UNIVERSITY OF CALIFORNIA
RIVERSIDE

New Classes of Moving Anchor Extragradient Algorithms for Saddlepoint Problems

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Mathematics

by

James K. Alcala

June 2024

Dissertation Committee:

Dr. Yat Tin Chow, Chairperson
Dr. Weitao Chen
Dr. Heyrim Cho

The Dissertation of James K. Alcala is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First and foremost, I am grateful to my advisor Yat Tin Chow, whose guidance, patience, understanding, wisdom, humor, and scientific expertise have guided this project since the beginning and seen it to its fruition. Similarly, I am grateful to Mahesh S., our collaborator, for his insights regarding TeX, expertise in MATLAB, knowledge in common mathematical writing practices, guidance during the seemingly endless hours working in the trenches on the specifics of this work, and friendliness in working alongside me. I am also grateful to Ernest K. R. and his students TaeHo Y., Jongmin L., Jisun P., and Jaewok S. whose works have greatly inspired this project, as well as for their friendliness and continued support. I am grateful to Donghwan K., whose work and discussion have also been significant driving forces in the present work. I am grateful to Wotao Y., whose optimization research introduced me to the field, and who graciously has assisted me in these early stages in my career. I am in debt to Stanley J. O. and his group, especially Siting L., for their invitation to present on this work in its early stages and for their discussion. I am thankful for Yat Sun P. inviting me to join the Microtutorials Project, and for the continued guidance in these early stages of my career. I would like to acknowledge Zhenghe Z., whose undergraduate real analysis course was one of the defining points in my choice to pursue a mathematical career, and for his enjoyable reading course. I am forever grateful to Marlén R.-H. for so much, but most of all, her friendship. Jamal M., Liz H., and Chris J., for their early guidance and mentorship. Lastly, I must acknowledge Fernando L.-G., my first mathematical mentor who took me in as a new math major and within a year, had me confidently presenting on Poincaré inequalities in research-level seminars.

“ [22] *The stone the builders rejected has become the capstone;*

[23] *the* LORD *has done this, and it is marvelous in our eyes.*

[24] *This is the day the* LORD *has made; let us rejoice and be glad in it.*”

Psalm 118:22-24, NIV.

To my brother Jacob, my mother Loretta, my father Jim, and Frosty, for everything. To Alex and Leo, and our exciting and limitless future together. To UCR’s Society of Hispanic Professional Engineers chapter, to the Verbal Coliseum, and to anyone who ever sat down to play music with me. To the communities at One and All Church and Crest Community Church, for sharing God’s love with me when I needed it most. To our Creator, whose love and faithfulness continually transform and guide me. Finally, to all the friends and family who got me here, and especially those who are no longer here.

ABSTRACT OF THE DISSERTATION

New Classes of Moving Anchor Extragradient Algorithms for Saddlepoint Problems

by

James K. Alcala

Doctor of Philosophy, Graduate Program in Mathematics
University of California, Riverside, June 2024
Dr. Yat Tin Chow, Chairperson

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 [46], 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 also extend the moving anchor to a more general nonconvex-nonconcave class of saddle point problems using the framework of [23], which generalizes [46]. We obtain similar order-optimal complexity results in this extended case. A preconditioned version of our algorithms is also introduced and analyzed to match optimal theoretical convergence rates. Our final theoretical contribution is the development and analysis of a moving anchor with a stochastic oracle, which matches accelerated convergence rates for convex-concave problems. Underlying these theoretical contributions is a selection of robust new Lyapunov/energy functional techniques that account for the moving anchor structure while maintaining the optimal order of complexity

under minimal assumptions. Various numerical experiments demonstrate the efficacy of the moving anchor extragradient algorithms compared to their fixed anchor variants, and in many cases suggest a more optimal constant in the big O notation that may surpass the traditional fixed anchor methods. We conclude by discussing current challenges and future directions this work may lead.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

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 [15], [5], adversarial learning [28], and reinforcement learning [12]. 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 the convergence guarantees for this 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 [46], combines extragradient and the more recently developed anchoring methods in a single framework 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 [22], [42], [46]. To show optimality, the authors of [46] adapt arguments from [33], [34] 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 [24], [44], [43]. 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 [23] introduced the Fast ExtraGradient Method, or FEG. This method generalizes the results of EAG and EG + [11] 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 by 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, called 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. 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.

3. We develop a stochastic moving anchor EAG-V variant that retains order-optimal convergence guarantees for this problem setting. Many modern optimization problems involve computations in very high dimensions, so this development is especially valuable for many applications.

4. 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. We also perform some experiments in the stochastic setting. These numerical examples demonstrate the efficacy of our algorithms, as in all deterministic cases, one of the moving anchor algorithm versions in each example is the fastest algorithm by a constant. The stochastic algorithms also demonstrate favorable convergence behavior across anchor variants.

1.2 Literature Review

1.2.1 Halpern Iteration and Anchoring

Introduced in 1967 and inspired by Browder's classical fixed point theorem, the Halpern iteration [18] is an algorithm built for approximating fixed point(s) of nonexpanding maps in a Hilbert space. Its convergence has been studied in [25], and it is extensively used in monotone inclusion-type problem settings [10], [44], [4]. A recent paper [43] draws an explicit connection between Halpern-inspired methods and Nesterov's AGM [35], linking two very active strains of acceleration literature.

Directly inspired by Halpern, algorithmic anchoring was recently introduced in the literature [40] and has since been utilized to establish optimal $O(1/k^2)$ convergence rates for smooth-structured convex-concave minimax problems [46]. Since then, these methods have been extended to the nonconvex-nonconcave, negative comonotone problem setting [23] and analogous settings for composite problems in a multi-step framework [24]. 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 [44], the authors develop an anchored Popov's scheme and a splitting version of the EAG developed in [46], with a similar analysis.

1.2.2 Extragradient Methods

The extragradient method first appeared in [21] and has since been an important acceleration method extensively studied in the optimization literature [2], [45], [27], especially in the context of generative adversarial networks [15], [5] and adversarial training [28]. 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 [30], [36] have an $O(1/k)$ rate, which is order-optimal [37], [32]. Recently, it was shown that the last iterate convergence rate for extragradient also attains $O(1/k)$ convergence [17], with *only* monotonicity and Lipschitz assumptions. This closes the gap between the last-iterate and ergodic-iterate convergence rates for extragradient discussed in [14]. Another recent interesting result was attained in [11], 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 $O(1/k^2)$ [46], [23], thanks in part to the synthesis with anchoring. This breaks the bound of the SCLI class of algorithms discussed in [14], 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 [46] 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 [46] and [14] 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 work and these recent works, has room for numerical improvement.

1.2.3 Stochastic Algorithms

Stochastic methods in optimization have a long and celebrated history owing to their ability to reduce the computational bottlenecks commonly encountered in modern, high-dimensional problems [3], and the literature regarding such methods in closely aligned fields is rich [23], [19], [20], [29], [31], [40], [41], [48]. We remark that [23] and [40] presented the first results regarding stochastic anchoring algorithms. [4] studies monotone inclusions with a stochastic Halpern iteration, combining these techniques with variance reduction to achieve optimal results in terms of stochastic oracle complexity. Crucially, the former work develops deterministic methods for certain nonmonotone operators, however, it develops stochastic methods for monotone operators associated to convex-concave problems only. Whereas [4] attains stronger stochastic oracle complexity results, but only in the case of monotone operators. In general, developing theory for the accelerated stochastic versions of algorithms in these nonconvex-nonconcave settings (and the associated nonmonotone operators) seems to be a hard problem. In [48] the authors develop stochastic methods for a class of nonconvex minimization problems referred to as *variationally co-*

herent, equivalent to the Minty Variational Inequality condition, which includes convexity and other interesting problem classes as sub-classes. Referring to this problem class as *coherent non-monotone variational inequalities*, the authors in [41] develop similarly powerful stochastic methods in the variational inequality setting (ie, a problem setting that contains saddle point problems). Until fairly recently, the most prominent article on the subject of stochastic nonconvex-nonconcave saddle point problems was [10], which develops the Extragradient+ method for a large class of nonmonotone problems, up to and including the very broad weak Minty Variational Inequaltiy class of problems. More recently, [13] provides an analysis of stochastic extragradient algorithms in the strongly monotone, affine, and monotone settings when random reshuffling, a popular computational technique for stochastic algorithms, is involved. [16] provides a very general framework to analyze stochastic extragradient algorithms and their variants. To conclude this section on a more optimistic note, both [7] and [26] have very recently developed stochastic extragradient methods with robust analyses for settings beyond monotone operators.

1.3 Notation and Basic Results

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,

$$G_L(z) = \begin{bmatrix} \nabla_x L(x, y) \\ -\nabla y L(x, y) \end{bmatrix} \tag{1.2}$$

where the L subscript is omitted when the underlying saddle function is known. When our problem is convex-concave, the operator (1.2) is known to be monotone [39], 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 (1.2) 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 [23]

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.

Chapter 2

Previous Fixed Anchor Methods

In this chapter, we outline the specifications and convergence results of the two original fixed anchor methods, the Extra Anchored Gradient (EAG) [46] and Fast Extra-Gradient (FEG) [23].

2.1 The Extra Anchored Gradient

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

$$z^{k+1/2} = z^k + \beta_k(z^0 - z^k) - \alpha_k G(z^k) \tag{2.1}$$

$$z^{k+1} = z^k + \beta_k(z^0 - z^k) - \alpha_k G(z^{k+1/2}) \tag{2.2}$$

$$\begin{aligned} \alpha_{k+1} &= \frac{\alpha_k}{1-\alpha_k^2 R^2}\left(1 - \frac{(k+2)^2}{(k+1)(k+3)}\alpha_k^2 R^2\right) \\ &= \alpha_k\left(1 - \frac{1}{(k+1)(k+3)}\frac{\alpha_k^2 R^2}{1-\alpha_k^2 R^2}\right) \end{aligned} \tag{2.3}$$

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 [46]. 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 [46].

Theorem 1 (EAG-V convergence rate [46]) *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 2 (EAG Lyapunov Functional [46]) *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 $k \geq 0$:*

$$A_k = \frac{\alpha_k}{2\beta_k} B_k \tag{2.4}$$

$$B_{k+1} = \frac{B_k}{1 - \beta_k} \tag{2.5}$$

$$\alpha_{k+1} = \frac{\alpha_k \beta_{k+1}(1 - \alpha_k^2 R^2 - \beta_k^2)}{\beta_k(1 - \beta_k)(1 - \alpha_k^2 R^2)} \tag{2.6}$$

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 $\{V_k\}_{k\geq 0}$ (2.7) defined below is nonincreasing in k.

$$V_k := A_k \|G(z^k)\|^2 + B_k \langle G(z^k), z^k - z^0 \rangle \tag{2.7}$$

Remark 3 *Within* (2.7)*, 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* (2.6)*.*

We also need the clarify the behavior of (2.6), as this will be needed in multiple analyses later.

Lemma 4 *If $\alpha_0 \in (0, \frac{3}{4R})$, then the sequence $\{\alpha_k\}_{k=0}^{\infty}$ of* (2.6) *monotonically decreases to a positive limit.*

Proof. This is proved as a corollary of Theorem 23. ■

2.2 The Fast Extra Gradient

In [23], the methods in [46] are expanded to a broader class of smooth structured nonconvex-nonconcave minimax problems, bringing an $O(1/k^2)$ rate of convergence rate to a larger class of problems in the setting of minimax games. This algorithm class is called the FEG, or Fast Extra Gradient method.

$$z^{k+1/2} = z^k + \beta_k (z^0 - z^k) - (1 - \beta_k)(\alpha_k + 2\rho_k) G(z^k) \tag{2.8}$$

$$z^{k+1} = z^k + \beta_k (z^0 - z^k) - \alpha_k G(z^{k+1/2}) - (1 - \beta_k) 2\rho_k G(z^k) \tag{2.9}$$

First, we remark that each update (2.8), (2.9) is mostly very similar to the corresponding updates in EAG-V, (2.1), (2.2). The reuse of the term $G(z^k)$ is useful for handling the negative comonotonicity that is represented by the parameter ρ_k; in (strongly) monotone problems, $\rho_k(>) = 0$. To obtain the accelerated convergence results, one may choose $\alpha_k = \frac{1}{R}, \beta_k = \frac{1}{k+1}, \rho_k = \rho$ for all $k \geq 0$, and we have the following theorem resulting.

Theorem 5 (Fixed Anchor FEG Convergence) *For the $R-$Lipschitz continuous and ρ-comonotone operator G with $\rho \geq 0$ and for any z^* in the set of points fixed by G, the sequence $\{z^k\}_{k\geq 0}$ generated by FEG satisfies, for all $k \geq 1$,*

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

Notably, this bound given by [23] is a significant improvement over the bound in [46], and the authors also provide a stochastic version with analogous convergence guarantees. The analysis of this algorithm is also completed using a powerful Lyapunov descent lemma. We direct interested readers to Section 7 of [23] to see its details, in particular Lemma 7.1, and to Theorem 12 for the more general moving anchor version of this descent lemma.

Chapter 3

Moving Anchor EAG-V

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

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

$$z^{k+1/2} = z^k + \frac{1}{k+2}(\bar{z}^k - z^k) - \alpha_k G(z^k) \tag{3.1}$$

$$z^{k+1} = z^k + \frac{1}{k+2}(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2}) \tag{3.2}$$

$$\bar{z}^{k+1} = \bar{z}^k + \gamma_{k+1} G(z^{k+1}) \tag{3.3}$$

The major structural difference 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. (3.3) is the regular update for this anchor; it depends on the algorithm update (3.2) rather than exclusively on itself. However, the fact that the anchor is now moving requires some additional machinery to

ensure that a new, more general Lyapunov functional is still nonincreasing. To these ends, the previously defined sequences remain the same, with new additions in the following sequences.

$$c_{k+1} \leq \frac{c_k}{1+\delta_k}, \tag{3.4}$$

$$\gamma_{k+1} \leq \frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}. \tag{3.5}$$

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 chapter and the chapters relating to the moving anchor FEG and the stochastic moving anchor EAG-V. 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 work, we take (3.4) and (3.5) to be given with equal signs instead of inequalities. As in the fixed anchor case, we will take $\beta_k = \frac{1}{k+2}$, resulting in similar sequences (2.4), (2.5), (2.6) for the moving anchor. 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 6 *The Lyapunov functional*

$$V_k := A_k\|G(z^k)\|^2 + B_k\langle G(z^k), z^k - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2, \tag{3.6}$$

corresponding to the moving anchor EAG-V algorithm (3.1) *through* (3.3) *with constants* A_k, B_k, c_k, β_k *defined as in* (2.4), (2.5), (2.6), *and Theorem 2, along with sequences* c_k, γ_k *defined in* (3.4), (3.5), *is non increasing.*

Proof.

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

$$z^k - z^{k+1} = \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}) \tag{3.7}$$

$$z^{k+1/2} - z^{k+1} = \alpha_k(G(z^{k+1/2}) - G(z^k)) \tag{3.8}$$

$$\bar{z}^k - z^{k+1} = (1-\beta_k)(\bar{z}^k - z^k) + \alpha_k G(z^{k+1/2}) \tag{3.9}$$

$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1} G(z^{k+1}) \tag{3.10}$$

(3.7) comes from rearranging (3.2), (3.8) comes from taking the difference between (3.1) and (3.2), (3.9) is $\bar{z}^k$ minus (3.2), and (3.10) is (3.3) rearranged. The overall goal of this proof is to show that the difference $V_k - V_{k+1}$ is nonnegative.

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

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 IV below. We may take advantage of the previously established identities (3.7) through (3.10).

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

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

$$\underbrace{-B_{k+1}\langle z^{k+1} - \bar{z}^{k+1}, G(z^{k+1})\rangle}_{\text{II}}$$

$$= B_{k+1}\langle \bar{z}^k - z^{k+1}, G(z^{k+1})\rangle - B_{k+1}\langle \bar{z}^k - \bar{z}^{k+1}, G(z^{k+1})\rangle \tag{3.11}$$

$$= B_{k+1}\langle (1-\beta_k)(\bar{z}^k - z^k) + \alpha_k G(z^{k+1/2}), G(z^{k+1})\rangle - B_{k+1}\langle -\gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle \tag{3.12}$$

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 (3.9) and (3.10). For III,

$$\underbrace{-\frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^k) - G(z^{k+1})\rangle}_{\text{III}}$$

$$= -\frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^k)\rangle + \frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^{k+1})\rangle \tag{3.13}$$

$$= -\frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^k)\rangle + \frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^{k+1})\rangle,$$

where the last equality is a result of substituting in (3.7) in each of the first arguments of the two terms in (3.13). Now, we can begin simplify everything we've done to obtain IV.

$$\underbrace{\langle z^k - \bar{z}^k, G(z^k)\rangle}_{\text{I}} \tag{3.14}$$

$$\underbrace{\langle (1-\beta_k)(z^k - \bar{z}^k) - \alpha_k G(z^{k+1/2}) - \gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle}_{\text{II}} \tag{3.15}$$

$$\underbrace{-\frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^k)\rangle}_{\text{III}} \tag{3.16}$$

$$\underbrace{+\frac{B_k}{\beta_k}\langle \beta_k(z^k - \bar{z}^k) + \alpha_k G(z^{k+1/2}), G(z^{k+1})\rangle}_{\text{III}} \tag{3.17}$$

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 (3.15) with the $\beta_k(z^k - \bar{z}^k)$ component of (3.17). Additionally, (3.14) cancels with the $\beta_k(z^k - \bar{z}^k)$ component of (3.16). 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

$$
\begin{aligned}
&A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 \\
&+B_k\langle z^k - \bar{z}^k, G(z^k)\rangle - B_{k+1}\langle z^{k+1} - \bar{z}^{k+1}, G(z^{k+1})\rangle \\
&+c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 - \frac{B_k}{\beta_k}\langle z^k - z^{k+1}, G(z^k) - G(z^{k+1})\rangle \\
&=A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \alpha_k B_{k+1}\langle G(z^{k+1/2}), G(z^{k+1})\rangle \qquad (3.18)\\
&-\frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle \qquad (3.19)\\
&+c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 + \frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2 \qquad (3.20)
\end{aligned}
$$

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

$$
||z^* - \bar{z}^{k+1}\|^2 \leq (1+\delta_k)\|z^* - \bar{z}^k\|^2 + (1+\frac{1}{\delta_k})\|\bar{z}^k - \bar{z}^{k+1}\|^2 \qquad (3.21)
$$

and from the algorithm definition,

$$
c_{k+1} = \frac{c_k}{1+\delta_k}, \quad \gamma_{k+1} = \frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}. \qquad (3.22)
$$

We apply (3.21) to (3.20) 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 (3.22):

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

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 [46]. Specifically, following (29) onwards in [46], 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 7 *The EAG-V algorithm with moving anchor* (3.1), (3.2), *and* (3.3) *together with the Lyapunov functional* V_k (3.6) *described in Theorem 6, has convergence rate*

$$
\|G(z^k)\|^2 \leq \frac{4(\alpha_0 R^2 + c_0)\|z^0 - z^*\|^2 2V_0}{\frac{\alpha_\infty}{4}(k+1)(k+2)} \tag{3.23}
$$

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

Proof. For the most part, this argument parallels the analogous argument found in [46]. We use the Lyapunov functional to isolate and bound $\|G(z^k)\|^2$.

$$V_k \leq V_0 = \alpha_0 \|G(z^0)\|^2 + c_0 \|z^0 - z^*\|^2 \tag{3.24}$$
$$\leq (\alpha_0 R^2 + c_0)\|z^0 - z^*\|^2$$

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. ■

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 8 *Replacing γ_k with $-\gamma_k$ in the definition of the EAG-V algorithm with moving anchor* (3.1), (3.2), (3.3), *and suppose* $\gamma_{k+1} = \min \dfrac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}, \dfrac{e_{k+1}}{2B_{k+1}\|G(z^{k+1})\|^2}$, *where* $\sum e_k < \infty$. *Then the Lyapunov functional* (3.6) *is nonincreasing, and has the same order of convergence* $O(1/k^2)$ *as in the positive* γ_k *moving anchor EAG-V algorithm* (3.23).

Proof. First, note that the anchor update (3.3) has been modified to become

$$-\gamma_{k+1} = -\frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}, \tag{3.25}$$

resulting in the following modification to (3.10):

$$\bar{z}^k - \bar{z}^{k+1} = \gamma_{k+1} G(z^{k+1}). \tag{3.26}$$

We see the first adjustment in the previous lemma in the transition from line (3.11) to (3.12); note that we focus only on the terms dependent on (3.26):

$$\begin{aligned}
&-B_{k+1}\langle \bar{z}^k - \bar{z}^{k+1}, G(z^{k+1})\rangle \\
&\quad = -B_{k+1}\langle \gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle \\
&\quad = -B_{k+1}\langle (2\gamma_{k+1} - \gamma_{k+1}) G(z^{k+1}), G(z^{k+1})\rangle \\
&\quad = -B_{k+1}\langle 2\gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle + B_{k+1}\langle \gamma_{k+1} G(z^{k+1}), G(z^{k+1})\rangle. \qquad (3.27)
\end{aligned}$$

The latter term in line (3.27) will go on and cancel in a quadratic form as in the proof of the original lemma. We are left 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 Theorem 6, we end up with the inequality

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

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

By construction, the left-hand side of the inequality should remain nonnegative. Now, by the new construction of γ_k we have

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

when we proceed as in the proof of Theorem 7 to show convergence, getting to the line (3.24), we get the chain of inequalities

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

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 Theorem 8 may not be too restrictive to work with. Our 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 the tuning of γ_k benefits numerical convergence rates. We leave the theoretical exploration of this phenomena to future work.

Chapter 4

Moving Anchor FEG

We introduce the moving anchor to the expanded framework of [23] involving certain nonconvex-nonconcave objective functions with associated negative comonotone saddle gradients. We show that a moving anchor with the same conditions to the convex-concave setting is a feasible approach in this broader 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 [23], is given as

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

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

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

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

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

where $\{\delta_k\}$ is chosen so that $\sum_{i=0}^{\infty} \log(1+\delta_i) < \infty$, with $\{\gamma_k\}$, and $\{c_k\}$, 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 in order:

Remark 9 *An additional assumption on the saddle-gradient operator G is needed: for some $\rho \in \left(-\frac{1}{2R}, \infty\right)$, $\langle G(z) - G(z'), z - z'\rangle \geq \rho\|G(z) - G(z')\|^2 \; \forall z, z' \in \mathbb{R}^m \times \mathbb{R}^n$. (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 10 *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.*

Remark 11 *Many of the sequences defined in the following lemma have similar naming conventions to those defined in Theorem 6. However, the instance of the moving anchor*

FEG class (4.1), (4.2), (4.3) *for which we state convergence results utilizes* $\alpha_k = \frac{1}{R}, \beta_k = \frac{1}{k+1}, \rho_k = \rho, R_k = R$ *for* $k \geq 0$. *To be clear, Theorem 12 is more general and does NOT require these definitions, while Theorem 13 uses these definitions for explicit convergence results.*

Lemma 12 *Suppose that the sequences* $\{c_k\}_{k\geq 0}, \{\gamma_k\}_{k\geq 0}$, *are defined as in* (4.4), (4.5), *and 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 a sequence* $\{\rho_k\} \subset \mathbb{R}$ *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

$$V_k := A_k\|G(z^k)\|^2 - B_k\langle G(z^k), \bar{z}^k - z^k\rangle + c_k\|z^* - \bar{z}^k\|^2, \tag{4.6}$$

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:

$$z^{k+1} - z^k = \frac{\beta_k}{1-\beta_k}(\bar{z}^k - z^{k+1}) - \frac{\alpha_k}{1-\beta_k}G(z^{k+1/2}) - 2\rho_k G(z^k) \tag{4.7}$$

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

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

$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1}G(z^{k+1}) \tag{4.10}$$

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:

$$\begin{aligned}
&V_k - V_{k+1} \\
\geq& A_k\|G(z^k)\|^2 - B_k\langle G(z^k), \bar{z}^k - z^k\rangle - A_{k+1}\|G(z^{k+1})\|^2 + B_{k+1}\langle G(z^{k+1}), \bar{z}^{k+1} - z^{k+1}\rangle \\
&+ c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 \\
&- \frac{B_k}{\beta_k}\big(\langle G(z^{k+1}) - G(z^k), z^{k+1} - z^k\rangle - \rho_k\|G(z^{k+1}) - G(z^k)\|^2\big) \\
=& A_k\|G(z^k)\|^2 - B_k\langle G(z^k), \bar{z}^k - z^k\rangle \qquad (4.11) \\
&- A_{k+1}\|G(z^{k+1})\|^2 + B_{k+1}\langle G(z^{k+1}), \bar{z}^{k+1} - z^{k+1}\rangle \qquad (4.12) \\
&+ c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 \\
&- \frac{B_k}{\beta_k}\langle G(z^{k+1}), z^{k+1} - z^k\rangle + \frac{B_k}{\beta_k}\langle G(z^k), z^{k+1} - z^k\rangle + \frac{B_k\rho_k}{\beta_k}\|G(z^{k+1}) - G(z^k)\|^2
\end{aligned}$$

From here, we first simplify the introduced term further and then substitute (4.7) into the inner product which has a B_k out front, and then substitute (4.8) into the inner product with a B_{k+1} out front; these are in lines (4.11) and (4.12), respectively. After some computation,

$$
\begin{aligned}
&V_k - V_{k+1} \\
&\geq \big(A_k - \frac{2B_k\rho_k(1-\beta_k)}{\beta_k}\big)\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \frac{\alpha_k B_k}{\beta_k(1-\beta_k)}\langle G(z^{k+1}), G(z^{k+1/2})\rangle \\
&+\frac{2\rho_k B_k}{\beta_k}\langle G(z^{k+1}), G(z^k)\rangle - \frac{\alpha_k B_k}{\beta_k}\langle G(z^k), G(z^{k+1/2})\rangle + B_{k+1}\langle G(z^k), \bar{z}^{k+1} - \bar{z}^k\rangle \\
&+\frac{B_k\rho_k}{\beta_k}\|G(z^{k+1}) - G(z^k)\|^2 + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 \\
&=\big(A_k - \frac{B_k\rho_k(1-2\beta_k)}{\beta_k}\big)\|G(z^k)\|^2 \\
&-\big(A_k - \frac{B_k\rho_k}{\beta_k}\big)\|G(z^{k+1})\|^2 + \frac{\alpha_k B_k}{\beta_k(1-\beta_k)}\langle G(z^{k+1}), G(z^{k+1/2})\rangle \qquad (4.13) \\
&-\frac{\alpha_k B_k}{\beta_k}\langle G(z^k), G(z^{k+1/2})\rangle + B_{k+1}\langle G(z^{k+1}), \bar{z}^{k+1} - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2. \qquad (4.14)
\end{aligned}
$$

Next, let's focus on the last three terms in (4.14): $B_{k+1}\langle G(z^{k+1}), \bar{z}^{k+1} - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2$. 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

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

Now we may judiciously apply these facts to the three terms in (4.14) under consideration in the following manner.

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

While this takes care of the latter three terms in lines (4.13) to (4.14), that everything else is nonnegative is a nontrivial argument. However, it directly follows the proof of Lemma 7.1 in [23], so as before we refer to their proof, and then our Lyapunov functional is also nonincreasing. ■

Theorem 13 ($O(1/k^2)$ convergence rate for FEG with moving anchor)

For the $R-$Lipschitz continuous and $\rho-$comonotone operator G where $\rho > -\frac{1}{2R}$, and $z^* \in Z_*(G), Z_*(G) := \{z^* \in \mathbb{R}^d : G(z^*) = 0\}$, 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 Theorem 12, we take $\alpha_k = 1/R, \beta_k = \frac{1}{k+1}, R_k = R, \rho_k = \rho$, 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 [23]'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 14 *In the setting of Theorem 12, 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 Theorem 12 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 Theorem 8. ■

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.

Chapter 5

Preconditioned Versions of Moving Anchor Algorithms

In this chapter, we introduce a sort of 'preconditioned' version of the previously developed moving anchor algorithms. The insight here is that the algorithms developed in previous chapters may utilize a proximal update for the anchoring step without any major modifications to the convergence theory. In particular, the descent lemmas and optimal order of convergence may be retained. The motivation for doing this is twofold: first, introducing a proximal update breaks these algorithms out of the class of deterministic algorithms where optimal complexity results are well-established. This motivates these preconditioned algorithms as objects of interest for both theoretical and computational purposes in future study.

5.1 Modified EAG-V with moving anchor

We begin with developing the proximal version of the EAG-V with moving anchor.

Definition 15 (Modified EAG-V with moving anchor) *In the setting of EAG-V with moving anchor, consider equation* (3.10) *from the proof of Theorem 6:*

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

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

$$\bar{z}^k - \bar{z}^{k+1} = -\gamma_{k+1} G(z^{k+1}) - t_k(H(\bar{z}^k) - H(\bar{z}^{k+1})), \tag{5.1}$$

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:

$$\bar{z}^{k+1} = (I + t_k H)^{-1}(\bar{z}^k + \gamma_{k+1} G(z^{k+1}) + t_k H(\bar{z}^k)). \tag{5.2}$$

This is the modified EAG-V with moving anchor.

Lemma 16 *Under the same conditions as Theorem 6 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* (5.2) *still results in a nonincreasing Lyapunov functional.*

Proof. Within the proof of Theorem 6 recall the following line:

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

Within this proof that the functional is nonincreasing, the primary change is that we must use equation (5.1) for substituting $G(z^{k+1})$. This results in

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

The term $\frac{B_{k+1}}{\gamma_{k+1}}\|\bar{z}^k - \bar{z}^{k+1}\|^2$ will be utilized elsewhere (see Theorem 6) 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 17 *The modified EAG-V algorithm with moving anchor has convergence rate* $O(1/k^2)$.

Remark 18 *While* H *may be any monotone operator, in practice one may wish to take* $H = G$.

5.2 Modified FEG with moving anchor

Definition 19 (Proximal FEG with moving anchor) *In the setting of FEG with moving anchor, consider* (4.10) *from the proof of Theorem 12:*

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

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:

$$\bar{z}^{k+1} = (I + t_k H)^{-1}(\bar{z}^k + \gamma_{k+1} G(z^{k+1}) + t_k H(\bar{z}^k)) \tag{5.3}$$

Lemma 20 *Under the same conditions as Theorem 12 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* (5.3) *still results in a nonincreasing Lyapunov functional.*

Proof. The proof proceeds in the same manner as in that of Theorem 16. 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 21 *The modified FEG algorithm with moving anchor has convergence rate $O(1/k^2)$.*

Chapter 6

Stochastic Moving Anchor EAG-V

Let G be an $R-$Lipschitz, monotone operator on $\mathbb{R}^n \times \mathbb{R}^m$, and let $N = n + m$. To develop the stochastic moving anchor EAG-V algorithm, the following additional clarifications and assumptions are necessary.

1. $\frac{1}{N}\sum_{i=1}^{N} G_i(z) = \mathbb{E}[G_\theta(z)|z] = G(z)$, or the expectation of $G(z)$ given z is $G(z)$ for θ iid on $1, \ldots, N$.

2. G has condition number $\overline{C_G}(z)$ which depends on the point z currently being evaluated by G, such that $\frac{1}{N}\sum_{i=1}^{N} \|G_i(z)\|^2 \leq \overline{C_G}(z)\|G(z)\|^2$ and $1 \leq \overline{C_G}(z) \leq N$ are true for all $z \in \mathbb{R}^n \times \mathbb{R}^m$, with the constant $\overline{K_G}(z) := N\overline{C_G}(z)$ resulting in $1 \leq \overline{K_G}(z) \leq N^2$. Note that this also gives us an a priori bound on certain variance terms. Additionally, given three indices i, j, k, (whose meaning will become apparent below), we have

$$
\begin{aligned}
Var(z) &:= \mathbb{E}[\langle G(z) - G_{i_j^k}(z), G(z) - G_{i_j^k}(z)\rangle | z] \\
&= \frac{1}{N}\sum_{i=1}^{N} \|G_i(z)\|^2 - \|G(z)\|^2 \\
&\leq (\overline{C_G}(z) - 1)\|G(z)\|^2 \qquad (6.1)
\end{aligned}
$$

no matter what values i, j, k, may take. Note that the condition number, as defined, has this property (6.1) that holds for any z; however, we are particularly interested in the behavior of $Var(z^k), Var(z^{k+1/2})$, where z^k is the $k-$th iteration of a stochastic algorithm and $z^{k+1/2}$ is a sort of half-way, 'interpolation' step. Therefore, we impose one other useful bound regarding this condition number $\overline{C_G}(z)$ as it relates to the stochastic iterates and half-iterates in the stochastic algorithm we define below.

Condition 22 *The function $\overline{C_G}$ depends on the local value z being evaluated by the operator G in such a way that the following inequalities hold for all k, where k is the iteration count of a stochastic algorithm:*

$$
\big(\overline{C_G}(z^k) - 1\big)\|G(z^k)\|^2 \leq \frac{C_1}{(k+1)^4} \qquad (6.2)
$$

$$
\mathbb{E}[\big(\overline{C_G}(z^{k+1/2}) - 1\big)\|G(z^{k+1/2})\|^2 | \bar{z}^k, z^k] \leq \frac{C_2}{(k+1)^4}, \qquad (6.3)
$$

where C_1, C_2 are fixed positive constants.

With Theorem 22 in mind, we will henceforth use the notation $C_G(z^k), K_G(z^k)$ to indicate two nonnegative, real-valued functions from $\mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}$ that behave according to (6.2), (6.3).

In particular, we note that for any z^k coming from a stochastic algorithm, we have that eventually,

$$C_G(z^k) \leq \overline{C_G}, \tag{6.4}$$

where we define $\overline{C_G}$ as the supremum of appropriate condition numbers, independent of z.

3. For all $z_1, z_2 \in \mathbb{R}^n \times \mathbb{R}^m$, $\|G_i(z_1) - G_i(z_2)\|^2 \leq R_i^2 \|z_1 - z_2\|^2$ with $R = \sqrt{\sum R_i}$.

Furthermore, define $\{i_k^1, i_k^2, i_k^3\}_{k=1}^{\infty}$ to be uniformly iid random on $\{1, \ldots, N\}$. Then the **stochastic EAG-V with moving anchor** is defined as

$$z^{k+1/2} = z^k + \frac{1}{k+2}(\bar{z}^k - z^k) - \alpha_k G_{i_k^1}(z^k) \tag{6.5}$$

$$z^{k+1} = z^k + \frac{1}{k+2}(\bar{z}^k - z^k) - \alpha_k G_{i_k^2}(z^{k+1/2}) \tag{6.6}$$

$$\bar{z}^{k+1} = \bar{z}^k + \tilde{\gamma}_{k+1} G_{i_k^3}(z^{k+1}) \tag{6.7}$$

$$\tilde{\gamma}_{k+1} = \frac{B_{k+1}}{c_{k+1}\overline{K_G}(1 + \frac{1}{\delta_k})} \tag{6.8}$$

where each $G_{i_k^j}, j = 1, 2, 3$, is assumed to be an unbiased estimator of $G(z)$, meaning that for $\xi_{i,j,k}(z) := G(z) - G_{i_k^j}(z)$, $\mathbb{E}[\xi_{i,j,k}(z)|z] = 0$. With these modifications, we may keep the update (3.4) the same as in the stochastic setting. First, we offer a lemma that clarifies the behavior of α_k; our primary modification to the original version of this lemma, due to [46], is an updated bound for α_0.

Lemma 23 *The sequence α_k (2.6) starting at $\alpha_0 \in (0, \eta)$, $\eta := \min\{\frac{3}{4R\sqrt{\overline{K_G}}}, \frac{1}{\sqrt{2}R}\}$, monotonically decreases to a positive limit. In particular, when $\sqrt{\overline{K_G}} = 1$, we recover Theorem 4.*

Remark 24 *Squeezing down the interval where α_0 may start is a choice made to force the positivity of the term $(1 - \alpha_k^2 R^2 - \beta_k)$ in (6.43) for ease of analysis. With a different choice*

of β_k, one may wish to modify the upper bound η by choosing the second term to be $\dfrac{\sqrt{1-\hat{\beta}}}{R}$, where $\hat{\beta} := \sup \beta_k$ is not equal to 1.

Proof. We assume $R = 1$ and $\sqrt{\overline{\overline{K_G}}} = 1$ without loss of generality. We may rewrite (2.6) as

$$\alpha_k - \alpha_{k+1} = \frac{\alpha_k^3}{(k+1)(k+3)(1-\alpha_k^2)}. \tag{6.9}$$

Suppose that we have established that for some $N \geq 0,\ 0 < \alpha_N < \rho$ for some $\rho \in (0,1)$ that satisfies

$$\eta := \frac{1}{2}\Big(\frac{1}{N+1} + \frac{1}{N+2}\Big)\frac{\rho^2}{1-\rho^2} < 1. \tag{6.10}$$

(6.10) holds for all $N \geq 0$ if $\rho < \frac{3}{4}$. We now show that with (6.10),

$$\alpha_N > \alpha_{N+1} > \cdots > \alpha_{N+k} > (1-\eta)\alpha_N \text{ for all } k > 0,$$

allowing us to obtain α_k as a monotonically decreasing sequence to some α such that $\alpha \geq (1-\eta)\alpha_N$. It suffices to prove $(1-\eta)\alpha_N < \alpha_{N+k} < \rho$ for all $k \geq 0$, as (6.9) indicates that $\{\alpha_k\}_{k=0}^{\infty}$ is decreasing. Use induction on k to prove that $\alpha_{N+k} \in ((1-\eta)\alpha_N, \rho)$. The case $k = 0$ is trivial. Now suppose that $(1-\eta)\alpha_N < \alpha_{N+j} < \rho$ holds true for $j = 0, \ldots, k$. Then by (6.9), for each $0 \leq j \leq k$ we have

$$\begin{aligned}
0 < \alpha_N - \alpha_{N+k+1} &< \sum_{j=0}^{k} \frac{1}{(N+j+1)(N+j+3)} \frac{\rho^2 \alpha_N}{1-\rho^2} \\
&< \frac{\rho^2 \alpha_N}{1-\rho^2} \sum_{j=0}^{\infty} \frac{1}{(N+j+1)(N+j+3)} \\
&= \frac{\rho^2 \alpha_N}{1-\rho^2} \frac{1}{2}\Big(\frac{1}{N+1}\frac{1}{N+2}\Big) = \eta \alpha_N,
\end{aligned}$$

which gives $(1-\eta)\alpha_N < \alpha_{N+k+1} < \alpha_N < \rho$, completing the induction. ∎

We need a careful Lyapunov analysis to handle the newly introduced stochasticity. Rather than focusing on making a *nonincreasing* Lyapunov functional as was previously the goal, we aim to control how negative the differences between subsequent terms may be via variances. The analysis here is inspired by the analogous stochastic Lyapunov lemma in [23].

Lemma 25 (Stochastic Lyapunov Functional, Moving Anchor EAG-V) *Consider the stochastic EAG-V with moving anchor* (6.5), (6.6), (6.7), (6.8), (3.4) *along with conditions 1, 2, 3, and* $\{i_k^1, i_k^2, i_k^3\}_{k=1}^{\infty}$ *as previously described. Suppose we are given the sequences* $\{A_k\}_{k=0}^{\infty}, \{B_k\}_{k=0}^{\infty}$ *as described in Theorem 2 and the sequence* $\{\alpha_k\}_{k=0}^{\infty}$ *described in Theorem 23. Define the stochastic Lyapunov functional as*

$$V_k = A_k\|G(z^k)\|^2 + B_k\langle G(z^k), z^k - \bar{z}^k\rangle + c_k\|z^* - \bar{z}^k\|^2, \tag{6.11}$$

Then, (6.11) *satisfies the following:*

$$\mathbb{E}[V_k - V_{k+1}|\bar{z}^k, z^k] \geq$$
$$-2A_k\alpha_k RVar(z^k) - \frac{2A_k}{1-\beta_k}\alpha_k R\mathbb{E}[Var(z^{k+1/2})|\bar{z}^k, z^k]$$

Proof. With our Lyapunov functional (6.11) in mind, we derive the following useful relations:

$$z^k - z^{k+1} = \beta_k(z^k - \bar{z}^k) + \alpha_k G_{i_k^2}(z^{k+1/2}) \tag{6.12}$$

$$z^{k+1/2} - z^{k+1} = \alpha_k\big(G_{i_k^2}(z^{k+1/2}) - G_{i_k^1}(z^k)\big) \tag{6.13}$$

$$\bar{z}^k - z^{k+1} = (1-\beta_k)(\bar{z}^k - z^k) + \alpha_k G_{i_k^2}(z^{k+1/2}) \tag{6.14}$$

$$\bar{z}^k - \bar{z}^{k+1} = -\tilde{\gamma}_{k+1} G_{i_k^3}(z^{k+1}). \tag{6.15}$$

(6.12) is z^k subtract (6.5), (6.13) is (6.6) subtract (6.5), (6.14) is $\bar{z}^k$ subtract (6.6), and (6.15) is (6.7) rearranged. As already evidenced, much of this proof will parallel the previous descending Lyapunov lemmas from the deterministic cases, but our end goal is to capture how negative the differences can be rather than force positivity. We introduce a nonnegative inner product to begin the process of simplifying:

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

After some additional computation utilizing (6.12) through (6.15), we obtain

$$\begin{aligned} V_k - V_{k+1} \geq \\ & \underbrace{A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle}_{\text{I}} \\ & \underbrace{+\alpha_k B_{k+1}\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2})\rangle}_{\text{I}} \underbrace{+c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2}_{\text{II}} \\ & \underbrace{+B_{k+1}\tilde{\gamma}_{k+1}\langle G_{i_k^3}(z^{k+1}), G(z^{k+1})\rangle}_{\text{II}}. \end{aligned}$$

From here, we will deal with I and II separately. We will deal with II first. To begin, let's analyze the inner product contained within II under expectation:

$$\mathbb{E}[B_{k+1}\tilde{\gamma}_{k+1}\langle G_{i_k^3}(z^{k+1}), G(z^{k+1})\rangle|\bar{z}^k, z^k] \tag{6.16}$$

$$=B_{k+1}\tilde{\gamma}_{k+1}\mathbb{E}\big[\mathbb{E}[\langle G_{i_k^3}(z^{k+1}), G(z^{k+1})\rangle|z^{k+1}, \bar{z}^k, z^k]\big|\bar{z}^k, z^k\big] \tag{6.17}$$

$$=B_{k+1}\tilde{\gamma}_{k+1}\mathbb{E}[\|G(z^{k+1})\|^2|\bar{z}^k, z^k] \tag{6.18}$$

$$\geq B_{k+1}\tilde{\gamma}_{k+1}\mathbb{E}\Big[\frac{\|G_{i_k^3}(z^{k+1})\|^2}{K_G(z^{k+1})}|\bar{z}^k, z^k\Big] \tag{6.19}$$

$$=B_{k+1}\mathbb{E}\Big[\frac{\|\bar{z}^k - \bar{z}^{k+1}\|^2}{\tilde{\gamma}_{k+1}K_G(z^{k+1})}|\bar{z}^k, z^k\Big]. \tag{6.20}$$

From (6.16) to (6.17) to (6.18), we apply the law of iterated expectation to get $G(z^{k+1})$. Knowing z^{k+1}, we recall that $G_{i_k^3}$ is an unbiased estimator of G to get (6.18). The inequality (6.19) results from 2, and (6.20) results from (6.15).

Thus after taking expectation, II changes into the following:

$$\mathbb{E}[\text{II}|\bar{z}^k, z^k] \tag{6.21}$$

$$\geq\mathbb{E}[c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\|z^* - \bar{z}^{k+1}\|^2 + \frac{B_{k+1}}{\tilde{\gamma}_{k+1}K_G}\|\bar{z}^k - \bar{z}^{k+1}\|^2|\bar{z}^k, z^k] \tag{6.22}$$

$$\geq\mathbb{E}[c_k\|z^* - \bar{z}^k\|^2 - c_{k+1}\big((1+\delta_k)\|z^* - \bar{z}^k\|^2 + (1+\frac{1}{\delta_k})\|\bar{z}^k - \bar{z}^{k+1}\|^2\big)$$

$$+\frac{B_{k+1}}{\tilde{\gamma}_{k+1}K_G(z^{k+1})}\|\bar{z}^k - \bar{z}^{k+1}\|^2|\bar{z}^k, z^k] \tag{6.23}$$

where (6.23) is an application of Cauchy-Schwartz to $\|z^* - \bar{z}^{k+1}\|$. Because $\tilde{\gamma}_{k+1} = \frac{1}{K_G}\frac{B_{k+1}}{c_{k+1}(1+\frac{1}{\delta_k})}$, we find

$$\mathbb{E}[\text{II}|\bar{z}^k, z^k] \geq 0,$$

and now we are left with I:

$$\mathbb{E}[V_k - V_{k+1}|\bar{z}^k, z^k] \geq \mathbb{E}[I|\bar{z}^k, z^k]$$

$$= \mathbb{E}\big[A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle$$

$$+ \alpha_k B_{k+1}\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2})\rangle\big|\bar{z}^k, z^k\big].$$

First, we note that

$$\|G(z^{k+1/2}) - G(z^{k+1})\|^2 \leq R^2\|z^{k+1/2} - z^{k+1}\|^2$$

$$= R^2\alpha_k^2\|G_{i_k^2}(z^{k+1/2}) - G_{i_k^1}(z^k)\|^2 \tag{6.24}$$

$$\Leftrightarrow$$

$$-A_k\|G_{i_k^2}(z^{k+1/2}) - G_{i_k^1}(z^k)\|^2 + \frac{A_k}{R^2\alpha_k^2}\|G(z^{k+1/2}) - G(z^{k+1})\|^2 \leq 0 \tag{6.25}$$

by $R-$smoothness, so that

$$\mathbb{E}[\mathrm{I}|\bar{z}^k, z^k] \geq$$

$$\mathbb{E}\big[A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle$$

$$+ \alpha_k B_{k+1}\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2})\rangle - A_k\|G_{i_k^2}(z^{k+1/2}) - G_{i_k^1}(z^k)\|^2$$

$$+ \frac{A_k}{R^2\alpha_k^2}\|G(z^{k+1/2}) - G(z^{k+1})\|^2\big|\bar{z}^k, z^k\big] \tag{6.26}$$

$$= \mathbb{E}\big[A_k\|G(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k) - G(z^{k+1})\rangle$$

$$+ \alpha_k B_{k+1}\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2})\rangle$$

$$- A_k\big(\|G_{i_k^2}(z^{k+1/2})\|^2 - 2\langle G_{i_k^2}(z^{k+1/2}), G_{i_k^1}(z^k)\rangle + \|G_{i_k^1}(z^k)\|^2\big)$$

$$+ \frac{A_k}{R^2\alpha_k^2}\big(\|G(z^{k+1/2})\|^2 - 2\langle G(z^{k+1/2}), G(z^{k+1})\rangle + \|G(z^{k+1})\|^2\big)\big|\bar{z}^k, z^k\big] \tag{6.27}$$

To be clear, (6.26) is I subtract (6.25), and (6.27) is (6.26) with the terms introduced from (6.25) expanded. Rearrange (6.27) to visualize cancellations and groupings of terms:

$$
\begin{aligned}
&\mathbb{E}\big[A_k\|G(z^k)\|^2 - A_k\|G_{i_k^1}(z^k)\|^2 - A_{k+1}\|G(z^{k+1})\|^2 + \frac{A_k}{R^2\alpha_k^2}\|G(z^{k+1})\|^2 \\
&+(\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1})\langle G_{i_k^2}(z^{k+1/2}), G(z^{k+1})\rangle - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k)\rangle \\
&-A_k\|G_{i_k^2}(z^{k+1/2}))\|^2 + \frac{A_k}{R^2\alpha_k^2}\|G(z^{k+1/2})\|^2 + 2A_k\langle G_{i_k^2}(z^{k+1/2}), G_{i_k^1}(z^k)\rangle \\
&-\frac{2A_k}{R^2\alpha_k^2}\langle G(z^{k+1/2}), G(z^{k+1})\rangle\big|\bar{z}^k, z^k\big] \qquad (6.28)\\
\geq\ &\mathbb{E}\big[(\frac{A_k}{R^2\alpha_k^2} - A_k K_G(z^{k+1/2}))\|G(z^{k+1/2})\|^2 + (\frac{A_k}{R^2\alpha_k^2} - A_{k+1})\|G(z^{k+1})\|^2 \\
&+2A_k\langle G_{i_k^2}(z^{k+1/2}), G_{i_k^1}(z^k)\rangle - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k)\rangle \\
&+(\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1})\langle G_{i_k^2}(z^{k+1/2}), G(z^{k+1})\rangle - \frac{2A_k}{R^2\alpha_k^2}\langle G(z^{k+1/2}), G(z^{k+1})\rangle\big|\bar{z}^k, z^k\big], \\
&\qquad (6.29)
\end{aligned}
$$

where the first two terms of (6.28) cancel by the law of iterated expectation applied to $\|G_{i_k^1}(z^k)\|^2$ and the first term in (6.29) comes from applying 2 to $-A_k\|G_{i_k^2}(z^{k+1/2}))\|^2 + \frac{A_k}{R^2\alpha_k^2}\|G(z^{k+1/2})\|^2$. Now, we may apply the law of iterated expectation to modify the two terms in the second line of (6.29):

$$
\begin{aligned}
&\mathbb{E}\big[2A_k\langle G_{i_k^2}(z^{k+1/2}), G_{i_k^1}(z^k)\rangle - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k)\rangle\big|\bar{z}^k, z^k\big] \\
=&\mathbb{E}\big[\mathbb{E}[2A_k\langle G_{i_k^2}(z^{k+1/2}), G_{i_k^1}(z^k)\rangle - \frac{\alpha_k B_k}{\beta_k}\langle G_{i_k^2}(z^{k+1/2}), G(z^k)\rangle|\bar{z}^k, z^k, i_k^1]\big|\bar{z}^k, z^k\big] \\
=&\mathbb{E}\big[2A_k\langle G(z^{k+1/2}), G_{i_k^1}(z^k)\rangle - \frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k)\rangle\big|\bar{z}^k, z^k\big]. \qquad (6.30)
\end{aligned}
$$

With observation (6.30) under our belts, we continue by introducing some terms at the tail end of an updated (6.29):

$$
\begin{aligned}
&\mathbb{E}\big[(\frac{A_k}{R^2\alpha_k^2} - A_k K_G(z^{k+1/2}))\|G(z^{k+1/2})\|^2 + (\frac{A_k}{R^2\alpha_k^2} - A_{k+1})\|G(z^{k+1})\|^2 \\
&\quad + 2A_k\langle G(z^{k+1/2}), G_{i_k^1}(z^k)\rangle - \frac{\alpha_k B_k}{\beta_k}\langle G(z^{k+1/2}), G(z^k)\rangle \\
&\quad + (\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1})\langle G_{i_k^2}(z^{k+1/2}), G(z^{k+1})\rangle - \frac{2A_k}{R^2\alpha_k^2}\langle G(z^{k+1/2}), G(z^{k+1})\rangle \\
&\quad + 2A_k\langle G(z^{k+1/2}), G(z^k)\rangle - 2A_k\langle G(z^{k+1/2}), G(z^k)\rangle \\
&\quad + (\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1})\langle G(z^{k+1}), G(z^{k+1/2})\rangle \\
&\quad - (\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1})\langle G(z^{k+1}), G(z^{k+1/2})\rangle \big| \bar{z}^k, z^k\big] \\
\geq &\mathbb{E}\big[(\frac{A_k}{R^2\alpha_k^2} - A_k K_G(z^{k+1/2}))\|G(z^{k+1/2})\|^2 + (\frac{A_k}{R^2\alpha_k^2} - A_{k+1})\|G(z^{k+1})\|^2 \qquad (6.31) \\
&\quad + (2A_k - \frac{\alpha_k B_k}{\beta_k})\langle G(z^{k+1/2}), G(z^k)\rangle \qquad (6.32) \\
&\quad + (\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1} - \frac{2A_k}{R^2\alpha_k^2})\langle G(z^{k+1}), G(z^{k+1/2})\rangle \qquad (6.33) \\
&\quad + 2A_k\langle G(z^{k+1/2}), G_{i_1^k}(z^k) - G(z^k)\rangle \qquad (6.34) \\
&\quad + (\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1})\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\rangle \big| \bar{z}^k, z^k\big] \qquad (6.35)
\end{aligned}
$$

Let us momentarily ignore the latter two terms (6.34), (6.35). Note (6.32) is zero by the definition of A_k, and for the other coefficients,

$$\frac{A_k}{R^2\alpha_k^2} - A_k K_G(z^{k+1/2}) = \frac{A_k(1-\alpha_k^2R^2K_G(z^{k+1/2}))}{\alpha_k^2R^2} \tag{6.36}$$

$$A_{k+1} = \frac{A_k(1-\alpha_k^2R^2-\beta_k^2)}{(1-\alpha_k^2R^2)(1-\beta_k)^2}$$

$$\implies \frac{A_k}{\alpha_k^2R^2} - A_{k+1} = \frac{A_k(1-\alpha_k^2R^2-\beta_k)^2}{\alpha_k^2R^2(1-\alpha_k^2R^2)(1-\beta_k)^2} \tag{6.37}$$

$$\alpha_k B_{k+1} + \frac{\alpha_k B_k}{\beta_k} - \frac{2A_k}{\alpha_k^2R^2} = \frac{2A_k}{1-\beta_k} - \frac{2A_k}{\alpha_k^2R^2}$$

$$= -\frac{2A_k(1-\alpha_k^2R^2-\beta_k)}{\alpha_k^2R^2(1-\beta_k)} \tag{6.38}$$

which, if we continue ignoring (6.34) and (6.35) while substituting in (6.36), (6.37), and (6.38), yields

$$V_k - V_{k+1} \geq$$

$$\mathbb{E}[\mathrm{I}|\bar{z}^k, z^k] \geq$$

$$\mathbb{E}\big[\big(\frac{A_k(1-\alpha_k^2R^2K_G(z^{k+1/2}))}{\alpha_k^2R^2}\big)\|G(z^{k+1/2})\|^2 + \big(\frac{A_k(1-\alpha_k^2R^2-\beta_k)^2}{\alpha_k^2R^2(1-\alpha_k^2R^2)(1-\beta_k)^2}\big)\|G(z^{k+1})\|^2 \tag{6.39}$$

$$- \frac{2A_k(1-\alpha_k^2R^2-\beta_k)}{\alpha_k^2R^2(1-\beta_k)}\langle G(z^{k+1}), G(z^{k+1/2})\rangle\big|\bar{z}^k, z^k\big]. \tag{6.40}$$

Our aim is to complete the square via Young's inequality to demonstrate the nonnegativity of these terms. A slight complicating factor exists in the extra $K_G(z^{k+1/2})$ in a coefficient within (6.39), which we deal with in the following way.

$$
\begin{aligned}
&\big(\frac{A_k(1-\alpha_k^2R^2K_G(z^{k+1/2}))}{\alpha_k^2R^2}\big)\|G(z^{k+1/2})\|^2+\big(\frac{A_k(1-\alpha_k^2R^2-\beta_k)^2}{\alpha_k^2R^2(1-\alpha_k^2R^2)(1-\beta_k)^2}\big)\|G(z^{k+1})\|^2\\
&-\frac{2A_k(1-\alpha_k^2R^2-\beta_k)}{\alpha_k^2R^2(1-\beta_k)}\langle G(z^{k+1}),G(z^{k+1/2})\rangle\\
=&\big(1-\alpha_k^2R^2K_G(z^{k+1/2})\big)\|G(z^{k+1/2})\|^2+\frac{(1-\alpha_k^2R^2-\beta_k)^2}{(1-\alpha_k^2R^2)(1-\beta_k)^2}\|G(z^{k+1})\|^2\\
&-\frac{2(1-\alpha_k^2R^2-\beta_k)}{(1-\beta_k)}\langle G(z^{k+1}),G(z^{k+1/2})\rangle \qquad (6.41)\\
=&\|\sqrt{1-\alpha_k^2R^2K_G(z^{k+1/2})}G(z^{k+1/2})\|^2+\frac{(1-\alpha_k^2R^2-\beta_k)^2}{(1-\alpha_k^2R^2)}\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2\\
&-2\Big\langle\frac{(1-\alpha_k^2R^2-\beta_k)}{(1-\beta_k)\sqrt{1-\alpha_k^2R^2K_G(z^{k+1/2})}}G(z^{k+1}),\sqrt{1-\alpha_k^2R^2K_G(z^{k+1/2})}G(z^{k+1/2})\Big\rangle\\
&+(1-\alpha_k^2R^2-\beta_k)^2\Big\|\frac{G(z^{k+1})}{(1-\beta_k)\sqrt{1-\alpha_k^2R^2K_G(z^{k+1/2})}}\Big\|^2\\
&-(1-\alpha_k^2R^2-\beta_k)^2\Big\|\frac{G(z^{k+1})}{(1-\beta_k)\sqrt{1-\alpha_k^2R^2K_G(z^{k+1/2})}}\Big\|^2\\
\geq&\frac{(1-\alpha_k^2R^2-\beta_k)^2}{(1-\alpha_k^2R^2)}\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2-\frac{(1-\alpha_k^2R^2-\beta_k)^2}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2 \qquad (6.42)\\
=&(1-\alpha_k^2R^2-\beta_k)\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2\left(\frac{(1-\alpha_k^2R^2-\beta_k)}{(1-\alpha_k^2R^2)}-\frac{(1-\alpha_k^2R^2-\beta_k)}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}\right) \qquad (6.43)
\end{aligned}
$$

where at (6.41), we drop the common factor of $\frac{A_k}{\alpha_k^2R^2}$ since it is positive and won't affect the latter computations. The inequality (6.42) is due to the 'Peter-Paul' variant of Young's inequality. Continuing on, we see that

$$
\begin{aligned}
=&(1-\alpha_k^2R^2-\beta_k)\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2 \\
&\cdot\left(1-\frac{(1-\alpha_k^2R^2)}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}+\beta_k\Big(\frac{1}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}-\frac{1}{(1-\alpha_k^2R^2)}\Big)\right) \\
=&(1-\alpha_k^2R^2-\beta_k)\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2 \\
&\cdot\left(\frac{\alpha_k^2R^2(1-K_G(z^{k+1/2}))}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}+\frac{\alpha_k^2R^2\beta_k(K_G(z^{k+1/2})-1)}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))(1-\alpha_k^2R^2)}\right) \\
>&(1-\alpha_k^2R^2-\beta_k)\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2 \\
&\cdot\left(\frac{\alpha_k^2R^2\beta_k(1-K_G(z^{k+1/2}))}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}+\frac{\alpha_k^2R^2\beta_k(K_G(z^{k+1/2})-1)}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))(1-\alpha_k^2R^2)}\right) \qquad (6.44) \\
>&(1-\alpha_k^2R^2-\beta_k)\Big\|\frac{G(z^{k+1})}{(1-\beta_k)}\Big\|^2\left(\frac{\alpha_k^2R^2\beta_k(1-K_G(z^{k+1/2}))}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}+\frac{\alpha_k^2R^2\beta_k(K_G(z^{k+1/2})-1)}{(1-\alpha_k^2R^2K_G(z^{k+1/2}))}\right) \qquad (6.45) \\
=&0,
\end{aligned}
$$

and we note (6.44) and (6.45) are due to $0<\beta_k<1$ and $1<\frac{1}{(1-\alpha_k^2R^2)}$, respectively. For clarity, the term $(1-\alpha_k^2R^2-\beta_k)$ is positive because of the starting point of α_0 in Theorem 23, so there are no issues with bringing this factor to the front of the expression for our analysis. Bringing back our other terms, this demonstrates that

$$
\begin{aligned}
&V_k - V_{k+1} \geq \\
&\mathbb{E}[\mathrm{I}|\bar{z}^k, z^k] \geq \\
&\quad + 2A_k\langle G(z^{k+1/2}), G_{i_1^k}(z^k) - G(z^k)\rangle && (6.46)\\
&\quad + (\frac{\alpha_k B_k}{\beta_k} + \alpha_k B_{k+1})\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\rangle \big| \bar{z}^k, z^k\big] && (6.47)\\
&= + 2A_k\langle G(z^{k+1/2}), G_{i_1^k}(z^k) - G(z^k)\rangle \\
&\quad + \frac{2A_k}{1-\beta_k}\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\rangle \big| \bar{z}^k, z^k\big]
\end{aligned}
$$

For (6.46), we note that

$$
0 = \langle -G\big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^k)\big), \mathbb{E}[G_{i_k^1}(z^k) - G(z^k)|\bar{z}^k, z^k]\rangle, \tag{6.48}
$$

which allows us to compute

$$
\begin{aligned}
&\big|\mathbb{E}[\langle G(z^{k+1/2}), G_{i_k^1}(z^k) - G(z^k)\rangle|\bar{z}^k, z^k]\big| \\
=&\big|\mathbb{E}[\langle G(z^{k+1/2}) - G\big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^k)\big), G_{i_k^1}(z^k) - G(z^k)\rangle|\bar{z}^k, z^k]\big| \\
\leq&\mathbb{E}\big[\big\|G(z^{k+1/2}) - G\big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^k)\big)\big\| \cdot \big\|G_{i_k^1}(z^k) - G(z^k)\big\||\bar{z}^k, z^k\big] \\
\leq&\mathbb{E}\big[R\big\|z^{k+1/2} - \big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^k)\big)\big\| \cdot \big\|G_{i_k^1}(z^k) - G(z^k)\big\||\bar{z}^k, z^k\big] \\
=&\mathbb{E}[R\alpha_k\|G_{i_k^1}(z^k) - G(z^k)\|^2|\bar{z}^k, z^k] \\
=&R\alpha_k Var(z^k),
\end{aligned} \tag{6.49}
$$

via an application of Cauchy-Schwartz, the Lipschitz property, and the definition of the algorithm, where $V_i(z) := \mathbb{E}[\langle G_i(z) - G(z), G_i(z) - G(z)\rangle|z]$ is the variance of the difference. Similarly, for (6.47) we obtain

$$
\begin{aligned}
&\big|\mathbb{E}[\langle G(z^{k+1}), G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\rangle | \bar{z}^k, z^k]\big| \\
=&\big|\mathbb{E}[\langle G(z^{k+1}) - G\big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2})\big), G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\rangle | \bar{z}^k, z^k]\big| \\
\leq&\mathbb{E}\big[\big\|G(z^{k+1}) - G\big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2})\big)\big\| \cdot \big\|G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\big\| | \bar{z}^k, z^k\big] \\
\leq&\mathbb{E}\big[R\big\|z^{k+1} - \big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2})\big)\big\| \cdot \big\|G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\big\| | \bar{z}^k, z^k\big] \\
=&\mathbb{E}\Big[\mathbb{E}\big[R\big\|z^{k+1} - \big(z^k + \beta_k(\bar{z}^k - z^k) - \alpha_k G(z^{k+1/2})\big)\big\| \\
&\cdot \big\|G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\big\| | z^{k+1/2}, \bar{z}^k, z^k\big] | \bar{z}^k, z^k\Big] \\
=&\mathbb{E}[R\alpha_k \|G_{i_k^2}(z^{k+1/2}) - G(z^{k+1/2})\|^2 | z^{k+1/2}, \bar{z}^k, z^k] \\
=&R\alpha_k \mathbb{E}[Var(z^{k+1/2}) | \bar{z}^k, z^k] \qquad (6.50)
\end{aligned}
$$

by the law of iterated expectation and reasoning similar to that of (6.46). Equipped with (6.49) and (6.50), we get

$$
V_k - V_{k+1} \geq -2A_k\alpha_k R Var(z^k) - \frac{2A_k}{1-\beta_k}\alpha_k R \mathbb{E}[Var(z^{k+1/2}) | \bar{z}^k, z^k] \qquad (6.51)
$$

and the lemma is proved. ■

To proceed towards convergence, we need the following.

Definition 26 *The filtration*

$$
\mathcal{F}^k = \sigma(z^0, \bar{z}^0, z^1, \bar{z}^1, \ldots, z^k, \bar{z}^k, i_1^0, i_2^0, i_3^0, \ldots, i_1^k, i_2^k, i_3^k)
$$

represents the history of of iterates, anchors, and choices of component i up through the current step k.

Theorem 27 (Supermartingale Convergence Theorem [8]) *Let $P^k, J^k,$ and W^k be positive sequences adapted to $\mathcal{F}^k$, and suppose W^k is summable with probability* 1. *If*

$$\mathbb{E}[P^{k+1}|\mathcal{F}^k] + J^k \leq P^k + W^k,$$

then with probability 1, *P^k converges to a $[0,\infty)-$valued random variable and $\sum_{j=1}^{\infty} J^k < \infty$.*

Now, we may apply Theorem 22 to satisfy the conditions of Theorem 27.

Lemma 28 (Summability of Variances) *Consider the stochastic Lyapunov functional V_k* (6.11) *discussed in Theorem 25 along with conditions* (1), (2), (3), *the choice $\beta_k = \frac{1}{k+2}$ made in Theorem 3, and Theorem 22. Given* (6.51), *the extraneous sequence of terms*

$$2A_k\alpha_k R\Big(Var(z^k) + \frac{\mathbb{E}[Var(z^{k+1/2})|\bar{z}^k, z^k]}{1-\beta_k}\Big)$$

is summable.

Proof. Because of (6.1) in (2) and Theorem 22, it is sufficient to demonstrate that

$$\sum_{k=0}^{\infty} 2A_k\alpha_k R\Big((C_G(z^k)-1)\|G(z^k)\|^2 + \frac{1}{1-\beta_k}\mathbb{E}[(\overline{C_G}(z^{k+1/2})-1)\|G(z^{k+1/2})\|^2|\bar{z}^k, z^k]\Big) \quad (6.52)$$

is finite. By construction, α_k is a nonnegative term bounded above by the choice α_0, so the following bound for each summand exists if we substitute in the definition of A_k made in Theorem 3:

$$
\begin{aligned}
&\alpha_0^2 R(k+1)(k+2)\Big((C_G(z^k)-1)\|G(z^k)\|^2 \\
&\quad + \frac{1}{1-\beta_k}\mathbb{E}[\big(\overline{C_G}(z^{k+1/2})-1\big)\|G(z^{k+1/2})\|^2|\bar{z}^k, z^k]\Big) \\
&\leq \alpha_0^2 R(k+1)(k+2)\Big(\frac{C_1}{(k+1)^4} + \frac{C_2}{(k+1)^4(1-\beta_k)}\Big) \qquad (6.53)\\
&= \frac{\alpha_0^2 R(k+1)(k+2)}{1-\beta_k}\Big(\frac{C_1(1-\beta_k)+C_2}{(k+1)^4}\Big) \qquad (6.54)\\
&< 2\alpha_0^2 R(k+1)(k+2)\Big(\frac{C_1+C_2}{(k+1)^4}\Big) \qquad (6.55)
\end{aligned}
$$

where (6.53) results from Theorem 22, (6.54) results from rationalizing the denominator with $(1-\beta_k)$, and the last inequality (6.55) is a result of the facts that $1-\beta_k < 1$, $\frac{1}{1-\beta_k} < 2$. For the summation, these facts result in

$$
\begin{aligned}
&\sum_{k=0}^{\infty} 2A_k\alpha_k R\Big((C_G(z^k)-1)\|G(z^k)\|^2 + \frac{1}{1-\beta_k}\mathbb{E}[\big(\overline{C_G}(z^{k+1/2})-1\big)\|G(z^{k+1/2})\|^2|\bar{z}^k, z^k]\Big) \\
&< 2\alpha_0^2 R(C_1+C_2)\sum_{k=0}^{\infty}\frac{k^2+3k+2}{(k+1)^4} \\
&= 2\alpha_0^2 R(C_1+C_2)\sum_{k=0}^{\infty}\frac{k^2}{k^4} + \text{ other terms more summable than } \frac{1}{k^2} < \infty.
\end{aligned}
$$

■

Theorem 29 (Stochastic Lyapunov Functional Convergence)

Consider the stochastic Lyapunov functional V_k (6.11) *along with conditions* (1), (2), (3), *the choice* $\beta_k = \frac{1}{k+2}$ *made in Theorem 3, and Theorem 22. Then with probability* 1, V_k *converges to a nonnegative, finite-valued random variable.*

Proof. Apply Theorem 27 and Theorem 28 with

$W^k = 2A_k\alpha_k R\Big(Var(z^k) + \frac{\mathbb{E}[Var(z^{k+1/2})|\bar{z}^k, z^k])}{1-\beta_k}\Big), P^k = V_k, J^k = 0.$ ■ Equipped with these results, we may now state the convergence of stochastic moving anchor methods.

Theorem 30 (Stochastic Moving Anchor EAG-V Convergence)

Consider the stochastic moving anchor EAG-V algorithm (6.5), (6.6), (6.7) *along with conditions* (1), (2), (3), *the choice* $\beta_k = \frac{1}{k+2}$ *made in Theorem 3, and Theorem 22 for the Lyapunov function* (6.11). *If* $c_\infty \geq \frac{1}{\alpha_\infty}$, *then the stochastic moving anchor EAG-V algorithm converges with rate*

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

where $sum(k-1) := \sum_{j=0}^{k-1} 2A_j\alpha_j R\big(Var(z^j) + \frac{\mathbb{E}[Var(z^{j+1/2})|\bar{z}^j, z^j]}{1-\beta_j}\big)$.

Proof. By (6.51), we see that

$$\begin{aligned} V_k &\leq V_{k-1} + 2A_{k-1}\alpha_{k-1}RVar(z^{k-1}) + \frac{2A_{k-1}}{1-\beta_{k-1}}\alpha_{k-1}R\mathbb{E}[Var(z^{(k-1)+1/2})|\bar{z}^{k-1}, z^{k-1}] \\ &\leq V_0 + \text{sum}(k-1) \\ &= \alpha_0\|G(z^0)\|^2 + c_0\|z^0 - z_*\|^2 + \text{sum}(k-1) \\ &\leq (\alpha_0 R + c_0)\|z^0 - z_*\|^2 + \text{sum}(k-1). \end{aligned}$$

Going in the opposite direction, we see that

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

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:

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

Finally, one divides both sides by the constant $\frac{\alpha_\infty}{4}(k+1)(k+2)$ to achieve the result. ■

Chapter 7

Numerical Experiments

7.1 Deterministic Examples

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 negative comonotone FEG example, δ_k is chosen to be $\exp(k^2) - 1$.

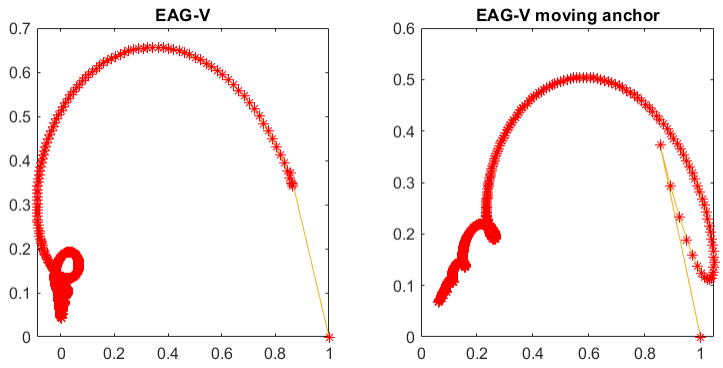

Figure 7.1: The first two thousand iterations of the EAG algorithm with varying step-size, or EAG-V, compared to the first two thousand iterations of the moving anchor EAG-V algorithm.

Figure 7.1 compares the iterations of EAG-V with a fixed anchor to the iterations for the moving anchor EAG-V. Figure 7.1, Figure 7.2, and Figure 7.3 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 $(0, 0)$.

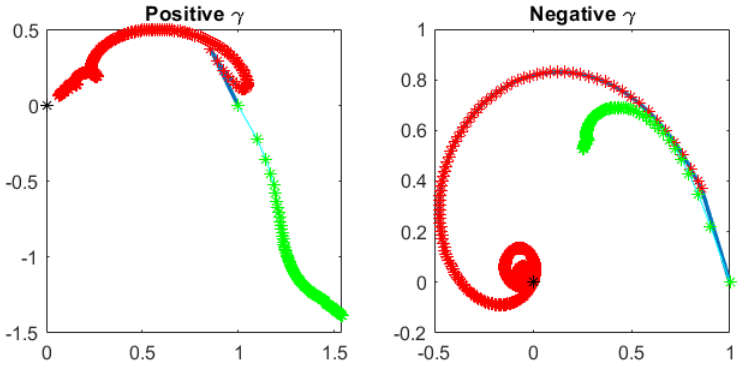

Figure 7.2: The two moving anchor EAG-V variants compared in red, along with their anchors in green.

Figure 7.2 compares, via the same function as Figure 7.1, 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.

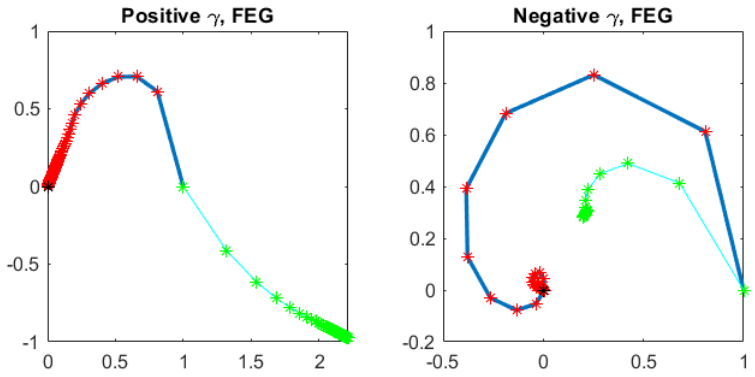

Figure 7.3: The two moving anchor FEG variants compared in red, along with their anchors in green.

Figure 7.3 compares the two moving anchor versions of the FEG method, in the same manner as the comparison shown in Figure 7.2: red dots are the algorithm updates, green dots are the anchor updates, and the function is the 'almost bilinear' one previously described. In [23], 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.

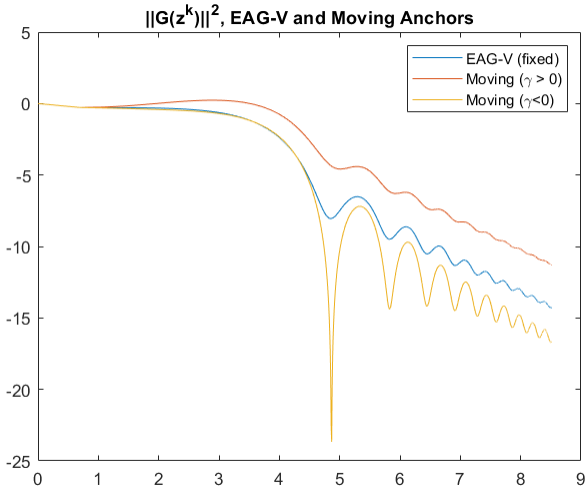

Figure 7.4: Comparison of the grad-norm squared of three EAG-V variants of interest on a toy 'almost bilinear' problem.

Figure 7.4 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 occurrsed 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 7.5 captures the error of FEG across all three anchor variants in a numerical example that is explicitly negative comonotone with a nonconvex-nonconcave objective:

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

with $L : \mathbb{R}^2 \to \mathbb{R}, R = 1, \rho = -1/3$ 1−smooth and $-1/3$-negative comonotone. Interestingly, in this scenario a variant of the positive γ_k moving anchor algorithm is the fastest when δ_k is

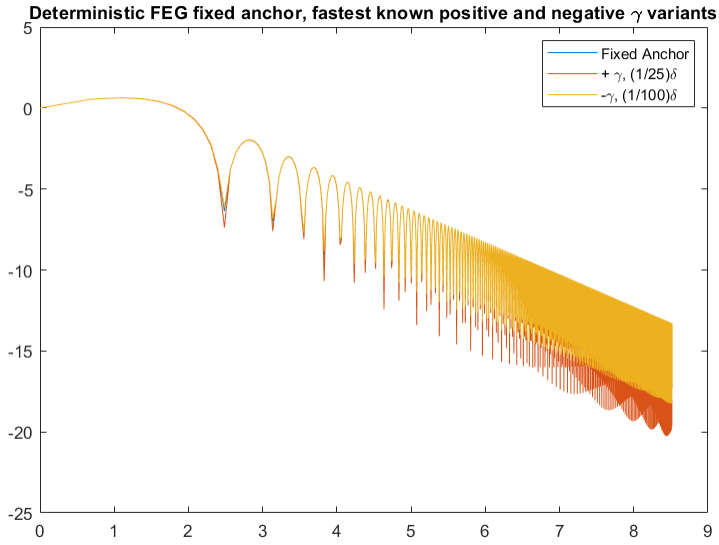

Figure 7.5: Comparison of the errors of three FEG variants in a nonconvex-nonconcave setting. Note the positive γ with δ scaled by $1/25$ converges fastest.

scaled by $1/25$, in sharp contrast to the result displayed in Figure 7.4. The fixed anchor and the negative γ_k values seem to almost coincide. This result suggests that different problem settings offer different optimal anchoring choices.

Finally, Figure 7.6 and Figure 7.7 compare three different monotone FEG variants on a particular nonlinear game that was studied extensively in [6]:

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

where $Q=A^TA$ 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, while the second player attempt to maximize their gain among m strategies with strategy j chosen with probability y_j $(j = 1, \ldots, m)$. The payoff is a quadratic function that depends on the strategy of both players. This was implemented following the 3 operator splitting scheme in [9], with parameter $\lambda = 0.25$

For this example, we used FEG fixed and moving anchor variants in the monotone (that is, $\rho = 0$) setting of the algorithm. For the high dimensional setting, we compute $20,000$ iterations and view the log of the grad norm squared of the fixed anchor versus the positive and negative γ_k moving anchor variants. Here, $m = 2500, n = 500$, resulting in an operator with 9 million entries. For the low dimensional setting, we compare the same algorithms but only compare this on $8,000$ iterations, as the convergence behavior in this lower dimensional setting is more quickly distinguished. In the lower dimensional setting, $m = 25, n = 5$. In both settings, the positive γ_k variant is the fastest (ie, most accelerated) algorithm by a significant margin, even with different random seeds selected across test runs. This is a significant contrast to the EAG-V toy example, also a convex-concave problem, where the $-\gamma_k$ variant of the moving anchor is the fastest algorithm. Taken together, both results are promising for the moving anchor framework, but suggest that more theoretical work is necessary to understand the acceleration mechanism offered by anchoring variants, and what variant will be fastest in a given problem.

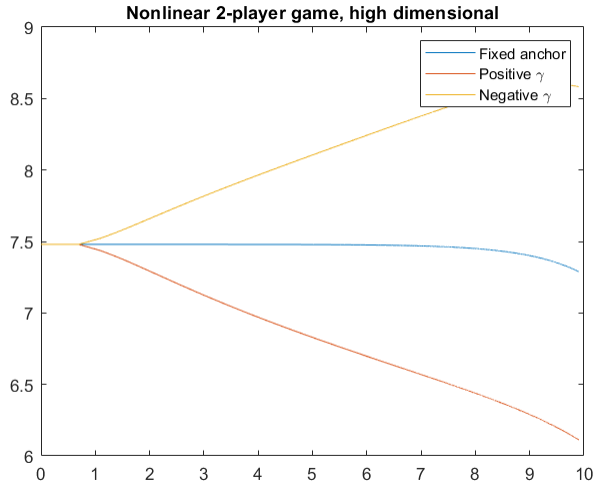

Figure 7.6: High dimensional nonlinear game.

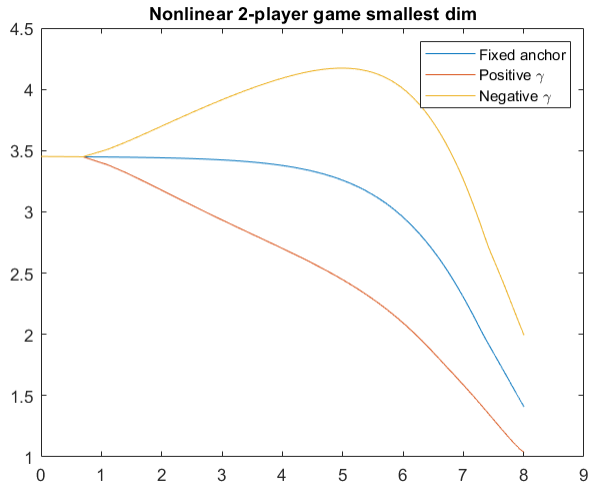

Figure 7.7: Low dimensional nonlinear game.

7.2 Stochastic Examples

In this section, the choices made for α_0 and c_0 change significantly to accommodate our stochastic theory. Specifically, we choose $\alpha_0 = 0.9(3/4)(1/R)(1/\sqrt{K_G})$ and then $c_0 = (1.01)(4/3)e^{\frac{\pi^2}{6}} R\sqrt{K_G}$.

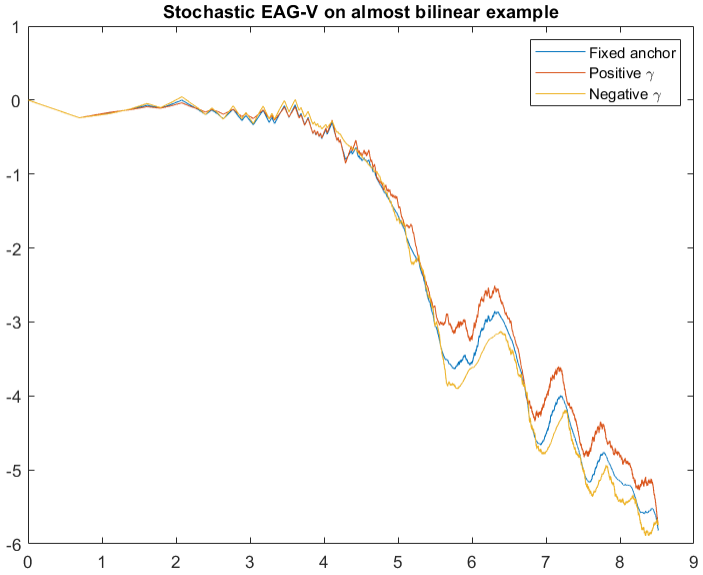

Figure 7.8: Comparison of the grad-norm squared of three stochastic EAG-V variants of interest on a toy 'almost bilinear' problem.

In Figure 7.8, one observes that, on a toy example, the behavior of the $-\gamma_k$ and $+\gamma_k$ variants of the moving anchor seem to parallel the deterministic setting of the same problem [1], in that the negative version is the fastest and the positive version is the slowest.

Next, Figure 7.9 showcases a stochastic moving anchor on a smooth nonconvex nonconcave operator that results in a negative comonotone operator [1], [23] using the theory we develop in this work. While this example is outside the scope of the theory developed,

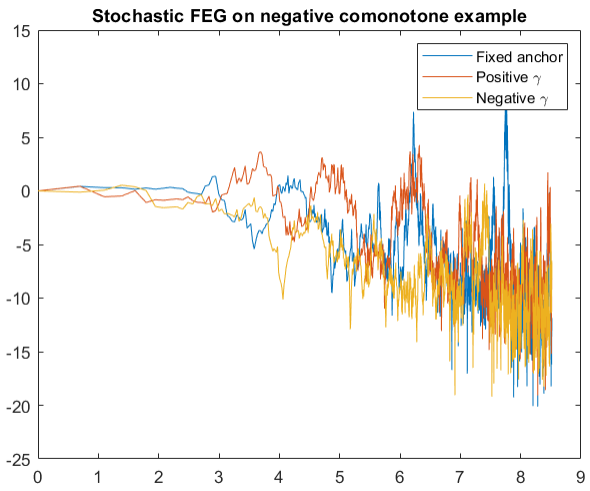

Figure 7.9: Comparison of the grad-norm squared of three negative comonotone stochastic FEG variants of interest on a test problem.

we note that the starting constants c_0 and α_0 were informed by the stochastic theory and across all three variants, this still seems to result in the average of the squared gradient norm decreasing significantly. Our numerical result is significant, then, as it suggests that our framework hints at a theory for such problems, even with a fixed anchor, despite the obvious noise.

In Figure 7.10, the 2 player nonlinear game studied in Figure 7.6 and Figure 7.7 is studied with our stochastic moving anchor EAG-V algorithm and compared to a stochastic fixed anchor EAG-V algorithm. We set $m = 2$ and $n = 48$, for a more moderately sized problem with a well-behaved condition number.

A few remarks are in order. This game was previously studied in both [6] and [1], where in the latter case the authors encountered favorable results using deterministic moving anchor algorithms. However, in both high and low dimensional examples, the choice

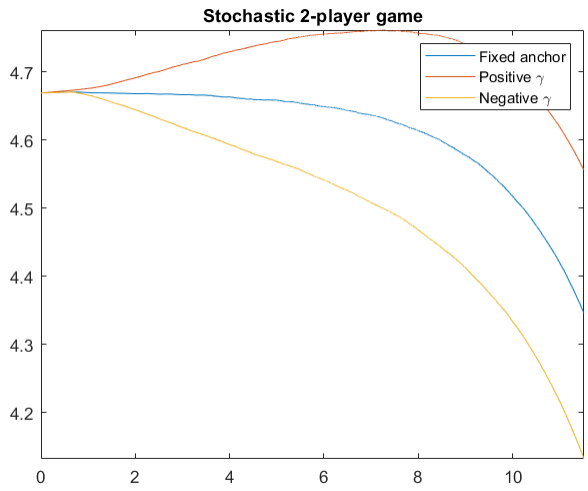

Figure 7.10: Comparison of the grad-norm squared of three stochastic EAG-V anchoring variants on a nonlinear game.

of positive γ_k resulted in the most significant acceleration beating the fixed anchor in the deterministic case. Here in our stochastic variant, we encounter the opposite behavior: a moving anchor variant indeed provides the most significant acceleration, but it is the negative γ_k that does so. One possibility is that the EAG-V algorithm structure somehow favors the negative γ_k variant of moving anchor algorithms, while the FEG algorithm structure - whether on a convex-concave problem or a nonconvex-nonconcave problem Secondly, although our theory calls for setting K_G to be the operator's condition number times the dimensions in the domain of the objective function, setting $K_G = 1$ resulted in significant numerical improvements, especially as even with $m = 2, n = 48$, the condition number in this type of problem can be very large. Finally, the choice of the parameter λ relating to the three operator splitting structure differs a large amount from that used in [1], and it is unclear why this is the case.

Chapter 8

Conclusion

The moving anchor variants of anchored acceleration methods retain optimal convergence rates and also demonstrate superior-to-comparable numerical performance with parameter tuning. The optimal order of convergence is obtained across different problem settings, from convex-concave to negative co-monotone problems and also in stochastic convex-concave 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. However, future work may explore further the exact conditions and parameters for optimal choice of anchor in one's anchored extragradient algorithm. 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 moving anchor with a sort of preconditioned anchoring step in both the convex-concave and negative co-monotone problem settings and demonstrate its convergence. The last theoretical contribution made in this work is the development and

implementation of a stochastic moving anchor algorithm, developed for convex-concave functions but with some numerical indications of a theory for nonconvex-nonconcave functions. Some immediate considerations for future work are the following:

- Numerical examples exploring the 'proximal' moving anchor variants;
- More practical numerical experiments such as image processing or neural network performance;
- Parallelized/asynchronous implementations of moving anchor algorithms for further computational ease;
- Tighter analyses of the convergence of $-\gamma_k$ variants of the moving anchor;
- Extensions of the moving anchor algorithmic framework to other anchoring and Halpern adjacent techniques such as [44], [43];
- Theoretical analyses that determine what problem/algorithm settings enable the fastest convergence of anchored algorithms;
- Analyses that clarify the discrepancy between the behavior of the various moving anchor variants parametrized by γ_k and δ_k.

More ambitious, long-term goals may include extending these methods (and their associated convergence guarantees) to broader classes of problems such as the (weak) Minty Variational Inequality classes, and their stochastic extensions to such problem classes.

Bibliography

[1] James K Alcala, Yat Tin Chow, and Mahesh Sunkula. Moving anchor extragradient methods for smooth structured minimax problems. *arXiv preprint arXiv:2308.12359*, 2023.

[2] 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. In *International conference on artificial intelligence and statistics*, pages 2863–2873. PMLR, 2020.

[3] Léon Bottou. *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[4] 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.

[5] 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.

[6] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[7] Sayantan Choudhury, Eduard Gorbunov, and Nicolas Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Patrick L Combettes and Jean-Christophe Pesquet. Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

[9] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25:829–858, 2017.

[10] Jelena Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pages 1428–1451. PMLR, 2020.

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

[12] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2017.

[13] Konstantinos Emmanouilidis, René Vidal, and Nicolas Loizou. Stochastic extragradient with random reshuffling: Improved convergence for variational inequalities. In *International Conference on Artificial Intelligence and Statistics*, pages 3682–3690. PMLR, 2024.

[14] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems, 2020.

[15] 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*, 27:2672–2680, 2014.

[16] Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022.

[17] Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pages 366–402. PMLR, 2022.

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

[19] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[20] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[21] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[22] Jongmin Lee and Ernest K Ryu. Accelerating value iteration with anchoring. *arXiv preprint arXiv:2305.16569*, 2023.

[23] Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. 2021.

[24] 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.

[25] Felix Lieder. On the convergence rate of the halpern-iteration. *Optimization letters*, 15(2):405–418, 2021.

[26] Liya Liu and Xiaolong Qin. An accelerated stochastic extragradient-like algorithm with new stepsize rules for stochastic variational inequalities. *Computers & Mathematics with Applications*, 163:117–135, 2024.

[27] 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.

[28] 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.

[29] 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. *arXiv preprint arXiv:1807.02629*, 2018.

[30] 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.

[31] 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.

[32] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[33] Arkadi S Nemirovsky. On optimality of krylov's information when solving linear operator equations. *Journal of Complexity*, 7(2):121–130, 1991.

[34] Arkadi S Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

[35] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pages 543–547, 1983.

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

[37] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2):1–35, 2021.

[38] 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*, 28:845–848, 1980.

[39] R Tyrrell Rockafellar. Monotone operators associated with saddle-functions and minimax problems. In *Proceedings of Symposia in Pure Mathematics*, volume 18, pages 241–250. American Mathematical Society, 1970.

[40] Ernest K Ryu, Kun Yuan, and Wotao Yin. Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems and gans. 2019.

[41] Chaobing Song, Zhengyuan Zhou, Yichao Zhou, Yong Jiang, and Yi Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33:14303–14314, 2020.

[42] Jaewook J Suh, Jisun Park, and Ernest K Ryu. Continuous-time analysis of anchor acceleration. *arXiv preprint arXiv:2304.00771*, 2023.

[43] Quoc Tran-Dinh. The connection between nesterov's accelerated methods and halpern fixed-point iterations. *arXiv preprint arXiv:2203.04869*, 2022.

[44] Quoc Tran-Dinh and Yang Luo. Halpern-type accelerated and splitting algorithms for monotone inclusions. *arXiv preprint arXiv:2110.08150*, 2021.

[45] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[46] 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*, 139:12098–12109, 18–24 Jul 2021.

[47] TaeHo Yoon and Ernest K Ryu. Accelerated minimax algorithms flock together. *arXiv preprint arXiv:2205.11093*, 2022.

[48] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. *Advances in Neural Information Processing Systems*, 30, 2017.