On the Suboptimality of Negative Momentum for Minimax Optimization

Guodong Zhang
University of Toronto, Vector Institute

Yuanhao Wang
Princeton University

Abstract

Smooth game optimization has recently attracted great interest in machine learning as it generalizes the single-objective optimization paradigm. However, game dynamics is more complex due to the interaction between different players and is therefore fundamentally different from minimization, posing new challenges for algorithm design. Notably, it has been shown that negative momentum is preferred due to its ability to reduce oscillation in game dynamics. Nevertheless, the convergence rate of negative momentum was only established in simple bilinear games. In this paper, we extend the analysis to smooth and strongly-convex strongly-concave minimax games by taking the variational inequality formulation. By connecting Polyak's momentum with Chebyshev polynomials, we show that negative momentum accelerates convergence of game dynamics locally, though with a suboptimal rate. To the best of our knowledge, this is the *first work* that provides an explicit convergence rate for negative momentum in this setting.

1 Introduction

Due to the increasing popularity of generative adversarial networks (Goodfellow et al., 2014; Radford et al., 2015; Arjovsky et al., 2017), adversarial training (Madry et al., 2018) and primal-dual reinforcement learning (Du et al., 2017; Dai et al., 2018), minimax optimization (or generally game optimization) has gained significant attention as it offers a flexible paradigm that goes beyond ordinary loss function minimization. In particular, our problem of interest is the following minimax optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}). \tag{1}$$

We are usually interested in finding a *Nash equilibrium* (Von Neumann and Morgenstern, 1944): a set of parameters from which no player can (locally and unilaterally) improve its objective function. Though the dynamics of gradient based methods are well understood for minimization problems, new issues and challenges appear in minimax games. For example, the naïve extension of gradient descent can fail to converge (Letcher et al., 2019; Mescheder et al., 2017) or converge to undesirable stationary points (Mazumdar et al., 2019; Adolphs et al., 2019; Wang et al., 2019).

Another important difference between minimax games and minimization problems is that negative momentum value is preferred for improving convergence (Gidel et al., 2019b). To be specific, for the blinear case $f(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\mathbf{A}\mathbf{y}$, negative momentum with alternating updates converges to ϵ-optimal solution with an iteration complexity of $\mathcal{O}(\kappa)$ where the condition number κ is defined as $\kappa = \frac{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})}{\lambda_{\min}(\mathbf{A}^\top\mathbf{A})}$, whereas Gradient Descent Ascent (GDA) fails to converge. Moreover, the rate of negative momentum matches the optimal rate of Extra-gradient (EG) (Korpelevich, 1976) and Optimistic Gradient Descent Ascent (OGDA) (Daskalakis et al., 2018; Mertikopoulos et al., 2019). A natural question to ask then is:

Does negative momentum improve on GDA for other settings?

In this paper, we extend the analysis of negative momentum[1] to the strongly-convex strongly-concave setting and answer the above question in the affirmative. In particular, we observe that momentum methods (Polyak, 1964), either positive or negative, can be connected to Chebyshev iteration (Manteuffel, 1977) in solving linear systems, which enables us to derive the optimal momentum parameter and asymptotic

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] Throughout the paper, negative momentum represents gradient descent-ascent with negative momentum.

convergence rate. With optimally tuned parameters, negative momentum achieves an acceleration *locally* with an improved iteration complexity $\mathcal{O}(\kappa^{1.5})$ as opposed to the $\mathcal{O}(\kappa^2)$ complexity of Gradient Descent Ascent (GDA). Following on that, we further ask:

Is negative momentum optimal in the same setting? Does it match the iteration complexity of EG and OGDA again?

We answer these questions in the negative. Particularly, our analysis implies that the iteration complexity lower bound for negative momentum is $\Omega(\kappa^{1.5})$. Nevertheless, the optimal iteration complexity for this family of problems under first-order oracle is $\Omega(\kappa)$ (Ibrahim et al., 2019; Zhang et al., 2019), which can be achieved by EG and OGDA. Therefore, we *for the first time* show that negative momentum alone is suboptimal for strongly-convex strongly-concave minimax games. To the best of our knowledge, this is the first work that provides an explicit convergence rate for negative momentum in this setting.

Organization. In Section 2, we define our notation and formulate minimax optimization as a variational inequality problem. Under the variational inequality framework, we further write first-order methods as discrete dynamical systems and show that we can safely linearize the dynamics for proving local convergence rates (thus simplifying the problem to that of solving linear systems). In Section 3, we discuss the connection between first-order methods and polynomial approximation and show that we can analyze the convergence of a first-order method through the sequence of polynomials it defines. In Section 4, we prove the local convergence rate of negative momentum for minimax games by connecting it with Chebyshev polynomials, showing that it has a suboptimal rate locally. Finally, in Section 6, we validate our claims in simulation.

2 Preliminaries

Notation. In this paper, scalars are denoted by lower-case letters (e.g., λ), vectors by lower-case bold letters (e.g., $\mathbf{z}$), matrices by upper-case bold letters (e.g., $\mathbf{J}$) and operators by upper-case letters (e.g., F). The superscript $^\top$ represents the transpose of a vector or a matrix. The spectrum of a square matrix $\mathbf{A}$ is denoted by $\mathrm{Sp}(\mathbf{A})$, and its eigenvalue by λ. We use $\Re$ and $\Im$ to denote the real part and imaginary part of a complex scalar respectively. We use $\mathbb{R}$ and $\mathbb{C}$ to denote the set of real numbers and complex numbers, respectively. We use $\rho(\mathbf{A}) = \lim_{n\to\infty} \|\mathbf{A}^n\|^{1/n}$ to denote the spectral radius of matrix $\mathbf{A}$. $\mathcal{O}$, Ω and Θ are standard asymptotic notations. We use Π_t to denote the set of real polynomials with degree no more than t.

2.1 Variational Inequality Formulation of Minimax Optimization

We begin by presenting the basic variational inequality framework that we will consider throughout the paper. To that end, let $\mathcal{Z}$ be a nonempty convex subset of $\mathbb{R}^d$, and let $F: \mathbb{R}^d \to \mathbb{R}^d$ be a continuous mapping on $\mathbb{R}^d$. In its most general form, the variational inequality (VI) problem (Harker and Pang, 1990) associated to F and $\mathcal{Z}$ can be stated as:

$$\text{find } \mathbf{z}^* \in \mathcal{Z} \text{ s.t. } F(\mathbf{z}^*)^\top(\mathbf{z} - \mathbf{z}^*) \geq 0 \ \ \forall \mathbf{z} \in \mathcal{Z}. \quad (2)$$

In the case of $\mathcal{Z} = \mathbb{R}^d$, the problem is reduced to finding $\mathbf{z}^*$ such that $F(\mathbf{z}^*) = 0$. To provide some intuition about the variational inequality problem, we discuss two important examples below:

Example 1 (Minimization). *Suppose that $F = \nabla_{\mathbf{z}} f$ for a smooth function f on $\mathbb{R}^d$, then the varitional inequality problem is essentially finding the critical points of f. In the case where f is convex, any solution of (2) would be a global minimum.*

Example 2 (Minimax Optimization). *Consider the convex-concave minimax optimization (or saddle-point optimization) problem, where the objective is to solve the following problem*

$$\min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}), \text{ where } f \text{ is a smooth function.} \quad (3)$$

One can show that it is a special case of (2) with $F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})^\top, -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})^\top]^\top$.

Notably, the vector field F in Example 2 is not necessarily conservative, i.e., it might not be the gradient of any function. In addition, since f in minimax problem happens to be convex-concave, any solution $\mathbf{z}^* = [\mathbf{x^*}^\top, \mathbf{y^*}^\top]^\top$ of (2) is a global *Nash Equilibrium* (Von Neumann and Morgenstern, 1944), i.e.,

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq f(\mathbf{x}, \mathbf{y}^*) \quad \forall \mathbf{x} \text{ and } \mathbf{y} \in \mathbb{R}^d.$$

In this work, we are particularly interested in the case of f being a strongly-convex-strongly-concave and smooth function, which essentially assumes that the vector field F is strongly-monotone and Lipschitz (see Fallah et al. (2020, Lemma 2.6)). Here we state our assumptions formally.

Assumption 1 (Strongly Monotone). *The vector field F is μ-strongly-monotone:*

$$(F(\mathbf{z}_1) - F(\mathbf{z}_2))^\top(\mathbf{z}_1 - \mathbf{z}_2) \geq \mu\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2, \ \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d. \quad (4)$$

Assumption 2 (Lipschitz). *The vector field F is L-Lipschitz if the following holds:*

$$\|F(\mathbf{z}_1) - F(\mathbf{z}_2)\|_2 \leq L\|\mathbf{z}_1 - \mathbf{z}_2\|_2, \ \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d. \quad (5)$$

Table 1: First-order algorithms for smooth and strongly-monotone games.

Method	Parameter Choice	Complexity	Reference
GDA	$\alpha = 0,\ \beta = 0$	$\mathcal{O}(\kappa^2)$	Ryu and Boyd (2016); Azizian et al. (2020a)
OGDA	$\alpha = 1,\ \beta = 0$	$\mathcal{O}(\kappa)$	Gidel et al. (2019a); Mokhtari et al. (2020)
NM	$\alpha = 0,\ \beta < 0$	$\Theta(\kappa^{1.5})$	**This paper** (Theorem 2)

In the context of variational inequalites, Lipschitzness and (strong) monotonicity are fairly standard and have been used in many classical works (Tseng, 1995; Chen and Rockafellar, 1997; Nesterov, 2007; Nemirovski, 2004). With these two assumptions in hand, we define the condition number $\kappa \triangleq L/\mu$, which measures the hardness of the problem. In the following, we turn to suitable optimization techniques for the variational inequality problem.

2.2 First-order methods for Minimax Optimization

The dynamics of optimization algorithms are often described by a vector field, F, and local convergence behavior can be understood in terms of the spectrum of its Jacobian. In minimization, the Jacobian coincides with the Hessian of the loss with all eigenvalues real. In minimax optimization, the Jacobian is nonsymmetric and can have complex eigenvalues, making it harder to analyze.

In the case of strongly-convex strongly-concave minimax games, finding the Nash equilibrium is equivalent to solving the fixed point equation $F(\mathbf{z}^*) = \mathbf{0}$. Here, we mainly focus on first-order methods (Nesterov, 1983) to find the stationary point $\mathbf{z}^*$:

Definition 1 (First-order methods)**.** *A first-order method generates*

$$\mathbf{z}_t \in \mathbf{z}_0 + \boldsymbol{Span}\{F(\mathbf{z}_0), ..., F(\mathbf{z}_{t-1})\}. \tag{6}$$

This wide class includes most gradient-based optimization methods we are interested in, such as GDA, OGDA and momentum method. All three methods are special case of the following update:

$$\mathbf{z}_{t+1} = (1+\beta)\mathbf{z}_t - \beta\mathbf{z}_{t-1} - \eta F((1+\alpha)\mathbf{z}_t - \alpha\mathbf{z}_{t-1}), \tag{7}$$

where β is the momentum parameter, α the extrapolation parameter and η the step size. With proper choices of parameters, we can recover GDA, OGDA and negative momentum (see Table 1). For instance, the update rule of negative momentum is given by

$$\mathbf{z}_{t+1} = (1+\beta)\mathbf{z}_t - \beta\mathbf{z}_{t-1} - \eta F(\mathbf{z}_t). \tag{8}$$

For smooth and strongly-monotone games, Azizian et al. (2020b, Corollary 1) showed a lower bound on convergence rate for *any* algorithm of the form (6):

$$\begin{aligned} &\|\mathbf{z}_k - \mathbf{z}^*\|_2 \geq \rho_{\text{opt}}^k \|\mathbf{z}_0 - \mathbf{z}^*\|_2 \\ &\text{with} \quad \rho_{\text{opt}} = 1 - \frac{2\mu}{\mu + L}. \end{aligned} \tag{9}$$

2.3 Dynamical System Viewpoint and Local Convergence

With a first-order algorithm defined, we study local convergence rates from the viewpoint of dynamical system. It is well-known that gradient-based methods can reliably find local stable fixed points (i.e., local minima) in single-objective optimization. Here, we generalize the concept of stability to games by taking game dynamics as a discrete dynamical system. An iteration of the form $\mathbf{z}_{t+1} = G(\mathbf{z}_t)$ can be viewed as a discrete dynamical system. If $G(\mathbf{z}^*) = \mathbf{z}^*$, then $\mathbf{z}^*$ is called a fixed point. We study the stability of fixed points as a proxy to local convergence of game dynamics.

Definition 2. *Let $\mathbf{J}_G$ denote the Jacobian of G at a fixed point $\mathbf{z}^*$. If it has spectral radius $\rho(\mathbf{J}_G) \leq 1$, then we call $\mathbf{z}^*$ a stable fixed point. If $\rho(\mathbf{J}_G) < 1$, then we call $\mathbf{z}^*$ a strictly stable fixed point.*

It has been shown that strict stability implies local convergence (see Galor (2007)). In other words, if $\mathbf{z}^*$ is a strictly stable fixed point, there exists a neighborhood U of $\mathbf{z}^*$ such that when initialized in U, the iteration steps always converge to $\mathbf{z}^*$.

Remark 1. *Because we focus on local convergence rates, we can safely take the Jacobian $\mathbf{J}$ as constant locally, which essentially linearizes the vector field $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}, \mathbf{A} = \mathbf{J}_F(\mathbf{z}^*)$. Therefore, locally solving the minimax game becomes arguably as easy as solving the linear system $\mathbf{A}\mathbf{z} + \mathbf{b} = \mathbf{0}$.*

2.4 Chebyshev Polynomials

The Chebyshev polynomials were discovered a century ago by the mathematician Chebyshev. Since then, they have found many uses in numerical analysis (Fox and Parker, 1968; Mason and Handscomb, 2002). The Chebyshev polynomials can be defined recursively as

$$\begin{aligned} T_0(z) &= 1,\ T_1(z) = z, \\ T_{n+1}(z) &= 2zT_n(z) - T_{n-1}(z). \end{aligned} \tag{10}$$

They may also be written as

$$T_n(z) = \begin{cases} \cos(n \arccos(z)) & \text{if } -1 \le z \le 1 \\ \cosh(n \cosh^{-1}(z)) & \text{otherwise} \end{cases}. \tag{11}$$

For the case of $z \notin [-1, 1]$, we have the Chebyshev polynomials $T_n(z) = \cosh(n \cosh^{-1}(z))$. Consider the map $\eta = \cosh(\sigma)$, let $\sigma = x + yi$ and $\eta = u + vi$. Then $\cosh(\sigma) = \cosh(x + yi) = u + vi = \eta$. By the property of cosh, we have

$$\cosh(x + yi) = \cosh(x)\cos(y) + \sinh(x)\sin(y)i. \tag{12}$$

If we fix $x = \text{const}$ (with varying y), then we have $\frac{u^2}{\cosh(x)^2} + \frac{v^2}{\sinh(x)^2} = 1$. That is, cosh maps the vertical line $x = \text{const}$ to an ellipse with semi-major axis $|\cosh(x)|$, semi-minor axis $|\sinh(x)|$ and foci at $+1$ and -1. This map has the period $2\pi i$.

3 Polynomial-based Iterative Methods

In the background section, we showed that solving minimax games locally boils down to solving a linear system. Here, we leverage the well-established theory of polynomial approximation for efficiently solving linear systems. The next lemma shows that when the vector field F is linear, first-order algorithms defined in (6) can be written as polynomials.

Lemma 1 (Fischer (2011)). *If the vector field is of the form of $F(\mathbf{z}) = \mathbf{A}\mathbf{z}+\mathbf{b}$, then $\mathbf{z}_t$ generated by first-order methods can be written as*

$$\mathbf{z}_t - \mathbf{z}^* = p_t(\mathbf{A})(\mathbf{z}_0 - \mathbf{z}^*), \tag{13}$$

where $\mathbf{z}^*$ satisfies $\mathbf{A}\mathbf{z}^* + \mathbf{b} = \mathbf{0}$ and $p_t \in \Pi_t$ is a polynomial with degree at most t that satisfies $p_t(0) = 1$.

To gain better intuition about the above Lemma, we provide an example of gradient descent below.

Example 3 (Gradient Descent). *For gradient descent with constant learning rate η, the corresponding polynomials are given by*

$$p_t(\mathbf{A}) = (\mathbf{I} - \eta\mathbf{A})^t. \tag{14}$$

Hence, the convergence of a first-order method can be analyzed through the sequence of polynomials p_t it generates. Specifically, we can bound the error of $\|\mathbf{z}_t - \mathbf{z}^*\|_2$ in a as follows:

$$\|\mathbf{z}_t - \mathbf{z}^*\|_2 \le \|p_t(\mathbf{A})\|_2 \|\mathbf{z}_0 - \mathbf{z}^*\|_2, \tag{15}$$

where $\|p_t(\mathbf{A})\|_2^{1/t} = (\max_{\lambda \in \mathrm{Sp}(\mathbf{A})} |p_t(\lambda)|)^{1/t}$ as $t \to \infty$. Importantly, the error depends on two factors: the polynomial (algorithm) p_t and the matrix (problem) $\mathbf{A}$. In practice, we are interested in the performance of a algorithm on a broad class of problem, therefore we instead consider a set $\mathcal{S}_K$ of matrices $\mathbf{A}$:

$$\mathcal{S}_K := \{\mathbf{A} \in \mathbb{R}^{d\times d} : \mathrm{Sp}(\mathbf{A}) \in K \in \mathbb{C}\} \tag{16}$$

Clearly, an obvious choice for the residual polynomial p_t is the one which minimizes the upper bound in (15). This optimal polynomial $\mathcal{P}_t(\lambda, K)$ is the solution of the following Chebyshev approximation problem

$$\max_{\lambda\in K} |\mathcal{P}_t(\lambda; K)| = \min\left\{ \max_{\lambda \in K} |p(\lambda)| : \ p \in \Pi_t, p(0) = 1 \right\}.$$

To measure the performance, we define the *asymptotic convergence factor* (Eiermann and Niethammer, 1983) with the following form:

$$\rho(K) = \lim_{t\to\infty} \left(\max_{\lambda \in K} |\mathcal{P}_t(\lambda; K)| \right)^{1/t}. \tag{17}$$

It was shown that the asymptotic convergence factor also serves as a **lower bound** in the worse-case (Nevanlinna, 1993) depending on the set K[2].

Proposition 1 (Nevanlinna (1993)). *Let K be a subset of $\mathbb{C}$ symmetric w.r.t the real axis, that does not contain the origin. Then, any oblivious first-order method (whose coefficients only depend on K, see Arjevani and Shamir (2016)) satisfies the following*

$$\forall t > 0, \exists \mathbf{z}_0, \exists \mathbf{A} : \|\mathbf{z}_t - \mathbf{z}^*\|_2 \ge \rho(K) \|\mathbf{z}_0 - \mathbf{z}^*\|_2.$$

Interestingly, if the set K is simple enough, we can compute the asymptotic convergence factor and the optimal polynomial. In particular, when K is a complex ellipse in the complex plane which does not contain the origin in its interior, the following result is known in the literature (Clayton, 1963; Wrigley, 1963; Manteuffel, 1977; Azizian et al., 2020b).

Theorem 1. *If the set of K is a complex ellipse with the following form:*

$$K = \left\{ \lambda \in \mathbb{C} : E_{a,b,d}(\lambda) \triangleq \frac{(\Re\lambda - d)^2}{a^2} + \frac{(\Im\lambda)^2}{b^2} \le 1 \right\}, \quad \text{with } d > a > 0, b > 0. \tag{18}$$

then one can show its asymptotically optimal polynomial is a rescaled and translated Chebyshev polynomial:

$$\mathcal{P}_t(\lambda; K) = \frac{T_t\left(\frac{d-\lambda}{c}\right)}{T_t\left(\frac{d}{c}\right)}, \ c^2 = a^2 - b^2. \tag{19}$$

Moreover, the asymptotic convergence factor (i.e., the maximum modulus) is achieved on the boundary:

$$\rho(K) = \begin{cases} a/d & \text{if } a = b \\ \frac{d - \sqrt{d^2 + b^2 - a^2}}{a - b} & \text{otherwise} \end{cases}. \tag{20}$$

[2]It is the fastest possible asymptotic convergence rate a first-order method can achieve for all linear systems with spectrum in the set.

Remark 2. *We note that c can be either pure real or pure imaginary. In either case, c^2 is real and throughout the paper c only appears as c^2.*

Notably, the first-order method corresponding to the optimal polynomial (19) is Polyak momentum.

Corollary 1. *The optimal first-order methods for K in the form of (18) iterates as follows:*

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t F(\mathbf{z}_t) + \beta_t(\mathbf{z}_t - \mathbf{z}_{t-1}), \tag{21}$$

where η_t and β_t are not constant over time. However, by choosing constants $\eta = 2\frac{d-\sqrt{d^2-c^2}}{c^2}$ and $\beta = d\eta - 1$, we can obtain the same asymptotic rate.

Remark 3. *The optimal momentum parameter β is*

$$\beta = 2\frac{d^2 - d\sqrt{d^2-c^2}}{c^2} - 1 = 2\frac{d^2 - \frac{1}{2}c^2 - \sqrt{d^2(d^2-c^2)}}{c^2}. \tag{22}$$

In particular, we note that the numerator of (22) is always non-negative. Therefore, we conclude that the optimal momentum parameter is negative when $c^2 = a^2 - b^2 < 0$ and positive when $c^2 > 0$.

We note that the eigenvalues of the Jacobian for the minimization problem lie on the real axis, which is a special case of a complex ellipse with $b = 0$. In that case, it is known that Polyak momentum has an optimal worst-case convergence rate over the class of first order methods (Polyak, 1987). In the special case of K being a disc, we have the optimal algorithm being gradient descent (Eiermann et al., 1985).

Corollary 2. *For the case of $a = b$, i.e., K is a disc in the complex plane, the optimal polynomial is*

$$\mathcal{P}_t(\lambda; K) = (1 - \lambda/d)^t, \tag{23}$$

and the optimal algorithm is gradient descent.

Corollary 3. *The modulus of $\mathcal{P}_t(\lambda; K)$ is constant on the boundary of (18) for large t.*

4 Suboptimality of Negative Momentum

In the previous section, we have shown that Polyak momentum with properly chosen parameters is asymptotically optimal for linear systems with the spectrum enclosed in the region of complex ellipse. In this section, we shift our attention back to minimax games. In particular, we analyze minimax optimization with the framework of variational inequality.

Obviously, under Assumptions 1 and 2, the eigenvalue of $\mathbf{J}_F(\mathbf{z}^*)$ will not tightly fall within a complex ellipse. It can be shown that it instead lies within the following set (Azizian et al., 2020b):

$$\hat{K} = \{\lambda \in \mathbb{C} : |\lambda| < L, \Re\lambda > \mu > 0\}. \tag{24}$$

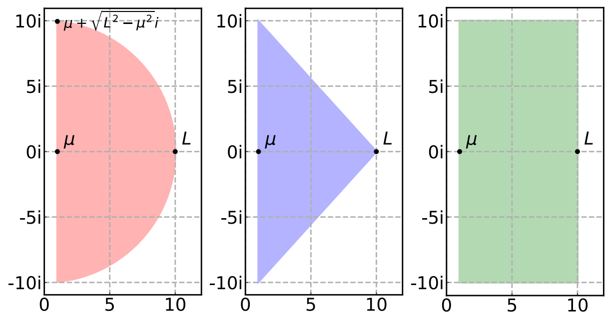

Figure 1: Left: The red region corresponds to $\hat{K}$, the set of strongly monotone problems. **Middle:** The blue triangle corresponds to $\hat{K}_1$, which is enclosed within $\hat{K}$. **Right:** The green rectangle is $\hat{K}_2$, which includes $\hat{K}$ in its interior. For visualization, we set $\mu = 1$ and $L = 10$.

This set is the intersection between a circle and a half-plane (see Figure 1).

Recall that our goal is to search for the best achievable convergence rate of negative momentum (or generally Polyak momentum) for linear systems with spectrum enclosed within $\hat{K}$. By linearizing the vector field locally $F(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b} = \mathbf{A}(\mathbf{z} - \mathbf{z}^*)$ and expanding the state space to $[\mathbf{z}_{t+1}^\top, \mathbf{z}_t^\top]^\top$, we can write (8) in matrix form

$$\begin{bmatrix} \mathbf{z}_{t+1} - \mathbf{z}^* \\ \mathbf{z}_t - \mathbf{z}^* \end{bmatrix} = \bar{\mathbf{J}} \begin{bmatrix} \mathbf{z}_t - \mathbf{z}^* \\ \mathbf{z}_{t-1} - \mathbf{z}^* \end{bmatrix}, \tag{25}$$

where the matrix $\bar{\mathbf{J}}$ has the following form:

$$\bar{\mathbf{J}} = \begin{bmatrix} (1+\beta)\mathbf{I} - \eta\mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}. \tag{26}$$

Thus, finding the asymptotic convergence rate boils down to the following min-max problem

$$\hat{\rho}(\hat{K}) \triangleq \min_{\eta,\beta} \max_{\lambda \in \hat{K}} \rho\left(\begin{bmatrix} 1+\beta-\eta\lambda & -\beta \\ 1 & 0 \end{bmatrix}\right). \tag{27}$$

Essentially, we would like to find the optimal step size η and momentum parameter β that minimize the spectral radius which determines the asymptotic convergence rate. However, due to the fact that the spectrum is in the complex plane and involves complex eigenvalues, bounding the spectral radius directly becomes challenging. Nevertheless, by Theorem 1 and Corollary 1, we have the following equivalence:

Lemma 2 (Asymptotic Equivalence between Polyak momentum and Chebyshev Iteration)**.** *For any $K \in \mathbb{C}$ that is symmetric w.r.t the real axis and does not contain the origin, if Polyak momentum with parameters η, β converges with rate $\rho < 1$, then there exists a rescaled and translated Chebyshev polynomial parameterized by $d, c^2 \in \mathbb{R}$ converging with the same asymptotic rate, and vice versa.*

Hence the min-max problem (27) is equivalent to:

$$\hat{\rho}(\hat{K}) = \min_{d,c^2\in\mathbb{R}} \max_{\lambda\in\hat{K}} r(\lambda; d, c^2),$$
$$\text{where } r(\lambda; d, c^2) \triangleq \lim_{t\to\infty} \left| \frac{T_t(\frac{d-\lambda}{c})}{T_t(\frac{d}{c})} \right|^{1/t}. \tag{28}$$

We term $r(\lambda; d, c^2)$ the **convergence factor** of Chebyshev polynomial $T_t(\frac{d-\lambda}{c})/T_t(\frac{d}{c})$ at the point λ. The reason why we can do such a reduction is that momentum method is equivalent to the rescaled and translated Chebyshev polynomial (19) asymptotically, and different parameters η, β exactly corresponds to different choices of d, c^2 in (19).

However, the equivalent min-max problem (28) is not easy to solve directly and some reductions have to be done. Our very first step is to use the the sandwich technique, which is inspired by Azizian et al. (2020b). Let $\hat{K}_1$ and $\hat{K}_2$ be the two regions tightly lower bounding and upper bounding $\hat{K}$ (see Figure 1).

$$\begin{aligned} \hat{K}_1 = \{&\lambda \in \mathbb{C} : \Re\lambda \geq \mu, \tfrac{1}{L}\Re\lambda + \tfrac{L-\mu}{L\sqrt{L^2-\mu}}\Im\lambda \leq 1, \\ &\tfrac{1}{L}\Re\lambda - \tfrac{L-\mu}{L\sqrt{L^2-\mu}}\Im\lambda \leq 1\}; \\ \hat{K}_2 = \{&\lambda \in \mathbb{C} : \mu \leq \Re\lambda \leq L, \\ &-\sqrt{L^2-\mu^2} \leq \Im\lambda \leq \sqrt{L^2-\mu^2}\}. \end{aligned} \tag{29}$$

One can see that both $\hat{K}_1$ and $\hat{K}_2$ are convex polygons and particularly $\hat{K}_1 \subset \hat{K} \subset \hat{K}_2$. Therefore, we have

$$\hat{\rho}(\hat{K}_1) \leq \hat{\rho}(\hat{K}) \leq \hat{\rho}(\hat{K}_2). \tag{30}$$

Now, the main challenge is to compute $\hat{\rho}(\hat{K}_1)$ and $\hat{\rho}(\hat{K}_2)$. Ideally, we would hope that they are close to each other and thus we can bound $\hat{\rho}(\hat{K})$ tightly. Given that $\hat{K}_1$ and $\hat{K}_2$ are convex polygons, we have the following results:

Lemma 3 (Manteuffel (1977, Lemma 3.2))**.** *Defining H_1 and H_2 to be the sets of vertices of $\hat{K}_1$ and $\hat{K}_2$, we have*

$$\begin{aligned} \hat{\rho}(\hat{K}_1) &= \min_{d,c^2} \max_{\lambda\in H_1} r(\lambda; d, c^2), \\ \hat{\rho}(\hat{K}_2) &= \min_{d,c^2} \max_{\lambda\in H_2} r(\lambda; d, c^2). \end{aligned} \tag{31}$$

Both H_1 and H_2 are symmetric w.r.t the real axis, we can therefore reduce them to $H_1 = \{L, \mu + \sqrt{L^2-\mu^2}i\}$ and $H_2 = \{L + \sqrt{L^2-\mu^2}i, \mu + \sqrt{L^2-\mu^2}i\}$.

Essentially, Lemma 3 says that, for optimal d and c^2, the largest convergence factor occurs on the *hull* of $\hat{K}_1$ and $\hat{K}_2$, i.e., the set including all the vertices. Therefore, we do not need to maximize over all elements of $\hat{K}_1$ and $\hat{K}_2$, which makes the problem much simpler. Next, we apply the powerful Alternative theorem in functional analysis (Bartle, 1964) to further simplify the min-max problem.

Lemma 4. *For optimal parameters d_i^*, c_i^{2*} in min-max problem (31), all points in H_i has the same rate*

$$r(L; d_1^*, c_1^{2*}) = r(\mu + \sqrt{L^2-\mu^2}i; d_1^*, c_1^{2*});$$
$$r(L + \sqrt{L^2-\mu^2}i; d_2^*, c_2^{2*}) = r(\mu + \sqrt{L^2-\mu^2}i; d_2^*, c_2^{2*}).$$

Intuitively, Lemma 4 suggests that vertices of $\hat{K}_i$ have the same convergence factor. As a consequence, one can show that all vertices of $\hat{K}_i$ are on the boundary of the same complex ellipse.

Lemma 5 (Manteuffel (1977))**.** *Let $\mathcal{E}(d, c^2)$ be the family of complex ellipse in the complex plane centered at d with foci at $d-c$ and $d+c$. Further let $E(d, c^2) \in \mathcal{E}(d, c^2)$ be a member of this family that not include the origin in its interior. Then for any two points $\lambda_i \in E_i(d, c^2)$ and $\lambda_j \in E_j(d, c^2)$, we have*

$$\begin{aligned} r(\lambda_i; d, c^2) = r(\lambda_j; d, c^2) &\iff E_i(d, c^2) = E_j(d, c^2) \\ r(\lambda_i; d, c^2) < r(\lambda_j; d, c^2) &\iff E_i(d, c^2) \subset E_j(d, c^2) \end{aligned}$$

To understand this Lemma, one shall realize that the convergence factor can be further written as

$$r(\lambda; d, c^2) = e^{\cosh^{-1}(\frac{d-\lambda}{c}) - \cosh^{-1}(\frac{d}{c})} \propto e^{\cosh^{-1}(\frac{d-\lambda}{c})}$$

In particular, the transformation $\lambda \mapsto \frac{d-\lambda}{c}$ maps the points in $E(d, c^2)$ to $E(0, 1)$. By the property of $\cosh^{-1}$ (see Section 2.4), $\cosh^{-1}(\frac{d-\lambda}{c})$ maps $E(d, c^2)$ to a vertical line $x = \cosh^{-1}(a)$ where a is the semi-major axis of the specific $E(0, 1)$. So to compare the convergence factors of two points λ_i, λ_j, we only need to compare the semi-major axis of $E_i(d, c^2)$ and $E_j(d, c^2)$.

Finally, we are ready to present our main result.

Theorem 2 (Suboptimality of Negative Momentum)**.** *Under Assumptions 1 and 2, we have the optimal momentum parameter β to be negative and*

$$\hat{\rho}(\hat{K}_1) = 1 - \Theta(\kappa^{-1.5}), \quad \hat{\rho}(\hat{K}_2) = 1 - \Theta(\kappa^{-1.5}).$$

By the sandwich argument, we therefore get $\hat{\rho}(\hat{K}) = 1 - \Theta(\kappa^{-1.5})$. Assuming the vector field F is continuously differentiable, for $\mathbf{z}_0$ close to $\mathbf{z}^*$, negative momentum can converge to $\mathbf{z}^*$ asymptotically with the rate $1 - \Theta(\kappa^{-1.5})$.

Proof sketch. Here we give a short proof sketch with detailed proof deferred to the supplement. Let's first prove the result for $\hat{\rho}(\hat{K}_1)$. By Lemma 4, we have

$$r(L; d_1^*, c_1^{2*}) = r(\mu + \sqrt{L^2-\mu^2}i; d_1^*, c_1^{2*}),$$

which implies that both L and $\mu + \sqrt{L^2-\mu^2}i$ are on the boundary of the same complex ellipse with the center d_1^* and foci at $d_1^* - c_1^*$ and $d_1^* + c_1^*$ according to

Lemma 5. Then by Theorem 1 and Corollary 3, we can reduce the computation of $\hat{\rho}(\hat{K}_1)$ to the following constrained problem:

$$\begin{aligned}&\hat{\rho}(\hat{K}_1) := \min_{a,b,d} \tfrac{d-\sqrt{d^2+b^2-a^2}}{a-b},\\ &\text{s.t. } E_{a,b,d}(L) = E_{a,b,d}(\mu + \sqrt{L^2-\mu^2}i) = 1\end{aligned} \tag{32}$$

which involves three free variables and two constraints. The two constraint equations imply

$$b^2 = \frac{(L+\mu)(L-d)^2}{L+\mu-2d} > (L-d)^2 = a^2.$$

Therefore, the optimal momentum β for $\hat{K}_1$ is negative. For $\hat{K}_2$, we follow the same procedure and have

$$b^2 = \frac{(L^2-\mu^2)a^2}{a^2-(\frac{L-\mu}{2})^2},\ a \in \left[\frac{L-\mu}{2}, \frac{L+\mu}{2}\right].$$

In the case of $L^2 > \mu^2 + \mu L$, we have $b^2 > a^2$ and therefore the optimal momentum is also negative since $c^2 = a^2 - b^2 < 0$ (see Remark 3). Hence, we conclude that the optimal momentum for $\hat{K}$ is negative.

Next, we bound $\hat{\rho}(\hat{K}_1)$ and $\hat{\rho}(\hat{K}_2)$ so as to estimate $\hat{\rho}(\hat{K})$. Towards this end, one can further simplify the problem (32) to a single variable minimization task:

$$\min_{d\in[\frac{L}{2},\frac{\mu+L}{2}]} \frac{d-\sqrt{\frac{2d(L-d)^2}{L+\mu-2d}+d^2}}{(L-d)(1-\sqrt{\frac{L+\mu}{L+\mu-2d}})}. \tag{33}$$

We can repeat the same process for $\hat{\rho}(\hat{K}_2)$, getting the following problem:

$$\min_{a\in[\frac{L-\mu}{2},\frac{L+\mu}{2}]} \frac{\frac{L+\mu}{2} - \sqrt{(\frac{L+\mu}{2})^2 + \frac{a^2L^2}{a^2-(\frac{L-\mu}{2})^2} - a^2}}{a - \frac{aL}{\sqrt{a^2-(\frac{L-\mu}{2})^2}}}. \tag{34}$$

Particularly, one can show that both (33) and (34) are approximately $1-\Theta(\kappa^{-1.5})$ (see the supplement for details). Hence, we have $\hat{\rho}(\hat{K}) = 1-\Theta(\kappa^{-1.5})$ by the sandwich argument. Together with the assumption that the vector field F is continuously differentiable, we proved that negative momentum converges locally with this rate. This completes the proof. □

This shows that the optimal momentum parameter for minimax games is indeed negative and negative momentum with optimally tuned parameter does speed up the convergence of GDA locally, whose iteration complexity is $\mathcal{O}(\kappa^2)$ (Azizian et al., 2020a). However, the best existing lower bound on $\hat{K}$ is $\Omega(\kappa)$ iteration complexity (Azizian et al., 2020b; Zhang et al., 2019). Furthermore, the lower bound is tight as it is already achieved by EG and OGDA (Mokhtari et al., 2020). Thus we conclude that negative momentum is indeed a suboptimal algorithm.

5 Related Works

Polynomial-based iterative methods have long been used in solving linear systems. Two classical examples are the conjugate gradient method (Hestenes et al., 1952) and the Chebyshev iteration (Lanczos, 1952; Golub and Varga, 1961), which forms the basis of some of the most used optimization methods such as Polyak momentum. For symmetric linear systems, Fischer (2011) provides a comprehensive study over the state of art on polynomial-based iterative methods. For non-symmetric linear systems, Manteuffel (1977) discussed Chebyshev polynomial and showed that the iteration converges whenever the eigenvalues of the linear system lie in the open right half complex plane. Particularly, it was shown by (Manteuffel, 1977) that Chebyshev polynomial is optimal when the eigenvalues of the linear system lie within a complex ellipse, which inspires our work. For general non-symmetric linear systems, Eiermann and Niethammer (1983) used complex analysis tools to define, for a given compact set, its asymptotic convergence factor: it is the optimal asymptotic convergence rate a first-order method can achieve for all linear systems with spectrum in the set. Recently, Azizian et al. (2020b) used the tool of polynomial approximation to characterize acceleration in smooth games. Pedregosa and Scieur (2020) and Scieur and Pedregosa (2020) used these ideas to develop methods that are optimal for the average-case.

In the context of minimax optimization, a line of recent work has studied various algorithms under different assumptions. For the strongly-convex strongly-concave case, Tseng (1995) and Nesterov and Scrimali (2006) proved that their algorithms find an ϵ-saddle point with a gradient complexity of $\mathcal{O}(\kappa \ln(1/\epsilon))$ using a variational inequality approach. Using a different approach, Gidel et al. (2019a) and Mokhtari et al. (2020) derived the same convergence results for OGDA. Particularly, Mokhtari et al. (2020) unified the algorithm of OGDA and EG from the perspective of proximal point method, which gives sharp analysis. Notably, this convergence rate is known to be optimal to some extent (Azizian et al., 2020b). Very recently, Ibrahim et al. (2019); Zhang et al. (2019) established fine-grained lower complexity bound among all the first-order algorithms in this setting, which was later achieved by the algorithms in Lin et al. (2020); Wang and Li (2020). To our knowledge, the convergence rate of negative momentum has not been established in this setting before. The only known rate of negative momentum was for simple bilinear games (Gidel et al., 2019b). Particularly, they showed that negative momentum with alternating updates achieves linear convergence, matching the rate of EG and OGDA. In this sense, we are the first to give an explicit rate of neg-

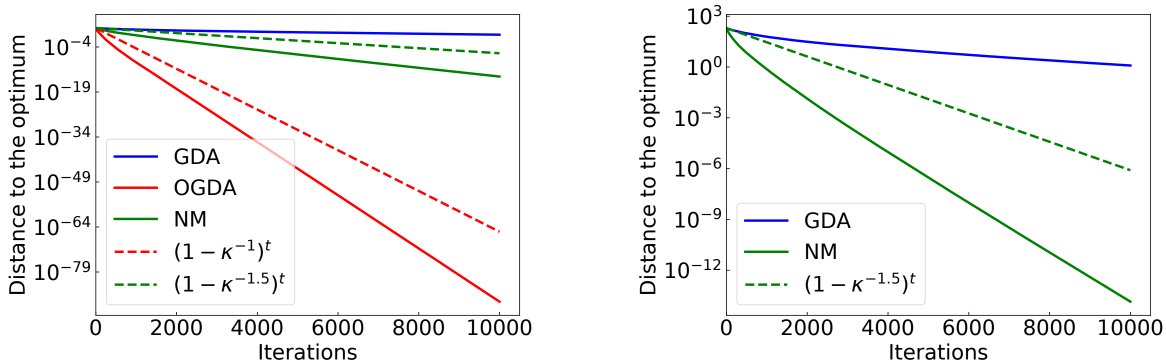

Figure 2: Distance to the optimum as a function of training iterations. Negative momentum accelerates GDA significantly on this quadratic minimax game. In particular, its convergence rate is slightly better than the worst-case rate of $1-\kappa^{-1.5}$. However, negative momentum is outperformed by OGDA, whose convergence rate is approximately $1-\kappa^{-1}$.

ative momentum for strongly-convex strongly-concave setting, though the rate is just local convergence rate.

More broadly, nonconvex-nonconcave problem has gained more attention due to its generality. However, there might be no *Nash* (or even local Nash) equilibrium in that setting due to the loss of strong duality. To overcome that, different notations of equilibrium were introduced by taking into account the sequential structure of games (Jin et al., 2019; Fiez et al., 2019; Zhang et al., 2020b; Farnia and Ozdaglar, 2020). In that setting, the main challenge is to find the right equilibrium and some algorithms (Wang et al., 2019; Adolphs et al., 2019; Mazumdar et al., 2019) have been proposed to achieve that.

6 Numerical Simulations

In this section, we compare the performance of negative momentum with Gradient-Descent-Ascent (GDA) and Optimistic Gradient-Descent-Ascent (OGDA) so as to verify our theoretical result on the convergence rate of negative momentum. In particular, we focus on the following quadratic minimax problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\max_{\mathbf{y}\in\mathbb{R}^d} f(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{B}\mathbf{y} - \frac{1}{2}\mathbf{y}^\top\mathbf{C}\mathbf{y} \quad (35)$$

where we set the dimension $d = 100$. The matrix $\mathbf{A}$ and $\mathbf{C}$ have eigenvalues $\left\{\frac{1}{i}\right\}_{i=1}^{d}$, giving a condition number of 100. For matrix $\mathbf{B}$, we set it to be a random diagonal matrix with entries sampling from $[0,1]$. For all algorithms, the iterates start with $\mathbf{x}_0 = \mathbf{1}$ and $\mathbf{y}_0 = \mathbf{1}$. Figure 2 shows that the distance to the optimum of negative momentum, GDA and OGDA versus the number of iterations for this quadratic minimax problem. For all methods, we tune their hyperparameters by grid-search. We can observe that all three methods converge linearly to the optimum. As expected, negative momentum performs better than GDA, but worse than OGDA. Moreover, both negative momentum and OGDA yield convergences rates that are slightly better than their worst-case rates.

7 Discussion

Although it may seem tempting to directly apply algorithmic techniques for minimization to minimax optimization, they can be provably suboptimal, as shown in this paper. The reason is that the dynamics of minimax optimization is different and considerably more complex. Thus we believe it is important to delve deeper and understand such game dynamics with multiple interacting objectives better. Despite an existing line of work on accelerating GDA in smooth games, previously negative momentum was only analyzed for bilinear games. Due to the fact that negative momentum enjoys the same convergence rate as OGDA does in bilinear games, researchers are often confused with the difference between them and even call OGDA as "negative momentum" (see Mokhtari et al. (2020) for example). Therefore, we believe our analysis of negative momentum is crucial as it highlights that negative momentum is fundamentally different from OGDA.

It is important to emphasize that we only provide local convergence rate of negative momentum in the paper. It is currently unknown whether negative momentum can attain the same geometric rate globally[3]. We left this analysis for future work. In addition, it would be interesting to derive the optimal polynomial (hence optimal first-order algorithm) for smooth and strongly-monotone games. One promising way to achieve that is to finding the conformal mapping between the complement of $\hat{K}$ and the complement of unit disk, then Fabor polynomial (Curtiss, 1971) can be adopted to derive the optimal polynomial.

[3] It is now proved by Zhang et al. (2020a) that negative momentum attains the same rate globally.

Acknowledgements

We thank Shengyang Sun, Jenny Bao, Ricky Chen and Roger Grosse for detailed comments on early drafts. We thank all the anonymous reviewers (especially reviewer 2) for their careful reading of our manuscript and their many insightful comments and suggestions. GZ would also like to thank for the support from Borealis AI fellowship.

References

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 486–495, 2019.

Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *International Conference on Machine Learning*, pages 908–916. PMLR, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873, 2020a.

Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1705–1715, 2020b.

Robert Gardner Bartle. *The elements of real analysis*, volume 2. Wiley New York, 1964.

Alan F Beardon. *Complex analysis: The argument principle in analysis and topology*. Courier Dover Publications, 2019.

George HG Chen and R Tyrrell Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Ao J Clayton. *Further results on polynomials having least maximum modulus over an ellipse in the complex plane*. UKAEA, 1963.

JH Curtiss. Faber polynomials and the faber series. *American Mathematical Monthly*, pages 577–596, 1971.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134, 2018.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations*, 2018.

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

Michael Eiermann and Wilhelm Niethammer. On the construction of semi-iterative methods. *SIAM journal on numerical analysis*, 20(6):1153–1160, 1983.

Michael Eiermann, Wilhelm Niethammer, and Richard S Varga. A study of semiiterative methods for nonsymmetric systems of linear equations. *Numerische Mathematik*, 47(4):505–533, 1985.

Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An optimal multistage stochastic gradient method for minimax problems. *arXiv preprint arXiv:2002.05683*, 2020.

Farzan Farnia and Asuman Ozdaglar. Gans may have no nash equilibria. *arXiv preprint arXiv:2002.09124*, 2020.

Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Bernd Fischer. *Polynomial based iteration methods for symmetric linear systems*. SIAM, 2011.

Leslie Fox and Ian Bax Parker. Chebyshev polynomials in numerical analysis. Technical report, 1968.

Oded Galor. *Discrete dynamical systems*. Springer Science & Business Media, 2007.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. 2019a. URL https://openreview.net/forum?id=r1laEnA5Ym.

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811, 2019b.

Gene H Golub and Richard S Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order richardson iterative methods. *Numerische Mathematik*, 3(1):147–156, 1961.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3):161–220, 1990.

Magnus R Hestenes, Eduard Stiefel, et al. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49(6):409–436, 1952.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Adam Ibrahim, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. *arXiv preprint arXiv:1906.07300*, 2019.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. 1976.

Cornelius Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952.

Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob N Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *J. Mach. Learn. Res.*, 20:84–1, 2019.

Tianyi Lin, Chi Jin, Michael Jordan, et al. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Thomas A Manteuffel. The tchebychev iteration for nonsymmetric linear systems. *Numerische Mathematik*, 28(3):307–327, 1977.

John C Mason and David C Handscomb. *Chebyshev polynomials*. CRC press, 2002.

Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg8jjC9KQ.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. In *Advances in Neural Information Processing Systems*, pages 1825–1835, 2017.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507, 2020.

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

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

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Available at SSRN 970903*, 2006.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Olavi Nevanlinna. *Convergence of Iterations for Linear Equations*. Springer Science & Business Media, 1993.

Wilhelm Niethammer and Richard S Varga. The analysis ofk-step iterative methods for linear systems from summability theory. *Numerische Mathematik*, 41(2):177–206, 1983.

Peter J Olver. Nonlinear systems. http://www-users.math.umn.edu/~olver/ln_/nls.pdf, 2015.

Fabian Pedregosa and Damien Scieur. Average-case acceleration through spectral density estimation. In *International Conference on Machine Learning*, 2020.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computa-*

tional Mathematics and Mathematical Physics, 4(5): 1–17, 1964.

Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Ernest K Ryu and Stephen Boyd. Primer on monotone operator methods. *Appl. Comput. Math*, 15(1):3–43, 2016.

Damien Scieur and Fabian Pedregosa. Universal average-case optimality of polyak momentum. In *International Conference on Machine Learning*, 2020.

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

J Von Neumann and O Morgenstern. Theory of games and economic behavior. 1944.

Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. *arXiv preprint arXiv:2006.06359*, 2020.

Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In *International Conference on Learning Representations*, 2019.

HE Wrigley. Accelerating the jacobi method for solving simultaneous equations by chebyshev extrapolation when the eigenvalues of the iteration matrix are complex. *The Computer Journal*, 6(2):169–176, 1963.

Guodong Zhang, Xuchao Bao, Laurent Lessard, and Roger Grosse. A unified analysis of first-order methods for smooth games via integral quadratic constraints. *arXiv preprint arXiv:2009.11359*, 2020a.

Guojun Zhang, Pascal Poupart, and Yaoliang Yu. Optimality and stability in non-convex-non-concave min-max optimization. *arXiv preprint arXiv:2002.11875*, 2020b.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.