Optimal Accelerated Extra Anchored Gradient Methods for Smooth Convex-Concave Minimax Optimization

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

October 1, 2024

Abstract

In this paper, we propose and analyze a family of accelerated Extra Anchored Gradient (EAG) algorithms designed to solve smooth convex-concave minimax problems. Our approach achieves an optimal $O(1/k^2)$ convergence rate in terms of the squared gradient norm, which improves upon the previously established $O(1/k)$ rates of extragradient-type methods. We also provide a matching lower bound, proving the optimality of our method. Our analysis highlights that different suboptimality measures lead to distinct acceleration mechanisms, offering insights into the convergence behavior of minimax optimization algorithms. Experimental results validate the effectiveness of EAG in accelerating convergence compared to existing methods.

Keywords: Convex-concave optimization, minimax problems, accelerated gradient methods, convergence rates.

1 Introduction

Minimax optimization problems, where one seeks to minimize a function over one set of variables while maximizing over another, have become a central topic in optimization and machine learning. These problems arise in various applications, such as adversarial training [Goodfellow et al.(2015), Madry et al.(2018)], generative adversarial networks (GANs) [Goodfellow et al.(2014)], and robust optimization. In these settings, the optimization dynamics are driven by the interplay between two competing objectives.

Formally, a minimax problem can be written as:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}},, \underset{\mathbf{y}\in\mathbb{R}^m}{\text{maximize}},, \mathbf{L}(\mathbf{x},\mathbf{y}) \tag{1}$$

where $\mathbf{L}(\mathbf{x},\mathbf{y})$ is typically a smooth convex-concave function. Solving such problems efficiently, especially in large-scale settings, has led to the development of various gradient-based algorithms. Among these, extragradient methods have gained particular attention due to their effectiveness in handling the non-symmetric nature of minimax problems.

Despite significant theoretical progress, many existing methods only achieve a suboptimal $O(1/k)$ rate with respect to the squared gradient norm. In this work, we aim to advance the state of minimax optimization by introducing an accelerated Extra Anchored Gradient (EAG) method, which achieves a faster $O(1/k^2)$ convergence rate. This acceleration closes the gap between theoretical guarantees and practical performance in minimax optimization.

Prior works on minimax optimization often consider compact domains X, Y for $\mathbf{x}, \mathbf{y}$ and use the *duality gap*

$$\mathrm{Err}_{\mathrm{gap}}(\mathbf{x},\mathbf{y}) := \sup_{\widetilde{\mathbf{y}}\in Y} \mathbf{L}(\mathbf{x},\widetilde{\mathbf{y}}) - \inf_{\widetilde{\mathbf{x}}\in X} \mathbf{L}(\widetilde{\mathbf{x}},\mathbf{y})$$

to quantify suboptimality of algorithms' iterates in solving (1). However, while it is a natural analog of minimization error for minimax problems, the duality gap can be difficult to measure directly in practice, and it is unclear how to generalize the notion to non-convex-concave problems.

In contrast, the squared gradient magnitude $\|\nabla \mathbf{L}(\mathbf{x}, \mathbf{y})\|^2$, when $\mathbf{L}$ is differentiable, is a more directly observable value for quantifying suboptimality. Moreover, the notion is meaningful for differentiable non-convex-concave minimax games. Interestingly, very few prior works have analyzed convergence rates on the gradient norm for minimax problems, and the optimal convergence rate or corresponding algorithms were hitherto unknown.

Main results In this work, we introduce the *extra anchored gradient (EAG)* algorithms for smooth convex-concave minimax problems and establish an accelerated $\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \mathcal{O}(R^2/k^2)$ rate, where R is the Lipschitz constant of $\nabla \mathbf{L}$. The rate improves upon the $\mathcal{O}(R^2/k)$ rates of prior algorithms and is the first $\mathcal{O}(R^2/k^2)$ rate in this setup. We then provide a matching $\Omega(R^2/k^2)$ complexity lower bound for gradient-based algorithms and thereby establish optimality of EAG.

Beyond establishing the optimal complexity, our results provide the following observations. First, different suboptimality measures lead to materially different acceleration mechanisms, since reducing the duality gap is done optimally by the extragradient algorithm [Nemirovski(2004), Nemirovsky(1992)]. Also, since our optimal accelerated convergence rate is on the non-ergodic last iterate, neither averaging nor keeping track of the best iterate is necessary for optimally reducing the gradient magnitude in the deterministic setup.

Contributions Our primary contributions are as follows:

- We introduce a new class of Extra Anchored Gradient (EAG) algorithms that attain an optimal $O(1/k^2)$ convergence rate for smooth convex-concave minimax problems.
- We establish a matching lower bound, proving the optimality of our proposed method in the minimax setting.
- We demonstrate that our algorithm outperforms existing methods in terms of both theoretical complexity and empirical performance.

Organization The remainder of this paper is organized as follows. In Section 2, we provide a brief overview of prior work on minimax optimization. Section 3 introduces our EAG algorithm and its convergence analysis. Section 4 presents experimental results validating the efficacy of our method. Finally, Section 5 concludes with a discussion of potential future research directions.

1.1 Preliminaries and notation

We say a saddle function $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is convex-concave if $\mathbf{L}(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x} \in \mathbb{R}^n$ for all fixed $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{L}(\mathbf{x}, \mathbf{y})$ is concave in $\mathbf{y} \in \mathbb{R}^m$ for all fixed $\mathbf{x} \in \mathbb{R}^n$. We say $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a saddle point of $\mathbf{L}$ if $\mathbf{L}(\mathbf{x}^\star, \mathbf{y}) \leq \mathbf{L}(\mathbf{x}^\star, \mathbf{y}^\star) \leq \mathbf{L}(\mathbf{x}, \mathbf{y}^\star)$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Solutions to the minimax problem (1) are defined to be saddle points of $\mathbf{L}$. For notational conciseness, write $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. When $\mathbf{L}$ is differentiable, define the *saddle operator* of $\mathbf{L}$ at $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ by

$$\mathbf{G}_{\mathbf{L}}(\mathbf{z}) = \begin{bmatrix} \nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}, \mathbf{y}) \end{bmatrix} \tag{2}$$

(When clear from the context, we drop the subscript $\mathbf{L}$.) The saddle operator is *monotone* [Rockafellar(1970)], i.e., $\langle \mathbf{G}(\mathbf{z}_1) - \mathbf{G}(\mathbf{z}_2), \mathbf{z}_1 - \mathbf{z}_2 \rangle \geq 0$ for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n \times \mathbb{R}^m$. We say $\mathbf{L}$ is R-smooth if $\mathbf{G}_\mathbf{L}$ is R-Lipschitz continuous. Note that $\nabla \mathbf{L} \neq \mathbf{G}_\mathbf{L}$ due to the sign change in the $\mathbf{y}$ gradient, but $\|\nabla \mathbf{L}\| = \|\mathbf{G}_\mathbf{L}\|$, and we use the two forms interchangeably. Because $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ is a saddle point of $\mathbf{L}$ if and only if $0 = \mathbf{G}_\mathbf{L}(\mathbf{z}^\star)$, the squared gradient magnitude is a natural measure of suboptimality at a given point for smooth convex-concave problems.

1.2 Prior work

Extragradient-type algorithms. The first main component of our proposed algorithm is the extragradient (EG) algorithm of [Korpelevich(1977)]. EG and its variants, including the algorithm of [Popov(1980)], have been studied in the context of saddle point and variational inequality problems and have appeared in the mathematical programming literature [Solodov & Svaiter(1999), Tseng(2000), Noor(2003), Censor et al.(2011), Lyashko et al.(2011), Malitsky & Semenov(2014), Malitsky(2015), Malitsky(2020)]. More recently in the machine learning literature, similar ideas such as optimism [Chiang et al.(2012), Rakhlin & Sridharan(2013a)], prediction [Yadav et al.(2018)], and negative momentum [Gidel et al.(2019), Zhang et al.(2020)] have been presented and used in the context of multi-player games [Daskalakis et al.(2011), Rakhlin & Sridharan(2013b), Syrgkanis et al.(2015), Antonakopoulos et al.(2021)] and GANs [Gidel et al.(2018), Mertikopoulos et al.(2019), Liang & Stokes(2019), Peng et al.(2020)].

$\mathcal{O}(R/k)$ rates on duality gap. For minimax problems with an R-smooth $\mathbf{L}$ and bounded domains for $\mathbf{x}$ and $\mathbf{y}$, [Nemirovski(2004)] presented the mirror-prox algorithm generalizing EG and established ergodic $\mathcal{O}(R/k)$ convergence rates on $\mathrm{Err}_{\mathrm{gap}}$. [Nesterov(2007), Monteiro & Svaiter(2010), Monteiro & Svaiter(2011)] extended the $\mathcal{O}(R/k)$ complexity analysis to the case of unbounded domains. [Mokhtari et al.(2020b)] showed that the optimistic descent converges at $\mathcal{O}(R/k)$ rate with respect to $\mathrm{Err}_{\mathrm{gap}}$. Since there exists $\Omega(R/k)$ complexity lower bound on $\mathrm{Err}_{\mathrm{gap}}$ for black-box gradient-based minimax optimization algorithms [Nemirovsky(1992), Nemirovski(2004)], in terms of duality gap, these algorithms are order-optimal.

Convergence rates on squared gradient norm. Using standard arguments (e.g. [Solodov & Svaiter(1999), Lemma 2.3]), one can show $\min_{i=0,\ldots,k} \|\mathbf{G}(\mathbf{z}^i)\|^2 \leq \mathcal{O}(R^2/k)$ convergence rate of EG, provided that $\mathbf{L}$ is R-smooth. [Ryu et al.(2019)] showed that optimistic descent algorithms also attain $\mathcal{O}(R^2/k)$ convergence in terms of the best iterate and proposed simultaneous gradient descent with *anchoring*, which pulls iterates toward the initial point $\mathbf{z}^0$, and established $\mathcal{O}(R^2/k^{2-2p})$ convergence rates in terms of squared gradient norm of the last iterate (where $p > \frac{1}{2}$ is an algorithm parameter; see Section A). Notably, anchoring resembles the Halpern iteration [Halpern(1967), Lieder(2020)], which was used in [Diakonikolas(2020)] to develop a regularization-based algorithm with near-optimal (optimal up to logarithmic factors) complexity with respect to the gradient norm of the last iterate. Anchoring turns out to be the second main component of the acceleration; combining EG steps with anchoring, we obtain the optimal last-iterate convergence rate of $\mathcal{O}(R^2/k^2)$.

Structured minimax problems. For structured minimax problems of the form

$$\mathbf{L}(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle - g(\mathbf{y})$$

where f, g are convex and $\mathbf{A}$ is a linear operator, primal-dual splitting algorithms [Chambolle & Pock(2011), Condat(2013), Vũ(2013), Yan(2018), Ryu & Yin(2021)] and Nesterov's smoothing technique [Nesterov(2005a), Nesterov(2005b)] have also been extensively studied [Chen et al.(2014), He & Monteiro(2016)]. Notably, when g is of "simple" form, Neterov's smoothing framework achieves an accelerated rate $\mathcal{O}\left(\frac{\|\mathbf{A}\|}{k} + \frac{L_f}{k^2}\right)$ on duality gap. Additionally, [Chambolle & Pock(2016)] have shown that splitting algorithms can achieve $\mathcal{O}(1/k^2)$ or linear convergence rates under appropriate strong convexity and smoothness assumptions on f and g, although they rely on proximal operations. [Kolossoski & Monteiro(2017), Hamedani & Aybat(2018), Zhao(2019), Alkousa et al.(2020)] generalized these accelerated algorithms to the setting where the coupling term $\langle \mathbf{Ax}, \mathbf{y} \rangle$ is replaced by non-bilinear convex-concave function $\Phi(\mathbf{x}, \mathbf{y})$.

Complexity lower bounds. [Ouyang & Xu(2021)] presented a $\Omega\left(\frac{\|\mathbf{A}\|}{k} + \frac{L_f}{k^2}\right)$ complexity lower bound on duality gap for gradient-based algorithms solving bilinear minimax problems with proximable g, establishing optimality of Nesterov's smoothing. [Zhang et al.(2019)] presented lower bounds for strongly-convex-strongly-concave problems. [Golowich et al.(2020)] proved that with the narrower class of 1-*SCLI* algorithms, which includes EG but not EAG, the squared gradient norm of the last iterate cannot be reduced beyond $\mathcal{O}(R^2/k)$ in R-smooth minimax problems. These approaches are aligned with the information-based complexity analysis, introduced in [Nemirovsky & Yudin(1983)] and thoroughly studied in [Nemirovsky(1991), Nemirovsky(1992)] for the special case of linear equations.

Other problem setups. [Nesterov(2009)] and [Nedić & Ozdaglar(2009)] proposed subgradient algorithms for non-smooth minimax problems. Stochastic minimax and variational inequality problems were studied in [Nemirovski et al.(2009), Juditsky et al.(2011), Lan(2012), Ghadimi & Lan(2012), Ghadimi & Lan(2013), Chen et al.(2014), Chen et al.(2017), Hsieh et al.(2019)]. Strongly monotone variational inequality problems or strongly-convex-strongly-concave minimax problems were studied in [Tseng(1995), Nesterov & Scrimali(2011), Gidel et al.(2018), Mokhtari et al.(2020a), Lin et al.(2020b), Wang & Li(2020), Zhang et al.(2020), Azizian et al.(2020)]. Recently, minimax problems with objectives that are either strongly convex or nonconvex in one variable were studied in [Rafique et al.(2018), Thekumparampil et al.(2019), Jin et al.(2019), Nouiehed et al.(2019), Ostrovskii et al.(2020), Lin et al.(2020a), Lin et al.(2020b), Lu et al.(2020), Wang & Li(2020), Yang et al.(2020), Chen et al.(2021)]. Minimax optimization of composite objectives with smooth and nonsmooth-but-proximable convex-concave functions were studied in [Tseng(2000), Csetnek et al.(2019), Malitsky & Tam(2020), Bùi & Combettes(20

2 Accelerated algorithms: Extra anchored gradient

We now present two accelerated EAG algorithms that are qualitatively very similar but differ in the choice of step-sizes. The two algorithms present a tradeoff between the simplicity of the step-size and the simplicity of the convergence proof; one algorithm has a varying step-size but a simpler convergence proof, while the other algorithm has a simpler constant step-size but has a more complicated proof.

2.1 Description of the algorithms

The proposed extra anchored gradient (EAG) algorithms have the following general form:

$$\begin{aligned} \mathbf{z}^{k+1/2} &= \mathbf{z}^k + \beta_k(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^k) \\ \mathbf{z}^{k+1} &= \mathbf{z}^k + \beta_k(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^{k+1/2}) \end{aligned} \tag{3}$$

for $k \geq 0$, where $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^m$ is the starting point. We use $\mathbf{G}$ defined in (2) rather than describing the $\mathbf{x}$- and $\mathbf{y}$- updates separately to keep the notation concise. We call $\alpha_k > 0$ *step-sizes* and $\beta_k \in [0, 1)$ *anchoring coefficients*. Note that when $\beta_k = 0$, EAG coincides with the unconstrained extragradient algorithm.

The simplest choice of $\{\alpha_k\}_{k\geq 0}$ is the constant one. Together with the choice $\beta_k = \frac{1}{k+2}$ (which we clarify later), we get the following simpler algorithm.

EAG with constant step-size (EAG-C)

$$\begin{aligned} \mathbf{z}^{k+1/2} &= \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha \mathbf{G}(\mathbf{z}^k) \\ \mathbf{z}^{k+1} &= \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha \mathbf{G}(\mathbf{z}^{k+1/2}) \end{aligned}$$

where $\alpha > 0$ is fixed.

Theorem 1. *Assume $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an R-smooth convex-concave function with a saddle point $\mathbf{z}^\star$. Assume $\alpha > 0$ satisfies*

$$\begin{aligned} 1 - 3\alpha R - \alpha^2 R^2 - \alpha^3 R^3 &\geq 0 \\ 1 - 8\alpha R + \alpha^2 R^2 - 2\alpha^3 R^3 &\geq 0 \end{aligned} \tag{4}$$

Then EAG-C converges with rate

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \frac{4(1 + \alpha R + \alpha^2 R^2)}{\alpha^2(1 + \alpha R)} \frac{\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(k+1)^2}$$

for $k \geq 0$.

Corollary 1. *In the setup of Theorem 1, $\alpha \in \left(0, \frac{1}{8R}\right]$ satisfies (4), and the particular choice $\alpha = \frac{1}{8R}$ yields*

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \frac{260 R^2 \|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(k+1)^2}$$

for $k \geq 0$.

While EAG-C is simple in its form, its convergence proof (presented in the appendix) is complicated. Furthermore, the constant 260 in Corollary 1 seems large and raises the question of whether it could be reduced. These issues, to some extent, are addressed by the following alternative version of EAG.

EAG with varying step-size (EAG-V)

$$\mathbf{z}^{k+1/2} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^k)$$
$$\mathbf{z}^{k+1} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^{k+1/2})$$

where $\alpha_0 \in \left(0, \frac{1}{R}\right)$ and

$$\begin{aligned}\alpha_{k+1} &= \frac{\alpha_k}{1-\alpha_k^2 R^2}\left(1 - \frac{(k+2)^2}{(k+1)(k+3)}\alpha_k^2 R^2\right) \\ &= \alpha_k\left(1 - \frac{1}{(k+1)(k+3)}\frac{\alpha_k^2 R^2}{1-\alpha_k^2 R^2}\right)\end{aligned} \tag{5}$$

for $k \geq 0$.

As the recurrence relation (5) may seem unfamiliar, we provide the following lemma describing the behavior of the resulting sequence.

Lemma 1. *If $\alpha_0 \in \left(0, \frac{3}{4R}\right)$, then the sequence $\{\alpha_k\}_{k\geq 0}$ of (5) monotonically decreases to a positive limit. In particular, when $\alpha_0 = \frac{0.618}{R}$, we have $\lim_{k\to\infty} \alpha_k \approx \frac{0.437}{R}$.*

We now state the convergence results for EAG-V.

Theorem 2. *Assume $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is an R-smooth convex-concave function with a saddle point $\mathbf{z}^\star$. Assume $\alpha_0 \in \left(0, \frac{3}{4R}\right)$, and define $\alpha_\infty = \lim_{k\to\infty} \alpha_k$. Then EAG-V converges with rate*

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \frac{4\left(1 + \alpha_0\alpha_\infty R^2\right)}{\alpha_\infty^2}\frac{\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(k+1)(k+2)}$$

for $k \geq 0$.

Corollary 2. *EAG-V with $\alpha_0 = \frac{0.618}{R}$ satisfies*

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \frac{27R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(k+1)(k+2)}$$

for $k \geq 0$.

2.2 Proof outline

We now outline the convergence analysis for EAG-V, whose proof is simpler than that of EAG-C. The key ingredient of the proof is a Lyapunov analysis with a nonincreasing Lyapunov function, the V_k of the following lemma.

Lemma 2. *Let $\{\beta_k\}_{k\geq 0} \subseteq (0,1)$ and $\alpha_0 \in \left(0, \frac{1}{R}\right)$ be given. Define the sequences $\{A_k\}_{k\geq 0}, \{B_k\}_{k\geq 0}$ and $\{\alpha_k\}_{\geq 0}$ by the recurrence relations*

$$A_k = \frac{\alpha_k}{2\beta_k} B_k \tag{6}$$
$$B_{k+1} = \frac{B_k}{1-\beta_k} \tag{7}$$

$$\alpha_{k+1} = \frac{\alpha_k \beta_{k+1}(1 - \alpha_k^2 R^2 - \beta_k^2)}{\beta_k(1-\beta_k)(1-\alpha_k^2 R^2)} \tag{8}$$

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

$$V_k := A_k \|\mathbf{G}(\mathbf{z}^k)\|^2 + B_k \langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \rangle \tag{9}$$

for EAG iterations in (3) is nonincreasing.

In Lemma 2, the choice of $\beta_k = \frac{1}{k+2}$ leads to $B_k = k+1$, $A_k = \frac{\alpha_k(k+2)(k+1)}{2}$, and (5). Why the Lyapunov function of Lemma 2 leads to the convergence guarantee of Theorem 2 may not be immediately obvious. The following proof provides the analysis.

Proof of Theorem 2. Let $\beta_k = \frac{1}{k+2}$ as specified by the definition of EAG-V. By Lemma 2, the quantity V_k defined by (9) is nonincreasing in k. Therefore,

$$V_k \leq \cdots \leq V_0 = \alpha_0 \|\mathbf{G}(\mathbf{z}^0)\|^2 \leq \alpha_0 R^2 \|\mathbf{z}^0 - \mathbf{z}^\star\|^2$$

Next, we have

$$\begin{aligned}
V_k = \; & A_k \|\mathbf{G}(\mathbf{z}^k)\|^2 + B_k \langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \rangle \\
\overset{(a)}{\geq} \; & A_k \|\mathbf{G}(\mathbf{z}^k)\|^2 + B_k \langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^\star - \mathbf{z}^0 \rangle \\
\overset{(b)}{\geq} \; & A_k \|\mathbf{G}(\mathbf{z}^k)\|^2 - \frac{A_k}{2}\|\mathbf{G}(\mathbf{z}^k)\|^2 - \frac{B_k^2}{2A_k}\|\mathbf{z}^0 - \mathbf{z}^\star\|^2 \\
\overset{(c)}{=} \; & \frac{\alpha_k}{4}(k+1)(k+2)\|\mathbf{G}(\mathbf{z}^k)\|^2 \\
& - \frac{k+1}{\alpha_k(k+2)}\|\mathbf{z}^0 - \mathbf{z}^\star\|^2 \\
\overset{(d)}{\geq} \; & \frac{\alpha_\infty}{4}(k+1)(k+2)\|\mathbf{G}(\mathbf{z}^k)\|^2 - \frac{1}{\alpha_\infty}\|\mathbf{z}^0 - \mathbf{z}^\star\|^2
\end{aligned}$$

where (a) follows from the monotonicity inequality $\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^\star \rangle \geq 0$, (b) follows from Young's inequality, (c) follows from plugging in $A_k = \frac{\alpha_k(k+1)(k+2)}{2}$ and $B_k = k+1$, and (d) follows from Lemma 1 ($\alpha_k \downarrow \alpha_\infty$). Reorganize to get

$$\begin{aligned}
\frac{\alpha_\infty}{4}(k+1)(k+2)\|\mathbf{G}(\mathbf{z}^k)\|^2 &\leq V_k + \frac{1}{\alpha_\infty}\|\mathbf{z}^0 - \mathbf{z}^\star\|^2 \\
&\leq \left(\alpha_0 R^2 + \frac{1}{\alpha_\infty}\right)\|\mathbf{z}^0 - \mathbf{z}^\star\|^2
\end{aligned}$$

and divide both sides by $\frac{\alpha_\infty}{4}(k+1)(k+2)$. □

2.3 Discussion of further generalizations

The algorithms and results of Sections 2.1 and 2.2 remain valid when we replace $\mathbf{G}$ with an R-Lipschitz continuous monotone operator; neither the definition of the EAG algorithms nor any part

of the proofs of Theorems 1 and 2 utilize properties of saddle functions beyond the monotonicity of their subdifferentials.

For EAG-C, the step-size conditions (4) in Theorem 1 can be relaxed to accommodate larger values of α. However, we do not pursue such generalizations to keep the already complicated and arduous analysis of EAG-C manageable. Also, larger step-sizes are more naturally allowed in EAG-V and Theorem 2. Finally, although (4) holds for values of α up to $\frac{0.1265}{R}$, we present a slightly smaller range $\left(0, \frac{1}{8R}\right]$ in Corollary 1 for simplicity.

For EAG-V, the choice $\beta_k = \frac{1}{k+2}$ was obtained by roughly, but not fully, optimizing the bound on EAG-V originating from Lemma 2. If one chooses $\beta_k = \frac{1}{k+\delta}$ with $\delta > 1$, then (6) and (7) become

$$A_k = \frac{\alpha_k(k+\delta)(k+\delta-1)}{2(\delta-1)}, \quad B_k = \frac{k+\delta-1}{\delta-1}$$

As the proof of Theorem 2 illustrates, linear growth of B_k and quadratic growth of A_k leads to $\mathcal{O}(1/k^2)$ convergence of $\|\mathbf{G}(\mathbf{z}^k)\|^2$. The value $\alpha_0 = \frac{0.618}{R}$ in Lemma 1 and Corollary 2 was obtained by numerically minimizing the constant $\frac{4}{\alpha_\infty^2}\left(1 + \alpha_0\alpha_\infty R^2\right)$ in Theorem 2 in the case of $\delta = 2$. The choice $\delta = 2$, however, is not optimal. Indeed, the constant 27 of Corollary 2 can be reduced to 24.44 with $(\delta^\star, \alpha_0^\star) \approx (2.697, 0.690/R)$, which was obtained by numerically optimizing over δ and α_0. Finally, there is a possibility that a choice of β_k not in the form of $\beta_k = \frac{1}{k+\delta}$ leads to an improved constant.

In the end, we choose to present EAG-C and EAG-V with the simple choice $\beta_k = \frac{1}{k+2}$. As we establish in Section 3, the EAG algorithms are optimal up to a constant.

3 Optimality of EAG via a matching complexity lower bound

Upon seeing an accelerated algorithm, it is natural to ask whether the algorithm is optimal. In this section, we present a $\Omega(R^2/k^2)$ complexity lower bound for the class of deterministic gradient-based algorithms for smooth convex-concave minimax problems. This result establishes that EAG is indeed optimal.

For the class of smooth minimax optimization problems, a deterministic *algorithm* $\mathcal{A}$ produces iterates $(\mathbf{x}^k, \mathbf{y}^k) = \mathbf{z}^k$ for $k \geq 1$ given a starting point $(\mathbf{x}^0, \mathbf{y}^0) = \mathbf{z}^0$ and a saddle function $\mathbf{L}$, and we write $\mathbf{z}^k = \mathcal{A}(\mathbf{z}^0, \ldots, \mathbf{z}^{k-1}; \mathbf{L})$ for $k \geq 1$. Define $\mathfrak{A}_{\text{sim}}$ as the class of algorithms satisfying

$$\mathbf{z}^k \in \mathbf{z}^0 + \operatorname{span}\{\mathbf{G}_{\mathbf{L}}(\mathbf{z}^0), \ldots, \mathbf{G}_{\mathbf{L}}(\mathbf{z}^{k-1})\}, \tag{10}$$

and $\mathfrak{A}_{\text{sep}}$ as the class of algorithms satisfying

$$\begin{aligned} \mathbf{x}^k &\in \mathbf{x}^0 + \operatorname{span}\left\{\nabla_{\mathbf{x}}\mathbf{L}(\mathbf{x}^0, \mathbf{y}^0), \ldots, \nabla_{\mathbf{x}}\mathbf{L}(\mathbf{x}^{k-1}, \mathbf{y}^{k-1})\right\} \\ \mathbf{y}^k &\in \mathbf{y}^0 + \operatorname{span}\left\{\nabla_{\mathbf{y}}\mathbf{L}(\mathbf{x}^0, \mathbf{y}^0), \ldots, \nabla_{\mathbf{y}}\mathbf{L}(\mathbf{x}^{k-1}, \mathbf{y}^{k-1})\right\}. \end{aligned} \tag{11}$$

To clarify, algorithms in $\mathfrak{A}_{\text{sim}}$ access and utilize the $\mathbf{x}$- and $\mathbf{y}$-subgradients *simultaneously*. So $\mathfrak{A}_{\text{sim}}$ contains simultaneous gradient descent, extragradient, Popov, and EAG (if we also count intermediate sequences $\mathbf{z}^{k+1/2}$ as algorithms' iterates). On the other hand, algorithms in $\mathfrak{A}_{\text{sep}}$ can access and utilize the $\mathbf{x}$- and $\mathbf{y}$-subgradients *separately*. So $\mathfrak{A}_{\text{sim}} \subset \mathfrak{A}_{\text{sep}}$, and alternating gradient descent-ascent belongs to $\mathfrak{A}_{\text{sep}}$ but not to $\mathfrak{A}_{\text{sim}}$.

In this section, we present a complexity lower bound that applies to all algorithms in $\mathfrak{A}_{\text{sep}}$, not just the algorithms in $\mathfrak{A}_{\text{sim}}$. Although EAG-C and EAG-V are in $\mathfrak{A}_{\text{sim}}$, we consider the broader

class $\mathfrak{A}_{\text{sep}}$ to rule out the possibility that separately updating the $\mathbf{x}$- and $\mathbf{y}$-variables provides an improvement beyond a constant factor.

We say $\mathbf{L}(\mathbf{x}, \mathbf{y})$ is biaffine if it is an affine function of $\mathbf{x}$ for any fixed $\mathbf{y}$ and an affine function of $\mathbf{y}$ for any fixed $\mathbf{x}$. Biaffine functions are, of course, convex-concave. We first establish a complexity lower bound on minimiax optimization problems with biaffine loss functions.

Theorem 3. *Let $k \geq 0$ be fixed. For any $n \geq k+2$, there exists an R-smooth biaffine function $\mathbf{L}$ on $\mathbb{R}^n \times \mathbb{R}^n$ for which*

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \geq \frac{R^2 \|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(2\lfloor k/2 \rfloor + 1)^2} \tag{12}$$

holds for any algorithm in $\mathfrak{A}_{sep}$, where $\lfloor \cdot \rfloor$ is the floor function and $\mathbf{z}^\star$ is the saddle point of $\mathbf{L}$ closest to $\mathbf{z}^0$. Moreover, this lower bound is optimal in the sense that it cannot be improved with biaffine functions.

Since smooth biaffine functions are special cases of smooth convex-concave functions, Theorem 3 implies the optimality of EAG applied to smooth convex-concave mimimax optimization problems.

Corollary 3. *For R-smooth convex-concave minimax problems, an algorithm in $\mathfrak{A}_{sep}$ cannot attain a worst-case convergence rate better than*

$$\frac{R^2 \|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(2\lfloor k/2 \rfloor + 1)^2}$$

with respect to $\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2$. Since EAG-C and EAG-V have rates $\mathcal{O}(R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2/k^2)$, they are optimal, up to a constant factor, in $\mathfrak{A}_{sep}$.

3.1 Outline of the worst-case biaffine construction

Consider biaffine functions of the form

$$\mathbf{L}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} - \mathbf{c} \rangle$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$. Then, $\nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}, \mathbf{y}) = \mathbf{A}^\intercal(\mathbf{y} - \mathbf{c})$, $\nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}, \mathbf{y}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $\mathbf{G}$ is $\|\mathbf{A}\|$-Lipschitz, and solutions to

$$\underset{\mathbf{x} \in X}{\text{minimize}} \ \underset{\mathbf{y} \in Y}{\text{maximize}} \ \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} - \mathbf{c} \rangle$$

are characterized by $\mathbf{A}\mathbf{x} - \mathbf{b} = 0$ and $\mathbf{A}^\intercal(\mathbf{y} - \mathbf{c}) = 0$.

Through translation, we may assume without loss of generality that $\mathbf{x}^0 = 0, \mathbf{y}^0 = 0$. In this case, (11) becomes

$$\begin{aligned}
\mathbf{x}^k \in \ & \operatorname{span}\{\mathbf{A}^\intercal \mathbf{c}, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)\mathbf{c}, \ldots, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k-1}{2} \rfloor}\mathbf{c}\} \\
& + \operatorname{span}\{\mathbf{A}^\intercal \mathbf{b}, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)\mathbf{b}, \ldots, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k}{2} \rfloor - 1}\mathbf{b}\} \\
\mathbf{y}^k \in \ & \operatorname{span}\{b, (\mathbf{A}\mathbf{A}^\intercal)\mathbf{b}, \ldots, (\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k-1}{2} \rfloor}\mathbf{b}\} \\
& + \operatorname{span}\{\mathbf{A}\mathbf{A}^\intercal \mathbf{c}, \ldots, (\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k}{2} \rfloor}\mathbf{c}\}
\end{aligned} \tag{13}$$

for $k \geq 2$. (We detail these arguments in the appendix.) Furthermore let $\mathbf{A} = \mathbf{A}^\intercal$ and $\mathbf{b} = \mathbf{A}^\intercal \mathbf{c} = \mathbf{A}\mathbf{c}$. Then the characterization of $\mathfrak{A}_{\text{sep}}$ further simplifies to

$$\mathbf{x}^k, \mathbf{y}^k \in \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b}) := \text{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\}$$

Note that $\mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b})$ is the order-$(k-1)$ Krylov subspace.

Consider the following lemma. Its proof, deferred to the appendix, combines arguments from [Nemirovsky(1991), Nemirovsky(1992)].

Lemma 3. *Let $R > 0$, $k \geq 0$, and $n \geq k+2$. Then there exists $\mathbf{A} = \mathbf{A}^\intercal \in \mathbb{R}^{n \times n}$ such that $\|\mathbf{A}\| \leq R$ and $\mathbf{b} \in \mathcal{R}(\mathbf{A})$, satisfying*

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \geq \frac{R^2 \|\mathbf{x}^\star\|^2}{(2\lfloor k/2 \rfloor + 1)^2} \tag{14}$$

for any $\mathbf{x} \in \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b})$, where $\mathbf{x}^\star$ is the minimum norm solution to the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$.

Take $\mathbf{A}$ and $\mathbf{b}$ as in Lemma 3 and $\mathbf{c} = \mathbf{x}^\star$. Then $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{x}^\star)$ is the saddle point of $\mathbf{L}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} - \mathbf{c} \rangle$ with minimum norm. Finally,

$$\begin{aligned}
\|\nabla \mathbf{L}(\mathbf{x}^k, \mathbf{y}^k)\|^2 &= \|\mathbf{A}^\intercal(\mathbf{y}^k - \mathbf{c})\|^2 + \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|^2 \\
&= \|\mathbf{A}\mathbf{y}^k - \mathbf{b}\|^2 + \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|^2 \\
&\geq \frac{R^2 \|\mathbf{x}^\star\|^2}{(2\lfloor k/2 \rfloor + 1)^2} + \frac{R^2 \|\mathbf{x}^\star\|^2}{(2\lfloor k/2 \rfloor + 1)^2} \\
&= \frac{R^2 \|\mathbf{z}^\star - \mathbf{z}^0\|^2}{(2\lfloor k/2 \rfloor + 1)^2}
\end{aligned}$$

for any $\mathbf{x}^k, \mathbf{y}^k \in \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b})$. This completes the construction of the biaffine $\mathbf{L}$ of Theorem 3.

3.2 Optimal complexity lower bound

We now formalize the notion of complexity lower bounds. This formulation will allow us to precisely state and prove the second statement of Theorem 3 regarding the optimality of the lower bound.

Let $\mathcal{F}$ be a function class, $\mathcal{P}_\mathcal{F} = \{\mathcal{P}_f\}_{f \in \mathcal{F}}$ a class of optimization problems (with some common form), and $\mathcal{E}(\cdot; \mathcal{P}_f)$ a suboptimality measure for the problem $\mathcal{P}_f$. Define the *worst-case complexity* of an algorithm $\mathcal{A}$ for $\mathcal{P}_\mathcal{F}$ at the k-th iteration given the initial condition $\|\mathbf{z}^0 - \mathbf{z}^\star\| \leq D$, as

$$\mathcal{C}\left(\mathcal{A}; \mathcal{P}_\mathcal{F}, D, k\right) := \sup_{\substack{\mathbf{z}^0 \in B(\mathbf{z}^\star; D) \\ f \in \mathcal{F}}} \mathcal{E}\left(\mathbf{z}^k; \mathcal{P}_f\right)$$

where $\mathbf{z}^j = \mathcal{A}(\mathbf{z}^0, \ldots, \mathbf{z}^{j-1}; f)$ for $j = 1, \ldots, k$ and $B(\mathbf{z}; D)$ denotes the closed ball of radius D centered at $\mathbf{z}$. The *optimal complexity lower bound* with respect to an algorithm class $\mathfrak{A}$ is

$$\begin{aligned}
\mathcal{C}\left(\mathfrak{A}; \mathcal{P}_\mathcal{F}, D, k\right) &:= \inf_{\mathcal{A} \in \mathfrak{A}} \mathcal{C}\left(\mathcal{A}; \mathcal{P}_\mathcal{F}, D, k\right) \\
&= \inf_{\mathcal{A} \in \mathfrak{A}} \sup_{\substack{\mathbf{z}^0 \in B(\mathbf{z}^\star; D) \\ f \in \mathcal{F}}} \mathcal{E}\left(\mathbf{z}^k; \mathcal{P}_f\right)
\end{aligned}$$

A *complexity lower bound* is a lower bound on the optimal complexity lower bound.

Let $\mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^m)$ be the class of R-smooth convex-concave functions on $\mathbb{R}^n \times \mathbb{R}^m$, $\mathcal{P}_{\mathbf{L}}$ the minimax problem (1), and $\mathcal{E}(\mathbf{z}; \mathcal{P}_{\mathbf{L}}) = \|\nabla \mathbf{L}(\mathbf{z})\|^2$. With this notation, the results of Section 2 can be expressed as

$$\mathcal{C}\left(\text{EAG}; \mathcal{P}_{\mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^m)}, D, k\right) = \mathcal{O}\left(\frac{R^2 D^2}{k^2}\right)$$

Let $\mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^m)$ be the class of R-smooth biaffine functions on $\mathbb{R}^n \times \mathbb{R}^m$. Then the first statement of Theorem 3, the existence of $\mathbf{L}$, can be expressed as

$$\mathcal{C}\left(\mathfrak{A}_{\text{sep}}; \mathcal{P}_{\mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)}, D, k\right) \geq \frac{R^2 D^2}{(2\lfloor k/2 \rfloor + 1)^2} \tag{15}$$

for $n \geq k + 2$.

As an aside, the argument of Corollary 3 can be expressed as: for any $\mathcal{A} \in \mathfrak{A}_{\text{sep}}$, we have

$$\begin{aligned}
\mathcal{C}\left(\mathcal{A}; \mathcal{P}_{\mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^n)}, D, k\right) &\geq \mathcal{C}\left(\mathfrak{A}_{\text{sep}}; \mathcal{P}_{\mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^n)}, D, k\right) \\
&\geq \mathcal{C}\left(\mathfrak{A}_{\text{sep}}; \mathcal{P}_{\mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)}, D, k\right) \\
&\geq \frac{R^2 D^2}{(2\lfloor k/2 \rfloor + 1)^2}
\end{aligned}$$

The first inequality follows from $\mathcal{A} \in \mathfrak{A}_{\text{sep}}$, the second from $\mathcal{L}_R^{\text{biaff}} \subset \mathcal{L}_R$, and the third from Theorem 3.

Optimality of lower bound of Theorem 3. Using above notations, our goal is to prove that for $n \geq k + 2$,

$$\mathcal{C}\left(\mathfrak{A}_{\text{sep}}; \mathcal{P}_{\mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)}, D, k\right) = \frac{R^2 D^2}{(2\lfloor k/2 \rfloor + 1)^2} \tag{16}$$

We establish this claim with the chain of inequalities:

$$\begin{aligned}
\frac{R^2 D^2}{(2\lfloor k/2 \rfloor + 1)^2} &\leq \mathcal{C}\left(\mathfrak{A}_{\text{sep}}; \mathcal{P}_{\mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)}, D, k\right) && (17) \\
&\leq \mathcal{C}\left(\mathfrak{A}_{\text{sim}}; \mathcal{P}_{\mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)}, D, k\right) && (18) \\
&\leq \mathcal{C}\left(\mathfrak{A}_{\text{lin}}; \mathcal{P}_{R,D}^{2n,\text{skew}}, k\right) && (19) \\
&\leq \mathcal{C}\left(\mathfrak{A}_{\text{lin}}; \mathcal{P}_{R,D}^{2n}, k\right) && (20) \\
&\leq \frac{R^2 D^2}{(2\lfloor k/2 \rfloor + 1)^2} && (21)
\end{aligned}$$

Inequality (17) is what we established in Section 3.1. Inequality (18) follows from $\mathfrak{A}_{\text{sim}} \subset \mathfrak{A}_{\text{sep}}$ and the fact that the infimum over a larger class is smaller. Roughly speaking, the quantities in lines (19) and (20) are the complexity lower bounds for solving linear equations using only matrix-vector products, which were studied thoroughly in [Nemirovsky(1991), Nemirovsky(1992)]. We will show inequalities (19), (20), and (21) by establishing the connection of Nemirovsky's work with our setup of biaffine saddle problems. Once this is done, equality holds throughout and (16) is proved.

We first provide the definitions. Let $\mathcal{P}_{R,D}^{2n}$ be the collection of linear equations with $2n \times 2n$ matrices $\mathbf{B}$ satisfying $\|\mathbf{B}\| \le R$ and $\mathbf{v} = \mathbf{B}\mathbf{z}^\star$ for some $\mathbf{z}^\star \in B(0; D)$. Let $\mathcal{P}_{R,D}^{2n,\text{skew}} \subset \mathcal{P}_{R,D}^{2n}$ be the subclass of equations with skew-symmetric $\mathbf{B}$. Let $\mathfrak{A}_{\text{lin}}$ be the class of iterative algorithms solving linear equations $\mathbf{B}\mathbf{z} = \mathbf{v}$ using only matrix multiplication by $\mathbf{B}$ and $\mathbf{B}^\intercal$ in the sense that

$$\mathbf{z}^k \in \text{span}\{\mathbf{v}^0, \ldots, \mathbf{v}^k\} \tag{22}$$

where $\mathbf{v}^0 = 0$, $\mathbf{v}^1 = \mathbf{v}$, and for $k \ge 2$,

$$\mathbf{v}^k = \mathbf{B}\mathbf{v}^j \text{ or } \mathbf{B}^\intercal\mathbf{v}^j \text{ for some } j = 0, \ldots, k-1$$

The optimal complexity lower bound for a class of linear equation instances is defined as

$$\mathcal{C}\left(\mathfrak{A}_{\text{lin}}; \mathcal{P}_{R,D}^{2n}, k\right) := \inf_{\mathcal{A} \in \mathfrak{A}_{\text{lin}}} \sup_{\substack{\|\mathbf{B}\| \le R \\ \mathbf{v} = \mathbf{B}\mathbf{z}^\star, \|\mathbf{z}^\star\| \le D}} \left\|\mathbf{B}\mathbf{z}^k - \mathbf{v}\right\|^2$$

Define $\mathcal{C}\left(\mathfrak{A}_{\text{lin}}; \mathcal{P}_{R,D}^{2n,\text{skew}}, k\right)$ analogously.

Now we relate the optimal complexity lower bounds for biaffine minimax problems to those for linear equations. For $\mathbf{L}(\mathbf{x}, \mathbf{y}) = \mathbf{b}^\intercal\mathbf{x} + \mathbf{x}^\intercal\mathbf{A}\mathbf{y} - \mathbf{c}^\intercal\mathbf{y}$, we have

$$\mathbf{G}_{\mathbf{L}}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \mathbf{O} & \mathbf{A} \\ -\mathbf{A}^\intercal & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix}$$

Therefore, the minimax problem $\mathcal{P}_{\mathbf{L}}$ for $\mathbf{L} \in \mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)$ is equivalent to solving the linear equation $\mathbf{B}\mathbf{z} = \mathbf{v}$ with $\mathbf{B} = \begin{bmatrix} \mathbf{O} & -\mathbf{A} \\ \mathbf{A}^\intercal & \mathbf{O} \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} \in \mathbb{R}^{2n}$, which belongs to $\mathcal{P}_{R,D}^{2n,\text{skew}}$ with $D = \|z^\star\|$.

For both algorithm classes $\mathfrak{A}_{\text{sim}}$ and $\mathfrak{A}_{\text{lin}}$, we may assume without loss of generality that $\mathbf{z}^0 = 0$ through translation. Then, the span condition (10) for $\mathfrak{A}_{\text{sim}}$ becomes

$$\mathbf{z}^k \in \mathcal{K}_{k-1}(\mathbf{B}; \mathbf{v}) \tag{23}$$

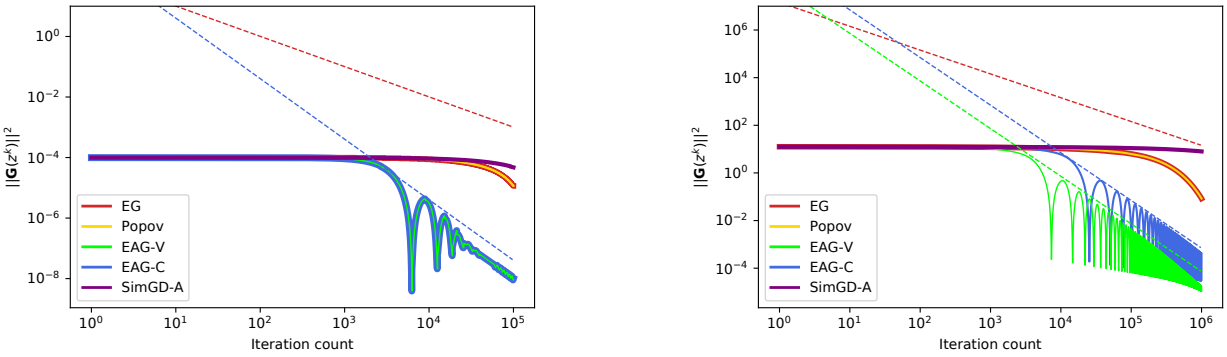
Note that (22) reduces to (23) as $\mathbf{B}$ is skew-symmetric, so $\mathfrak{A}_{\text{sim}}$ and $\mathfrak{A}_{\text{lin}}$ are effectively the same class of algorithms under the identification $\mathcal{P}_{\mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)} \subset \mathcal{P}_{R,D}^{2n,\text{skew}}$.

Since the supremum over a larger class of problems is larger, inequality (19) holds. Similarly, inequality (20) follows from $\mathcal{P}_{R,D}^{2n,\text{skew}} \subset \mathcal{P}_{R,D}^{2n}$.

Finally, (21) follows from the following lemma, using arguments based on Chebyshev-type matrix polynomials from [Nemirovsky(1992)]. Its proof is deferred to the appendix.

Lemma 4. *Let $R > 0$ and $k \ge 0$. Then there exists $\mathcal{A} \in \mathfrak{A}_{\text{lin}}$ such that for any $m \ge 1$, $\mathbf{B} \in \mathbb{R}^{m \times m}$, and $\mathbf{v} = \mathbf{B}\mathbf{z}^\star$ satisfying $\|\mathbf{B}\| \le R$ and $\|\mathbf{z}^\star\| \le D$, the $\mathbf{z}^k$-iterate produced by $\mathcal{A}$ satisfies*

$$\left\|\mathbf{B}\mathbf{z}^k - \mathbf{v}\right\|^2 \le \frac{R^2 D^2}{(2\lfloor k/2 \rfloor + 1)^2}$$



(a) Two-dimensional example $\mathbf{L}_{\delta,\epsilon}$ of (24)

(b) Lagrangian of linearly constrained QP of (25)

Figure 1. Plots of $\|\mathbf{G}(\mathbf{z}^k)\|^2$ versus iteration count. Dashed lines indicate corresponding theoretical upper bounds.

3.3 Broader algorithm classes via resisting oracles

In (10) and (11), we assumed the subgradient queries are made within the span of the gradients at the previous iterates. This requirement (the *linear span assumption*) can be removed, i.e., a similar analysis can be done on general deterministic black-box gradient-based algorithms (formally defined in the appendix, Section C.5), using the resisting oracle technique [Nemirovsky & Yudin(1983)] at the cost of slightly enlarging the required problem dimension. We informally state the generalized result below and provide details in the appendix.

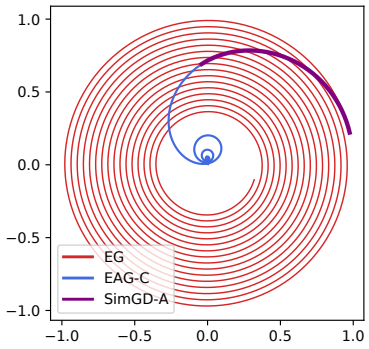
Theorem 4 (Informal)**.** *Let $n \geq 3k+2$. For any gradient-based deterministic algorithm, there exists an R-smooth biaffine function $\mathbf{L}$ on $\mathbb{R}^n \times \mathbb{R}^n$ such that* (12) *holds.*

Although we do not formally pursue this, the requirement that the algorithm is not randomized can also be removed using the techniques of [Woodworth & Srebro(2016)], which exploit near-orthogonality of random vectors in high dimensions.

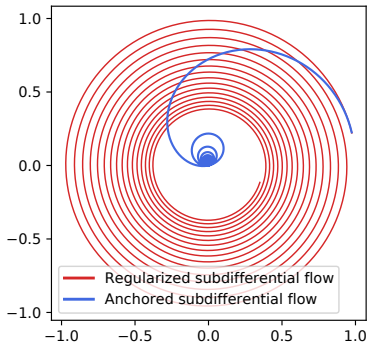
3.4 Discussion

We established that one cannot improve the lower bound of Theorem 3 using biaffine functions, arguably the simplest family of convex-concave functions. Furthermore, this optimality statement holds for both algorithm classes $\mathfrak{A}_{\text{sep}}$ and $\mathfrak{A}_{\text{sim}}$ as established through the chain of inequalities in Section 3.2. However, as demonstrated by [Drori(2017)], who introduced a non-quadratic lower bound for smooth convex minimization that improves upon the classical quadratic lower bounds of [Nemirovsky(1992)] and [Nesterov(2013)], a non-biaffine construction may improve the constant. In our setup, there is a factor-near-100 difference between the upper and lower bounds. (Note that each EAG iteration requires 2 evaluations of the saddle subdifferential oracle.) We suspect that both the algorithm and the lower bound can be improved upon, but we leave this to future work.

[Golowich et al.(2020)] establishes that for the class of 1-SCLI algorithms (S is for *stationary*), a subclass of $\mathfrak{A}_{\text{sim}}$ for biaffine objectives, one cannot achieve a rate faster than $\|\nabla\mathbf{L}(\mathbf{z}^k)\|^2 \leq \mathcal{O}(1/k)$. This lower bound applies to EG but not EAG; EAG is not 1-SCLI, as its anchoring coefficients $\frac{1}{k+2}$ vary over iterations, and its convergence rate breaks the 1-SCLI lower bound. On the other hand, we can view EAG as a non-stationary CLI algorithm [Arjevani & Shamir(2016), Definition 2]. We further discuss these connections in the appendix, Section E.



(a) Discrete trajectories with $\mathbf{L}_{\delta,\epsilon}$



(b) Moreau–Yosida regularized flow with $\lambda = 0.01$ and the anchored flow with $\mathbf{L}(x, y) = xy$

Figure 2. Comparison of the discrete trajectories and their corresponding continuous-time flow. Trajectories from EAG-C and SimGD-A virtually coincide and resemble the anchored flow. However, SimGD-A progresses slower due to its diminishing step-sizes.

4 Experiments

We now present experiments illustrating the accelerated rate of EAG. We compare EAG-C and EAG-V against the prior algorithms with convergence guarantees: EG, Popov's algorithm (or optimistic descent) and simultaneous gradient descent with anchoring (SimGD-A). The precise forms of the algorithms are restated in the appendix.

Figure 1(a) presents experiments on our first example, constructed as follows. For $\epsilon > 0$, define

$$f_\epsilon(u) = \begin{cases} \epsilon|u| - \frac{1}{2}\epsilon^2 & \text{if } |u| \geq \epsilon, \\ \frac{1}{2}u^2 & \text{if } |u| < \epsilon \end{cases}$$

Next, for $0 < \epsilon \ll \delta \ll 1$, define

$$\mathbf{L}_{\delta,\epsilon}(x, y) = (1 - \delta)f_\epsilon(x) + \delta xy - (1 - \delta)f_\epsilon(y) \tag{24}$$

where $x, y \in \mathbb{R}$. Since f_ϵ is a 1-smooth convex function, $\mathbf{L}_{\delta,\epsilon}$ has smoothness parameter 1, which is almost tight due to the quadratic behavior of $\mathbf{L}_{\delta,\epsilon}$ within the region $|x|, |y| \leq \epsilon$. This construction was inspired by [Drori & Teboulle(2014)], who presented f_ϵ as the worst-case instance for gradient descent. We choose the step-size $\alpha = 0.1$ as this value is comfortably within the theoretical range of convergent parameters for EG, EAG-C, and Popov. For EAG-V, we set $\alpha_0 = 0.1$. We use $N = 10^5$, $\delta = 10^{-2}$, and $\epsilon = 5 \times 10^{-5}$, and the initial point $\mathbf{z}^0$ has norm 1.

Figure 1(b) presents experiments on our second example

$$\mathbf{L}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^\intercal \mathbf{H}\mathbf{x} - \mathbf{h}^\intercal \mathbf{x} - \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} \rangle \tag{25}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{H} \in \mathbb{R}^{n \times n}$ is positive semidefinite, and $\mathbf{h} \in \mathbb{R}^n$. Note that this is the Lagrangian of a linearly constrained quadratic minimization problem. We adopted this saddle function from [Ouyang & Xu(2021)], where the authors constructed $\mathbf{H}, \mathbf{h}, \mathbf{A}$ and $\mathbf{b}$ to provide a lower bound on duality gap. The exact forms of $\mathbf{H}$, $\mathbf{h}$, $\mathbf{A}$, and $\mathbf{b}$ are restated in the appendix. We use $n = 200$, $N = 10^6$, $\alpha = 0.5$ for EG and Popov, $\alpha = 0.1265$ for EAG-C and $\alpha_0 = 0.618$ for EAG-V. Finally, we use the initial point $\mathbf{z}^0 = 0$.

ODE Interpretation Figure 2(a) illustrates the algorithms applied to (24). For $|x|, |y| \gg \epsilon$,

$$\mathbf{G}_{\mathbf{L}_{\delta,\epsilon}}(x, y) = \begin{bmatrix}(1-\delta)\epsilon + \delta y \\ (1-\delta)\epsilon - \delta x\end{bmatrix} \approx \delta \begin{bmatrix} y \\ -x\end{bmatrix}$$

so the algorithms roughly behave as if the objective is the bilinear function δxy. When δ is sufficiently small, trajectories of the algorithms closely resemble the corresponding continuous-time flows with $\mathbf{L}(x, y) = xy$.

[Csetnek et al.(2019)] demonstrated that Popov's algorithm can be viewed as discretization of the Moreau–Yosida regularized flow $\dot{\mathbf{z}}(t) = -\frac{\mathbf{G}-(\mathrm{Id}+\lambda\mathbf{G})^{-1}}{\lambda}(\mathbf{z}(t))$ for some $\lambda > 0$, and a similar analysis can be performed with EG. This connection explains why EG's trajectory in Figure 2(a) and the regularized flow depicted in Figure 2(b) are similar.

On the other hand, EAG and SimGD-A can be viewed as a discretization of the anchored flow ODE

$$\dot{\mathbf{z}}(t) = -\mathbf{G}(\mathbf{z}(t)) + \frac{1}{t}(\mathbf{z}^0 - \mathbf{z}(t))$$

The anchored flow depicted in Figure 2(b) approaches the solution much more quickly due to the anchoring term dampening the cycling behavior. The trajectories of EAG and SimGD-A iterates in Figure 2(a) are very similar to the anchored flow. However, SimGD-A requires diminishing step-sizes $\frac{1-p}{(k+1)^p}$ (both theoretically and experimentally) and therefore progresses much slower.

5 Conclusion

In this work, we presented the Extra Anchored Gradient (EAG) algorithms, which achieve an accelerated $\mathcal{O}(1/k^2)$ convergence rate on the squared gradient magnitude for smooth convex-concave minimax problems. This rate improves upon the existing $\mathcal{O}(1/k)$ or slower rates achieved by extragradient, Popov, and gradient descent with anchoring. The acceleration mechanism combines extragradient steps with anchoring, which is distinct from Nesterov's momentum-based acceleration. We complement the $\mathcal{O}(1/k^2)$ rate with a matching $\Omega(1/k^2)$ complexity lower bound, establishing the optimality of EAG.

While anchoring dampens oscillations and momentum amplifies them, an open question remains: ***are the two acceleration mechanisms entirely unrelated?*** Exploring potential connections between these phenomena presents an interesting avenue for future research.

References

[Alkousa et al.(2020)] Alkousa, M., Gasnikov, A., Dvinskikh, D., Kovalev, D., and Stonyakin, F. Accelerated methods for saddle-point problem. *Computational Mathematics and Mathematical Physics*, 60(11):1787–1809, 2020.

[Antonakopoulos et al.(2021)] Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. Adaptive extragradient methods for min-max optimization and games. *ICLR*, 2021.

[Arjevani & Shamir(2016)] Arjevani, Y. and Shamir, O. On the iteration complexity of oblivious first-order optimization algorithms. *ICML*, 2016.

[Arjevani et al.(2016)] Arjevani, Y., Shalev-Shwartz, S., and Shamir, O. On lower and upper bounds in smooth and strongly convex optimization. *Journal of Machine Learning Research*, 17(1):4303–4353, 2016.

[Azizian et al.(2020)] Azizian, W., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. *AISTATS*, 2020.

[Bùi & Combettes(2021)] Bùi, M. N. and Combettes, P. L. A warped resolvent algorithm to construct nash equilibria. *arXiv preprint arXiv:2101.00532*, 2021.

[Censor et al.(2011)] Censor, Y., Gibali, A., and Reich, S. The subgradient extragradient method for solving variational inequalities in Hilbert space. *Journal of Optimization Theory and Applications*, 148(2):318–335, 2011.

[Chambolle & Pock(2011)] Chambolle, A. and Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[Chambolle & Pock(2016)] Chambolle, A. and Pock, T. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

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

[Chen et al.(2017)] Chen, Y., Lan, G., and Ouyang, Y. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

[Chen et al.(2021)] Chen, Z., Zhou, Y., Xu, T., and Liang, Y. Proximal gradient descent-ascent: Variable convergence under KŁ geometry. *ICLR*, 2021.

[Chiang et al.(2012)] Chiang, C.-K., Yang, T., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., and Zhu, S. Online optimization with gradual variations. *COLT*, 2012.

[Condat(2013)] Condat, L. A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2): 460–479, 2013.

[Csetnek et al.(2019)] Csetnek, E. R., Malitsky, Y., and Tam, M. K. Shadow Douglas–Rachford splitting for monotone inclusions. *Applied Mathematics & Optimization*, 80(3):665–678, 2019.

[Daskalakis et al.(2011)] Daskalakis, C., Deckelbaum, A., and Kim, A. Near-optimal no-regret algorithms for zero-sum games. *SODA*, 2011.

[Diakonikolas(2020)] Diakonikolas, J. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. *COLT*, 2020.

[Drori(2017)] Drori, Y. The exact information-based complexity of smooth convex minimization. *Journal of Complexity*, 39:1–16, 2017.

[Drori & Teboulle(2014)] Drori, Y. and Teboulle, M. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

[Ghadimi & Lan(2012)] Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[Ghadimi & Lan(2013)] Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[Gidel et al.(2018)] Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *ICLR*, 2018.

[Gidel et al.(2019)] Gidel, G., Hemmat, R. A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. *AISTATS*, 2019.

[Golowich et al.(2020)] Golowich, N., Pattathil, S., Daskalakis, C., and Ozdaglar, A. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. *COLT*, 2020.

[Goodfellow et al.(2014)] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NeurIPS*, 2014.

[Goodfellow et al.(2015)] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *ICLR*, 2015.

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

[Hamedani & Aybat(2018)] Hamedani, E. Y. and Aybat, N. S. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

[He & Monteiro(2016)] He, Y. and Monteiro, R. D. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26(1):29–56, 2016.

[Hsieh et al.(2019)] Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. *NeurIPS*, 2019.

[Jin et al.(2019)] Jin, C., Netrapalli, P., and Jordan, M. I. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

[Juditsky et al.(2011)] Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[Kolossoski & Monteiro(2017)] Kolossoski, O. and Monteiro, R. D. An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems. *Optimization Methods and Software*, 32(6):1244–1272, 2017.

[Korpelevich(1977)] Korpelevich, G. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

[Lan(2012)] Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[Liang & Stokes(2019)] Liang, T. and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *AISTATS*, 2019.

[Lieder(2020)] Lieder, F. On the convergence rate of the halpern-iteration. *Optimization Letters*, pp. 1–14, 2020.

[Lin et al.(2020a)] Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. *ICML*, 2020a.

[Lin et al.(2020b)] Lin, T., Jin, C., Jordan, M., et al. Near-optimal algorithms for minimax optimization. *COLT*, 2020b.

[Lu et al.(2020)] Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

[Lyashko et al.(2011)] Lyashko, S., Semenov, V., and Voitova, T. Low-cost modification of Korpelevich's methods for monotone equilibrium problems. *Cybernetics and Systems Analysis*, 47(4):631, 2011.

[Madry et al.(2018)] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

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

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

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

[Malitsky & Semenov(2014)] Malitsky, Y. V. and Semenov, V. An extragradient algorithm for monotone variational inequalities. *Cybernetics and Systems Analysis*, 50(2):271–277, 2014.

[Mason & Handscomb(2002)] Mason, J. C. and Handscomb, D. C. *Chebyshev Polynomials*. 2002.

[Mertikopoulos et al.(2019)] Mertikopoulos, P., Zenati, H., Lecouat, B., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *ICLR*, 2019.

[Mokhtari et al.(2020a)] Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *AISTATS*, 2020a.

[Mokhtari et al.(2020b)] Mokhtari, A., Ozdaglar, A. E., and Pattathil, S. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

[Monteiro & Svaiter(2010)] Monteiro, R. D. and Svaiter, B. F. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6): 2755–2787, 2010.

[Monteiro & Svaiter(2011)] Monteiro, R. D. and Svaiter, B. F. Complexity of variants of Tseng's modified FB splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

[Nedić & Ozdaglar(2009)] Nedić, A. and Ozdaglar, A. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

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

[Nemirovski et al.(2009)] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

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

[Nemirovsky & Yudin(1983)] Nemirovsky, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. 1983.

[Nesterov(2005a)] Nesterov, Y. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005a.

[Nesterov(2005b)] Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005b.

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

[Nesterov(2009)] Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[Nesterov(2013)] Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. 2013.

[Nesterov & Scrimali(2011)] Nesterov, Y. and Scrimali, L. Solving strongly monotone variational and quasi-variational inequalities. *Discrete & Continuous Dynamical Systems – A*, 31(4):1383–1396, 2011.

[Noor(2003)] Noor, M. A. New extragradient-type methods for general variational inequalities. *Journal of Mathematical Analysis and Applications*, 277(2):379–394, 2003.

[Nouiehed et al.(2019)] Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. *NeurIPS*, 2019.

[Ostrovskii et al.(2020)] Ostrovskii, D. M., Lowy, A., and Razaviyayn, M. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*, 2020.

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

[Peng et al.(2020)] Peng, W., Dai, Y.-H., Zhang, H., and Cheng, L. Training GANs with centripetal acceleration. *Optimization Methods and Software*, 35(5):955–973, 2020.

[Popov(1980)] Popov, L. D. A modification of the Arrow–Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[Rafique et al.(2018)] Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

[Rakhlin & Sridharan(2013a)] Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. *COLT*, 2013a.

[Rakhlin & Sridharan(2013b)] Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. *NeurIPS*, 2013b.

[Rockafellar(1970)] Rockafellar, R. T. Monotone operators associated with saddle-functions and minimax problems. *Nonlinear Functional Analysis*, 18(part 1):397–407, 1970.

[Ryu & Yin(2021)] Ryu, E. K. and Yin, W. *Large-Scale Convex Optimization via Monotone Operators*. Draft, 2021.

[Ryu et al.(2019)] Ryu, E. K., Yuan, K., and Yin, W. ODE analysis of stochastic gradient methods with optimism and anchoring for minimax problems and GANs. *arXiv preprint arXiv:1905.10899*, 2019.

[Solodov & Svaiter(1999)] Solodov, M. V. and Svaiter, B. F. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4): 323–345, 1999.

[Syrgkanis et al.(2015)] Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast convergence of regularized learning in games. *NeurIPS*, 2015.

[Taylor & Bach(2019)] Taylor, A. and Bach, F. Stochastic first-order methods: Non-asymptotic and computer-aided analyses via potential functions. *COLT*, 2019.

[Taylor et al.(2017)] Taylor, A. B., Hendrickx, J. M., and Glineur, F. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1-2):307–345, 2017.

[Thekumparampil et al.(2019)] Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. *NeurIPS*, 2019.

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

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

[Vũ(2013)] Vũ, B. C. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[Wang & Li(2020)] Wang, Y. and Li, J. Improved algorithms for convex-concave minimax optimization. *NeurIPS*, 2020.

[Woodworth & Srebro(2016)] Woodworth, B. and Srebro, N. Tight complexity bounds for optimizing composite objectives. *NeurIPS*, 2016.

[Yadav et al.(2018)] Yadav, A., Shah, S., Xu, Z., Jacobs, D., and Goldstein, T. Stabilizing adversarial nets with prediction methods. *ICLR*, 2018.

[Yan(2018)] Yan, M. A new primal–dual algorithm for minimizing the sum of three functions with a linear operator. *Journal of Scientific Computing*, 76(3):1698–1717, 2018.

[Yang et al.(2020)] Yang, J., Zhang, S., Kiyavash, N., and He, N. A catalyst framework for minimax optimization. *NeurIPS*, 2020.

[Zhang et al.(2020)] Zhang, G., Bao, X., Lessard, L., and Grosse, R. A unified analysis of first-order methods for smooth games via integral quadratic constraints. *arXiv preprint arXiv:2009.11359*, 2020.

[Zhang et al.(2019)] Zhang, J., Hong, M., and Zhang, S. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.

[Zhao(2019)] Zhao, R. Optimal stochastic algorithms for convex-concave saddle-point problems. *arXiv preprint arXiv:1903.01687*, 2019.

A Algorithm specifications

For the sake of clarity, we precisely specify all the algorithms discussed in this work.

Simultaneous gradient descent for smooth minimax optimization is defined as

$$\begin{aligned}\mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha\nabla_{\mathbf{x}}\mathbf{L}(\mathbf{x}^k, \mathbf{y}^k)\\ \mathbf{y}^{k+1} &= \mathbf{y}^k + \alpha\nabla_{\mathbf{y}}\mathbf{L}(\mathbf{x}^k, \mathbf{y}^k)\end{aligned}$$

The notation becomes more concise with the joint variable notation $\mathbf{z}^k = (\mathbf{x}^k, \mathbf{y}^k)$ and the saddle operator (2), where the sign change in $\mathbf{y}$-gradient is already included:

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha\mathbf{G}(\mathbf{z}^k)$$

Alternating gradient descent-ascent is defined as

$$\begin{aligned}\mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha\nabla_{\mathbf{x}}\mathbf{L}(\mathbf{x}^k, \mathbf{y}^k)\\ \mathbf{y}^{k+1} &= \mathbf{y}^k + \alpha\nabla_{\mathbf{y}}\mathbf{L}(\mathbf{x}^{k+1}, \mathbf{y}^k)\end{aligned}$$

Note that we update the $\mathbf{x}$ variable first and then use it to update the $\mathbf{y}$-iterate.

The *extragradient (EG) algorithm* is defined as

$$\begin{aligned}\mathbf{z}^{k+1/2} &= \mathbf{z}^k - \alpha\mathbf{G}(\mathbf{z}^k),\\ \mathbf{z}^{k+1} &= \mathbf{z}^k - \alpha\mathbf{G}(\mathbf{z}^{k+1/2})\end{aligned}$$

Popov's algorithm, or *optimistic descent*, is defined as

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha\mathbf{G}(\mathbf{z}^k) - \alpha\left(\mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k-1})\right)$$

Simultaneous gradient descent with anchoring (SimGD-A) [Ryu et al.(2019)] is defined as

$$\mathbf{z}^{k+1} = \mathbf{z}^k - \frac{1-p}{(k+1)^p}\mathbf{G}(\mathbf{z}^k) + \frac{(1-p)\gamma}{k+1}(\mathbf{z}^0 - \mathbf{z}^k)$$

where $p \in (1/2, 1)$ and $\gamma > 0$. It has been proved in [Ryu et al.(2019)] that SimGD-A converges at $\mathcal{O}(1/k^{2-2p})$ rate. In this paper, we always used $\gamma = 1$ and $p = \frac{1}{2} + 10^{-2}$.

B Omitted proofs of Section 2

The following identities follow directly from the definition of EAG iterates:

$$\mathbf{z}^k - \mathbf{z}^{k+1} = \beta_k(\mathbf{z}^k - \mathbf{z}^0) + \alpha_k\,\mathbf{G}(\mathbf{z}^{k+1/2}) \tag{26}$$

$$\mathbf{z}^{k+1/2} - \mathbf{z}^{k+1} = \alpha_k\left(\mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^k)\right) \tag{27}$$

$$\mathbf{z}^0 - \mathbf{z}^{k+1} = (1-\beta_k)(\mathbf{z}^0 - \mathbf{z}^k) + \alpha_k\,\mathbf{G}(\mathbf{z}^{k+1/2}) \tag{28}$$

B.1 Proof of Lemma 2

Recall that $\mathbf{G}$ is a monotone operator, so that

$$0 \leq \left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle$$

Therefore,

$$\begin{aligned}
&V_k - V_{k+1} \\
&\geq V_k - V_{k+1} - \frac{B_k}{\beta_k} \left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\
&= A_k \left\| \mathbf{G}(\mathbf{z}^k) \right\|^2 + B_k \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \right\rangle \\
&\quad - A_{k+1} \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 - B_{k+1} \left\langle \mathbf{G}(\mathbf{z}^{k+1}), \mathbf{z}^{k+1} - \mathbf{z}^0 \right\rangle - \frac{B_k}{\beta_k} \left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\
&\overset{\text{(a)}}{=} A_k \left\| \mathbf{G}(\mathbf{z}^k) \right\|^2 + B_k \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \right\rangle \\
&\quad - A_{k+1} \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 + B_{k+1} \left\langle \mathbf{G}(\mathbf{z}^{k+1}), (1-\beta_k)(\mathbf{z}^0 - \mathbf{z}^k) + \alpha_k \mathbf{G}(\mathbf{z}^{k+1/2}) \right\rangle \\
&\quad - B_k \left\langle \mathbf{z}^k - \mathbf{z}^0, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle - \frac{\alpha_k B_k}{\beta_k} \left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\
&\overset{\text{(b)}}{=} A_k \left\| \mathbf{G}(\mathbf{z}^k) \right\|^2 - A_{k+1} \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 + \alpha_k B_{k+1} \left\langle \mathbf{G}(\mathbf{z}^{k+1}), \mathbf{G}(\mathbf{z}^{k+1/2}) \right\rangle \\
&\quad - \frac{\alpha_k B_k}{\beta_k} \left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle
\end{aligned} \tag{29}$$

where (a) follows from (26) and (28), and (b) results from cancellation and collection of terms using (7). Next, we have

$$\begin{aligned}
0 &\leq R^2 \left\| \mathbf{z}^{k+1/2} - \mathbf{z}^{k+1} \right\|^2 - \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \\
&= \alpha_k^2 R^2 \left\| \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1/2}) \right\|^2 - \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2
\end{aligned} \tag{30}$$

from R-Lipschitzness of $\mathbf{G}$ and (27). Now multiplying the factor $\frac{A_k}{\alpha_k^2 R^2}$ to (30) and subtracting from (29) gives

$$\begin{aligned}
&V_k - V_{k+1} \\
&\geq A_k \left\| \mathbf{G}(\mathbf{z}^k) \right\|^2 - A_{k+1} \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 + \alpha_k B_{k+1} \left\langle \mathbf{G}(\mathbf{z}^{k+1}), \mathbf{G}(\mathbf{z}^{k+1/2}) \right\rangle \\
&\quad - \frac{\alpha_k B_k}{\beta_k} \left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\
&\quad - A_k \left\| \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1/2}) \right\|^2 + \frac{A_k}{\alpha_k^2 R^2} \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \\
&= \frac{A_k(1 - \alpha_k^2 R^2)}{\alpha_k^2 R^2} \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) \right\|^2 + \left(\frac{A_k}{\alpha_k^2 R^2} - A_{k+1} \right) \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \\
&\quad + \left(2A_k - \frac{\alpha_k B_k}{\beta_k} \right) \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{G}(\mathbf{z}^{k+1/2}) \right\rangle \\
&\quad + \left(\alpha_k B_{k+1} + \frac{\alpha_k B_k}{\beta_k} - \frac{2A_k}{\alpha_k^2 R^2} \right) \left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle
\end{aligned} \tag{31}$$

Observe that the $\left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{G}(\mathbf{z}^{k+1/2}) \right\rangle$ term vanishes because of (6), and that

$$\alpha_k B_{k+1} + \frac{\alpha_k B_k}{\beta_k} = \alpha_k \left(\frac{B_k}{1-\beta_k} + \frac{B_k}{\beta_k} \right) = \frac{\alpha_k B_k}{\beta_k(1-\beta_k)} = \frac{2A_k}{1-\beta_k}$$

Furthermore, by (8), we have

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

Plugging these identities into (31) and simplifying, we get

$$\begin{aligned}
&V_k - V_{k+1} \\
&\geq \frac{A_k(1-\alpha_k^2 R^2)}{\alpha_k^2 R^2} \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) \right\|^2 + \frac{A_k(1-\alpha_k^2 R^2 - \beta_k)^2}{\alpha_k^2 R^2 (1-\alpha_k^2 R^2)(1-\beta_k)^2} \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \\
&\quad - \frac{2A_k(1-\alpha_k^2 R^2 - \beta_k)}{\alpha_k^2 R^2 (1-\beta_k)} \left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\
&\geq 0
\end{aligned}$$

where the last inequality is an application of Young's inequality.

B.2 Proof of Lemma 1

We may assume $R = 1$ without loss of generality because we can recover the general case by replacing α_k with $\alpha_k R$. Rewrite (5) as

$$\alpha_k - \alpha_{k+1} = \frac{\alpha_k^3}{(k+1)(k+3)(1-\alpha_k^2)} \tag{32}$$

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

$$\gamma := \frac{1}{2} \left(\frac{1}{N+1} + \frac{1}{N+2} \right) \frac{\rho^2}{1-\rho^2} < 1 \tag{33}$$

Note that (33) holds true for all $N \geq 0$ if $\rho < \frac{3}{4}$. Now we will show that given (33),

$$\alpha_N > \alpha_{N+1} > \cdots > \alpha_{N+k} > (1-\gamma)\alpha_N \quad \text{for all } k \geq 0$$

so that $\alpha_k \downarrow \alpha$ for some $\alpha \geq (1-\gamma)\alpha_N$. It suffices to prove that $(1-\gamma)\alpha_N < \alpha_{N+k} < \rho$ for all $k \geq 0$, because it is clear from (32) that $\{\alpha_k\}_{k \geq 0}$ is decreasing.

We use induction on k to prove that $\alpha_{N+k} \in ((1-\gamma)\alpha_N, \rho)$. The case $k = 0$ is trivial. Now suppose that $(1-\gamma)\alpha_N < \alpha_{N+j} < \rho$ holds true for all $j = 0, \ldots, k$. Then by (32), for each $0 \leq j \leq k$ we have

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

Summing up the inequalities for $j = 0, \ldots, k$, we obtain

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

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

In particular, when $\alpha_0 = 0.618$, direct calculation gives $0.437 > \alpha_N > 0.4366$ when $N = 1000$. With $\rho = 0.437$ and $N = 1000$, we have $\gamma = \frac{1}{2}\left(\frac{1}{N+1} + \frac{1}{N+2}\right)\frac{\rho^2}{1-\rho^2} < 2.5 \times 10^{-4}$, which gives $\alpha \geq (1-\gamma)\alpha_N \approx 0.4365$.

B.3 Proof of Theorem 1

As in the proof of Theorem 2, assume without loss of generality that $R = 1$. The strategy of the proof is basically the same as in Theorem 2; we construct a nonincreasing Lyapunov function by combining the same set of inequalities, but with different (more intricate) coefficients. For $k \geq 0$, let

$$V_k = A_k \left\| \mathbf{G}(\mathbf{z}^k) \right\|^2 + B_k \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \right\rangle$$

As in Lemma 2, we will use $B_k = \frac{1}{1-\beta_k} = k+1$, and $a_k \geq 0$ will be specified later. Because we have the fixed step-size α, the identities (26), (27), and (28) become

$$\begin{aligned} \mathbf{z}^{k+1/2} - \mathbf{z}^{k+1} &= \alpha \left(\mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^k) \right) \\ \mathbf{z}^k - \mathbf{z}^{k+1} &= \frac{1}{k+2}(\mathbf{z}^k - \mathbf{z}^0) + \alpha\, \mathbf{G}(\mathbf{z}^{k+1/2}) \\ \mathbf{z}^{k+1} - \mathbf{z}^0 &= \frac{k+1}{k+2}(\mathbf{z}^k - \mathbf{z}^0) - \alpha\, \mathbf{G}(\mathbf{z}^{k+1/2}) \end{aligned}$$

Now, subtracting the same inequalities from monotonicity and Lipschitzness from $V_k - V_{k+1}$ as in Lemma 2, each with coefficients $(k+1)(k+2)$ and $\tau_k \geq 0$ (to be specified later), we obtain

$$\begin{aligned} & V_k - V_{k+1} \\ & \geq V_k - V_{k+1} - (k+1)(k+2) \left\langle \mathbf{z}^k - \mathbf{z}^{k+1}, \mathbf{G}(\mathbf{z}^k) - \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\ & \quad - \tau_k \left(\left\| \mathbf{z}^{k+1/2} - \mathbf{z}^{k+1} \right\|^2 - \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) - \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \right) \\ & = (A_k - \alpha^2 \tau_k) \left\| \mathbf{G}(\mathbf{z}^k) \right\|^2 + \tau_k (1-\alpha^2) \left\| \mathbf{G}(\mathbf{z}^{k+1/2}) \right\|^2 + (\tau_k - A_{k+1}) \left\| \mathbf{G}(\mathbf{z}^{k+1}) \right\|^2 \\ & \quad + \left(2\alpha^2 \tau_k - \alpha(k+1)(k+2) \right) \left\langle \mathbf{G}(\mathbf{z}^k), \mathbf{G}(\mathbf{z}^{k+1/2}) \right\rangle + \left(\alpha(k+2)^2 - 2\tau_k \right) \left\langle \mathbf{G}(\mathbf{z}^{k+1/2}), \mathbf{G}(\mathbf{z}^{k+1}) \right\rangle \\ & = \mathrm{Tr}\left(\mathbf{M}_k \mathbf{S}_k \mathbf{M}_k^\intercal \right) \end{aligned}$$

where we define $\mathbf{M}_k := \begin{bmatrix}\mathbf{G}(\mathbf{z}^k) & \mathbf{G}(\mathbf{z}^{k+1/2}) & \mathbf{G}(\mathbf{z}^{k+1})\end{bmatrix}$ and

$$\mathbf{S}_k := \begin{bmatrix} A_k - \alpha^2\tau_k & \alpha^2\tau_k - \frac{\alpha}{2}(k+1)(k+2) & 0 \\ \alpha^2\tau_k - \frac{\alpha}{2}(k+1)(k+2) & \tau_k(1-\alpha^2) & \frac{\alpha}{2}(k+2)^2 - \tau_k \\ 0 & \frac{\alpha}{2}(k+2)^2 - \tau_k & \tau_k - A_{k+1} \end{bmatrix} \tag{34}$$

If $\mathbf{S}_k \succeq \mathbf{O}$, then $\mathrm{Tr}\left(\mathbf{M}_k\mathbf{S}_k\mathbf{M}_k^\intercal\right) = \mathrm{Tr}\left(\mathbf{S}_k\mathbf{M}_k^\intercal\mathbf{M}_k\right) \geq 0$ because the positive semidefinite cone is self-dual with respect to the matrix inner product $\langle\mathbf{A},\mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}^\intercal\mathbf{B})$. Because $B_k = k+1$ grows linearly, provided that the sequence $\{A_k\}$ grows quadratically, we can derive $\mathcal{O}(1/k^2)$ convergence by using similar line of arguments as in the proof of Theorem 2. This reduction of the proof into a search of appropriate parameters (i.e., τ_k) that meet semidefiniteness constraints ($\mathbf{S}_k \succeq \mathbf{O}$ in our case) while allowing for desired rate of growth in Lyapunov function coefficients (A_k in our case) was inspired by works of [Taylor et al.(2017)] and [Taylor & Bach(2019)]. In the following, we demonstrate that careful choices of A_0 and τ_k make A_k asymptotically close to $\frac{\alpha(k+1)(k+2)}{2}$, so quadratic growth is guaranteed. We begin with the following lemma, which will be used throughout the proof.

Lemma 5. *Let $k \in \mathbb{N}_{\geq 0}$ and $\alpha \in \left(0, \frac{1}{2}\right]$ be fixed, and define*

$$\ell_k := \frac{\alpha(k+2)(k+1+k\alpha)}{2(1+\alpha)}, \quad u_k := \frac{\alpha(k+2)(k+1-k\alpha)}{2(1-\alpha)}$$

Then,

$$\begin{aligned} u_k > \frac{\alpha(k+1)(k+2)}{2} &> \ell_k && (35)\\ &\geq \frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)} && (36)\\ &\geq \frac{\alpha(k+1)^2 - \alpha^3 k(k+2)}{2(1-\alpha^2)} && (37)\\ &\geq \max\left\{\frac{\alpha(k+1)(k+1-\alpha(k+2))}{2(1-\alpha)}, \frac{\alpha^2(k+1)(k+2)}{1+\alpha}\right\} && (38)\\ &\geq \frac{\alpha^2(k+1)(k+2) + \alpha^3(k+2)^2}{2(1+\alpha)} && (39) \end{aligned}$$

We shall prove Lemma 5 after the proof of the main theorem and for now, focus on why we need such results. Observe that all the quantities within the lines (35) through (37) are asymptotically close to $\frac{\alpha k^2}{2}$. We show that $A_k \in I_k := [\ell_k, u_k]$ for all $k \geq 0$, which implies the quadratic growth. The quantities in Lemma 5 are used for choosing the right τ_k and for showing the positive semidefiniteness of $\mathbf{S}_k$.

Subdivide the interval I_k into two parts:

$$I_k^- = \left[\ell_k, \frac{\alpha(k+1)(k+2)}{2}\right], \quad I_k^+ = \left[\frac{\alpha(k+1)(k+2)}{2}, u_k\right]$$

We divide cases: $A_k \in I_k^-$ and $A_k \in I_k^+$. However, the latter case is in fact not needed unless we wish to extend the proof for α beyond $\frac{0.1265}{R}$. If that is not the case, we recommend the readers to refer to Case 1 only. Nevertheless, we exhibit analysis of both cases because Case 2 might provide

useful data for enlarging or even completely determining the range of convergent step-sizes for EAG-C.

Case 1. Suppose that $A_k \in I_k^-$. In this case, we choose

$$\tau_k = \frac{(k+2)^2\left(2(1-\alpha)A_k - \alpha(k+1)(k+1-\alpha(k+2))\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{40}$$

The denominator and numerator of (40) are both positive because $u_k > A_k > \frac{\alpha(k+1)(k+1-\alpha(k+2))}{2(1-\alpha)}$ (see (38)). Thus, $\tau_k > 0$. Next, define A_{k+1} as

$$\begin{aligned} A_{k+1} &= \frac{\alpha(k+2)^2\left(4(1-\alpha)A_k - \alpha(k+1-\alpha(k+2))^2\right)}{4(1-\alpha)\left((1-\alpha)A_k + \alpha^2(k+1)(k+2)\right)} \\ &= \frac{\alpha(k+2)^2}{1-\alpha}\left(1 - \frac{\alpha(k+1+\alpha(k+2))^2}{4((1-\alpha)A_k + \alpha^2(k+1)(k+2))}\right) \end{aligned} \tag{41}$$

Then (34) can be rewritten as

$$\mathbf{S}_k = \begin{bmatrix} s_{11} & s_{12} & 0 \\ s_{12} & s_{22} & s_{23} \\ 0 & s_{23} & s_{33} \end{bmatrix}$$

where

$$s_{11} = \frac{\left(\alpha(k+1)(k+2) - 2A_k\right)\left(2(1-\alpha)A_k + \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{42}$$

$$s_{12} = -\frac{\alpha(1-\alpha)(k+2)(k+1+\alpha(k+2))(\alpha(k+1)(k+2) - 2A_k)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{43}$$

$$s_{22} = \frac{(1-\alpha^2)(k+2)^2\left(2(1-\alpha)A_k - \alpha(k+1)(k+1-\alpha(k+2))\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{44}$$

$$s_{23} = -\frac{(k+2)^2\left(2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3 k(k+2)\right)}{2\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)} \tag{45}$$

$$s_{33} = \frac{(k+2)^2\left(2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3 k(k+2)\right)\left(2(1-\alpha)A_k + \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2\right)}{4(1-\alpha)\left(\alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k\right)\left((1-\alpha)A_k + \alpha^2(k+1)(k+2)\right)} \tag{46}$$

The expressions seem ridiculously complicated, but there are a number of repeating terms. Let

$$\begin{aligned} E_1 &= \alpha(k+2)(k+1-k\alpha) - 2(1-\alpha)A_k \\ E_2 &= \alpha(k+1)(k+2) - 2A_k \end{aligned}$$

Because $A_k \leq \frac{\alpha(k+1)(k+2)}{2} < u_k$ (see (35)), we have $E_1 > 0, E_2 \geq 0$. (Note that $E_2 = 0$ only in the boundary case $A_k = \sup I_k^-$.) Next, put

$$E_3 = 2(1-\alpha)A_k - \alpha(k+1)(k+1-\alpha(k+2))$$

which is a factor that appears within the definition of τ_k (40); we have already seen that $E_3 > 0$. Further, let

$$E_4 = 2(1-\alpha)A_k + \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2$$

$$E_5 = (1-\alpha)A_k + \alpha^2(k+1)(k+2)$$
$$E_6 = 2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3 k(k+2)$$
$$E_7 = k+1+\alpha(k+2).$$

It is obvious that $E_5, E_7 > 0$, and $E_6 > 0$ follows directly from (37). To see that $E_4 > 0$, observe that $k+1-\alpha(k+2) = (k+2)\left(\frac{k+1}{k+2} - \alpha\right) \geq (k+2)\left(\frac{1}{2} - \alpha\right) \geq 0$, provided that $\alpha \leq \frac{1}{2}$. This implies

$$E_4 = 2(1-\alpha)A_k + \alpha^2(k+2)\left(k+1-(k+2)\alpha\right) > 0$$

Now we can rewrite (42) through (46) as

$$s_{11} = \frac{E_2 E_4}{2E_1}$$
$$s_{12} = -\frac{\alpha(1-\alpha)(k+2)E_2 E_7}{2E_1}$$
$$s_{22} = \frac{(1-\alpha^2)(k+2)^2 E_3}{2E_1}$$
$$s_{23} = -\frac{(k+2)^2 E_6}{2E_1}$$
$$s_{33} = \frac{(k+2)^2 E_4 E_6}{4(1-\alpha)E_1 E_5}$$

This immediately shows that the diagonal entries s_{ii} are nonnegative for $i = 1, 2, 3$. By brute-force calculation, it is not difficult to verify the identity

$$(1+\alpha)E_3 E_4 = \alpha^2(1-\alpha)E_2 E_7^2 + 2E_5 E_6$$

Using this, we see that $\mathbf{v} := \begin{bmatrix} \frac{\alpha(k+2)E_7}{2E_5} & \frac{E_4}{2(1-\alpha)E_5} & 1 \end{bmatrix}^\mathsf{T}$ satisfies $\mathbf{S}_k\mathbf{v} = 0$, and this implies $\det \mathbf{S}_k = 0$. The cofactor-expansion of $\det \mathbf{S}_k$ along the first row gives

$$0 = \det \mathbf{S}_k = s_{11}\begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix} - s_{12}\begin{vmatrix} s_{12} & s_{23} \\ 0 & s_{33} \end{vmatrix} \iff \begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix} = \frac{s_{12}^2 s_{33}}{s_{11}} > 0$$

when $s_{11} > 0$, and via continuity argument we can argue that $\begin{vmatrix} s_{22} & s_{23} \\ s_{23} & s_{33} \end{vmatrix} \geq 0$ even in the boundary case $s_{11} = 0$. Similarly one can show that $\begin{vmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{vmatrix} \geq 0$. Therefore, we have shown that all diagonal submatrices of $\mathbf{S}_k$ (including the trivial case $\begin{vmatrix} s_{11} & 0 \\ 0 & s_{33} \end{vmatrix} = s_{11}s_{33} \geq 0$) have nonnegative determinants, that is, $\mathbf{S}_k \succeq \mathbf{O}$.

Finally, (41) shows that A_{k+1} is increasing with respect to A_k. We see that

$$A_{k+1}\Big|_{A_k = \frac{\alpha(k+1)(k+2)}{2}} = \frac{\alpha(k+2)\left((k+1)(k+3) - \alpha^2(k+2)^2\right)}{2(1-\alpha^2)(k+1)} < \frac{\alpha(k+2)(k+3)}{2} \tag{47}$$

and

$$A_{k+1}|_{A_k=\ell_k} - \ell_{k+1} = \frac{\alpha^2\left((1-3\alpha-\alpha^2-\alpha^3)k + 1 - 8\alpha + \alpha^2 - 2\alpha^3\right)}{2(1-\alpha^2)\left((1+\alpha)^2 k + 1 + \alpha + 2\alpha^2\right)}$$

and the last expression is nonnegative because of the assumption (4), which we restate here for the case $R = 1$ for convenience: $1 - 3\alpha - \alpha^2 - \alpha^3 \geq 0$ and $1 - 8\alpha + \alpha^2 - 2\alpha^3 \geq 0$. This proves that $A_{k+1} \in I_{k+1}^- \subset I_{k+1}$, as desired.

Case 2. Suppose that $A_k \in I_k^+$. The proof would be similar to Case 1, but choices of τ_k and A_{k+1} are different. We let

$$\tau_k = \frac{(k+2)^2\left(2(1+\alpha)A_k - \alpha(k+1)(k+1+\alpha(k+2))\right)}{4(1+\alpha)A_k - 2\alpha(k+2)(k+1+k\alpha)} \tag{48}$$

Since $A_k > \ell_k > \frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)}$, the denominator and numerator of (48) are both positive and thus $\tau_k > 0$. Next, let

$$\begin{aligned} A_{k+1} &= \frac{\alpha(k+2)^2\left(4(1+\alpha)A_k - \alpha(k+1+\alpha(k+2))^2\right)}{4(1+\alpha)\left((1+\alpha)A_k - \alpha^2(k+1)(k+2)\right)} \\ &= \frac{\alpha(k+2)^2}{1+\alpha}\left(1 - \frac{\alpha(k+1-\alpha(k+2))^2}{4((1+\alpha)A_k - \alpha^2(k+1)(k+2))}\right) \end{aligned} \tag{49}$$

Then we can check that

$$s_{11} = \frac{(2A_k - \alpha(k+1)(k+2))(2(1+\alpha)A_k - \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2)}{4(1+\alpha)A_k - 2\alpha(k+2)(k+1+k\alpha)}$$

$$s_{33} = \frac{(k+2)^2\left(2(1+\alpha)A_k - \alpha^2(k+1)(k+2) - \alpha^3(k+2)^2\right)\left(2(1-\alpha^2)A_k - \alpha(k+1)^2 + \alpha^3 k(k+2)\right)}{4(1+\alpha)\left(2(1+\alpha)A_k - \alpha(k+2)(k+1+k\alpha)\right)\left(2(1+\alpha)A_k - \alpha^2(k+1)(k+2)\right)}$$

and so on. (Note that $2A_k - \alpha(k+1)(k+2) \geq 0$ because now we are assuming that $A_k \in I_k^+$.) We omit further details of calculations, but with the above choices of τ_k and A_{k+1} it can be shown that $\det \mathbf{S}_k = 0$ and $s_{11}, s_{33} \geq 0$, using (36) through (39). As in Case 1, this implies $\mathbf{S}_k \succeq \mathbf{O}$.

The identity (49) shows that A_{k+1} is increasing with respect to A_k. Interestingly, although (41) and (49) have distinct forms, for the boundary value $A_k = \frac{\alpha(k+1)(k+2)}{2}$, they evaluate to the same expression (47) and thus arguments from Case 1 readily show that $A_{k+1} > \ell_{k+1}$. On the other hand, we have

$$u_{k+1} - A_{k+1}|_{A_k = u_k} = \frac{\alpha^2\left(\left(1 + 3\alpha - \alpha^2 + \alpha^3\right)k + 1 + 8\alpha + \alpha^2 + 2\alpha^3\right)}{2(1-\alpha^2)\left((1-\alpha)^2 k + 1 - \alpha + 2\alpha^2\right)}$$

and the last term is positive for any $\alpha \in (0,1)$, i.e., $A_{k+1} < u_{k+1}$. This completes Case 2.

Proof of the theorem statement. Given that $A_k \in I_k^-$ implies $A_{k+1} \in I_{k+1}^-$ (which has been proved in Case 1), the rest is easy. If we take $A_0 = \ell_0 = \frac{\alpha}{1+\alpha}$, then because $\mathbf{S}_k \succeq \mathbf{O}$ for all $k \geq 0$, we see that V_k is nonincreasing:

$$\frac{\alpha}{1+\alpha}\|\mathbf{z}^0 - \mathbf{z}^\star\|^2 \geq \frac{\alpha}{1+\alpha}\left\|\mathbf{G}(\mathbf{z}^0)\right\|^2 = V_0 \geq \cdots \geq V_k = A_k\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2 + (k+1)\left\langle \mathbf{z}^k - \mathbf{z}^0, \mathbf{G}(\mathbf{z}^k)\right\rangle$$

where the first inequality follows from Lipschitzness of $\mathbf{G}$ (recall that we are assuming that $R = 1$). Also by (35) and (36),

$$A_k \geq \ell_k > \frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)} = \frac{\alpha(k+1)}{2}\frac{(1+\alpha)(k+1)+\alpha}{1+\alpha} > \frac{\alpha(k+1)^2}{2} \tag{50}$$

Hence, we obtain

$$\begin{aligned}\frac{\alpha}{1+\alpha}\|\mathbf{z}^0-\mathbf{z}^\star\|^2 \geq V_k &\geq \ell_k\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2+(k+1)\left\langle \mathbf{z}^k-\mathbf{z}^0,\mathbf{G}(\mathbf{z}^k)\right\rangle\\ &\overset{\text{(a)}}{\geq} \frac{\alpha(k+1)^2}{2}\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2+(k+1)\left\langle \mathbf{z}^\star-\mathbf{z}^0,\mathbf{G}(\mathbf{z}^k)\right\rangle\\ &\overset{\text{(b)}}{\geq} \frac{\alpha(k+1)^2}{2}\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2-(k+1)\left(\frac{1}{\alpha(k+1)}\|\mathbf{z}^\star-\mathbf{z}^0\|^2+\frac{\alpha(k+1)}{4}\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2\right)\end{aligned}$$

where (a) follows from (50) and the monotonicity inequality $\langle \mathbf{z}^k-\mathbf{z}^\star,\mathbf{G}(\mathbf{z}^k)\rangle\geq 0$, and (b) follows from Young's inequality. Rearranging terms, we conclude that

$$\left\|\mathbf{G}(\mathbf{z}^k)\right\|^2\leq\frac{4}{\alpha(k+1)^2}\left(\frac{\alpha}{1+\alpha}+\frac{1}{\alpha}\right)\|\mathbf{z}^0-\mathbf{z}^\star\|^2=\frac{C\|\mathbf{z}^0-\mathbf{z}^\star\|^2}{(k+1)^2}$$

where $C=\frac{4(1+\alpha+\alpha^2)}{\alpha^2(1+\alpha)}$.

Proof of Lemma 5. Direct calculation gives

$$\begin{aligned}u_k-\frac{\alpha(k+1)(k+2)}{2}&=\frac{\alpha^2(k+2)}{2(1-\alpha)}>0\\ \frac{\alpha(k+1)(k+2)}{2}-\ell_k&=\frac{\alpha^2(k+2)}{2(1+\alpha)}>0\end{aligned}$$

showing (35). Next,

$$\ell_k-\frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)}=\frac{\alpha(k+1-\alpha(k+2))}{2(1+\alpha)}\geq 0$$

because $k+1-\alpha(k+2)=(k+2)(\frac{k+1}{k+2}-\alpha)\geq(k+2)(\frac{1}{2}-\alpha)\geq 0$, which shows (36). Similarly, we observe that

$$\begin{aligned}\frac{\alpha(k+1)(k+1+\alpha(k+2))}{2(1+\alpha)}-\frac{\alpha(k+1)^2-\alpha^3k(k+2)}{2(1-\alpha^2)}&=\frac{\alpha^2(k+1-\alpha(k+2))}{2(1-\alpha^2)}\geq 0\\ \frac{\alpha(k+1)^2-\alpha^3k(k+2)}{2(1-\alpha^2)}-\frac{\alpha(k+1)(k+1-\alpha(k+2))}{2(1-\alpha)}&=\frac{\alpha^2(k+1+\alpha(k+2))}{2(1-\alpha^2)}>0\\ \frac{\alpha(k+1)^2-\alpha^3k(k+2)}{2(1-\alpha^2)}-\frac{\alpha^2(k+1)(k+2)}{1+\alpha}&=\frac{\alpha(k+1-\alpha(k+2))^2}{2(1-\alpha^2)}\geq 0\\ \frac{\alpha^2(k+1)(k+2)}{1+\alpha}-\frac{\alpha^2(k+1)(k+2)+\alpha^3(k+2)^2}{2(1+\alpha)}&=\frac{\alpha^2(k+2)(k+1-\alpha(k+2))}{2(1+\alpha)}\geq 0\end{aligned}$$

and each line corresponds to an inequality within (37), (38) and (39).

C Omitted proofs of Section 3

In this section, we provide a self-contained discussion on the complexity lower bound results for linear operator equations from [Nemirovsky(1991), Nemirovsky(1992)].

C.1 Proof of Theorem 3

The proof of Theorem 3 was essentially completed in the main body of the paper, except the argument regarding translation, (13), and the proof of Lemma 3.

We first provide the precise meaning of the translation invariance that we are to prove. Given a saddle function $\mathbf{L}$ and $\mathbf{z} \in \mathbb{R}^n \times \mathbb{R}^n$, let $\mathbf{z}^\star_{\mathbf{L}}(\mathbf{z})$ be the saddle point of $\mathbf{L}$ nearest to $\mathbf{z}$. For any $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^n$, $k \geq 0$ and $D > 0$, define

$$\mathfrak{T}\left(\mathbf{z}^0; k, D\right) := \left\{ \mathbf{z}^k \;\middle|\; \begin{array}{l} \mathbf{L}(\mathbf{x},\mathbf{y}) = \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} - \mathbf{c} \rangle,\ \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{b},\mathbf{c} \in \mathbb{R}^n,\ \left\|\mathbf{z}^\star_{\mathbf{L}}\left(\mathbf{z}^0\right) - \mathbf{z}^0\right\| \leq D, \\ \mathbf{z}^j = \mathcal{A}(\mathbf{z}^0, \ldots, \mathbf{z}^{j-1}; \mathbf{L}),\ j = 1, \ldots, k,\ \mathcal{A} \in \mathfrak{A}_{\text{sep}} \end{array} \right\}$$

We will show that

$$\mathfrak{T}\left(\mathbf{z}^0; k, D\right) = \mathbf{z}^0 + \mathfrak{T}\left(0; k, D\right)$$

holds for any $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^n$.

Let $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0)$ and $\mathbf{L}(\mathbf{x},\mathbf{y}) = \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y} - \mathbf{c} \rangle$ be given, and assume that $\|\mathbf{z}^\star_{\mathbf{L}}(\mathbf{z}^0) - \mathbf{z}^0\| \leq D$. Let $\mathbf{b}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}^0$ and $\mathbf{c}_0 = \mathbf{c} - \mathbf{y}^0$. Then

$$\begin{aligned} \nabla_{\mathbf{x}} \mathbf{L}_0(\mathbf{x}^0, \mathbf{y}^0) &= \mathbf{A}^\intercal(\mathbf{y}^0 - \mathbf{c}) = -\mathbf{A}^\intercal \mathbf{c}_0 \\ \nabla_{\mathbf{y}} \mathbf{L}_0(\mathbf{x}^0, \mathbf{y}^0) &= \mathbf{A}\mathbf{x}^0 - \mathbf{b} = -\mathbf{b}_0 \end{aligned}$$

Hence, (11) with $k = 1$ reads as

$$\begin{aligned} \mathbf{x}^1 - \mathbf{x}^0 &\in \operatorname{span}\{\mathbf{A}^\intercal \mathbf{c}_0\} \triangleq \mathcal{X}_1(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \\ \mathbf{y}^1 - \mathbf{y}^0 &\in \operatorname{span}\{\mathbf{b}_0\} \triangleq \mathcal{Y}_1(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \end{aligned}$$

This further shows that

$$\begin{aligned} \nabla_{\mathbf{x}} \mathbf{L}_0(\mathbf{x}^1, \mathbf{y}^1) &= \mathbf{A}^\intercal(\mathbf{y}^1 - \mathbf{c}) = \mathbf{A}^\intercal(\mathbf{y}^1 - \mathbf{y}^0) - \mathbf{A}^\intercal \mathbf{c}_0 \in \operatorname{span}\{\mathbf{A}^\intercal \mathbf{b}_0, \mathbf{A}^\intercal \mathbf{c}_0\} \\ \nabla_{\mathbf{y}} \mathbf{L}_0(\mathbf{x}^1, \mathbf{y}^1) &= \mathbf{A}\mathbf{x}^1 - \mathbf{b} = \mathbf{A}(\mathbf{x}^1 - \mathbf{x}^0) - \mathbf{b}_0 \in \operatorname{span}\{\mathbf{A}(\mathbf{A}^\intercal \mathbf{c}_0), \mathbf{b}_0\} \end{aligned}$$

and (11) with $k = 2$ becomes

$$\begin{aligned} \mathbf{x}^2 - \mathbf{x}^0 &\in \operatorname{span}\{\mathbf{A}^\intercal \mathbf{c}_0, \mathbf{A}^\intercal \mathbf{b}_0\} \triangleq \mathcal{X}_2(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \\ \mathbf{y}^2 - \mathbf{y}^0 &\in \operatorname{span}\{\mathbf{b}_0, \mathbf{A}\mathbf{A}^\intercal \mathbf{c}_0\} \triangleq \mathcal{Y}_2(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \end{aligned}$$

As one can see, we have $\mathbf{x}^k - \mathbf{x}^0 \in \mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0)$ and $\mathbf{y}^k - \mathbf{y}^0 \in \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0)$, where we inductively define

$$\begin{aligned} \mathcal{X}_{k+1}(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\{\mathbf{A}^\intercal \mathbf{c}_0\} + \mathbf{A}^\intercal \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \\ \mathcal{Y}_{k+1}(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\{\mathbf{b}_0\} + \mathbf{A}\mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) \end{aligned}$$

Then it is not difficult to see that for $k \geq 2$,

$$\begin{aligned} \mathcal{X}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\left\{\mathbf{A}^\intercal \mathbf{c}_0, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)\mathbf{c}_0, \ldots, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k-1}{2} \rfloor} \mathbf{c}_0\right\} + \operatorname{span}\left\{\mathbf{A}^\intercal \mathbf{b}_0, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)\mathbf{b}_0, \ldots, \mathbf{A}^\intercal(\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k}{2} \rfloor - 1} \mathbf{b}_0\right\} \\ \mathcal{Y}_k(\mathbf{A}; \mathbf{b}_0, \mathbf{c}_0) &= \operatorname{span}\left\{\mathbf{b}_0, (\mathbf{A}\mathbf{A}^\intercal)\mathbf{b}_0, \ldots, (\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k-1}{2} \rfloor} \mathbf{b}_0\right\} + \operatorname{span}\left\{\mathbf{A}\mathbf{A}^\intercal \mathbf{c}_0, \ldots, (\mathbf{A}\mathbf{A}^\intercal)^{\lfloor \frac{k}{2} \rfloor} \mathbf{c}_0\right\} \end{aligned}$$

Now consider $\mathbf{L}_0(\mathbf{x},\mathbf{y}) := \langle \mathbf{A}\mathbf{x}-\mathbf{b}_0, \mathbf{y}-\mathbf{c}_0\rangle = \big\langle \mathbf{A}(\mathbf{x}+\mathbf{x}^0)-\mathbf{b}, \mathbf{y}+\mathbf{y}^0-\mathbf{c}\big\rangle$. Because $\mathbf{z}^\star_{\mathbf{L}_0}$ is a saddle point of $\mathbf{L}_0$ if and only if $\mathbf{z}^\star_{\mathbf{L}_0}+\mathbf{z}^0$ is a saddle point of $\mathbf{L}$, we have $\mathbf{z}^\star_{\mathbf{L}_0}(0) = \mathbf{z}^\star_{\mathbf{L}}(\mathbf{z}^0) - \mathbf{z}^0$, and thus $\|\mathbf{z}^\star_{\mathbf{L}_0}(0)\| \le D$. Therefore, if we let

$$\mathcal{S}(\mathbf{A}; D) \triangleq \left\{ (\widetilde{\mathbf{b}}, \widetilde{\mathbf{c}}) \in \mathbb{R}^n \times \mathbb{R}^n \;\middle|\; \left\| \mathbf{z}^\star_{\widetilde{\mathbf{L}}}(0) \right\| \le D, \text{ where } \widetilde{\mathbf{L}}(\mathbf{x},\mathbf{y}) = \langle \mathbf{A}\mathbf{x}-\widetilde{\mathbf{b}}, \mathbf{y}-\widetilde{\mathbf{c}}\rangle \right\}$$

then

$$\mathfrak{T}\left(\mathbf{z}^0; k, D\right) = \bigcup_{\substack{\mathbf{A}\in\mathbb{R}^{n\times n} \\ (\mathbf{b}_0,\mathbf{c}_0)\in\mathcal{S}(\mathbf{A};D)}} \mathbf{z}^0 + (\mathcal{X}_k(\mathbf{A};\mathbf{b}_0,\mathbf{c}_0) \times \mathcal{Y}_k(\mathbf{A};\mathbf{b}_0,\mathbf{c}_0))$$

This proves that the translation invariance holds with $\mathfrak{T}(0;k,D) = \bigcup_{\substack{\mathbf{A}\in\mathbb{R}^{n\times n} \\ (\mathbf{b}_0,\mathbf{c}_0)\in\mathcal{S}(\mathbf{A};D)}} (\mathcal{X}_k(\mathbf{A};\mathbf{b}_0,\mathbf{c}_0) \times \mathcal{Y}_k(\mathbf{A};\mathbf{b}_0,\mathbf{c}_0))$ and in particular, shows (13).

C.2 Complexity of solving linear operator equations and minimax polynomials

We first make some general observations. Suppose that we are given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, and an integer $k \ge 1$. Then any $\mathbf{x} \in \mathcal{K}_{k-1}(\mathbf{A};\mathbf{b}) = \operatorname{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\}$ can be expressed in the form

$$\mathbf{x} = q(\mathbf{A})\mathbf{b}, \quad \text{where } q(t) = q_0 + q_1 t + \cdots + q_{k-1}t^{k-1}$$

for some $q_0, \ldots, q_{k-1} \in \mathbb{R}$. Then we can write

$$\mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{b} - \mathbf{A}q(\mathbf{A})\mathbf{b} = (\mathbf{I} - \mathbf{A}q(A))\mathbf{b} = p(\mathbf{A})\mathbf{b} \tag{51}$$

where $p(t) = 1 - tq(t)$ is a polynomial of degree at most k satisfying $p(0) = 1$. Note that conversely, given any polynomial $\widetilde{p}(t)$ with degree $\le k$ and constant term 1, one can decompose it as $\widetilde{p}(t) = 1 - t\widetilde{q}(t)$ and recover a polynomial $\widetilde{q}$ of degree $\le k-1$ corresponding to $\mathbf{x}$.

Now suppose further there exists $\mathbf{x}^\star \in \mathbb{R}^n$ such that $\mathbf{b} = \mathbf{A}\mathbf{x}^\star$ and $\|\mathbf{x}^\star\| \le D$. The symmetric matrix $\mathbf{A}$ has an orthonormal eigenbasis $\mathbf{v}_1, \ldots, \mathbf{v}_n$, corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$, so we can write $\mathbf{x}^\star = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$ for some $c_1, \ldots, c_n \in \mathbb{R}$. Using (51), we obtain

$$\begin{aligned} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 = \|p(\mathbf{A})\mathbf{A}\mathbf{x}^\star\|^2 = \left\| \sum_{j=1}^n c_j \mathbf{A}p(\mathbf{A})\mathbf{v}_j \right\|^2 &= \left\| \sum_{j=1}^n c_j \lambda_j p(\lambda_j)\mathbf{v}_j \right\|^2 \\ &= \sum_{j=1}^n c_j^2 \lambda_j^2 p(\lambda_j)^2 \le D^2 \left(\max_{j=1,\ldots,n} \lambda_j^2 p(\lambda_j)^2 \right) \end{aligned} \tag{52}$$

We define the problem class by $\|\mathbf{A}\| \le R$, which is equivalent to $\lambda_j \in [-R, R]$ for all $j = 1, \ldots, n$. Therefore, we consider a method corresponding to a polynomial $q(t)$ such that $p(t) = 1 - tq(t)$ minimizes

$$\max_{\lambda\in[-R,R]} \lambda^2 p(\lambda)^2 = \left(\max_{\lambda\in[-R,R]} |\lambda p(\lambda)| \right)^2$$

More precisely, if $p_k^\star(t) = 1 - tq_k^\star(t)$ minimizes the last quantity among all $p(t)$ such that $\deg p \le k$ and $p(0) = 1$, and if we put $\mathbf{x}^k = q_k^\star(\mathbf{A})\mathbf{b}$, then (52) implies

$$\left\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\right\|^2 = \sum_{j=1}^{n} c_j^2 \lambda_j^2 \left(p_k^\star(\lambda_j)\right)^2 \le D^2 M^\star(k,R)^2$$

$$M^\star(k,R) \triangleq \min_{\substack{\deg p \le k \\ p(0)=1}} \max_{\lambda \in [-R,R]} |\lambda p(\lambda)| \tag{53}$$

for all $\mathbf{A}$ whose spectrum belongs to $[-R,R]$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\star$ with $\|\mathbf{x}^\star\| \le D$. As $p_k^\star$ solves (53), it is called a *minimax polynomial*.

In order to establish Lemma 3, we present a two-fold analysis in the following. First, we compute the quantity (53) by explicitly naming $p_k^\star$ for each $k \ge 1$. (This was given by [Nemirovsky(1992)], but without a proof.) Then, following the exposition from [Nemirovsky(1991)], we show that there exists an instance of $(\mathbf{A}, \mathbf{b})$ such that

$$\|\mathbf{A}q(\mathbf{A})\mathbf{b} - \mathbf{b}\|^2 \ge D^2 M^\star(k,R)^2$$

holds for any polynomial q of degree $\le k-1$.

C.3 Proof of Lemma 3

The solutions to (53) are characterized using the *Chebyshev polynomials of first kind*, defined by

$$T_N(\cos\theta) = \cos(N\theta), \quad N \ge 1$$

or equivalently by $T_N(t) = \cos(N \arccos t)$. If $N = 2d$ for some nonnegative integer d, then T_N is an even polynomial satisfying $T_N(0) = \cos(d\pi) = (-1)^d$. On the other hand, if $N = 2d+1$, then T_N is an odd polynomial of the form

$$T_{2d+1}(t) = (-1)^d(2d+1)t + \cdots \tag{54}$$

which can be shown via induction using the recurrence relation $T_{N+1}(t) = 2tT_N(t) - T_{N-1}(t)$, which follows from the trigonometric identity

$$\cos((N+1)\theta) + \cos((N-1)\theta) = 2\cos(N\theta)\cos\theta$$

Based on arguments from [Nemirovsky(1992), Mason & Handscomb(2002)], we will show that given $k \ge 1$ and $m := \lfloor \frac{k}{2} \rfloor$,

$$p_k^\star(t) := \frac{(-1)^m}{2m+1} \left(\frac{R}{t}\right) T_{2m+1}\left(\frac{t}{R}\right)$$

solves (53).

The Chebyshev polynomials satisfy the *equioscillation property* which makes them so special: the extrema of T_N within $[-1,1]$ occur at $t_j = \cos\frac{(N-j)\pi}{N}$ for $j = 0, \ldots, N$, and the signs of the extremal values alternate. Indeed, we have $|T_N(t) = \cos(N \arccos t)| \le 1$ for all $t \in [-1,1]$, and for each $j = 0, \ldots, N$,

$$T_N(t_j) = \cos\left(N\frac{(N-j)\pi}{N}\right) = \cos(N-j)\pi = (-1)^{N-j}$$

Also, we have $T_N(t_j) = -T_N(t_{j-1})$ for each $j = 1, \ldots, n$.

Given $k \geq 1$, we denote by $\mathcal{P}_k$ the collection of all polynomials p of degree $\leq k$ with $p(0) = 1$. Recall that we are to minimize

$$M(p, R) := \max_{\lambda \in [-R,R]} |\lambda\, p(\lambda)| \tag{55}$$

over $p \in \mathcal{P}_k$. If $p \in \mathcal{P}_k$ minimizes (55), then so does $p_{\mathrm{ev}}(t) := \frac{p(t)+p(-t)}{2}$, since for all $\lambda \in [-R, R]$

$$|\lambda p_{\mathrm{ev}}(\lambda)| = |\lambda| \cdot \left|\frac{p(\lambda) + p(-\lambda)}{2}\right| \leq \frac{|\lambda p(\lambda)|}{2} + \frac{|(-\lambda)p(-\lambda)|}{2} \leq \frac{M(p,R)}{2} + \frac{M(p,R)}{2} = M(p,R) \tag{56}$$

holds, which implies that $M(p_{\mathrm{ev}}, R) \leq M(p, R)$.

Observe that $p_k^\star \in \mathcal{P}_k$ due to (54). Next, note that $\lambda p_k^\star(\lambda) = \frac{(-1)^m R}{2m+1} T_{2m+1}(\frac{\lambda}{R})$ has extrema of alternating signs and same magnitude within $[-R, R]$, which occur precisely at $\lambda_j := R\cos\frac{(2m+1-j)\pi}{2m+1}$, where $j = 0, \ldots, 2m+1$. Suppose that $p_k^\star$ is not a minimizer of $M(p, R)$ over $\mathcal{P}_k$, so that there exists $p \in \mathcal{P}_k$ such that

$$|\lambda_j p(\lambda_j)| \leq M(p,R) < M(p_k^\star, R) = |\lambda_j p_k^\star(\lambda_j)| \quad (j = 0, \ldots, 2m+1) \tag{57}$$

Due to (56), by replacing p with p_{ev} if necessary, we may assume that p is even and has degree $\leq 2m$. Since $\lambda_j \neq 0$ for all $j = 0, \ldots, 2m+1$, the condition (57) reduces to $|p(\lambda_j)| < |p_k^\star(\lambda_j)|$.

As p and $p_k^\star$ are both polynomials of degree $\leq 2m$ and constant terms 1, we can write

$$p_k^\star(\lambda) - p(\lambda) = \lambda q(\lambda)$$

for some polynomial q of degree $\leq 2m-1$. But then $|p(\lambda_j)| = |p_k^\star(\lambda_j) - \lambda_j q(\lambda_j)| < |p_k^\star(\lambda_j)|$, which implies that $p_k^\star(\lambda_j)$ and $\lambda_j q(\lambda_j)$ have same signs for $j = 0, \ldots, 2m+1$. Now, because $p_k^\star(\lambda_j)$ have alternating signs and

$$\lambda_0 < \cdots < \lambda_m < 0 < \lambda_{m+1} < \cdots < \lambda_{2m+1}$$

we see that the signs of $q(\lambda_j)$ alternate over $j = 0, \ldots, m$ and over $j = m+1, \ldots, 2m+1$, respectively. Therefore, q must have at least one zero in each open interval $(\lambda_j, \lambda_{j+1})$ for $j = 0, \ldots, m-1, m+1, \ldots, 2m$. This implies that $q(t) \equiv 0$ since $\deg q \leq 2m-1$, while q has at least $2m$ zeros. Therefore, we arrive at $p_k^\star = p$, which is a contradiction.

We have established that

$$M^\star(k,R) = M(p_k^\star, R) = |\lambda_j p_k^\star(\lambda_j)| = \frac{R}{2m+1} = \frac{R}{2\lfloor k/2 \rfloor + 1} \quad (j = 0, \ldots, 2m+1) \tag{58}$$

Furthermore, the above arguments show that the minimization of (55) over $p \in \mathcal{P}_k$ is in fact the same as the minimization of

$$\max_{j=0,\ldots,2m+1} |\lambda_j p(\lambda_j)| = \max_{\lambda \in \Lambda} |\lambda p(\lambda)|, \quad \Lambda := \{\lambda_0, \lambda_1, \ldots, \lambda_{2m+1}\} \tag{59}$$

Note that the trick of replacing p by p_{ev} is still applicable to (59), but only because the set Λ is symmetric with respect to the origin. Now we can write

$$M^\star(k,R)^2 = \left(\min_{p \in \mathcal{P}_k} \max_{\lambda \in [-R,R]} |\lambda p(\lambda)|\right)^2 = \left(\min_{p \in \mathcal{P}_k} \max_{\lambda \in \Lambda} |\lambda p(\lambda)|\right)^2 = \min_{p \in \mathcal{P}_k} \max_{\lambda \in \Lambda} \lambda^2 p(\lambda)^2 \tag{60}$$

and the final problem from the line (60) is equivalent to

$$\begin{array}{ll} \underset{\nu\in\mathbb{R},\, p\in\mathcal{P}_k}{\text{minimize}} & \nu \\ \text{subject to} & \lambda_j^2 p(\lambda_j)^2 \le \nu, \quad j = 0, \ldots, 2m+1. \end{array} \tag{61}$$

We can identify any $p(t) = 1 + p_1 t + \cdots + p_k t^k \in \mathcal{P}_k$ as the vector $(p_1, \ldots, p_k) \in \mathbb{R}^k$. Under this identification, (61) is a second order cone program (as the constraints are convex quadratic in $p_1, \ldots, p_k$), and Slater's constraint qualification is clearly satisfied. Hence $M^\star(k, R)^2$ equals the optimal value of the dual problem

$$\begin{array}{lll} \underset{\boldsymbol{\mu}\in\mathbb{R}^{2m+2}}{\text{maximize}} & \underset{p\in\mathcal{P}_k}{\text{minimize}} & \sum_{j=0}^{2m+1} \mu_j \lambda_j^2 p(\lambda_j)^2 \\ \text{subject to} & & \sum_{j=0}^{2m+1} \mu_j = 1, \\ & & \boldsymbol{\mu} \ge 0. \end{array} \tag{62}$$

Let $\boldsymbol{\mu}^\star = (\mu_0^\star, \ldots, \mu_{2m+1}^\star)$ be the dual optimal solution to (62). Provided that $n \ge k+2 \ge 2m+2$, we can take standard basis vectors (with 0-indexing) $\mathbf{e}_0, \ldots, \mathbf{e}_{2m+1} \in \mathbb{R}^n$. Define $\mathbf{A}$ by

$$\mathbf{A}\mathbf{e}_j = \lambda_j \mathbf{e}_j \quad (j = 0, \ldots, 2m+1), \quad \mathbf{A}\mathbf{v} = 0 \quad (\mathbf{v} \perp \operatorname{span}\{\mathbf{e}_0, \ldots, \mathbf{e}_{2m+1}\})$$

and let

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\star, \quad \mathbf{x}^\star = D \sum_{j=0}^{2m+1} \left(\mu_j^\star\right)^{1/2} \mathbf{e}_j$$

so that $\|\mathbf{x}^\star\| = D$. For any given $\mathbf{x} = q(\mathbf{A})\mathbf{b}$ with $\deg q \le k-1$, we use (52) to rewrite $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ as

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 \left(1 - \lambda_j q(\lambda_j)\right)^2 = D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p(\lambda_j)^2$$

where $p(t) = 1 - tq(t) \in \mathcal{P}_k$. But since $(p_k^\star, \boldsymbol{\mu}^\star)$ is the primal-dual solution pair to the problems (61) and (62), $p_k^\star$ minimizes $\sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p(\lambda_j)^2$ within $\mathcal{P}_k$. Therefore,

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p(\lambda_j)^2 \ge D^2 \sum_{j=0}^{2m+1} \mu_j^\star \lambda_j^2 p_k^\star(\lambda_j)^2 = D^2 M^\star(k, R)^2 = \frac{R^2 D^2}{2(\lfloor k/2 \rfloor + 1)^2}$$

which establishes (14).

C.4 Proof of Lemma 4

Let $k \ge 0$ be a given (fixed) integer. Consider the polynomial $p_k^\star$ we defined in the previous section. It is an even polynomial of degree $2\lfloor \frac{k}{2} \rfloor$, and thus $p_k^\star\left(\sqrt{t}\right)$ is a polynomial in t of degree $\lfloor \frac{k}{2} \rfloor$, whose constant term is $p_k^\star(0) = 1$. Therefore, we can write $p_k^\star\left(\sqrt{t}\right) = 1 - tq_k(t)$ for some polynomial q_k. We will show that

$$\mathbf{z}^k = q_k\left(\mathbf{B}^\mathsf{T}\mathbf{B}\right)\mathbf{B}^\mathsf{T}\mathbf{v} \tag{63}$$

satisfies $\|\mathbf{B}\mathbf{z}^k - \mathbf{v}\|^2 \le \frac{R^2D^2}{2(\lfloor k/2 \rfloor + 1)^2}$ for any (possibly non-symmetric) $\mathbf{B} \in \mathbb{R}^{m\times m}$ and $\mathbf{v} = \mathbf{B}\mathbf{z}^\star$ satisfying $\|\mathbf{B}\| \le R$ and $\|\mathbf{z}^\star\| \le D$. The equation (63) defines an algorithm within the class $\mathfrak{A}_{\text{lin}}$, as q_k is of degree $\lfloor \frac{k}{2} \rfloor - 1$, so that $\mathbf{z}^k$ is determined by $2\lfloor \frac{k}{2} \rfloor - 1 \le k - 1$ queries to the matrix multiplication oracle.

We proceed via arguments similar to derivations in C.2. First, observe that

$$\left\|\mathbf{B}\mathbf{z}^k - \mathbf{v}\right\|^2 = \left\|\mathbf{B}\mathbf{z}^k - \mathbf{B}\mathbf{z}^\star\right\|^2 = (\mathbf{z}^k - \mathbf{z}^\star)^\intercal \mathbf{B}^\intercal \mathbf{B}(\mathbf{z}^k - \mathbf{z}^\star) = (\mathbf{z}^k - \mathbf{z}^\star)^\intercal |\mathbf{B}|^2 (\mathbf{z}^k - \mathbf{z}^\star) = \left\||\mathbf{B}|\mathbf{z}^k - |\mathbf{B}|\mathbf{z}^\star\right\|^2 \tag{64}$$

where $|\mathbf{B}|$ is the matrix square root of the positive semidefinite matrix $\mathbf{B}^\intercal\mathbf{B}$. Rewriting (63) in terms of $|\mathbf{B}|$, we obtain

$$\mathbf{z}^k = q_k\left(\mathbf{B}^\intercal\mathbf{B}\right)\mathbf{B}^\intercal\mathbf{B}\mathbf{z}^\star = q_k\left(|\mathbf{B}|^2\right)|\mathbf{B}|^2\mathbf{z}^\star$$

Plugging the last equation into (64) gives

$$\left\||\mathbf{B}|\mathbf{z}^\star - |\mathbf{B}|\mathbf{z}^k\right\|^2 = \left\|\left(\mathbf{I} - |\mathbf{B}|^2 q_k\left(|\mathbf{B}|^2\right)\right)|\mathbf{B}|\mathbf{z}^\star\right\|^2 = \left\|p_k^\star\left(|\mathbf{B}|\right)|\mathbf{B}|\mathbf{z}^\star\right\|^2$$

Finally, because $|\mathbf{B}|$ is a symmetric matrix whose eigenvalues are within $[0, R]$, we can apply (52) with $|\mathbf{B}|, \mathbf{z}^\star$ in places of $\mathbf{A}, \mathbf{x}^\star$, and use (58) to conclude that

$$\left\||\mathbf{B}|\mathbf{z}^\star - |\mathbf{B}|\mathbf{z}^k\right\|^2 \le D^2\left(\max_{\lambda\in[0,R]} \lambda^2 p_k^\star(\lambda)^2\right) \le D^2\left(\max_{\lambda\in[-R,R]} \lambda^2 p_k^\star(\lambda)^2\right) = \frac{R^2D^2}{(2\lfloor k/2 \rfloor + 1)^2}$$

C.5 Proof of Theorem 4

We first describe the general class $\mathfrak{A}$ of algorithms without the linear span assumption. An algorithm $\mathcal{A}$ within $\mathfrak{A}$ is a sequence of deterministic functions $\mathcal{A}_1, \mathcal{A}_2, \ldots$, each of which having the form

$$(\mathbf{z}^i, \overline{\mathbf{z}}^i) = \mathcal{A}_i\left(\mathbf{z}^0, \mathcal{O}(\mathbf{z}^0; \mathbf{L}), \ldots, \mathcal{O}(\mathbf{z}^{i-1}; \mathbf{L}); \mathbf{L}\right)$$

for $i \ge 1$, where $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^n \times \mathbb{R}^m$ is an initial point and $\mathcal{O}\colon (\mathbb{R}^n \times \mathbb{R}^m) \times \mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^m) \to \mathbb{R}^n \times \mathbb{R}^m$ is the gradient oracle defined as

$$\mathcal{O}((\mathbf{x}, \mathbf{y}); \mathbf{L}) = (\nabla_{\mathbf{x}}\mathbf{L}(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}}\mathbf{L}(\mathbf{x}, \mathbf{y}))$$

The sequence $\{\mathbf{z}^i\}_{i\ge 0}$ are the *inquiry points*, and $\{\overline{\mathbf{z}}^i\}_{i\ge 0}$ are the *approximate solutions* produced by $\mathcal{A}$. When $k \ge 1$ is the predefined maximum number of iterations, then we assume $\overline{\mathbf{z}}^k = \mathbf{z}^k$ without loss of generality. Similar definitions for deterministic algorithms have been considered in [Nemirovsky(1991), Ouyang & Xu(2021)].

To clarify, given $\mathbf{L} \in \mathcal{L}_R(\mathbb{R}^n \times \mathbb{R}^m)$, an algorithm $\mathcal{A}$ uses only the previous oracle information to choose the next inquiry point and approximate solution. Therefore, if $\mathcal{O}(\mathbf{z}^i; \mathbf{L}_1) = \mathcal{O}(\mathbf{z}^i; \mathbf{L}_2)$ for all $i = 0, \ldots, k-1$, then the algorithm output $(\mathbf{z}^k, \overline{\mathbf{z}}^k)$ for the two functions will coincide, even if $\mathbf{L}_1 \ne \mathbf{L}_2$. In that sense, $\mathcal{A}$ is *deterministic*, *black-box*, and *gradient-based*.

Now we precisely restate Theorem 4.

. *Let $k \ge 1$ and $n \ge 3k + 2$. Let $\mathcal{A} \in \mathfrak{A}$ be a deterministic black-box gradient-based algorithm for solving convex-concave minimax problems on $\mathbb{R}^n \times \mathbb{R}^n$. Then for any initial point $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^n$, there exists $\mathbf{L} \in \mathcal{L}_R^{\text{biaff}}(\mathbb{R}^n \times \mathbb{R}^n)$ with a saddle point $\mathbf{z}^\star$, for which $\mathbf{z}^k$, the k-th iterate produced by $\mathcal{A}$, satisfies*

$$\|\nabla\mathbf{L}(\mathbf{z}^k)\| \ge \frac{\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(2\lfloor k/2 \rfloor + 1)^2}$$

Proof. Let $\mathbf{z}^0 = (\mathbf{x}^0, \mathbf{y}^0) \in \mathbb{R}^n \times \mathbb{R}^n$ be given. Take $\mathbf{A}$ and $\mathbf{b}$ as in Lemma 3. Denote by $\mathbf{x}^{\min}$ the minimum norm solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$. Recall the construction of $\mathbf{A}$ and $\mathbf{b}$, where $\mathcal{R}(\mathbf{A}) = \operatorname{span}\{\mathbf{e}_0, \ldots, \mathbf{e}_{2m+1}\} \perp \ker(\mathbf{A})$. Define

$$\mathbf{L}_0(\mathbf{x}^0, \mathbf{y}^0) = -\mathbf{b}^\intercal(\mathbf{x} - \mathbf{x}^0) + (\mathbf{x} - \mathbf{x}^0)^\intercal \mathbf{A}(\mathbf{y} - \mathbf{y}^0) - \mathbf{b}^\intercal(\mathbf{y} - \mathbf{y}^0)$$

Then $(\nabla_{\mathbf{x}} \mathbf{L}_0(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} \mathbf{L}_0(\mathbf{x}, \mathbf{y})) = \big(\mathbf{A}(\mathbf{y} - \mathbf{y}^0) - \mathbf{b}, \mathbf{A}(\mathbf{x} - \mathbf{x}^0) - \mathbf{b}\big)$, and $\mathbf{z}^0 + \big(\mathbf{x}^{\min}, \mathbf{x}^{\min}\big)$ is a saddle point of $\mathbf{L}_0$.

We follow the oracle-resisting proof strategy of [Nemirovsky(1991)], described as follows. For each $i = 1, \ldots, k$, we inductively define a *rotated* biaffine function

$$\mathbf{L}_i(\mathbf{x}^0, \mathbf{y}^0) = -\mathbf{b}^\intercal(\mathbf{x} - \mathbf{x}^0) + (\mathbf{x} - \mathbf{x}^0)^\intercal \mathbf{A}_i(\mathbf{y} - \mathbf{y}^0) - \mathbf{b}^\intercal(\mathbf{y} - \mathbf{y}^0)$$

where $\mathbf{A}_i = \mathbf{U}_i \mathbf{A} \mathbf{U}_i^\intercal$ for an orthogonal matrix $\mathbf{U}_i \in \mathbb{R}^{n \times n}$. We will show that U_i can be chosen to satisfy $\mathbf{U}_i \mathbf{b} = \mathbf{b}$,

$$\mathcal{O}(\mathbf{z}^j; \mathbf{L}_i) = \mathcal{O}(\mathbf{z}^j; \mathbf{L}_{i-1}) \tag{65}$$

for $j = 0, \ldots, i-1$, and

$$\mathbf{x}^j - \mathbf{x}^0, \mathbf{y}^j - \mathbf{y}^0 \in \mathcal{K}_{j-1}(\mathbf{A}_i; \mathbf{b}) \oplus \mathbf{U}_i \mathcal{N}_i = \mathbf{U}_i \mathcal{K}_{j-1}(\mathbf{A}; \mathbf{b}) \oplus \mathbf{U}_i \mathcal{N}_i \tag{66}$$

for $j = 0, \ldots, i$, where $\mathcal{N}_i$ is a subspace of $\ker(\mathbf{A})$ such that $\dim(\mathcal{N}_i) \le 2i$. Note that (65) implies that the algorithm iterates $(\mathbf{z}^j, \overline{\mathbf{z}}^j)$ for $j = 1, \ldots, i$ do not change when $\mathbf{L}_{i-1}$ is replaced by $\mathbf{L}_i$. Hence, this process sequentially adjusts the objective function $\mathbf{L}$ upon observing an iterate $\mathbf{z}^i$ to resist the algorithm from optimizing it efficiently. Indeed, if (66) holds with $i = j = k$, then

$$\begin{aligned} \mathbf{x}^k - \mathbf{x}^0 &= \mathbf{U}_k q_{\mathbf{x}}(\mathbf{A})\mathbf{b} + \mathbf{U}_k \mathbf{v}_{\mathbf{x}}^k \\ \mathbf{y}^k - \mathbf{y}^0 &= \mathbf{U}_k q_{\mathbf{y}}(\mathbf{A})\mathbf{b} + \mathbf{U}_k \mathbf{v}_{\mathbf{y}}^k \end{aligned}$$

for some polynomials $q_{\mathbf{x}}, q_{\mathbf{y}}$ of degree $\le k-1$ and $\mathbf{v}_{\mathbf{x}}^k, \mathbf{v}_{\mathbf{y}}^k \in \mathcal{N}_i \subseteq \ker(\mathbf{A})$. Thus

$$\nabla_{\mathbf{x}} \mathbf{L}_k(\mathbf{x}^k, \mathbf{y}^k) = \mathbf{A}_k(\mathbf{y}^k - \mathbf{y}^0) - \mathbf{b} = \mathbf{U}_k \mathbf{A} \mathbf{U}_k^\intercal \left(\mathbf{U}_k q_y(\mathbf{A})\mathbf{b} + \mathbf{U}_k \mathbf{v}_{\mathbf{y}}^k\right) - \mathbf{b} = \mathbf{U}_k \left(\mathbf{A} q_y(\mathbf{A}) - \mathbf{I}\right) \mathbf{b}$$

and similarly

$$\nabla_{\mathbf{y}} \mathbf{L}_k(\mathbf{x}^k, \mathbf{y}^k) = \mathbf{U}_k \left(\mathbf{A} q_x(\mathbf{A}) - \mathbf{I}\right) \mathbf{b}$$

showing that

$$\|\nabla \mathbf{L}_k(\mathbf{z}^k)\|^2 = \|\mathbf{U}_k \left(\mathbf{A} q_y(\mathbf{A}) - \mathbf{I}\right) \mathbf{b}\|^2 + \|\mathbf{U}_k \left(\mathbf{A} q_x(\mathbf{A}) - \mathbf{I}\right) \mathbf{b}\|^2 \ge \frac{2\|\mathbf{x}^{\min}\|^2}{(2\lfloor k/2 \rfloor + 1)^2}$$

Then the theorem statement follows from the fact that $\mathbf{z}^\star = \mathbf{z}^0 + (\mathbf{U}_k \mathbf{x}^{\min}, \mathbf{U}_k \mathbf{x}^{\min})$ is a saddle point of $\mathbf{L}_k$.

It remains to provide an inductive scheme for choosing $\mathbf{U}_i$. We set $\mathbf{U}_0 = \mathbf{I}$ (so that $\mathbf{A}_0 = \mathbf{A}$), $\mathcal{N}_0 = \{0\}$, and define $\mathcal{K}_{-1}(\mathbf{A}; \mathbf{b}) = \{0\}$ for convenience. Let $1 \le i \le k$, and suppose that we already have an orthogonal matrix $\mathbf{U}_{i-1}$ and $\mathcal{N}_{i-1} \subseteq \ker(\mathbf{A})$ for which $\mathbf{U}_{i-1}\mathbf{b} = \mathbf{b}$, $\dim(\mathcal{N}_{i-1}) \le 2i-2$, and (66) holds with $i-1$ (which is vacuously true when $i = 1$). Let

$$(\mathbf{z}^i, \overline{\mathbf{z}}^i) = \mathcal{A}_i\left(\mathbf{z}^0, \mathcal{O}(\mathbf{z}^0; \mathbf{L}_{i-1}), \ldots, \mathcal{O}(\mathbf{z}^{i-1}; \mathbf{L}_{i-1})\right)$$

We want $\mathbf{U}_i$ (to be defined) to satisfy $\mathbf{s}_\mathbf{x}^i, \mathbf{s}_\mathbf{y}^i \in \mathbf{U}_i \ker(\mathbf{A})$ while $\mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}) = \mathcal{K}_{i-1}(\mathbf{A}_i; \mathbf{b})$. The latter condition is satisfied if $\mathbf{U}_i = \mathbf{Q}_i \mathbf{U}_{i-1}$ for some orthogonal matrix $\mathbf{Q}_i$ which preserves every element within

$$\mathcal{J}_{i-1} = \mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}) \oplus \mathbf{U}_{i-1}\mathcal{N}_{i-1}$$

because then it follows that $\mathbf{U}_i\mathbf{b} = \mathbf{Q}_i\mathbf{U}_{i-1}\mathbf{b} = \mathbf{Q}_i\mathbf{b} = \mathbf{b}$ and

$$\mathcal{K}_{i-1}(\mathbf{A}_i; \mathbf{b}) = \mathbf{U}_i\mathcal{K}_{i-1}(\mathbf{A}; \mathbf{b}) = \mathbf{Q}_i\mathbf{U}_{i-1}\mathcal{K}_{i-1}(\mathbf{A}; \mathbf{b}) = \mathbf{Q}_i\mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}) = \mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b})$$

Consider the decomposition

$$\begin{aligned}\mathbf{x}^i - \mathbf{x}^0 &= \Pi_{\mathcal{K}_{i-1}(\mathbf{A}_{i-1};\mathbf{b})}(\mathbf{x}^i - \mathbf{x}^0) + \mathbf{U}_{i-1}\mathbf{r}_\mathbf{x}^i + \mathbf{s}_\mathbf{x}^i \\ \mathbf{y}^i - \mathbf{y}^0 &= \Pi_{\mathcal{K}_{i-1}(\mathbf{A}_{i-1};\mathbf{b})}(\mathbf{y}^i - \mathbf{y}^0) + \mathbf{U}_{i-1}\mathbf{r}_\mathbf{y}^i + \mathbf{s}_\mathbf{y}^i\end{aligned}$$

where Π denotes the orthogonal projection, $\mathbf{r}_\mathbf{x}^i, \mathbf{r}_\mathbf{y}^i \in \mathcal{N}_{i-1}$ and $\mathbf{s}_\mathbf{x}^i, \mathbf{s}_\mathbf{y}^i \in \mathcal{J}_{i-1}^\perp$. Since $\dim\ker(\mathbf{A}) = n - 2m - 2 \geq n - k - 2$ and $\dim(\mathcal{N}_{i-1})^\perp \geq n - (2i-2) \geq n - 2k + 2$, we have

$$\dim\left(\ker(\mathbf{A}) \cap (\mathcal{N}_{i-1})^\perp\right) \geq n - 3k \geq 2$$

so there exist $\widetilde{\mathbf{s}}_\mathbf{x}^i, \widetilde{\mathbf{s}}_\mathbf{y}^i \in \ker(\mathbf{A}) \cap (\mathcal{N}_{i-1})^\perp$ such that $\|\widetilde{\mathbf{s}}_\mathbf{x}^i\| = \|\mathbf{s}_\mathbf{x}^i\|$, $\|\widetilde{\mathbf{s}}_\mathbf{y}^i\| = \|\mathbf{s}_\mathbf{y}^i\|$, and $\langle \widetilde{\mathbf{s}}_\mathbf{x}^i, \widetilde{\mathbf{s}}_\mathbf{y}^i \rangle = \langle \mathbf{s}_\mathbf{x}^i, \mathbf{s}_\mathbf{y}^i \rangle$. Also, because $\ker(\mathbf{A}) \perp \mathcal{K}_{i-1}(\mathbf{A}; \mathbf{b})$,

$$\mathcal{J}_{i-1} = \mathbf{U}_{i-1}\left(\mathcal{K}_{i-1}(\mathbf{A}; \mathbf{b}) + \mathcal{N}_{i-1}\right) \perp \mathbf{U}_{i-1}\left(\ker(\mathbf{A}) \cap (\mathcal{N}_{i-1})^\perp\right)$$

This implies that there exists an orthogonal $\mathbf{Q}_i \in \mathbb{R}^{n\times n}$ satisfying

$$\begin{aligned}\mathbf{Q}_i\big|_{\mathcal{J}_{i-1}} &= \mathrm{Id}_{\mathcal{J}_{i-1}} \\ \mathbf{Q}_i\left(\mathbf{U}_{i-1}\widetilde{\mathbf{s}}_\mathbf{x}^i\right) &= \mathbf{s}_\mathbf{x}^i \\ \mathbf{Q}_i\left(\mathbf{U}_{i-1}\widetilde{\mathbf{s}}_\mathbf{y}^i\right) &= \mathbf{s}_\mathbf{y}^i\end{aligned}$$

Now let $\mathbf{v}_\mathbf{x}^i = \mathbf{r}_\mathbf{x}^i + \widetilde{\mathbf{s}}_\mathbf{x}^i \in \ker(\mathbf{A})$, $\mathbf{v}_\mathbf{y}^i = \mathbf{r}_\mathbf{y}^i + \widetilde{\mathbf{s}}_\mathbf{y}^i \in \ker(\mathbf{A})$, and

$$\begin{aligned}\mathbf{U}_i &\triangleq \mathbf{Q}_i\mathbf{U}_{i-1} \\ \mathcal{N}_i &\triangleq \mathcal{N}_{i-1} + \mathrm{span}\{\mathbf{v}_\mathbf{x}^i, \mathbf{v}_\mathbf{y}^i\}\end{aligned}$$

Then clearly $\mathbf{U}_i\mathbf{b} = \mathbf{b}$, $\mathcal{N}_i \subseteq \ker(\mathbf{A})$, and $\dim\mathcal{N}_i \leq 2i$. Next, for each $j = 0, \ldots, i-1$, we have

$$\mathbf{x}^j - \mathbf{x}^0, \mathbf{y}^j - \mathbf{y}^0 \in \mathcal{K}_{j-1}(\mathbf{A}_{i-1}; \mathbf{b}) \oplus \mathbf{U}_{i-1}\mathcal{N}_{i-1} \subseteq \mathcal{K}_{j-1}(\mathbf{A}_i; \mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i$$

since $\mathbf{Q}_i$ preserves $\mathcal{J}_{i-1}$ and $\mathcal{N}_{i-1} \subseteq \mathcal{N}_i$. Moreover, because $\mathbf{U}_{i-1}\mathbf{r}_\mathbf{x}^i = \mathbf{Q}_i\mathbf{U}_{i-1}\mathbf{r}_\mathbf{x}^i = \mathbf{U}_i\mathbf{r}_\mathbf{x}^i$ and $\mathbf{s}_\mathbf{x}^i = \mathbf{Q}_i\mathbf{U}_{i-1}\widetilde{\mathbf{s}}_\mathbf{x}^i = \mathbf{U}_i\widetilde{\mathbf{s}}_\mathbf{x}^i$,

$$\mathbf{x}^i - \mathbf{x}^0 = \Pi_{\mathcal{K}_{i-1}(\mathbf{A}_{i-1};\mathbf{b})}(\mathbf{x}^i - \mathbf{x}^0) + \mathbf{U}_i(\mathbf{r}_\mathbf{x}^i + \widetilde{\mathbf{s}}_\mathbf{x}^i) \in \mathcal{K}_{i-1}(\mathbf{A}_{i-1}; \mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i = \mathcal{K}_{i-1}(\mathbf{A}_i; \mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i$$

and similarly $\mathbf{y}^i - \mathbf{y}^0 \in \mathcal{K}_{i-1}(\mathbf{A}_i; \mathbf{b}) \oplus \mathbf{U}_i\mathcal{N}_i$. This proves (66).

Finally, for $j = 0, \ldots, i-1$,

$$\nabla_\mathbf{x}\mathbf{L}_i(\mathbf{x}^j, \mathbf{y}^j) = \mathbf{A}_i(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b} = \mathbf{Q}_i\mathbf{A}_{i-1}\mathbf{Q}_i^\intercal(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b}$$

But $\mathbf{Q}_i^\intercal(\mathbf{y}^j - \mathbf{y}^0) = \mathbf{y}^j - \mathbf{y}^0$ because $\mathbf{y}^j - \mathbf{y}^0 \in \mathcal{K}_{j-1}(\mathbf{A}_{i-1}; \mathbf{b}) \oplus \mathbf{U}_{i-1}\mathcal{N}_{i-1} \subseteq \mathcal{J}_{i-1}$, and

$$\mathbf{A}_{i-1}(\mathbf{y}^j - \mathbf{y}^0) \in \mathbf{A}_{i-1}\mathcal{K}_{j-1}(\mathbf{A}_{i-1}; \mathbf{b}) \oplus \mathbf{A}_{i-1}\mathbf{U}_{i-1}\mathcal{N}_{i-1} = \mathcal{K}_j(\mathbf{A}_{i-1}; \mathbf{b}) \subseteq \mathcal{J}_{i-1}$$

which shows that $\nabla_\mathbf{x}\mathbf{L}_i(\mathbf{x}^j, \mathbf{y}^j) = \mathbf{Q}_i\mathbf{A}_{i-1}\mathbf{Q}_i^\intercal(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b} = \mathbf{A}_{i-1}(\mathbf{y}^j - \mathbf{y}^0) - \mathbf{b} = \nabla_\mathbf{x}\mathbf{L}_{i-1}(\mathbf{x}^j, \mathbf{y}^j)$. Arguing analogously for the $\mathbf{y}$-variable gives $\nabla_\mathbf{y}\mathbf{L}_i(\mathbf{x}^j, \mathbf{y}^j) = \nabla_\mathbf{y}\mathbf{L}_{i-1}(\mathbf{x}^j, \mathbf{y}^j)$, proving (65). This completes the induction step, and hence the proof. □

D Experimental details

D.1 Exact forms of the construction from [Ouyang & Xu(2021)]

Following [Ouyang & Xu(2021)], we use

$$\mathbf{A} = \frac{1}{4}\begin{bmatrix} & & & & -1 & 1 \\ & & & \iddots & \iddots & \\ & & -1 & 1 & & \\ -1 & 1 & & & & \\ 1 & & & & & \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad \mathbf{b} = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \in \mathbb{R}^n, \quad \mathbf{h} = \frac{1}{4}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

and $\mathbf{H} = 2\mathbf{A}^\intercal\mathbf{A}$. [Ouyang & Xu(2021)] shows that $\|\mathbf{A}\| \le \frac{1}{2}$, which implies $\|\mathbf{H}\| \le \frac{1}{2}$. Therefore (25) is a 1-smooth saddle function.

D.2 Best-iterate gradient norm bound for EG

In Figure 1, we indicated theoretical upper bounds for EG. To clarify, there is no known last-iterate convergence result for EG with respect to $\|\mathbf{G}(\cdot)\|^2$. However, it is straightforward to derive $\mathcal{O}(R^2/k)$ *best-iterate* convergence via standard summability arguments in weak convergence proofs for EG. Although there is no theoretical guarantee that $\|\mathbf{G}(\mathbf{z}^k)\|^2$ will monotonically decrease with EG, in our experiments on both examples, they did monotonically decrease (see Figures 1(a), 1(b)). Therefore, we safely used the best-iterate bounds to visualize the upper bound for EG in Figure 1. For the sake of completeness, we derive the best-iterate bound below.

Lemma 6. *Let $\mathbf{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be an R-smooth convex-concave saddle function with a saddle point $\mathbf{z}^\star$. Let $\mathbf{z} \in \mathbb{R}^n \times \mathbb{R}^m$ and $\alpha \in \left(0, \frac{1}{R}\right)$. Then $\mathbf{w} = \mathbf{z} - \alpha\mathbf{G}(\mathbf{z})$ and $\mathbf{z}^+ = \mathbf{z} - \alpha\mathbf{G}(\mathbf{w})$ satisfy*

$$\|\mathbf{z} - \mathbf{z}^\star\|^2 - \|\mathbf{z}^+ - \mathbf{z}^\star\|^2 \ge (1 - \alpha^2R^2)\|\mathbf{z} - \mathbf{w}\|^2$$

Proof.

$$\begin{aligned} \|\mathbf{z} - \mathbf{z}^\star\|^2 - \|\mathbf{z}^+ - \mathbf{z}^\star\|^2 &= \left(\|\mathbf{z} - \mathbf{w}\|^2 + 2\langle \mathbf{z} - \mathbf{w}, \mathbf{w} - \mathbf{z}^\star\rangle + \|\mathbf{w} - \mathbf{z}^\star\|^2\right) \\ &\quad - \left(\|\mathbf{z}^+ - \mathbf{w}\|^2 + 2\langle \mathbf{z}^+ - \mathbf{w}, \mathbf{w} - \mathbf{z}^\star\rangle + \|\mathbf{w} - \mathbf{z}^\star\|^2\right) \\ &= \|\mathbf{z} - \mathbf{w}\|^2 - \|\mathbf{z}^+ - \mathbf{w}\|^2 + 2\langle \mathbf{z} - \mathbf{z}^+, \mathbf{w} - \mathbf{z}^\star\rangle \\ &\ge \|\mathbf{z} - \mathbf{w}\|^2 - \|\mathbf{z}^+ - \mathbf{w}\|^2 \end{aligned}$$

The last inequality is just monotonicity: $\langle \mathbf{z} - \mathbf{z}^+, \mathbf{w} - \mathbf{z}^\star\rangle = \alpha\langle \mathbf{G}(\mathbf{w}), \mathbf{w} - \mathbf{z}^\star\rangle \ge 0$. Now the conclusion follows from

$$\|\mathbf{z}^+ - \mathbf{w}\|^2 = \|(\mathbf{z} - \alpha\mathbf{G}(\mathbf{w})) - (\mathbf{z} - \alpha\mathbf{G}(\mathbf{z}))\|^2 = \alpha^2\|\mathbf{G}(\mathbf{z}) - \mathbf{G}(\mathbf{w})\|^2 \le \alpha^2R^2\|\mathbf{z} - \mathbf{w}\|^2$$

where the last inequality follows from R-Lipschitzness of $\mathbf{G}$. □

Now fix an integer $k \ge 0$, and consider the EG iterations

$$\begin{aligned} \mathbf{z}^{i+1/2} &= \mathbf{z}^i - \alpha\mathbf{G}(\mathbf{z}^i) \\ \mathbf{z}^{i+1} &= \mathbf{z}^i - \alpha\mathbf{G}(\mathbf{z}^{i+1/2}) \end{aligned}$$

for $i = 0, \ldots, k$. Applying Lemma 6 with $\mathbf{z} = \mathbf{z}^i$, $\mathbf{w} = \mathbf{z}^{i+1/2}$ and $\mathbf{z}^+ = \mathbf{z}^{i+1}$, we have

$$\|\mathbf{z}^i - \mathbf{z}^\star\|^2 - \|\mathbf{z}^{i+1} - \mathbf{z}^\star\|^2 \geq (1 - \alpha^2 R^2)\|\mathbf{z}^i - \mathbf{z}^{i+1/2}\|^2 = (1 - \alpha^2 R^2)\alpha^2 \|\mathbf{G}(\mathbf{z}^i)\|^2 \tag{67}$$

for $i = 0, \ldots, k$. Summing up the inequalities (67) for all $i = 0, \ldots, k$, we obtain

$$\|\mathbf{z}^0 - \mathbf{z}^\star\|^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^\star\|^2 \geq (1 - \alpha^2 R^2)\alpha^2 \sum_{i=0}^{k} \|\mathbf{G}(\mathbf{z}^i)\|^2$$

The left hand side is at most $\|\mathbf{z}^0 - \mathbf{z}^\star\|^2$, while the right hand side is lower bounded by

$$(1 - \alpha^2 R^2)\alpha^2 \, (k+1) \min_{i=0,\ldots,k} \|\mathbf{G}(\mathbf{z}^i)\|^2$$

Therefore we conclude that

$$\min_{i=0,\ldots,k} \|\mathbf{G}(\mathbf{z}^i)\|^2 \leq \frac{C\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k+1}$$

where $C = \frac{1}{\alpha^2(1-\alpha^2 R^2)}$.

D.3 ODE flows for $\mathbf{L}(x, y) = xy$

Interestingly, the continuous-time flows with $\mathbf{L}(x, y) = xy$ have exact closed-form solutions.

Note that $\mathbf{G}(x, y) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$. Therefore,

$$\mathbf{G}_\lambda(x, y) = \frac{1}{\lambda}\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & \lambda \\ -\lambda & 1 \end{bmatrix}^{-1} \right) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{\lambda}{1+\lambda^2} & \frac{1}{1+\lambda^2} \\ -\frac{1}{1+\lambda^2} & \frac{\lambda}{1+\lambda^2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

The solution to the Moreau–Yosida regularized flow

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\frac{\lambda}{1+\lambda^2} & -\frac{1}{1+\lambda^2} \\ \frac{1}{1+\lambda^2} & -\frac{\lambda}{1+\lambda^2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

can be obtained with the matrix exponent. The results are

$$x(t) = \exp\left(-\frac{\lambda}{1+\lambda^2}t\right)\left(x^0 \cos\frac{t}{1+\lambda^2} - y^0 \sin\frac{t}{1+\lambda^2}\right)$$
$$y(t) = \exp\left(-\frac{\lambda}{1+\lambda^2}t\right)\left(y^0 \cos\frac{t}{1+\lambda^2} + x^0 \sin\frac{t}{1+\lambda^2}\right)$$

The anchored flow ODE for $\mathbf{L}(x, y) = xy$ is given by

$$\dot{x}(t) = -y(t) + \frac{1}{t}\left(x^0 - x(t)\right)$$
$$\dot{y}(t) = x(t) + \frac{1}{t}\left(y^0 - y(t)\right)$$

From the first equation, we have $\frac{d}{dt}(tx(t)) = t\dot{x}(t) + x(t) = -ty(t) + x^0$, while similar manipulation of the second equation gives $\frac{d}{dt}(ty(t)) = tx(t) + y^0$. Therefore,

$$\frac{d^2}{dt^2}(tx(t)) = -\frac{d}{dt}(ty(t)) = -tx(t) - y^0$$

$$\frac{d^2}{dt^2}(ty(t)) = \frac{d}{dt}(tx(t)) = -ty(t) + x^0$$

which gives

$$tx(t) = c_1 \cos t - c_2 \sin t - y^0$$
$$ty(t) = c_1 \sin t + c_2 \cos t + x^0$$

Using the initial conditions to determine the coefficients c_1, c_2, we obtain

$$x(t) = \frac{y^0 \cos t + x^0 \sin t - y^0}{t}$$
$$y(t) = \frac{y^0 \sin t - x^0 \cos t + x^0}{t}$$

E Connection to CLI lower bounds

In this section, we discuss how EAG relates to the prior work on complexity lower bounds on the class of CLI and SCLI algorithms, introduced and studied in [Arjevani et al.(2016), Arjevani & Shamir(2016), Azizian et al.(2020), Golowich et al.(2020)]. Specifically, we show that EAG is not SCLI, so it can break the $\Omega(R^2/k)$ lower bound on squared gradient norm for the 1-SCLI class derived by [Golowich et al.(2020)]. On the other hand, we show that EAG is 2-CLI in the sense of [Golowich et al.(2020)], and that EAG belongs to an extended class of 1-CLI algorithms.

E.1 Lower bounds for 1-SCLI and non-stationarity of EAG

We start with the notion of 1-SCLI algorithms by [Golowich et al.(2020)]. Consider an algorithm $\mathcal{A}$ for finding saddle points of biaffine functions of the form

$$\mathbf{L}(\mathbf{x}, \mathbf{y}) = \mathbf{b}^\intercal \mathbf{x} + \mathbf{x}^\intercal \mathbf{A} \mathbf{y} - \mathbf{c}^\intercal \mathbf{y}$$

where $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$. We say $\mathcal{A}$ is *1-stationary canonical linear iterative (1-SCLI)* if there exist some fixed matrix mappings $\mathbf{C}, \mathbf{N} : \mathbb{R}^{2n \times 2n} \to \mathbb{R}^{2n \times 2n}$ such that

$$\mathbf{z}^{k+1} = \mathbf{C}\left(\begin{bmatrix} \mathbf{O} & \mathbf{A} \\ -\mathbf{A}^\intercal & \mathbf{O} \end{bmatrix}\right) \mathbf{z}^k + \mathbf{N}\left(\begin{bmatrix} \mathbf{O} & \mathbf{A} \\ -\mathbf{A}^\intercal & \mathbf{O} \end{bmatrix}\right) \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} = \mathbf{C}(\mathbf{B})\mathbf{z}^k + \mathbf{N}(\mathbf{B})\mathbf{v} \tag{68}$$

for $k \geq 0$, where

$$\mathbf{B} = \begin{bmatrix} \mathbf{O} & \mathbf{A} \\ -\mathbf{A}^\intercal & \mathbf{O} \end{bmatrix} \in \mathbb{R}^{2n \times 2n}, \quad \mathbf{v} = \begin{bmatrix} \mathbf{b} \\ \mathbf{c} \end{bmatrix} \in \mathbb{R}^{2n}$$

Following the convention of [Azizian et al.(2020)] and [Golowich et al.(2020)], we also require that $\mathbf{C}, \mathbf{N}$ are matrix polynomials. The classical extragradient method (EG) is an 1-SCLI algorithm: with $\mathbf{G}(\mathbf{z}) = \mathbf{B}\mathbf{z} + \mathbf{v}$, we can express EG as

$$\begin{aligned} \mathbf{z}^{k+1} &= \mathbf{z}^k - \alpha \mathbf{G}\left(\mathbf{z}^k - \alpha \mathbf{G}(\mathbf{z}^k)\right) \\ &= \mathbf{z}^k - \alpha\left(\mathbf{B}\left(\mathbf{z}^k - \alpha \mathbf{B}\mathbf{z}^k - \alpha \mathbf{v}\right) + \mathbf{v}\right) \\ &= \left(\mathbf{I} - \alpha \mathbf{B} + \alpha^2 \mathbf{B}^2\right) \mathbf{z}^k - \alpha(\mathbf{I} - \alpha \mathbf{B})\mathbf{v} \end{aligned}$$

which is of the 1-SCLI form.

A 1-SCLI algorithm $\mathcal{A}$ is *consistent* with respect to an invertible matrix $\mathbf{B}$ if for any $\mathbf{v} \in \mathbb{R}^{2n}$, iterates $\{\mathbf{z}^k\}_{k\geq 0}$ produced by $\mathcal{A}$ satisfy

$$\mathbf{z}^k \to \mathbf{z}^\star = -\mathbf{B}^{-1}\mathbf{v}$$

If $\mathcal{A}$ is consistent with respect to $\mathbf{B}$, then for any $\mathbf{w} = \mathbf{B}^{-1}\mathbf{v} \in \mathbb{R}^{2n}$, we have

$$\begin{aligned} -\mathbf{w} = -\mathbf{B}^{-1}\mathbf{v} &= \lim_{k\to\infty} \mathbf{z}^{k+1} \\ &= \lim_{k\to\infty} \mathbf{C}(\mathbf{B})\mathbf{z}^k + \mathbf{N}(\mathbf{B})\mathbf{v} \\ &= \mathbf{C}(\mathbf{B})(-\mathbf{B}^{-1}\mathbf{v}) + \mathbf{N}(\mathbf{B})\mathbf{v} \\ &= (-\mathbf{C}(\mathbf{B}) + \mathbf{N}(\mathbf{B})\mathbf{B})\,\mathbf{w} \end{aligned}$$

As this holds for all $\mathbf{w} \in \mathbb{R}^{2n}$, we have the following result.

Lemma 7 ([Arjevani et al.(2016)])**.** *If a 1-SCLI algorithm $\mathcal{A}$ described by* (68) *is consistent with respect to $\mathbf{B}$, then*

$$\mathbf{I} + \mathbf{N}(\mathbf{B})\mathbf{B} = \mathbf{C}(\mathbf{B}) \tag{69}$$

Indeed, the 1-SCLI formulation of EG satisfies (69).

For the class of consistent 1-SCLI algorithms, [Golowich et al.(2020)] established $\Omega(1/k)$ a complexity lower bound on squared gradient norm.

Theorem 5 ([Golowich et al.(2020)])**.** *Let $k \geq 0$ and $n \geq 1$. Then for any consistent 1-SCLI algorithm of the form* (68) *with* $\deg \mathbf{N} = d_{\mathbf{N}}$*, there exist a biaffine function $\mathbf{L}(\mathbf{x}, \mathbf{y}) = \mathbf{b}^\intercal\mathbf{x} + \mathbf{x}^\intercal\mathbf{A}\mathbf{y} - \mathbf{c}^\intercal\mathbf{y}$ on $\mathbb{R}^n \times \mathbb{R}^n$ with invertible $\mathbf{A}$, for which*

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \geq \frac{R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{20(d_{\mathbf{N}}+1)^2 k} = \Omega\left(\frac{R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k}\right)$$

where $\mathbf{z}^\star$ is the unique saddle point of $\mathbf{L}$.

To clarify, $\deg \mathbf{N}$ refers to the degree of the matrix polynomial defining $\mathbf{N}$. 1-SCLI algorithms with $d_{\mathbf{C}} = \deg \mathbf{C} = 1$ forms a subclass of $\mathfrak{A}_{\text{sim}}$ and $\mathfrak{A}_{\text{sep}}$. (Even if $d_{\mathbf{C}} > 1$, one can still view 1-SCLI algorithms as instances of $\mathfrak{A}_{\text{sim}}$ or $\mathfrak{A}_{\text{sep}}$ by introducing $d_{\mathbf{C}} - 1$ dummy iterates for each 1-SCLI iteration.) However, EAG is an algorithm that belongs to $\mathfrak{A}_{\text{sim}}$ but is not 1-SCLI; if it was, a contradiction would occur, as $\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \mathcal{O}(1/k^2)$ for EAG. In fact, it is intuitively clear that EAG is not 1-SCLI; the S in 1-**S**CLI stands for stationary, but EAG has anchoring coefficients $\frac{1}{k+2}$ that vary over iterations.

E.2 Understanding EAG as a CLI algorithm

In this section, we show that EAG algorithms are (non-stationary) 2-CLI, and that we can expand the definition of 1-CLI algorithms to accommodate EAG.

First, we state the definition of m-CLI algorithms introduced by [Arjevani & Shamir(2016)] adapted to the case of biaffine saddle functions. For $m \geq 1$, an m-CLI algorithm $\mathcal{A}$ takes m initial points $\mathbf{z}_1^0, \ldots, \mathbf{z}_m^0$ and at each iteration $k \geq 0$, outputs

$$\mathbf{z}_i^{k+1} = \sum_{j=1}^{m} \mathbf{C}_{ij}^{(k)}(\mathbf{B})\,\mathbf{z}_j^k + \mathbf{N}_i^{(k)}(\mathbf{B})\,\mathbf{v} \tag{70}$$

for $i = 1, \dots, m$, where $\mathbf{C}_{ij}^{(k)}, \mathbf{N}_i^{(k)} : \mathbb{R}^{2n \times 2n} \to \mathbb{R}^{2n \times 2n}$ for $i, j = 1, \dots, m$ are matrix polynomials that depend on k but not on $\{\mathbf{z}_1^k, \dots, \mathbf{z}_m^k\}_{k \geq 0}$. In the case where $\mathbf{C}_{ij}^{(k)} \equiv \mathbf{C}_{ij}$ and $\mathbf{N}_i^{(k)} \equiv \mathbf{N}_i$ for all $i, j = 1, \dots, m$ and $k \geq 0$, we say $\mathcal{A}$ is stationary. Indeed, when $m = 1$, this definition of stationary 1-CLI coincides with that of 1-SCLI given in Section E.1. Also note that the definition (70) includes algorithms that obtain $\mathbf{z}^{k+1}$ with m previous iterates $\mathbf{z}^k, \mathbf{z}^{k-1}, \dots, \mathbf{z}^{k-m+1}$, by letting $\mathbf{z}_i^k = \mathbf{z}^{k+1-i}$ for $i = 1, \dots, m$.

[Golowich et al.(2020)] showed that the averaged EG iterates, which have rate $\mathcal{O}(1/k)$ on duality gap, can be written in 2-CLI form; hence, the $\Omega(1/\sqrt{k})$ 1-SCLI lower bound on duality gap therein cannot be generalized to m-CLI algorithms for $m \geq 2$. They then posed the open problem of whether the $\Omega(1/\sqrt{k})$ 1-SCLI lower bound on duality gap can be generalized to 1-CLI algorithms. Below, we provide a similar discussion on rates on squared gradient norm.

It is straightforward to see that EAG is 2-CLI; define $\mathbf{z}_2^{k+1} = \mathbf{z}_2^k = \cdots = \mathbf{z}_2^0 = \mathbf{z}^0 = \mathbf{z}_1^0$ for all $k \geq 0$, and

$$\begin{aligned}\mathbf{z}_1^{k+1} &= \mathbf{z}_1^k - \alpha_k \mathbf{G}\left(\mathbf{z}_1^k - \alpha_k \mathbf{G}(\mathbf{z}_1^k) + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}_1^k)\right) + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}_1^k) \\ &= \left(\frac{k+1}{k+2}\mathbf{I} - \frac{k+1}{k+2}\alpha_k \mathbf{B} + \alpha_k^2 \mathbf{B}^2\right)\mathbf{z}_1^k + \frac{1}{k+2}(\mathbf{I} - \alpha_k \mathbf{B})\mathbf{z}_2^k - \alpha_k(\mathbf{I} - \alpha_k \mathbf{B})\mathbf{v} \qquad (71)\end{aligned}$$

For EAG-C, one can alternatively eliminate the dependency on $\mathbf{z}^0$ to define $\mathbf{z}^{k+1}$ in terms of $\mathbf{z}^k$, $\mathbf{z}^{k-1}$, and $\mathbf{v}$; respectively multiply $(k+2)$ and $(k+1)$ to the following identities

$$\begin{aligned}\mathbf{z}^{k+1} &= \left(\frac{k+1}{k+2}\mathbf{I} - \frac{k+1}{k+2}\alpha \mathbf{B} + \alpha^2 \mathbf{B}^2\right)\mathbf{z}^k + \frac{1}{k+2}(\mathbf{I} - \alpha \mathbf{B})\mathbf{z}^0 - \alpha(\mathbf{I} - \alpha \mathbf{B})\mathbf{v} \\ \mathbf{z}^k &= \left(\frac{k}{k+1}\mathbf{I} - \frac{k}{k+1}\alpha \mathbf{B} + \alpha^2 \mathbf{B}^2\right)\mathbf{z}^{k-1} + \frac{1}{k+1}(\mathbf{I} - \alpha \mathbf{B})\mathbf{z}^0 - \alpha(\mathbf{I} - \alpha \mathbf{B})\mathbf{v}\end{aligned}$$

and subtract to eliminate $\mathbf{z}^0$. Since EAG has $\mathcal{O}(1/k^2)$ rate, this reformulation shows that the $\Theta(1/k)$ 1-SCLI lower bound on the squared gradient norm cannot be generalized to 2-CLI algorithms.

Furthermore, EAG also provides a partial resolution, in the negative, of the open problem of whether the $\Theta(1/k)$ 1-SCLI lower bound on the squared gradient norm can be generalized to 1-CLI algorithms. Observe that if we translate the given problem to set $\mathbf{z}^0 = 0$, keeping the sequence $\mathbf{z}_2^k$ is no longer necessary, and (71) reduces to 1-CLI form. Such translation is not allowed in the definition (70), but it is reasonable to consider an expanded class of algorithms that are 1-CLI up to translation. Precisely, define an algorithm $\mathcal{A}$ to be *translated 1-CLI* if it takes the form

$$\mathbf{z}^{k+1} = \mathbf{C}^{(k)}(\mathbf{B})(\mathbf{z}^k) + \mathbf{N}^{(k)}(\mathbf{B})(\mathbf{v})$$

when $\mathbf{z}^0 = 0$, and is *translation invariant* in the sense that

$$\mathbf{z}^k = \mathcal{A}(\mathbf{z}^0, \mathbf{z}^1, \dots, \mathbf{z}^{k-1}; \mathbf{L}) = \mathbf{z}^0 + \mathcal{A}(0, \mathbf{z}^1 - \mathbf{z}^0, \dots, \mathbf{z}^{k-1} - \mathbf{z}^0; \mathbf{L}_{\mathbf{z}^0})$$

when $\mathbf{z}^0 \neq 0$, where $\mathbf{L}_{\mathbf{z}^0}(\mathbf{x}, \mathbf{y}) = \mathbf{L}(\mathbf{x} + \mathbf{x}^0, \mathbf{y} + \mathbf{y}^0)$. That is, the iterates of $\mathcal{A}$ are generated equivalently by starting with $\mathbf{z}^0 = 0$ and applying $\mathcal{A}$ to the translated objective $\mathbf{L}_{\mathbf{z}^0}$. The concept of translated 1-CLI can be viewed as a generalization of consistent 1-SCLI algorithms; observe that we can rewrite (68) as

$$\mathbf{z}^{k+1} - \mathbf{z}^0 = \mathbf{C}(\mathbf{B})(\mathbf{z}^k - \mathbf{z}^0) + \mathbf{N}(\mathbf{B})(\mathbf{B}\mathbf{z}^0 + \mathbf{v}) - (\mathbf{I} + \mathbf{N}(\mathbf{B})\mathbf{B} - \mathbf{C}(\mathbf{B}))\mathbf{z}^0$$

which shows that a 1-SCLI algorithm is translation invariant if and only if it satisfies the consistency formula (69). Since EAG has $\mathcal{O}(1/k^2)$ rate and is a translated 1-CLI algorithm, our results prove that the $\Theta(1/k)$ 1-SCLI lower bound on the squared gradient norm can be generalized to translated 1-CLI algorithms.

Performance measure	Algorithm class	Lower bound	Best known rate	Order-optimality
Duality gap (Last iterate)	1-SCLI	$\Omega\left(\frac{R}{\sqrt{k}}\right)$ [Golowich et al.(2020)]	$\mathcal{O}\left(\frac{R}{\sqrt{k}}\right)$ [Golowich et al.(2020)]*	Established*
	1-CLI	$\Omega\left(\frac{R}{k}\right)$ ([Nemirovsky(1992)], [Nemirovski(2004)])	$\mathcal{O}\left(\frac{R}{\sqrt{k}}\right)$ [Golowich et al.(2020)]*	Unknown
	m-CLI ($m \geq 2$)	$\Omega\left(\frac{R}{k}\right)$ ([Nemirovsky(1992)], [Nemirovski(2004)])	$\mathcal{O}\left(\frac{R}{k}\right)$ ([Nemirovski(2004)], [Golowich et al.(2020)])	Established
Squared gradient norm (Last iterate)	1-SCLI	$\Omega\left(\frac{R^2}{k}\right)$ [Golowich et al.(2020)]	$\mathcal{O}\left(\frac{R^2}{k}\right)$ [Golowich et al.(2020)]*	Established*
	1-CLI	$\Omega\left(\frac{R^2}{k^2}\right)$ [Nemirovsky(1992)]	$\mathcal{O}\left(\frac{R^2}{k}\right)$ [Golowich et al.(2020)]*	Unknown
	Translated 1-CLI	$\Omega\left(\frac{R^2}{k^2}\right)$ [Nemirovsky(1992)]	$\mathcal{O}\left(\frac{R^2}{k^2}\right)$ (This paper)	Established
	m-CLI ($m \geq 2$)	$\Omega\left(\frac{R^2}{k^2}\right)$ [Nemirovsky(1992)]	$\mathcal{O}\left(\frac{R^2}{k^2}\right)$ (This paper)	Established

Table 1. Lower bounds and best known rates for CLI algorithm classes (* means that the result holds with the additional assumption that the derivative of $\mathbf{G}$ is Lipschitz continuous).