Unified Spectral Analysis and Accelerated Convergence in Multi-Player Differentiable Games

Chris Junchi Li$^\diamond$

Department of Electrical Engineering and Computer Sciences$^\diamond$
University of California, Berkeley

September 27, 2024

Abstract

In recent years, gradient-based optimization methods have become central to machine learning, particularly in complex frameworks such as multi-agent systems, generative adversarial networks (GANs), and reinforcement learning. These settings often involve multi-player games, where each player optimizes its own objective, leading to challenges such as rotational dynamics and slow convergence in standard methods like gradient descent (GD). This paper presents a unified spectral and convergence analysis of three key gradient-based algorithms—Gradient Descent (GD), Extragradient (EG), and Optimistic Gradient (OG)—in the context of differentiable games. By leveraging spectral techniques, we offer a comprehensive analysis that encompasses bilinear, strongly monotone, and intermediate game structures. Our results establish global and local convergence guarantees with tighter rates than previous studies, while introducing a novel geometric framework to extend momentum-based acceleration techniques to games with complex eigenvalue distributions. These findings enhance the theoretical understanding of optimization in multi-agent systems and provide practical insights for improving convergence in adversarial and cooperative games.

Keywords: Multi-player differentiable games, gradient-based optimization, Nash equilibrium, spectral analysis, extragradient method, optimistic gradient, convergence rates

1 Introduction

Optimization algorithms, particularly gradient-based methods, have been crucial to the advancement of machine learning, particularly in areas like neural network training, generative modeling, and reinforcement learning. Traditionally, these methods were designed for single-objective convex optimization. However, as machine learning frameworks evolve, many modern applications involve multi-agent systems where multiple interacting players aim to optimize their own objectives. These problems, often modeled as differentiable games, arise in areas such as generative adversarial networks (GANs), multi-agent reinforcement learning, and actor-critic models.

In multi-player differentiable games, each player's strategy affects the others, making the optimization task more complex than standard single-objective problems. The goal is often to reach a Nash equilibrium, where no player can unilaterally improve their outcome. However, these games introduce unique challenges, particularly in terms of dynamics. Standard optimization algorithms like gradient descent (GD) struggle to converge due to rotational dynamics, leading to oscillations rather than convergence, as seen in bilinear game examples.

To address these issues, alternative algorithms such as the extragradient (EG) and optimistic gradient (OG) methods have been developed. These methods adjust the gradient updates to handle adversarial and cooperative interactions between players, improving convergence behavior. However, despite these advances, the performance of these methods across the full spectrum of game types, from bilinear to strongly monotone, remains inadequately explored. Much of the existing literature focuses on specific cases, leaving gaps in a unified understanding of these algorithms in more general game settings.

This paper seeks to bridge this gap by providing a comprehensive spectral analysis of gradient-based methods in differentiable games. We analyze the convergence properties of GD, EG, and OG, offering both local and global convergence guarantees. Our work introduces a novel geometric framework to examine the spectral shape of smooth games, revealing how the distribution of eigenvalues influences the difficulty of the game and the potential for acceleration. By leveraging matrix iteration theory, we offer insights into how these methods can be accelerated, particularly in games dominated by bilinear structures or with small imaginary perturbations.

Contributions Our key contributions are as follows:

- We provide a unified spectral analysis of gradient-based methods, encompassing bilinear, strongly monotone, and intermediate game structures. This analysis reveals improved convergence rates for the extragradient (EG) and optimistic gradient (OG) methods.

- We introduce a geometric framework for understanding game conditioning via the concept of *spectral shape*, linking the difficulty of a game to the eigenvalue distribution of the Jacobian. This approach enables us to design an optimal algorithm for bilinear games and derive new lower bounds, demonstrating the near-optimality of EG among first-order methods.

- We establish global convergence guarantees for GD, EG, and OG, and propose an accelerated EG method for bilinear games with imaginary perturbations. We also enhance consensus optimization by incorporating momentum, achieving near-accelerated convergence rates.

Organization The remainder of the paper is organized as follows: Section 2 discusses related work on gradient-based methods in differentiable games and their convergence properties. Section 3 presents the theoretical framework for analyzing gradient-based methods using spectral techniques. Section 4 provides a detailed spectral analysis of extragradient and optimistic gradient methods, with a comparison to gradient descent. In Section 5, we offer global convergence results and prove new lower bounds for the considered methods. Finally, Section 6 concludes the paper and discusses potential future research directions.

2 Unified Convergence and Spectral Analysis of Gradient-Based Methods in Differentiable Multi-Player Games

An increasing number of frameworks rely on optimization problems that involve multiple players and objectives. For instance, actor-critic models [PV16], generative adversarial networks (GANs) [GPM$^+$14] and automatic curricula [SKSF18] can be cast as two-player games.

Hence games are a generalization of the standard single-objective framework. The aim of the optimization is to find *Nash equilibria*, that is to say situations where no player can unilaterally decrease their loss. However, new issues that were not present for single-objective problems arise. The presence of rotational dynamics prevent standard algorithms such as the gradient method to converge on simple bilinear examples [Goo16, BRM$^+$18]. Furthermore, stationary points of the gradient dynamics are not necessarily Nash equilibria [ADLH18, MJS19].

	[Tse95]	[GBV$^+$19]	[MOP19]	[ALW19]	This work §2.4
EG	$c\frac{\mu^2}{L^2}$	-	$\frac{\mu}{4L}$	-	$\frac{1}{4}\left(\frac{\mu}{L}+\frac{\gamma^2}{16L^2}\right)$
OG	-	$\frac{\mu}{4L}$	$\frac{\mu}{4L}$	-	$\frac{1}{4}\left(\frac{\mu}{L}+\frac{\gamma^2}{32L^2}\right)$
CO	-	-	-	$\frac{\gamma^2}{4L_H^2}$	$\frac{\mu^2}{2L_H^2}+\frac{\gamma^2}{2L_H^2}$

Table 1. Summary of the global convergence results presented in §2.4 for extragradient (EG), optimistic gradient (OMD) and consensus optimization (CO) methods. If a result shows that the iterates converge as $\mathcal{O}((1-r)^t)$, the quantity r is reported (the larger the better). The letter c indicates that the numerical constant was not reported by the authors. μ is the strong monotonicity of the vector field, γ is a global lower bound on the singular values of ∇v , L is the Lipschitz constant of the vector field and L_H^2 the Lipschitz-smoothness of $\frac{1}{2}\|v\|_2^2$. For instance, for the so-called bilinear example (Ex. 1), we have $\mu = 0$ and $\gamma = \sigma_{\min}(A)$. Note that for this particular example, previous papers developed a specific analysis that breaks when a small regularization is added (see Ex. 3).

Some recent progress has been made by introducing new methods specifically designed with games or variational inequalities in mind. The main example are the optimistic gradient method (OG) introduced by [Rakhlin and Sridharan(2013)] initially for online learning, consensus optimization (CO) which adds a regularization term to the optimization problem and the extragradient method (EG) originally introduced by [Kor76]. Though these news methods and the gradient method (GD) have similar performance in convex optimization, their behaviour seems to differ when applied to games: unlike gradient, they converge on the so-called bilinear example [Tse95, GBV$^+$19, MOP19, ALW19].

However, linear convergence results for EG and OG (a.k.a extrapolation from the past) in particular have only been proven for either strongly monotone variational inequalities problems, which include strongly convex-concave saddle point problems, or in the bilinear setting separately [Tse95, GBV$^+$19, MOP19].

In this section, we study the dynamics of such gradient-based methods and in particular GD, EG and more generally multi-step extrapolations methods for unconstrained games. Our objective is three-fold. First, taking inspiration from the analysis of GD by [GHP$^+$19], we aim at providing a single precise analysis of EG which covers both the bilinear and the strongly monotone settings and their intermediate cases. Second, we are interested in theoretically comparing EG to GD and general multi-step extrapolations through upper and lower bounds on convergence rates. Third, we provide a framework to extend the unifying results of spectral analysis in global guarantees and leverage it to prove tighter convergence rates for OG and CO.

The remainder of this section is organized as follows: In Section 2, we discuss related work on gradient-based methods in differentiable games and their convergence properties. Section 3 presents the theoretical framework for analyzing gradient-based methods using spectral techniques. Section 4 provides a detailed spectral analysis of extragradient and optimistic gradient methods, with a comparison to gradient descent. In Section 5, we offer global convergence results and prove new lower bounds for the considered methods. Finally, Section 6 concludes the paper and discusses potential future research directions.

2.1 Background and motivation

2.1.1 *n*-player differentiable games

Following [BRM+18], a *n-player differentiable game* can be defined as a family of twice continuously differentiable losses $l_i : \mathbb{R}^d \to \mathbb{R}$ for $i = 1, \ldots, n$. The parameters for player i are $\omega^i \in \mathbb{R}^{d_i}$ and we note $\omega = (\omega^1, \ldots, \omega^n) \in \mathbb{R}^d$ with $d = \sum_{i=1}^n d_i$. Ideally, we are interested in finding an *unconstrained Nash equilibrium* [VNM44]: that is to say a point $\omega^* \in \mathbb{R}^d$ such that

$$\forall i \in \{1, \ldots, n\}, \quad (\omega^i)^* \in \operatorname*{arg\,min}_{\omega^i \in \mathbb{R}^{d_i}} l_i((\omega^{-i})^*, \omega^i)$$

where the vector $(\omega^{-i})^*$ contains all the coordinates of ω^* except the i^{th} one. Moreover, we say that a game is *zero-sum* if $\sum_{i=1}^n l_i = 0$. For instance, following [Mescheder et al.(2017), GHP+19], the standard formulation of GANs from [GPM+14] can be cast as a two-player zero-sum game. The Nash equilibrium corresponds to the desired situation where the generator exactly capture the data distribution, completely confusing a perfect discriminator.

Let us now define the *vector field*

$$v(\omega) = \left(\nabla_{\omega^1} l_1(\omega)\ , \quad \cdots \quad , \ \nabla_{\omega^n} l_n(\omega)\right)$$

associated to a n-player game and its Jacobian:

$$\nabla v(\omega) = \begin{pmatrix} \nabla^2_{\omega^1} l_1(\omega) & \ldots & \nabla_{\omega^n}\nabla_{\omega^1} l_1(\omega) \\ \vdots & & \vdots \\ \nabla_{\omega^1}\nabla_{\omega^n} l_n(\omega) & \ldots & \nabla^2_{\omega^n} l_n(\omega) \end{pmatrix}$$

We say that v is *L-Lipschitz* for some $L \geq 0$ if $\|v(\omega) - v(\omega')\| \leq L\|\omega - \omega'\| \ \forall \omega, \omega' \in \mathbb{R}^d$, that v is *μ-strongly monotone* for some $\mu \geq 0$, if $\mu\|\omega - \omega'\|^2 \leq (v(\omega) - v(\omega'))^T(\omega - \omega') \ \forall \omega, \omega' \in \mathbb{R}^d$.

A Nash equilibrium is always a *stationary* point of the gradient dynamics, i.e. a point $\omega \in \mathbb{R}^d$ such that $v(\omega) = 0$. However, as shown by [ADLH18, MJS19, Berard et al.(2019)], in general, being a Nash equilibrium is neither necessary nor sufficient for being a locally stable stationary point, but if v is monotone, these two notions are equivalent. Hence, in this work we focus on finding stationary points. One important class of games is *saddle-point problems*: two-player games with $l_1 = -l_2$. If v is monotone, or equivalently f is convex-concave, stationary points correspond to the solutions of the min-max problem

$$\min_{\omega_1 \in \mathbb{R}^{d_1}} \max_{\omega_2 \in \mathbb{R}^{d_2}} l_1(\omega_1, \omega_2)$$

[GHP+19] and [BRM+18] mentioned two particular classes of games, which can be seen as the two opposite ends of a spectrum. As the definitions vary, we only give the intuition for these two categories. The first one is *adversarial games*, where the Jacobian has eigenvalues with small real parts and large imaginary parts and the cross terms $\nabla_{\omega_i}\nabla_{\omega_j} l_j(\omega)$, for $i \neq j$, are dominant. Ex. 1 gives a prime example of such game that has been heavily studied: a simple bilinear game whose Jacobian is anti-symmetric and so only has imaginary eigenvalues (see Lem. 23 in App. B.1.5):

Example 1 (Bilinear game)**.**

$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^m} x^T A y + b^T x + c^T y$$

with $A \in \mathbb{R}^{m \times m}$ non-singular, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^m$.

If A is non-singular, there is an unique stationary point which is also the unique Nash equilibrium. The gradient method is known not to converge in such game while the proximal point and extragradient methods converge [Roc76, Tse95].

Bilinear games are of particular interest to us as they are seen as models of the convergence problems that arise during the training of GANs. Indeed, [Mescheder et al.(2017)] showed that eigenvalues of the Jacobian of the vector field with small real parts and large imaginary parts could be at the origin of these problems. Bilinear games have pure imaginary eigenvalues and so are limiting models of this situation. Moreover, they can also be seen as a very simple type of WGAN, with the generator and the discriminator being both linear, as explained in [GBV$^+$19, MGN18].

The other category is *cooperative games*, where the Jacobian has eigenvalues with large positive real parts and small imaginary parts and the diagonal terms $\nabla^2_{\omega_i} l_i$ are dominant. Convex minimization problems are the archetype of such games. Our hypotheses, for both the local and the global analyses, encompass these settings.

2.1.2 Methods and convergence analysis

Convergence theory of fixed-point iterations. Seeing optimization algorithms as the repeated application of some operator allows us to deduce their convergence properties from the spectrum of this operator. This point of view was presented by [Pol87, Ber99] and recently used by [ASS16, Mescheder et al.(2017), GHP$^+$19] for instance. The idea is that the iterates of a method $(\omega_t)_t$ are generated by a scheme of the form:

$$\omega_{t+1} = F(\omega_t) \quad \forall t \geq 0$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is an operator representing the method. Near a stationary point ω^*, the behavior of the iterates is mainly governed by the properties of $\nabla F(\omega^*)$ as $F(\omega) - \omega^* \approx \nabla F(\omega^*)(\omega - \omega^*)$. This is formalized by the following classical result:

Theorem 1 ([Pol87])**.** *Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be continuously differentiable and let $\omega^* \in \mathbb{R}^d$ be a fixed point of F. If $\rho(\nabla F(\omega^*)) < 1$, then for ω_0 in a neighborhood of ω^*, the iterates $(\omega_t)_t$ defined by $\omega_{t+1} = F(\omega_t)$ for all $t \geq 0$ converge linearly to ω^* at a rate of $\mathcal{O}((\rho(\nabla F(\omega^*)) + \epsilon)^t)$ for all $\epsilon > 0$.*

This theorem means that to derive a local rate of convergence for a given method, one needs only to focus on the eigenvalues of $\nabla F(\omega^*)$. Note that if the operator F is linear, there exists slightly stronger results such as Thm. 11 in appendix B.1.3.

Gradient method. Following [GHP$^+$19], we define GD as the application of the operator $F_\eta(\omega) := \omega - \eta v(\omega)$, for $\omega \in \mathbb{R}^d$. Thus we have:

$$\omega_{t+1} = F_\eta(\omega_t) = \omega_t - \eta v(\omega_t) \tag{GD}$$

Proximal point. For v monotone [Min62, Roc76], the proximal point operator can be defined as $P_\eta(\omega) = (\mathrm{Id} + \eta v)^{-1}(\omega)$ and therefore can be seen as an implicit scheme: $\omega_{t+1} = \omega_t - \eta v(\omega_{t+1})$.

Extragradient. EG was introduced by [Kor76] in the context of variational inequalities. Its update rule is

$$\omega_{t+1} = \omega_t - \eta v(\omega_t - \eta v(\omega_t)) \tag{EG}$$

It can be seen as an approximation of the implicit update of the proximal point method. Indeed [Nem04] showed a rate of $\mathcal{O}(1/t)$ for extragradient by treating it as a "good enough" approximation of the proximal point method. To see this, fix $\omega \in \mathbb{R}^d$. Then $P_\eta(\omega)$ is the solution of $z = \omega - \eta v(z)$. Equivalently, $P_\eta(\omega)$ is the fixed point of

$$\varphi_{\eta,\omega} : z \longmapsto \omega - \eta v(z) \tag{1}$$

which is a contraction for $\eta > 0$ small enough. From Picard's fixed point theorem, one gets that the proximal point operator $P_\eta(\omega)$ can be obtained as the limit of $\varphi^k_{\eta,\omega}(\omega)$ when k goes to infinity. What [Nem04] showed is that $\varphi^2_{\eta,\omega}(\omega)$, that is to say the extragradient update, is close enough to the result of the fixed point computation to be used in place of the proximal point update without affecting the sublinear convergence speed. Our analysis of multi-step extrapolation methods will encompass all the iterates $\varphi^k_{\eta,\omega}$ and we will show that a similar phenomenon happens for linear convergence rates.

Optimistic gradient. Originally introduced in the online learning literature [Chiang et al.(2012), Rakhlin and Sridharan(2013)] as a two-steps method, [DISZ18] reformulated it with only one step in the unconstrained case:

$$w_{t+1} = w_t - 2\eta v(w_t) + \eta v(w_{t-1}) \tag{OG}$$

Consensus optimization. Introduced by [Mescheder et al.(2017)] in the context of games, consensus optimization is a second-order yet efficient method, as it only uses a Hessian-vector multiplication whose cost is the same as two gradient evaluations [Pearlmutter(1994)]. We define the CO update as:

$$\omega_{t+1} = \omega_t - (\alpha v(\omega_t) + \beta \nabla H(\omega_t)) \tag{CO}$$

where $H(\omega) = \frac{1}{2}\|v(\omega)\|_2^2$ and $\alpha, \beta > 0$ are step sizes.

2.1.3 *p*-SCLI framework for game optimization

In this section, we present an extension of the framework of [ASS16] to derive lower bounds for game optimization (also see §B.3). The idea of this framework is to see algorithms as the iterated application of an operator. If the vector field is linear, this transformation is linear too and so its behavior when iterated is mainly governed by its spectral radius. This way, showing a lower bound for a class of algorithms is reduced to lower bounding a class of spectral radii.

We consider $\mathcal{V}_d$ the set of linear vector fields $v : \mathbb{R}^d \to \mathbb{R}^d$, i.e., vector fields v whose Jacobian ∇v is a constant $d \times d$ matrix.[1] The class of algorithms we consider is the class of 1-*Stationary Canonical Linear Iterative algorithms (*1-*SCLI)*. Such an algorithm is defined by a mapping $\mathcal{N} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$. The associated update rule can be defined through,

$$F_{\mathcal{N}}(\omega) = w + \mathcal{N}(\nabla v)v(\omega) \quad \forall \omega \in \mathbb{R}^d \tag{2}$$

This form of the update rule is required by the consistency condition of [ASS16] which is necessary for the algorithm to converge to stationary points, as discussed in §B.3. Also note that 1-SCLI are first-order methods that use only the last iterate to compute the next one. Accelerated methods such as accelerated gradient descent [Nes04] or the heavy ball method [Pol64] belong in fact to the class of 2-SCLI, which encompass methods which uses the last two iterates.

[1] With a slight abuse of notation, we also denote by ∇v this matrix.

As announced above, the spectral radius of the operator gives a lower bound on the speed of convergence of the iterates of the method on affine vector fields, which is sufficient to include bilinear games, quadratics and so strongly monotone settings too.

Theorem 2 ([ASS16]). *For all $v \in \mathcal{V}_d$, for almost every*[2] *initialization point $\omega_0 \in \mathbb{R}^d$, if $(\omega_t)_t$ are the iterates of $F_{\mathcal{N}}$ starting from ω_0,*

$$\|\omega_t - \omega^*\| \geq \Omega(\rho(\nabla F_{\mathcal{N}})^t \|\omega_0 - \omega^*\|)$$

2.2 Revisiting GD for games

In this section, our goal is to illustrate the precision of the spectral bounds and the complexity of the interactions between players in games. We first give a simplified version of the bound on the spectral radius from [GHP$^+$19] and show that their results also imply that this rate is tight.

Theorem 3. *Let ω^* be a stationary point of v and denote by σ^* the spectrum of $\nabla v(\omega^*)$. If the eigenvalues of $\nabla v(\omega^*)$ all have positive real parts, then*

(i). [GHP$^+$19] For $\eta = \min_{\lambda \in \sigma^*} \Re(1/\lambda)$, the spectral radius of F_η can be upper-bounded as

$$\rho(\nabla F_\eta(\omega^*))^2 \leq 1 - \min_{\lambda \in \sigma^*} \Re(1/\lambda) \min_{\lambda \in \sigma^*} \Re(\lambda)$$

(ii). For all $\eta > 0$, the spectral radius of the gradient operator F_η at ω^* is lower bounded by

$$\rho(\nabla F_\eta(\omega^*))^2 \geq 1 - 4 \min_{\lambda \in \sigma^*} \Re(1/\lambda) \min_{\lambda \in \sigma^*} \Re(\lambda)$$

This result is stronger than what we need for a standard lower bound: using Thm. 2, this yields a lower bound on the convergence of the iterates for all games with affine vector fields.

We then consider a saddle-point problem, and under some assumptions presented below, one can interpret the spectral rate of the gradient method mentioned earlier in terms of the standard strong convexity and Lipschitz-smoothness constants. There are several cases, but one of them is of special interest to us as it demonstrates the precision of spectral bounds.

Example 2 (Highly adversarial saddle-point problem)**.** Consider $\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^m} f(x, y)$ with f twice differentiable such that

(i). f satisfies, with μ_1, μ_2 and μ_{12} non-negative,

$$\mu_1 I \preccurlyeq \nabla_x^2 f \preccurlyeq L_1 I \quad \mu_2 I \preccurlyeq -\nabla_y^2 f \preccurlyeq L_2 I$$
$$\mu_{12}^2 I \preccurlyeq (\nabla_x \nabla_y f)^T (\nabla_x \nabla_y f) \preccurlyeq L_{12}^2 I$$

such that $\mu_{12} > 2 \max(L_1 - \mu_2, L_2 - \mu_1)$.

(ii). There exists a stationary point $\omega^* = (x^*, y^*)$ and at this point, $\nabla_y^2 f(\omega^*)$ and $\nabla_x \nabla_y f(\omega^*)$ commute and $\nabla_x^2 f(\omega^*)$, $\nabla_y^2 f(\omega^*)$ and $(\nabla_x \nabla_y f(\omega^*))^T (\nabla_x \nabla_y f(\omega^*))$ commute.

[2] For any measure absolutely continuous w.r.t. the Lebesgue measure.

Assumption *(i)* corresponds to a highly adversarial setting as the coupling (represented by the cross derivatives) is much bigger than the Hessians of each player. Assumption *(ii)* is a technical assumption needed to compute a precise bound on the spectral radius and holds if, for instance, the objective is separable, i.e. $f(x, y) = \sum_{i=1}^{m} f_i(x_i, y_i)$. Using these assumptions, we can upper bound the rate of Thm. 3 as follows:

Corollary 1. *Under the assumptions of Thm. 3 and Ex. 2,*

$$\rho(\nabla F_\eta(\omega^*))^2 \leq 1 - \frac{1}{4}\frac{(\mu_1+\mu_2)^2}{L_{12}^2 + L_1 L_2} \tag{3}$$

What is surprising is that, in some regimes, this result induces faster local convergence rates than the existing upper-bound for EG [Tse95]:

$$1 - \frac{\min(\mu_1,\mu_2)}{4L_{max}} \quad \text{where} \quad L_{max} = \max(L_1, L_2, L_{12}) \tag{4}$$

If, say, μ_2 goes to zero, that is to say the game becomes unbalanced, the rate of EG goes to 1 while the one of (3) stays bounded by a constant which is strictly less than 1. Indeed, the rate of Cor. 1 involves the arithmetic mean of μ_1 and μ_2, which is roughly the maximum of them, while (4) makes only the minimum of the two appear. This adaptivity to the best strong convexity constant is not present in the standard convergence rates of the EG method. We remedy this situation with a new analysis of EG in the following section.

2.3 Spectral analysis of multi-step EG

In this section, we study the local dynamics of EG and, more generally, of extrapolation methods. Define a *k-extrapolation method* (k-EG) by the operator

$$F_{k,\eta} : \omega \mapsto \varphi_{\eta,\omega}^k(\omega) \quad \text{with} \quad \varphi_{\eta,\omega} : z \mapsto \omega - \eta v(z) \tag{5}$$

We are essentially considering all the iterates of the fixed point computation discussed in §2.1.2. Note that $F_{1,\eta}$ is GD while $F_{2,\eta}$ is EG. We aim at studying the local behavior of these methods at stationary points of the gradient dynamics, so fix ω^* s.t. $v(\omega^*) = 0$ and let $\sigma^* = \operatorname{Sp} \nabla v(\omega^*)$. We compute the spectra of these operators at this point and this immediately yields the spectral radius on the proximal point operator:

Lemma 1. *The spectra of the k-extrapolation operator and the proximal point operator are given by:*

$$\operatorname{Sp} \nabla F_{\eta,k}(\omega^*) = \left\{ \textstyle\sum_{j=0}^{k} (-\eta\lambda)^j \mid \lambda \in \sigma^* \right\}$$
$$\textit{and} \quad \operatorname{Sp} \nabla P_\eta(\omega^*) = \left\{ (1+\eta\lambda)^{-1} \mid \lambda \in \sigma^* \right\}$$

Hence, for all $\eta > 0$, the spectral radius of the operator of the proximal point method is equal to:

$$\rho(\nabla P_\eta(\omega^*))^2 = 1 - \min_{\lambda \in \sigma^*} \frac{2\eta \Re\lambda + \eta^2|\lambda|^2}{|1+\eta\lambda|^2} \tag{6}$$

Again, this shows that a k-EG is essentially an approximation of proximal point for small step sizes as $(1+\eta\lambda)^{-1} = \sum_{j=0}^{k} (-\eta\lambda)^j + \mathcal{O}\left(|\eta\lambda|^{k+1}\right)$. This could suggest that increasing the number of extrapolations might yield better methods but we will actually see that $k = 2$ is enough to achieve a similar rate to proximal. We then bound the spectral radius of $\nabla F_{\eta,k}(\omega^*)$:

Theorem 4. *Let $\sigma^* = \mathrm{Sp}\,\nabla v(\omega^*)$. If the eigenvalues of $\nabla v(\omega^*)$ all have non-negative real parts, the spectral radius of the k-extrapolation method for $k \geq 2$ satisfies:*

$$\rho(\nabla F_{\eta,k}(\omega^*))^2 \leq 1 - \min_{\lambda \in \sigma^*} \frac{2\eta \Re\lambda + \frac{7}{16}\eta^2|\lambda|^2}{|1+\eta\lambda|^2} \tag{7}$$

$\forall \eta \leq \frac{1}{4^{\frac{1}{k-1}}} \frac{1}{\max_{\lambda\in\sigma^*}|\lambda|}$. *For $\eta = (4\max_{\lambda\in\sigma^*}|\lambda|)^{-1}$, this can be simplified as (noting $\rho := \rho(\nabla F_{\eta,k}(\omega^*))$):*

$$\rho^2 \leq 1 - \frac{1}{4}\left(\frac{\min_{\lambda\in\sigma^*}\Re\lambda}{\max_{\lambda\in\sigma^*}|\lambda|} + \frac{1}{16}\frac{\min_{\lambda\in\sigma^*}|\lambda|^2}{\max_{\lambda\in\sigma^*}|\lambda|^2}\right) \tag{8}$$

The zone of convergence of extragradient as provided by this theorem is illustrated in Fig. 1.

The bound of (8) involves two terms: the first term can be seen as the strong monotonicity of the problem, which is predominant in convex minimization problems, while the second shows that even in the absence of it, this method still converges, such as in bilinear games. Furthermore, in situation in between, this bound shows that the extragradient method exploits the biggest of these quantities as they appear as a sum as illustrated by the following simple example.

Example 3 ("In between" example)**.**

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} \frac{\epsilon}{2}(x^2 - y^2) + xy \quad \text{for } 1 \geq \epsilon > 0$$

Though for ϵ close to zero, the dynamics will behave as such, this is not a purely bilinear game. The associated vector field is only ϵ-strongly monotone and convergence guarantees relying only on strong monotonicity would give a rate of roughly $1 - \epsilon/4$. However Thm. 4 yields a convergence rate of roughly $1 - 1/64$ for extragradient.

Similarity to the proximal point method. First, note that the bound (7) is surprisingly close to the one of the proximal method (6). However, one can wonder why the proximal point converges with any step size — and so arbitrarily fast — while it is not the case for the k-EG, even as k goes to infinity. The reason for this difference is that for the fixed point iterates to converge to the proximal point operator, one needs $\varphi_{\eta,\omega}$ to be a contraction and so to have η small enough, at least $\eta < (\max_{\lambda\in\sigma^*}|\lambda|)^{-1}$ for local guarantees. This explains the bound on the step size for k-EG .

Comparison with the gradient method. We can now compare this result for EG with the convergence rate of the gradient method Thm. 3 which was shown to be tight. In general $\min_{\lambda\in\sigma^*}\Re(1/\lambda) \leq (\max_{\lambda\in\sigma^*}|\lambda|)^{-1}$ and, for adversarial games, the first term can be arbitrarily smaller than the second one. Hence, in this setting which is of special interest to us, EG has a much faster convergence speed than GD.

Recovery of known rates. If v is μ-strongly monotone and L-Lipschitz, this bound is at least as precise as the standard one $1 - \mu/(4L)$ as μ lower bounds the real part of the eigenvalues of the Jacobian, and L upper bounds their magnitude, as shown in Lem. 10 in §B.2.2. On the other hand, Thm. 4 also recovers the standard rates for the bilinear problem,[3] as shown below:

[3] Note that by exploiting the special structure of the bilinear game and the fact that $k = 2$, one could derive a better constant in the rate. Moreover, our current spectral tools cannot handle the singularity which arises if the two players have a different number of parameters. We provide sharper results to handle this difficulty in appendix B.5.

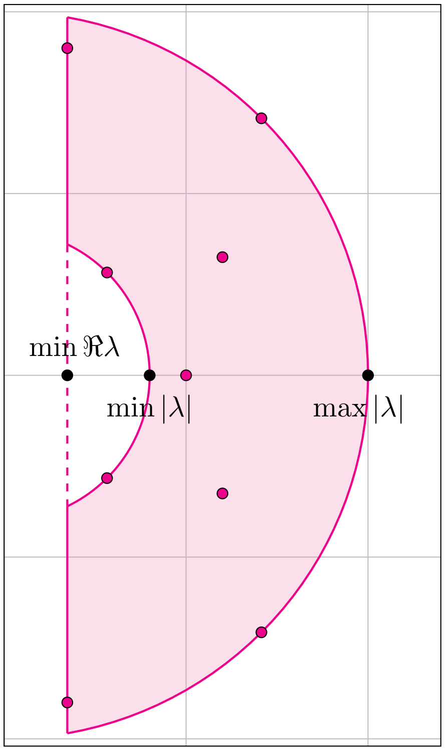

Figure 1. Illustration of the three quantities involved in Thm. 4. The magenta dots are an example of eigenvalues belonging to σ^*. Note that σ^* is always symmetric with respect to the real axis because the Jacobian is a real matrix (and thus non-real eigenvalues are complex conjugates). Note how $\min \Re\lambda$ may be significantly smaller that $\min|\lambda|$.

Corollary 2 (Bilinear game)**.** *Consider Ex. 1. The iterates of the k-extrapolation method with $k \geq 2$ converge globally to ω^* at a linear rate of* $\mathcal{O}\big(\big(1 - \frac{1}{64}\frac{\sigma_{min}(A)^2}{\sigma_{max}(A)^2}\big)^t\big)$.

Note that this rate is similar to the one derived by [GHP+19] for alternating gradient descent with negative momentum. This raises the question of whether general acceleration exists for games, as we would expect the quantity playing the role of the condition number in Cor. 2 to appear without the square in the convergence rate of a method using momentum.

Finally it is also worth mentioning that the bound of Thm. 4 also displays the adaptivity discussed in §2.2. Hence, the bound of Thm. 4 can be arbitrarily better than the rate (4) for EG from the literature and also better than the global convergence rate we prove below.

Lower bounds for extrapolation methods. We now show that the rates we proved for EG are tight and optimal by deriving lower bounds of convergence for general extrapolation methods. As described in §2.1.3, a 1-SCLI method is parametrized by a polynomial $\mathcal{N}$. We consider the class of methods where $\mathcal{N}$ is any polynomial of degree at most $k-1$, and we will derive lower bounds for this class. This class is large enough to include all the k'-extrapolation methods for $k' \leq k$ with possibly different step sizes for each extrapolation step (see §B.4 for more examples).

Our main result is that no method of this class can significantly beat the convergence speed of EG of Thm. 4 and Thm. 6. We proceed in two steps: for each of the two terms of these bounds, we provide an example matching it up to a factor. In (i) of the following theorem, we give an example of convex optimization problem which matches the real part, or strong monotonicity, term. Note that this example is already an extension of [ASS16] as the authors only considered constant $\mathcal{N}$. Next, in (ii), we match the other term with a bilinear game example.

Theorem 5. *Let $0 < \mu, \gamma < L$. (i) If $d - 2 \geq k \geq 3$, there exists $v \in \mathcal{V}_d$ with a symmetric positive Jacobian whose spectrum is in $[\mu, L]$, such that for any $\mathcal{N}$real polynomial of degree at most $k-1$, $\rho(F_{\mathcal{N}}) \geq 1 - \frac{4k^3}{\pi}\frac{\mu}{L}$.*

(ii) If $d/2 - 2 \geq k/2 \geq 3$ and d is even, there exists $v \in \mathcal{V}_d$ L-Lipschitz with $\min_{\lambda \in \operatorname{Sp} \nabla v} |\lambda| = \sigma_{min}(\nabla v) \geq \gamma$ corresponding to a bilinear game of Example 1 with $m = d/2$, such that, for any $\mathcal{N}$real polynomial of degree at most $k-1$, $\rho(F_{\mathcal{N}}) \geq 1 - \frac{k^3}{2\pi}\frac{\gamma^2}{L^2}$.

First, these lower bounds show that both our convergence analyses of EG are tight, by looking at them for $k = 3$ for instance. Then, though these bounds become looser as k grows, they still show that the potential improvements are not significant in terms of conditioning, especially compared to the change of regime between GD and EG . Hence, they still essentially match the convergence speed of EG of Thm. 4 or Thm. 6. Therefore, EG can be considered as optimal among the general class of algorithms which uses at most a fixed number of composed gradient evaluations and only the last iterate. In particular, there is no need to consider algorithms with more extrapolation steps or with different step sizes for each of them as it only yields a constant factor improvement.

2.4 Unified global proofs of convergence

We have shown in the previous section that a spectral analysis of EG yields tight and unified convergence guarantees. We now demonstrate how, combining the strong monotonicity assumption and Tseng's error bound, global convergence guarantees with the same unifying properties might be achieved.

2.4.1 Global Assumptions

[Tse95] proved linear convergence results for EG by using the projection-type error bound [Tse95, Eq. 5] which, in the unconstrained case, i.e. for $v(\omega^*) = 0$, can be written as,

$$\gamma \|\omega - \omega^*\|_2 \leq \|v(\omega)\|_2 \quad \forall \omega \in \mathbb{R}^d. \tag{9}$$

The author then shows that this condition holds for the bilinear game of Example 1 and that it induces a convergence rate of $1 - c\sigma_{min}(A)^2/\sigma_{max}(A)^2$ for some constant $c > 0$. He also shows that this condition is implied by strong monotonicity with $\gamma = \mu$. Our analysis builds on the results from [Tse95] and extends them to cover the whole range of games and recover the optimal rates.

To be able to interpret Tseng's error bound (9), as a property of the Jacobian ∇v, we slightly relax it to,

$$\gamma \|\omega - \omega'\|_2 \leq \|v(\omega) - v(\omega')\|_2, \quad \forall \omega, \omega' \in \mathbb{R}^d \tag{10}$$

This condition can indeed be related to the properties of ∇v as follows:

Lemma 2. *Let v be continuously differentiable and $\gamma > 0$: (10) holds if and only if $\sigma_{min}(\nabla v) \geq \gamma$.*

Hence, γ corresponds to a lower bound on the singular values of ∇v. This can be seen as a weaker "strong monotonicity" as it is implied by strong monotonicity, with $\gamma = \mu$, but it also holds for a square non-singular bilinear example of Example 1 with $\gamma = \sigma_{min}(A)$.

As announced, we will combine this assumption with the strong monotonicity to derive unified global convergence guarantees. Before that, note that this quantities can be related to the spectrum of $\operatorname{Sp} \nabla v(\omega^*)$ as follows – see Lem. 10 in appendix B.2.1,

$$\mu \leq \Re(\lambda), \quad \gamma \leq |\lambda| \leq L \quad \forall \lambda \in \operatorname{Sp} \nabla v(\omega^*) \tag{11}$$

Hence, theses global quantities are less precise than the spectral ones used in Thm. 4, so the following global results will be less precise than the previous ones.

2.4.2 Global analysis EG and OG

We can now state our global convergence result for EG:

Theorem 6. *Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and*

(i) μ-strongly monotone for some $\mu \geq 0$,

(ii) L-Lipschitz,

(iii) such that $\sigma_{min}(\nabla v) \geq \gamma$ for some $\gamma > 0$.

Then, for $\eta \leq (4L)^{-1}$, the iterates $(\omega_t)_t$ of (EG) *converge linearly to ω^* as, for all $t \geq 0$,*

$$\|\omega_t - \omega^*\|_2^2 \leq \left(1 - \eta\mu - \frac{7}{16}\eta^2\gamma^2\right)^t \|\omega_0 - \omega^*\|_2^2$$

As for Thm. 4, this result not only recovers both the bilinear and the strongly monotone case, but shows that EG actually gets the best of both world when in between. Furthermore this rate is surprisingly similar to the result of Thm. 4 though less precise, as discussed.

Combining our new proof technique and the analysis provided by [GBV$^+$19], we can derive a similar convergence rate for the optimistic gradient method.

Theorem 7. *Under the same assumptions as in Thm. 6, for $\eta \leq (4L)^{-1}$, the iterates $(\omega_t)_t$ of* (OG) *converge linearly to ω^* as, for all $t \geq 0$,*

$$\|\omega_t - \omega^*\|_2^2 \leq 2\left(1 - \eta\mu - \frac{1}{8}\eta^2\gamma^2\right)^{t+1} \|\omega_0 - \omega^*\|_2^2$$

Interpretation of the condition numbers. As in the previous section, this rate of convergence for EG is similar to the rate of the proximal point method for a small enough step size, as shown by Prop. 9 in §B.2.2. Moreover, the proof of the latter gives insight into the two quantities appearing in the rate of Thm. 6. Indeed, the convergence result for the proximal point method is obtained by bounding the singular values of ∇P_η, and so we compute,[4]

$$(\nabla P_\eta)^T \nabla P_\eta = \left(I_d + \eta\mathcal{H}(\nabla v) + \eta^2 \nabla v \nabla v^T\right)^{-1}$$

where $\mathcal{H}(\nabla v) := \frac{\nabla v + \nabla v^T}{2}$. This explains the quantities L/μ and L^2/γ^2 appear in the convergence rate, as the first corresponds to the condition number of $\mathcal{H}(\nabla v)$ and the second to the condition number of $\nabla v \nabla v^T$. Thus, the proximal point method uses information from both matrices to converge, and so does EG, explaining why it takes advantage of the best conditioning.

[4] We dropped the dependence on ω for compactness.

2.4.3 Global analysis of consensus optimization

In this section, we give a unified proof of CO. A global convergence rate for this method was proven by [ALW19]. However it used a perturbation analysis of HGD. The drawbacks are that it required that the CO update be sufficiently close to the one of HGD and could not take advantage of strong monotonicity. Here, we combine the monotonicity μ with the lower bound on the singular value γ.

As this scheme uses second-order[5] information, we need to replace the Lipschitz hypothesis with one that also controls the variations of the Jacobian of v: we use L_H^2, the Lispchitz smoothness of H. See [ALW19] for how it might be instantiated.

Theorem 8. *Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable such that*

(i) v is μ- strongly monotone for some $\mu \geq 0$,

(ii) $\sigma_{min}(\nabla v) \geq \gamma$ for some $\gamma > 0$

(iii) H is L_H^2 Lipschitz-smooth.

Then, for $\alpha = (\mu + \sqrt{\mu^2 + 2\gamma^2})/(4L_H^2)$, $\beta = (2L_H^2)^{-1}$ the iterates of CO defined by (CO) *satisfy, for all $t \geq 0$,*

$$H(\omega_t) \leq \left(1 - \frac{\mu^2}{2L_H^2} - \left(1 + \frac{\mu}{\gamma}\right)\frac{\gamma^2}{2L_H^2}\right)^t H(\omega_0)\,.$$

This result shows that CO has the same unifying properties as EG, though the dependence on μ is worse.

This result also encompasses the rate of HGD [ALW19, Lem. 4.7]. The dependance in μ is on par with the standard rate for the gradient method (see [NS06, Eq. 2.12] for instance). However, this can be improved using a sharper assumption, as discussed in Remark 1 in appendix B.2.3, and so our result is not optimal in this regard.

3 Spectral Analysis and Acceleration in Smooth Games

Recent successes of multi-agent formulations in various areas of deep learning [GPM+14, PV16] have caused a surge of interest in the theoretical understanding of first-order methods for the solution of differentiable multi-player games [PB16, GBV+19, BRM+18, MNG17, MGN18, MJS19]. This exploration hinges on a key question: ***How fast can a first-order method be?*** In convex minimization, [Nes83, Nes04] answered this question with lower bounds for the rate of convergence and an accelerated, momentum-based algorithm matching that optimal lower bound.

The dynamics of numerical methods is often described by a vector field, F, and summarized in the spectrum of its Jacobian. In minimization problems, the eigenvalues of the Jacobian lie on the real line. On strongly convex problems, the *condition number* (the dynamic range of eigenvalues) is at the heart of Nesterov's upper and lower bound results, characterizing the hardness of an minimization problem.

Our understanding of differentiable games is nowhere close to this point. There, the eigenvalues of the Jacobian at the solution are distributed on the complex plane, suggesting a richer, more complex set of dynamics [MNG17, BRM+18]. Some old papers [Kor76, Tse95] and many recent ones [Nem04, CLO14, PB16, MNG17, GBV+19, GHP+19, DISZ18, MOP19, AMLJG19] suggest new methods and provide better upper bounds.

[5] W.r.t. the losses.

All of the above work relies on bounding the magnitude or the real part of the eigenvalues of submatrices of the Jacobian. This coarse-grain approach can be oblivious to the dependence of upper and lower bounds on the exact distribution of eigenvalues on the complex plane. More importantly, the questions of acceleration and optimality have not been answered for smooth games.

In this section, we take a different approach. We use matrix iteration theory to characterize acceleration in smooth games. Our analysis framework revolves around the *spectral shape* of a family of games, defined as the set containing all eigenvalues of the Jacobians of natural gradient dynamics in the family (cf. §3.1.2). This fine-grained analysis framework can captures the dependence of upper and lower bounds on the specific shape of the spectrum. Critically, it allows us to establish acceleration in specific families of smooth games.

The remainder of this section is organized as follows. In Section 2, we review related work on game optimization and matrix iteration theory. Section 3 introduces our spectral shape framework and applies it to bilinear games. In Section 4, we present our main acceleration results for games with small imaginary perturbations. Section 5 extends these results to consensus optimization. Finally, we conclude in Section 6 with a discussion of future research directions and applications to machine learning problems.

3.1 Setting and notation

We consider the problem of finding a stationary point $\omega^* \in \mathbb{R}^d$ of a vector field $F : \mathbb{R}^d \to \mathbb{R}^d$, i.e., $F(\omega^*) = 0.$, the solution of an unconstrained *variational inequality* problem [HP90]. A relevant special case is a n-player convex game, where ω^* corresponds to a Nash equilibrium [VNM44, BRM+18]. Consider n players $i = 1, \dots, n$ who want to minimize their loss $l_i(\omega^{(i)}, \omega^{(-i)})$. The notation $\cdot^{(-i)}$ means all indexes but i. A Nash equilibrium satisfies

$$(\omega^*)^{(i)} \in \mathop{\arg\min}_{\omega^{(i)} \in \mathbb{R}^{d_i}} l_i(\omega^{(i)}, (\omega^*)^{(-i)}) \quad \forall i \in \{1, \dots, n\}.$$

In this situation no player can unilaterally reduce its loss. The vector field of the game is

$$F(\omega) = \left[\nabla_{\omega_1} l_1^T(\omega^{(1)}, \omega^{(-1)}), ..., \nabla_{\omega_n} l_n^T(\omega^{(n)}, \omega^{(-n)})\right]^T .$$

3.1.1 First-order methods

To study lower bounds of convergence, we need a class of algorithms. We consider the classic definition[6] of first-order methods from [NY83].

Definition 1. *A* first-order method *generates*

$$\omega_t \in \omega_0 + \boldsymbol{Span}\{F(\omega_0), \dots, F(\omega_{t-1})\}, \quad t \geq 1 .$$

This class is widely used in large-scale optimization, as it involves only gradient computation. For instance, Nesterov's acceleration belongs to the class of first-order methods. On the contrary, this definition does not cover Adagrad [DHS11], that could conceptually be also considered as first-order. This is due to the diagonal re-scaling, so ω_t can go *outside* the span of gradients. The next proposition gives a way to easily identify first-order methods that fit our definition.

[6] Technically, first-order algorithms are more generally methods that have access only to first-order oracles.

Proposition 1. *[AS16] first-order methods can be written as*

$$\omega_{t+1} = \textstyle\sum_{k=0}^{t} \alpha_k^{(t)} F(\omega_k) + \beta_k^{(t)} \omega_k, \tag{12}$$

where $\sum_{k=0}^{t} \beta_k^{(t)} = 1$. The method is called oblivious if the coefficients $\alpha_k^{(t)}$ and $\beta_k^{(t)}$ are known in advance.

Oblivious methods allow the knowledge of "side information" on the function, like its smoothness constant. Most of first-order methods belong to this class, but it excludes for instance methods with adaptive step-sizes. We show how standard methods fit into this framework.

Gradient method. Consider the gradient method with time-dependant step-size: $\omega_{t+1} = \omega_t - \eta_t F(\omega_t)$. This is a first-order method, where $\alpha_t^{(t)} = -\eta_t$, $\beta_t^{(t)} = 1$ and all the other coefficients set to zero.

Momentum method. The momentum method defines iterates as $\omega_{t+1} = \omega_t - \alpha F(\omega_t) + \beta(\omega_t - \omega_{t-1})$. It fits into the previous framework with $\alpha_t^{(t)} = -\alpha$, $\beta_t^{(t)} = 1 + \beta$, $\beta_{t-1}^{(t)} = -\beta$.

Extragradient method. Though slightly trickier, the extragradient method (EG) is also encompassed by this definition. The iterates of EG are defined by $\omega_{t+1} = \omega_t - \eta F(\omega_t - \eta F(\omega_t))$ where

$$\begin{cases} \beta_t^{(t)} = 0,\ \beta_{t-1}^{(t)} = 1 & \text{if } t \text{ is odd (update)}\,, \\ \beta_t^{(t)} = 1,\ \beta_{t-1}^{(t)} = 0 & \text{if } t \text{ is even (extrapolation)}\,, \end{cases}$$

and $\alpha_t^{(t)} = -\eta$ the step size.

The next (known) lemma shows that when F is linear, first-order methods can be written using *polynomials.*

Lemma 3. *[e.g., [Chi11]] If $F(\omega) = A\omega + b$,*

$$\omega_t - \omega^* = p_t(A)(\omega_0 - \omega^*)\,, \tag{13}$$

where ω^* satisfies $A\omega^* + b = 0$ and p_t is a real polynomial of degree at most t such that $p_t(0) = 1$.

We denote by $\mathcal{P}_t$ the set of real polynomials of degree at most t such that $p_t(0) = 1$. Hence, the convergence of a first-order method can be analyzed through the sequence of polynomials $(p_t)_t$ it defines.

3.1.2 Problem class

In the previous section, when F is the linear function $F = Ax + b$, the iterates ω_t follow the relation (13) involving the polynomial p_t. Since all first-order methods can be written using polynomials (12), they follow

$$\|\omega_t - \omega^*\|_2 = \|p_t(A)(\omega_0 - \omega^*)\|_2\,. \tag{14}$$

This gives the rate of convergence of the method for a specific matrix A. Instead, we consider a larger class of problems. It consists of a set $\mathcal{M}_K$ of matrices A whose eigenvalues belong to a set K on the complex plane,

$$\mathcal{M}_K \coloneqq \{A \in \mathbb{R}^d :\ \mathrm{Sp}(A) \subset K \subset \mathbb{C}_+\}, \tag{15}$$

where $\mathrm{Sp}(A)$ is the set of eigenvalues of A and $\mathbb{C}_+$ is the set of complex numbers with positive real part. Moreover, we assume that $d \geq 2$ to avoid trivial cases.

3.1.3 Geometric intuition

This section is entirely based on the study of the support K of the eigenvalues of the Jacobian of the operator F, denoted by $\mathbf{J}_F(\omega^*)$. Before detailing our theoretical results, we give a high-level explanation of our objectives. This geometric intuition comes from the fact that the standard assumptions made in the literature correspond to particular problem classes $\mathcal{M}_K$.

Smooth and strongly convex minimization. Consider the minimization of a twice-differentiable, L-smooth and μ-strongly convex function f,

$$\mu \mathbf{I} \preceq \nabla^2 f(\omega) \preceq L\mathbf{I} \quad \forall \omega \in \mathbb{R}^d.$$

There is a link between minimization problems and games, since the vector field F becomes the gradient of the objective, and its Jacobian $\mathbf{J}_F(\omega)$ is the Hessian $\nabla^2 f(\omega)$. Thus, the class corresponding to the minimization of smooth, strongly convex functions is

$$\{F : \forall \omega \in \mathbb{R}^d,\ \operatorname{Sp} \mathbf{J}_F(\omega) \subset [\mu, L]\}\,,\ \ 0 < \mu \leq L\}.$$

Bilinear games. Consider the following problem,

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} x^\top A y\,.$$

Its Jacobian $\mathbf{J}_F(\omega)$ is constant and skew-symmetric. It is a standard linear algebra result (see Lem. 23) to show that $\operatorname{Sp} \mathbf{J}_F(\omega) \in \pm[i\sigma_{\min}(A), i\sigma_{\max}(A)]$.

Variational inequalities. The Lipchitz assumption

$$\|F(\omega) - F(\omega')\|_2^2 \leq L\|\omega - \omega'\|_2^2 \tag{16}$$

implies an upper bound on the magnitude of the eigenvalues of $\mathbf{J}_F(\omega^*)$. The strong monotonicity assumption

$$(\omega - \omega')^T(F(\omega) - F(\omega')) \geq \mu\|\omega - \omega'\|_2^2 \tag{17}$$

implies a lower bound on the real part of the eigenvalues of $\mathbf{J}_F(\omega^*)$ (see Lem. 21 in §C.2) which thus belong to

$$K = \{\lambda \in \mathbb{C} : 0 < \mu \leq \Re\lambda,\ |\lambda| \leq L\}$$

This set is the intersection between a circle and a half-plane, as shown in Figure 2 (left).

Fine-grained bounds. [Nem04] provides a lower-bound for the class of strongly monotone and Lipschitz operators (see §3.2.2) excluding the possibility of acceleration in that general setting. It motivates the adoption of more refined assumptions on the eigenvalues of $\mathbf{J}_F(\omega^*)$. We consider the class of games where these eigenvalues belong to a specified set K. Since $\mathbf{J}_F(\omega^*)$ is real, its spectrum is symmetric w.r.t. the real axis, so we assume that K is too. For this class of problem, we have a simple method to compute lower and upper convergence bounds using a class of well studied shapes: ellipses.

Proposition 2 (Ellipse method for lower and upper bound (Informal))**.** *Let $K \subset \mathbb{C}_+$ be a compact set, then any ellipse symmetric w.r.t. the real axis that includes (resp. is included in) K provides an upper (resp. lower) convergence bound for the class of problem $\mathcal{M}_K$ using Polyak momentum with a step-size and a momentum depending on the ellipse.*

See appendix C.3.2, Thm. 21 for the precise result on ellipses. The proposition extends to any shape whose optimal algorithm (resp. lower bound) is known. This propositionheavily relies on the fact that, the optimal method for ellipses is Polyak momentum [NV83].

Any first-order method can be seen as a way to transform the set K. In order to illustrate that we consider Lemma 3: since a first-order method update for a linear operator $F = Ax + b$ can be written using a polynomial p, the eigenvalues to consider are not the ones of A but the ones of $p(A)$. Thus, the set of interest is $p(K)$.

As an example, consider EG with momentum. This consists in applying the momentum method to the transformed vector field $\omega \mapsto F(\omega - \eta F(\omega))$. From a spectral point of view, this is equivalent to first transforming the shape K into $\varphi(K)$ with the extragradient mapping $\varphi_\eta : \lambda \mapsto \lambda(1 - \eta\lambda)$, then study the effect of momentum on $\varphi(K)$. This example of transformation is used in §3.3.4.

3.2 Asymptotic convergence factor

We recall known results that compute lower bounds for some classes of games using the *asymptotic convergence factor* [EN83, ENV85, Nev93]. Then, we illustrate them on two particular classes of problems.

3.2.1 Lower bounds for a class of problems

We now show how to lower bound the worst-case rate of convergence of a *specific* method over the class $\mathcal{M}_K$ (15), with the worst possible initialisation ω_0. We start with equation (14), but this time we pick the worst-case over all matrices $A \in \mathcal{M}_K$, i.e.,

$$\max_{A \in \mathcal{M}_K} \|p_t(A)(\omega_0 - \omega^*)\|_2.$$

Now, we can pick an arbitrary bad initialisation ω_0, in particular, the one that corresponds to the largest eigenvalue of $p_t(A)$ in magnitude. This gives

$$\begin{aligned} \exists \omega_0 : \|\omega_t - \omega^*\|_2 &\geq \max_{A \in \mathcal{M}_K} \rho\Big(p_t(A)\Big)\|\omega_0 - \omega^*\|_2 \\ &= \max_{\lambda \in K} |p_t(\lambda)| \|\omega_0 - \omega^*\|_2 \,. \end{aligned} \tag{18}$$

It remains to lower bound $\max_{\lambda \in K} |p_t(\lambda)|$ over *all possible* first-order methods. This is called the *asymptotic convergence factor*, presented in the next section.

3.2.2 Asymptotic convergence factor

Here we recall the definition of the *asymptotic convergence factor* [EN83], which gives a lower bound for the rate of convergence over matrices which belong to the class $\mathcal{M}_k$ (15), for all possible first-order methods. We mainly follow the definition of [Nev93] (see Rmk. 4 in §C.2 for details).

The simplest way to lower bound $\|\omega_t - \omega^*\|_2$ is given by minimizing (18) over all polynomials corresponding to a first-order method. By Lemma 3, this class of polynomials is given by $\mathcal{P}_t$. Thus, for some ω_0,

$$\|\omega_t - \omega^*\| \geq \min_{p_t \in \mathcal{P}_t} \max_{\lambda \in K} |p_t(\lambda)| \cdot \|\omega_0 - \omega^*\|_2.$$

The *asymptotic convergence factor* $\underline{\rho}(K)$ for the class K is given by taking the *minimum average* rate of convergence over t for any t, i.e.,

$$\underline{\rho}(K) = \inf_{t>0} \min_{p_t \in \mathcal{P}_t} \max_{\lambda \in K} \sqrt[t]{|p_t(\lambda)|} \,. \tag{19}$$

This way, by construction, $\underline{\rho}(K)$ gives a lower-bound on the *worst-case* rate of convergence for the class $\mathcal{M}_K$. We formalize this statement in the proposition below.

Proposition 3. *[Nev93] Let $K \subset \mathbb{C}$ be a subset of $\mathbb{C}$ symmetric w.r.t. the real axis, which does not contain 0 and such that $\mathcal{M}_K \neq \emptyset$. Then, any oblivious first-order method (whose coefficients only depend on K) satisfies,*

$$\forall t \geq 0,\, \exists A \in \mathcal{M}_K,\, \exists \omega_0 : \|\omega_t - \omega^*\|_2 \geq \underline{\rho}(K)^t \|\omega_0 - \omega^*\|_2.$$

However, the object $\underline{\rho}(K)$ may be complicated to obtain as it depends on the solution of a minimax problem *over a set $K \subset \mathbb{C}_+$*. If the set is simple enough, we can lower-bound the asymptotic rate of convergence. We start by giving the two extreme cases: when K is a segment on the real line (convex and smooth minimization) or K is a disc (monotone and smooth games).

3.2.3 Extreme cases: real segments and discs

Smooth and strongly convex minimization. In the case where we are interested in lower-bounds, we can consider the restricted class of functions where $J_F(\omega)(= \nabla^2 f(\omega))$ is constant, i.e., independent of ω. This corresponds to quadratic minimization, and our restricted class becomes

$$\mathcal{M}_K \quad \text{where} \;\; K = [\mu, L].$$

For this specific class, where K is a segment in the real line, the solution to the subproblem associated to the *asymptotic rate of convergence* (19), i.e.,

$$\min_{p \in \mathcal{P}_t} \max_{\lambda \in [\mu, L]} |p(\lambda)| \tag{20}$$

is well-known. The optimal polynomial p_t^* is a properly scaled and translated Chebyshev polynomial of the first kind of degree t [GV61, Man77]. The rate of convergence of p_t evolves with t, but asymptotically converges to

$$\underline{\rho}([\mu, L]) = \tfrac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\,.$$

This is the lower bound of [Nes04, Thm. 2.1.13], which corresponds to an accelerated linear rate. The condition number L/μ appears as a square root unlike for the rate of the plain gradient descent, which implies a huge (asymptotic) improvement.

In this section, we have seen that when the spectrum is constrained to be on a segment in the real line, one can expect acceleration. The next section shows that this is not the case for the class of discs.

Discs and strongly monotone vector fields Consider a disc with a real positive center

$$K = \{z \in \mathbb{C} : |z - c| \leq r\}, \quad \text{with } 0 < c < r$$

This time again, the shape is simple enough to have an explicit solution for the optimal polynomials

$$p_t^*(\lambda) = \arg\min_{p_t \in \mathcal{P}_t} \max_{\lambda \in K} |p_t(\lambda)|.$$

In this case, the optimal polynomial reads $p_t^*(\omega) = (1 - \omega/c)^t$, and this corresponds to gradient descent with step-size $\eta = 1/c$. Hence, with this specific shape, gradient method is optimal [ENV85,

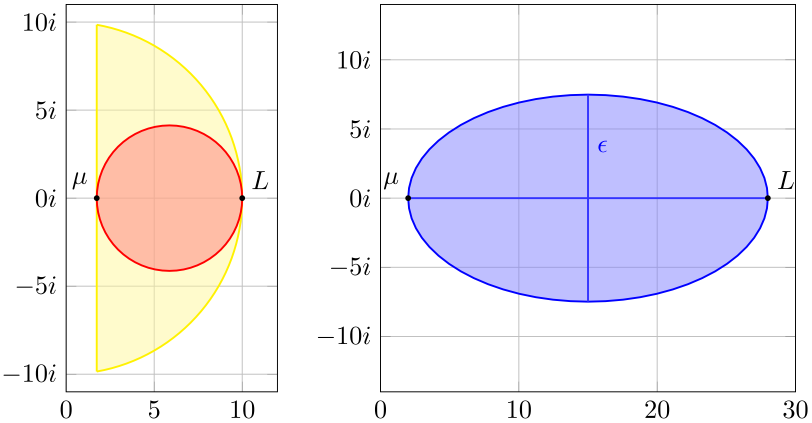

Figure 2. Left: Illustration of the proof of Cor. 3. The yellow set correspond to K, the set of strongly monotone problems while the red disc is the disc of center $\frac{1}{2}(\mu + L)$ and radius $\frac{1}{2}(L - \mu)$ which fits inside. **Right:** Illustration of K_ϵ of Prop. 6 with $\epsilon = \sqrt{\mu L}$.

§6.2]; [Nev93, Example 3.8.2]. A direct consequence of this result is a lower bound of convergence for the class of Lipshitz, strongly monotone vector fields, i.e., vector fields F that satisfies (16)-(17). For linear vector fields parameterized by the matrix A as in Lemma 3, this is included in the set

$$\mathcal{M}_K, \;\; K = \{\lambda \in \mathbb{C} : \; 0 < \mu \leq \Re\lambda, \; |\lambda| \leq L\} \tag{21}$$

This set is the intersection between a circle and a half-plane, as shown in Figure 2 (left). Notice that the disc of center $\frac{\mu+L}{2}$ and radius $\frac{L-\mu}{2}$ actually fits in K, as illustrated by Fig. 2. Since this disc in *included* in K, a lower bound for the disc also gives a lower bound for K, as stated in the following corollary.

Corollary 3. *Let K be defined in* (21)*. Then,*

$$\underline{\rho}(K) > \tfrac{L-\mu}{L+\mu} = 1 - \tfrac{2\mu}{L+\mu}\,.$$

The rate of Cor. 3 is already achieved by first-order methods, without momentum or acceleration, such as EG. Thus, acceleration is *not possible* for the general class of smooth, strongly monotone games.

3.3 Acceleration in games

We present our contributions in this section. The previous section highlights a big contrast between optimization and games. In the former, acceleration is possible, but this does not generalize for the latter. Here, we explore acceleration via a sharp analysis of intermediate cases, like imaginary segments (bilinear games) or thin ellipses (perturbed acceleration), via lower and upper bounds. Since we use spectral arguments, the convergence guarantees of our algorithms are local, but lower bounds remain valid globally.

3.3.1 Local convergence of optimization methods for nonlinear vector fields

Before presenting our result, we recall the classical local convergence theorem from [Pol64]. In this section, we are interested in finding the fixed point ω^* of a vector field V, i.e, $V(\omega^*) = \omega^*$. V here plays the role of an iterative optimization methods and defines iterates according to the fixed-point iteration

$$\omega_{t+1} = V(\omega_t). \tag{22}$$

Analysing the properties of the vector field V is usually challenging, as V can be any nonlinear function. However, under mild assumption, we can simplify the analysis by considering the linearization $V(\omega) \approx V(\omega^*) + \mathbf{J}_V(\omega^*)(\omega - \omega^*)$, where $\mathbf{J}_V(\omega)$ is the Jacobian of V evaluated at ω^*. The next theorem shows we can deduce the rate of convergence of (22) using the spectral radius of $\mathbf{J}_V(\omega^*)$, denoted by $\rho(\mathbf{J}_V(\omega^*))$.

Theorem 9 ([Pol87])**.** *Let $V : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be continuously differentiable and let ω^* one of its fixed-points. Assume that there exists $\rho^* > 0$ such that,*

$$\rho(\mathbf{J}_V(\omega^*)) \leq \rho^* < 1$$

For ω_0 close to ω^*, (22) converges linearly to ω^* at a rate $\mathcal{O}((\rho^* + \epsilon)^t)$. If V is linear, then $\epsilon = 0$.

Recent works such as [MNG17, GHP$^+$19, DP18] used this connection to study game optimization methods.

Thm. 9 can be applied directly on methods which use only the last iterate, such as gradient or EG. For methods that do not fall into this category, such as momentum, a small adjustment is required, called *system augmentation.*

Consider that $V : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ follows the recursion

$$\omega_{t+1} = V(\omega_t, \omega_{t-1}). \tag{23}$$

Instead we consider its *augmented operator*

$$\begin{bmatrix} \omega_t \\ \omega_{t+1} \end{bmatrix} = V_{\text{augm}}(\omega_t, \omega_{t-1}) = \begin{bmatrix} \omega_t \\ V(\omega_t, \omega_{t-1}) \end{bmatrix},$$

to which we can now apply the previous theorem. This technique is summarized in the following lemma.

Lemma 4. *Let $V : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be continuously differentiable and let ω^* satisfies $V(\omega^*, \omega^*) = \omega^*$. Assume there exists $\rho^* > 0$ such that, $\rho(\mathbf{J}_{V_{augm}}(\omega^*)) \leq \rho^* < 1$. If ω_0 and ω_1 are close to ω^*, then (22) converges linearly to ω^* at rate $\big(\rho^* + \epsilon\big)^t$. If V is linear, then $\epsilon = 0$.*

3.3.2 Acceleration for bilinear games

For convex minimization, adding momentum results in an accelerated rate for strongly convex functions we have discuss above. For instance, if $\operatorname{Sp} \nabla F(\omega^*) \subset [\mu, L]$, the Polyak's Heavy-ball method (see the full statement inappendix C.3.1), [Pol64, Thm. 9]

$$\begin{aligned} \omega_{t+1} &= V^{\text{Polyak}}(\omega_t, \omega_{t-1}) \\ &:= \omega_t - \alpha F(\omega_t) + \beta(\omega_t - \omega_{t-1}) \end{aligned} \tag{24}$$

converges (locally) with the accelerated rate

$$\rho(\mathbf{J}_{V^{\text{Polyak}}}(\omega^*, \omega^*)) \leq \tfrac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\,.$$

Another example are bilinear games. Most known methods converge at a rate of $(1 - c\sigma_{min}(A)^2/\sigma_{\max}(A)^2)^t$ for some $c > 0$ [DISZ18, MNG17, GBV$^+$19, GHP$^+$19, LS18, ALW19]. Using results from [ELV89], we show that this rate is suboptimal.

For bilinear games, the eigenvalues of the Jacobian $\mathbf{J}_F$ are purely imaginary (see Lem. 23 inappendix C.3.1), i.e.,

$$K = [i\sigma_{\min}(A), i\sigma_{\max}(A)] \cup [-i\sigma_{\min}(A), -i\sigma_{\max}(A)].$$

A method that follows strictly the vector field F does not converge, as its flow is composed by only concentric circles, thus leading to oscillations. This problem is avoided if we transform the vector field into another one with better properties. For example, the transformation

$$F^{\text{real}}(\omega) = \frac{1}{\eta}(F(\omega - \eta F(\omega)) - F(\omega)) \tag{25}$$

can be seen as a finite-difference approximation of $\nabla\left(\frac{1}{2}\|F\|_2^2\right)$. It is easier to find the equilibrium of V since the eigenvalues of $\mathbf{J}_V(\omega) = -\mathbf{J}_F^2(\omega)$ are located on a real segment. Thus, we can use standard minimization methods like the Polyak Heavy-Ball method.

Proposition 4. *Let F be a vector field such that $\operatorname{Sp}\nabla F(\omega^*) \subset [ia, ib] \cup [-ia, -ib]$, for $0 < a < b$. Setting $\sqrt{\alpha} = \frac{2}{a+b}$, $\sqrt{\beta} = \frac{b-a}{b+a}$, the Polyak Heavy-Ball method (24) on the transformation (25), i.e.,*

$$\omega_{t+1} = \omega_t - \alpha F^{real}(\omega_t) + \beta(\omega_t - \omega_{t-1})\,.$$

converges locally at a linear rate $O\big((1 - \frac{2a}{a+b})^t\big)$.

Using results from [ELV89], we show that this method is optimal. Indeed, for this set, we can compute explicitly $\underline{\rho}(K)$ from (19), the lower bound for the local convergence factor.

Proposition 5. *Let $K = [ia, ib] \cup [-ia, -ib]$ for $0 < a < b$. Then, $\underline{\rho}(K) = \sqrt{\frac{b-a}{b+a}}$.*

Proof. (Sketch). The transformation that we have applied, i.e. $\lambda \mapsto -\lambda^2$, preserves the asymptotic convergence factor $\underline{\rho}$ (up to a square root), as it satisfies the assumptions of [ELV89, Thm. 6]. □

The difference of a square root between the lower bound and the bound on the spectral radius is explained by the fact that the method presented here queries two gradient per iteration and so one of its iterations actually corresponds to two steps of a first-order method as defined in Definition 1.

In this subsection, we showed that when the eigenvalues of the Jacobian are purely real or imaginary, acceleration is possible using momentum on the right vector field. Yet the previous subsection shows it is not the case for general smooth, strongly monotone games. The question of acceleration remains for intermediate shapes, like ellipses. The next subsection shows how to recover an accelerated rate of convergence in this case.

3.3.3 Perturbed acceleration

As we cannot compute $\underline{\rho}$ explicitly for most sets K, we focus on ellipses to answer this question. They have been well studied, and optimal methods are again based on Chebyshev polynomials [Man77].

In this section we study games whose eigenvalues of the Jacobian belong to a thin ellipse. These ellipses correspond to the real segments $[\mu, L]$ perturbed in an elliptic way, see Fig. 2 (right). Mathematically, we have for $0 < \mu < L$ and $\epsilon > 0$, the equation

$$K_\epsilon = \left\{ z \in \mathbb{C} : \left(\frac{\Re z - \frac{\mu+L}{2}}{\frac{L-\mu}{2}} \right)^2 + \left(\frac{\Im z}{\epsilon} \right)^2 \leq 1 \right\}$$

When $\epsilon = 0$ (with the convention that $0/0 = 0$), Polyak momentum achieves the rate of $1 - 2\frac{\sqrt{\mu}}{\sqrt{\mu}+\sqrt{L}}$. However, when $\epsilon = \frac{L-\mu}{2}$, we showed the lower bound of $1 - 2\frac{\mu}{\mu+L}$ in Cor. 3. To check if acceleration still persists for intermediate cases, we study the behaviour of the asymptotic convergence factor (when $L/\mu \to +\infty$) as a function of ϵ. The next proposition uses results from [NV83, ENV85] to show that acceleration is still possible on K_ϵ.

Proposition 6. *Define $\epsilon(\mu, L)$ as $\frac{\epsilon(\mu,L)}{L} = \left(\frac{\mu}{L}\right)^\theta$ with $\theta > 0$ and $a \wedge b = \min(a, b)$. Then, when $\frac{\mu}{L} \to 0$,*

$$\underline{\rho}(K_\epsilon) = \begin{cases} 1 - 2\sqrt{\frac{\mu}{L}} + \mathcal{O}\left(\left(\frac{\mu}{L}\right)^{\theta \wedge 1} \right), & \text{if } \theta > \frac{1}{2} \\ 1 - 2(\sqrt{2}-1)\sqrt{\frac{\mu}{L}} + \mathcal{O}\left(\frac{\mu}{L}\right), & \text{if } \theta = \frac{1}{2} \\ 1 - \left(\frac{\mu}{L}\right)^{1-\theta} + \mathcal{O}\left(\left(\frac{\mu}{L}\right)^{1 \wedge (2-3\theta)} \right), & \text{if } \theta < \frac{1}{2}. \end{cases}$$

Moreover, the momentum method is optimal for K_ϵ. This means there exists $\alpha > 0$ and $\beta > 0$ (function of μ, L and ϵ only) such that if $\operatorname{Sp} \boldsymbol{J}_F(\omega^*) \subset K_\epsilon$, then, $\rho(\boldsymbol{J}_{V^{Polyak}}(\omega^*, \omega^*)) \leq \underline{\rho}(K_\epsilon)$.

This shows that the convergence rate interpolates continuously between the accelerated rate and the non-accelerated one. Crucially, for small perturbations, that is to say if the ellipse is thin enough, acceleration persists until $\theta = \frac{1}{2}$ or equivalently $\epsilon \sim \sqrt{\mu L}$. That's why Prop. 6 plays a central role in our forthcoming analyses of accelerated EG and CO.

3.3.4 Accelerating extragradient

We now consider the acceleration of EG using momentum. Its main appealing property is its convergence on bilinear games, unlike the gradient method. On the class of bilinear problems, EG achieves a convergence rate of $(1 - ca^2/b^2)$ for some constant $c > 0$.

In the previous section, we achieved an accelerated rate on bilinear games by applying momentum to the operator $F^{\text{real}}(\omega)$ instead of F, as the Jacobian of F^{real} has real eigenvalues when $\mathbf{J}_F(\omega^*)$ has its spectrum in K. Here we try to apply momentum to the EG operator $F^{\text{e-g}}(\omega)$, defined as

$$F^{\text{e-g}}(\omega) = F(\omega - \eta F(\omega)) \,. \tag{26}$$

Unfortunately, when $\operatorname{Sp} \mathbf{J}_F \subset K$, the spectrum of $F^{\text{e-g}}(\omega^*)$ is never purely real. Using the insight from Prop. 6, we can choose $\eta > 0$ such that we are in the first case of Prop. 6, making acceleration possible.

Proposition 7. *Consider the vector field F, where $\operatorname{Sp} \boldsymbol{J}_F(\omega^*) \subset [ia, ib] \cup [-ia, -ib]$ for $0 < a < b$. There exists $\alpha, \beta, \eta > 0$ such that, the operator defined by*

$$\omega_{t+1} = \omega_t - \alpha F(\omega_t - \eta F(\omega_t)) + \beta(\omega_t - \omega_{t-1}) \,,$$

converges locally at a linear rate $O\big(\big(1 - c\frac{a}{b} + M\frac{a^2}{b^2}\big)^t\big)$ where $c = \sqrt{2} - 1$ and M is an absolute constant.

One drawback is that, to achieve fast convergence on bilinear games, one has to tune the two step-sizes α, η of EG precisely and separately. They actually differ by a factor $\frac{b^2}{a^2}$: η is roughly proportional to $\frac{1}{a}$ while α behaves like $\frac{a}{b^2}$ (see Lem. 25 in appendix C.3.4).

3.4 Beyond typical first-order methods

In the previous section, we achieved acceleration with first-order methods for specific problem classes. However, the lower bound from Cor. 3 still prevents us from doing so for the larger problem classes for smooth and strongly monotone games. To bypass this limitation, we can consider going *beyond* first-order methods. In this section, we consider two different approaches. The first one is adaptive acceleration, which is a *non-oblivious* first-order method. The second is consensus optimization, an inversion-free second order method.

3.4.1 Adaptive acceleration

In previous sections, we considered shapes whose optimal polynomial is known. This optimal polynomial lead to an optimal first-order method. However, when the shape is *unknown*, we cannot use better methods than EG with an appropriate stepsize.

Recent work in optimization analysed adaptive algorithms, such as *Anderson Acceleration* [WN11], that are adaptive to the problem constants. They can be seen as an automatic way to find the optimal combination of the previous iterates. Recent works on Anderson Acceleration extended the theory for non-quadratic minimization, by using regularisation [SdB16] (RNA method). The theory has also been extended to "non symmetric operators" [BSd18], and this setting fits perfectly the one of games, as $\mathbf{J}_F(\omega^*)$ is not symmetric.

Anderson acceleration and its extension RNA are similar to quasi-Newton [FS09], but remains first-order methods. Even if they find the optimal first-order method (for linear F), they cannot beat a lower bound similar to Cor. 3, when the number of iterations is smaller than the dimension of the problem. The next section shows how to use *cheap* second-order information to improve the convergence rate.

3.4.2 Momentum consensus optimization

CO [MNG17] iterates as follow:

$$\omega_{t+1} = \omega_t - \alpha\big(F(\omega_t) + \tau \mathbf{J}_F^T(\omega)F(\omega)\big).$$

Albeit being a second-order method, each iteration requires only one Jacobian-vector multiplication. This operation can be computed efficiently by modern machine learning frameworks, with automatic differentiation and back-propagation. For instance, for neural networks, the computation time of this product or the gradient is comparable. Moreover, unlike Newton's method, CO does *not* require a matrix inversion.

Though CO is a second-order method, its analysis can still be reduced to our framework by considering the following transformation of the initial operator $F(\omega)$,

$$F^{\text{cons.}}(\omega) = F(\omega) + \tau\nabla\left(\frac{1}{2}\|F\|^2\right)(\omega)\,. \tag{27}$$

Though the eigenvalues of $\mathbf{J}_{F^{\text{cons.}}}$ are not purely real in general, their imaginary to real part ratio can be controlled by [MNG17, Lem. 9] as,

$$\max_{\lambda\in \operatorname{Sp}\mathbf{J}_{F^{\text{cons.}}}(\omega^*)} \tfrac{|\Im\lambda|}{|\Re\lambda|} = O\left(\tfrac{1}{\tau}\right)\,.$$

Therefore, if τ increases, these eigenvalues move closer to the real axis and can be included in a thin ellipse as described by §3.3.3. We then show that, if τ is large enough, this ellipse can be chosen thin enough to fall into the accelerated regime of Prop. 6 and therefore, adding momentum achieves acceleration.

Proposition 8. *Let σ_i be the singular values of $\boldsymbol{J}_F(\omega^*)$. Assume that*

$$\gamma \leq \sigma_i \leq L, \quad \tau = \tfrac{L}{\gamma^2}.$$

There exists α, β, s.t., momentum applied to $F^{cons.}$,

$$\omega_{t+1} = \omega_t - \alpha F^{cons.}(\omega_t) + \beta(\omega_t - \omega_{t-1})$$

converges locally at a rate $O\Big(\big(1 - c\frac{\gamma}{L} + M\frac{\gamma^2}{L^2}\big)^t\Big)$ where $c = \sqrt{2} - 1$ and M is an absolute constant.

Hence, adding momentum to CO yields an accelerated rate. The assumption on the Jacobian encompasses both strongly monotone and bilinear games. On these two classes of problems, CO is at least as fast as any oblivious first-order method as its rate roughly matches the lower bounds of Prop. 3 and 5.

Note that, choosing τ of this order is what is done by [ALW19] for (non-accelerated) CO. They claim that this point of view – seeing consensus as a perturbation of gradient descent on $\frac{1}{2}\|F\|^2$ – is justified by practice as in the experiments of [MNG17], τ is set to 10.

4 Conclusion

In this paper, we presented a comprehensive spectral analysis of gradient-based methods in multiplayer differentiable games, covering a wide range of game types, from bilinear to strongly monotone structures. By leveraging matrix iteration theory, we introduced the concept of spectral shape, linking game difficulty to the geometric distribution of eigenvalues in the Jacobian. This perspective allowed us to propose an optimal algorithm for bilinear games and derive new lower bounds, demonstrating the near-optimality of the extragradient (EG) method among first-order algorithms.

Our analysis unified the behavior of EG, optimistic gradient (OG), and consensus optimization (CO), providing global convergence guarantees and highlighting that these methods leverage different convergence mechanisms depending on the game's structure. Notably, EG outperforms gradient descent (GD) in adversarial settings, while OG and CO achieve similar improvements in specific regimes.

We also explored the potential for acceleration in game dynamics, drawing parallels to Polyak and Nesterov momentum methods in convex optimization. While the possibility of acceleration in adversarial games remains an open question, we proposed momentum-based enhancements to consensus optimization, achieving near-accelerated rates.

Future work will explore extending these findings to stochastic and high-dimensional settings, as well as refining the spectral analysis to handle more intricate eigenvalue distributions, especially in real-world applications like generative adversarial networks (GANs). These directions hold promise for advancing the performance of optimization methods in complex machine learning environments.

References

[ADLH18] Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local Saddle Point Optimization: A Curvature Exploitation Approach. *arXiv:1805.05751 [cs, math, stat]*, 2018.

[ALW19] Jacob Abernethy, Kevin A. Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *arXiv*, 2019.

[AMLJG19] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of extragradient for a whole spectrum of differentiable games. *arXiv*, 2019.

[AS16] Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *ICML*, 2016.

[ASS16] Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On Lower and Upper Bounds in Smooth and Strongly Convex Optimization. *JMLR*, 2016.

[Atk89] Kendall E. Atkinson. *An Introduction to Numerical Analysis.* Wiley, 1989.

[Ber99] Dimitri P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[BN10] Joseph Bak and Donald J. Newman. *Complex Analysis.* Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2010.

[BRM$^+$18] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The Mechanics of n-Player Differentiable Games. In *ICML*, 2018.

[BSd18] Raghu Bollapragada, Damien Scieur, and Alexandre d'Aspremont. Nonlinear acceleration of momentum and primal-dual algorithms. *arXiv*, 2018.

[BT04] Jean-Paul Berrut and Lloyd N. Trefethen. Barycentric Lagrange Interpolation. *SIAM Review*, 46, 2004.

[BV04] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[Chi11] Theodore S Chihara. *An introduction to orthogonal polynomials.* Courier Corporation, 2011.

[CLO14] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Accelerated Schemes For A Class of Variational Inequalities. *arXiv*, 2014.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 2011.

[DISZ18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. In *ICLR*, 2018.

[DP18] Constantinos Daskalakis and Ioannis Panageas. The Limit Points of (Optimistic) Gradient Descent in Min-Max Optimization. *arXiv*, 2018.

[ELV89] Michael Eiermann, X Li, and Richard Varga. On Hybrid Semi-Iterative Methods. *Siam Journal on Numerical Analysis*, 1989.

[EN83] Michael Eiermann and Wilhelm Niethammer. On the Construction of Semiterative Methods. *SIAM Journal on Numerical Analysis*, 1983.

[ENV85] M. Eiermann, W. Niethammer, and R. S. Varga. A study of semiiterative methods for nonsymmetric systems of linear equations. *Numerische Mathematik*, 1985.

[FP03] Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems Vol I.* Springer Series in Operations Research and Financial Engineering, Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer-Verlag, 2003.

[FS09] Haw-ren Fang and Yousef Saad. Two classes of multisecant methods for nonlinear acceleration. *Numerical Linear Algebra with Applications*, 2009.

[GBV$^+$19] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In *ICLR*, 2019.

[GHP$^+$19] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Remi Lepriol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative Momentum for Improved Game Dynamics. In *AISTATS*, 2019.

[Goo16] Ian Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv:1701.00160 [cs]*, 2016.

[GPM$^+$14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NeurIPS*. 2014.

[GV61] Gene H. Golub and Richard S. Varga. Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods. *Numerische Mathematik*, 1961.

[Had06] Jacques Hadamard. Sur les transformations ponctuelles. *Bull. Soc. Math. France*, 34:71–84, 1906.

[HP90] Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 1990.

[IAGM19] Adam Ibrahim, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. *arXiv preprint arXiv:1906.07300*, 2019.

[Kor76] G.M. Korpelevich. The extragradient method for finding saddle points and other problems., 1976.

[Lax07] Peter D. Lax. *Linear Algebra and Its Applications*. John Wiley & Sons, 2007.

[LBR$^+$19] Alistair Letcher, David Balduzzi, Sébastien Racanière, James Martens, Jakob N. Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *JMLR.*, 2019.

[Lev20] M. P. Levy. Sur les fonctions de lignes implicites. *Bull. Soc. Math. France*, 48:13–27, 1920.

[LMH15] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *NeurIPS*, 2015.

[LS18] Tengyuan Liang and James Stokes. Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks. In *AISTATS*, 2018.

[Man77] Thomas A. Manteuffel. The Tchebychev iteration for nonsymmetric linear systems. *Numerische Mathematik*, 1977.

[MGN18] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs do actually Converge? In *ICML*, 2018.

[Min62] George J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29:341–346, 1962.

[MJS19] 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:1901.00838 [cs, math, stat]*, 2019.

[MLZ$^+$19] 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 *ICLR*, 2019.

[MNG17] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of GANs. In *NeurIPS*, 2017.

[MOP19] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. *arXiv*, 2019.

[Neh52] Zeev Nehari. *Conformal Mapping.* McGraw-Hill, 1952.

[Nem92] A.S Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8:153–175, 1992.

[Nem04] 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*, 2004.

[Nes83] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 1983.

[Nes03] Yurii Nesterov. Dual Extrapolation and its Applications for Solving Variational Inequalities and Related Problems'. SSRN Scholarly Paper, Social Science Research Network, 2003.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Springer, 2004.

[Nev93] Olavi Nevanlinna. *Convergence of Iterations for Linear Equations.* Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 1993.

[NS06] Yurii Nesterov and Laura Scrimali. Solving Strongly Monotone Variational and Quasi-Variational Inequalities. SSRN Scholarly Paper, Social Science Research Network, 2006.

[NV83] W. Niethammer and R. S. Varga. The analysis ofk-step iterative methods for linear systems from summability theory. *Numerische Mathematik*, 1983.

[NY83] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[OS84] Gerhard Opfer and Glenn Schober. Richardson's iteration for nonsymmetric matrices. *Linear Algebra and its Applications*, 58:343–361, 1984.

[OX18] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *arXiv:1808.02901 [math]*, 2018.

[PB16] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *NeurIPS*, 2016.

[Pol64] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 1964.

[Pol87] Boris T Polyak. *Introduction to Optimization.* Optimization Software, 1987.

[PV16] David Pfau and Oriol Vinyals. Connecting Generative Adversarial Networks and Actor-Critic Methods. *arXiv*, 2016.

[Ran95] Thomas Ransford. *Potential Theory in the Complex Plane.* Cambridge University Press, 1995.

[Rhe69] Werner C. Rheinboldt. Local mapping relations and global implicit function theorems. 1969.

[Roc76] R. Tyrrell Rockafellar. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis.* McGraw-Hill, 1976.

[Sal71] H E Salzer. Lagrangian interpolation at the Chebyshev points xn, v = cos (VTT/AZ), V = 0(1)//; some unnoted advantages. 1971.

[Sau64] V. K. Saul'yev. *Integration of Equations of Parabolic Type by the Method of Nets.* Elsevier, 1964.

[SdB16] Damien Scieur, Alexandre d'Aspremont, and Francis Bach. Regularized nonlinear acceleration. In *NeurIPS*, 2016.

[SKSF18] Sainbayar Sukhbaatar, Ilya Kostrikov, Arthur Szlam, and Rob Fergus. Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play, 2018.

[SMDH13] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

[Tse95] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 1995.

[VNM44] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. 1944.

[Wal35] J. L. Walsh. *Interpolation and Approximation by Rational Functions in the Complex Domain*. American Mathematical Soc., 1935.

[WN11] Homer F Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 2011.

[Zha05] Fuzhen Zhang, editor. *The Schur Complement and Its Applications*. Numerical Methods and Algorithms. Springer-Verlag, 2005.