Accelerated Algorithms for Smooth Convex-Concave Minimax Problems with $\mathcal{O}(1/k^2)$ Rate on Squared Gradient Norm

TaeHo Yoon [1] **Ernest K. Ryu** [1]

Abstract

In this work, we study the computational complexity of reducing the squared gradient magnitude for smooth minimax optimization problems. First, we present algorithms with accelerated $\mathcal{O}(1/k^2)$ last-iterate rates, faster than the existing $\mathcal{O}(1/k)$ or slower rates for extragradient, Popov, and gradient descent with anchoring. The acceleration mechanism combines extragradient steps with anchoring and is distinct from Nesterov's acceleration. We then establish optimality of the $\mathcal{O}(1/k^2)$ rate through a matching lower bound.

1. Introduction

Minimax optimization problems, or minimax games, of the form

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

have recently gained significant interest in the optimization and machine learning communities due to their application in adversarial training (Goodfellow et al., 2015; Madry et al., 2018) and generative adversarial networks (GANs) (Goodfellow et al., 2014).

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

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

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.

Contributions. In this work, we introduce the *extra anchored gradient (EAG)* algorithms for smooth convex-concave minimax problems and establish an accelerated $\|\nabla\mathbf{L}(\mathbf{z}^k)\|^2 \le \mathcal{O}(R^2/k^2)$ 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.

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 $\mathbf{L}(\mathbf{x}^\star,\mathbf{y}) \le \mathbf{L}(\mathbf{x}^\star,\mathbf{y}^\star) \le \mathbf{L}(\mathbf{x},\mathbf{y}^\star)$ 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.,

[1]Department of Mathematical Sciences, Seoul National University, Seoul, Korea. Correspondence to: Ernest K. Ryu <ernestryu@snu.ac.kr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

$\langle \mathbf{G}(\mathbf{z}_1) - \mathbf{G}(\mathbf{z}_2), \mathbf{z}_1 - \mathbf{z}_2 \rangle \geq 0$ 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; 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; 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 $\min_{i=0,\dots,k} \|\mathbf{G}(\mathbf{z}^i)\|^2 \leq \mathcal{O}(R^2/k)$ 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;b) 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 $\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle$ 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; 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; 2013; Chen et al., 2014; 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;b; 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, 2021).

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)

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

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

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

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 $\mathbf{z}^k = \mathcal{A}(\mathbf{z}^0, \dots, \mathbf{z}^{k-1}; \mathbf{L})$ for $k \geq 1$. Define $\mathfrak{A}_{\text{sim}}$ as the class of algorithms satisfying

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

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

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

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

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

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

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

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; 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; 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 $\|\mathbf{B}\| \leq R$ 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 $k \geq 2$,

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

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*

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

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 $n \ge 3k+2$. 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 $\|\nabla\mathbf{L}(\mathbf{z}^k)\|^2 \le \mathcal{O}(1/k)$. 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.

4. Experiments

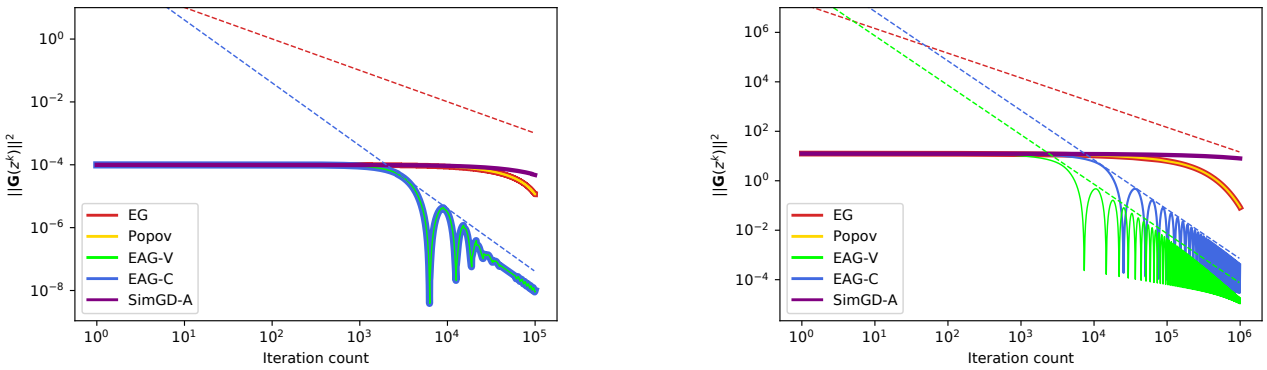
We now present experiments illustrating the accelerated rate of EAG. We compare EAG-C and EAG-V against the prior algorithms with convergence guarantees: EG, Popov's algorithm (or optimistic descent) and simultaneous gradient descent with anchoring (SimGD-A). The precise forms of the algorithms are restated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

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

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

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

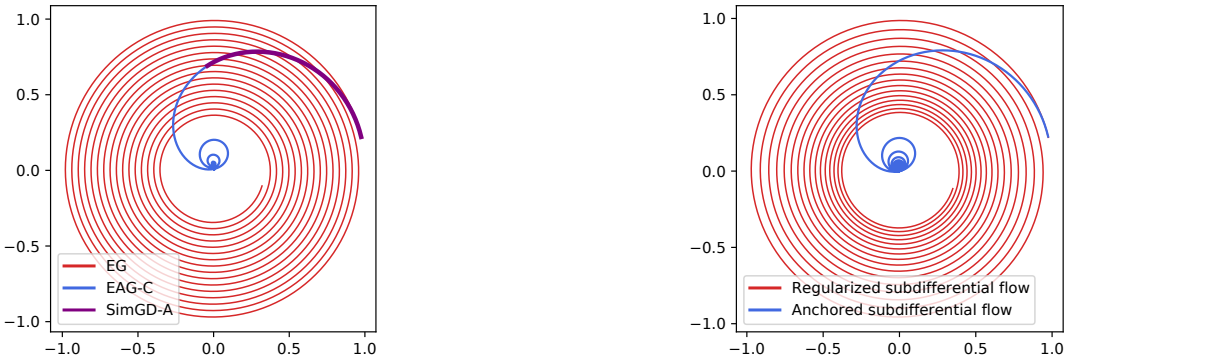
$$\mathbf{L}_{\delta,\epsilon}(x,y) = (1-\delta)f_\epsilon(x) + \delta xy - (1-\delta)f_\epsilon(y), \tag{24}$$



(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.



(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.

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 $|x|, |y| \le \epsilon$. 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

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

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

This work presents the extra anchored gradient (EAG) algorithms, which exhibit accelerated $\mathcal{O}(1/k^2)$ rates on the squared gradient magnitude for smooth convex-concave minimax problems. The acceleration combines the extragradient and anchoring mechanisms, which separately achieve $\mathcal{O}(1/k)$ or slower rates. We complement the $\mathcal{O}(1/k^2)$ rate with a matching $\Omega(1/k^2)$ complexity lower bound, thereby establishing optimality of EAG.

At a superficial level, the acceleration mechanism of EAG seems to be distinct from that of Nesterov; anchoring dampens oscillations, but momentum provides the opposite effect of dampening. However, are the two accelerations phenomena entirely unrelated? Finding a common structure, a connection, between the two acceleration phenomena would be an interesting direction of future work.

Acknowledgements

TY and EKR were supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2020R1F1A1A01072877], the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2017R1A5A1015626], by the New Faculty Startup Fund from Seoul National University, and by the AI Institute of Seoul National University (AIIS) through its AI Frontier Research Grant (No. 0670-20200015) in 2020. We thank Jaewook Suh and Jongmin Lee for reviewing the manuscript and providing valuable feedback. We thank Jelena Diakonikolas for the discussion on the prior work on parameter-free near-optimal methods for the smooth minimax setup. Finally, we thank the anonymous referees for bringing to our attention the recent complexity lower bound on the class of 1-SCLI algorithms by Golowich et al. (2020).

References

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, K., Belmega, E. V., and Mertikopoulos, P. Adaptive extra-gradient methods for min-max optimization and games. *ICLR*, 2021.

Arjevani, Y. and Shamir, O. On the iteration complexity of oblivious first-order optimization algorithms. *ICML*, 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, 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, M. N. and Combettes, P. L. A warped resolvent algorithm to construct nash equilibria. *arXiv preprint arXiv:2101.00532*, 2021.

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

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

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

Daskalakis, C., Deckelbaum, A., and Kim, A. Near-optimal no-regret algorithms for zero-sum games. *SODA*, 2011.

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

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

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, 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, 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, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *ICLR*, 2018.

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, 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, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NeurIPS*, 2014.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *ICLR*, 2015.

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

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, 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, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. *NeurIPS*, 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, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

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

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

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

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

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. *ICML*, 2020a.

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

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, 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, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

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

Malitsky, Y. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184:383–410, 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, Y. V. and Semenov, V. An extragradient algorithm for monotone variational inequalities. *Cybernetics and Systems Analysis*, 50(2):271–277, 2014.

Mason, J. C. and Handscomb, D. C. *Chebyshev Polynomials*. 2002.

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

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

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

Nemirovsky, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. 1983.

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

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

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

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

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. 2013.

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

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

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

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, 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, A. and Sridharan, K. Online learning with predictable sequences. *COLT*, 2013a.

Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. *NeurIPS*, 2013b.

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

Ryu, E. K. and Yin, W. *Large-Scale Convex Optimization via Monotone Operators*. Draft, 2021.

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, 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, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast convergence of regularized learning in games. *NeurIPS*, 2015.

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

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, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. *NeurIPS*, 2019.

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

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

Wang, Y. and Li, J. Improved algorithms for convex-concave minimax optimization. *NeurIPS*, 2020.

Woodworth, B. and Srebro, N. Tight complexity bounds for optimizing composite objectives. *NeurIPS*, 2016.

Yadav, A., Shah, S., Xu, Z., Jacobs, D., and Goldstein, T. Stabilizing adversarial nets with prediction methods. *ICLR*, 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, J., Zhang, S., Kiyavash, N., and He, N. A catalyst framework for minimax optimization. *NeurIPS*, 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, J., Hong, M., and Zhang, S. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.

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