A Tight and Unified Analysis of Gradient-Based Methods for a Whole Spectrum of Differentiable Games

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

Abstract

We consider differentiable games where the goal is to find a Nash equilibrium. The machine learning community has recently started using variants of the gradient method (GD). Prime examples are extragradient (EG), the optimistic gradient method (OG) and consensus optimization (CO), which enjoy linear convergence in cases like bilinear games, where the standard GD fails. The full benefits of theses relatively new methods are not known as there is no unified analysis for both strongly monotone and bilinear games. We provide new analyses of the EG's local and global convergence properties and use is to get a tighter global convergence rate for OG and CO. Our analysis covers the whole range of settings between bilinear and strongly monotone games. It reveals that these methods converges via different mechanisms at these extremes; in between, it exploits the most favorable mechanism for the given problem. We then prove that EG achieves the optimal rate for a wide class of algorithms with any number of extrapolations. Our tight analysis of EG's convergence rate in games shows that, unlike in convex minimization, EG may be much faster than GD.

1 Introduction

Gradient-based optimization methods have underpinned many of the recent successes of machine learning. The training of many models is indeed formulated as the minimization of a loss involving the data. However, a growing number of frameworks rely on optimization problems that involve multiple players and objectives. For instance, actor-critic models (Pfau and Vinyals, 2016), generative adversarial networks (GANs) (Goodfellow et al., 2014) and automatic curricula (Sukhbaatar et al., 2018) can be cast as two-player games.

Hence games are a generalization of the standard single-objective framework. The aim of the optimization is to find *Nash equilibria*, that is to say situations where no player can unilaterally decrease their loss. However, new issues that were not present for single-objective problems arise. The presence of rotational dynamics prevent standard algorithms such as the gradient method to converge on simple bilinear examples (Goodfellow, 2016; Balduzzi et al., 2018). Furthermore, stationary points of the gradient dynamics are not necessarily Nash equilibria (Adolphs et al., 2019; Mazumdar et al., 2019).

Some recent progress has been made by introducing new methods specifically designed with games or variational inequalities in mind. The main example are the optimistic gradient method (OG) introduced by Rakhlin and Sridharan (2013) initially for online learning, consensus optimization (CO) which adds a regularization term to the optimization problem and the extragradient method (EG) originally introduced by Korpelevich (1976). Though these news methods and the gradient method (GD) have similar performance in convex optimization, their behaviour seems to differ when applied to games: unlike gradient, they converge on the so-called bilinear example (Tseng, 1995; Gidel et al., 2019a; Mokhtari et al., 2019; Abernethy et al., 2019).

However, linear convergence results for EG and OG (a.k.a extrapolation from the past) in particular have only been proven for either strongly monotone variational inequalities problems, which include strongly convex-concave saddle point problems, or in the bilinear setting separately (Tseng, 1995; Gidel et al., 2019a; Mokhtari et al., 2019).

In this paper, we study the dynamics of such gradient-

†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.

based methods and in particular GD, EG and more generally multi-step extrapolations methods for unconstrained games. Our objective is three-fold. First, taking inspiration from the analysis of GD by Gidel et al. (2019b), we aim at providing a single precise analysis of EG which covers both the bilinear and the strongly monotone settings and their intermediate cases. Second, we are interested in theoretically comparing EG to GD and general multi-step extrapolations through upper and lower bounds on convergence rates. Third, we provide a framework to extend the unifying results of spectral analysis in global guarantees and leverage it to prove tighter convergence rates for OG and CO. Our contributions can be summarized as follows:

- We perform a spectral analysis of EG in §5. We derive a local rate of convergence which covers the whole range of settings between purely bilinear and strongly monotone games and which is faster than existing rates in some regimes. Our analysis also encompasses multi-step extrapolation methods and highlights the similarity between EG and the proximal point methods.
- We use and extend the framework from Arjevani et al. (2016) to derive lower bounds for specific classes of algorithms. (i) We show in §4 that the previous spectral analysis of GD by Gidel et al. (2019b) is tight, confirming the difference of behaviors with EG. (ii) We prove lower bounds for 1-Stationary Canonical Linear Iterative methods with any number of extrapolation steps in §5. As expected, this shows that increasing this number or choosing different step sizes for each does not yield significant improvements and hence EG can be considered as optimal among this class.
- In §6, we derive a global convergence rate for the EG with the same unifying properties as the local analysis. We then leverage our approach to derive global convergence guarantees for OG and CO with similar unifying properies. It shows that, while these methods converges for different reasons in the convex and bilinear settings, in between they actually take advantage of the most favorable one.

2 Related Work

Extragradient was first introduced by Korpelevich (1976) in the context of variational inequalities. Tseng (1995) proves results which induce linear convergence rates for this method in the bilinear and strongly monotone cases. We recover both rates with our analysis. The extragradient method was generalized to arbitrary geometries by Nemirovski (2004) as the mirror-prox method. A sublinear rate of $\mathcal{O}(1/t)$ was proven for monotone variational inequalities by treating this method as an approximation of the proximal point method as we will discuss later. More recently, Mertikopoulos et al. (2019) proved that, for a broad class of saddle-point problems, its stochastic version converges almost surely to a solution.

	Tseng (1995)	Gidel et al. (2019a)	Mokhtari et al. (2019)	Abernethy et al. (2019)	This work §6
EG	$c\frac{\mu}{L}$	-	$\frac{\mu}{4L}$	-	$\frac{1}{4}(\frac{\mu}{L}+\frac{\gamma^2}{16L^2})$
OG	-	$\frac{\mu}{4L}$	$\frac{\mu}{4L}$	-	$\frac{1}{4}(\frac{\mu}{L}+\frac{\gamma^2}{32L^2})$
CO	-	-	-	$\frac{\gamma^2}{4L_H^2}$	$\frac{\mu^2}{2L_H^2}+\frac{\gamma^2}{2L_H^2}$

Table 1: Summary of the global convergence results presented in §6 for extragradient (EG), optimistic gradient (OMD) and consensus optimization (CO) methods. If a result shows that the iterates converge as $\mathcal{O}((1-r)^t)$, the quantity r is reported (the larger the better). The letter c indicates that the numerical constant was not reported by the authors. μ is the strong monotonicity of the vector field, γ is a global lower bound on the singular values of ∇v , L is the Lipschitz constant of the vector field and L_H^2 the Lipschitz-smoothness of $\frac{1}{2}\|v\|_2^2$. For instance, for the so-called bilinear example (Ex. 1), we have $\mu=0$ and $\gamma=\sigma_{\min}(A)$. Note that for this particular example, previous papers developed a specific analysis that breaks when a small regularization is added (see Ex. 3).

Optimistic gradient method is slightly different from EG and can be seen as a kind of extrapolation from the past (Gidel et al., 2019a). It was initially introduced for online learning (Chiang et al., 2012; Rakhlin and Sridharan, 2013) and subsequently studied in the context of games by Daskalakis et al. (2018), who proved that this method converges on bilinear games. Gidel et al. (2019a) interpreted GANs as a variational inequality problem and derived OG as a variant of EG which avoids "wasting" a gradient. They prove a linear convergence rate for strongly monotone variational inequality problems. Treating EG and OG as perturbations of the proximal point method, Mokhtari et al. (2019) gave new but still separate derivations for the standard linear rates in the bilinear and the strongly convex-concave settings. Liang and Stokes (2019) mentioned the potential impact of the interaction between the players, but they only formally show this on bilinear examples: our results show that this conclusion extends to general nonlinear games.

Consensus optimization has been motivated by the use of gradient penalty objectives for the practical training of GANs (Gulrajani et al., 2017; Mescheder et al., 2017). It has been analysed by Abernethy et al. (2019) as a perturbation of Hamiltonian gradient descent.

We provide a unified and tighter analysis for these three algorithms leading to faster rates (cf. Tab. 1).

Lower bounds in optimization date back to Nemirovsky and Yudin (1983) and were popularized by Nesterov (2004). One issue with these results is that they are either only valid for a finite number of iterations depending on the dimension of the problem or are proven in infinite dimensional spaces. To avoid this issue, Arjevani et al. (2016) introduced a new framework called p-Stationary Canonical Linear Iterative algorithms (p-SCLI). It encompasses methods which, applied on quadratics, compute the next iterate as fixed linear transformation of the p last iterates, for some fixed $p \geq 1$. We build on and extend this framework to derive lower bounds for games for 1-SCLI. Note that sublinear lower bounds have been proven for saddle-point problems by Nemirovsky (1992); Nemirovski (2004); Chen et al. (2014); Ouyang and Xu (2018), but they are outside the scope of this paper since we focus on linear convergence bounds.

Our notation is presented in §A. The proofs can be found in the subsequent appendix sections.

3 Background and motivation

3.1 n-player differentiable games

Following Balduzzi et al. (2018), a *n-player differentiable game* can be defined as a family of twice continuously differentiable losses $l_i : \mathbb{R}^d \to \mathbb{R}$ for $i = 1, \dots, n$. The parameters for player i are $\omega^i \in \mathbb{R}^{d_i}$ and we note $\omega = (\omega^1, \dots, \omega^n) \in \mathbb{R}^d$ with $d = \sum_{i=1}^n d_i$. Ideally, we are interested in finding an *unconstrained Nash equilibrium* (Von Neumann and Morgenstern, 1944): that is to say a point $\omega^* \in \mathbb{R}^d$ such that

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

where the vector $(\omega^{-i})^*$ contains all the coordinates of ω^* except the i^{th} one. Moreover, we say that a game is *zero-sum* if $\sum_{i=1}^n l_i = 0$. For instance, following Mescheder et al. (2017); Gidel et al. (2019b), the standard formulation of GANs from Goodfellow et al. (2014) can be cast as a two-player zero-sum game. The Nash equilibrium corresponds to the desired situation where the generator exactly capture the data distribution, completely confusing a perfect discriminator.

Let us now define the *vector field*

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

associated to a n-player game and its Jacobian:

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

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

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

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

Gidel et al. (2019b) and Balduzzi et al. (2018) mentioned two particular classes of games, which can be seen as the two opposite ends of a spectrum. As the definitions vary, we only give the intuition for these two categories. The first one is *adversarial games*, where the Jacobian has eigenvalues with small real parts and large imaginary parts and the cross terms $\nabla_{\omega_i}\nabla_{\omega_j} l_j(\omega)$, for $i \neq j$, are dominant. Ex. 1 gives a prime example of such game that has been heavily studied: a simple bilinear game whose Jacobian is anti-symmetric and so only has imaginary eigenvalues (see Lem. 7 in App. E):

Example 1 (Bilinear game)**.**

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

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

If A is non-singular, there is an unique stationary point which is also the unique Nash equilibrium. The gradient method is known not to converge in such game while the proximal point and extragradient methods converge Rockafellar (1976); Tseng (1995).

Bilinear games are of particular interest to us as they are seen as models of the convergence problems that arise during the training of GANs. Indeed, Mescheder et al. (2017) showed that eigenvalues of the Jacobian of the vector field with small real parts and large imaginary parts could be at the origin of these problems. Bilinear games have pure imaginary eigenvalues and so are limiting models of this situation. Moreover, they can also be seen as a very simple type of WGAN, with the generator and the discriminator being both linear, as explained in Gidel et al. (2019a); Mescheder et al..

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

3.2 Methods and convergence analysis

Convergence theory of fixed-point iterations. Seeing optimization algorithms as the repeated application of some operator allows us to deduce their convergence properties from the spectrum of this operator. This point of view was presented by Polyak (1987); Bertsekas (1999) and recently used by Arjevani et al. (2016); Mescheder et al. (2017); Gidel et al. (2019b) for instance. The idea is that the iterates of a method $(\omega_t)_t$ are generated by a scheme of the form:

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

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

Theorem 1 (Polyak (1987))**.** *Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be continuously differentiable and let $\omega^* \in \mathbb{R}^d$ be a fixed point of F. If $\rho(\nabla F(\omega^*)) < 1$, then for ω_0 in a neighborhood of ω^*, the iterates $(\omega_t)_t$ defined by $\omega_{t+1} = F(\omega_t)$ for all $t \geq 0$ converge linearly to ω^* at a rate of $\mathcal{O}((\rho(\nabla F(\omega^*)) + \epsilon)^t)$ for all $\epsilon > 0$.*

This theorem means that to derive a local rate of convergence for a given method, one needs only to focus on the eigenvalues of $\nabla F(\omega^*)$. Note that if the operator F is linear, there exists slightly stronger results such as Thm. 10 in Appendix C.

Gradient method. Following Gidel et al. (2019b), we define GD as the application of the operator $F_\eta(\omega) := \omega - \eta v(\omega)$, for $\omega \in \mathbb{R}^d$. Thus we have:

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

Proximal point. For v monotone (Minty, 1962; Rockafellar, 1976), the proximal point operator can be defined as $P_\eta(\omega) = (\mathrm{Id} + \eta v)^{-1}(\omega)$ and therefore can be seen as an implicit scheme: $\omega_{t+1} = \omega_t - \eta v(\omega_{t+1})$.

Extragradient. EG was introduced by Korpelevich (1976) in the context of variational inequalities. Its update rule is

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

It can be seen as an approximation of the implicit update of the proximal point method. Indeed Nemirovski (2004) showed a rate of $\mathcal{O}(1/t)$ for extragradient by treating it as a "good enough" approximation of the proximal point method. To see this, fix $\omega \in \mathbb{R}^d$. Then $P_\eta(\omega)$ is the solution of $z = \omega - \eta v(z)$. Equivalently, $P_\eta(\omega)$ is the fixed point of

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

which is a contraction for $\eta > 0$ small enough. From Picard's fixed point theorem, one gets that the proximal point operator $P_\eta(\omega)$ can be obtained as the limit of $\varphi^k_{\eta,\omega}(\omega)$ when k goes to infinity. What Nemirovski (2004) showed is that $\varphi^2_{\eta,\omega}(\omega)$, that is to say the extragradient update, is close enough to the result of the fixed point computation to be used in place of the proximal point update without affecting the sublinear convergence speed. Our analysis of multi-step extrapolation methods will encompass all the iterates $\varphi^k_{\eta,\omega}$ and we will show that a similar phenomenon happens for linear convergence rates.

Optimistic gradient. Originally introduced in the online learning literature (Chiang et al., 2012; Rakhlin and Sridharan, 2013) as a two-steps method, Daskalakis et al. (2018) reformulated it with only one step in the unconstrained case:

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

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

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

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

3.3 *p*-SCLI framework for game optimization

In this section, we present an extension of the framework of Arjevani et al. (2016) to derive lower bounds for game optimization (also see §G). The idea of this framework is to see algorithms as the iterated application of an operator. If the vector field is linear, this transformation is linear too and so its behavior when iterated is mainly governed by its spectral radius. This way, showing a lower bound for a class of algorithms is reduced to lower bounding a class of spectral radii.

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

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

a mapping $\mathcal{N} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$. The associated update rule can be defined through,

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

This form of the update rule is required by the consistency condition of Arjevani et al. (2016) which is necessary for the algorithm to converge to stationary points, as discussed in §G. Also note that 1-SCLI are first-order methods that use only the last iterate to compute the next one. Accelerated methods such as accelerated gradient descent (Nesterov, 2004) or the heavy ball method (Polyak, 1964) belong in fact to the class of 2-SCLI, which encompass methods which uses the last two iterates.

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

Theorem 2 (Arjevani et al. (2016))**.** *For all $v \in \mathcal{V}_d$, for almost every*[2] *initialization point $\omega_0 \in \mathbb{R}^d$, if $(\omega_t)_t$ are the iterates of $F_{\mathcal{N}}$ starting from ω_0,*

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

4 Revisiting GD for games

In this section, our goal is to illustrate the precision of the spectral bounds and the complexity of the interactions between players in games. We first give a simplified version of the bound on the spectral radius from Gidel et al. (2019b) and show that their results also imply that this rate is tight.

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

(i). *(Gidel et al., 2019b) For $\eta = \min_{\lambda\in\sigma^*} \Re(1/\lambda)$, the spectral radius of F_η can be upper-bounded as*

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

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

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

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

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

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

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

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

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

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

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

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

$$\rho(\nabla F_\eta(\omega^*))^2 \leq 1 - \tfrac{1}{4}\tfrac{(\mu_1+\mu_2)^2}{L_{12}^2 + L_1L_2}\,. \tag{3}$$

What is surprising is that, in some regimes, this result induces faster local convergence rates than the existing upper-bound for EG (Tseng, 1995):

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

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

5 Spectral analysis of multi-step EG

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

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

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

We are essentially considering all the iterates of the fixed point computation discussed in §3.2. Note that $F_{1,\eta}$ is GD while $F_{2,\eta}$ is EG. We aim at studying the local behavior of these methods at stationary points of the gradient dynamics, so fix ω^* s.t. $v(\omega^*) = 0$ and let $\sigma^* = \operatorname{Sp} \nabla v(\omega^*)$. We compute the spectra of these operators at this point and this immediately yields the spectral radius on the proximal point operator:

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

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

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

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

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

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

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

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

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

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

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

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

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

Though for ϵ close to zero, the dynamics will behave as such, this is not a purely bilinear game. The associated

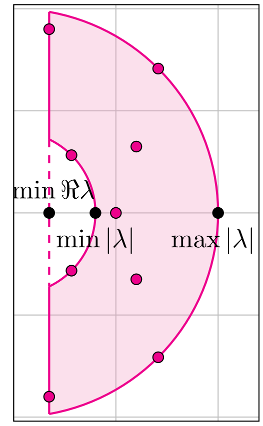


Figure 1: Illustration of the three quantities involved in Thm. 4. The magenta dots are an example of eigenvalues belonging to σ^*. Note that σ^* is always symmetric with respect to the real axis because the Jacobian is a real matrix (and thus non-real eigenvalues are complex conjugates). Note how $\min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

vector field is only ϵ-strongly monotone and convergence guarantees relying only on strong monotonicity would give a rate of roughly $1 - \epsilon/4$. However Thm. 4 yields a convergence rate of roughly $1 - 1/64$ for extragradient.

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

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

Recovery of known rates. If v is μ-strongly monotone and L-Lipschitz, this bound is at least as precise as the standard one $1 - \mu/(4L)$ as μ lower bounds the real part of the eigenvalues of the Jacobian, and L upper bounds their magnitude, as shown in Lem. 8 in §F.2. We empirically evaluate the improvement over this standard rate on synthetic examples in Appendix J. On the other hand, Thm. 4 also recovers the standard rates for the bilinear problem,[3] as shown below:

Corollary 2 (Bilinear game)**.** *Consider Ex. 1. The iterates of the k-extrapolation method with $k \geq 2$ converge*

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

globally to ω^* at a linear rate of $\mathcal{O}\Big(\big(1 - \frac{1}{64}\frac{\sigma_{min}(A)^2}{\sigma_{max}(A)^2}\big)^t\Big)$.

Note that this rate is similar to the one derived by Gidel et al. (2019b) for alternating gradient descent with negative momentum. This raises the question of whether general acceleration exists for games, as we would expect the quantity playing the role of the condition number in Cor. 2 to appear without the square in the convergence rate of a method using momentum.

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

Lower bounds for extrapolation methods. We now show that the rates we proved for EG are tight and optimal by deriving lower bounds of convergence for general extrapolation methods. As described in §3.3, a 1-SCLI method is parametrized by a polynomial $\mathcal{N}$. We consider the class of methods where $\mathcal{N}$ is any polynomial of degree at most $k-1$, and we will derive lower bounds for this class. This class is large enough to include all the k'-extrapolation methods for $k' \leq k$ with possibly different step sizes for each extrapolation step (see §H for more examples).

Our main result is that no method of this class can significantly beat the convergence speed of EG of Thm. 4 and Thm. 6. We proceed in two steps: for each of the two terms of these bounds, we provide an example matching it up to a factor. In (i) of the following theorem, we give an example of convex optimization problem which matches the real part, or strong monotonicity, term. Note that this example is already an extension of Arjevani et al. (2016) as the authors only considered constant $\mathcal{N}$. Next, in (ii), we match the other term with a bilinear game example.

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

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

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

6 Unified global proofs of convergence

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

6.1 Global Assumptions

Tseng (1995) proved linear convergence results for EG by using the projection-type error bound Tseng (1995, Eq. 5) which, in the unconstrained case, i.e. for $v(\omega^*) = 0$, can be written as,

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

The author then shows that this condition holds for the bilinear game of Example 1 and that it induces a convergence rate of $1 - c\sigma_{min}(A)^2/\sigma_{max}(A)^2$ for some constant $c > 0$. He also shows that this condition is implied by strong monotonicity with $\gamma = \mu$. Our analysis builds on the results from Tseng (1995) and extends them to cover the whole range of games and recover the optimal rates.

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

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

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

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

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

As announced, we will combine this assumption with the strong monotonicity to derive unified global convergence guarantees. Before that, note that this quantities can be related to the spectrum of $\mathrm{Sp}\,\nabla v(\omega^*)$ as follows – see Lem. 8 in Appendix F.1,

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

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

6.2 Global analysis EG and OG

We can now state our global convergence result for EG:

Theorem 6. *Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and (i) μ-strongly monotone for some $\mu \geq 0$, (ii) L-Lipschitz, (iii) such that $\sigma_{min}(\nabla v) \geq \gamma$ for some $\gamma > 0$. Then, for $\eta \leq (4L)^{-1}$, the iterates $(\omega_t)_t$ of (EG) converge linearly to ω^* as, for all $t \geq 0$,*

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

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

Combining our new proof technique and the analysis provided by Gidel et al. (2019a), we can derive a similar convergence rate for the optimistic gradient method.

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

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

Interpretation of the condition numbers. As in the previous section, this rate of convergence for EG is similar to the rate of the proximal point method for a small enough step size, as shown by Prop. 1 in §F.2. Moreover, the proof of the latter gives insight into the two quantities appearing in the rate of Thm. 6. Indeed, the convergence result for the proximal point method is obtained by bounding the singular values of ∇P_η, and so we compute,[4]

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

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

6.3 Global analysis of consensus optimization

In this section, we give a unified proof of CO. A global convergence rate for this method was proven by Abernethy et al. (2019). However it used a perturbation analysis of HGD. The drawbacks are that it required that the CO update be sufficiently close to the one of HGD and could not take advantage of strong monotonicity. Here, we combine the monotonicity μ with the lower bound on the singular value γ.

As this scheme uses second-order[5] information, we need to replace the Lipschitz hypothesis with one that also controls the variations of the Jacobian of v: we use L_H^2, the Lispchitz smoothness of H. See Abernethy et al. (2019) for how it might be instantiated.

Theorem 8. *Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable such that (i) v is μ- strongly monotone for some $\mu \geq 0$, (ii) $\sigma_{min}(\nabla v) \geq \gamma$ for some $\gamma > 0$ (iii) H is L_H^2 Lipschitz-smooth. Then, for $\alpha = (\mu + \sqrt{\mu^2 + 2\gamma^2})/(4L_H^2)$, $\beta = (2L_H^2)^{-1}$ the iterates of CO defined by (CO) satisfy, for all $t \geq 0$,*

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

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

This result also encompasses the rate of HGD (Abernethy et al., 2019, Lem. 4.7). The dependance in μ is on par with the standard rate for the gradient method (see Nesterov and Scrimali (2006, Eq. 2.12) for instance). However, this can be improved using a sharper assumption, as discussed in Remark 1 in Appendix F.3, and so our result is not optimal in this regard.

7 Conclusion

In this paper, we studied the dynamics of EG, both locally and globally and extended our global guarantees to other promising methods such as OG and CO. Our analysis is tight for EG and unified as they cover the whole spectrum of games from bilinear to purely cooperative settings. They show that in between, these methods enjoy the best of both worlds. We confirm that, unlike in convex minimization, the behaviors of EG and GD differ significantly. The other lower bounds show that EG can be considered as optimal among first-order methods that use only the last iterate.

Finally, as mentioned in §5, the rate of alternating gradient descent with negative momentum from Gidel et al. (2019b) on the bilinear example essentially matches the rate of EG in Cor. 2. Thus the question of an acceleration for adversarial games similar to the one in the convex case using Polyak (Polyak, 1964) or Nesterov's (Nesterov, 2004) momentum remains open.

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

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

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.

Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local Saddle Point Optimization: A Curvature Exploitation Approach. *AISTATS*, 2019.

Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On Lower and Upper Bounds in Smooth and Strongly Convex Optimization. *Journal of Machine Learning Research*, 17, 2016.

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

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

Hugo Berard, Gauthier Gidel, Amjad Almahairi, Pascal Vincent, and Simon Lacoste-Julien. A closer look at the optimization landscapes of generative adversarial networks. *arXiv*, 2019.

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

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

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

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

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *COLT*, 2012.

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

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.

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. *AISTATS*, 2019b.

Ian Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. 2016.

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.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

Jacques Hadamard. Sur les transformations ponctuelles. *Bull. Soc. Math. France*, 34, 1906.

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.

M. P. Levy. Sur les fonctions de lignes implicites. *Bull. Soc. Math. France*, 48, 1920.

Tengyuan Liang and James Stokes. Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks. *AISTATS*, 2019.

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*, 2019.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *ICLR*, 2019.

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

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The Numerics of GANs. *NIPS*, 2017.

George J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Mathematical Journal*, 29, 1962.

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

Vaishnavh Nagarajan and J. Zico Kolter. Gradient descent GAN optimization is locally stable. *NIPS*, June 2017.

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.

A.S Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8, 1992.

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

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

Gerhard Opfer and Glenn Schober. Richardson's iteration for nonsymmetric matrices. *Linear Algebra and its Applications*, 1984.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. 2018.

Barak A. Pearlmutter. Fast Exact Multiplication by the Hessian. *Neural Computation*, 1994.

David Pfau and Oriol Vinyals. Connecting Generative Adversarial Networks and Actor-Critic Methods. 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.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NeurIPS*, 2013.

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

R. Tyrrell Rockafellar. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14, 1976.

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

H E Salzer. Lagrangian interpolation at the Chebyshev points $x_{n,\nu} = \cos(\nu\pi/n), \nu = 0(1)n$; some unnoted advantages. 1971.

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

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. OCLC: 1629708.

Fuzhen Zhang, editor. *The Schur Complement and Its Applications*, volume 4 of *Numerical Methods and Algorithms.* Springer-Verlag, 2005.

A Notation

We denote by $\mathrm{Sp}(A)$ the spectrum of a matrix A. Its spectral radius is defined by $\rho(A) = \max\{|\lambda| \mid \lambda \in \mathrm{Sp}(A)\}$. We write $\sigma_{min}(A)$ for the smallest singular value of A, and $\sigma_{max}(A)$ for the largest. $\Re$ and $\Im$ denote respectively the real part and the imaginary part of a complex number. We write $A \preccurlyeq B$ for two symmetric real matrices if and only if $B - A$ is positive semi-definite. For a vector $X \in \mathbb{C}^d$, denote its transpose by X^T and its conjugate transpose by X^H. $\|.\|$ denotes an arbitrary norm on $\mathbb{R}^d$ unless specified. We sometimes denote $\min(a, b)$ by $a \wedge b$ and $\max(a, b)$ by $a \vee b$. For $f : \mathbb{R}^d \to \mathbb{R}^d$, we denote by f^k the composition of f with itself k times, i.e. $f^k(\omega) = \underbrace{f \circ f \circ \cdots \circ f}_{k \text{ times}}(\omega)$.

B Interpretation of spectral quantities in a two-player zero-sum game

In this appendix section, we are interested in interpreting spectral bounds in terms of the usual strong convexity and Lipschitz continuity constants in a two-player zero-sum game:

$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^p} f(x, y) \tag{12}$$

with f is two times continuously differentiable.

Assume,

$$\mu_1 I_m \preccurlyeq \nabla_x^2 f \preccurlyeq L_1 I_m \tag{13}$$

$$\mu_2 I_p \preccurlyeq -\nabla_y^2 f \preccurlyeq L_2 I_p \tag{14}$$

$$\mu_{12}^2 I_p \preccurlyeq (\nabla_x \nabla_y f)^T (\nabla_x \nabla_y f) \preccurlyeq L_{12}^2 I_p \tag{15}$$

where μ_1, μ_2 and μ_{12} are non-negative constants. Let $\omega^* = (x^*, y^*)$ be a stationary point. To ease the presentation, let,

$$\nabla v(\omega^*) = \begin{pmatrix} \nabla_x^2 f(\omega^*) & (\nabla_x \nabla_y f(\omega^*))^T \\ -(\nabla_x \nabla_y f(\omega^*)) & \nabla_y^2 f(\omega^*) \end{pmatrix} = \begin{pmatrix} S_1 & A \\ -A^T & S_2 \end{pmatrix}. \tag{16}$$

Now, more precisely, we are interested in lower bounding $\Re(\lambda)$ and $|\lambda|$ and upper bounding $|\lambda|$ for $\lambda \in \mathrm{Sp}\,\nabla v(\omega^*)$.

B.1 Commutative and square case

In this subsection we focus on the square and commutative case as formalized by the following assumptions:

Assumption 1 (Square and commutative case)**.** The following holds: *(i)* $p = m = \frac{d}{2}$; *(ii)* S_2 and A^T commute; *(iii)* S_1, S_2 and AA^T commute.

Assumption 1 holds if, for instance, the objective is separable, i.e. $f(x, y) = \sum_{i=1}^m f_i(x_i, y_i)$. Then, using a well-known linear algebra theorem, Assumption 1 implies that there exists $U \in \mathbb{R}^{d \times d}$ unitary such that $S_1 = U \operatorname{diag}(\alpha_1, \ldots, \alpha_m) U^T$, $S_2 = U \operatorname{diag}(\beta_1, \ldots, \beta_m) V^T$ and $AA^T = U \operatorname{diag}(\sigma_1^2, \ldots, \sigma_d^2) U^T$ where $\alpha_1, \ldots, \alpha_m$ are the eigenvalues of S_1, $\beta_1, \ldots, \beta_m$ are the eigenvalues of S_2 and $\sigma_1, \ldots, \sigma_p$ are the singular values of A. See Lax (2007, p. 74) for instance.

Define,

$$\mu = \begin{pmatrix} \mu_1 & \mu_{12} \\ -\mu_{12} & \mu_2 \end{pmatrix}$$

$$L = \begin{pmatrix} L_1 & L_{12} \\ -L_{12} & L_2 \end{pmatrix}.$$

Denote by $|\mu|$ and $|L|$ the determinants of theses matrices, and by $\operatorname{Tr}\mu$ and $\operatorname{Tr} L$ their traces.

In this case we get an exact characterization of the spectrum $\nabla v(\omega^*)$, which we denote by $\sigma^* = \mathrm{Sp}\,\nabla v(\omega^*)$:

Lemma 3. *Under Assumption 1, $\lambda \in sigma^*$ if and only if there exists some $i \leq d$ such that λ is a root of*

$$P_i = X^2 - (\alpha_i + \beta_i) X + \alpha_i \beta_i + \sigma_i^2$$

Proof. We compute the characteristic polynomial of $\nabla v(\omega^*)$ using that S_2 and A^T commute, using the formula for the determinant of a block matrix, which can be found in Zhang (2005, Section 0.3) for instance.

$$\begin{aligned}\begin{vmatrix} XI - S_1 & -A \\ A^T & XI - S_2 \end{vmatrix} &= |(XI - S_1)(XI - S_2) + AA^T| \\ &= |X^2 I - X(S_1 + S_2) + S_1 S_2 + AA^T| \\ &= \prod_i \left(X^2 - (\alpha_i + \beta_i)X + \alpha_i\beta_i + \sigma_i^2\right)\end{aligned}$$

□

Theorem 9. *Under Assumption 1, we have the following results on the eigenvalues of $\nabla v(\omega^*)$.*

(a) For $i \leq m$, if $(\alpha_i - \beta_i)^2 < 4\sigma_i^2$, the roots of P_i satisfy:

$$\frac{\operatorname{Tr}\mu}{2} \leq \Re(\lambda), \quad \det\mu \leq |\lambda|^2 \leq \det L\,, \quad \forall \lambda \in \mathbb{C} \;s.t.\; P_i(\lambda) = 0\,. \tag{17}$$

(b) For $i \leq m$, if $(\alpha_i - \beta_i)^2 \geq 4\sigma_i^2$, the roots of P_i are real non-negative and satisfy :

$$\max\left(\mu_1 \wedge \mu_2, \frac{\det\mu}{\operatorname{Tr}L}\right) \leq \lambda \leq L_1 \vee L_2\,, \quad \forall \lambda \in \mathbb{C} \;s.t.\; P_i(\lambda) = 0\,. \tag{18}$$

(c) Hence, in general,

$$\mu_1 \wedge \mu_2 \leq \Re\lambda, \quad |\lambda|^2 \leq 2L_{max}^2 \quad \forall \lambda \in \sigma^*\,, \tag{19}$$

where $L_{max} = \max(L_1, L_2, L_{12})$.

Proof. (a) Assume that $(\alpha_i - \beta_i)^2 < 4\sigma_i^2$, i.e. the discriminant of the polynomial P_i of Lem. 3 is negative. Consider λ a root of P_i. Then $\Re\lambda = \frac{\alpha_i+\beta_i}{2}$ and $|\lambda|^2 = \alpha_i\beta_i + \sigma_i^2$. Hence $\Re\lambda \geq \frac{1}{2}\operatorname{Tr}\mu$ and $\det\mu \leq |\lambda|^2 \leq \det L$.

(b) Assume that $(\alpha_i - \beta_i)^2 \geq 4\sigma_i^2$, i.e. the discriminant of the polynomial P_i of Lem. 3 is non-negative. This implies that $\Delta = (\operatorname{Tr}L)^2 - 4\det\mu \geq 0$.

Denote by λ_+ and λ_- the two real roots of P_i. Then

$$\lambda_\pm = \frac{\alpha_i + \beta_i \pm \sqrt{(\alpha_i + \beta_i)^2 - 4(\alpha_i\beta_i + \sigma_i^2)}}{2}$$

Hence

$$\lambda_+ \geq \lambda_- \geq \min_{\max(\operatorname{Tr}\mu,\; 4\det\mu) \leq x \leq \operatorname{Tr}L} \frac{x - \sqrt{x^2 - 4\det\mu}}{2}$$

As $x \mapsto x - \sqrt{x^2 - 4\det\mu}$ is decreasing on its domain, the minimum is reached at $\operatorname{Tr}L$ and is $\frac{\operatorname{Tr}L - \sqrt{\Delta}}{2} \geq 0$. However this lower bound is quite loose when $A = 0$. So note that

$$\begin{aligned}\lambda_- &= \frac{\alpha_i + \beta_i - \sqrt{(\alpha_i + \beta_i)^2 - 4(\alpha_i\beta_i + \sigma_i^2)}}{2} &(20)\\ &\geq \frac{\alpha_i + \beta_i - \sqrt{(\alpha_i - \beta_i)^2}}{2} = \alpha_i \wedge \beta_i &(21)\\ &\geq \mu_1 \wedge \mu_2 &(22)\\ & &(23)\end{aligned}$$

Similarly,

$$\lambda_+ \leq \frac{\alpha_i + \beta_i + \sqrt{(\alpha_i - \beta_i)^2}}{2} = \alpha_i \vee \beta_i \leq L_1 \vee L_2 \tag{24}$$

Finally:

$$L_1 \vee L_2 \geq \lambda_+ \geq \lambda_- \geq \max\left(\frac{\operatorname{Tr} L - \sqrt{\Delta}}{2}, \mu_1 \wedge \mu_2\right). \tag{25}$$

Moreover,

$$\operatorname{Tr} L - \sqrt{\Delta} = \frac{\left(\operatorname{Tr} L - \sqrt{\Delta}\right)\left(\operatorname{Tr} L + \sqrt{\Delta}\right)}{\operatorname{Tr} L + \sqrt{\Delta}} \tag{26}$$

$$= \frac{4 \det \mu}{\operatorname{Tr} L + \sqrt{\Delta}} \tag{27}$$

$$\geq \frac{2 \det \mu}{\operatorname{Tr} L}, \tag{28}$$

which yields the result.

(c) These assertions are immediate corollaries of the two previous ones.

□

We need the following lemma to be able to interpret Thm. 6 in the context of Example 2, whose assumptions imply Assumption 1.

Lemma 4. *Under Assumption 1, the singular values of $\nabla v(\omega^*)$ can be lower bounded as:*

$$\mu_{12}(\mu_{12} - \max(L_1 - \mu_2, L_2 - \mu_1)) \leq \sigma_{min}(\nabla v(\omega^*))^2. \tag{29}$$

In particular, if $\mu_{12} > 2\max(L_1 - \mu_2, L_2 - \mu_1)$, this becomes

$$\frac{1}{2}\mu_{12}^2 \leq \sigma_{min}(\nabla v(\omega^*))^2. \tag{30}$$

Proof. To prove this we compute the eigenvalues of $(\nabla v(\omega^*))^T \nabla v(\omega^*)$. We have that,

$$(\nabla v(\omega^*))^T \nabla v(\omega^*) = \begin{pmatrix} S_1^2 + AA^T & S_1 A - AS_2 \\ A^T S_1 - S_2 A^T & A^T A + S_2^2 \end{pmatrix}. \tag{31}$$

As in the proof of Lem. 3, as Assumption 1 implies that $A^T S_1 - S_2 A^T$ and $A^T A + S_2^2$ commute,

$$|XI - (\nabla v(\omega^*))^T \nabla v(\omega^*)| = \left|(XI - S_1^2 - AA^T)(XI - S_2^2 - A^T A) - (S_1 - S_2)^2 AA^T\right| \tag{32}$$

$$= \prod_i \left((XI - \alpha_i^2 - \sigma_i^2)(XI - \beta_i^2 - \sigma_i^2) - (\alpha_i - \beta_i)^2 \sigma_i^2\right). \tag{33}$$

Let $Q_i(X) = (XI - \alpha_i^2 - \sigma_i^2)(XI - \beta_i^2 - \sigma_i^2) - (\alpha_i - \beta_i)^2 \sigma_i^2$. Its discriminant is

$$\Delta_i' = (\alpha_i^2 + \beta_i^2 + 2\sigma_i^2)^2 - 4\left((\alpha_i^2 + \sigma_i^2)(\beta_i^2 + \sigma_i^2) - (\alpha_i - \beta_i)^2)\sigma_i^2\right) \tag{34}$$

$$= (\alpha_i - \beta_i)^2((\alpha_i + \beta_i)^2 + 4\sigma_i^2) \geq 0. \tag{35}$$

Hence the roots of Q_i are:

$$\lambda_{i\pm} = \frac{1}{2}\left(\alpha_i^2 + \beta_i^2 + 2\sigma_i^2 \pm \sqrt{(\alpha_i - \beta_i)^2((\alpha_i + \beta_i)^2 + 4\sigma_i^2)}\right). \tag{36}$$

The smallest is λ_{i-} which can be lower bounded by

$$\lambda_{i-} = \frac{1}{2}\left(\alpha_i^2 + \beta_i^2 + 2\sigma_i^2 - \sqrt{(\alpha_i + \beta_i)^2((\alpha_i + \beta_i)^2 + 4\sigma_i^2)}\right) \tag{37}$$

$$\geq \frac{1}{2}\left(\alpha_i^2 + \beta_i^2 - |\alpha_i^2 - \beta_i^2| + 2\sigma_i(\sigma_i - |\alpha_i - \beta_i|)\right) \tag{38}$$

$$\geq \sigma_i(\sigma_i - |\alpha_i - \beta_i|) \tag{39}$$

$$\geq \mu_{12}(\mu_{12} - \max(L_1 - \mu_2, L_2 - \mu_1)). \tag{40}$$

□

C Complement for §3

The convergence result of Thm. 1 can be strengthened if the Jacobian is constant as shown below. A proof of this classical result in linear algebra can be found in Arjevani et al. (2016) for instance.

Theorem 10. *Let $F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ be a linear operator. If $\rho(\nabla F) < 1$, then for all $\omega_0 \in \mathbb{R}^d$, the iterates $(\omega_t)_t$ defined as above converge linearly to ω^* at a rate of $\mathcal{O}((\rho(\nabla F))^t)$.*

D Convergence results of §4

Let us restate Thm. 3 for clarity.

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

(i). (Gidel et al., 2019b) For $\eta = \min_{\lambda \in \sigma^*} \Re(1/\lambda)$, the spectral radius of F_η can be upper-bounded as

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

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

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

In this subsection, we quickly show how to obtain (i) of Thm. 3 from Theorem 2 of Gidel et al. (2019b), whose part which interests us now is the following:

Theorem (Gidel et al. (2019b, part of Theorem 2))**.** *If the eigenvalues of $\nabla v(\omega^*)$ all have positive real parts, then for $\eta = \Re(1/\lambda_1)$ one has*

$$\rho(\nabla F_\eta(\omega^*))^2 \leq 1 - \Re(1/\lambda_1)\delta \tag{41}$$

where $\delta = \min_{1 \leq j \leq m} |\lambda_j|^2(2\Re(1/\lambda_j) - \Re(1/\lambda_1))$ and $\operatorname{Sp} \nabla v(\omega^*) = \{\lambda_1, \ldots, \lambda_m\}$ sorted such that $0 < \Re(1/\lambda_1) \leq \Re(1/\lambda_2) \leq \cdots \leq \Re(1/\lambda_m)$.

Proof of (i) of Thm. 3. By definition of the order on the eigenvalues,

$$\delta = \min_{1 \leq j \leq m} |\lambda_j|^2(\Re(1/\lambda_j) + \Re(1/\lambda_j) - \Re(1/\lambda_1)) \tag{42}$$

$$\geq \min_{1 \leq j \leq m} |\lambda_j|^2(\Re(1/\lambda_j)) \tag{43}$$

$$= \min_{1 \leq j \leq m} \Re(\lambda_j) \tag{44}$$

□

To prove the second part of Thm. 3, we rely on a different part of Gidel et al. (2019b, Theorem 2) which we recall below:

Theorem (Gidel et al. (2019b, part of Theorem 2))**.** *The best step-size η^*, that is to say the solution of the optimization problem*

$$\min_\eta \rho(\nabla F_\eta(\omega^*))^2\,, \tag{45}$$

satisfy:

$$\min_{\lambda \in \sigma^*} \Re(1/\lambda) \leq \eta^* \leq 2 \min_{\lambda \in \sigma^*} \Re(1/\lambda)\,. \tag{46}$$

(ii) of Thm. 3 is now immediate.

Proof of (ii) of Thm. 3. By definition of the spectral radius,

$$\rho(\nabla F_{\eta^*}(\omega^*))^2 = \max_{\lambda \in \operatorname{Sp}(\nabla v(\omega^*))} |1 - \eta^* \lambda|^2 \tag{47}$$

$$= 1 - \min_{\lambda \in \operatorname{Sp}(\nabla v(\omega^*))} 2\eta^* \Re\lambda - |\eta^* \lambda|^2 \tag{48}$$

$$\geq 1 - \min_{\lambda \in \operatorname{Sp}(\nabla v(\omega^*))} 2\eta^* \Re\lambda \tag{49}$$

$$\geq 1 - 4 \min_{\lambda \in \operatorname{Sp}(\nabla v(\omega^*))} \Re\lambda \min_{\lambda \in \sigma^*} \Re(1/\lambda) \tag{50}$$

□

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

$$\rho(\nabla F_\eta(\omega^*))^2 \leq 1 - \tfrac{1}{4}\tfrac{(\mu_1+\mu_2)^2}{L_{12}^2+L_1L_2}\,. \tag{3}$$

Proof. Note that the hypotheses stated in §4 correspond to the assumptions of §B.1. Moreover, with the notations of this subsection, one has that $4\sigma_i^2 \geq 4\mu_{12}^2$ and $\max(L_1, L_2)^2 \geq (\alpha_i - \beta_i)^2$. Hence the condition $2\mu_{12} \geq \max(L_1, L_2)$ implies that all the eigenvalues of $\nabla v(\omega^*)$ satisfy the case (a) of Thm. 9. Then, using Thm. 3,

$$\begin{aligned}
\rho(\nabla F_\eta(\omega^*))^2 &\leq 1 - \min_{\lambda\in\sigma^*} \Re(1/\lambda) \min_{\lambda\in\sigma^*} \Re(\lambda) && (51)\\
&\leq 1 - \left(\frac{\min_{\lambda\in\sigma^*}\Re(\lambda)}{\max_{\lambda\in\sigma^*}|\lambda|}\right)^2 && (52)\\
&\leq 1 - \frac{1}{4}\frac{(\mu_1+\mu_2)^2}{L_{12}^2+L_1L_2}\,. && (53)
\end{aligned}$$

□

E Spectral analysis of §5

We prove Lem. 1.

Lemma 5. *Assuming that the eigenvalues of $\nabla v(\omega^*)$ all have non-negative real parts, the proximal point operator P_η is continuously differentiable in a neighborhood of ω^* . Moreover, the spectra of the k-extrapolation operator and the proximal point operator are given by:*

$$\operatorname{Sp}\nabla F_{\eta,k}(\omega^*) = \left\{\textstyle\sum_{j=0}^k(-\eta\lambda)^j \;\middle|\; \lambda\in\sigma^*\right\} \tag{54}$$

$$\textit{and}\quad \operatorname{Sp}\nabla P_\eta(\omega^*) = \left\{(1+\eta\lambda)^{-1} \;\middle|\; \lambda\in\sigma^*\right\}\,. \tag{55}$$

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

$$\rho(\nabla P_\eta(\omega^*))^2 = 1 - \min_{\lambda\in\sigma^*} \tfrac{2\eta\Re\lambda+\eta^2|\lambda|^2}{|1+\eta\lambda|^2}\,. \tag{56}$$

To prove the result about the k-extrapolation operator, we first show the following lemma, which will be used again later.

Recall that we defined $\varphi_{\eta,\omega} : z \mapsto \omega - \eta v(z)$. We drop the dependence on η in $\varphi_{\eta,\omega}$ for compactness.

Lemma 6. *The Jacobians of $\varphi_\omega^k(z)$ with respect to z and ω can be written as*

$$\nabla_z\varphi_\omega^k(z) = (-\eta)^k\nabla v(\varphi_\omega^{k-1}(z))\nabla v(\varphi_\omega^{k-2}(z))\dots\nabla v(\varphi_\omega^0(z)) \tag{57}$$

$$\nabla_\omega\varphi_\omega^k(z) = \sum_{j=0}^{k-1}(-\eta)^j\nabla v(\varphi_\omega^{k-1}(z))\nabla v(\varphi_\omega^{k-2}(z))\dots\nabla v(\varphi_\omega^{k-j}(z))\,. \tag{58}$$

Proof. We prove the result by induction:

- For $k = 1$, $\varphi_\omega(z) = \omega - \eta v(z)$ and the result holds.
- Assume this result holds for $k \geq 0$. Then,

$$\begin{aligned}
\nabla_z\varphi_\omega^{k+1}(z) &= \nabla_z\varphi_\omega(\varphi_\omega^k(z))\nabla_z\varphi_\omega^k(z) && (59)\\
&= -\eta\nabla v(\varphi_\omega^k(z))(-\eta)^k\nabla v(\varphi_\omega^{k-1}(z))\dots\nabla v(\varphi_\omega^0(z)) && (60)\\
&= (-\eta)^{k+1}\nabla v(\varphi_\omega^k(z))\nabla v(\varphi_\omega^{k-1}(z))\dots\nabla v(\varphi_\omega^0(z))\,. && (61)
\end{aligned}$$

For the derivative with respect to ω, we use the chain rule:

$$\nabla_\omega \varphi_\omega^{k+1}(z) = \nabla_\omega \varphi_\omega(\varphi_\omega^k(z)) + \nabla_z \varphi_\omega(\varphi_\omega^k(z)) \nabla_\omega \varphi_\omega^k(z) \tag{62}$$

$$= I_d - \eta v(\varphi_\omega^k(z)) \sum_{j=0}^{k-1} (-\eta)^j \nabla v(\varphi_\omega^{k-1}(z)) \dots \nabla v(\varphi_\omega^{k-j}(z)) \tag{63}$$

$$= I_d + \sum_{j=0}^{k-1} (-\eta)^{j+1} \nabla v(\varphi_\omega^k(z)) \nabla v(\varphi_\omega^{k-1}(z)) \dots \nabla v(\varphi_\omega^{k-j}(z)) \tag{64}$$

$$= I_d + \sum_{j=1}^{k} (-\eta)^{j} \nabla v(\varphi_\omega^k(z)) \nabla v(\varphi_\omega^{k-1}(z)) \dots \nabla v(\varphi_\omega^{k+1-j}(z)) \tag{65}$$

$$= \sum_{j=0}^{k} (-\eta)^{j} \nabla v(\varphi_\omega^k(z)) \nabla v(\varphi_\omega^{k-1}(z)) \dots \nabla v(\varphi_\omega^{k+1-j}(z)) \tag{66}$$

□

In the proof of Lem. 1 and later we will use the spectral mapping theorem, which we state below for reference:

Theorem 11 (Spectral Mapping Theorem)**.** *Let $A \in \mathbb{C}^{d\times d}$ be a square matrix, and P be a polynomial. Then,*

$$\operatorname{Sp} P(A) = \{P(\lambda) \mid \lambda \in \operatorname{Sp} A\} . \tag{67}$$

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

Proof of Lem. 1. First we compute $\nabla F_{\eta,k}(\omega^*)$. As ω^* is a stationary point, it is a fixed point of the extrapolation operators, i.e. $\varphi_{\omega^*}^j(\omega^*) = \omega^*$ for all $j \geq 0$. Then, by the chain rule,

$$\nabla F_{\eta,k}(\omega^*) = \nabla_z \varphi_{\omega^*}^k(\omega^*) + \nabla_\omega \varphi_{\omega^*}^k(\omega^*) \tag{68}$$

$$= (-\eta \nabla v(\omega^*))^k + \sum_{j=0}^{k-1} (-\eta \nabla v(\omega^*))^j \tag{69}$$

$$= \sum_{j=0}^{k} (-\eta \nabla v(\omega^*))^j . \tag{70}$$

Hence $\nabla F_{\eta,k}(\omega^*)$ is a polynomial in $\nabla v(\omega^*)$. Using the spectral mapping theorem (Thm. 11), one gets that

$$\operatorname{Sp} \nabla F_{\eta,k}(\omega^*) = \left\{ \sum_{j=0}^{k} (-\eta)^j \lambda^j \mid \lambda \in \operatorname{Sp} \nabla v(\omega^*) \right\} \tag{71}$$

For the proximal point operator, first let us prove that it is differentiable in a neighborhood of ω^*. First notice that,

$$\operatorname{Sp}(I_d + \eta \nabla v(\omega^*)) = \{1 + \eta\lambda \mid \lambda \in \operatorname{Sp} \nabla v(\omega^*)\} . \tag{72}$$

If the eigenvalues of $\nabla v(\omega^*)$ all have non-negative real parts, this spectrum does not contain zero. Hence $\omega \mapsto \omega + \eta v(\omega)$ is continuously differentiable and has a non-singular differential at ω^*. By the inverse function theorem (see for instance Rudin (1976)), $\omega \mapsto \omega + \eta v(\omega)$ is invertible in a neighborhood of ω^* and its inverse, which is P_η, is continuously differentiable there. Moreover,

$$\nabla P_\eta(\omega^*) = (I_d + \eta \nabla v(\omega^*))^{-1} . \tag{73}$$

Recall that the eigenvalues of a non-singular matrix are exactly the inverses of the eigenvalues of its inverse. Hence,

$$\operatorname{Sp} \nabla P_\eta(\omega^*) = \{\lambda^{-1} \mid \lambda \in \operatorname{Sp}(I_d + \eta \nabla v(\omega^*))\} = \left\{(1+\eta\lambda)^{-1} \mid \lambda \in \operatorname{Sp} \nabla v(\omega^*)\right\} , \tag{74}$$

where the last equality follows from the spectral mapping theorem applied to $I_d + \eta \nabla v(\omega^*)$. Now, the bound on the spectral radius of the proximal point operator is immediate. Indeed, its spectral radius is:

$$\rho(\nabla P_\eta(\omega^*))^2 = \max_{\lambda \in \sigma^*} \frac{1}{|1+\eta\lambda|^2} \tag{75}$$

$$= 1 - \min_{\lambda \in \sigma^*} \left(\frac{2\eta \Re\lambda + \eta^2 |\lambda|^2}{|1+\eta\lambda|^2} \right) \tag{76}$$

which yields the result. □

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

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

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

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

Proof. Let $L = \max_{\lambda\in\sigma^*}|\lambda|$ and $\eta = \frac{\tau}{L}$ for some $\tau > 0$. For $\lambda \in \sigma^*$,

$$\left| \sum_{j=0}^{k} (-\eta)^j \lambda^j \right|^2 = \frac{|1 - (-\eta)^{k+1}\lambda^{k+1}|^2}{|1+\eta\lambda|^2} \tag{77}$$

$$= \frac{1 + 2(-1)^k \eta^{k+1} \Re(\lambda^{k+1}) + \eta^{2(k+1)}|\lambda|^{2(k+1)}}{|1+\eta\lambda|^2} \tag{78}$$

$$= 1 - \frac{2\eta\Re\lambda + \eta^2|\lambda|^2 - 2(-1)^k\eta^{k+1}\Re(\lambda^{k+1}) - \eta^{2(k+1)}|\lambda|^{2(k+1)}}{|1+\eta\lambda|^2} \tag{79}$$

$$= 1 - \frac{2\eta\Re\lambda + \eta^2|\lambda|^2 \left(1 - 2(-1)^k\eta^{k-1}\frac{\Re(\lambda^{k+1})}{|\lambda|^2} - \eta^{2(k-1)}|\lambda|^{2(k-1)}\right)}{|1+\eta\lambda|^2} \tag{80}$$

Now we focus on lower bounding the terms in between the parentheses. By definition of η, we have $\eta^{k-1}\frac{|\Re(\lambda^{k+1})|}{|\lambda|^2} \leq \tau^{k-1}$ and $\eta^{2(k-1)}|\lambda|^{2(k-1)} \leq \tau^{2(k-1)}$. Hence

$$1 + 2(-1)^k\eta^{k-1}\frac{\Re(\lambda^{k+1})}{|\lambda|^2} + \eta^{2(k-1)}|\lambda|^{2(k-1)}) \geq 1 - 2\eta^{k-1}\frac{|\Re(\lambda^{k+1})|}{|\lambda|^2} - \eta^{2(k-1)}|\lambda|^{2(k-1)} \tag{81}$$

$$\geq 1 - 2\tau^{k-1} - \tau^{2(k-1)} \tag{82}$$

$$\tag{83}$$

Notice that if $k = 1$, i.e. for the gradient method, we cannot control this quantity. However, for $k \geq 2$, if $\tau \leq (\frac{1}{4})^{\frac{1}{k-1}}$, one gets that

$$1 - 2\tau^{k-1} - \tau^{2(k-1)} \geq 1 - \frac{1}{2} - \frac{1}{16} = \frac{7}{16} \tag{84}$$

which yields the first assertion of the theorem. For the second one, take $\eta = \frac{1}{4L}$, i.e. the maximum step-size authorized for extragradient, and one gets that

$$|1+\eta\lambda|^2 = 1 + 2\eta\Re\lambda + \eta^2|\lambda|^2 \tag{85}$$

$$\leq 1 + 2\frac{1}{4} + \frac{1}{16} = \frac{25}{16}\,. \tag{86}$$

Then,

$$\frac{2\eta\Re\lambda + \frac{7}{16}\eta^2|\lambda|^2}{|1+\eta\lambda|^2} \geq \frac{1}{4}\left(2\frac{16}{25}\frac{\Re\lambda}{L} + \frac{7}{100}\frac{|\lambda|^2}{L^2}\right) \tag{87}$$

$$\geq \frac{1}{4}\left(\frac{\Re\lambda}{L} + \frac{7}{112}\frac{|\lambda|^2}{L^2}\right) \tag{88}$$

$$\geq \frac{1}{4}\left(\frac{\Re\lambda}{L} + \frac{1}{16}\frac{|\lambda|^2}{L^2}\right), \tag{89}$$

which yields the desired result. □

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

First we need to compute the eigenvalues of ∇v.

Lemma 7. *Let $A \in \mathbb{R}^{m\times m}$ and*

$$M = \begin{pmatrix} 0_m & A \\ -A^T & 0_m \end{pmatrix}. \tag{90}$$

Then,

$$\operatorname{Sp} M = \{\pm i\sigma \mid \sigma^2 \in \operatorname{Sp} AA^T\}. \tag{91}$$

Proof. Assumption 1 of Appendix B.1 holds so we can apply Lem. 3 which yields the result.

□

Proof of Cor. 2. The Jacobian is constant here and has following the form:

$$\nabla v = \begin{pmatrix} 0_m & A \\ -A^T & 0_m \end{pmatrix}. \tag{92}$$

Applying Lem. 7 yields

$$\operatorname{Sp} \nabla v = \{\pm i\sigma \mid \sigma^2 \in \operatorname{Sp} AA^T\}. \tag{93}$$

Hence $\min_{\lambda\in \operatorname{Sp}\nabla v}|\lambda|^2 = \sigma_{min}(A)^2$ and $\max_{\lambda\in \operatorname{Sp}\nabla v}|\lambda|^2 = \sigma_{max}(A)^2$. Using Thm. 4, we have that,

$$\rho(\nabla F_{\eta,k}(\omega^*))^2 \leq \left(1 - \frac{1}{64}\frac{\sigma_{min}(A)^2}{\sigma_{max}(A)^2}\right). \tag{94}$$

Finally, Thm. 10 implies that the iterates of the k-extrapolation converge globally at the desired rate. □

Corollary 3. *Under the assumptions of Cor. 1, the spectral radius of the n-extrapolation method operator is bounded by*

$$\rho(\nabla F_{\eta,k}(\omega^*))^2 \leq 1 - \frac{1}{4}\Big(\frac{1}{2}\frac{\mu_1+\mu_2}{\sqrt{L_{12}^2 + L_1L_2}} + \frac{1}{16}\frac{\mu_{12}^2 + \mu_1\mu_2}{L_{12}^2 + L_1L_2}\Big). \tag{95}$$

Proof. This is a direct consequence of Thm. 4 and Thm. 9, as the latter gives that for any $\lambda \in \operatorname{Sp}\nabla v(\omega^*)$,

$$\frac{\operatorname{Tr}\mu}{2} \leq \Re\lambda, \quad |\mu| \leq |\lambda|^2 \leq |L|, \tag{96}$$

as discussed in the proof of Cor. 1. □

F Global convergence proofs

In this section, $\|.\|$ denotes the Euclidean norm.

F.1 Alternative characterizations and properties of the assumptions

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

Let us recall (10) here for simplicity:

$$\|\omega - \omega'\| \leq \gamma^{-1}\|v(\omega) - v(\omega')\| \quad \forall \omega, \omega' \in \mathbb{R}^d\,. \tag{10}$$

The proof of this lemma is an immediate consequence of a global inverse theorem from Hadamard (1906); Levy (1920). Let us recall its statement here:

Theorem 12 (Hadamard (1906); Levy (1920))**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable map. Assume that, for all $\omega \in \mathbb{R}^d$, ∇f is non-singular and $\sigma_{min}(\nabla f) \geq \gamma > 0$. Then f is a C^1-diffeomorphism, i.e. a one-to-one map whose inverse is also continuously differentiable.*

A proof of this theorem can be found in Rheinboldt (1969, Theorem 3.11). We now proceed to prove the lemma.

Proof of Lem. 2. First we prove the direct implication. By the theorem stated above, v is a bijection from $\mathbb{R}^d$ to $\mathbb{R}^d$, its inverse is continuously differentiable on $\mathbb{R}^d$ and so we have, for all $\omega \in \mathbb{R}^d$:

$$\nabla v^{-1}(v(\omega)) = (\nabla v(\omega))^{-1}\,. \tag{97}$$

Hence $\|\nabla v^{-1}(v(\omega))\| = (\sigma_{min}(\nabla v(\omega)))^{-1} \leq \gamma^{-1}$.

Consider $\omega, \omega' \in \mathbb{R}^d$ and let $u = v(\omega)$ and $u' = v(\omega')$. Then

$$\begin{aligned}
\|\omega - \omega'\| &= \|v^{-1}(u) - v^{-1}(u')\| && (98)\\
&= \left\|\int_0^1 \nabla v^{-1}(tu + (1-t)u')(u - u')\right\| && (99)\\
&\leq \gamma^{-1}\|u - u'\| && (100)\\
&= \gamma^{-1}\|v(\omega) - v(\omega')\| && (101)
\end{aligned}$$

which proves the result.

Conversely, if (10) holds, fix $u \in \mathbb{R}^d$ with $\|u\| = 1$. Taking $\omega' = \omega + tu$ in (10) with $t \neq 0$ and rearranging yields:

$$\gamma \leq \left\|\frac{v(\omega + tu) - v(\omega)}{t}\right\|\,.$$

Taking the limit when t goes to 0 gives that $\gamma \leq \|\nabla v(\omega)u\|$. As it holds for all u such that $\|u\| = 1$ this implies that $\gamma \leq \sigma_{min}(\nabla v)$. □

With the next lemma, we relate the quantities appearing in Thm. 6 to the spectrum of ∇v. Note that the first part of the proof is standard — it can be found in Facchinei and Pang (2003, Prop. 2.3.2) for instance — and we include it only for completeness.

Lemma 8. *Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and (i) μ-strongly monotone for some $\mu \geq 0$, (ii) L-Lispchitz, (iii) such that $\sigma_{min}(\nabla v) \geq \gamma$ for some $\gamma \geq 0$.. Then, for all $\omega \in \mathbb{R}^d$,*

$$\mu\|u\|^2 \leq (\nabla v(\omega)u)^T u\,, \quad \gamma\|u\| \leq \|\nabla v(\omega)u\| \leq L\|u\|\,, \quad \forall u \in \mathbb{R}^d\,, \tag{102}$$

and

$$\mu \leq \Re(\lambda), \quad \gamma \leq |\lambda| \leq L\,, \quad \forall \lambda \in \operatorname{Sp} \nabla v(\omega)\,. \tag{103}$$

Proof. By definition of μ-strong monotonicity, and L-Lispchitz one has that, for any $\omega, \omega' \in \mathbb{R}^d$,

$$\begin{aligned}
&\mu\|\omega - \omega'\|^2 \leq (v(\omega) - v(\omega'))^T(\omega - \omega') && (104)\\
&\|v(\omega) - v(\omega')\| \leq L\|\omega - \omega'\|\,. && (105)
\end{aligned}$$

Fix $\omega \in \mathbb{R}^d$, $u \in \mathbb{R}^d$ such that $\|u\| = 1$. Taking $\omega' = \omega + tu$ for $t > 0$ in the previous inequalities and dividing by t yields

$$\mu \leq \frac{1}{t}(v(\omega) - v(\omega + tu))^T u \tag{106}$$
$$\frac{1}{t}\|v(\omega) - v(\omega + tu)\| \leq L\,. \tag{107}$$

Letting t goes to 0 gives

$$\mu \leq (\nabla v(\omega)u)^T u \tag{108}$$
$$\|\nabla v(\omega)u\| \leq L\,. \tag{109}$$

Furthermore, by the properties of the singular values,

$$\|\nabla v(\omega)u\| \geq \gamma\,. \tag{110}$$

Hence, by homogeneity, we have that, for all $u \in \mathbb{R}^d$,

$$\mu\|u\|^2 \leq (\nabla v(\omega)u)^T u\,, \quad \gamma\|u\| \leq \|\nabla v(\omega)u\| \leq L\|u\|\,. \tag{111}$$

Now, take $\lambda \in \operatorname{Sp} \nabla v(\omega)$ an eigenvalue of $\nabla v(\omega)$ and let $Z \in \mathbb{C}^d \setminus \{0\}$ be one of its associated eigenvectors. Note that Z can be written as $Z = X + iY$ with $X, Y \in \mathbb{R}^d$. By definition of Z, we have

$$\nabla v(\omega)Z = \lambda Z\,. \tag{112}$$

Now, taking the real and imaginary part yields:

$$\begin{cases} \nabla v(\omega)X & = \Re(\lambda)X - \Im(\lambda)Y \\ \nabla v(\omega)Y & = \Im(\lambda)X + \Re(\lambda)Y \end{cases} \tag{113}$$

Taking the squared norm and developing the right-hand sides yields

$$\begin{cases} \|\nabla v(\omega)X\|^2 & = \Re(\lambda)^2\|X\|^2 + \Im(\lambda)^2\|Y\|^2 - 2\Re(\lambda)\Im(\lambda)X^TY \\ \|\nabla v(\omega)Y\|^2 & = \Im(\lambda)^2\|X\|^2 + \Re(\lambda)^2\|Y\|^2 + 2\Re(\lambda)\Im(\lambda)X^TY\,. \end{cases} \tag{114}$$

Now summing these two equations gives

$$\|\nabla v(\omega)X\|^2 + \|\nabla v(\omega)Y\|^2 = |\lambda|^2(\|X\|^2 + \|Y\|^2)\,. \tag{115}$$

Finally, apply (111) for $u = X$ and $u = Y$:

$$\gamma^2(\|X\|^2 + \|Y\|^2) \leq |\lambda|^2(\|X\|^2 + \|Y\|^2) \leq L^2(\|X\|^2 + \|Y\|^2)\,. \tag{116}$$

As $Z \neq 0$, $\|X\|^2 + \|Y\|^2 > 0$ and this yields $\gamma \leq |\lambda| \leq L$. To get the inequality concerning γ, multiply on the left the first line of (113) by X^T and the second one by Y^T:

$$\begin{cases} X^T(\nabla v(\omega)X) & = \Re(\lambda)\|X\|^2 - \Im(\lambda)X^TY \\ Y^T(\nabla v(\omega)Y) & = \Im(\lambda)Y^TX + \Re(\lambda)\|Y\|^2\,. \end{cases} \tag{117}$$

Again, summing these two lines and using (111) yields:

$$\mu(\|X\|^2 + \|Y\|^2) \leq \Re(\lambda)(\|X\|^2 + \|Y\|^2)\,. \tag{118}$$

As $Z \neq 0$, $\|X\|^2 + \|Y\|^2 > 0$ and so $\mu \leq \Re(\lambda)$. □

F.2 Proofs of §6: extragradient, optimistic and proximal point methods

We now prove a slightly more detailed version of Thm. 6.

Theorem 13. *Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable and (i) μ-strongly monotone for some $\mu \geq 0$, (ii) L-Lipschitz, (iii) such that $\sigma_{min}(\nabla v) \geq \gamma$ for some $\gamma > 0$. Then, for $\eta \leq (4L)^{-1}$, the iterates of the extragradient method $(\omega_t)_t$ converge linearly to ω^* the unique stationary point of v,*

$$\|\omega_t - \omega^*\|_2^2 \leq \left(1 - \left(\eta\mu + \tfrac{7}{16}\eta^2\gamma^2\right)\right)^t \|\omega_0 - \omega^*\|_2^2 . \tag{119}$$

For $\eta = (4L)^{-1}$, this can be simplified as: $\|\omega_t - \omega^*\|_2^2 \leq \left(1 - \frac{1}{4}\left(\frac{\mu}{L} + \frac{1}{16}\frac{\gamma^2}{L^2}\right)\right)^t \|\omega_0 - \omega^*\|_2^2$.

The proof is inspired from the ones of Gidel et al. (2019a); Tseng (1995).

We will use the following well-known identity. It can be found in Gidel et al. (2019a) for instance but we state it for reference.

Lemma 9. *Let $\omega, \omega', u \in \mathbb{R}^d$. Then*

$$\|\omega + u - \omega'\|^2 = \|\omega - \omega'\|^2 + 2u^T(\omega + u - \omega') - \|u\|^2 \tag{120}$$

Proof.

$$\begin{aligned} \|\omega + u - \omega'\|^2 &= \|\omega - \omega'\|^2 + 2u^T(\omega - \omega') + \|u\|^2 && (121)\\ &= \|\omega - \omega'\|^2 + 2u^T(\omega + u - \omega') - \|u\|^2 && (122) \end{aligned}$$

□

Proof Thm. 13. First note that as $\gamma > 0$, by Thm. 12, v has a stationary point ω^* and it is unique.

Fix any $\omega_0 \in \mathbb{R}^d$, and denote $\omega_1 = \omega_0 - \eta v(\omega_0)$ and $\omega_2 = \omega_0 - \eta v(\omega_1)$. Applying Lem. 9 for $(\omega, \omega', u) = (\omega_0, \omega^*, -\eta v(\omega_1))$ and $(\omega, \omega', u) = (\omega_0, \omega_2, -\eta v(\omega_0))$ yields:

$$\begin{aligned} \|\omega_2 - \omega^*\|^2 &= \|\omega_0 - \omega^*\|^2 - 2\eta v(\omega_1)^T(\omega_2 - \omega^*) - \|\omega_2 - \omega_0\|^2 && (123)\\ \|\omega_1 - \omega_2\|^2 &= \|\omega_0 - \omega_2\|^2 - 2\eta v(\omega_0)^T(\omega_1 - \omega_2) - \|\omega_1 - \omega_0\|^2 && (124) \end{aligned}$$

Summing these two equations gives:

$$\begin{aligned} &\|\omega_2 - \omega^*\|^2 = && (125)\\ &\|\omega_0 - \omega^*\|^2 - 2\eta v(\omega_1)^T(\omega_2 - \omega^*) - 2\eta v(\omega_0)^T(\omega_1 - \omega_2) - \|\omega_1 - \omega_0\|^2 - \|\omega_1 - \omega_2\|^2 && (126) \end{aligned}$$

Then, rearranging and using that $v(\omega^*) = 0$ yields that,

$$\begin{aligned} &2\eta v(\omega_1)^T(\omega_2 - \omega^*) + 2\eta v(\omega_0)^T(\omega_1 - \omega_2) && (127)\\ &= 2\eta(v(\omega_1))^T(\omega_1 - \omega^*) + 2\eta(v(\omega_0) - v(\omega_1))^T(\omega_1 - \omega_2) && (128)\\ &= 2\eta(v(\omega_1) - v(\omega^*))^T(\omega_1 - \omega^*) + 2\eta(v(\omega_0) - v(\omega_1))^T(\omega_1 - \omega_2) && (129)\\ &\geq 2\eta\mu\|\omega_1 - \omega^*\|^2 - 2\eta\|v(\omega_0) - v(\omega_1)\|\|\omega_1 - \omega_2\| && (130) \end{aligned}$$

where the first term is lower bounded using strong monotonicity and the second one using Cauchy-Schwarz's inequality. Using in addition the fact that v is Lipschitz continuous we obtain:

$$\begin{aligned} &2\eta v(\omega_1)^T(\omega_2 - \omega^*) + 2\eta v(\omega_0)^T(\omega_1 - \omega_2) && (131)\\ &\geq 2\eta\mu\|\omega_1 - \omega^*\|^2 - 2\eta L\|\omega_0 - \omega_1\|\|\omega_1 - \omega_2\| && (132)\\ &\geq 2\eta\mu\|\omega_1 - \omega^*\|^2 - (\eta^2 L^2\|\omega_0 - \omega_1\|^2 + \|\omega_1 - \omega_2\|^2) && (133) \end{aligned}$$

where the last inequality comes from Young's inequality. Using this inequality in (125) yields:

$$\|\omega_2 - \omega^*\|^2 \leq \|\omega_0 - \omega^*\|^2 - 2\eta\mu\|\omega_1 - \omega^*\|^2 + (\eta^2 L^2 - 1)\|\omega_0 - \omega_1\|^2 . \tag{134}$$

Now we lower bound $\|\omega_1 - \omega^*\|$ using $\|\omega_0 - \omega^*\|$. Indeed, from Young's inequality we obtain

$$2\|\omega_1 - \omega^*\|^2 \geq \|\omega_0 - \omega^*\|^2 - 2\|\omega_0 - \omega_1\|^2 \,. \tag{135}$$

Hence, we have that,

$$\|\omega_2 - \omega^*\|^2 \leq (1 - \eta\mu)\|\omega_0 - \omega^*\|^2 + (\eta^2 L^2 + 2\eta\mu - 1)\|\omega_0 - \omega_1\|^2 \,. \tag{136}$$

Note that if $\eta \leq \frac{1}{4L}$, as $\mu \leq L$, $\eta^2 L^2 + 2\eta\mu - 1 \leq -\frac{7}{16}$. Therefore, with $c = \frac{7}{16}$,

$$\|\omega_2 - \omega^*\|^2 \leq (1 - \eta\mu)\|\omega_0 - \omega^*\|^2 - c\|\omega_0 - \omega_1\|^2 \tag{137}$$

$$= (1 - \eta\mu)\|\omega_0 - \omega^*\|^2 - c\eta^2\|v(\omega_0)\|^2 \,. \tag{138}$$

Finally, using (iii) and Lem. 2, we obtain:

$$\|\omega_2 - \omega^*\|^2 \leq (1 - \eta\mu - c\eta^2\gamma^2)\|\omega_0 - \omega^*\|^2 \tag{139}$$

which yields the result. □

Proposition 1. *Under the assumptions of Thm. 6, the iterates of the proximal point method method* $(\omega_t)_t$ *with* $\eta > 0$ *converge linearly to* ω^* *the unique stationary point of* v,

$$\|\omega_t - \omega^*\|^2 \leq \left(1 - \frac{2\eta\mu + \eta^2\gamma^2}{1 + 2\eta\mu + \eta^2\gamma^2}\right)^t \|\omega_0 - \omega^*\|^2 \quad \forall t \geq 0 \,. \tag{140}$$

Proof. To proof this convergence result, we upper bound the singular values of the proximal point operator P_η. As v is monotone, by Lem. 8, the eigenvalues of ∇v have all non-negative real parts everywhere. As in the proof of Lem. 1, $\omega \mapsto \omega + \eta v(\omega)$ is continuously differentiable and has a non-singular differential at every $\omega_0 \in \mathbb{R}^d$. By the inverse function theorem, $\omega \mapsto \omega + \eta v(\omega)$ has a continuously differentiable inverse in a neighborhood of ω_0. Its inverse is exactly P_η and it also satisfies

$$\nabla P_\eta(\omega_0) = (I_d + \eta\nabla v(\omega_0))^{-1} \,. \tag{141}$$

The singular values $\nabla P_\eta(\omega_0)$ are the eigenvalues of $(\nabla P_\eta(\omega_0))^T(\nabla P_\eta(\omega_0))$. The latter is equal to:

$$(\nabla P_\eta(\omega_0))^T(\nabla P_\eta(\omega_0)) = \left(I_d + \eta\nabla v(\omega_0) + \eta(\nabla v(\omega_0))^T + \eta^2(\nabla v(\omega_0))^T(\nabla v(\omega_0))\right)^{-1} \,. \tag{142}$$

Now, let $\lambda \in \mathbb{R}$ be an eigenvalue of $(\nabla P_\eta(\omega_0))^T(\nabla P_\eta(\omega_0))$ and let $X \neq 0$ be one of its associated eigenvectors. As $\nabla P_\eta(\omega_0)$ is non-singular, $\lambda \neq 0$ and applying the previous equation yields:

$$\lambda^{-1}X = \left(I_d + \eta\nabla v(\omega_0) + \eta(\nabla v(\omega_0))^T + \eta^2(\nabla v(\omega_0))^T(\nabla v(\omega_0))\right) X \,. \tag{143}$$

Finally, multiply this equation on the left by X^T:

$$\lambda^{-1}\|X\|^2 = \|X\|^2 + \eta X^T(\nabla v(\omega_0) + (\nabla v(\omega_0))^T)X + \eta^2\|\nabla v(\omega_0))X\|^2 \,. \tag{144}$$

Applying the first part of Lem. 8 yields

$$\lambda^{-1}\|X\|^2 \geq (1 + 2\eta\mu + \eta^2\gamma^2)\|X\|^2 \,. \tag{145}$$

Hence, as $X \neq 0$, we have proven that,

$$\sigma_{max}(\nabla v(\omega_0)) \leq (1 + 2\eta\mu + \eta^2\gamma^2)^{-1} \,. \tag{146}$$

This implies that, for all $\omega, \omega' \in \mathbb{R}^d$,

$$\|P_\eta(\omega) - P_\eta(\omega')\|^2 = \left\|\int_0^1 \nabla v(\omega' + t(\omega - \omega'))(\omega - \omega')\right\|^2 \tag{147}$$

$$\leq (1 + 2\eta\mu + \eta^2\gamma^2)^{-1}\|\omega - \omega'\|^2 \,. \tag{148}$$

Hence, as $P_\eta(\omega^*) = \omega^*$, taking $\omega' = \omega^*$ gives the desired global convergence rate. □

Now let us prove the result Thm. 7 regarding Optimistic method.

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

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

Proof. For the beginning of this proof we follow the proof of Gidel et al. (2019a, Theorem 1) using their notation:

$$\omega'_t = \omega_t - \eta v(\omega'_{t-1}) \tag{149}$$

$$\omega_{t+1} = \omega_t - \eta v(\omega'_t) \tag{150}$$

Note that, with this notation, summing the two upates steps, we recover (OG)

$$\omega'_{t+1} = \omega'_t - 2\eta v(\omega'_t) + \eta v(\omega'_{t-1}) \,. \tag{151}$$

Let us now recall Gidel et al. (2019a, Equation 88) for a constant step-size $\eta_t = \eta$,

$$\begin{aligned} \|\omega_{t+1} - \omega^*\|_2^2 \leq (1-\eta\mu)\, \|\omega_t - \omega^*\|_2^2 &+ \eta^2 L^2 (4\eta^2 L^2 \|\omega'_{t-1} - \omega'_{t-2}\|_2^2 - \|\omega'_{t-1} - \omega'_t\|_2^2) \\ &- (1 - 2\eta\mu - 4\eta^2 L^2)\|\omega'_t - \omega_t\|_2^2 \end{aligned} \tag{152}$$

we refer the reader to the proof of Gidel et al. (2019a, Theorem 1) for the details on how to get to this equation. Thus with $\eta \leq (4L)^{-1}$, using the update rule $\omega'_t = \omega_t - \eta v(\omega'_{t-1})$, we get,

$$(1 - 2\eta_t\mu - 4\eta_t^2 L^2)\|\omega'_t - \omega_t\|_2^2 \geq \frac{1}{4}\|\omega'_t - \omega_t\|_2^2 = \frac{\eta^2}{4}\|v(\omega'_{t-1})\|_2^2 \geq \frac{\eta^2\gamma^2}{4}\|\omega'_{t-1} - \omega^*\|_2^2 \tag{153}$$

where for the last inequality we used that $\sigma_{\min}(\nabla v) \geq \gamma$ and Lemma 2. Using Young's inequality, the update rule and the Lipchitzness of v, we get that,

$$\begin{aligned} 2\|\omega'_{t-1} - \omega^*\|_2^2 &\geq \|\omega_t - \omega^*\|_2^2 - 2\|\omega'_{t-1} - \omega_t\|_2^2 && (154)\\ &= \|\omega_t - \omega^*\|_2^2 - 2\eta^2\|v(\omega'_{t-1}) - v(\omega'_{t-2})\|_2^2 && (155)\\ &\geq \|\omega_t - \omega^*\|_2^2 - 2\eta^2 L^2\|\omega'_{t-1} - \omega'_{t-2}\|_2^2 && (156) \end{aligned}$$

Thus combining (152), (153) and (156), we get with a constant step-size $\eta \leq (4L)^{-1}$,

$$\|\omega_{t+1} - \omega^*\|_2^2 \leq \left(1 - \eta\mu - \frac{\eta^2\gamma^2}{8}\right)\|\omega_t - \omega^*\|_2^2 + \eta^2 L^2\big((4\eta^2 L^2 + \tfrac{\eta^2\gamma^2}{4})\|\omega'_{t-1} - \omega'_{t-2}\|_2^2 - \|\omega'_{t-1} - \omega'_t\|_2^2\big)$$

This leads to,

$$\|\omega_{t+1} - \omega^*\|_2^2 + \eta^2 L^2\|\omega'_{t-1} - \omega'_t\|_2^2 \leq \left(1 - \eta\mu - \frac{\eta^2\gamma^2}{8}\right)\|\omega_t - \omega^*\|_2^2 + \eta^2(4L^2 + \tfrac{\gamma^2}{4})\eta^2 L^2\|\omega'_{t-1} - \omega'_{t-2}\|_2^2 \tag{157}$$

In order to get the theorem statement we need a rate on ω'_t. We first unroll this geometric decrease and notice that

$$\begin{aligned} \|\omega'_t - \omega^*\|_2^2 &\leq 2\|\omega_{t+1} - \omega^*\|_2^2 + 2\|\omega'_t - \omega_{t+1}\|_2^2 && (158)\\ &= 2\|\omega_{t+1} - \omega^*\|_2^2 + 2\eta^2\|v(\omega'_{t-1}) - v(\omega'_t)\|_2^2 && (159)\\ &= 2\|\omega_{t+1} - \omega^*\|_2^2 + 2\eta^2 L^2\|\omega'_{t-1} - \omega'_t\|_2^2 && (160) \end{aligned}$$

to get (using the fact that $\omega'_0 = \omega'_{-1}$),

$$\begin{aligned} \|\omega'_t - \omega^*\|_2^2 &\leq 2\|\omega_{t+1} - \omega^*\|_2^2 + 2\eta^2 L^2\|\omega'_{t-1} - \omega'_t\|_2^2 && (161)\\ &\leq 2\max\left\{1 - \eta\mu - \frac{\eta^2\gamma^2}{8L^2}, 4\eta^2 L^2 + \frac{\eta^2\gamma^2}{4}\right\}^{t+1}\|\omega_0 - \omega^*\|_2^2 \,. && (162) \end{aligned}$$

With $\eta \leq (4L)^{-1}$ we can use the fact that $\max(\mu, \gamma) \leq L$ to get,

$$1 - \eta\mu - \frac{\eta^2\gamma^2}{8} \geq 1 - \frac{1}{4}\Big(\frac{\mu}{L} + \frac{\gamma^2}{32L^2}\Big) \geq \frac{3}{4} - \frac{1}{32} \geq \frac{1}{4} + \frac{1}{64} \geq 4\eta^2 L^2 + \frac{\eta^2\gamma^2}{4} \tag{163}$$

Thus,

$$\max\left\{1-\eta\mu-\frac{\eta^2\gamma^2}{8}, 4\eta^2L^2+\frac{\eta^2\gamma^2}{4}\right\} = 1-\eta\mu-\frac{\eta^2\gamma^2}{8}\,, \qquad \forall \eta \le (4L)^{-1} \tag{164}$$

leading to the statement of the theorem. Finally, for $\eta = (4L)^{-1}$ that can be simplified into,

$$\|\omega'_t-\omega^*\|_2^2 \le 2\max\left\{1-\frac{1}{4}\Big(\frac{\mu}{L}+\frac{\gamma^2}{32L^2}\Big), \frac{1}{4}+\frac{\gamma^2}{64L^2}\right\}^{t+1}\|\omega_0-\omega^*\|_2^2\,. \tag{165}$$

$$= 2\Big(1-\frac{1}{4}\Big(\frac{\mu}{L}+\frac{\gamma^2}{32L^2}\Big)\Big)^{t+1}\|\omega_0-\omega^*\|_2^2 \tag{166}$$

□

F.3 Proof of §6.3: consensus optimization

Let us recall (CO) here,

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

where H is the squared norm of v. We prove a more detailed version Thm. 8.

Theorem 14. *Let $v : \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable such that (i) v is μ- strongly monotone for some $\mu \ge 0$, (ii) $\sigma_{min}(\nabla v) \ge \gamma$ for some $\gamma > 0$ (iii) H is L_H^2 Lipschitz-smooth. Then, for*

$$\alpha^2 \le \frac{1}{2}\left(\frac{\alpha\mu}{L_H}+\frac{\beta\gamma^2}{2L_H}\right)\,,$$

and $\beta \le (2L_H)^{-1}$, the iterates of CO defined by (CO) satisfy,

$$H(\omega_t) \le \left(1-\alpha\mu-\tfrac{1}{2}\beta\gamma^2\right)H(\omega_0)\,. \tag{167}$$

In particular, for

$$\alpha = \frac{\mu+\sqrt{\mu^2+2\gamma^2}}{4L_H}\,,$$

and $\beta = (2L_H)^{-1}$,

$$H(\omega_t) \le \left(1-\frac{\mu^2}{2L_H}-\frac{\gamma^2}{2L_H^2}\left(1+\frac{\mu}{\gamma}\right)\right)H(\omega_0)\,. \tag{168}$$

Proof. As H is L_H^2 Lipschitz smooth, we have,

$$H(\omega_{t+1})-H(\omega_t) \le \nabla H(\omega_t)^T(\omega_{t+1}-\omega_t)+\frac{L_H^2}{2}\|\omega_{t+1}-\omega_t\|^2\,.$$

Then, replacing $\omega_{t+1}-\omega_t$ by its expression and using Young's inequality,

$$H(\omega_{t+1})-H(\omega_t) \le -\alpha\nabla H(\omega_t)^Tv(\omega_t)-\beta\|\nabla H(\omega_t)\|^2+L_H^2\alpha^2\|v(\omega_t)\|^2+L_H^2\beta^2\|\nabla H(\omega_t)\|^2\,.$$

Note that, crucially, $\nabla H(\omega_t) = \nabla v(\omega_t)^Tv(\omega_t)$. Using the first part of Lem. 8 to introduce μ and assuming $\beta \le (2L_H^2)^{-1}$,

$$H(\omega_{t+1})-H(\omega_t) \le -\alpha\mu\|v(\omega_t)\|^2-\frac{\beta}{2}\|\nabla H(\omega_t)\|^2+L_H^2\alpha^2\|v(\omega_t)\|^2\,.$$

Finally, using Lem. 8 to introduce γ,

$$\begin{aligned} H(\omega_{t+1})-H(\omega_t) &\le -\alpha\mu\|v(\omega_t)\|^2-\frac{\beta\gamma^2}{2}\|v(\omega_t)\|^2+L_H^2\alpha^2\|v(\omega_t)\|^2 \\ &= -2\left(\alpha\mu+\frac{\beta\gamma^2}{2}-L_H^2\alpha^2\right)H(\omega_t)\,. \end{aligned}$$

Hence, if

$$\alpha^2 \leq \frac{1}{2}\left(\frac{\alpha\mu}{L_H^2} + \frac{\beta\gamma^2}{2L_H^2}\right), \tag{169}$$

then the decrease of H becomes,

$$H(\omega_t) \leq \left(1 - \alpha\mu - \tfrac{1}{2}\beta\gamma^2\right) H(\omega_0).$$

Now, note that (169) is a second-order polynomial condition on α, so we can compute the biggest α which satisfies this condition. This yields,

$$\begin{aligned}\alpha &= \frac{\frac{\mu}{2} + \sqrt{\frac{\mu^2}{4} + L_H^2\beta\gamma^2}}{2L_H^2} \\ &= \frac{\mu + \sqrt{\mu^2 + 2\gamma^2}}{4L_H^2},\end{aligned}$$

where in the second line we defined $\beta = (2L_H^2)^{-1}$. Then the rate becomes,

$$\begin{aligned}\alpha\mu + \frac{1}{2}\beta\gamma^2 &= \frac{\mu^2}{4L_H^2} + \frac{\mu\sqrt{\mu^2 + 2\gamma^2}}{4L_H^2} + \frac{\gamma^2}{4L_H^2} \\ &\geq \frac{\mu^2}{4L_H^2} + \frac{\mu^2}{4\sqrt{2}L_H^2} + \frac{\mu\gamma}{4L_H^2} + \frac{\gamma^2}{4L_H^2},\end{aligned}$$

where we use Young's inequality: $\sqrt{2}\sqrt{a+b} \geq \sqrt{a} + \sqrt{b}$. Noting that $\frac{1}{2}(1 + \frac{1}{\sqrt{2}}) \geq 1$ yields the result. □

Remark 1. A common convergence result for the gradient method for variational inequalities problem – see Nesterov and Scrimali (2006) for instance – is that the iterates convergence as $O\left(\left(1 - \frac{\mu^2}{L^2}\right)^t\right)$ where μ is the monotonicty constant of v and L its Lipschitz constant. However, this rate is not optimal, and also not satisfying as it does not recover the convergence rate of the gradient method for strongly convex optimization. One way to remedy this situation is to use the *co-coercivity* or *inverse strong monotonicity* assumption:

$$\ell(v(\omega) - v(\omega'))^T(\omega - \omega') \geq \|v(\omega) - v(\omega')\|^2 \quad \forall \omega, \omega'.$$

This yields a convergence rate of $O\left(\left(1 - \frac{\mu}{\ell}\right)^t\right)$ which can be significantly better than the former since ℓ can take all the values of $[L, L^2/\mu]$ (Facchinei and Pang, 2003, §12.1.1). On one hand, if v is the gradient of a convex function, $\ell = L$ and so we recover the standard rate in this case. On the other, one can consider for example the operator $v(w) = Aw$ with $A = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ with $a > 0$ and $b \neq 0$ for which $\mu = a, L = \sqrt{a^2 + b^2}$ anf $\ell = \mu/L^2$.

G The p-SCLI framework for game optimization

The approach we use to prove our lower bounds comes from Arjevani et al. (2016). Though their whole framework was developed for convex optimization, a careful reading of their proof shows that most of their results carry on to games, at least those in their first three sections. However, we work only in the restricted setting of 1-SCLI and so we actually rely on a very small subset of their results, more exactly two of them.

The first one is Thm. 2 and is crucially used in the derivation of our lower bounds. We state it again for clarity.

Theorem 2 (Arjevani et al. (2016))**.** *For all $v \in \mathcal{V}_d$, for almost every[6] initialization point $\omega_0 \in \mathbb{R}^d$, if $(\omega_t)_t$ are the iterates of $F_\mathcal{N}$ starting from ω_0,*

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

Actually, as $F_\mathcal{N} : \mathbb{R}^d \to \mathbb{R}^d$ is an affine operator and ω^* is one of its fixed point, this theorem is only a reformulation of Arjevani et al. (2016, Lemma 10), which is a standard result in linear algebra. So we actually do not rely on their most advanced results which were proven only for convex optimization problems. For completeness, we state this lemma below and show how to derive Thm. 2 from it.

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

Lemma 10 (Arjevani et al. (2016, Lemma 10)). *Let $A \in \mathbb{R}^{d\times d}$. There exists $c > 0$, $d \geq m \geq 1$ integer and $r \in \mathbb{R}^d$, $r \neq 0$ such that for any $u \in \mathbb{R}^d$ such that $u^T r \neq 0$, for sufficiently large $t \geq 1$ one has:*

$$\|A^t u\| \geq ct^{m-1}\rho(A)^t \|u\| \,. \tag{170}$$

Proof of Thm. 2. $F_\mathcal{N}$ is affine so it can be written as $F_\mathcal{N}(\omega) = \nabla F_\mathcal{N}\omega + F_\mathcal{N}(0)$.

Moreover, as $v(\omega^*) = 0$, $F_\mathcal{N}(\omega^*) = \omega^* + \mathcal{N}(\nabla v)v(\omega^*) = \omega^*$. Hence, for all $\omega \in \mathbb{R}^d$,

$$F_\mathcal{N}(\omega) - \omega^* = F_\mathcal{N}(\omega) - F_\mathcal{N}(\omega^*) = \nabla F_\mathcal{N}(\omega - \omega^*) \,. \tag{171}$$

Therefore, for $t \geq 0$,

$$\|\omega_t - w^*\| = \|(\nabla F_\mathcal{N})^t(\omega - \omega^*)\| \,. \tag{172}$$

Finally, apply the lemma above to $A = \nabla F_\mathcal{N}$. The condition $(\omega_0 - \omega^*)^T r \neq 0$ is not satisfied only on an affine subset of dimension 1, which is of measure zero for any measure absolutely continuous w.r.t. the Lebesgue measure. Hence for almost every $\omega_0 \in \mathbb{R}^d$ w.r.t. to such measure, $(\omega_0 - \omega^*)^T r \neq 0$ and so one has, for $t \geq 1$ large enough,

$$\begin{aligned} \|\omega_t - w^*\| &\geq ct^{m-1}\rho(\nabla F_\mathcal{N})^t \|\omega_t - \omega^*\| && (173)\\ &\geq c\rho(\nabla F_\mathcal{N})^t \|\omega_t - \omega^*\| \,, && (174) \end{aligned}$$

which is the desired result. □

The other result we use is more anecdotal : it is their consistency condition, which is a necessary condition for an p-SCLI method to converge to a stationary point of the gradient dynamics. Indeed, general 1-SCLI as defined in Arjevani et al. (2016) are given not by one but by two mappings $\mathcal{C}, \mathcal{N} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ and the update rule is

$$F_\mathcal{N}(\omega) = \mathcal{C}(\nabla v)w + \mathcal{N}(\nabla v)v(0) \quad \forall \omega \in \mathbb{R}^d \,. \tag{175}$$

However, they show in Arjevani et al. (2016, Thm. 5) that, for a method to converge to a stationary point of v, at least for convex problems, that is to say symmetric positive semi-definite ∇v, $\mathcal{C}$ and $\mathcal{N}$ need to satisfy:

$$I_d - \mathcal{C}(\nabla v) = -\mathcal{N}(\nabla v)\nabla v \,. \tag{176}$$

If $\mathcal{C}$ and $\mathcal{N}$ are polynomials, this equality for all symmetric positive semi-definite ∇v implies the equality on all matrices. Injecting this result in (175) yields the definition of 1-SCLI we used.

H Proofs of lower bounds

The class of methods we consider, that is to say the methods whose coefficient mappings $\mathcal{N}$ are any polynomial of degree at most $k-1$, is very general. It includes:

- the k'-extrapolation methods $F_{k',\eta}$ for $k' \leq k$ as defined by (5).
- extrapolation methods with different step sizes for each extrapolation:

$$\omega \longmapsto \varphi_{\eta_1,\omega} \circ \varphi_{\eta_2,\omega} \circ \cdots \circ \varphi_{\eta_k,\omega}(\omega) \,, \tag{177}$$

- cyclic Richardson iterations (Opfer and Schober, 1984): methods whose update is composed of successive gradient steps with possibly different step sizes for each

$$\omega \longmapsto F_{\eta_1} \circ F_{\eta_2} \circ \cdots \circ F_{\eta_k}(\omega) \,, \tag{178}$$

and any combination of these with at most k composed gradient evaluations.

The lemma below shows how k-extrapolation algorithms fit into the definition of 1-SCLI:

Lemma 11. *For a k-extrapolation method, $\mathcal{N}(\nabla v) = -\eta \sum_{j=0}^{k-1}(-\eta\nabla v)^k$.*

Proof. This result is a direct consequence of Lem. 6. For $\omega \in \mathbb{R}^d$, one gets, by the chain rule,

$$\nabla F_{\eta,k}(\omega) = \nabla_z \varphi_\omega^k(\omega) + \nabla_\omega \varphi_\omega^k(\omega) \tag{179}$$

$$= (-\eta \nabla v)^k + \sum_{j=0}^{k-1} (-\eta \nabla v)^j \tag{180}$$

$$= \sum_{j=0}^{k} (-\eta \nabla v)^j \,. \tag{181}$$

as ∇v is constant. Hence, as expected, $F_{\eta,k}$ is linear so write that, for all $\omega \in \mathbb{R}^d$,

$$F_{\eta,k}(\omega) = \nabla F_{\eta,k}\omega + b\,. \tag{182}$$

If v has a stationary point ω^*, evaluating at ω^* yields

$$\omega^* = \sum_{j=0}^{k} (-\eta \nabla v)^j \omega^* + b\,. \tag{183}$$

Using that $v(\omega^*) = 0$ and so $(\nabla v)\omega^* = -v(0)$, one gets that

$$b = -\eta \sum_{j=1}^{k} (-\eta \nabla v)^{j-1} v(0)\,, \tag{184}$$

and so

$$F_{\eta,k}(\omega) = \omega - \eta \sum_{j=1}^{k} (-\eta \nabla v)^{j-1} v(\omega)\,, \tag{185}$$

which yields the result for affine vector fields with a stationary point. In particular it holds for vector fields such that ∇v is non-singular. As the previous equality is continuous in ∇v, by density of non-singular matrices, the result holds for all affine vector fields.

□

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

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

To ease the presentation of the proof of the theorem, we rely on several lemmas. We first prove (i) and (ii) will follow as a consequence.

In the following, we denote by $\mathbb{R}_{k-1}[X]$ the set of real polynomials of degree at most $k-1$.

Lemma 12. *For, $v \in \mathcal{V}_d$,*

$$\min_{N \in \mathbb{R}_{k-1}[X]} \frac{1}{2}\rho(F_\mathcal{N})^2 = \min_{a_0,\dots,a_{k-1}\in\mathbb{R}} \max_{\lambda \in \operatorname{Sp}\nabla v} \frac{1}{2}|1 + \sum_{l=0}^{k-1} a_l \lambda^{l+1}|^2\,. \tag{186}$$

Proof. Recall the definition of $F_\mathcal{N}$, which is affine by assumption,

$$\forall \omega \in \mathbb{R}^d,\; F_\mathcal{N}(\omega) = w + \mathcal{N}(\nabla v)v(\omega)\,. \tag{187}$$

Then $\nabla F_\mathcal{N} = I_d + \mathcal{N}(\nabla v)\nabla v$. As $\mathcal{N}$ is a polynomial, by the spectral mapping theorem (Thm. 11),

$$\operatorname{Sp}\nabla F_\mathcal{N} = \{1 + \mathcal{N}(\lambda)\lambda \mid \lambda \in \operatorname{Sp}\nabla v\}\,, \tag{188}$$

which yields the result. □

Lemma 13. *Assume that* $\mathrm{Sp}\,\nabla v = \{\lambda_1, \dots, \lambda_m\} \subset \mathbb{R}$. *Then* (186) *can be lower bounded by the value of the following problem:*

$$\begin{aligned} &\max \sum_{j=1}^m \nu_j (\xi_j - \frac{1}{2}\xi_j^2) \\ &s.t. \ \ \nu_j \geq 0, \quad \xi_j \in \mathbb{R}, \ \forall 1 \leq j \leq m \\ &\qquad \sum_{j=1}^m \nu_j \xi_j \lambda_j^l = 0, \ \forall 1 \leq l \leq k \\ &\qquad \sum_{j=1}^m \nu_j = 1 \end{aligned} \tag{189}$$

Proof. The right-hand side of (186) can be written as a constrained optimization problem as follows:

$$\begin{aligned} &\min_{t, a_0, \dots, a_{k-1}, z_1, \dots, z_m \in \mathbb{R}} t \\ &\text{s.t.} \ \ t \geq \frac{1}{2} z_j^2, \ \forall\ 1 \leq j \leq m \\ &z_j = 1 + \sum_{l=0}^{k-1} a_l \lambda_j^{l+1}, \ \forall\ 1 \leq j \leq m\,. \end{aligned} \tag{190}$$

By weak duality, see Boyd and Vandenberghe (2004) for instance, we can lower bound the value of this problem by the value of its dual. So let us write the Lagrangian of this problem:

$$\begin{aligned} &\mathcal{L}(t, a_0, \dots, a_{k-1}, z_1, \dots, z_m, \nu_1, \dots, \nu_m, \chi_1, \dots, \chi_m) \\ &= t + \sum_{j=0}^m \nu_j (\frac{1}{2} z_j^2 - t) + \chi_j (1 + \sum_{l=0}^{k-1} a_l \lambda_j^{l+1} - z_j)\,. \end{aligned} \tag{191}$$

The Lagrangian is convex and quadratic so its minimum with respect to $t, a_0, \dots, a_{k-1}, z_1, \dots, z_m$ is characterized by the first order condition. Moreover, if there is no solution to the first order condition, its minimum is $-\infty$ (see for instance Boyd and Vandenberghe (2004, Example 4.5)).

One has that, for any $1 \leq j \leq m$ and $0 \leq l \leq k-1$,

$$\partial_t \mathcal{L} = 1 - \sum_{j=0}^m \nu_j \tag{192}$$

$$\partial_{a_l} \mathcal{L} = \sum_{j=0}^m \chi_j \lambda_j^{l+1} \tag{193}$$

$$\partial_{z_j} \mathcal{L} = \nu_j z_j - \chi_j\,. \tag{194}$$

Setting these quantities to zero yields the following dual problem:

$$\begin{aligned} &\max \sum_{j=1,\ \nu_j \neq 0}^m \chi_j - \frac{1}{2\nu_j} \chi_j^2 \\ &\text{s.t.} \ \ \nu_j \geq 0, \quad \chi_j \in \mathbb{R}, \ \forall 1 \leq j \leq m \\ &\qquad \sum_{j=1}^m \chi_j \lambda_j^l = 0, \ \forall 1 \leq l \leq k \\ &\qquad \nu_j = 0 \implies \chi_j = 0 \\ &\qquad \sum_{j=1}^m \nu_j = 1 \end{aligned} \tag{195}$$

Taking $\nu_j \xi_j = \chi_j$ yields the result:

$$\begin{aligned} &\max \sum_{j=1}^m \nu_j (\xi_j - \frac{1}{2}\xi_j^2) \\ &\text{s.t. } \ \nu_j \geq 0, \ \ \xi_j \in \mathbb{R}, \ \forall 1 \leq j \leq m \\ &\qquad \sum_{j=1}^m \nu_j \xi_j \lambda_j^l = 0, \ \forall 1 \leq l \leq k \\ &\qquad \sum_{j=1}^m \nu_j = 1\,. \end{aligned} \tag{196}$$

□

The next lemma concerns Vandermonde matrices and Lagrange polynomials.

Lemma 14. *Let $\lambda_1, \dots, \lambda_d$ be distinct reals. Denote the Vandermonde matrix by*

$$V(\lambda_1, \dots, \lambda_d) = \begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{d-1} \\ 1 & \lambda_2 & \lambda_2^2 & \dots & \lambda_2^{d-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{d-1} \end{pmatrix}. \tag{197}$$

Then

$$V(\lambda_1, \dots, \lambda_d)^{-1} = \begin{pmatrix} L_1^{(0)} & L_2^{(0)} & \dots & L_d^{(0)} \\ L_1^{(1)} & L_2^{(1)} & \dots & L_d^{(1)} \\ \vdots & \vdots & & \vdots \\ L_1^{(d-1)} & L_2^{(d-1)} & \dots & L_d^{(d-1)} \end{pmatrix} \tag{198}$$

where $L_1, L_2, \dots, L_d$ are the Lagrange interpolation polynomials associated to $\lambda_1, \dots, \lambda_d$ and $L_j = \sum_{l=0}^{d-1} L_j^{(l)} X^l$ for $1 \leq j \leq d$.

A proof of this result can be found at Atkinson (1989, Theorem 3.1).

The next lemma is the last one before we finally prove the theorem. Recall that in Thm. 5 we assume that $k+1 \leq d$.

Lemma 15. *Assume that $\operatorname{Sp} \nabla v = \{\lambda_1, \dots, \lambda_{k+1}\}$ where $\lambda_1, \dots, \lambda_{k+1}$ are distinct non-zero reals. Then the problem of (186) is lower bounded by*

$$\frac{1}{2} \left(\frac{1 - \sum_{j=1}^k \frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})}{1 + \sum_{j=1}^k |\frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})|} \right)^2, \tag{199}$$

where $L_1, \dots, L_k$ are the Lagrange interpolation polynomials associated to $\lambda_1, \dots, \lambda_k$.

Proof. To prove this lemma, we start from the result of Lem. 13 and we provide feasible $(\nu_j)_j$ and $(\xi_j)_j$. First, any feasible $(\nu_j)_j$ and $(\xi_j)_j$ must satisfy the k constraints involving the powers of the eigenvalues, which can be rewritten as:

$$V(\lambda_1, \dots, \lambda_k)^T \begin{pmatrix} \nu_1 \xi_1 \lambda_1 \\ \nu_2 \xi_2 \lambda_2 \\ \vdots \\ \nu_k \xi_k \lambda_k \end{pmatrix} = -\nu_{k+1} \xi_{k+1} \begin{pmatrix} \lambda_{k+1} \\ \lambda_{k+1}^2 \\ \dots \\ \lambda_{k+1}^k \end{pmatrix}. \tag{200}$$

Using the previous lemma yields, for $1 \leq j \leq k$,

$$\nu_j \xi_j = -\nu_{k+1}\xi_{k+1}\frac{1}{\lambda_j}\begin{pmatrix} L_j^{(0)} & L_j^{(1)} & \dots & L_j^{(k-1)} \end{pmatrix}\begin{pmatrix} \lambda_{k+1} \\ \lambda_{k+1}^2 \\ \dots \\ \lambda_{k+1}^k \end{pmatrix} \tag{201}$$

$$= -\nu_{k+1}\xi_{k+1}\frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})\,. \tag{202}$$

Hence the problem can be rewritten only in terms of the $(\nu_j)_j$ and ξ_{k+1}. Let $c_j = \frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})$. The objective becomes:

$$\sum_{j=1}^m \nu_j \left(\xi_j - \frac{1}{2}\xi_j^2\right) = \nu_{k+1}\xi_{k+1}\left(1 - \sum_{j=1}^k c_j\right) - \frac{1}{2}\nu_{k+1}\xi_{k+1}^2\left(1 + \sum_{j=0}^k \frac{\nu_{k+1}}{\nu_j} c_j^2\right)\,. \tag{203}$$

Choosing $\xi_{k+1} = \frac{1-\sum_{j=1}^k c_j}{1+\sum_{j=0}^k \frac{\nu_{k+1}}{\nu_j}c_j^2}$ to maximize this quadratic yields:

$$\sum_{j=1}^m \nu_j \left(\xi_j - \frac{1}{2}\xi_j^2\right) = \frac{1}{2}\nu_{k+1}\frac{\left(1 - \sum_{j=1}^k c_j\right)^2}{1 + \sum_{j=0}^k \frac{\nu_{k+1}}{\nu_j}c_j^2}\,. \tag{204}$$

Finally take $\nu_j = \frac{|c_j|}{1+\sum_{j=1}^k |c_j|}$ for $j \leq k$ and $\nu_{k+1} = \frac{1}{1+\sum_{j=1}^k |c_j|}$which satisfy the hypotheses of the problem of Lem. 13. With the feasible $(\nu_j)_j$ and $(\xi_j)_j$ defined this way, the value of the objective is

$$\frac{1}{2}\left(\frac{1 - \sum_{j=1}^k c_j}{1 + \sum_{j=1}^k |c_j|}\right)^2\,, \tag{205}$$

which is the desired result. □

We finally prove (i) of Thm. 5.

Proof of (i) of Thm. 5. To prove the theorem, we build on the result of Lem. 15. We have to choose $\lambda_1, \dots, \lambda_{k+1} \in [\mu, L]$ positive distinct such that (199) is big. One could try to distribute the eigenvalues uniformly across the interval but this leads to a lower bound which decreases exponentially in k. To make things a bit better, we use Chebyshev points of the second kind studied by Salzer (1971). However we will actually refer to the more recent presentation of Berrut and Trefethen (2004).

For now, assume that k is even and so $k \geq 4$. We will only use that $d-1 \geq k$ (and not that $d-2 \geq k$). Define, for $1 \leq j \leq k$, $\lambda_j = \frac{\mu+L}{2} - \frac{L-\mu}{2}\cos\frac{j-1}{k-1}\pi$. Using the barycentric formula of Berrut and Trefethen (2004, Eq. 4.2), the polynomial which interpolates $f_1, \dots, f_k$ at the points $\lambda_1, \dots, \lambda_k$ can be written as:

$$P(X) = \frac{\sum_{j=1}^k \frac{w_j}{X-\lambda_j} f_j}{\sum_{j=1}^k \frac{w_j}{X-\lambda_j}}\,, \tag{206}$$

where

$$w_j = \begin{cases} (-1)^{j-1} & \text{if} \quad 2 \leq j \leq k-1 \\ \frac{1}{2}(-1)^{j-1} & \text{if} \quad j \in \{1, k\}\,. \end{cases} \tag{207}$$

Define $Z(X) = \sum_{j=1}^k \frac{w_j}{X-\lambda_j}$.

Now, $\sum_{j=1}^k \frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})$ can be seen as the polynomial interpolating $\frac{\lambda_{k+1}}{\lambda_1}, \dots, \frac{\lambda_{k+1}}{\lambda_k}$ at the points $\lambda_1, \dots, \lambda_j$ evaluated at λ_{k+1}. Hence, using the barycentric formula,

$$\sum_{j=1}^k \frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1}) = \frac{1}{Z(\lambda_{k+1})}\sum_{j=1}^k \frac{w_j}{\lambda_{k+1}-\lambda_j}\frac{\lambda_{k+1}}{\lambda_j}\,. \tag{208}$$

Similarly, $\sum_{j=1}^k |\frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})|$ can be seen as the polynomial interpolating $|\frac{\lambda_{k+1}}{\lambda_1}| \operatorname{sign}(L_1(\lambda_{k+1})), \dots, |\frac{\lambda_{k+1}}{\lambda_k}| \operatorname{sign}(L_k(\lambda_{k+1}))$ at the points $\lambda_1, \dots, \lambda_j$ evaluated at λ_{k+1}. However, from Berrut and Trefethen (2004, Section 3),

$$L_j(\lambda_{k+1}) = \left(\prod_{j=1}^k (\lambda_{k+1} - \lambda_j)\right) \frac{w_j}{\lambda_{k+1} - \lambda_j}\,, \tag{209}$$

and by Berrut and Trefethen (2004, Eq. 4.1),

$$1 = \left(\prod_{j=1}^k (\lambda_{k+1} - \lambda_j)\right) Z(\lambda_{k+1})\,. \tag{210}$$

Hence

$$\operatorname{sign}(L_j(\lambda_{k+1})) = \operatorname{sign} Z(\lambda_{k+1}) \operatorname{sign}\left(\frac{w_j}{\lambda_{k+1} - \lambda_j}\right)\,. \tag{211}$$

Therefore, using the barycentric formula again,

$$\sum_{j=1}^k \frac{\lambda_{k+1}}{\lambda_j} |L_j(\lambda_{k+1})| = \frac{1}{|Z(\lambda_{k+1})|} \sum_{j=1}^k \left|\frac{w_j}{\lambda_{k+1} - \lambda_j}\right| \frac{\lambda_{k+1}}{\lambda_j}\,. \tag{212}$$

Hence, (199) becomes:

$$\frac{1}{2}\left(\frac{1 - \sum_{j=1}^k \frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})}{1 + \sum_{j=1}^k |\frac{\lambda_{k+1}}{\lambda_j} L_j(\lambda_{k+1})|}\right)^2 \tag{213}$$

$$= \frac{1}{2}\left(\frac{1 - \frac{1}{Z(\lambda_{k+1})} \sum_{j=1}^k \frac{w_j}{\lambda_{k+1} - \lambda_j} \frac{\lambda_{k+1}}{\lambda_j}}{1 + \frac{1}{|Z(\lambda_{k+1})|} \sum_{j=1}^k \left|\frac{w_j}{\lambda_{k+1} - \lambda_j}\right| \frac{\lambda_{k+1}}{\lambda_j}}\right)^2 \tag{214}$$

$$= \frac{1}{2}\left(1 - \frac{\frac{1}{|Z(\lambda_{k+1})|} \sum_{j=1}^k \left|\frac{w_j}{\lambda_{k+1} - \lambda_j}\right| \frac{\lambda_{k+1}}{\lambda_j} \left(1 + \operatorname{sign} Z(\lambda_{k+1}) \operatorname{sign}\left(\frac{w_j}{\lambda_{k+1} - \lambda_j}\right)\right)}{1 + \frac{1}{|Z(\lambda_{k+1})|} \sum_{j=1}^k \left|\frac{w_j}{\lambda_{k+1} - \lambda_j}\right| \frac{\lambda_{k+1}}{\lambda_j}}\right)^2\,. \tag{215}$$

Now take any λ_{k+1} such that $\lambda_1 < \lambda_{k+1} < \lambda_2$. Then, from (210), $\operatorname{sign} Z(\lambda_{k+1}) = (-1)^{k+1} = -1$ as we assume that k is even. By definition of the coefficients w_j, $\operatorname{sign} \frac{w_1}{\lambda_{k+1} - \lambda_1} = +1$. Hence $1 + \operatorname{sign} Z(\lambda_{k+1}) \operatorname{sign} \frac{w_1}{\lambda_{k+1} - \lambda_1} = 0$. Similarly, $\operatorname{sign} \frac{w_2}{\lambda_{k+1} - \lambda_2} = +1$ and so $1 + \operatorname{sign} Z(\lambda_{k+1}) \operatorname{sign} \frac{w_2}{\lambda_{k+1} - \lambda_2} = 0$ too[7].

As the quantity inside the parentheses of (215) is non-negative, we can focus on lower bounding it. Using the

[7] We could do without this, but it is free and gives slightly better constants.

considerations on signs we get:

$$\frac{\frac{1}{|Z(\lambda_{k+1})|}\sum_{j=1}^{k}\left|\frac{w_j}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j}\left(1+\operatorname{sign} Z(\lambda_{k+1})\operatorname{sign}\left(\frac{w_j}{\lambda_{k+1}-\lambda_j}\right)\right)}{1+\frac{1}{|Z(\lambda_{k+1})|}\sum_{j=1}^{k}\left|\frac{w_j}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j}} \tag{216}$$

$$= \frac{\frac{1}{|Z(\lambda_{k+1})|}\sum_{j=3}^{k}\left|\frac{w_j}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j}\left(1+\operatorname{sign} Z(\lambda_{k+1})\operatorname{sign}\left(\frac{w_j}{\lambda_{k+1}-\lambda_j}\right)\right)}{1+\frac{1}{|Z(\lambda_{k+1})|}\sum_{j=1}^{k}\left|\frac{w_j}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j}} \tag{217}$$

$$\leq 2\frac{\sum_{j=3}^{k}\left|\frac{1}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j}}{|Z(\lambda_{k+1})|+\sum_{j=1}^{k}\left|\frac{w_j}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j}} \tag{218}$$

$$\leq 2\frac{\sum_{j=3}^{k}\left|\frac{1}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j}}{\frac{1}{2}\left|\frac{1}{\lambda_{k+1}-\lambda_1}\right|\frac{\lambda_{k+1}}{\lambda_1}} \tag{219}$$

$$\leq 2\frac{(k-2)\left|\frac{1}{\lambda_{k+1}-\lambda_3}\right|\frac{\lambda_{k+1}}{\lambda_3}}{\frac{1}{2}\left|\frac{1}{\lambda_{k+1}-\lambda_1}\right|\frac{\lambda_{k+1}}{\lambda_1}} \tag{220}$$

$$\tag{221}$$

where we used that, for $j \geq 3$, $\left|\frac{1}{\lambda_{k+1}-\lambda_j}\right|\frac{\lambda_{k+1}}{\lambda_j} \leq \left|\frac{1}{\lambda_{k+1}-\lambda_3}\right|\frac{\lambda_{k+1}}{\lambda_3}$ as $\lambda_1 < \lambda_{k+1} < \lambda_2 < \lambda_3 < \cdots < \lambda_k$. Now, recalling that $\lambda_1 = \mu$, and using that $\lambda_1 < \lambda_{k+1} < \lambda_2 < \lambda_3$ for the inequality,

$$2\frac{(k-2)\left|\frac{1}{\lambda_{k+1}-\lambda_3}\right|\frac{\lambda_{k+1}}{\lambda_3}}{\frac{1}{2}\left|\frac{1}{\lambda_{k+1}-\lambda_1}\right|\frac{\lambda_{k+1}}{\lambda_1}} = 4(k-2)\frac{\mu}{\lambda_3}\frac{|\lambda_{k+1}-\lambda_1|}{|\lambda_{k+1}-\lambda_3|} \tag{222}$$

$$\leq 4(k-2)\frac{\mu}{\lambda_3}\frac{|\lambda_2-\lambda_1|}{|\lambda_2-\lambda_3|} \tag{223}$$

$$= 4(k-2)\frac{\mu}{\frac{1}{2}L(1-\cos\frac{2\pi}{k-1})+\frac{1}{2}\mu(1+\cos\frac{2\pi}{k-1})}\frac{|\cos\frac{\pi}{k-1}-1|}{|\cos\frac{\pi}{k-1}-\cos\frac{2\pi}{k-1}|} \tag{224}$$

$$\leq 8(k-2)\frac{\mu}{L(1-\cos\frac{\pi}{k-1})}\frac{|\cos\frac{\pi}{k-1}-1|}{|\cos\frac{\pi}{k-1}-\cos\frac{2\pi}{k-1}|} \tag{225}$$

$$= 8(k-2)\frac{\mu}{L}\frac{1}{|\cos\frac{\pi}{k-1}-\cos\frac{2\pi}{k-1}|} \tag{226}$$

by definition of the interpolation points. Now, for $k \geq 4$, the sinus is non-negative on $[\frac{\pi}{k-1}, \frac{2\pi}{k-1}]$ and reaches its minimum at $\frac{\pi}{k-1}$. Hence,

$$\left|\cos\frac{\pi}{k-1}-\cos\frac{2\pi}{k-1}\right| = \left|\int_{\pi/(k-1)}^{2\pi/(k-1)}\sin t dt\right| \tag{227}$$

$$= \int_{\pi/(k-1)}^{2\pi/(k-1)}\sin t dt \tag{228}$$

$$\geq \frac{\pi}{k-1}\sin\frac{\pi}{k-1} \tag{229}$$

$$\geq 2\frac{\pi}{(k-1)^2} \tag{230}$$

as $0 \geq \frac{\pi}{k-1} \geq \frac{\pi}{2}$. Putting everything together yields,

$$\frac{1}{2}\left(\frac{1-\sum_{j=1}^{k}\frac{\lambda_{k+1}}{\lambda_j}L_j(\lambda_{k+1})}{1+\sum_{j=1}^{k}|\frac{\lambda_{k+1}}{\lambda_j}L_j(\lambda_{k+1})|}\right)^2 \geq \frac{1}{2}\left(1-\frac{4(k-1)^2(k-2)}{\pi}\frac{\mu}{L}\right)^2 \tag{231}$$

$$\geq \frac{1}{2}\left(1-\frac{4(k-1)^3)}{\pi}\frac{\mu}{L}\right)^2, \tag{232}$$

which yields the desired result by the definition of the problem of (186).

The lower bound holds for any v such that $\operatorname{Sp}\nabla v = \{\lambda_1, \ldots, \lambda_{k+1}\}$. As $\{\lambda_1, \ldots, \lambda_{k+1}\} \subset [\mu, L]$, one can choose v of the form $v = \nabla f$ where $f : \mathbb{R}^d \to \mathbb{R}$ is a μ-strongly convex and L-smooth quadratic function with $\operatorname{Sp}\nabla^2 f = \{\lambda_1, \ldots, \lambda_{k+1}\}$.

Now, we tackle the case k odd, with $k \geq 3$ and $d-1 \geq k+1$. Note that if $\mathcal{N}$ is a real polynomial of degree at most $k-1$, it is also a polynomial of degree at most $(k+1)-1$. Applying the result above yields that there exists $v \in \mathcal{V}_d$ with the desired properties such that,

$$\rho(F_{\mathcal{N}}) \geq 1 - \frac{k^3}{2\pi}\frac{\gamma^2}{L^2}. \tag{233}$$

Hence, (i) holds for any $d-2 \geq k \geq 3$. □

Then, (ii) is essentially a corollary of (i).

Proof of (ii) of Thm. 5. For a square zero-sum two player game, the Jacobian of the vector field can be written as,

$$\nabla v = \begin{pmatrix} 0_m & A \\ -A^T & 0_m \end{pmatrix} \tag{234}$$

where $A \in \mathbb{R}^{m\times m}$. By Lem. 3,

$$\operatorname{Sp}\nabla v = \{i\sqrt{\lambda} \mid \lambda \in \operatorname{Sp} AA^T\} \cup \{-i\sqrt{\lambda} \mid \lambda \in \operatorname{Sp} AA^T\}. \tag{235}$$

Using Lem. 12, one gets that:

$$\min_{N\in\mathbb{R}_{k-1}[X]} \frac{1}{2}\rho(F_{\mathcal{N}})^2 = \min_{a_0,\ldots,a_{k-1}\in\mathbb{R}} \max_{\lambda\in\operatorname{Sp} AA^T} \frac{1}{2}\max\left(\left|1+\sum_{l=0}^{k-1} a_l(i\sqrt{\lambda})^{l+1}\right|^2, \left|1+\sum_{l=0}^{k-1} a_l(-i\sqrt{\lambda})^{l+1}\right|^2\right) \tag{236}$$

$$\geq \min_{a_0,\ldots,a_{k-1}\in\mathbb{R}} \max_{\lambda\in\operatorname{Sp} AA^T} \frac{1}{2}\left|1+\sum_{l=0}^{k-1} a_l(i\sqrt{\lambda})^{l+1}\right|^2 \tag{237}$$

$$\geq \min_{a_0,\ldots,a_{k-1}\in\mathbb{R}} \max_{\lambda\in\operatorname{Sp} AA^T} \frac{1}{2}\left(\Re\left(1+\sum_{l=0}^{k-1} a_l(i\sqrt{\lambda})^{l+1}\right)\right)^2 \tag{238}$$

$$= \min_{a_0,\ldots,a_{k-1}\in\mathbb{R}} \max_{\lambda\in\operatorname{Sp} AA^T} \frac{1}{2}\left(1+\sum_{l=1}^{\lfloor k/2\rfloor} a_{2l-1}(-1)^l\lambda^l\right)^2 \tag{239}$$

$$= \min_{a_0,\ldots,a_{\lfloor k/2\rfloor-1}\in\mathbb{R}} \max_{\lambda\in\operatorname{Sp} AA^T} \frac{1}{2}\left(1+\sum_{l=1}^{\lfloor k/2\rfloor} a_{l-1}\lambda^l\right)^2. \tag{240}$$

Using Lem. 12 again,

$$\min_{a_0,\ldots,a_{\lfloor k/2\rfloor-1}\in\mathbb{R}} \max_{\lambda\in\operatorname{Sp} AA^T} \frac{1}{2}\left(1+\sum_{l=1}^{\lfloor k/2\rfloor} a_{l-1}\lambda^l\right)^2 = \min_{N\in\mathbb{R}_{\lfloor k/2\rfloor-1}[X]} \frac{1}{2}\rho(\tilde{F}_{\mathcal{N}})^2. \tag{241}$$

where $\tilde{F}_{\mathcal{N}}$ is the 1-SCLI operator of $\mathcal{N}$, as defined by (2) applied to the vector field $\omega \mapsto AA^T\omega$. Let $S \in \mathbb{R}^{m\times m}$ be a symmetric positive definite matrix given by (i) of this theorem applied with $(\mu, L) = (\gamma^2, L^2)$ and $\lfloor \frac{k}{2} \rfloor$ instead of k and so such that $\operatorname{Sp} S \subset [\gamma^2, L^2]$. Now choose $A \in \mathbb{R}^{m\times m}$ such that $A^T A = S$, for instance by taking a square root of S (see Lax (2007, Chapter 10)). Then,

$$\min_{N\in\mathbb{R}_{\lfloor k/2\rfloor -1}[X]} \frac{1}{2}\rho(\tilde{F}_{\mathcal{N}})^2 \geq \frac{1}{2}\left(1 - \frac{k^3}{2\pi}\frac{\mu}{L}\right) . \tag{242}$$

Moreover, by computing $\nabla v^T \nabla v$ and using that $\operatorname{Sp} AA^T = \operatorname{Sp} A^T A$, one gets that $\min_{\lambda\in\operatorname{Sp}\nabla v} |\lambda| = \sigma_{min}(\nabla v) = \sigma_{min}(A) \geq \gamma$ and $\sigma_{max}(\nabla v) = \sigma_{max}(A) \leq L$. $\square$

Remark 2. Interestingly, the examples we end up using have a spectrum similar to the one of the matrix Nesterov uses in the proofs of his lower bounds in Nesterov (2004). The choice of the spectrum of the Jacobian of the vector field was indeed the choice of interpolation points. Following Salzer (1971); Berrut and Trefethen (2004) we used points distributed across the interval as a cosinus as it minimizes oscillations near the edge of the interval. Therefore, this links the hardness Nesterov's examples to the well-conditioning of families of interpolation points.

I Handling singularity

The following theorem is a way to use spectral techniques to obtain geometric convergence rates even if the Jacobian of the vector field at the stationary point is singular. We only need to ensure that the vector field is locally null along these directions of singularity.

In this subsection, for $A \in \mathbb{R}^{m\times p}$, $\operatorname{Ker} A = \{x \in \mathbb{R}^p \mid Ax = 0\}$ denotes the kernel (or the nullspace) of A.

The following theorem is actually a combination of the proof of Nagarajan and Kolter (2017, Thm. A.4), which only proves asymptotic stability in continuous time with no concern for the rate, and the classic Thm. 1.

Theorem 15. *Consider $h : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^m \times \mathbb{R}^p$ twice continuously differentiable vector field and write $h(\theta, \varphi) = (h_\theta(\theta, \varphi), h_\varphi(\theta, \varphi))$. Assume that $(0, 0)$ is a stationary point, i.e. $h(0, 0) = (0, 0)$ and that there exists $\tau > 0$ such that,*

$$\forall \varphi \in \mathbb{R}^p \cap B(0, \tau), \quad h(0, \varphi) = (0, 0) . \tag{243}$$

Let $\rho^* = \rho(\nabla_\theta(\operatorname{Id} + h_\theta)(0, 0))$ and define the iterates $(\theta_t, \varphi_t)_t$ by

$$(\theta_{t+1}, \varphi_{t+1}) = (\theta_t, \varphi_t) + h(\theta_t, \varphi_t) . \tag{244}$$

Then, if $\rho^* < 1$, for all $\epsilon > 0$, there exists a neighborhood of $(0, 0)$ such that for any initial point in this neighborhood, the distance of the iterates $(\theta_t, \varphi_t)_t$ to a stationary point of h decreases as $\mathcal{O}((\rho^* + \epsilon)^t)$. If v is linear, this is satisfied with the whole space as a neighborhood for all $\epsilon > 0$.

The following proof is inspired from the ones of Nagarajan and Kolter (2017, Thm. 4) and Gidel et al. (2019b, Thm. 1).

Proof. Let $\mathbf{J} = \nabla_\theta h(0, 0) \in \mathbb{R}^{(m+p)\times m}$, $\mathbf{J}_\theta = \nabla_\theta h_\theta(0, 0) \in \mathbb{R}^{m\times m}$ and $\mathbf{J}_\varphi = \nabla_\theta h_\varphi(0, 0) \in \mathbb{R}^{p\times m}$. Let $\epsilon > 0$ and suppose $\rho^* + \epsilon < 1$. By Bertsekas (1999, Prop. A.15) there exists a norm $\|.\|$ on $\mathbb{R}^m$ such that the induced matrix norm on $\mathbb{R}^{m\times m}$ satisfy:

$$\| \operatorname{Id} + J_\theta\| \leq \rho^* + \frac{\epsilon}{2} . \tag{245}$$

On the contrary the norm on $\mathbb{R}^p$ can be chosen arbitrarily.

The extension of these norms to $\mathbb{R}^m \times \mathbb{R}^p$ is chosen such that $\|\theta, \varphi\| = \|\theta\| + \|\varphi\|$ for simplicity (but this is without loss of generality). In this proof, we denote the d-dimensional balls by $B_d(x, r) = \{y \in \mathbb{R}^d \mid \|x - y\| \leq r\}$ with $x \in \mathbb{R}^d$, $r > 0$.

- Let $\mathbf{J} = \nabla_\theta h(0, 0) \in \mathbb{R}^{(m+p)\times m}$.

  We first show that, for all $\eta > 0$ there exists $\tau \geq \delta > 0$ such that,

$$\forall (\theta, \varphi) \in B_{m+p}((0, 0), \delta) : \ \|h(\theta, \varphi) - \mathbf{J}\theta\| \leq \eta\|\theta\| . \tag{246}$$

The interesting thing here is that we are completely getting rid of the dependence on φ, both in the linearization and in the bound.

Let $\varphi \in B_p(0,\tau)$. Then, using that $h(0,\varphi)=0$, the Taylor development of $h(\theta,\varphi)$ w.r.t. to θ yields:

$$
\begin{aligned}
h(\theta,\varphi) &= \nabla_\theta h(0,\varphi)\theta + R(\theta,\varphi) &(247)\\
&= J\theta + (\nabla_\theta h(0,\varphi) - \nabla_\theta h(0,0))\theta + R(\theta,\varphi)\,. &(248)\\
& &(249)
\end{aligned}
$$

We now deal with the last two terms. First the rest $R(\theta,\varphi)$. As v is assumed to be continuously differentiable, there exists $c>0$ constant (which depends on τ) such that, for any $\theta \in \mathbb{R}^m$:

$$
\forall \varphi \in B_p(0,\delta) : \|R(\theta,\varphi)\| \le c\|\theta\|^2\,, \tag{250}
$$

Hence, for $\theta \in B(0,\frac{\eta}{2c})$, we get that:

$$
\forall \varphi \in B_p(0,\delta) : \|R(\theta,\varphi)\| \le \frac{\eta}{2}\|\theta\|\,. \tag{251}
$$

Concerning the other term, by continuity, $\nabla_\theta h(0,\varphi) - \nabla_\theta h(0,0)$ goes to zero as φ goes to zero. Hence, there exists $\delta>0$, $\delta \le \min(\tau,\frac{\eta}{2c})$ such that for any $\varphi \in B_p(0,\delta)$, $\|(\nabla_\theta h(0,\varphi) - \nabla_\theta h(0,0))\theta\| \le \frac{\eta}{2}\|\theta\|$. Combining the two bounds yields the desired result.

- We now apply the previous result with $\eta = \epsilon/2$. We first examine what this means for $(\theta_{t+1},\varphi_{t+1})$ when (θ_t,φ_t) is in $B_{m+p}((0,0),\delta)$. However, the neighborhood $B_{m+p}((0,0),\delta)$ is not necessarily stable, so we will again restrict it afterwards. See the proof (Nagarajan and Kolter, 2017, Thm. 4) for a more detailed discussion on this. Assume for now that $(\theta_t,\varphi_t) \in B_{m+p}((0,0),\delta)$. Then,

$$
\begin{aligned}
\|\theta_{t+1}\| &= \|(\mathrm{Id} + \mathbf{J}_\theta)\theta_t + (h_\theta(\theta_t,\varphi_t) - \mathbf{J}_\theta\theta_t)\| &(252)\\
&\le \|(\mathrm{Id} + \mathbf{J}_\theta)\theta_t\| + \|(h_\theta(\theta_t,\varphi_t) - \mathbf{J}_\theta\theta_t)\| &(253)\\
&\le (\rho^* + \epsilon)\|\theta_t\|\,. &(254)\\
& &(255)
\end{aligned}
$$

Consider now the other coordinate φ_{t+1}, still under the assumption that $(\theta_t,\varphi_t) \in B_{m+p}((0,0),\delta)$. Then,

$$
\begin{aligned}
\|\varphi_{t+1} - \varphi_t\| &= \|\mathbf{J}_\varphi\theta_t + (h_\varphi(\theta_t,\varphi_t) - \mathbf{J}_\varphi\theta_t)\| &(256)\\
&\le \|\mathbf{J}_\varphi\theta_t\| + \|(h_\varphi(\theta_t,\varphi_t) - \mathbf{J}_\varphi\theta_t)\| &(257)\\
&\le (\|\mathbf{J}_\varphi\| + \frac{\epsilon}{2})\|\theta_t\|\,. &(258)\\
& &(259)
\end{aligned}
$$

Now let $V = \{(\theta,\varphi) \in \mathbb{R}^m \times \mathbb{R}^p \;|(\theta,\varphi) \in B((0,0),\delta),\, (1 + \frac{\|\mathbf{J}_\varphi\| + \frac{\epsilon}{2}}{1-\rho^*-\epsilon})\|\theta\| + \|\varphi\| < \delta\}$ neighborhood of $(0,0)$. We show, by induction, that if $(\theta_0,\varphi_0) \in V$, then the iterates stay in $B_{m+p}((0,0),\delta)$.

Assume $(\theta_0,\varphi_0) \in V$. By construction, $(\theta_0,\varphi_0) \in B_{m+p}((0,0),\delta)$. Now assume that $(\theta_0,\varphi_0),(\theta_1,\varphi_1),\dots,(\theta_t,\varphi_t)$ are in $B_{m+p}((0,0),\delta)$ for some $t \ge 0$. By what has been proven above, first, $\|\theta_{t+1}\| \le (\rho^*+\epsilon)^{t+1}\|\theta_0\| \le \|\theta_0\|$. Then,

$$
\begin{aligned}
\|\varphi_{t+1}\| &\le \|\varphi_0\| + \sum_{k=0}^{t} \|\varphi_{k+1} - \varphi_k\| &(260)\\
&\le \|\varphi_0\| + (\|\mathbf{J}_\varphi\| + \frac{\epsilon}{2})\sum_{k=0}^{t}\|\theta_k\| &(261)\\
&\le \|\varphi_0\| + (\|\mathbf{J}_\varphi\| + \frac{\epsilon}{2})\sum_{k=0}^{t}(\rho^*+\epsilon)^k\|\theta_0\| &(262)\\
&\le \|\varphi_0\| + \frac{\|\mathbf{J}_\varphi\| + \frac{\epsilon}{2}}{1-\rho^*-\epsilon}\|\theta_0\| &(263)\\
& &(264)
\end{aligned}
$$

Hence, putting the two coordinates together,

$$\|(\theta_{t+1}, \varphi_{t+1})\| = \|\theta_{t+1}\| + \|\varphi_{t+1}\| \tag{265}$$

$$\leq \|\varphi_0\| + \left(1 + \frac{\|\mathbf{J}_\varphi\| + \frac{\epsilon}{2}}{1 - \rho^* - \epsilon}\right) \|\theta_0\| \tag{266}$$

$$< \delta\,. \tag{267}$$

by definition of V. Hence, $(\theta_{t+1}, \varphi_{t+1}) \in B_{m+p}((0,0),\delta)$ which concludes the induction and the proof.

For the linear operator case, note that we can choose $\tau = +\infty$, $c = 0$, $\eta = 0$ and $\delta = +\infty$. Then we have $V = \mathbb{R}^m \times \mathbb{R}^p$. □

By a linear base change, we get the more practical corollary:

Corollary 4. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be twice continuously differentiable and $\omega^* \in \mathbb{R}^d$ be a fixed point. Assume that there exists $\delta > 0$ such that for all $\xi \in \operatorname{Ker}(\nabla F(\omega^*) - I_d) \cap B(0,\delta)$, $\omega^* + \xi$ is still a fixed point and that $\operatorname{Ker}(\nabla F(\omega^*) - I_d)^2 = \operatorname{Ker}(\nabla F(\omega^*) - I_d)$. Define*

$$\rho^* = \max\{|\lambda| \mid \lambda \in \operatorname{Sp} \nabla F(\omega^*),\, \lambda \neq 1\}\,, \tag{268}$$

and assume $\rho^* < 1$. Consider the iterates $(\omega_t)_t$ built from $\omega_0 \in \mathbb{R}^d$ as:

$$\omega_{t+1} = F(\omega_t) \quad \forall t \geq 0\,. \tag{269}$$

Then, for all $\epsilon > 0$, for any ω_0 in a neighborhood of ω^*, the distance of the iterates $(\omega_t)^t$ to fixed points of F decreases in $\mathcal{O}((\rho^* + \epsilon)^t)$.

Moreover, if F is linear, we can take this neighborhood to be the whole space and $\epsilon = 0$.

Proof. We consider the spaces $\operatorname{Ker}(\nabla F(\omega^*) - \lambda I_d)^{m_\lambda}$, $\lambda \in \operatorname{Sp} \nabla F(\omega^*)$ where m_λ denotes the multiplicity of the eigenvalue λ as root of the characteristic polynomial of $\nabla F(\omega^*)$. Then, we have,

$$\mathbb{R}^d = \bigoplus_{\lambda \in \operatorname{Sp} \nabla F(\omega^*)} \operatorname{Ker}(\nabla F(\omega^*) - \lambda I_d)^{m_\lambda}\,,$$

see Lax (2007, Chap. 6) for instance.

Now, using that $\operatorname{Ker}(\nabla F(\omega^*) - I_d)^2 = \operatorname{Ker}(\nabla F(\omega^*) - I_d)$, we have that $\operatorname{Ker}(\nabla F(\omega^*) - I_d)^{m_1} = \operatorname{Ker}(\nabla F(\omega^*) - I_d)$. Hence, the whole space can be decomposed as $\mathbb{R}^d = \operatorname{Ker}(\nabla F(\omega^*) - I_d) \oplus E$ where $E = \bigoplus_{\lambda \in \operatorname{Sp} \nabla F(\omega^*)\setminus\{1\}} \operatorname{Ker}(\nabla F(\omega^*) - \lambda I_d)^{m_\lambda}$. Note that E is stable by $\nabla F(\omega^*)$ and so $\rho(\nabla F(\omega^*)|_E) = \rho^*$ as defined in the statement of the theorem. Denote by $P \in \mathbb{R}^d \times \mathbb{R}^d$ the (invertible) change of basis such that $\operatorname{Ker}(\nabla F(\omega^*) - I_d)$ is sent on the subspace $\mathbb{R}^m \times \{0\}^p$ and E on the subspace $\{0\}^m \times \mathbb{R}^p$, where m and p are the respective dimensions of $\operatorname{Ker}(\nabla F(\omega^*) - I_d)$ and E. Then, we apply the previous theorem Thm. 15 with h defined by,

$$(\theta, \varphi) + h(\theta, \varphi) = PF\left(\omega^* + P^{-1}(\theta, \varphi)\right)\,.$$

which concludes the proof. □

Remark 3. In general the condition $\operatorname{Ker}(\nabla F(\omega^*) - I_d)^2 = \operatorname{Ker}(\nabla F(\omega^*) - I_d)$ will be equivalent to $\operatorname{Ker} \nabla v(\omega^*)^2 = \operatorname{Ker} \nabla v(\omega^*)$ where v is the game vector field. We keep this remark informal but we prove this for extragradient below as an example. Indeed, as seen with $1-SCLI$ in §3.3 with (2), $\nabla F(\omega^*)$ is of the form $\operatorname{Id} + \mathcal{N}(\nabla v(\omega^*))\nabla v(\omega^*)$ where $\mathcal{N}$ is a polynomial. Hence, $(\nabla F(\omega^*) - I_d)^j = \mathcal{N}(\nabla v(\omega^*))^j \nabla v(\omega^*)^j$. Moreover, in practice, $\mathcal{N}(\nabla v(\omega^*))$ will be chosen — e.g. by the choice of the step-size — to be non-singular. Hence, $\operatorname{Ker}(\nabla F(\omega^*) - I_d)^j = \operatorname{Ker} \nabla v(\omega^*)^j$ and so $\operatorname{Ker}(\nabla F(\omega^*) - I_d)^2 = \operatorname{Ker}(\nabla F(\omega^*) - I_d)$ will be equivalent to $\operatorname{Ker} \nabla v(\omega^*)^2 = \operatorname{Ker} \nabla v(\omega^*)$.

We now prove a lemma concerning extragradient which as a first step before apply Cor. 4. We could have proven this result for k-extrapolation methods but we focus on extragradient for simplicity.

Lemma 16. *Let $F_{2,\eta} : \omega \to \omega - \eta v(\omega - \eta v(\omega))$ denote the extragradient operator. Assume that v is L-Lipschitz. Then, if $0 < \eta < \frac{1}{L}$, for ω^* stationary point of v,*

$$\operatorname{Ker}(\nabla F_{2,\eta}(\omega^*) - I_d) = \operatorname{Ker} \nabla v(\omega^*)\,,$$

and

$$\operatorname{Ker}(\nabla F_{2,\eta}(\omega^*) - I_d)^2 = \operatorname{Ker}(\nabla F_{2,\eta}(\omega^*) - I_d) \iff \operatorname{Ker} \nabla v(\omega^*)^2 = \operatorname{Ker} \nabla v(\omega^*)\,.$$

Proof. We have $\nabla F_{2,\eta}(\omega^*) = I_d - \eta \nabla v(\omega^*)(I_d - \eta \nabla v(\omega^*))$ and so $\nabla F_{2,\eta}(\omega^*) - I_d = -\eta \nabla v(\omega^*)(I_d - \eta \nabla v(\omega^*))$. As $\nabla v(\omega^*)$ and $I_d - \eta \nabla v(\omega^*))$ commute, for $j \in \{1, 2\}$,

$$(\nabla F_{2,\eta}(\omega^*) - I_d)^j = (-\eta\,(I_d - \eta \nabla v(\omega^*)))^j\, \nabla v(\omega^*)^j\,.$$

By the choice of η, $\eta(I_d - \eta \nabla v(\omega^*))$ is non-singular and so $\operatorname{Ker}(\nabla F_{2,\eta}(\omega^*) - I_d)^j = \operatorname{Ker} \nabla v(\omega^*)^j$ which yields the result. □

The whole framework developed implies in particular that Thm. 4 actually also yields convergence guarantees for extragradient on more general bilinear games than those considered in Example 2.

Example 4 (Bilinear game with potential singularity)**.** A saddle-point problem of the form:

$$\min_{x\in\mathbb{R}^m} \max_{y\in\mathbb{R}^p} x^T A y + b^T x + c^T y \tag{270}$$

with $A \in \mathbb{R}^{m\times p}$ not null, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^p$.

Corollary 5. *Consider the bilinear game of Example 4. The iterates of extragradient with $\eta = (4\sigma_{max}(A))$ converge globally to ω^* at a linear rate of $\mathcal{O}\big(\big(1 - \frac{1}{64}\frac{\tilde\sigma_{min}(A)^2}{\sigma_{max}(A)^2}\big)^t\big)$ where $\tilde\sigma_{min}(A)$ is the smallest non-zero singular value of A.*

Proof. Let ω^* be a stationary point of the associated vector field v. Then, $\nabla v(\omega^*) = \begin{pmatrix} O & A \\ -A^T & 0 \end{pmatrix}$ which is skew-symmetric. Note that if $\eta = (4\sigma_{max}(A))^{-1}$, then $0 < \eta < L$ where L is the Lipschitz constant of v.

We check that $\operatorname{Ker} \nabla v(\omega^*)^2 = \operatorname{Ker} \nabla v(\omega^*)$. Let $X \in \mathbb{R}^{m+p}$ such that $\nabla v(\omega^*)^2 X = 0$. As $\nabla v(\omega^*)$ is skew-symmetric, this is equivalent to $\nabla v(\omega^*)^T \nabla v(\omega^*) X = 0$ which implies that $\|\nabla v(\omega^*)\| = 0$ which implies our claim. By Lem. 16, this implies that $\operatorname{Ker}(\nabla F_{2,\eta}(\omega^*) - I_d)^2 = \operatorname{Ker}(\nabla F_{2,\eta}(\omega^*) - I_d)$. Moreover, if $\xi \in \operatorname{Ker}(\nabla F(\omega^*) - I_d)$ then by Lem. 16, $\xi \in \operatorname{Ker} \nabla v(\omega^*)$ and so $v(\omega^* + \xi) = 0$ too. Hence the hypothesises of Cor. 4 are satisfied. Then, by our choice of η and Lem. 1,

$$\begin{aligned}
\rho^* &= \max\{|\lambda| \mid \lambda \in \operatorname{Sp} \nabla F_{2,\eta}(\omega^*),\ \lambda \neq 1\} \\
&= \max\{|1 - \eta\lambda(1 - \eta\lambda)| \mid \lambda \in \operatorname{Sp} \nabla v(\omega^*),\ \lambda \neq 0\} \\
&= \max\{|1 - \eta\lambda(1 - \eta\lambda)| \mid \lambda = \pm i\sigma,\ \sigma^2 \in \operatorname{Sp} AA^T,\ \sigma \neq 0\}\,,
\end{aligned}$$

by a similar reasoning as Lem. 7 since $\operatorname{Sp} AA^T \setminus \{0\} = \operatorname{Sp} A^T A \setminus \{0\}$.

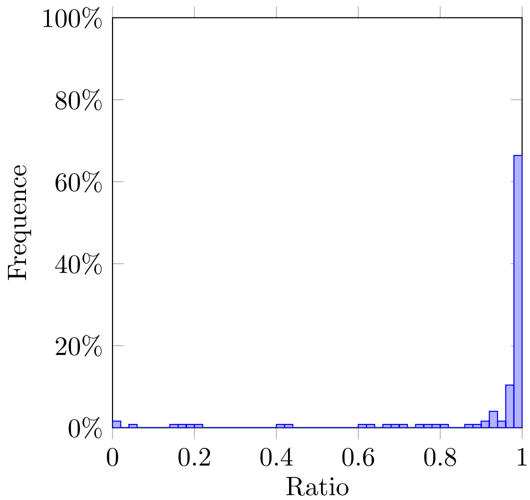
The result is now a consequence of the proof of Thm. 4. □

J Improvement of global rate

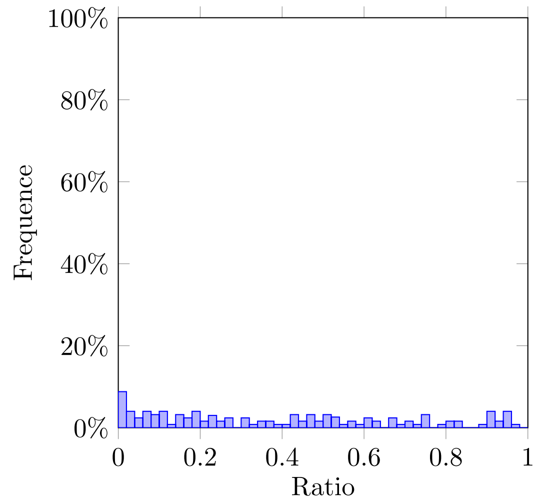
In this section we study experimentally the importance of the term $\eta^2\gamma^2$ in the global rate of Thm. 6. For this we generate two player zero-sum random montone matrix games, that is to say saddle-point problems of the form

$$\min_{\omega_1\in\mathbb{R}^{d_1}} \max_{\omega_2\in\mathbb{R}^{d_2}} \begin{pmatrix}\omega_1 & \omega_2\end{pmatrix} \begin{pmatrix} S_1 & A \\ -A^T & S_2 \end{pmatrix} \begin{pmatrix}\omega_1 \\ \omega_2\end{pmatrix}, \tag{271}$$

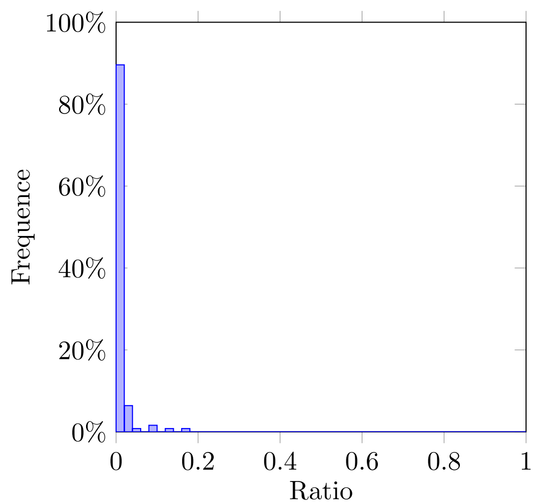
where S_1 and S_2 are symmetric semi-definite positive. To generate a symmetric semi-definite positive matrix of dimension m, we first draw independently m non-negative scalars $\lambda_1, \ldots, \lambda_m$ according to the chi-squared law.



(a) $d_1 = 450$ and $d_2 = 50$



(b) $d_1 = 350$ and $d_2 = 150$



(c) $d_1 = 250$ and $d_2 = 250$

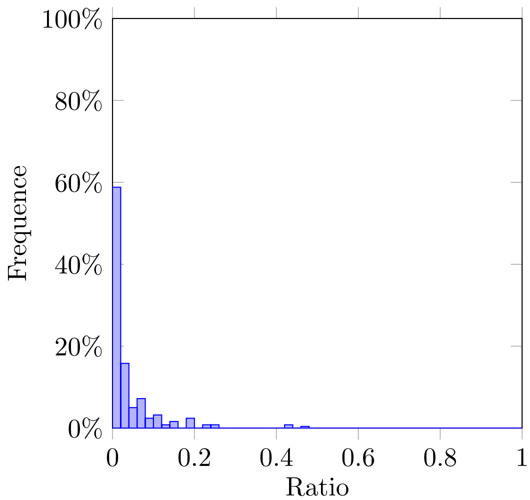
Figure 2: We keep $d_1 + d_2 = 500$ fixed and vary the balance between the two players.

Then, we draw an orthogonal matrix O according to the uniform distribution over the orthogonal group. The result is $S = O^T \operatorname{diag}(\lambda_1, \dots, \lambda_m)O$. The coefficients of A are chosen independently according to a normal law $\mathcal{N}(0, 1)$.

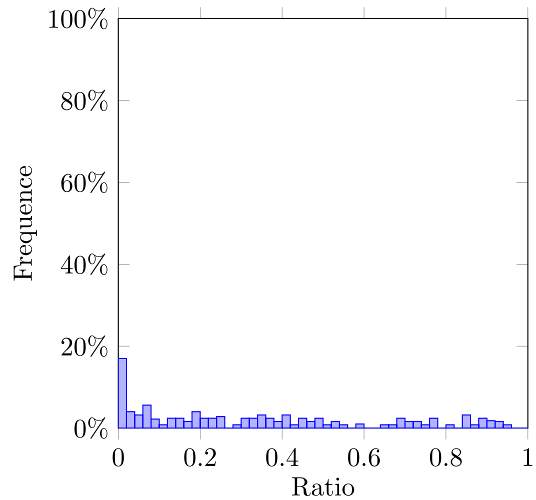
To study how the use of Tseng's error bound improves the standard rate which uses the strong monotonicty only, we compute, for each such matrix game, the ratio $\frac{\eta\mu}{\eta\mu + \frac{7}{16}\eta^2\gamma^2}$ with $\eta = (4L)^{-1}$ (with the same notations as Thm. 6). This ratio lies between 0 and 1: if it close to 0, it means that $\eta^2\gamma^2$ is much bigger than $\eta\mu$ so that our new convergence rate improves the previous one a lot, while if its near 1, it means that $\eta^2\gamma^2$ is much smaller than $\eta\mu$ and so that our new result does not improve much.

We realize two sets of graphics, each time keeping a different parameter fixed. These histograms are constructed from $N = 500$ samples.

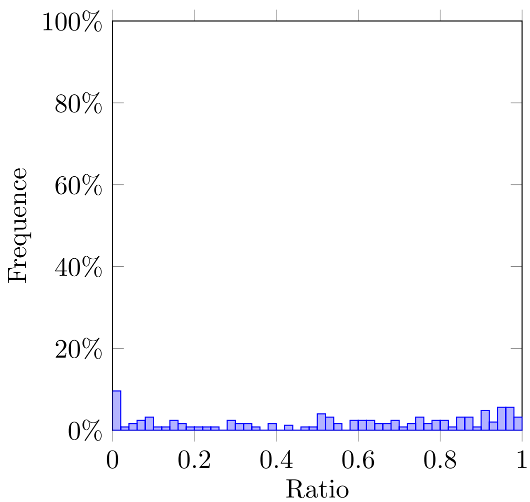
What observe is that, as soon as none of the dimensions are too small, our new rate improves the previous one in many situations. This is greatly amplified if d_1 and d_2 are similar.



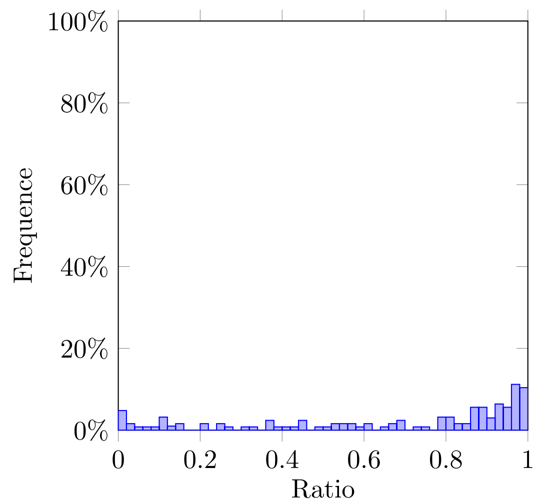
(a) $d_1 = 100$ and $d_2 = 100$



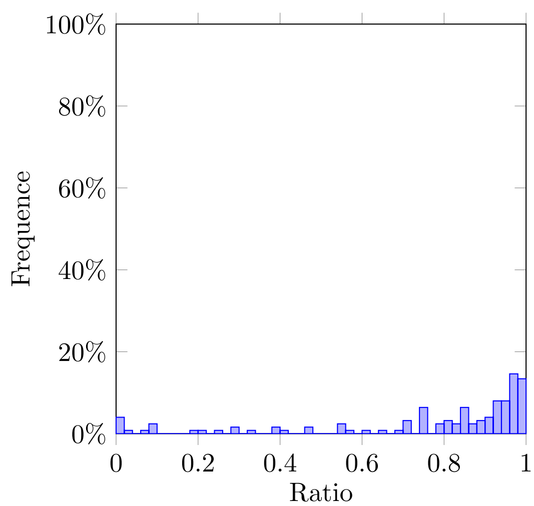
(b) $d_1 = 100$ and $d_2 = 200$



(c) $d_1 = 100$ and $d_2 = 300$



(d) $d_1 = 100$ and $d_2 = 400$



(e) $d_1 = 100$ and $d_2 = 500$

Figure 3: We keep $d_1 = 100$ fixed and make d_2 vary.