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. Copyright 2020 by the author(s).

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

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

where the vector $(\omega^{-i})^*$ contains all the coordinates of ω^* except the i^{th} one. Moreover, we say that a game is *zero-sum* if $\sum_{i=1}^n l_i = 0$. For instance, following Mescheder et al. (2017); 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*

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

associated to a n-player game and its Jacobian:

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

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,*

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

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

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

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)**.**

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

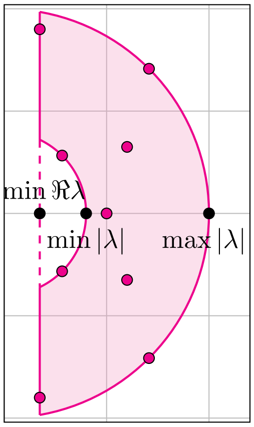

Figure 1: Illustration of the three quantities involved in Thm. 4. The magenta dots are an example of eigenvalues belonging to σ^*. Note that σ^* is always symmetric with respect to the real axis because the Jacobian is a real matrix (and thus non-real eigenvalues are complex conjugates). Note how $\min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

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

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

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

Recovery of known rates. If v is μ-strongly monotone and L-Lipschitz, this bound is at least as precise as the standard one $1 - \mu/(4L)$ as μ lower bounds the real part of the eigenvalues of the Jacobian, and L upper bounds their magnitude, as shown in Lem. 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.