Accelerating Smooth Games by Manipulating Spectral Shapes

Waïss Azizian[1,†] **Damien Scieur**[2] **Ioannis Mitliagkas**[3,‡] **Simon Lacoste-Julien**[2,3,‡] **Gauthier Gidel**[3]
[1]École Normale Supérieure, Paris [2]Mila & SAIT AI Lab, Montreal [3]Mila & DIRO, Université de Montréal

Abstract

We use matrix iteration theory to characterize acceleration in smooth games. We define the *spectral shape* of a family of games as the set containing all eigenvalues of the Jacobians of standard gradient dynamics in the family. Shapes restricted to the real line represent well-understood classes of problems, like minimization. Shapes spanning the complex plane capture the added numerical challenges in solving smooth games. In this framework, we describe gradient-based methods, such as extragradient, as transformations on the spectral shape. Using this perspective, we propose an optimal algorithm for bilinear games. For smooth and strongly monotone operators, we identify a continuum between convex minimization, where acceleration is possible using Polyak's momentum, and the worst case where gradient descent is optimal. Finally, going beyond first-order methods, we propose an accelerated version of consensus optimization.

1 Introduction

Recent successes of multi-agent formulations in various areas of deep learning (Goodfellow et al., 2014; Pfau and Vinyals, 2016) have caused a surge of interest in the theoretical understanding of first-order methods for the solution of differentiable multi-player games (Palaniappan and Bach, 2016; Gidel et al., 2019a; Balduzzi et al., 2018; Mescheder et al., 2017, 2018; Mazumdar et al., 2019). This exploration hinges on a key question:

How fast can a first-order method be?

In convex minimization, Nesterov (1983, 2004) 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 (Mescheder et al., 2017; Balduzzi et al., 2018). Some old papers (Korpelevich, 1976; Tseng, 1995) and many recent ones (Nemirovski, 2004; Chen et al., 2014; Palaniappan and Bach, 2016; Mescheder et al., 2017; Gidel et al., 2019a,b; Daskalakis et al., 2018; Mokhtari et al., 2019; Azizian et al., 2019) suggest new methods and provide better upper bounds.

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

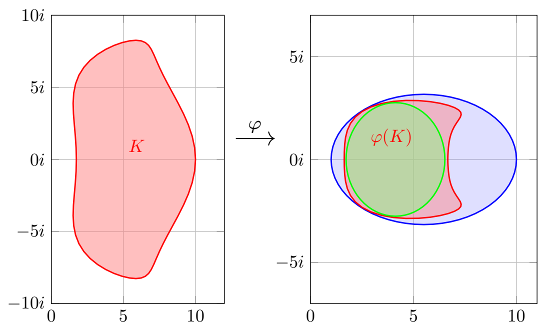

Figure 1: Transformation of the spectral shape K (in red from left to right) by the extragradient operator $\varphi : \lambda \mapsto \lambda(1 - \eta\lambda)$. Any ellipse (e.g. in blue) that contains the transformed red shape $\varphi(K)$ provides a upper convergence bound using extragradient with Polyak momentum (with step-size and momentum that depends on the ellipse parameters). Any ellipse included in it (e.g. in green) provides a lower bound. See §3.3.

[†]Work done during an internship at Mila.
[‡]Canada CIFAR AI Chair.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

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 paper, 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. §3.2). 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.

Contributions. Our main contribution is a geometric interpretation of the conditioning of a game (via its *spectral shape* as illustrated in Fig. 1, and discussed with more details in §3.3). Our result links the "hardness" of a game to the distribution of the eigenvalues of its Jacobian of the game at the optimum. Using our framework, we make the following contributions.

1. We show a reduction from bilinear games to games with *real* eigenvalues, where acceleration is possible through momentum. We provide lower bounds and design an optimal algorithm for this class.

2. Showing that acceleration persists even if there is an "imaginary perturbation", we propose an accelerated version of extragradient (EG) for bilinear games.

3. We accelerate *consensus optimization* (CO), a *cheap* second-order method. We combine it with momentum to achieve a nearly-accelerated rate, improving the best rate previously known for this method.

Organisation. We recall the definition of the *asymptotic convergence factor* in §4 and use it to show that acceleration is *not possible* for the general class of smooth and strongly monotone games. In §5 we show that bilinear games or games with a "small imaginary perturbation" can be accelerated. Finally, in §6 we improve the rate of CO by using momentum.

2 Related work

Matrix iteration theory. There is extensive literature on iterative methods for linear systems, due to their countless applications. An important line of work considers the design of iterative methods through the lens of approximation problems by polynomials on the complex plane. Eiermann and Niethammer (1983) then 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. Niethammer and Varga (1983) bring tools from summability theory to analyze multi-step iterative methods in this framework and provide optimal methods, in particular, the momentum method for ellipses. Eiermann et al. (1985) continued in this direction, summarizing and improving the previous results. Finally Eiermann et al. (1989) study how polynomial transformations of the spectrum help compute the asymptotic convergence factor and the optimal method for a given set, potentially yielding faster convergence.

Acceleration and lower bounds. Lower bounds of convergence are standard in convex optimization (Nesterov, 2004) but are often non-asymptotic or cast in an infinite-dimensional space. Arjevani et al. (2016); Arjevani and Shamir (2016) showed non-asymptotic lower bounds using a framework called p-SCLI close to matrix iteration theory. Ibrahim et al. (2019); Azizian et al. (2019) extended this framework to multi-player games, but they consider lower and upper-bounds on the eigenvalues of the Jacobian of the game rather than their distribution in the complex plane. Two main acceleration methods in convex optimization achieve these lower bounds, Polyak's momentum (Polyak, 1964) and Nesterov's acceleration Nesterov (1983). The latter is the only one that has global convergence guarantees for convex functions. Nevertheless, Polyak's momentum still plays a crucial role in the training of large scale machine learning models Sutskever et al. (2013).

Acceleration for games. Recent work applied acceleration techniques to game optimization. Gidel et al. (2019b) showed that negative momentum with alternating updates converges on bilinear games, but with the same geometrical rate as EG. Chen et al. (2014) provided a version of the mirror-prox method which improves the constant but not its rate. In the context of minimax optimization, Palaniappan and Bach (2016) used Catalyst (Lin et al., 2015), a generic acceleration method, to improve the convergence of variance-reduced algorithms for min-max problems. In the context of variational inequalities, the standard assumptions on the operator are Lipschitzness and (strong) monotonicity (Tseng, 1995; Nesterov, 2003). Nemirovski (2004) provided a lower bound in $\mathcal{O}(1/t)$ on the convergence rate for smooth monotone games, which suggests that EG is nearly optimal in the strongly monotone case. In our work, we show that acceleration is possible by substituting the smoothness and monotonicity assumptions on the operator into more precise assumptions on the *eigenvalues of its Jacobian*.

3 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 (Harker and Pang, 1990). A relevant special case is a n-player convex game, where ω^* corresponds to a Nash equilibrium (Von Neumann and Morgenstern, 1944; Balduzzi et al., 2018). 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

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

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 First-order methods

To study lower bounds of convergence, we need a class of algorithms. We consider the classic definition[1] of first-order methods from Nemirovsky and Yudin (1983).

Definition 1. A *first-order method* generates

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

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 (Duchi et al., 2011), 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.

Proposition 1. *(Arjevani and Shamir, 2016) first-order methods can be written as*

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

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 $\alpha_t^{(t)} = -\alpha$, $\beta_t^{(t)} = 1 + \beta$, $\beta_{t-1}^{(t)} = -\beta$.

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

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 1. *(e.g. Chihara, 2011) If* $F(\omega) = A\omega + b$,

$$\omega_t - \omega^* = p_t(A)(\omega_0 - \omega^*), \tag{2}$$

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.

3.2 Problem class

In the previous section, when F is the linear function $F = Ax + b$, the iterates ω_t follow the relation (2) involving the polynomial p_t. Since all first-order methods can be written using polynomials (1), they follow

$$\|\omega_t - \omega^*\|_2 = \|p_t(A)(\omega_0 - \omega^*)\|_2. \tag{3}$$

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,

$$\mathcal{M}_K := \{A \in \mathbb{R}^d : \operatorname{Sp}(A) \subset K \subset \mathbb{C}_+\}, \tag{4}$$

where $\operatorname{Sp}(A)$ 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.

3.3 Geometric intuition

Our paper 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

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

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. 7) to show that $\operatorname{Sp} \mathbf{J}_F(\omega) \in \pm[i\sigma_{\min}(A), i\sigma_{\max}(A)]$.

Variational inequalities. The Lipchitz assumption

$$\|F(\omega) - F(\omega')\|_2^2 \le L\|\omega - \omega'\|_2^2 \tag{5}$$

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

$$(\omega - \omega')^T(F(\omega) - F(\omega')) \ge \mu\|\omega - \omega'\|_2^2 \tag{6}$$

implies a lower bound on the real part of the eigenvalues of $\mathbf{J}_F(\omega^*)$ (see Lem. 5 in §B) which thus belong to

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

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

Fine-grained bounds. Nemirovski (2004) provides a lower-bound for the class of strongly monotone and Lipschitz operators (see §4.2) 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.2, Thm. 6 for the precise result on ellipses. The proposition extends to any shape whose optimal algorithm (resp. lower bound) is known. This proposition, illustrated in Fig. 1, heavily relies on the fact that, the optimal method for ellipses is Polyak momentum (Niethammer and Varga, 1983).

Any first-order method can be seen as a way to transform the set K. In order to illustrate that we consider Lemma 1: 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 illustrated in Fig. 1, and this idea is used in §5.4.

4 Asymptotic convergence factor

We recall known results that compute lower bounds for some classes of games using the *asymptotic convergence factor* (Eiermann and Niethammer, 1983; Eiermann et al., 1985; Nevanlinna, 1993). Then, we illustrate them on two particular classes of problems.

4.1 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 $\mathcal{M}_K$ (4), with the worst possible initialisation ω_0. We start with equation (3), 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

$$\begin{aligned}\exists\omega_0 : \|\omega_t - \omega^*\|_2 &\ge \max_{A\in\mathcal{M}_K} \rho\Big(p_t(A)\Big)\|\omega_0 - \omega^*\|_2 \\ &= \max_{\lambda\in K} |p_t(\lambda)|\|\omega_0 - \omega^*\|_2\,.\end{aligned} \tag{7}$$

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.

4.2 Asymptotic convergence factor

Here we recall the definition of the *asymptotic convergence factor* (Eiermann and Niethammer, 1983), which gives a lower bound for the rate of convergence over matrices which belong to the class $\mathcal{M}_k$ (4), for all possible first-order methods. We mainly follow the definition of Nevanlinna (1993) (see Rmk. 1 in §B for details).

The simplest way to lower bound $\|\omega_t - \omega^*\|_2$ is given by minimizing (7) over all polynomials corresponding to a first-order method. By Lemma 1, 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.,

$$\underline{\rho}(K) = \inf_{t>0} \min_{p_t \in \mathcal{P}_t} \max_{\lambda \in K} \sqrt[t]{|p_t(\lambda)|}\,. \tag{8}$$

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. *(Nevanlinna, 1993) 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).

4.3 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

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

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* (8), i.e.,

$$\min_{p \in \mathcal{P}_t} \max_{\lambda \in [\mu, L]} |p(\lambda)| \tag{9}$$

is well-known. The optimal polynomial p_t^* is a properly scaled and translated Chebyshev polynomial of the first kind of degree t (Golub and Varga, 1961; Manteuffel, 1977). 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 Nesterov (2004, Thm. 2.1.13), which corresponds to an accelerated linear rate. The condition number L/μ appears as a square root unlike

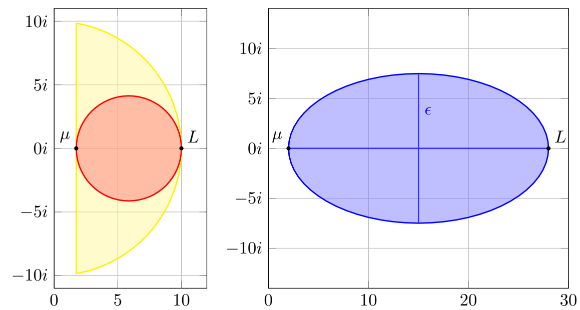

Figure 2: **Left:** Illustration of the proof of Cor. 1. The yellow set correspond to K, the set of strongly monotone problems while the red disc is the disc of center $\frac{1}{2}(\mu + L)$ and radius $\frac{1}{2}(L - \mu)$ which fits inside. **Right:** Illustration of K_ϵ of Prop. 6 with $\epsilon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

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 (Eiermann et al., 1985, §6.2); Nevanlinna (1993, 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 (5)-(6). For linear vector fields parameterized by the matrix A as in Lemma 1, this is included in the set

$$\mathcal{M}_K,\ K = \{\lambda \in \mathbb{C} :\ 0 < \mu \leq \Re\lambda,\ |\lambda| \leq L\}. \tag{10}$$

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 1. *Let K be defined in* (10)*. Then,*

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

The rate of Cor. 1 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.

5 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.

5.1 Local convergence of optimization methods for nonlinear vector fields

Before presenting our result, we recall the classical local convergence theorem from Polyak (1964). 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

$$\omega_{t+1} = V(\omega_t). \tag{11}$$

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 (11) using the spectral radius of $\mathbf{J}_V(\omega^*)$, denoted by $\rho(\mathbf{J}_V(\omega^*))$.

Theorem 1 (Polyak (1987))**.** *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^*, (11) converges linearly to ω^* at a rate $\mathcal{O}((\rho^* + \epsilon)^t)$. If V is linear, then $\epsilon = 0$.

Recent works such as Mescheder et al. (2017); Gidel et al. (2019b); Daskalakis and Panageas (2018) used this connection to study game optimization methods.

Thm. 1 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

$$\omega_{t+1} = V(\omega_t, \omega_{t-1}). \tag{12}$$

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 2. *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 (11) converges linearly to ω^* at rate $\left(\rho^* + \epsilon\right)^t$. If V is linear, then $\epsilon = 0$.*

5.2 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.1), Polyak (1964, Thm. 9)

$$\begin{aligned} \omega_{t+1} &= V^{\text{Polyak}}(\omega_t, \omega_{t-1}) \\ &:= \omega_t - \alpha F(\omega_t) + \beta(\omega_t - \omega_{t-1}) \end{aligned} \tag{13}$$

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 $c > 0$ (Daskalakis et al., 2018; Mescheder et al., 2017; Gidel et al., 2019a,b; Liang and Stokes, 2018; Abernethy et al., 2019). Using results from Eiermann et al. (1989), we show that this rate is suboptimal.

For bilinear games, the eigenvalues of the Jacobian $\mathbf{J}_F$ are purely imaginary (see Lem. 7 inAppendix C.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

$$F^{\text{real}}(\omega) = \tfrac{1}{\eta}(F(\omega - \eta F(\omega)) - F(\omega)) \tag{14}$$

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 (13) on the transformation (14), 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 Eiermann et al. (1989), we show that this method is optimal. Indeed, for this set, we can compute explicitly $\underline{\rho}(K)$ from (8), 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 Eiermann et al. (1989, 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.

5.3 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 (Manteuffel, 1977).

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. 1. 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 Niethammer and Varga (1983); Eiermann et al. (1985) 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 $\frac{\mu}{L} \to 0$,*

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

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.

5.4 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

$$F^{\text{e-g}}(\omega) = F(\omega - \eta F(\omega)) \,. \tag{15}$$

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. 9 in Appendix C.4).

6 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. 1 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.

6.1 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* (Walker and Ni, 2011), 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 (Scieur et al., 2016) (RNA method). The theory has also been extended to "non symmetric operators" (Bollapragada et al., 2018), 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 (Fang and Saad, 2009), 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. 1, 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.

6.2 Momentum consensus optimization

CO (Mescheder et al., 2017) 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 $F(\omega)$,

$$F^{\text{cons.}}(\omega) = F(\omega) + \tau \nabla \left(\tfrac{1}{2}\|F\|^2\right)(\omega)\,. \qquad (16)$$

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 Mescheder et al. (2017, Lem. 9) as,

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

Therefore, if τ increases, these eigenvalues move closer to the real axis and can be included in a thin ellipse as described by §5.3. 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*

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

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 Abernethy et al. (2019) 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 Mescheder et al. (2017), τ is set to 10.

7 Conclusion

This paper shows that a spectral perspective is fundamental to understand the conditioning of games. The latter is indeed linked to the geometric properties of the distribution of the spectrum of its Jacobian. In the light of this perspective, we demonstrate how several gradient-based methods transform the spectral shape of a game to achieve accelerated convergence when combined with Polyak momentum. Our main tool throughout this paper was the flexible and convenient class of ellipses; we left as future work the study of more intricate shapes, which – ideally – would fit the distribution of the eigenvalues of applications of challenging machine learning problems such as GANs.

Acknowledgments

This research was partially supported by the Canada CIFAR AI Chair Program, the Canada Excellence Research Chair in "Data Science for Realtime Decision-making", by the NSERC Discovery Grants RGPIN-2017-06936 and RGPIN-2019-06512, the FRQNT new researcher program 2019-NC-257943, by a Borealis AI fellowship, a Google Focused Research award and a startup grant by IVADO. Simon Lacoste-Julien is a CIFAR Associate Fellow in the Learning in Machines & Brains program.

Bibliography

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

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

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

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.

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

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The Mechanics of n-Player Differentiable Games. In *ICML*, 2018.

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

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

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

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

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

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

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

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

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

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.

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

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In *ICLR*, 2019a.

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*, 2019b.

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

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.

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.

Adam Ibrahim, Waïss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Lower Bounds and Conditioning of Differentiable Games. *arXiv*, 2019.

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

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

Alistair Letcher, David Balduzzi, Sébastien Racanière, James Martens, Jakob N. Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *JMLR.*, 2019.

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

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

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

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.

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

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

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.

Zeev Nehari. *Conformal Mapping*. McGraw-Hill, 1952.

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.

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

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

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

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

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

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

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

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

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

Boris T Polyak. *Introduction to Optimization*. Optimization Software, 1987.

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

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

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

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

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

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

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

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

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

A Linear algebra results

Theorem 2 (Spectral Mapping Theorem)**.** *Let $A \in \mathbb{C}^{d\times d}$ and P be a polynomial. Then,*

$$\operatorname{Sp} P(A) = \{P(\lambda) \mid \lambda \in \operatorname{Sp} A\}\,. \tag{17}$$

See for instance Lax (2007, Theorem 4, p. 66) for a proof.

B Proofs of general lemmas

Proposition 1. *(Arjevani and Shamir, 2016) first-order methods can be written as*

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

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.

Proof. The fact that any first-order method as defined by Definition 1 satisfies such relations is immediate. The converse can be shown by induction. Assume that $(\omega_t)_t$ are generated by the rule of Prop. 1. For $t = 0$, the condition of Definition 1 is trivial. Assume that for all $k \leq t$, $w_k \in \omega_0 + \mathbf{Span}\{F(\omega_0), \dots, F(\omega_{k-1})\}$. Then,

$$\begin{aligned}
\omega_{t+1} &= \sum_{k=0}^{t} \alpha_k^{(t)} F(\omega_k) + \beta_k^{(t)}\omega_k \\
&= \omega_0 + \sum_{k=0}^{t} \alpha_k^{(t)} F(\omega_k) + \beta_k^{(t)}(\omega_k - \omega_0) \\
&\in \omega_0 + \mathbf{Span}\{F(\omega_0), \dots, F(\omega_t)\}\,.
\end{aligned}$$

□

Lemma 1. *(e.g. Chihara, 2011) If $F(\omega) = A\omega + b$,*

$$\omega_t - \omega^* = p_t(A)(\omega_0 - \omega^*)\,, \tag{2}$$

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

Proof. We use Prop. 1 to prove this result by induction. For $t = 0$, the statement holds. Now assume that for all $k \leq t$, $\omega_k - \omega^* = p_k(A)(\omega_0 - \omega^*)$ with $p_{t'}$ a real polynomial of degree at most t' such that $p_{t'}(0) = 1$ (and which depends only on the coefficients of Prop. 1). Note that if $F(\omega^*) = 0$, then as F is linear, we can rewrite F as $F(\omega) = A(\omega - \omega^*)$. Then, by Prop. 1, as $\sum_{k=0}^{t}\beta_k^{(t)} = 1$,

$$\begin{aligned}
\omega_{t+1} - \omega^* &= \sum_{k=0}^{t} \alpha_k^{(t)} F(\omega_k) + \beta_k^{(t)}(\omega_k - \omega^*) \\
&= \sum_{k=0}^{t} \alpha_k^{(t)} A(\omega_k - \omega^*) + \beta_k^{(t)}(\omega_k - \omega^*) \\
&= \sum_{k=0}^{t} \alpha_k^{(t)} A p_k(A)(\omega_0 - \omega^*) + \beta_k^{(t)} p_k(A)(\omega_0 - \omega^*) \\
&= p_{t+1}(A)(\omega_0 - \omega^*)\,,
\end{aligned}$$

where $p_{t+1}(X) = \sum_{k=0}^{t}\alpha_k^{(t)} X p_k(X) + \beta_k^{(t)} p_k(X)$, which is a real polynomial of degree at most $t+1$. Then $p_{t+1}(0) = \sum_{k=0}^{t}\beta_k^{(t)} p_k(0) = 1$, which concludes the proof. □

Lemma 2. *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 (11) converges linearly to ω^* at rate $\left(\rho^* + \epsilon\right)^t$. If V is linear, then $\epsilon = 0$.*

Proof. This is a direct application of Thm. 1 to $V_{augm} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$. $\square$

Proposition 3. *(Nevanlinna, 1993) 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,*

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

Note. If we work in $\mathbb{C}^d$ this proposition is immediate. However, as we constrain ourselves to real vectors and matrices, this is slightly more difficult. This is why we need the matrix representation of complex numbers which is described in the following lemma.

Lemma 3. *Define, for $z \in \mathbb{C}$, the real 2×2 matrix $C(z) = \begin{pmatrix} \Re z & -\Im z \\ \Im z & \Re z \end{pmatrix}$. Then,*

(i). The spectrum of $C(z)$ is $\operatorname{Sp} C(z) = \{z, \bar{z}\}$.

(ii). C is $\mathbb{R}$-linear,

$$\forall z, z' \in \mathbb{C},\ a, a' \in \mathbb{R}, \quad C(az + a'z') = aC(z) + a'C(z')\,.$$

(iii). C is a multiplicative group homorphism,

$$\forall z, z' \in \mathbb{C}, \quad C(zz') = C(z)C(z')\,.$$

We now show a small lemma which will be useful to construct matrices in $\mathcal{M}_K$.

Lemma 4. *Let $K \subset \mathbb{C}$ be a subset of $\mathbb{C}$ symmetric w.r.t. the real axis, and such that $\mathcal{M}_K \neq \emptyset$. If $d \geq 3$, then,*

$$\{A \in \mathbb{R}^{d-2} :\ \operatorname{Sp}(A) \subset K\} \neq \emptyset\,.$$

Proof. We consider two cases, depending on the parity of d.

- Assume that d is odd. We show that this implies that K intersects the real axis. Let M be a matrix in $\mathcal{M}_K$ as it is non-empty by assumption. Then, as the dimension d is odd, M has at least one real eigenvalue, i.e. $\operatorname{Sp} M \cap \mathbb{R} \neq \emptyset$. Hence, $K \cap \mathbb{R} \neq \emptyset$ and let $\nu K \cap \mathbb{R}$ be such an element. Then, the matrix $\operatorname{diag}(\nu, \dots, \nu) \in \mathbb{R}^{(d-2)\times(d-2)}$, which is the square diagonal matrix of size $d-2$ with only ν on its diagonal, belongs to $\{A \in \mathbb{R}^{d-2} :\ \operatorname{Sp}(A) \subset K\}$ which proves the claim.
- Assume that d is even. As $\mathcal{M}_K \neq \emptyset$, $K \neq \emptyset$ and so take $\lambda \in K$. As K is assumed to be symmetric w.r.t. the real axis, $\bar{\lambda}$ belongs to K too. As d is even, we can then define the matrix $M = \operatorname{diag}(C(\lambda), \dots, C(\lambda)) \in \mathbb{R}^{(d-2)\times(d-2)}$ which a real block-diagonal matrix. Its spectrum is $\operatorname{Sp} M = \operatorname{Sp} C(\lambda) = \{\lambda, \bar{\lambda}\} \subset K$ so it proves the claim.

$\square$

Proof. We write this proof with $\omega^* = 0$ without loss of generality. Consider an oblivious first-order method, given by its sequence of polynomials $p_t \in \mathcal{P}_t$, $t \geq 0$. Fix $t \geq 0$ and take $\lambda \in \arg\max_{z \in K} |p_t(z)|$.

We now build $A \in \mathcal{M}_K$ which has λ as an eigenvalue. First assume that $d \geq 3$. Then, by Lem. 4, there exists $M \in \mathbb{R}^{(d-2)\times(d-2)}$ such that $\operatorname{Sp} M \subset K$. Now construct A as,

$$A = \left(\begin{array}{c|c} C(\lambda) & 0_{2\times(d-2)} \\ \hline 0_{(d-2)\times 2} & M \end{array}\right)\,.$$

If $d = 2$, simply take $A = C(\lambda)$.

As A is block-diagonal, $\operatorname{Sp} A = \operatorname{Sp} C(z) \cup \operatorname{Sp} M = \{\lambda, \bar{\lambda}\} \cup \operatorname{Sp} M$. By definition $\operatorname{Sp} M \subset K$ and, as $\lambda \in K$ and K is symmetric w.r.t. the real axis, $\{\lambda, \bar{\lambda}\} \subset K$ too. Hence $\operatorname{Sp} A \subset K$ and so $A \in \mathcal{M}_K$.

We now look at the iterates of the method applied to the vector field $x \mapsto Ax$ to prove the claim. As $\omega^* = 0$, $\|\omega_t - \omega^*\|_2 = \|\omega_t\|_2 = \|p_t(A)\omega_0\|_2$.

To explicit $p_t(A)$, we need to compute $p_t(C(\lambda))$. But, as p_t is a real polynomial, by Lem. 3, we have $p_t(C(\lambda)) = C(p_t(\lambda))$. Hence,

$$p_t(A) = \left(\begin{array}{c|c} C(p_t(\lambda)) & 0_{2\times(d-2)} \\ \hline 0_{(d-2)\times 2} & p_t(M) \end{array} \right) .$$

Now take $\omega_0 = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^T$. Then $\|p_t(A)\omega_0\|^2 = (\Re(p_t(\lambda)))^2 + (\Im(p_t(\lambda)))^2 = |p_t(\lambda)|^2$ and so $\|p_t(A)\omega_0\| = \max_{z\in K} |p_t(z)|\|\omega_0\| \geq \underline{\rho}(K)^t\|\omega_0\|$, which yields the result.

□

Remark 1 (Definition of the asymptotic convergence factor in the matrix iteration literature). The original definitions of the asymptotic convergence factor for linear systems iterations and in particular the one in Nevanlinna (1993) (which is called *optimal reduction factor* in their work), are actually different from the one we presented here. Indeed, the authors work with complex numbers all along so they consider methods with potentially non-real coefficients. Hence, they define the asymptotic convergence factor as,

$$\underline{\rho}(K)' = \inf_{t>0} \min_{q_t \in \mathcal{Q}_t} \max_{\lambda\in K} \sqrt[t]{|q_t(\lambda)|} , \tag{18}$$

where $\mathcal{Q}_t$ is the set of *complex* polynomials q_t of degree at most t such that $q_t(0) = 1$. However, for infinite K which are symmetric w.r.t. the real axis, these two definitions, the one with the complex polynomials and the one with the real polynomials, coincide, as, for all $t \geq 0$,

$$\min_{q_t\in\mathcal{Q}_t} \max_{\lambda\in K} |q_t(\lambda)| = \min_{p_t\in\mathcal{P}_t} \max_{\lambda\in K} |p_t(\lambda)| . \tag{19}$$

This is a consequence of the uniqueness of such minimizers, see Nevanlinna (1993, Cor. 3.5.4).

The following lemma justifies our choice of spectral problem class for strongly monotone and Lipschitz vector fields.

Lemma 5. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable vector field μ-strongly monotone and L-Lipschitz. Then, for all $\omega \in \mathbb{R}^d$,*

$$\mu \leq \Re\lambda , \quad |\lambda| \leq L , \quad \forall\lambda \in \operatorname{Sp} \boldsymbol{J}_F(\omega) . \tag{20}$$

Proof. Fix $\omega \in \mathbb{R}^d$. The first step is standard, see Facchinei and Pang (2003, Prop. 2.3.2) for instance. From the strong monotonicity and the Lipschitz assumptions, for any $\omega' \in \mathbb{R}^d$,

$$(\omega - \omega')^T(F(\omega) - F(\omega')) \geq \mu\|\omega - \omega'\|^2 , \quad \|F(\omega) - F(\omega')\| \leq L\|\omega - \omega'\| .$$

Take $u \in \mathbb{R}^d$. Letting $\omega' = \omega + tu$, dividing by, respectively, t^2 and t, and letting t goes to zero yields,

$$u^T\mathbf{J}_F(\omega)u \geq \mu\|u\|^2 , \quad \|\mathbf{J}_F(\omega)u\| \leq L\|u\| .$$

From the second inequality, we get that $\|\mathbf{J}_F(\omega)\| \leq L$ and so the magnitudes of the eigenvalues of $\mathbf{J}_F(\omega)$ are bounded by L. From the first one, we get that $\mathcal{H}(\mathbf{J}_F(\omega)) := \frac{\mathbf{J}_F(\omega)+\mathbf{J}_F(\omega)^T}{2} \succeq \mu I_d$. Now, for $\lambda \in \operatorname{Sp}\mathbf{J}_F(\omega)$, and $v \in \mathbb{C}^d$ associated eigenvector with $\|v\| = 1$, then $\mathbf{J}_F(\omega)v = \lambda v$ and so $\lambda = \bar{v}^T\mathbf{J}_F(\omega)v$. In particular $\Re\lambda = \frac{\lambda+\bar{\lambda}}{2} = \bar{v}^T\mathcal{H}(\mathbf{J}_F(\omega))\, v \geq \mu\|v\|^2 = \mu$, which yields the result. □

Lemma 6 (Eiermann et al. (1985, §6.2); Nevanlinna (1993, Example 3.8.2))**.** *Let $K = \{z \in \mathbb{C} : |z - c| \leq r\}$ with $c > r > 0$. Then, for all $t \geq 0$, the polynomial*

$$p_t^*(z) = \left(1 - \frac{z}{c}\right)^t .$$

is optimal, i.e.,

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

and so $\underline{\rho}(K) = \frac{r}{c}$. Moreover, the gradient method with step-size $1/c$ is optimal for K: for any vector field F such that $\operatorname{Sp}\boldsymbol{J}_F(\omega^*) \subset K$, the gradient operator defined by

$$\omega_{t+1} = V_{grad}(\omega_t) = \omega_t - \eta F(\omega_t) , \tag{21}$$

satisfy, for $\eta = \frac{1}{c}$,

$$\rho(\boldsymbol{J}_{V_{grad}}(\omega^*)) \leq \frac{r}{c} . \tag{22}$$

This result is only briefly discussed in the references above and as consequence of broader theories. For completeness and simplicity we give a simpler proof using Rouché's theorem. We recall a simplified version of this theorem, see Bak and Newman (2010, Thm. 10.10) for a proof.

Theorem 3 (Rouché)**.** *Let f and g be analytic functions, and $D = \{z \in \mathbb{C} \,|\, |z - z_c| < R\}$ for $z_C \in \mathbb{C}$ and $R > 0$. If for all $z \in \partial D$ the boundary of D it holds that $|f(z)| > |g(z)|$, then the number of zeroes of $f - g$ inside D (counted with multiplicity) is the same as the number of zeroes of f inside D.*

Proof of Lem. 6. Let $p_t^*(z) = \left(1 - \frac{z}{c}\right)^t$ which belongs to $\mathcal{P}_t$. For the sake of contradiction assume that p_t^* is not optimal, i.e. there exists $q_t \in \mathcal{P}_t$ different from p_t^* such that

$$\max_{z \in K} |p_t^*(z)| > \max_{z \in K} |q_t(z)| \,,$$

where K was defined in the statement as $K = \{z \in \mathbb{C} : |z - c| \leq r\}$ with $c > r > 0$. Observe that $|p_t^*|$ reaches its maximum $\left(\frac{r}{c}\right)^t$ on K everywhere on the boundary of K,

$$\max_{z \in K} |p_t^*(z)| = \left(\frac{r}{c}\right)^t = |p_t^*(z_b)| \quad \forall z_b \in \partial K \,.$$

□

Hence, for all $z_b \in \partial K$,

$$|q_t(z_b)| \leq \max_{z \in K} |q_t(z)| < \max_{z \in K} |p_t^*(z)| = |p_t^*(z_b)| \,.$$

Therefore, as q_t and p_t^* are polynomials and in particular analytic, we can apply Rouché's theorem with $D = \operatorname{int} K$ and this yields that the number of zeroes of $p_t^* - q_t$ in $\operatorname{int} K$ is the same as the number of zeroes of p_t^* in $\operatorname{int} K$. On the one hand, c, which belongs to the interior of K, is a zero of multiplicity t of p_t^*. On the other hand, as q_t and p_t^* are in $\mathcal{P}_t$, they satisfy $p_t^*(0) = 1 = q_t(0)$ and so $(p_t^* - q_t)(0) = 0$. However, as $c > r$, 0 is not in K. So, as $(p_t^* - q_t)$ is of degree at most t, it can have at most $t - 1$ remaining zeroes (counted with multiplicity) in $\operatorname{int} K$. This contradicts the conclusion of Rouché's theorem that $p_t^* - q_t$ must have exaclty t zeroes inside K. Therefore, there exists no such q_t and so $p_t^* \in \arg\min_{p_t \in \mathcal{P}_t} \max_{z \in K} |p_t(z)|$.

Moreover, this implies that $\min_{p_t \in \mathcal{P}_t} \max_{z \in K} |p_t(z)| = \left(\frac{r}{c}\right)^t$ and so that $\underline{\rho}(K) = \frac{r}{c}$.

What is left to check is the bound $\rho(\mathbf{J}_V(\omega^*)) \leq \frac{r}{c}$. Recall that $V_{grad}(\omega) = \omega - \eta F(\omega)$ and so that $\mathbf{J}_{V_{\mathrm{grad}}}(\omega) = I_d - \eta \mathbf{J}_F(\omega)$. By the spectral mapping theorem (Thm. 2),

$$\operatorname{Sp} \mathbf{J}_{V_{\mathrm{grad}}}(\omega) = \{1 - \eta\lambda \,|\, \lambda \in \operatorname{Sp} \mathbf{J}_F(\omega)\} \,.$$

Letting $\eta = \frac{1}{c}$ and using that $\operatorname{Sp} \mathbf{J}_F(\omega) \subset K$ yields the result.

C Acceleration related proofs

C.1 Bilinear games

We recall Polyak's theorem.

Theorem 4 (Polyak (1964, Thm. 9)). *Let $0 < \mu < L$. Define Polyak's Heavy-ball method as*

$$\omega_{t+1} = V_{\alpha,\beta}^{Polyak}(\omega_t, \omega_{t-1}) = \omega_t - \alpha F(\omega_t) + \beta(\omega_t - \omega_{t-1})\,. \tag{23}$$

For $\alpha = \frac{4}{(\sqrt{\mu}+\sqrt{L})^2}$ and $\beta = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$ and for any vector field F such that $\operatorname{Sp}\nabla F(\omega^*) \subset [\mu, L]$, then

$$\rho(\nabla V_{\alpha,\beta}^{Polyak}(\omega^*, \omega^*)) \leq \underline{\rho}([\mu, L]) = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\,. \tag{24}$$

In this subsection we first prove the following result.

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 (13) on the transformation (14), i.e.,*

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

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

Proof. This proposition follows from Thm. 4. Indeed, the Jacobian of V^{real} at ω^* is,

$$\begin{aligned}\mathbf{J}_{V^{real}}(\omega^*) &= \frac{1}{\eta}(\mathbf{J}_F(\omega^*)(\mathrm{Id} - \eta \mathbf{J}_F(\omega^*)) - \mathbf{J}_F(\omega^*)) \\ &= -\mathbf{J}_F(\omega^*)^2\,,\end{aligned}$$

where we used $F(\omega^*) = \omega^*$. Now, we can deduce the spectrum of $\mathbf{J}_{V^{real}}(\omega^*)$ from the one of $\mathbf{J}_F(\omega^*)$ using the spectral mapping theorem Thm. 2,

$$\begin{aligned}\operatorname{Sp}\mathbf{J}_{V^{real}}(\omega^*) &= \{-\lambda^2 \mid \lambda \in \operatorname{Sp}\mathbf{J}_F(\omega^*)\} \\ &\subset \{-\lambda^2 \mid \lambda \in \pm[ia, ib]\} \\ &\subset [a^2, b^2]\,.\end{aligned}$$

We can now apply Polyak's momentum method to V^{real} and we get the desired bound on the spectral radius by Thm. 4, with $\alpha = \frac{4}{(a+b)^2}$ and $\beta = \left(\frac{b-a}{b+a}\right)^2$. □

We now prove the following lemma, in order to use Prop. 4 on bilinear games. Note that as A is square, $\sigma_{min}(A)^2$ and $\sigma_{max}(A)^2$ actually correspond to, respectively, the smallest and the largest eigenvalues of AA^T

Lemma 7. *Consider the bilinear game*

$$\min_{x\in\mathbb{R}^m}\max_{y\in\mathbb{R}^m} x^T A y + b^T x + c^T y\,. \tag{25}$$

Let $F : \mathbb{R}^{2m} \to \mathbb{R}^{2m}$ be the associated vector field. Then,

$$\operatorname{Sp}\nabla F(\omega^*) \subset [i\sigma_{min}(A), i\sigma_{max}(A)] \cup [-i\sigma_{min}(A), -i\sigma_{max}(A)]\,. \tag{26}$$

Proof. We have $F(\omega) = \begin{pmatrix} Ay + b \\ -A^T x - c\end{pmatrix}$ and so

$$\nabla F(\omega) = \begin{pmatrix} 0_{m\times m} & A \\ -A^T & 0_{m\times m}\end{pmatrix}\,.$$

We compute the characteristic polynomial of $\nabla F(\omega)$ using the bloc determinant formula, which can be found in Zhang (2005, Section 0.3), as A^T and I_m commute,

$$\det(XI_{2m} - A) = \begin{vmatrix} XI_m & -A \\ A^T & XI_m \end{vmatrix} = \det(X^2 I_m + AA^T)\,.$$

Hence, $\operatorname{Sp} \nabla F(\omega) = \{\pm i\lambda \,|\, \lambda^2 \in \operatorname{Sp} AA^T\}$ which gives the result. □

We now prove the optimality of this method. For this we rely on Eiermann et al. (1989, Thm. 6), that we state below for completeness.

Theorem 5 (Eiermann et al. (1989, Thm. 6))**.** *Let $\Omega \subset \mathbb{C}$ be a compact set such that $0 \notin \Omega$, Ω has no isolated points and $\mathbb{C} \cup \infty \setminus \Omega$ is of finite connectivity. Consider t_n polynomial of degree n such that $t_n(0) = 0$ and define $\tilde{\Omega} = t_n(\Omega)$. If, $t_n^{-1}(\tilde{\Omega}) = \Omega$, then we have,*

$$\underline{\rho}(\Omega) = \underline{\rho}(\tilde{\Omega})^{1/n}$$

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

Proof. We use Thm. 5 with $\Omega = \pm[ia, ib]$, $t_2(X) = -X^2$ and $\tilde{\Omega} = [a^2, b^2]$. We get,

$$\underline{\rho}(\underbrace{\pm[ia, ib]}_{=K}) = \underline{\rho}([a^2, b^2])^{1/2} = \sqrt{\frac{b-a}{b+a}}\,.$$

□

C.2 Ellipses

Define, for $a, b, c \geq 0$, the ellipse

$$E(a,b,c) = \left\{\lambda \in \mathbb{C} : \frac{(\Re\lambda - c)^2}{a^2} + \frac{(\Im\lambda)^2}{b^2} \leq 1\right\}\,. \tag{27}$$

As mentioned earlier, we work with shapes symmetric w.r.t. the real axis and in $\mathbb{C}_+$ (the set of complex number with non-negative real part). So the ellipses we consider have their center on the positive real axis and we we will require below that $0 \notin E(a, b, c)$. Ellipses have been studied in the context of matrix iteration, due to their flexibility and their link to the momentum method. The next theorem can be considered as a summary and reinterpretation of the literature on the subject. The way to obtain it from the literature, and a partial proof, are deferred to the Appendix D.

Theorem 6. *Let $a, b \geq 0$, $c > 0$, $(a, b) \neq 0$, such that $0 \notin E(a, b, c)$. Then, if $\rho(a, b, c) < 1$,*

$$\underline{\rho}(E(a,b,c)) = \rho(a,b,c) \tag{28}$$

where

$$\rho(a,b,c) = \begin{cases} \frac{a}{c} & \text{if } a = b \\ \frac{c - \sqrt{b^2+c^2-a^2}}{a-b} & \text{otherwise} \end{cases} \tag{29}$$

Assume that F is any vector field $F : \mathbb{R}^d \to \mathbb{R}^d$ that satisfies $\operatorname{Sp} \nabla F(\omega^*) \subset E(a, b, c)$. There exists $\alpha(a, b, c) > 0$, $\beta(a, b, c) \in (-1, 1]$, whose signis the same as $a - b$ such that, for the momentum operator $V : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$,

$$V(\omega, \omega') = (\omega - \alpha F(\omega) + \beta(\omega - \omega'), \omega'), \tag{30}$$

we have

$$\rho(\boldsymbol{J}_V(\omega^*, \omega^*)) \leq \rho(a,b,c)\,. \tag{31}$$

More exactly the corresponding parameters are given by,

$$\beta(a,b,c) = \begin{cases} 0 & \text{if } a = b \\ 2c\frac{c-\sqrt{c^2+b^2-a^2}}{a^2-b^2} - 1 & \text{otherwise,} \end{cases} \qquad \alpha(a,b,c) = \frac{1+\beta}{c} = \begin{cases} \frac{1}{c} & \text{if } a = b \\ 2\frac{c-\sqrt{c^2+b^2-a^2}}{a^2-b^2} & \text{otherwise,} \end{cases} \tag{32}$$

and $\beta(a, b, c)$ can be written $\beta = \chi(a, b, c)(a - b)$ with $\chi(a, b, c) > 0$.

Remark 2 (On the sign of the momentum parameter). As briefly mentioned in the theorem, and detailed in Prop. 13, the optimal momentum parameter $\beta(a,b,c)$ has the same sign as $a-b$, i.e., more exactly, there exists $\chi(a,b,c)>0$ such that $\beta(a,b,c)=\chi(a,b,c)(a-b)$. Hence the sign of the optimal β has a nice geometric interpretation, which answers some of the questions left open by Gidel et al. (2019b).

- In the case where $a>b$, or equivalently $\beta>0$, the ellipse is more elongated in the direction of the real axis. The extreme case is a segment on the real line that corresponds to strongly convex optimization.
- In the case where $a<b$, or equivalently $\beta<0$, the ellipse is more elongated in the direction of the imaginary axis.
- Finally, when $a=b$ we have a disk instead of an ellipse. For such shape, we have no momentum, which means that gradient descent is optimal as seen in §4.3.

C.3 Perturbed acceleration

In this subsection, we prove Prop. 6. Note that the constants in the $\mathcal{O}(.)$ are absolute.

Proposition 9. *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 $\frac{\mu}{L}\to 0$,*

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

Moreover, the momentum method is optimal for K_ϵ. This means that with $\alpha\left(\frac{L-\mu}{2},\epsilon,\frac{L+\mu}{2}\right)>0$ and $\beta\left(\frac{L-\mu}{2},\epsilon,\frac{L+\mu}{2}\right)>0$ (as defined in Thm. 6), if $\operatorname{Sp}\boldsymbol{J}_F(\omega^*)\subset K_\epsilon$, then, $\rho(\boldsymbol{J}_{V^{Polyak}}(\omega^*,\omega^*))\leq\underline{\rho}(K_\epsilon)$.

where K_ϵ is the ellipse defined by,

$$K_\epsilon=\left\{z\in\mathbb{C}:\left(\frac{\Re z-\frac{\mu+L}{2}}{\frac{L-\mu}{2}}\right)^2+\left(\frac{\Im z}{\epsilon}\right)^2\leq 1\right\}. \tag{33}$$

Proof. A direct application of Thm. 6 using K_ϵ gives $\underline{\rho}(K_\epsilon)=\rho(\frac{L-\mu}{2},\epsilon(\mu,L),\frac{L+\mu}{2})$.
We now study $\rho(\frac{L-\mu}{2},\epsilon(\mu,L),\frac{L+\mu}{2})$. First note that

$$\rho(a,b,c)=1-\frac{a-b-c+\sqrt{b^2+c^2-a^2}}{a-b},$$

and so

$$1-\rho(a,b,c)=\frac{\sqrt{b^2+c^2-a^2}+a-b-c}{a-b}.$$

We now replace a, b and c by their expressions (and multiply the denominator and numerator by 2).

$$\begin{aligned}1-\underline{\rho}(K_\epsilon)&=\frac{\sqrt{4\epsilon^2+(L+\mu)^2-(L-\mu)^2}+(L-\mu)-2\epsilon-(L+\mu)}{L-\mu-2\epsilon}\\&=2\frac{\sqrt{\epsilon^2+\mu L}-2\epsilon-2\mu}{L-\mu-2\epsilon}.\end{aligned}$$

Define $t = \frac{\mu}{L}$, then $\frac{\epsilon}{L} = t^\theta$. We are now interested in studying the behaviour of $1 - \underline{\rho}(K_\epsilon)$ when t goes to zero.

$$\begin{aligned}
1 - \underline{\rho}(K_\epsilon) &= 2\frac{\sqrt{t^{2\theta} + t} - 2t^\theta - 2t}{1 - t - 2t^\theta} \\
&= 2(\sqrt{t^{2\theta} + t} - 2t^\theta - 2t)(1 + t + 2t^\theta + \mathcal{O}\left(t^{2(\theta \wedge 1)}\right) , \\
&= 2\left(\sqrt{t^{2\theta} + t} - 2t^\theta - 2t\right)\left(1 + \mathcal{O}\left(t^{\theta \wedge 1}\right)\right) ,
\end{aligned}$$

where $a \wedge b$ denotes $\min(a, b)$.

- If $\theta = \frac{1}{2}$. This is the smallest θ with which acceleration happens.

$$\begin{aligned}
1 - \underline{\rho}(K_\epsilon) &= 2\left(\sqrt{2t} - 2\sqrt{t} - 2t\right)\left(1 + \mathcal{O}\left(\sqrt{t}\right)\right) \\
&= 2\left(\sqrt{2} - 1\right)\sqrt{t}\left(1 + \mathcal{O}\left(\sqrt{t}\right)\right)\left(1 + \mathcal{O}\left(\sqrt{t}\right)\right) \\
&= 2\left(\sqrt{2} - 1\right)\sqrt{t} + \mathcal{O}\left(t\right) .
\end{aligned}$$

- If $\theta > \frac{1}{2}$. This regime is "better" than the previous one, i.e., the perturbation is even smaller so we get a similar asymptotic behavior, up to an improved constant.

$$\begin{aligned}
1 - \underline{\rho}(K_\epsilon) &= 2\left(\sqrt{t^{2\theta} + t} - 2t^\theta - 2t\right)\left(1 + \mathcal{O}\left(t^{\theta \wedge 1}\right)\right) \\
&= 2\sqrt{t}\left(\sqrt{t^{2\theta - 1} + 1} - 2t^{\theta - 1/2} - 2\sqrt{t}\right)\left(1 + \mathcal{O}\left(t^{\theta \wedge 1}\right)\right) \\
&= 2\sqrt{t}\left(1 + \mathcal{O}\left(t^{2\theta - 1}\right) - 2t^{\theta - 1/2} - 2\sqrt{t}\right)\left(1 + \mathcal{O}\left(t^{\theta \wedge 1}\right)\right) \\
&= 2\sqrt{t} + \mathcal{O}\left(t^{\theta \wedge 1}\right) .
\end{aligned}$$

- If $\theta < \frac{1}{2}$. In this regime, terms in t^θ are limiting as they are bigger than $\sqrt{t}$, so we do not get the rate in $\sqrt{t}$.

$$\begin{aligned}
1 - \underline{\rho}(K_\epsilon) &= 2\left(\sqrt{t^{2\theta} + t} - t^\theta - t\right)\left(1 + \mathcal{O}\left(t^\theta\right)\right) \\
&= 2\left(t^\theta\sqrt{1 + t^{1 - 2\theta}} - t^\theta - t\right)\left(1 + \mathcal{O}\left(t^\theta\right)\right) \\
&= 2\left(t^\theta\left(1 + \frac{1}{2}t^{1 - 2\theta} + \mathcal{O}\left(t^{2 - 4\theta}\right)\right) - t^\theta - t\right)\left(1 + \mathcal{O}\left(t^\theta\right)\right) \\
&= 2t^{1 - \theta}\left(\frac{1}{2} + \mathcal{O}\left(t^{1 - 2\theta}\right) - t^\theta\right)\left(1 + \mathcal{O}\left(t^\theta\right)\right) \\
&= t^{1 - \theta} + \mathcal{O}\left(t^{1 \wedge (2 - 3\theta)}\right) .
\end{aligned}$$

□

Before going further we need to introduce this technical lemma.

Lemma 8. *For any real $m \geq 2$, we have*

$$\forall x \in [0, 1] : \quad \frac{(1 - x)^2}{(1 - \frac{x}{m})^2} + x \leq 1 . \tag{34}$$

Proof. Indeed,

$$(1-x^2)+x\left(1-\frac{x}{m}\right)^2-\left(1-\frac{x}{m}\right)^2 = 1-2x+x^2+x-2\frac{x^2}{m}+\frac{x^3}{m^2}-1+2\frac{x}{m}-\frac{x^2}{m^2} \tag{35}$$

$$= \left(\frac{2}{m}-1\right)x+\left(1-\frac{2}{m}\right)x^2+\frac{x^3-x^2}{m^2} \tag{36}$$

$$= \left(1-\frac{2}{m}\right)x(x-1)+\frac{x^3-x^2}{m^2} \tag{37}$$

$$\tag{38}$$

Then, as $m \geq 2$, $1-\frac{2}{m} \geq 0$, and so, since $0 \leq x \leq 1$, $\left(1-\frac{2}{m}\right)x(x-1) \leq 0$. As $0 \leq x \leq 1$, we also have $\frac{x^3-x^2}{m^2} \leq 0$ which concludes the proof. □

C.4 Acceleration of extragradient on bilinear games

We now prove that we can accelerate EG on bilinear games.

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.

This proposition is a consequence of Lem. 2 combinded with the following result.

Proposition 10. *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*

$$\begin{aligned}\omega_{t+1} &= V^{Polyak+e-g}(\omega_t, \omega_{t-1}),\\ &= \omega_t - \alpha F(\omega_t - \eta F(\omega_t)) + \beta(\omega_t - \omega_{t-1}),\end{aligned}$$

satisfies, with $c = \sqrt{2}-1$, and absolute constants in the $\mathcal{O}(.)$,

$$\rho(\boldsymbol{J}_{V^{Polyak+e-g}}(\omega^*, \omega^*)) \leq 1 - c\tfrac{a}{b} + O(\tfrac{a^2}{b^2}). \tag{39}$$

More precisely, the parameters are chosen as:

$$\eta = \frac{b}{a\sqrt{2b^2 - \frac{a^2}{2}}} \quad \alpha = \alpha\left(\eta\left(b^2 - \frac{a^2}{2}\right), b, \eta b^2\right) \quad \beta = \beta\left(\eta\left(b^2 - \frac{a^2}{2}\right), b, \eta b^2\right),$$

with where the functions $\alpha(\cdot)$ and $\beta(\cdot)$ are the ones defined in Thm. 6.

Note that this proposition actually requires α and η to be tuned separately and they are actually very different. They actually differ by a factor $\frac{b}{a}$: η is roughly proportional to $\frac{1}{a}$ while α behaves like $\frac{b^2}{a}$.

Proof. Consider $F^{\text{e-g}}(\omega) = F(\omega - \eta F(\omega))$. Then, for ω^* such that $F(\omega^*) = 0$, we have,

$$\mathbf{J}_{F^{\text{e-g}}}(\omega^*) = \mathbf{J}_F(\omega^*) - \eta \mathbf{J}_F(\omega^*)^2.$$

Hence, by Thm. 2,

$$\operatorname{Sp} \nabla \mathbf{J}_{F^{\text{e-g}}}(\omega^*) \subset \{z - \eta z^2 \mid z \in \pm[ia, ib]\} = \{i\lambda + \eta\lambda^2 \mid \lambda \in \pm[a, b]\}.$$

As we want to apply Prop. 9, we now look for $\epsilon, \bar{\mu}, \bar{L}$ such that the ellipsis

$$K(\epsilon, \bar{\mu}, \bar{L}) = \left\{z \in \mathbb{C} : \left(\frac{\Re z - \frac{\bar{\mu}+\bar{L}}{2}}{\frac{\bar{L}-\bar{\mu}}{2}}\right)^2 + \left(\frac{\Im z}{\epsilon}\right)^2 \leq 1\right\} \tag{40}$$

contains $\operatorname{Sp} \nabla \mathbf{J}_{F^{\text{e-g}}}(\omega^*)$. For this we will choose $\epsilon, \bar{\mu}, \bar{L}$ such that

$$\{i\lambda + \eta\lambda^2 \,|\, \lambda \in \pm[a,b]\} \subset K(\epsilon, \bar{\mu}, \bar{L})\,,$$

which is equivalent to for all $\lambda \in [a,b]$,

$$\left(\frac{\eta\lambda^2 - \frac{\bar{\mu}+\bar{L}}{2}}{\frac{\bar{L}-\bar{\mu}}{2}}\right)^2 + \left(\frac{\lambda}{\epsilon}\right)^2 \leq 1\,. \tag{41}$$

Note that the left-hand side is convex in λ^2 so that we only need to check this inequality for the limit values $\lambda = a$ and $\lambda = b$. Hence, now we have reduced the problem to that of looking for $\bar{\mu}$, $\bar{L}$ and ϵ such that $\lambda = a$ and $\lambda = b$ satisfy (41). This is equivalent to look for an ellipse $K(\epsilon, \bar{\mu}, \bar{L})$ that contains $ib + \eta b^2$ and $ia + \eta a^2$.

We now construct this ellipsis explicitly. As we want to apply Prop. 9, we want ϵ as small as possible. So we start with $ib + \eta b^2$ as it is the one with the largest imaginary part, compared to $ia + \eta a^2$. We choose the center of the ellipse – which must lie on the real axis – such that it is placed at the same abscisse as $ib + \eta b^2$, i.e. the same real part. So we define $\frac{\bar{\mu}+\bar{L}}{2} = \eta b^2$. We need another condition to fix $\bar{\mu}$ and $\bar{L}$. To make sure that $ia + \eta a^2$ is also in the ellipsis, we need to choose $\bar{\mu}$ small enough. Define $\bar{\mu} = \frac{\eta a^2}{m}$ with $m > 0$ to be chosen later. This fixes the value of $\bar{L}$ as,

$$\bar{L} = 2\eta b^2 - \bar{\mu} = 2\eta b^2 - \eta\frac{a^2}{m}\,.$$

We choose ϵ so that $ib + \eta b^2$ is in the ellipsis, and as we chose the center to be $\frac{\bar{\mu}+\bar{L}}{2} = \eta b^2$, we define $\epsilon = b$. This way $ib + \eta b^2 \in K(\epsilon, \bar{\mu}, \bar{L})$. We must now check that $ia + \eta a^2 \in K(\epsilon, \bar{\mu}, \bar{L})$. For this we check that $\lambda = a$ satisfies (41),

$$\left(\frac{\eta a^2 - \frac{\bar{\mu}+\bar{L}}{2}}{\frac{\bar{L}-\bar{\mu}}{2}}\right)^2 + \left(\frac{a}{\epsilon}\right)^2 \tag{42}$$

$$= \left(\frac{\eta a^2 - \eta b^2}{\eta b^2 - \eta\frac{a^2}{m}}\right)^2 + \left(\frac{a}{b}\right)^2 \tag{43}$$

$$= \frac{(1-x)^2}{(1-\frac{x}{m})^2} + x\,, \tag{44}$$

$$\tag{45}$$

where $x = \frac{a^2}{b^2} \in [0,1]$. By Lem. 8, for $m = 2$, this quantity is smaller than one and so $ia + \eta a^2 \in K(\epsilon, \bar{\mu}, \bar{L})$. Hence, $K(\epsilon, \bar{\mu}, \bar{L})$ contains $\operatorname{Sp} \nabla \mathbf{J}_{F^{\text{e-g}}}(\omega^*)$.

Before we apply Prop. 9, we need to make sure that we are in the accelerated regime, that is to say ϵ is small enough compared to $\bar{\mu}$ and $\bar{L}$. Fortunately, we have not chosen η yet. So we define it so that we reach the accelerated regime,

$$\frac{\epsilon}{\bar{L}} = \sqrt{\frac{\bar{\mu}}{\bar{L}}} \iff \epsilon = \sqrt{\bar{\mu}\bar{L}} \iff \eta = \frac{b}{a\sqrt{2b^2 - \frac{a^2}{m}}}\,.$$

We now apply Prop. 9. As $\frac{\bar{\mu}}{\bar{L}}$ goes to zero, $\underline{\rho}(K(\epsilon, \bar{\mu}, \bar{L})) = 1 - 2(\sqrt{2}-1)\sqrt{\frac{\bar{\mu}}{\bar{L}}} + \mathcal{O}\left(\frac{\bar{\mu}}{\bar{L}}\right)$.

Now note that $\frac{\bar{\mu}}{\bar{L}} = \frac{a^2}{2mb^2 - a^2}$, so if $\frac{a}{b} \to 0$ then $\frac{\mu}{L} \to 0$. Moreover, $\frac{\bar{\mu}}{\bar{L}} = \frac{a^2}{2mb^2} + \mathcal{O}\left(\frac{a^4}{b^4}\right)$. Hence, as we chose $m = 2$,

$$\underline{\rho}(K(\epsilon, \bar{\mu}, \bar{L})) = 1 - (\sqrt{2}-1)\frac{a}{b} + \mathcal{O}\left(\frac{a^2}{b^2}\right)\,.$$

The parameters of the momentum method applied to $F^{\text{e-g}}$ are then chosen according to Prop. 6 so that we reach this rate locally. □

In the previous proposition we showed that EG can be accelerated but that it requires a careful choice of the parameters α, β, η. The following lemma describes the general behavior of these quantities when the condition number worsens.

Lemma 9. *In the context of the previous proposition, Prop. 10, it holds, when* $\frac{a}{b} \to 0$,

$$\begin{aligned}
\eta &= \frac{1}{a}\left(\frac{1}{\sqrt{2}} + \mathcal{O}\left(\frac{a^2}{b^2}\right)\right) \\
\alpha &= \frac{a}{b^2}\left(2\sqrt{2} + \mathcal{O}\left(\frac{a}{b}\right)\right) \\
\beta &= 1 - 2\sqrt{3}\frac{a}{b} + \mathcal{O}\left(\frac{a^2}{b^2}\right) .
\end{aligned}$$

Proof. For compactness, denote $t = \frac{a}{b} > 0$. So we study the asymptotic behavior of α and η when t goes to 0. By definition of η we have,

$$\eta = \frac{b}{a\sqrt{2b^2 - \frac{a^2}{2}}} = \frac{1}{\sqrt{2}a\sqrt{1 - \frac{t^2}{4}}} = \frac{1}{\sqrt{2}a}\left(1 + \frac{t^2}{8} + \mathcal{O}\left(t^4\right)\right) ,$$

which gives the first claim as $\frac{t^2}{8} + \mathcal{O}\left(t^4\right) = \mathcal{O}\left(t^2\right)$.

Before moving to the second claim, let us consider some consequences of this asymptotic expansion of η. Indeed, we have,

$$\eta b^2 = \eta\frac{a^2}{t^2} = \frac{a}{\sqrt{2}t^2}\left(1 + \frac{t^2}{8} + \mathcal{O}\left(t^4\right)\right)$$

and, using the expansion above,

$$\eta\left(b^2 - \frac{a^2}{2}\right) = \eta b^2\left(1 - \frac{t^2}{2}\right) = \frac{a}{\sqrt{2}t^2}\left(1 + \frac{t^2}{8} + \mathcal{O}\left(t^4\right)\right)\left(1 - \frac{t^2}{2}\right) = \frac{a}{\sqrt{2}t^2}\left(1 - \frac{3t^2}{8} + \mathcal{O}\left(t^4\right)\right) .$$

We can now study the behavior of $\alpha = \alpha\left(\eta\left(b^2 - \frac{a^2}{2}\right), b, \eta b^2\right)$. Recall that $\alpha(.,.,.)$ is the function defined in Thm. 6 and that its definition depends on whether its first and second arguments are equal. However, suing the expansion above, the ratio of its first and second arguments is

$$\frac{\eta\left(b^2 - \frac{a^2}{2}\right)}{b} = \frac{1}{\sqrt{2}t}\left(1 - \frac{3t^2}{8} + \mathcal{O}\left(t^4\right)\right) ,$$

which diverges to infinity as t goes to zero. Hence, when t is close enough to zero, $\eta\left(b^2 - \frac{a^2}{2}\right)$ and b are different and so we have,

$$\alpha = \alpha\left(\eta\left(b^2 - \frac{a^2}{2}\right), b, \eta b^2\right) = 2\frac{\eta b^2 - \sqrt{(\eta b^2)^2 + b^2 - \left(\eta\left(b^2 - \frac{a^2}{2}\right)\right)^2}}{\left(\eta\left(b^2 - \frac{a^2}{2}\right)\right)^2 - b^2} .$$

We first consider the term under the square root,

$$\begin{aligned}
\sqrt{(\eta b^2)^2 + b^2 - \left(\eta\left(b^2 - \frac{a^2}{2}\right)\right)^2} &= \sqrt{\frac{a^2}{2t^4}\left(1 + \frac{t^2}{8} + \mathcal{O}\left(t^4\right)\right)^2 + \frac{a^2}{t^2} - \frac{a^2}{2t^4}\left(1 - \frac{3t^2}{8} + \mathcal{O}\left(t^4\right)\right)^2} \\
&= \sqrt{\frac{a^2}{2t^4}\left(1 + \frac{t^2}{4} + \mathcal{O}\left(t^4\right)\right) + \frac{a^2}{t^2} - \frac{a^2}{2t^4}\left(1 - \frac{3t^2}{4} + \mathcal{O}\left(t^4\right)\right)} \\
&= \sqrt{\frac{3}{2}\frac{a^2}{t^2} + a^2 \times \mathcal{O}\left(1\right)} \\
&= \sqrt{\frac{3}{2}}\frac{a}{t}\sqrt{1 + \mathcal{O}\left(t^2\right)} \\
&= \sqrt{\frac{3}{2}}\frac{a}{t}\left(1 + \mathcal{O}\left(t^2\right)\right) .
\end{aligned}$$

We can now give the expansion of α when $t = \frac{a}{b}$ goes to zero,

$$\begin{aligned}
\alpha &= 2\frac{\eta b^2 - \sqrt{(\eta b^2)^2 + b^2 - \left(\eta\left(b^2 - \frac{a^2}{2}\right)\right)^2}}{\left(\eta\left(b^2 - \frac{a^2}{2}\right)\right)^2 - b^2} \\
&= 2\frac{\frac{a}{\sqrt{2}t^2}\left(1 + \mathcal{O}\left(t^2\right)\right) - \sqrt{\frac{3}{2}}\frac{a}{t}\left(1 + \mathcal{O}\left(t^2\right)\right)}{\frac{a^2}{2t^4}\left(1 + \mathcal{O}\left(t^2\right)\right) - \frac{a^2}{t^2}} \\
&= 2\frac{\frac{1}{\sqrt{2}}\left(1 + \mathcal{O}\left(t^2\right)\right) - \sqrt{\frac{3}{2}}t\left(1 + \mathcal{O}\left(t^2\right)\right)}{\frac{a}{2t^2}\left(1 + \mathcal{O}\left(t^2\right)\right) - a}
\end{aligned}$$

by multiplying both the numerator and the denominator by $\frac{t^2}{a}$. Then, if we factorize $\frac{a}{t^2}$ in the denominator, we get,

$$\begin{aligned}
\alpha &= 2\frac{t^2}{a}\frac{\frac{1}{\sqrt{2}}\left(1 + \mathcal{O}\left(t^2\right)\right) - \sqrt{\frac{3}{2}}t\left(1 + \mathcal{O}\left(t^2\right)\right)}{\frac{1}{2}\left(1 + \mathcal{O}\left(t^2\right)\right) - t^2} \\
&= 2\frac{t^2}{a}\frac{\frac{1}{\sqrt{2}} - \sqrt{\frac{3}{2}}t + \mathcal{O}\left(t^2\right)}{\frac{1}{2} + \mathcal{O}\left(t^2\right)} \\
&= 2\sqrt{2}\frac{t^2}{a}(1 - \sqrt{3}t + \mathcal{O}\left(t^2\right)),
\end{aligned}$$

which yields the result for α.

Recall that, from the definition of $\alpha(\cdot)$ and $\beta(\cdot)$ in Thm. 6, we have,

$$\beta = \eta b^2 \alpha - 1\,,$$

and so, when $t = \frac{a}{b}$ goes to zero,

$$\beta = \frac{a}{\sqrt{2}t^2}(1 + \mathcal{O}\left(t^2\right)) \times 2\sqrt{2}\frac{t^2}{a}(1 - \sqrt{3}t + \mathcal{O}\left(t^2\right)) - 1 = 1 - 2\sqrt{3}t + \mathcal{O}\left(t^2\right)\,.$$

□

C.5 Consensus optimization and momentum

The general idea behind the proof of Prop. 11 is illustrated by Fig. 3. Using the following lemma, we first prove that the eigenvalues of the consensus optimization operator $F^{cons.}$ are contained in a trapezoid (in blue on the figure). Then, we find a suitable ellipse of the form of Prop. 9 (in orange) such that the trapezoid, thus the spectrum of $\mathbf{J}_{F^{cons.}}$ as shown by Lem. 10, fits inside.

First we need to refine Mescheder et al. (2017, Lem. 9).

Lemma 10. *Let $A \in \mathbb{R}^{d\times d}$ be a square matrix. Let σ_i be the singular values of A. Assume that*

$$\gamma \le \sigma_i \le L, \quad \mathcal{H}(A) \coloneqq \tfrac{A+A^T}{2} \succeq \mu I_d\,,$$

with $\gamma > 0$ and $\mu \ge 0$. Then, for $\tau > 0$ such that $\tau\gamma^2 \ge \mu(1 + 2\tau\mu)$,

$$\max\left\{\frac{|\Im\lambda|}{|\Re\lambda|}\;\middle|\;\lambda \in \mathrm{Sp}(A + \tau A^T A)\right\} \le \frac{\gamma}{\mu + \tau\gamma^2}\,.$$

Moreover, for $\lambda \in \mathrm{Sp}(A + \tau A^T A)$, we have $\mu + \tau\gamma^2 \le \Re\lambda \le L + \tau L^2$.

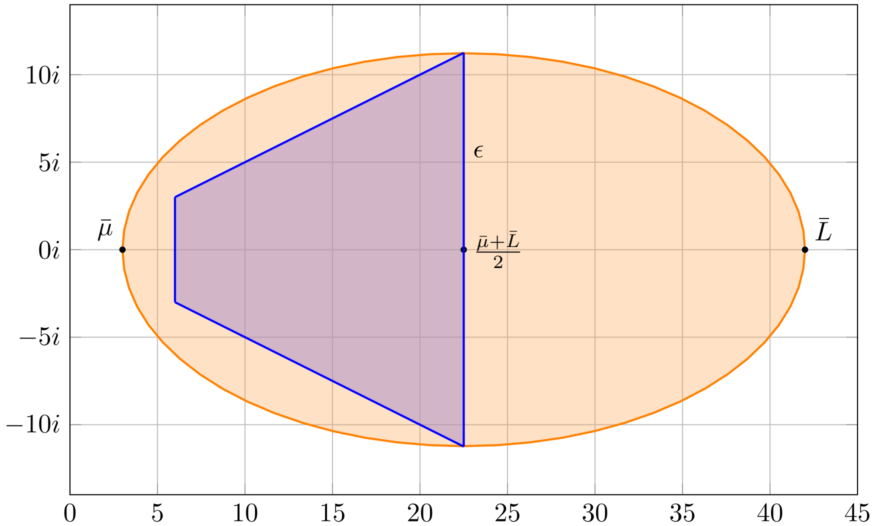

Figure 3: Illustration of the proof of Prop. 11. Lem. 10 guarantees that the spectrum of $\mathbf{J}_{F^{\text{cons.}}}$ is located inside a trapezoid (in blue). We then find a suitable ellipse of the form of Prop. 9 (in orange) which contains it.

Proof. In this proof, for M a real matrix, $\mathcal{H}(M) = \frac{M+M^T}{2}$ its Hermitian part and $\mathcal{S}(M = (M - M^T)/2$ its skew-symmetric part.

Let $B = A + \tau A^T A$. Let $\lambda \in \operatorname{Sp} B$ and let $v \in \mathbb{C}^d$ its associated eigenvector with $\|v\| = 1$. Then

$$\Re\lambda = \frac{\lambda + \bar{\lambda}}{2} = \bar{v}^T \mathcal{H}(B)v = \bar{v}^T \mathcal{H}(A)v + \tau\|Av\|^2 \geq \mu\|v\|^2 + \tau\|Av\|^2 = \mu + \tau\|Av\|^2\,, \tag{46}$$

by the assumption on $\mathcal{H}(A)$. We now deal with the imaginary part,

$$\Im\lambda = \frac{\lambda - \bar{\lambda}}{2i} = \frac{1}{i}\bar{v}^T \mathcal{S}(A)v\,.$$

However, this quantity is hard to bound. Thus, we rewrite it using $\bar{v}^T A\bar{v}$ and $\bar{v}^T \mathcal{H}(A)v$. We have that $\bar{v}^T \mathcal{H}(A)v$ and $\frac{1}{i}\bar{v}^T \mathcal{S}(A)v$ are real and correspond respectively to the real and imaginary parts of $\bar{v}^T Av$. Hence,

$$(\Im\lambda)^2 = \left(\Im\bar{v}^T Av)\right)^2 = |\bar{v}^T Av|^2 - (\Re\bar{v}^T Av)^2 \leq |\bar{v}^T Av|^2\,.$$

Using Cauchy-Schwarz inequality we get,

$$(\Im\lambda)^2 \leq \|Av\|^2\,.$$

Finally,

$$\frac{|\Im\lambda|}{|\Re\lambda|} \leq \frac{\|Av\|}{\mu + \tau\|Av\|^2} = \varphi(\|Av\|)\,.$$

with $\varphi : x \mapsto \frac{x}{\mu+\tau x^2}$. Using the derivative φ', we have that φ is non-decreasing before $x = \sqrt{\frac{\mu}{\tau}}$ and non-increasing after. Note that $\|Av\| \geq \sigma_{min}(A) \geq \gamma$. Hence, if $\tau\gamma^2 \geq \mu$, then $\varphi(\|Av\|) \leq \varphi(\gamma) = \frac{\gamma}{\mu+\tau\gamma^2}$ which concludes the proof of the first part of the lemma.

Now, take $\lambda \in \operatorname{Sp}(A + \tau A^T A)$. Then the inequality $\mu + \tau\gamma^2 \leq \Re\lambda$ comes from (46). The other one is

$$\Re\lambda \leq |\lambda| = \|Bv\| \leq \|Av\| + \tau\|Av\|^2 \leq L + \tau L^2\,.$$

□

We can now proceed to show Prop. 8 by proving the more detailed proposition below.

Proposition 11. *Let σ_i be the singular values and eigenvalues of $J_F(\omega^*)$. Assume that*

$$\gamma \leq \sigma_i \leq L, \quad \tfrac{\boldsymbol{J}_F(\omega^*)+\boldsymbol{J}_F(\omega^*)}{2} \succeq \mu I_d\,.$$

Define $F^{cons.}(\omega) = F(\omega) + \tau\nabla(\frac{1}{2}\|F\|^2)(\omega)$ with $\tau > 0$ and consider the momentum method applied to $F^{cons.}$,

$$\begin{aligned}\omega_{t+1} &= V^{mom+cons.}(\omega_t, \omega_{t-1})\\ &= \omega_t - \alpha F^{cons.}(\omega_t) + \beta(\omega_t - \omega_{t-1}).\end{aligned}$$

If $\tau\gamma^2 \geq \mu$ and

$$\frac{\gamma}{\mu+\tau\gamma^2} \leq \sqrt{\frac{3}{2}}\sqrt{\frac{\mu+\tau\gamma^2}{L+\tau L^2}}\,, \tag{47}$$

Then, one can choose $\alpha > 0$ and $\beta > 0$ such that,

$$\rho(\boldsymbol{J}_{V^{mom+cons.}}(\omega^*, \omega^*))) \leq 1 - c\sqrt{\frac{\mu+\tau\gamma^2}{L+\tau L^2}} + \mathcal{O}\left(\frac{\mu+\tau\gamma^2}{L+\tau L^2}\right)\,. \tag{48}$$

More precisely, the parameters are given by,

$$\alpha = \alpha\left(L+\tau L^2 - \frac{\mu+\tau\gamma^2}{2}, \frac{1}{2}\sqrt{\mu+\tau\gamma^2}\sqrt{4(L+\tau L^2)-(\mu+\tau\gamma^2)}, L+\tau L^2\right) \tag{49}$$

$$\beta = \beta\left(L+\tau L^2 - \frac{\mu+\tau\gamma^2}{2}, \frac{1}{2}\sqrt{\mu+\tau\gamma^2}\sqrt{4(L+\tau L^2)-(\mu+\tau\gamma^2)}, L+\tau L^2\right)\,, \tag{50}$$

where the functions $\alpha(\cdot)$ and $\beta(\cdot)$ are the ones defined in Thm. 6.

If $\tau = \frac{L}{\gamma^2}$ Then, $\rho(\boldsymbol{J}_{V^{mom+cons.}}(\omega^*, \omega^*)))$ is bounded by

$$\rho(\boldsymbol{J}_{V^{mom+cons.}}(\omega^*, \omega^*))) \leq 1 - (\sqrt{2}-1)\frac{\gamma}{L} + \mathcal{O}\left(\frac{\gamma^2}{L^2}\right)\,,$$

where the constants in the $\mathcal{O}(.)$ are absolute.

Proof. Similarly to the proof of Prop. 7, we want to apply Prop. 9. We now look for $\epsilon, \bar{\mu}, \bar{L}$ such that the ellipsis

$$K(\epsilon, \bar{\mu}, \bar{L}) = \left\{z \in \mathbb{C} : \left(\frac{\Re z - \frac{\bar{\mu}+\bar{L}}{2}}{\frac{\bar{L}-\bar{\mu}}{2}}\right)^2 + \left(\frac{\Im z}{\epsilon}\right)^2 \leq 1\right\} \tag{51}$$

contains $\operatorname{Sp}\nabla\mathbf{J}_{F^{\text{cons.}}}(\omega^*)$. First we compute $\mathbf{J}_{F^{\text{cons.}}}(\omega^*)$, for F twice differentiable. Note that $F^{\text{cons.}}$ can be written as $F^{\text{cons.}}(\omega) = F(\omega) + \tau\mathbf{J}_F^T(\omega)F(\omega)$, thus

$$\mathbf{J}_{F^{\text{cons.}}}(\omega^*) = \mathbf{J}_F(\omega^*) + \tau\mathbf{J}_F^T(\omega)\mathbf{J}_F(\omega)\,.$$

Note that the derivative of $\mathbf{J}_F$ does not appear as $F(\omega^*) = 0$. From Lem. 10 with $A = \mathbf{J}_F(\omega^*)$, we have a control on

$$q(\tau) := \frac{\gamma}{\mu+\tau\gamma^2} \geq \max\left\{\frac{|\Im\lambda|}{|\Re\lambda|} \;\middle|\; \lambda \in \operatorname{Sp}(\mathbf{J}_{F^{\text{cons.}}}(\omega^*))\right\} \geq 0\,.$$

Using $q(\tau)$ and the bounds on the real parts of Lem. 10, we have that the spectrum of $\mathbf{J}_{F^{\text{cons.}}}(\omega^*)$ is inside the following shape,

$$\operatorname{Sp}\mathbf{J}_{F^{\text{cons.}}}(\omega^*) \subset S(\tau) := \{\lambda \in \mathbb{C} \mid \mu+\tau\gamma^2 \leq \Re\lambda \leq L+\tau L^2,\, |\Im\lambda| \leq q(\tau)\Re\lambda\}\,.$$

We now only seek to include $S(\tau)$ in an ellipse $K(\epsilon, \bar{\mu}, \bar{L})$. First, we show that we can focus on two points, i.e. we prove that if $(1+iq(\tau))(\mu+\tau\gamma^2)$ and $(1+iq(\tau))(L+\tau L^2)$ belong to $K(\epsilon, \bar{\mu}, \bar{L})$, then $S(\tau) \subset K(\epsilon, \bar{\mu}, \bar{L})$.

We have that $S(\tau) \cap \{\Re z \geq 0\}$ is a trapezoid, the convex hull of the four points

$$(1+iq(\tau))(\mu+\tau\gamma^2)\,; \quad (1-iq(\tau)(\mu+\tau\gamma^2)\,; \quad (1+iq(\tau))(L+\tau L^2)\,; \quad (1-iq(\tau))(L+\tau L^2).$$

As $K(\epsilon, \bar{\mu}, \bar{L})$ is convex, we only need to show that these four points belong to the ellipsis. By horizontal symmetry, we can restrict our analysis to the two points $(1+iq(\tau))(\mu+\tau\gamma^2)$ and $(1+iq(\tau))(L+\tau L^2)$.

Therefore, we focus on choosing a symmetric ellipse $K(\epsilon, \bar{\mu}, \bar{L})$ to which $(1+iq(\tau))(\mu+\tau\gamma^2)$ and $(1+iq(\tau))(L+\tau L^2)$ belong. The construction of this ellipse is similar to the one of the proof of Prop. 7. Since $(1+iq(\tau))(L+\tau L^2)$ is the farthest point from the real axis, to be able to choose ϵ as small as possible, we put the center of the ellipse at $(L+\tau L^2)$. This way, we force $\frac{\bar{L}+\bar{\mu}}{2} = L+\tau L^2$.

To make sure $(1+iq(\tau))(\mu+\tau\gamma^2)$ is also in the ellipsis, we need to choose $\bar{\mu}$ small enough. Define $\bar{\mu} = \frac{\mu+\tau\gamma^2}{m}$ with $m \geq 2$. This fixes the value of $\bar{L}$ as,

$$\bar{L} = 2(L+\tau L^2) - \bar{\mu} = 2(L+\tau L^2) - \frac{\mu+\tau\gamma^2}{m}\,.$$

We now take $\epsilon > 0$ such that $(1+iq(\tau))(L+\tau L^2)$ is in the ellipsis. We thus have the condition $\epsilon \geq q(\tau)(L+\tau L^2)$. The precise value of ϵ will be chosen later.

We must now check that $(1+iq(\tau))(\mu+\tau\gamma^2) \in K(\epsilon, \bar{\mu}, \bar{L})$. For this we check that this point satisfies the equation of the ellipsis,

$$\left(\frac{\mu+\tau\gamma^2 - \frac{\bar{\mu}+\bar{L}}{2}}{\frac{\bar{L}-\bar{\mu}}{2}}\right)^2 + q(\tau)^2\left(\frac{\mu+\tau\gamma^2}{\epsilon}\right)^2 \tag{52}$$

$$= \left(\frac{\mu+\tau\gamma^2 - (L+\tau L^2)}{L+\tau L^2 - \frac{\mu+\tau\gamma^2}{m}}\right)^2 + q(\tau)^2\left(\frac{\mu+\tau\gamma^2}{\epsilon}\right)^2 \tag{53}$$

$$\leq \left(\frac{\mu+\tau\gamma^2 - (L+\tau L^2)}{L+\tau L^2 - \frac{\mu+\tau\gamma^2}{m}}\right)^2 + \left(\frac{\mu+\tau\gamma^2}{L+\tau L^2}\right)^2\,, \tag{54}$$

by the choice of ϵ. Now let $x = \frac{\mu+\tau\gamma^2}{L+\tau L^2} \in [0,1]$. We have,

$$\begin{aligned}\left(\frac{\mu+\tau\gamma^2 - \frac{\bar{\mu}+\bar{L}}{2}}{\frac{\bar{L}-\bar{\mu}}{2}}\right)^2 + q(\tau)^2\left(\frac{\mu+\tau\gamma^2}{\epsilon}\right)^2 &\leq \left(\frac{1-x}{1-\frac{x}{m}}\right)^2 + x^2 \\ &\leq \left(\frac{1-x}{1-\frac{x}{m}}\right)^2 + x \\ &\leq 1\,,\end{aligned}$$

by application of Lem. 8 as $m \geq 2$.

We now fix ϵ. We want to take $\epsilon = \sqrt{\bar{\mu}\bar{L}}$ so we can apply Prop. 9. Thus we need $q(\tau)(L+\tau L^2) \leq \sqrt{\bar{\mu}\bar{L}}$. Substituting for $\bar{\mu}$ and $\bar{L}$, as $\bar{L} = 2(L+\tau L^2) - (\mu+\tau\gamma^2)/2 \geq \frac{3}{2}(L+\tau L^2)$, this is implied by $q(\tau) \leq \sqrt{\frac{3}{2}}\sqrt{\frac{\mu+\tau\gamma^2}{L+\tau L^2}}$. Now, if $\tau\gamma^2 \geq \mu$, we can apply Lem. 10 and obtain the bound $q(\tau) \leq \frac{\mu+\tau\gamma^2}{L+\tau L^2}$. Hence, if

$$\frac{\gamma}{\mu+\tau\gamma^2} \leq \sqrt{\frac{3}{2}}\sqrt{\frac{\mu+\tau\gamma^2}{L+\tau L^2}}\,, \tag{55}$$

we can apply Prop. 9 with $\epsilon = \sqrt{\bar{\mu}\bar{L}}$. Hence, one can choose $\alpha, \beta > 0$ such that,

$$\begin{aligned}\rho(\mathbf{J}_{V^{\text{mom+cons.}}}(\omega^*, \omega^*))) &\leq 1 - 2(\sqrt{2}-1)\sqrt{\frac{\bar{\mu}}{\bar{L}}} + \mathcal{O}\left(\frac{\bar{\mu}}{\bar{L}}\right), \\ &= 1 - 2(\sqrt{2}-1)\sqrt{\frac{\mu+\tau\gamma^2}{2m(L+\tau L^2) - (\mu+\tau\gamma^2)}} + \mathcal{O}\left(\frac{\mu+\tau\gamma^2}{L+\tau L^2}\right), \\ &\leq 1 - (\sqrt{2}-1)\sqrt{\frac{\mu+\tau\gamma^2}{L+\tau L^2}} + \mathcal{O}\left(\frac{\mu+\tau\gamma^2}{L+\tau L^2}\right),\end{aligned}$$

as $m = 2$. This yields the first part of the proposition. We now need to find an admissible τ.

Assume $\tau L \geq 1$ and so $L \leq \tau L^2$. Then,

$$\begin{aligned}
&\frac{\gamma}{\mu+\tau\gamma^2} \leq \sqrt{\frac{3}{2}}\sqrt{\frac{\mu+\tau\gamma^2}{L+\tau L^2}},\\
\Longleftarrow\ &\frac{\gamma}{\mu+\tau\gamma^2} \leq \frac{\sqrt{3}}{2}\sqrt{\frac{\mu+\tau\gamma^2}{\tau L^2}},\\
\Longleftrightarrow\ &\frac{\gamma}{\mu+\tau\gamma^2} \leq \frac{\sqrt{3}}{2L}\sqrt{\frac{\mu+\tau\gamma^2}{\tau}},\\
\Longleftrightarrow\ &\frac{r}{\mu+\tau\gamma^2} \leq \frac{\sqrt{3}}{2L}\sqrt{\frac{\mu+\tau\gamma^2}{\tau\gamma^2}},\\
\Longleftarrow\ &\frac{1}{\mu+\tau\gamma^2} \leq \frac{\sqrt{3}}{2L}\sqrt{\frac{\mu+\tau\gamma^2}{\mu+\tau\gamma^2}},\\
\Longleftrightarrow\ &\frac{1}{\mu+\tau\gamma^2} \leq \frac{\sqrt{3}}{2L}\,.
\end{aligned}$$

After rearranging, we get that the last condition is equivalent to,

$$\tau \geq \frac{\frac{2}{\sqrt{3}}L-\mu}{\gamma^2}\,,$$

which is implied by $\tau \geq \frac{L}{\gamma^2}$.

Then, if $\tau \geq \frac{L}{\gamma^2}$, $\tau L \geq 1$ and $\tau\gamma^2 \geq \mu$ and so this condition implies $q(\tau)(L+\tau L^2) \leq \sqrt{\mu L}$, which is what we wanted.

Then, for $\tau = \frac{L}{\gamma^2}$, we have

$$\begin{aligned}
\frac{\mu+\tau\gamma^2}{L+\tau L^2} &= \frac{\gamma^2}{L^2}\frac{1+\mu/L}{1+\gamma^2/L^2}\\
&\geq \frac{\gamma^2}{L^2}\frac{1}{1+\gamma^2/L^2}\\
&= \frac{\gamma^2}{L^2} + \mathcal{O}\left(\frac{\gamma^4}{L^4}\right)
\end{aligned}$$

and also in particular $\mathcal{O}\left(\frac{\mu+\tau\gamma^2}{L+\tau L^2}\right) = \mathcal{O}\left(\frac{\gamma^2}{L^2}\right)$. Hence,

$$\rho(\mathbf{J}_{V^{\text{mom+cons.}}}(\omega^*,\omega^*))) \leq 1 - (\sqrt{2}-1)\frac{\gamma}{L} + \mathcal{O}\left(\frac{\gamma^2}{L^2}\right)\,.$$

□

Remark 3. Note that, this rate is roughly similar to the one that can be obtained with the standard momentum method applied to minimizing the objective

$$f(\omega) = \frac{1}{2}\|F\|^2\,.$$

Indeed, one can check, at a stationary point ω^*, the eigenvalues of of the Hessian of f are in $[\gamma^2, L^2]$ (with the notations of the previous proposition). So applying Thm. 4 would yield a local convergence rate of $\mathcal{O}\left(\left(1-\frac{2\gamma}{L+\gamma}\right)^t\right)$.

One could then wonder what is the advantage of Consensus Optimization over the latter. Actually a plain gradient descent on $\frac{1}{2}\|F\|^2$ does not behave well in practice unlike Consensus Optimization (Mescheder et al., 2017) and can be attracted to unstable equilibria in non-monotone landscapes (Letcher et al., 2019).

Remark 4. Though this is not the focus of this paper, similarly to the result of Abernethy et al. (2019) in the non-accelerated case, taking τ slightly higher, such as $\tau = \frac{2L}{\gamma^2}$, guarantees this same accelerated rate even in non-monotone setting. Indeed, all we need is that $\min_{\lambda \in \mathrm{Sp}\, \mathbf{J}_F(\omega^*)} \Re\lambda + \tau\gamma^2 > 0$, which is always satisfied by $\tau = \frac{2L}{\gamma^2}$ as the eigenvalues of $\mathbf{J}_F(\omega^*)$ are bounded by L.

D Ellipses

D.1 Main results

We recall the definition of the ellipses which interests us. Define, for $a, b, c \geq 0$, the ellipse:

$$E(a,b,c) = \left\{ \lambda \in \mathbb{C} : \frac{(\Re\lambda - c)^2}{a^2} + \frac{(\Im\lambda)^2}{b^2} \leq 1 \right\} . \tag{56}$$

We adopt the convention that $\frac{0}{0} = 0$ so that for $b = 0$ the ellipse $E(a,b,c)$ degenerates into a real segment.

We now need to define objects related to the momentum method, and in particular its *ρ-convergence region*. For $\alpha, \rho \geq 0$, $\beta \in \mathbb{R}$, define

$$S(\alpha,\beta,\rho) = \{\lambda \in \mathbb{C} : \forall z \in \mathbb{C},\ z^2 - (1 - \alpha\lambda + \beta)z + \beta = 0 \implies |z| \leq \rho\} . \tag{57}$$

We call it the ρ-convergence region of the momentum method as it corresponds to the maximal regions of the complex plane where the momentum method converges at speed $\mathcal{O}(\rho^t)$ if the operator has its eigenvalues in this zone. This is formalized by the following lemma,

Lemma 11 (Saul'yev (1964, II.7), Polyak (1964), Gidel et al. (2019b, Thm. 3))**.** *Denote the momentum operator applied to the vector field F by*

$$V(\omega, \omega') = (\omega - \alpha F(\omega) + \beta(\omega - \omega'), \omega') \tag{58}$$

with $\alpha \geq 0$ step size and $\beta \in \mathbb{R}$ momentum parameter. Then, for any $\rho \geq 0$,

$$\rho(\nabla V(\omega^*, \omega^*)) \leq \rho \tag{59}$$

if and only if $\operatorname{Sp} \nabla F(\omega^*) \subset S(\alpha,\beta,\rho)$.

For a proof of this lemma in the context of games, see the proof of Thm. 3 of Gidel et al. (2019b).

The next is a geometrical characterization of $S(\alpha,\beta,\rho)$: this is an ellipse, which is described in the following lemma.

Lemma 12 (Niethammer and Varga (1983, Cor. 6))**.** *If $|\beta| > \rho^2$, $S(\alpha,\beta,\rho) = \emptyset$. Otherwise, if $|\beta| \leq \rho^2$ and $\rho > 0$,*

$$S(\alpha,\beta,\rho) = \left\{ \lambda \in \mathbb{C} : \frac{(1 - \alpha\Re\lambda + \beta)^2}{(1+\tau)^2} + \frac{(\alpha\Im\lambda)^2}{(1-\tau)^2} \leq \rho^2 \right\} , \tag{60}$$

where $\tau = \frac{\beta}{\rho^2}$.

As indicated, this lemma is a consequence of the results of Niethammer and Varga (1983), and more exaclty their section §6. However, their notations are significantly different from ours. We give a few elements to help the readers translate their results into our setting. In §6 of Niethammer and Varga (1983), they study iterative methods of the form,

$$\omega_{t+1} = \mu_0(\operatorname{Id} - F(\omega_t)) + \mu_1\omega_t + \mu_2\omega_{t-1} ,$$

with $\mu_0 + \mu_1 + \mu_2 = 1$. Developing and using this relation, their iteration rule becomes,

$$\omega_{t+1} = \omega_t - \mu_0 F(\omega_t) + \mu_2(\omega_{t-1} - \omega_t) .$$

Identifying with (58), we get that $\alpha = \mu_0$, $\beta = -\mu_2$ and so $\mu_1 = 1 + \beta - \alpha$.

Moreover, what they denote by $S_\eta(p)$, where p is a variable encompassing the parameters μ_0,μ_1 and μ_2, actually corresponds to $1 - S(\alpha,\beta,\rho)$ with α, β linked to μ_0,μ_1, μ_2 as described above and $\eta = \frac{1}{\rho}$. Indeed[2], $S_\eta(p)$ is meant to be compared to the eigenvalues of $I_d - \nabla F(\omega^*)$ instead of $\nabla F(\omega^*)$. Hence, the center and the semiaxes of the ellipse $1 - S(\alpha,\beta,\gamma)$ are given by (6.3) of Niethammer and Varga (1983) and once translated in our notations yield Lem. 12.

[2]This is a standard convention in the linear system theory. They consider $\omega = T\omega + c$ instead of $A\omega + b$ as they use splittings of A.

Remark 5. This lemma actually does not require the complex analysis machinery of Niethammer and Varga (1983). This can be proven by hand using this remark on second-order equations. Let $0 < \rho \leq 1$ and let z_1, z_2 denote the two (possibly equal) roots of $X^2 + bX + c$. Then,

$$\max(|z_1|, |z_2|) \leq \rho \iff \begin{cases} |c| \leq \rho \\ |b|^2 + |\Delta| \leq 2\left(\rho^2 + \frac{c^2}{\rho^2}\right) \end{cases},$$

where $\Delta = b^2 - 4c$ denote the discriminant of the equation.

Now, we can introduce one of the main results of Niethammer and Varga (1983). This is an answer to the natural question: what is $\underline{\rho}(S(\alpha, \beta, \rho))$? In particular is it equal to ρ? In other words, is momentum optimal w.r.t. to its convergence sets? The answer is yes for the momentum method. Note however that this doe snot hold for all stationary methods, this is linked to tricky questions of existence of branch for the roots of some polynomial equations, see Nevanlinna (1993, §3.7) for a discussion on this.

Proposition 12 (Niethammer and Varga (1983, Cor. 10))**.** *Assume $|\beta| \leq \rho^2 < 1$ and $\alpha > 0$, then*

$$\underline{\rho}(S(\alpha, \beta, \rho)) = \rho\,. \tag{61}$$

Hence, momentum is optimal for the sets $S(\alpha, \beta, \rho)$. What is left to show is that the sets $S(\alpha, \beta, \gamma)$ can represent most ellipses $E(a, b, c)$.

Proposition 13. *Let $a, b \geq 0$, $c > 0$, $(a, b) \neq (0, 0)$. There exists $\alpha > 0$, $\rho > 0$, $\beta \in (-1, 1]$, with $|\beta| \leq \rho$ such that $E(a, b, c) = S(\alpha, \beta, \rho)$ if and only if $a^2 \leq b^2 + c^2$. If it is the case,*

1. *The triple (α, β, ρ) satisfying such conditions is unique.*
2. *The corresponding β can be written $\beta = \chi(a - b)$ with $\chi > 0$.*
3. *The corresponding ρ is equal to:*
$$\rho = \begin{cases} \frac{a}{c} & \text{if } a = b \\ \frac{c - \sqrt{b^2 + c^2 - a^2}}{a - b} & \text{otherwise.} \end{cases}$$
4. *The parameters $\alpha > 0$ and $\beta \in (-1, 1]$ are given by,*

$$\beta = \begin{cases} 0 & \text{if } a = b \\ 2c\frac{c - \sqrt{c^2 + b^2 - a^2}}{a^2 - b^2} - 1 & \text{otherwise,} \end{cases} \qquad \alpha = \frac{1 + \beta}{c} = \begin{cases} \frac{1}{c} & \text{if } a = b \\ 2\frac{c - \sqrt{c^2 + b^2 - a^2}}{a^2 - b^2} & \text{otherwise,} \end{cases} \tag{62}$$

Proof. Recall these two parametrizations of an ellipse,

$$E(a, b, c) = \left\{\lambda \in \mathbb{C} : \frac{(\Re\lambda - c)^2}{a^2} + \frac{(\Im\lambda)^2}{b^2} \leq 1\right\}$$
$$S(\alpha, \beta, \rho) = \left\{\lambda \in \mathbb{C} : \frac{(1 - \alpha\Re\lambda + \beta)^2}{(1 + \tau)^2} + \frac{(\alpha\Im\lambda)^2}{(1 - \tau)^2} \leq \rho^2\right\},$$

where $\tau = \frac{\beta}{\rho^2}$. Note that $(a, b) \neq (0, 0)$, $E(a, b, c)$ is not reduced to a point. So if these ellipses are equal, $\rho > 0$, and we also have $\alpha > 0$ These ellipses are characterised by their centers and their semiaxes so they are equal if and only if,

$$\begin{cases} \frac{1+\beta}{\alpha} = c \\ \rho + \frac{\beta}{\rho} = \alpha a \\ \rho - \frac{\beta}{\rho} = \alpha b \end{cases} \iff \begin{cases} \frac{1+\beta}{\alpha} = c \\ \rho = \alpha(a + b) \\ \frac{\beta}{\rho} = \alpha(b - a) \end{cases} \iff \begin{cases} \frac{1+\beta}{\alpha} = c \\ \rho = \frac{1+\beta}{2}(\tilde{a} + \tilde{b}) \\ \frac{\beta}{\rho} = \frac{1+\beta}{2}(\tilde{a} - \tilde{b}), \end{cases} \tag{63}$$

where $\tilde{a} = \frac{a}{c}$ and $\tilde{b} = \frac{b}{c}$. We further let $\tilde{\beta} = 1 + \beta$. Then, the last two equations imply the following equation on β,

$$\beta = \frac{(1 + \beta)^2}{4}(\tilde{a}^2 - \tilde{b}^2) \iff \tilde{\beta} - 1 = \frac{\tilde{\beta}^2}{4}(\tilde{a}^2 - \tilde{b}^2)\,. \tag{64}$$

Its discriminant is $\Delta = 1 - (\tilde{a}^2 - \tilde{b}^2)$, which is non-negative if and only if $b^2 + c^2 \geq a^2$.

Before solving this equation, we briefly discuss when it degenerates into a degree one equation. Indeed, if $a = b$, and so $\tilde{a} = \tilde{b}$, the unique solution of (64) is $\tilde{\beta} = 1$ and so $\beta = 0$. Moreover $\rho = \frac{\tilde{a}+\tilde{b}}{2} = \frac{a}{c}$.

We now assume that $\tilde{a}^2 - \tilde{b}^2 \neq 0$. The two solutions of (64) are,

$$\tilde{\beta}_\pm = 1 + \beta_\pm = 2\frac{1 \pm \sqrt{\Delta}}{\tilde{a}^2 - \tilde{b}^2}\,.$$

We distinguish three cases.

- If $\Delta = 0$. There is only one solution $\tilde{\beta} = 1 + \beta = 2$ to (64) and so $\beta = 1$.
- If $0 < \Delta < 1$ then in particular $\tilde{a} > \tilde{b}$. As $0 < \Delta < 1$, we also have $0 < \tilde{a}^2 - \tilde{b}^2 < 1$. This implies that $\tilde{\beta}_+ > 2(1 + \sqrt{\Delta}) > 2$ and so $\beta_+ > 1$ which do not satisfy the desired conditions on β. We show that β_- satisfy them instead. As $\Delta < 1$, $\tilde{\beta}_- > 0$. Moreover, $\sqrt{\Delta} \geq \Delta$ and so $\tilde{\beta}_- \leq 2\frac{1-\Delta}{\tilde{a}^2-\tilde{b}^2} = 2$ and so $\beta_- \in (-1, 1]$.
- If $\Delta > 1$ and so $\tilde{a} < \tilde{b}$. One has immediately that $\tilde{\beta}_+ < 0$ which disqualifies β_+. On the contrary as $\Delta > 1$, $\tilde{\beta}_- > 0$ and $\tilde{\beta}_- = 2\frac{\sqrt{1+\tilde{b}^2-\tilde{a}^2}-1}{\tilde{b}^2-\tilde{a}^2} \leq 2\frac{1+\sqrt{\tilde{b}^2-\tilde{a}^2}-1}{\tilde{b}^2-\tilde{a}^2} = 2$. And so $\beta_- \in (-1, 1]$.

Note that the case $\Delta = 1$ is prevented by the assumption $\tilde{a}^2 - \tilde{b}^2 \neq 0$.

In each of the three cases above we ended up with,

$$\beta = \beta_- = \tilde{\beta}_- - 1 = 2\frac{1 - \sqrt{1 + \tilde{b}^2 - \tilde{a}^2}}{\tilde{a}^2 - \tilde{b}^2} - 1 = 2c\frac{c - \sqrt{c^2 + b^2 - a^2}}{a^2 - b^2} - 1 \in (-1, 1]\,.$$

Note that the third equation of (63) easily gives that $\beta = \chi(a - b)$ with $\chi > 0$. We now define ρ with the second equation of (63),

$$\rho = \frac{1 + \beta}{2}(\tilde{a} + \tilde{b})\,.$$

As β satisfy (64), β and ρ also satisfy the third one of (63). α can then be defined by the first equation of (63). Finally note that the fact that $|\beta| \leq \rho^2$ comes from the combination of the second and the third equations of (63). □

Note that if $0 \notin E(a, b, c)$, then $c^2 > a^2$ and so the hypothesis of the proposition above is satisfied. Thm. 6 is now proven by simply combining all the results in this subsection.

D.2 Proof of optimality of momentum on its convergence zones

For this proof, we will need a characterization of $\underline{\rho}(K)$ using Green functions. We will follow the presentation of Nevanlinna (1993).

Definition 2. The Green function (with pole at ∞) of a non-empty, connected, unbounded open set $\Omega \subset \mathbb{C}$ is the unique function $g : \Omega \to \mathbb{R}$ such that:

1. g is harmonic on Ω.
2. $g(z) = \log|z| + \mathcal{O}(1)$ as $|z| \to \infty$.
3. $g(z) \xrightarrow[z \to \zeta]{} 0$ for every $\zeta \in \partial\Omega$.

For a compact $K \subset \mathbb{C}$, denote by G_∞ the unbounded connected component of $\bar{\mathbb{C}} \setminus K$. $\underline{\rho}(K)$ can then be obtained from the Green function of G_∞. This is not our concern here, but note that the Green function of G_∞ is guaranteed to exist if its boundary is sufficiently nice (see for instance Walsh (1935); Ransford (1995) for a thorough treatment of this classical question).

The following theorem is a deep result in complex analysis, which links the minimization problem over polynomial which defines $\underline{\rho}$ to the geometric properties of K through its Green function.

Theorem 7 (Nevanlinna (1993, Prop. 3.4.6, Thm. 3.4.9))**.** *If G_∞ has a Green function g and if $0 \in G_\infty$,*

$$\underline{\rho}(K) = \exp(-g(0))\,. \tag{65}$$

We will also need the following complex analysis lemma about the Joukowsky map, see Nehari (1952, Chap. VI) for instance.

Lemma 13. *Let $\psi(z) = z + \frac{1}{z}$. Then $\psi : \bar{\mathbb{C}} \setminus \{z : |z| \leq 1\} \to \bar{\mathbb{C}} \setminus [-1, 1]$ is a conformal mapping. Its inverse ϕ is characterized by: for any $z_0 \notin [-1, 1]$, $\phi(z_0)$ is the unique solution of*

$$z^2 - 2zz_0 + 1 = 0 \tag{66}$$

outside $\{z : |z| \leq 1\}$. Moreover, $\phi(z) = 2z + \mathcal{O}(1)$ when $z \to \infty$.

First we begin with a simple lemma about the convergence zones of momentum.

Lemma 14. *If $\rho^2 < 1$, $0 \notin S(\alpha, \beta, \rho)$.*

Proof. Consider the equation

$$z^2 - (1 + \beta)z + \beta = 0\,. \tag{67}$$

Its two roots are β and 1 which yields the result. □

As the boundary of the set plays a special role in the definition of the Green function, we need to have a precise characterization of it. This is done through the the next two lemmas.

Lemma 15. *If $0 < |\beta| \leq \rho^2$, then*

$$\mathrm{int}(S(\alpha, \beta, \rho)) = \bigcup_{\rho' > 0 : |\beta| < \rho' < \rho} S(\alpha, \beta, \rho')\,. \tag{68}$$

Proof. The functions $x \mapsto x \pm \frac{\beta}{x}$ are increasing positive on $]\sqrt{|\beta|}, +\infty[$. So their square is also increasing. By Lem. 12,

$$\mathrm{int}\, S(\alpha, \beta, \rho) = \left\{ \lambda \in \mathbb{C} : \frac{(1 - \alpha\Re\lambda + \beta)^2}{(1 + \tau)^2} + \frac{(\alpha\Im\lambda)^2}{(1 - \tau)^2} < \rho^2 \right\}\,. \tag{69}$$

Define for $x > \sqrt{|\beta|}$ the function $h_\lambda(x) = \frac{(1-\alpha\Re\lambda+\beta)^2}{(x+\frac{\beta}{x})^2} + \frac{(\alpha\Im\lambda)^2}{(x-\frac{\beta}{x})^2}$, which is continuous and non-increasing. We show the result by double inclusion.

- Let $\lambda \in \mathrm{int}\, S(\alpha, \beta, \rho)$. As $\rho > 0$, $h_\lambda(\rho) < 1$ by (69). As $\rho > \sqrt{|\beta|}$, by continuity of h_λ at ρ, there exists $\rho > \rho' > \sqrt{|\beta|}$ such that $h_\lambda(\rho') < 1$. As $\rho' > \sqrt{|\beta|} \geq 0$, this implies that $\lambda \in S(\alpha, \beta, \rho')$.
- Let $\rho' > 0$ such that $|\beta| < \rho' < \rho$ and take $\lambda \in S(\alpha, \beta, \rho')$. By Lem. 12, as $\rho' > 0$, this implies that $h_\lambda(\rho') \leq 1$. Note that if both $\Im\lambda = 0$ and $1 - \alpha\Re\lambda + \beta = 0$, $\lambda \in \mathrm{int}\, S(\alpha, \beta, \rho)$ as $\rho > 0$. Otherwise, if at least one of them is non-zero, this means that h_λ is actually decreasing on $]\sqrt{|\beta|}, +\infty[$. Hence, $h_\lambda(\rho) < h_\lambda(\rho') \leq 1$ and so $\lambda \in \mathrm{int}\, S(\alpha, \beta, \rho)$.

□

Lemma 16. *If $0 < |\beta| \leq \rho^2$,*

$$\partial S(\alpha, \beta, \rho) = S(\alpha, \beta, \rho) \cap \{\lambda \in \mathbb{C} : \exists z \in \mathbb{C},\ z^2 - (1 - \alpha\lambda + \beta)z + \beta = 0 \text{ and } |z| = \rho\} \tag{70}$$

Proof. This is a direct consequence of Lem. 15 and the definition of $S(\alpha, \beta, \rho)$. □

Lemma 17. *For $0 < \beta < \rho^2$,*

$$\{\lambda \in \mathbb{R} : (1 - \alpha\lambda + \beta)^2 \leq 4\beta\} \subset \mathrm{int}(S(\alpha, \beta, \lambda)) \tag{71}$$

For $0 < -\beta < \rho^2$,

$$\{\lambda \in \mathbb{C} : 1 - \alpha\Re\lambda + \beta = 0,\ (\alpha\Im\lambda)^2 \leq 4|\beta|\} \subset \mathrm{int}(S(\alpha, \beta, \lambda)) \tag{72}$$

Proof. First assume that $0 < \beta < \rho^2$. Consider $\lambda \in \mathbb{R}$ such that $(1 - \alpha\lambda + \beta)^2 \leq 4\beta$. Using the characterization of Lem. 12, we only need to show that $(1 - \alpha\lambda + \beta)^2 < \rho^2(1 + \tau)^2$ (as $-\rho^2 < \beta \implies \tau \neq -1$). But, from the definition of τ we get

$$\rho^2(1 + \tau)^2 - 4\beta = (\rho - \frac{\beta}{\rho})^2 > 0. \tag{73}$$

Hence $4\beta < \rho^2(1 + \tau)^2$ and the result follows from the choice of λ.

The proof for the second point is similar.

□

We can now prove the proposition which was the target of this subsection. Note that the following proof does not encompass the case particular $\rho^2 = |\beta|$ in which the ellipse is degenerate. This falls into the case of segments, which is much simpler, see the aforementioned references (or Nevanlinna (1993) for a didactic explanation).

Proposition 14. *Assume $0 < |\beta| < \rho^2 < 1$ and $\alpha > 0$, then*

$$\underline{\rho}(S(\alpha, \beta, \rho)) = \rho \tag{74}$$

Proof. We will build the Green function for $\bar{\mathbb{C}} \setminus S(\alpha, \beta, \rho)$ using Lem. 13.

First, we show that if $\lambda \notin \operatorname{int} S(\alpha, \beta, \rho)$, then $\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}} \notin [-1, 1]$ where $\sqrt{\beta}$ is a square root (with positive real part) of β. Indeed, assume for the sake of contradiction that it is not the case, i.e. there exists $\lambda \notin \operatorname{int} S(\alpha, \beta, \rho)$ such that $\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}} \in [-1, 1]$. Assume first that $\beta > 0$. This implies that $\Im(1 - \alpha\lambda + \beta) = 0$ and so that $\lambda \in \mathbb{R}$ as $\alpha \neq 0$. Moreover, as $\beta > 0$, λ satisfies $(1 - \alpha\lambda + \beta)^2 \leq 4\beta$. By Lem. 17, $\lambda \in \operatorname{int} S(\alpha, \beta, \rho)$ which is a contradiction. If $\beta < 0$, $\sqrt{\beta} = i\sqrt{|\beta|}$. This implies that $\Re(1 - \eta\lambda + \beta) = 0$. Moreover, λ satisfies $(\Im(1 - \alpha\lambda + \beta))^2 \leq 4|\beta|$. We get a similar contradiction using Lem. 17.

Take $\lambda \notin \operatorname{int} S(\alpha, \beta, \rho)$. Then, as $\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}} \notin [-1, 1]$, we can consider $\phi\left(\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}}\right)$. By Lem. 13, $\phi\left(\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}}\right)$ is the unique solution of modulus (strictly) greater than one of

$$z^2 - 2\frac{1 - \alpha\lambda + \beta}{2\sqrt{\beta}}z + 1 = 0\,. \tag{75}$$

Hence $\sqrt{\beta}\phi\left(\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}}\right)$ is the unique solution of modulus (strictly) greater than $\sqrt{|\beta|}$ of

$$\frac{z^2}{\beta} - 2\frac{1 - \alpha\lambda + \beta}{2\beta}z + 1 = 0 \tag{76}$$

$$\iff z^2 - (1 - \eta\lambda + \beta)z + \beta = 0\,. \tag{77}$$

Let $z_1 = \sqrt{\beta}\phi\left(\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}}\right)$ and let z_2 be the other root of (77). Then $z_1z_2 = \beta$ and so $|z_1z_2| = |\beta|$. Hence, as $|z_1| > \sqrt{|\beta|}$, we have $|z_2| < \sqrt{|\beta|}$. Hence $\sqrt{\beta}\phi\left(\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}}\right)$ is the solution of greatest magnitude of (77). We will see that this quantity is very regular as a function of λ outside $S(\alpha, \beta, \rho)$. Define

$$\chi : \begin{cases} \mathbb{C} \setminus \operatorname{int} S(\alpha, \beta, \rho) \longrightarrow \mathbb{C} \\ \qquad\qquad \lambda \qquad\quad \longmapsto \sqrt{\beta}\phi\left(\dfrac{1 - \alpha\lambda + \beta}{2\sqrt{\beta}}\right)\,. \end{cases} \tag{78}$$

We can now build our Green function using χ. Define,

$$g : \begin{cases} \mathbb{C} \setminus \operatorname{int} S(\alpha, \beta, \rho) \longrightarrow \mathbb{R} \\ \qquad\qquad \lambda \qquad\quad \longmapsto \log \dfrac{|\chi(\lambda)|}{\rho}\,. \end{cases} \tag{79}$$

Note that as ϕ is continuous on its domain of definition χ is continuous too. Moreover, as $\beta \neq 0$ and $\chi(\lambda)$ is a root of (77), $\chi(\lambda) \neq 0$ for $\lambda \notin \operatorname{int} S(\alpha, \beta, \rho)$. Hence g is well-defined and continuous too on $\mathbb{C} \setminus \operatorname{int} S(\alpha, \beta, \rho)$. We now show that g is the Green function of $G_\infty = \mathbb{C} \setminus S(\alpha, \beta, \gamma)$ according to Definition 2.

1. By Lem. 13, ϕ is analytic and so is χ on the open set $\mathbb{C} \setminus S(\alpha, \beta, \rho)$. Moreover, as mentioned above, $\chi(\lambda) \neq 0$ for $\lambda \notin S(\alpha, \beta, \rho)$ Hence, g is harmonic on $\mathbb{C} \setminus S(\alpha, \beta, \rho) = G_\infty$.

2. When $\lambda \to \infty$, $\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}} \to \infty$ too as $\alpha \neq 0$. Hence, by Lem. 13,

$$
\begin{aligned}
g(\lambda) &= \log \frac{|\chi(\lambda)|}{\rho} && (80)\\
&= \log \left| \phi\left(\frac{1-\alpha\lambda+\beta}{2\sqrt{\beta}} \right) \right| + \mathcal{O}(1) && (81)\\
&= \log |\lambda + \mathcal{O}(1)| + \mathcal{O}(1) && (82)\\
&= \log |\lambda| + \mathcal{O}(1)\,. && (83)\\
& && (84)
\end{aligned}
$$

3. Let $\zeta \in \partial\left(\mathbb{C} \setminus S(\alpha, \beta, \gamma)\right) = \partial S(\alpha, \beta, \gamma)$. Note that χ is defined on $\mathbb{C} \setminus \operatorname{int} S(\alpha, \beta, \rho)$ on so on $\partial S(\alpha, \beta, \gamma)$. Then, by Lem. 16 and the definition of χ, $|\chi(\zeta)| = \rho$. By continuity of g, $g(\lambda) \xrightarrow[\lambda \to \zeta]{} g(\zeta) = 0$. Hence g is the Green function for G_∞ by Definition 2. Moreover, by Lem. 14, $0 \in G_\infty$. We can now apply Thm. 7 to get that $\underline{\rho}(S(\alpha, \beta, \rho)) = \exp(-g(0))$. Finally, we compute $g(0)$. Recall that, as $0 \notin S(\alpha, \beta, \rho)$, $\chi(0)$ is the root of greatest magnitude of

$$z^2 - (1+\beta)z + \beta = 0\,. \tag{85}$$

The two roots of this equation are β and 1. As $0 < |\beta| < 1$, $\chi(0) = 1$, so $g(0) = \log \frac{1}{\rho}$ and $\underline{\rho}(S(\alpha, \beta, \rho)) = \rho$.

□

E Synthetic Experiments

In this section we evaluate the accelerated methods we studied on synthetic bilinear games.

We consider bilinear games of the form,

$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^m} (x - x^*)^\top A(y - y^*)\,.$$

A, x^*, y^* and the initial points are chosen randomly. More precisely, each of their coefficients is drawn from a standard normal distribution and A is normalized such that $\frac{\sigma_{max}(A)}{\sigma_{min}(A)} = 100$. The total dimension of the parameter space is $d = 2m$.

We compare the accelerated methods we presented to methods which are proved to converge on such games: EG, Hamiltonian gradient descent (HGD) (Abernethy et al., 2019), the alternating gradient method with negative momentum (Gidel et al., 2019b) and optimistic mirror descent (OMD) (Daskalakis et al., 2018).

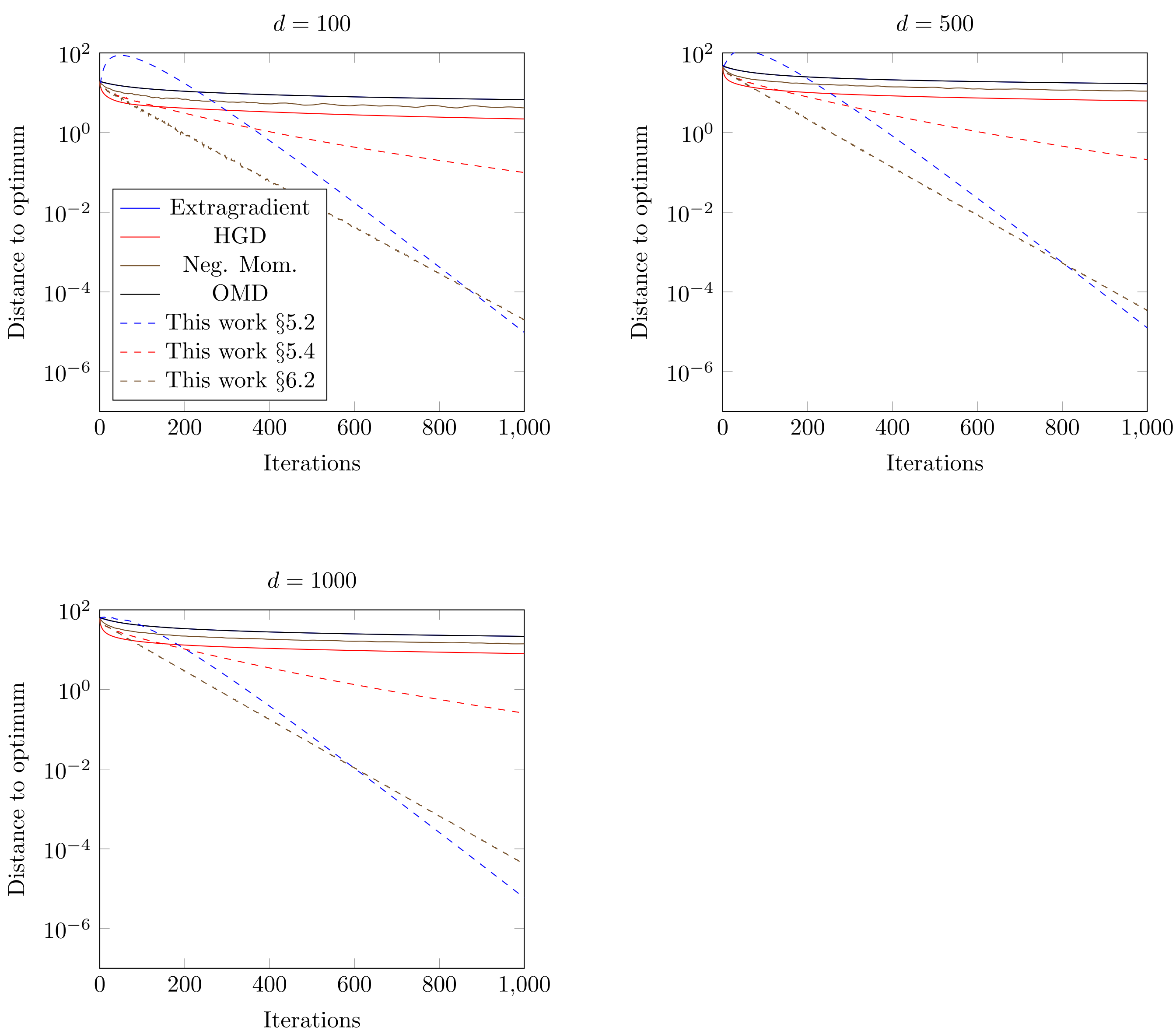

Figure 4: Distance to the optimum as a function of the number of iterations of the methods.