Linear Lower Bounds and Conditioning of Differentiable Games

Adam Ibrahim*
Mila, University of Montreal

Waïss Azizian†
École Normale Supérieure, Paris

Gauthier Gidel
Mila, University of Montreal

Ioannis Mitliagkas‡
Mila, University of Montreal

Abstract

Many recent machine learning tools rely on differentiable game formulations. While several numerical methods have been proposed for these types of games, most of the work has been on convergence proofs or on upper bounds for the rate of convergence of those methods. In this work, we approach the question of fundamental iteration complexity by providing lower bounds. We extend to games the p-SCLI framework used to derive spectral lower bounds for a large class of derivative-based single-objective optimisers. Additionally, we propose a definition of the condition number arising from our lower bound analysis that matches the conditioning observed in upper bounds. Our condition number is more expressive than previously used definitions, as it covers a wide range of games, including bilinear games that lack strong convex-concavity.

1 Introduction

Game formulations arise commonly in many fields, such as game theory [Harker and Pang, 1990] or machine learning [Kim and Boyd, 2008, Goodfellow et al., 2014, Chambolle and Pock, 2011, Wang et al., 2014] among others, and encompass saddle-point problems [Palaniappan and Bach, 2016, Chambolle and Pock, 2011, Chen et al., 2017]. The machine learning community has been overwhelmingly using gradient-based methods to train differentiable games [Goodfellow et al., 2014, Salimans et al., 2016]. These methods are not designed with game dynamics in mind [Mescheder et al., 2017], and to make matters worse, have been tuned suboptimally [Gidel et al., 2019b]. A recent series of publications in machine learning brings in tools from the minimax and game theory literature to offer better, faster alternatives [Gidel et al., 2019a, Daskalakis et al., 2018, Gidel et al., 2019b]. This exciting trend begs the question: how fast can we go? Knowing the fundamental limits of this class of problems is critical in steering future algorithmic research.

When studying lower bounds for convex-concave min-max, one is faced with a number of distinct challenges compared to the optimisation setting. In particular, there is no universally accepted definition of a condition number. Some commonly used definitions, like the one used in Chambolle and Pock [2011], Palaniappan and Bach [2016], give a practically infinite condition number for bilinear problems. This is problematic because we know that both extragradient and gradient methods with negative momentum achieve linear convergence in bilinear games [Korpelevich, 1976, Gidel et al., 2019b]. *Can we get a condition number that captures the fact that linear rates are possible even in the absence of strong convex-concavity?*

* <firstname>.<lastname>@umontreal.ca

† Work done during an internship at Mila.

‡ Canada CIFAR AI chair

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

We answer this question in the affirmative by providing new lower bounds yielding meaningful condition numbers even for the bilinear case, in the absence of strong convex-concavity. In the interest of brevity, we will omit here a lower bound presented in our full paper that is based on an infinite-dimensional counterexample and a generalisation of an argument by Nesterov [2004].

2 Background

Differentiable Games Following the definition of Balduzzi et al. [2018], a *differentiable game* is characterised by n players, each associated with a set of parameters $\mathbf{w}_i \in \mathbb{R}^{d_i}$ and a twice continuously differentiable objective function $l_i : \mathbb{R}^d \to \mathbb{R}$ of all the parameters $\mathbf{w} = (\mathbf{w}_1, ..., \mathbf{w}_n) \in \mathbb{R}^d$, where $d = \sum_{i=1}^n d_i$.

Often, we seek to minimise the objectives l_i, and look for *Nash equilibria* $\mathbf{w}^* = (\mathbf{w}_1^*, ..., \mathbf{w}_n^*)$, which satisfy for all i

$$\mathbf{w}_i^* \in \arg\min_{\mathbf{w}_i} l_i \left(\mathbf{w}_1^*, ..., \mathbf{w}_{i-1}^*, \mathbf{w}_i, \mathbf{w}_{i+1}^*, ..., \mathbf{w}_n^*\right) . \tag{1}$$

In order to find the Nash equilibria, we may look for stationary points, corresponding to the zeros of the vector field $\boldsymbol{v}(\mathbf{w}) = (\nabla_{\mathbf{w}_1} l_1(\mathbf{w}), ..., \nabla_{\mathbf{w}_n} l_n(\mathbf{w}))$. In single-objective optimisation, which corresponds to a 1-player game, we know that stationary points of $\boldsymbol{v}$ do not necessarily represent minima of the objective function, and additional information, such as the Hessian, is necessary to determine whether a stationary point is a minimum. The same is true for a game with several players [Balduzzi et al., 2018], where the *Jacobian* of $\boldsymbol{v}$, given by $\nabla \boldsymbol{v}(\mathbf{w})$, gives sufficient conditions to determine whether a stationary point is a Nash equilibrium.

In order to gain insight on general games, we focus on quadratic games, corresponding to games with quadratic losses. In such games, the vector field is given by:

$$\boldsymbol{v}(\mathbf{w}) = \boldsymbol{A}\mathbf{w} + \mathbf{b}, \qquad \boldsymbol{A} \triangleq \begin{pmatrix} \boldsymbol{M}_{11} & \dots & \boldsymbol{M}_{1n} \\ \vdots & & \vdots \\ \boldsymbol{M}_{n1} & \dots & \boldsymbol{M}_{nn} \end{pmatrix}, \tag{2}$$

where $\boldsymbol{A}$ is the Jacobian of $\boldsymbol{v}$ and where we may assume the $\boldsymbol{M}_{ii} \in \mathbb{R}^{d_i \times d_i}$ to be symmetric (shown in full paper). In particular, the optimisation of min-max problems of the form

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^{d_1}} \max_{\boldsymbol{y} \in \mathbb{R}^{d_2}} \quad & f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{y} + \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{S}_1 \boldsymbol{x} - \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{S}_2 \boldsymbol{y} \\ & \qquad + \boldsymbol{x}^\top \mathbf{b}_1 - \boldsymbol{y}^\top \mathbf{b}_2 + c \\ \text{where} \quad & \sigma(\boldsymbol{M}\boldsymbol{M}^\top),\ \sigma(\boldsymbol{S}_1),\ \sigma(\boldsymbol{S}_2) \subseteq [0, +\infty) \end{aligned} \tag{$\mathcal{P}$}$$

is equivalent, as we show in our full paper, to finding Nash equilibria of games with vector field

$$\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) = \begin{pmatrix} \boldsymbol{S}_1 & \boldsymbol{M} \\ -\boldsymbol{M}^\top & \boldsymbol{S}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} + \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} . \tag{3}$$

Therefore, solutions to problems in $\mathcal{P}$ exist if and only if the corresponding games with vector field given in eq. 3 admit a Nash equilibrium since $\boldsymbol{S}_1, \boldsymbol{S}_2 \succeq 0$. As we could also go from a quadratic game satisfying eq. 3 to a min-max formulation, any lower bound on quadratic games of the form of eq. 3 is a lower bound on min-max problems (in $\mathcal{P}$), and vice versa.

In this paper, we will denote the spectrum of a matrix $\boldsymbol{M}$ by $\sigma(\boldsymbol{M})$, and define the *block spectral bounds* $\mu_1, \mu_2, \mu_{12}, L_1, L_2, L_{12}$ as constants bounding the spectra of the blocks in the Jacobian of eq. 3 such that for $i \in \{1, 2\}$:

$$\mu_i \le |\sigma(\boldsymbol{S}_i)| \le L_i \qquad \mu_{12}^2 \le |\sigma(\boldsymbol{M}\boldsymbol{M}^\top)| \le L_{12}^2 \tag{4}$$

Existing bounds for 2-player quadratic games Some upper bounds for problems in $\mathcal{P}$ exist for certain optimisation algorithms. Letting $\kappa = \frac{L_{12}}{\sqrt{\mu_1 \mu_2}}$, Chen and Rockafellar [1997] find a rate of $\sqrt{1 - \frac{\min(\mu_1, \mu_2, \mu_{12})}{\max(L_1, L_2, L_{12})}}$ for the forward-backward algorithm. Chambolle and Pock [2011] give an algorithm for which the best rate is $\sqrt{1-\frac{2}{\kappa+2}}$. Palaniappan and Bach [2016] show an accelerated version of the forward-backward algorithm with variance reduction with rate $1-\frac{1}{1+2\kappa}$. Finally, Gidel et al. [2019b] give an upper bound $1-\frac{1}{4L_{12}^2/\mu_{12}^2}$ on the rate of convergence of alternating gradient descent with negative momentum for bilinear games, i.e. games satisfying eq. 3 with $\boldsymbol{S}_1=\boldsymbol{S}_2=0$.

On the other hand, relevant lower bounds for first-order methods on saddle-point problems are scarcer in the literature. Nemirovsky [1992] gives a lower bound in $\mathcal{O}(1/t)$ for a limited number of steps. Ouyang and Xu [2018] also leverage Krylov subspace techniques, and show lower bounds in $\mathcal{O}(1/t)$ and $\mathcal{O}(1/t^2)$, assuming the number of iterations is less than half the dimension of the parameters. A key issue is that since these bounds are only valid for a limited number of steps, they do not yield bounds that can be compared with the upper bounds previously mentioned. In contrast, the lower bounds presented in this work are valid for any number of steps and are linear, and therefore provide a direct limit to the acceleration of methods achieving linear convergence on two-player games. Additionally, our lower bounds also yield condition numbers that give intuition about the difficulty inherent to a problem, and can be computed in a plug-and-play fashion using either bounds on the spectrum of the full Jacobian, or on the spectra of its blocks.

Lower bounds in single-objective optimisation In single-objective optimisation, i.e. a 1-player game, the Jacobian of the vector field reduces to the Hessian of the objective, denoted $\boldsymbol{H}(\boldsymbol{x})$. In that case, if there exists $\mu, L \in \mathbb{R}^{++}$ such that for all $\boldsymbol{x}$ in the domain considered $\mu \preceq \boldsymbol{H}(\boldsymbol{x}) \preceq L$, the objective is μ-strongly convex and has L-Lipschitz gradients, and the convergence rates are known to be linear in the number of iterations for several classes of algorithms [Nemirovsky and Yudin, 1983, Nesterov, 2004]. In particular, Arjevani et al. [2016] introduce the p-SCLI framework to provide bounds for a large class of methods used for optimising such objectives. Roughly speaking, an algorithm is p-SCLI if its update rule on quadratics $f(\boldsymbol{x})=\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^\top \mathbf{b}$ with $\boldsymbol{A}$ symmetric is a linear combination of the p previous iterates and $\mathbf{b}$, where the coefficients are matrices that depend on $\boldsymbol{A}$ and are assumed to be simultaneously triangularisable. The spectral properties of the update rule are used to derive a lower bound on the rate of convergence of p-SCLI methods $\rho \geq 1-\frac{2}{\sqrt[p]{\kappa}+1}$ for $\kappa = L/\mu$. Thus, we recover lower bounds for gradient descent ($p=1$) or Nesterov [1983]'s accelerated gradient descent ($p=2$) that match the upper bounds.

When there are several players, however, the Jacobian is no longer symmetric, and its spectrum will generally be complex, and hence several of the arguments used in single-objective optimisation fail to apply.

3 p-SCLI-n for n-player games

Let $\mathcal{Q}^{d_1,\ldots,d_n}$ denote the set of vectors of n quadratic losses $l_i : \mathbb{R}^{d_i} \to \mathbb{R}$ corresponding to n-player quadratic games, and $f_{\boldsymbol{A},\mathbf{b}}(\mathbf{w}) \in \mathcal{Q}^{d_1,\ldots,d_n}$ be a game with vector field $\boldsymbol{A}\mathbf{w}+\mathbf{b}$ as indicated in eq. 2. The following definition is a generalisation of the definition of p-SCLI algorithms given by Arjevani et al. [2016] to n-player games.

Definition 1 (p-SCLI-n optimisation algorithms for n-player games)**.** *Let $\mathcal{A}$ be an optimisation algorithm for n-player quadratic games. Then $\mathcal{A}$ is a p-stationary canonical linear iterative method for n-player games (p-SCLI-n) if there exist functions $C_0, \ldots, C_{p-1}, N$ from $\mathbb{R}^{d\times d}$ to $\mathbb{R}^{d\times d}$-valued random variables, such that the following conditions are satisfied for all $f_{\boldsymbol{A},\mathbf{b}}(\mathbf{w}) \in \mathcal{Q}^{d_1,\ldots,d_n}$:*

1. *Given an initialisation $\mathbf{w}^0, \ldots, \mathbf{w}^{p-1} \in \mathbb{R}^d$, the update rule at iteration $t \geq p$ is given by*

$$\mathbf{w}^t = \sum_{i=0}^{p-1} C_i(\boldsymbol{A})\mathbf{w}^{t-p+i} + N(\boldsymbol{A})\mathbf{b} \tag{5}$$

2. *$C_0(\boldsymbol{A}), \ldots, C_{p-1}(\boldsymbol{A}), N(\boldsymbol{A})$ are independent from previous iterations*

3. *$\mathbb{E}C_i(\boldsymbol{A})$ are finite and simultaneously triangularisable*

We will refer to the C_i as the coefficient matrices and N as the inversion matrix.
Furthermore, given a Jacobian $\boldsymbol{A}$ as in eq. 2, $\mathcal{A}$ is said to be consistent with respect to $\boldsymbol{A}$ if for all $f_{\boldsymbol{A},\mathbf{b}}$, the sequence of iterates satisfies $(\mathbf{w}^t) \to -\boldsymbol{A}^{-1}\mathbf{b}$.

Examples of p-SCLI-n methods include simultaneous gradient descent (possibly with momentum if all the momentum parameters are equal) and extragradient.

We are now ready to introduce the lower bound for p-SCLI-n methods with scalar inversion matrix, such as simultaneous implementations of gradient descent where the step-size is the same for all objectives.

Proposition 2. *Let $\mathcal{A}$ be a p-SCLI-n algorithm with scalar inversion matrix for optimising games over $\mathbb{R}^{d_1} \times \ldots \times \mathbb{R}^{d_n}$. Then for quadratics $f_{\boldsymbol{A},\mathbf{b}} \in \mathcal{Q}^{d_1,\ldots,d_n}$, if $\mathcal{A}$ is consistent with respect to $\boldsymbol{A}$ and if $0 \notin \sigma(\boldsymbol{A})$, we have the following lower bound on the (linear) rate of convergence ρ:*

$$\rho_* \geq \frac{\sqrt[p]{\kappa}-1}{\sqrt[p]{\kappa}+1} = 1 - \frac{2}{\sqrt[p]{\kappa}+1} \tag{6}$$

where the condition number κ is defined as $\kappa \triangleq \frac{\max|\sigma(\boldsymbol{A})|}{\min|\sigma(\boldsymbol{A})|}$ where $\sigma(\boldsymbol{A})$ is the spectrum of $\boldsymbol{A}$.

Interestingly, this is close, and in fact, reduces for $n = 1$ to the bound obtained in the 1-player case for μ-strongly convex objectives with L-Lipschitz gradients, where $\kappa = \frac{\max \sigma(\boldsymbol{A})}{\min \sigma(\boldsymbol{A})} = L/\mu$. Moreover, this form is valid for n-player games (and min-max problems) in finite dimension, and κ arises naturally from the spectral properties of the update rules of the p-SCLI-n methods and is lower bounded by 1.

However, while the moduli in the $n > 1$ case allow us to handle complex spectra, several analyses have shown that not only the modulus, but also the relative size of the real and imaginary parts of elements of the spectrum matter [Mescheder et al., 2017, Gidel et al., 2019b]. Such an analysis is out of the scope of this work. We will nevertheless give a more explicit form of the bound for 2-player games for which $d_1 = d_2$ that will make the μ_i and L_i appear.

Some explicit bounds for p-SCLI-2 with $d_1 = d_2$ Prop. 2 may be used to derive lower bounds for 2-player games for which $d_1 = d_2$. These bounds depend on the value of the μ_i and L_i defined as in eq. 4. Namely, let

$$\Delta_\mu = (\mu_1 - \mu_2)^2 - 4\mu_{12}^2\,, \qquad\qquad \Delta_L = (L_1 - L_2)^2 - 4L_{12}^2 \tag{7}$$

Table 1 gives lower bounds on the condition number that may then be plugged into eq. 6 to get lower bounds on two-players games corresponding to min-max problems (and are therefore lower bounds for general 2-player games).

Table 1: Lower bounds on the condition number

	$\Delta_\mu < 0$	$\Delta_\mu \geq 0$
$\Delta_L < 0$	$\kappa = \sqrt{\frac{L_1L_2+L_{12}^2}{\mu_1\mu_2+\mu_{12}^2}}$	$\kappa \geq 2\frac{\sqrt{L_1L_2+L_{12}^2}}{\mu_1+\mu_2-\sqrt{\Delta_\mu}}$
$\Delta_L \geq 0$	$\kappa \geq \frac{1}{2}\frac{L_1+L_2+\sqrt{\Delta_L}}{\sqrt{\mu_1\mu_2+\mu_{12}^2}}$	$\kappa \geq \frac{L_1+L_2+\sqrt{\Delta_L}}{\mu_1+\mu_2-\sqrt{\Delta_\mu}}$

4 Conclusion

In this work, we generalise the framework of p-SCLI to provide bounds for a large class of optimisers for n-player games, and give explicit bounds for 2-player games and min-max problems. Moreover, we derived formulations for the condition number that encompass the existing ones in the upper bound literature. As in the single-objective case, our bounds and condition numbers suggest that optimisers may converge faster on games for which the eigenvalues are at a similar, remote distance from the origin (e.g. on a circle) than on games for which some eigenvalues are close to and others are far from 0.

Acknowledgements

This work was supported by the FRQNT new researcher program (2019-NC-257943), the NSERC Discovery grant (RGPIN-2019-06512), a startup grant by IVADO and a Canada CIFAR AI chair. GG was partially supported by a Borealis AI Graduate Fellowship.

References

Yossi Arjevani, Shai Shalev-Shwartz, and Ohad Shamir. On lower and upper bounds in smooth and strongly convex optimization. *The Journal of Machine Learning Research*, 17(1):4303–4353, 2016.

David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 363–372, 2018.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

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

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

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

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019a.

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

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

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

Seung-Jean Kim and Stephen Boyd. A minimax theorem with applications to machine learning, signal processing, and finance. *SIAM Journal on Optimization*, 19(3):1344–1367, 2008.

GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

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

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

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

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

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

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

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Hongxing Wang, Chaoqun Weng, and Junsong Yuan. Multi-feature spectral clustering with minimax optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4106–4113, 2014.

Fuzhen Zhang, editor. *The Schur Complement and Its Applications*, volume 4 of *Numerical Methods and Algorithms*. Springer-Verlag, New York, 2005. ISBN 978-0-387-24271-2.

Appendix A Proofs

A.1 Proof of Prop. 2

In this section, we follow Arjevani et al. [2016] to derive results for the p-SCLI-n methods. First, we reproduce several definitions and theorems that are proven in Arjevani et al. [2016] and that apply directly to the generalisation. Here, $\boldsymbol{A}$ will denote the Jacobian of some quadratic game with $f_{\boldsymbol{A},\mathbf{b}} \in \mathcal{Q}^{d_1,\ldots,d_n}$ such that 0 is not in the spectrum of $\boldsymbol{A}$.

Definition 3 (Characteristic polynomial of a p-SCLI-n)**.** *Let $\mathcal{A}$ be a p-SCLI-n optimisation algorithm with coefficient matrices C_i as defined in def. 1. Then for $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, the characteristic polynomial of $\mathcal{A}$ is given by*

$$\mathcal{L}(\lambda, \boldsymbol{X}) \triangleq \boldsymbol{I}_d \lambda^p - \sum_{i=0}^{p-1} \mathbb{E}C_i(\boldsymbol{X})\lambda^i \tag{8}$$

and its root radius is

$$\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{X})) = \rho(\det \mathcal{L}(\lambda, \boldsymbol{X})) = \max\{|\lambda| \mid \det \mathcal{L}(\lambda, \boldsymbol{X}) = 0\}$$

Theorem 4 (Consistency - characteristic polynomial (Based on Theorem 5 of Arjevani et al. [2016]))**.** *A p-SCLI-n algorithm $\mathcal{A}$ with characteristic polynomial $\mathcal{L}(\lambda, \boldsymbol{X})$ and inversion matrix $N(\boldsymbol{X})$ is consistent with respect to $\boldsymbol{A}$ if and only if the following two conditions hold:*

$$1.\ \mathcal{L}(1, \boldsymbol{A}) = -\mathbb{E}N(\boldsymbol{A})\boldsymbol{A} \tag{9}$$

$$2.\ \rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A})) < 1 \tag{10}$$

We may rephrase theorem 13 of Arjevani et al. [2016] (and lower bound t^{m-1} by 1 since $m \in \mathbb{N}$) as the following to use the root radius of the characteristic polynomial to show linear rates:

Theorem 5 (Based on Theorem 13 of Arjevani et al. [2016])**.** *If $\boldsymbol{A}$ is the Jacobian of a quadratic game and $\mathcal{A}$ is a p-SCLI-n, there exists an initialisation point $\mathbf{w}_0 \in \mathbb{R}^d$ such that*

$$\max_{i=0,\ldots,p-1} \left\|\mathbb{E}\mathbf{w}^{t+i} - \mathbb{E}\mathbf{w}^*\right\| \in \Omega(\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A}))^t) \tag{11}$$

In other words, this means that $\mathcal{A}$ cannot converge on $f_{\boldsymbol{A},\mathbf{b}}$ with linear rate faster than $\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A}))$, up to a constant. As Arjevani et al. [2016] argue, in both deterministic and stochastic settings, a lower bound on $\|\mathbb{E}[\mathbf{w}^t - \mathbf{w}^*]\|^2$ implies[4] a lower bound on $\mathbb{E}\|\mathbf{w}^t - \mathbf{w}^*\|^2$, since

$$\mathbb{E}\left[\left\|\mathbf{w}^t - \mathbf{w}^*\right\|^2\right] = \mathbb{E}\left[\left\|\mathbf{w}^t - \mathbb{E}\mathbf{w}^t\right\|^2\right] + \left\|\mathbb{E}\left[\mathbf{w}^t - \mathbf{w}^*\right]\right\|^2 \tag{12}$$

We can now focus on finding a lower bound on $\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A}))$.

Proposition 6. *Let $\mathcal{A}$ be a p-SCLI-n optimisation algorithm with inversion matrix $N(\boldsymbol{X})$ that is consistent with respect to $\boldsymbol{A}$. Then,*

$$\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A})) \geq \max_{j=1,\ldots,d} |\sqrt[p]{|\sigma_j(-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A})|} - 1| \tag{13}$$

where the $\sigma_j(-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A})$ are elements of the spectrum (eigenvalues) of $-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A}$.

A.1.1 Proof of Prop. 6

Our proof starts exactly as the one presented by Arjevani et al. [2016] for the $n = 1$ particular case, where the authors assume that $\boldsymbol{A}$ is symmetric with strictly positive spectrum. However, we will generalise the proof to cover non-symmetric matrices and matrices that may not have strictly positive spectrum, since the Jacobian of a quadratic n-player game generally does not have these properties.

Let $\mathcal{A}$ be a deterministic p-SCLI-n optimisation algorithm with characteristic polynomial $\mathcal{L}(\lambda, \boldsymbol{X})$ and inversion matrix $N(\boldsymbol{X})$, and $f_{A,\mathbf{b}}(\mathbf{w}) \in \mathcal{Q}^{d_1,\ldots,d_n}$ represent a quadratic n-player game. Since

[4] Note that since we only use in this paper a lower bound on the second term of the right hand-side of the equation to bound the left hand-side, one may derive in stochastic settings tighter lower bounds than the ones presented in this paper by factoring in the first term of the right hand-side. We leave this as future work.

$\mathcal{A}$ is p-SCLI-n, its (expected) coefficient matrices $\mathbb{E}C_i$ evaluated on $\boldsymbol{A}$ are simultaneously triangularisable, so $\exists \boldsymbol{Q} \in \mathbb{R}^{d\times d}$ such that for $i = 0, ..., p-1$, we have

$$\boldsymbol{T}_i \triangleq \boldsymbol{Q}^{-1}\mathbb{E}C_i(\boldsymbol{A})\boldsymbol{Q} \tag{14}$$

where $\boldsymbol{T}_i$ is triangular. Thus,

$$\begin{aligned}\det \mathcal{L}(\lambda, \boldsymbol{A}) &= \det\left(\boldsymbol{Q}^{-1}\mathcal{L}(\lambda, \boldsymbol{A})\boldsymbol{Q}\right)\\ &= \det\left(\boldsymbol{I}_d\lambda^p - \sum_{i=0}^{p-1}\boldsymbol{T}_i\lambda^i\right)\end{aligned} \tag{15}$$

Since $\boldsymbol{I}_d\lambda^p - \sum_{i=0}^{p-1}\boldsymbol{T}_i\lambda^i$ is a upper triangular matrix, its determinant is given by

$$\det \mathcal{L}(\lambda, \boldsymbol{A}) = \prod_{j=1}^{d} \ell_j(\lambda) \tag{16}$$

where

$$\ell_j(\lambda) = \lambda^p - \sum_{i=0}^{p-1}\sigma_j^i\lambda^i \tag{17}$$

and where $\sigma_1^i, \ldots, \sigma_d^i,\ i = 0, \ldots, p-1$ denote the elements on the diagonal of $\boldsymbol{T}_i$, which are just the eigenvalues of $\mathbb{E}C_i$ ordered according to $\boldsymbol{Q}$. Hence, the root radius of the characteristic polynomial of $\mathcal{A}$ is

$$\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A})) = \max\{|\lambda| \mid \ell_j(\lambda) = 0 \text{ for some } j = 1, ..., d\} \tag{18}$$

On the other hand, by consistency condition (9) we get that for all $j = 1, ..., d$

$$\ell_j(1) = \sigma_j(\mathcal{L}(1, \boldsymbol{A})) = \sigma_j(-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A}) \tag{19}$$

In the case of p-SCLI-1, the authors prove their Corollary 7 (i.e. our prop. 6 without taking the modulus of the eigenvalues) by using a lemma (see Lemma 6 in Arjevani et al. [2016]) that gives a lower bound on each $\rho(\ell_j(\lambda))$ by using the sign of $\ell_j(1) = \sigma_j(-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A})$. Lemma 6 of Arjevani et al. [2016] is proven using the following lemma, which we can in fact use to handle arbitrary eigenvalues (e.g. complex or negative).

Lemma 7 (Lemma 15 of Arjevani et al. [2016])**.** *Let $q_r^*(z) \triangleq (z - (1 - \sqrt[p]{r}))^p$ where r is some non-negative constant. Suppose $q(z)$ is a monic polynomial of degree p with complex coefficients. Then,*

$$\rho(q(z)) \leq |\sqrt[p]{|q(1)|} - 1| \iff q(z) = q^*_{|q(1)|}(z)$$

The proof of the lemma can be found in Arjevani et al. [2016]. Here, we can use the lemma directly on each ℓ_j with $q = \ell_j$ and $r = |q(1)| = |\ell_j(1)| = |\sigma_j(-\mathbb{E}[N(\boldsymbol{A})]A)|$. Indeed, since $r \geq 0$,

- if $q(z) = q_r^*(z) = (z - (1 - \sqrt[p]{r}))^p$ then clearly $\rho(q(z)) = |1 - \sqrt[p]{r}|$
- if $q(z) \neq q_r^*(z)$, then we have $\rho(q(z)) > |\sqrt[p]{|q(1)|} - 1|$

Which implies that for any j we have $\rho(\ell_j(\lambda)) \geq |\sqrt[p]{|\ell_j(1)|} - 1| = |\sqrt[p]{|\sigma_j(-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A})|} - 1|$. Using this in eq. 18 yields

$$\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A})) \geq \max_{j=1,\ldots,d} |\sqrt[p]{|\sigma_j(-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A})|} - 1| \tag{20}$$

A.1.2 Deriving the optimal ρ for scalar inversion matrices

We are now ready to obtain the general lower bound. Consider $f_{\boldsymbol{A},\mathbf{b}} \in \mathcal{Q}^{d_1,\ldots,d_n}$ with $0 \notin \sigma(\boldsymbol{A})$ and a consistent p-SCLI-n algorithm $\mathcal{A}$. Let $\mu = \min|\sigma(\boldsymbol{A})|$, $L = \max|\sigma(\boldsymbol{A})|$ where $\sigma(\boldsymbol{A})$ is the spectrum of $\boldsymbol{A}$. For a scalar inversion matrix i.e. $\mathbb{E}[N(\boldsymbol{A})] = \nu$ we have from eq. 20:

$$\begin{aligned}\rho_\lambda(\mathcal{L}(\lambda, \boldsymbol{A})) &\geq \max_{j=1,\ldots,d} |\sqrt[p]{|\sigma_j(-\mathbb{E}[N(\boldsymbol{A})]\boldsymbol{A})|} - 1| = \max_{j=1,\ldots,d} |\sqrt[p]{|\nu\sigma_j(\boldsymbol{A})|} - 1|\\ &= \max\left\{|\sqrt[p]{|\nu|\mu} - 1|, |\sqrt[p]{|\nu|L} - 1|\right\}\end{aligned} \tag{21}$$

Note that consistency (eq. 10) constrains $\nu \in \left(\frac{-2^p}{L}, \frac{2^p}{L}\right) \setminus \{0\}$. We proceed as Arjevani et al. [2016] in the p-SCLI-1 case, and study the ranges of $|\nu|$ by using $\max(a,b) = \frac{a+b+|a-b|}{2}$ to obtain table 2.

Table 2: Lower bound for ρ_* by subranges of $|\nu|$ and minimiser $|\nu^*|$

	$\sqrt[p]{|\nu|\mu} - 1 < 0$			$\sqrt[p]{|\nu|\mu} - 1 \geq 0$		
	Range	Minimiser	Bound	Range	Minimiser	Bound
$\sqrt[p]{|\nu|L} - 1 \leq 0$	$(0, 1/L]$	$1/L$	$1 - \sqrt[p]{\frac{\mu}{L}}$		N/A	
$\sqrt[p]{|\nu|L} - 1 > 0$	$(1/L, 1/\mu)$	$\left(\frac{2}{\sqrt[p]{L}+\sqrt[p]{\mu}}\right)^p$	$\frac{\sqrt[p]{L/\mu}-1}{\sqrt[p]{L/\mu}+1}$	$[1/\mu, 2^p/L)$	$1/\mu$	$\sqrt[p]{\frac{L}{\mu}} - 1$

Note that case 3 requires $p > \log_2 L/\mu$. Hence,

$$\rho_* \geq \min\left\{1 - \sqrt[p]{\frac{\mu}{L}}, \frac{\sqrt[p]{L/\mu}-1}{\sqrt[p]{L/\mu}+1}, \sqrt[p]{\frac{L}{\mu}} - 1\right\} = \frac{\sqrt[p]{L/\mu}-1}{\sqrt[p]{L/\mu}+1} \tag{22}$$

where $\mu = \min|\sigma(\boldsymbol{A})|$, $L = \max|\sigma(\boldsymbol{A})|$.

A.2 Finding a suitably hard example for 2-player with $d_1 = d_2$

We now only need to find a hard counterexample. We present the argument for $d_1 = d_2 = 2$, which can easily be generalised for arbitrary d. Consider the matrix

$$\boldsymbol{A} = \begin{pmatrix} \mu_1 & 0 & \mu_{12} & 0 \\ 0 & L_1 & 0 & L_{12} \\ -\mu_{12} & 0 & \mu_2 & 0 \\ 0 & -L_{12} & 0 & L_2 \end{pmatrix} \tag{23}$$

corresponding to the Jacobian of a quadratic game in $\mathcal{Q}^{d_1,d_2}$.

First we compute the characteristic polynomial of A, using the formula for the determinant of a block matrix (see Zhang [2005, Section 0.3] for instance):

$$\det(XI - A) = \det\begin{pmatrix} X-\mu_1 & 0 & -\mu_{12} & 0 \\ 0 & X-L_1 & 0 & -L_{12} \\ \mu_{12} & 0 & X-\mu_2 & 0 \\ 0 & L_{12} & 0 & X-L_2 \end{pmatrix} \tag{24}$$

$$= \det\left(\begin{pmatrix} (X-\mu_1)(X-\mu_2) & 0 \\ 0 & (X-L_1)(X-L_2) \end{pmatrix} + \begin{pmatrix} \mu_{12}^2 & 0 \\ 0 & L_{12}^2 \end{pmatrix}\right) \tag{25}$$

$$= (X^2 - (\mu_1+\mu_2)X + \mu_1\mu_2 + \mu_{12}^2)(X^2 - (L_1+L_2)X + L_1L_2 + L_{12}^2) \tag{26}$$

The discriminants of these two quadratic equations are, respectively:

$$\Delta_\mu = (\mu_1+\mu_2)^2 - 4(\mu_1\mu_2 + \mu_{12}^2) = (\mu_1-\mu_2)^2 - 4\mu_{12}^2 \tag{27}$$

$$\Delta_L = (L_1+L_2)^2 - 4(L_1L_2 + L_{12}^2) = (L_1-L_2)^2 - 4L_{12}^2 \tag{28}$$

which yields the following eigenvalues:

$$\lambda_{\mu\pm} = \frac{\mu_1+\mu_2}{2} \pm \sqrt{\left(\frac{\mu_1-\mu_2}{2}\right)^2 - \mu_{12}^2}$$

$$\lambda_{L\pm} = \frac{L_1+L_2}{2} \pm \sqrt{\left(\frac{L_1-L_2}{2}\right)^2 - L_{12}^2} \tag{29}$$

We distinguish four cases, which are presented in the following table:

Table 3: Lower bounds on the condition number

	$\Delta_\mu < 0$	$\Delta_\mu \geq 0$
$\Delta_L < 0$	$\kappa = \sqrt{\frac{L_1 L_2 + L_{12}^2}{\mu_1\mu_2 + \mu_{12}^2}}$	$\kappa \geq 2\frac{\sqrt{L_1 L_2 + L_{12}^2}}{\mu_1 + \mu_2 - \sqrt{\Delta_\mu}}$
$\Delta_L \geq 0$	$\kappa \geq \frac{1}{2}\frac{L_1 + L_2 + \sqrt{\Delta_L}}{\sqrt{\mu_1\mu_2 + \mu_{12}^2}}$	$\kappa \geq \frac{L_1 + L_2 + \sqrt{\Delta_L}}{\mu_1 + \mu_2 - \sqrt{\Delta_\mu}}$

where we used that $\kappa = \frac{\max|\sigma(\boldsymbol{A})|}{\min|\sigma(\boldsymbol{A})|}$.

We now discuss these four cases:

- If $\Delta_\mu < 0$ and $\Delta_L < 0$, we have that

$$
\begin{aligned}
|\lambda_{\mu\pm}| &= \left|\frac{\mu_1 + \mu_2}{2} \pm i\sqrt{\mu_{12}^2 - \left(\frac{\mu_1 - \mu_2}{2}\right)^2}\right| \\
&= \sqrt{\mu_1\mu_2 + \mu_{12}^2}
\end{aligned} \tag{30}
$$

  Similarly we get

$$
|\lambda_{L\pm}| = \sqrt{L_1 L_2 + L_{12}^2} \tag{31}
$$

  Clearly then $\min|\sigma(\boldsymbol{A})| = |\lambda_{\mu\pm}|$ and $\max|\sigma(\boldsymbol{A})| = |\lambda_{L\pm}|$, which yields $\kappa = \sqrt{\frac{L_1 L_2 + L_{12}^2}{\mu_1\mu_2 + \mu_{12}^2}}$.

- If $\Delta_\mu \geq 0$ and $\Delta_L \geq 0$, $\lambda_{L+},\ \lambda_{L-},\ \lambda_{\mu+}$ and $\lambda_{\mu-}$ are all real. We have that,

$$
\lambda_{\mu-} \geq \min|\sigma(\boldsymbol{A})|, \quad \text{and} \quad \lambda_{L+} \leq \max|\sigma(\boldsymbol{A})|, \tag{32}
$$

  which yields the result.

- If $\Delta_\mu < 0$ and $\Delta_L \geq 0$, it holds that,

$$
|\lambda_{\mu\pm}| = \min|\sigma(\boldsymbol{A})|, \quad \text{and} \quad \lambda_{L+} \leq \max|\sigma(\boldsymbol{A})|, \tag{33}
$$

  from which we obtain the result.

- Similarly, if $\Delta_\mu \geq 0$ and $\Delta_L < 0$, it holds that,

$$
\lambda_{\mu-} \geq \min|\sigma(\boldsymbol{A})|, \quad \text{and} \quad |\lambda_{L\pm}| = \max|\sigma(\boldsymbol{A})|. \tag{34}
$$

One could wonder whether our lower bounds on κ when at least one of the discriminant is non-negative are actually equalities. We provide an example showing that it is not the case when $\Delta_L \geq 0$ and $\Delta\mu \geq 0$. A similar one can be found when $\Delta_L < 0$ and $\Delta\mu \geq 0$.

Take $\mu_{12} = 0$ and $L_{12} = \frac{|L_1 - L_2|}{2}$. Then $\Delta_L \geq 0$ and $\Delta\mu \geq 0$. Then,

$$
\begin{aligned}
\lambda_{\mu+} &= \frac{\mu_1 + \mu_2}{2} + \sqrt{\left(\frac{\mu_1 - \mu_2}{2}\right)^2 - \mu_{12}^2} = \max(\mu_1, \mu_2) \\
\lambda_{L\pm} &= \frac{L_1 + L_2}{2} \pm \sqrt{\left(\frac{L_1 - L_2}{2}\right)^2 - L_{12}^2} = \frac{L_1 + L_2}{2}\,.
\end{aligned} \tag{35}
$$

Choose $\mu_1 = L_1$, $\mu_2 = L_2$ and $L_1 \neq L_2$. Then $\lambda_{\mu+} > \lambda_{L\pm}$. However we have $\lambda_{\mu-} = \min|\sigma(A)|$ and so in this case $\kappa = \lambda_{\mu+}/\lambda_{\mu-}$.