then is overtaken by the $\frac{1}{100}\delta$ variant and then the fixed anchor while the initially faster $\frac{1}{10}\delta$ 'flattens out' rather quickly. For much of this experiment, it appears that the $\frac{1}{100}\delta$ variant and the fixed anchor very closely parallel one another - there seems to be a slight advantage to the $\frac{1}{100}\delta$ moving anchor for the majority of the experiment after between two and three dozen iterations, and then at the tail end of the experiment the fixed anchor may have a marginal lead.

8. CONCLUSION

The moving anchor acceleration methods retain optimal convergence rates and also demonstrate superior-to-comparable numerical performance with some parameter tuning. The optimal order of convergence is obtained across different problem settings, from convex-concave to negative co-monotone problems. Interestingly, across numerous problem settings there exists a version of the moving anchor algorithm, parametrized by γ_k, that demonstrates superior numerical performance compared to other state-of-the-art algorithms. The variety of numerical examples demonstrates a wide array of applications for our algorithms in both theoretical and applied settings. In addition, we develop a 'proximal' version of the moving anchor in both the convex-concave and negative co-monotone problem settings and demonstrate its convergence. Of future interest, one may consider numerical and practical implementations of the proximal moving anchor, parallelized/asynchronous implementations of moving anchor saddle point algorithms, a tighter analysis of $-\gamma_k$ convergence, a theoretical understanding of how the γ_k and other parameters such as δ affect convergence rates, and the identification of problem settings which our moving anchor may exploit effectively among many other topics.

Acknowledgements

We thank Donghwan Kim, Ernest K. Ryu, and Taeho Yoon for their invaluable suggestions that have greatly improved and inspired this work.

References

1. Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel, *A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games*, International conference on artificial intelligence and statistics, PMLR, 2020, pp. 2863–2873.
2. Xufeng Cai, Chaobing Song, Cristóbal Guzmán, and Jelena Diakonikolas, *A stochastic halpern iteration with variance reduction for stochastic monotone inclusion problems*, arXiv preprint arXiv:2203.09436 (2022).
3. Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien, *Reducing noise in gan training with variance reduced extragradient*, Advances in Neural Information Processing Systems **32** (2019).
4. Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang, *Optimal primal-dual methods for a class of saddle point problems*, SIAM Journal on Optimization **24** (2014), no. 4, 1779–1814.
5. Jelena Diakonikolas, *Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities*, Conference on Learning Theory, PMLR, 2020, pp. 1428–1451.
6. Jelena Diakonikolas, Constantinos Daskalakis, and Michael I Jordan, *Efficient methods for structured nonconvex-nonconcave min-max optimization*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 2746–2754.
7. Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou, *Stochastic variance reduction methods for policy evaluation*, International Conference on Machine Learning, PMLR, 2017, pp. 1049–1058.
8. Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar, *Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems*, 2020.
9. Ian J Goodfellow, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, and Jean Pouget-Abadie, *Generative adversarial nets*, Advances in neural information processing systems **27** (2014), 2672–2680.
10. Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel, *Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity*, International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 366–402.
11. Benjamin Halpern, *Fixed points of nonexpanding maps*, Bulletin of the American Mathematical Society **73** (1967), no. 6, 957–961.
12. Galina M Korpelevich, *The extragradient method for finding saddle points and other problems*, Matecon **12** (1976), 747–756.
13. Jongmin Lee and Ernest K Ryu, *Accelerating value iteration with anchoring*, arXiv preprint arXiv:2305.16569 (2023).
14. Sucheol Lee and Donghwan Kim, *Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems*, (2021).
15. 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 (2021).
16. Felix Lieder, *On the convergence rate of the halpern-iteration*, Optimization letters **15** (2021), no. 2, 405–418.
17. Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang, *Towards better understanding of adaptive gradient algorithms in generative adversarial nets*, arXiv preprint arXiv:1912.11940 (2019).
18. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, *Towards deep learning models resistant to adversarial attacks*, arXiv preprint arXiv:1706.06083 (2017).
19. 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** (2004), no. 1, 229–251.
20. Arkadij Semenovič Nemirovskij and David Borisovich Yudin, *Problem complexity and method efficiency in optimization*, (1983).
21. Arkadi S Nemirovsky, *On optimality of krylov's information when solving linear operator equations*, Journal of Complexity **7** (1991), no. 2, 121–130.
22. ———, *Information-based complexity of linear operator equations*, Journal of Complexity **8** (1992), no. 2, 153–175.
23. Yurii Nesterov, *A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2)*, Doklady an ussr, vol. 269, 1983, pp. 543–547.
24. ———, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Mathematical Programming **109** (2007), no. 2-3, 319–344.

25. Yuyuan Ouyang and Yangyang Xu, *Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems*, Mathematical Programming **185** (2021), no. 1-2, 1–35.
26. R Tyrrell Rockafellar, *Monotone operators associated with saddle-functions and minimax problems*, Proceedings of Symposia in Pure Mathematics, vol. 18, American Mathematical Society, 1970, pp. 241–250.
27. Ernest K Ryu, Kun Yuan, and Wotao Yin, *Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems and gans*, (2019).
28. Jaewook J Suh, Jisun Park, and Ernest K Ryu, *Continuous-time analysis of anchor acceleration*, arXiv preprint arXiv:2304.00771 (2023).
29. Quoc Tran-Dinh, *The connection between nesterov's accelerated methods and halpern fixed-point iterations*, arXiv preprint arXiv:2203.04869 (2022).
30. Quoc Tran-Dinh and Yang Luo, *Halpern-type accelerated and splitting algorithms for monotone inclusions*, arXiv preprint arXiv:2110.08150 (2021).
31. Paul Tseng, *On linear convergence of iterative methods for the variational inequality problem*, Journal of Computational and Applied Mathematics **60** (1995), no. 1-2, 237–252.
32. Taeho Yoon and Ernest K Ryu, *Accelerated algorithms for smooth convex-concave minimax problems with $o(1/k^2)$ rate on squared gradient norm*, Proceedings of the 38th International Conference on Machine Learning **139** (2021), 12098–12109.

Department of Mathematics, University of California, Riverside, USA
Email address: `jalca014@ucr.edu`

Department of Mathematics, University of California, Riverside, USA
Email address: `yattinc@ucr.edu`

Department of Mathematics, Purdue University, West Lafayette, Indiana, USA
Email address: `msunkula@purdue.edu`