Momentum Extragradient is Optimal for Games with Cross-Shaped Spectrum

Junhyung Lyle Kim[*] JLYLEKIM@RICE.EDU
Rice University
Gauthier Gidel GAUTHIER.GIDEL@UMONTREAL.CA
Mila & Université de Montréal
Canada CIFAR AI Chair
Anastasios Kyrillidis ANASTASIOS@RICE.EDU
Rice University
Fabian Pedregosa PEDREGOSA@GOOGLE.COM
Google Research

Abstract

The extragradient method has recently gained a lot of attention, due to its convergence behavior on smooth games. In games, the eigenvalues of the Jacobian of the vector field are distributed on the complex plane, exhibiting more convoluted dynamics compared to minimization. In this work, we take a polynomial-based analysis of the extragradient with momentum for optimizing games with *cross-shaped* spectrum on the complex plane. We show two results: first, the extragradient with momentum exhibits three different modes of convergence based on the hyperparameter setup: when the eigenvalues are distributed (i) on the real line, (ii) both on the real line along with complex conjugates, and (iii) only as complex conjugates. Then, we focus on the case (ii), i.e., when the spectrum of the Jacobian has *cross-shaped* structure, as observed in training generative adversarial networks. For this problem class, we derive the optimal hyperparameters and show that the extragradient with momentum achieves accelerated convergence rate.

1. Introduction

Background. Following [4, 17], we define the n-player differentiable game as a family of twice continuously differentiable losses $\ell_i : \mathbb{R}^{d_i} \to \mathbb{R}$, for $i = 1, \ldots, n$. The player i controls the parameter $w^{(i)} \in \mathbb{R}^{d_i}$. We denote the concatenated parameters by $w = [w^{(1)}, \ldots, w^{(n)}] \in \mathbb{R}^d$, where $d = \sum_{i=1}^n d_i$. For this problem, a Nash equilibrium satisfies $w^{(i),\star} \in \arg\min_{w^{(i)} \in \mathbb{R}^{d_i}} \ell_i\big(w^{(i)}, w^{(\neg i),\star}\big)$, where the notation $\cdot^{(\neg i)}$ denotes all indices except for i. We also define the vector field of the game v as the concatenation of the individual gradients, and its associated Jacobian ∇v respectively as:

$$v(w) = [\nabla_{w^{(1)}} \ell_1(w), \ldots, \nabla_{w^{(n)}} \ell_n(w)]^\top, \text{ and } \nabla v(w) = \begin{bmatrix} \nabla^2_{w^{(1)}} \ell_1(w) & \ldots & \nabla_{w^{(n)}} \nabla_{w^{(1)}} \ell_1(w) \\ \vdots & & \vdots \\ \nabla_{w^{(1)}} \nabla_{w^{(n)}} \ell_n(w) & \ldots & \nabla^2_{w^{(n)}} \ell_1(w) \end{bmatrix}.$$

Unfortunately, finding Nash equilibria for general games is impractical to solve [17, 28].[1] Instead, we focus on finding a stationary point of the vector field v, since a Nash equilibrium is always a

[*] Authors after JLK are listed in alphabetical order.

[1] Finding Nash equilibira can be refomulated as a nonlinear complementarity problem, which is PPAD hard [7, 17].

stationary point of the gradient dynamics. That is, we want to solve:

$$\text{Find} \quad w^\star \in \mathbb{R}^d \quad \text{such that} \quad v(w^\star) = 0. \tag{1}$$

Related work. The above game formulation is strictly more general than minimization, and includes many problems such as saddle-point problems [25], bilinear games [18, 20], and variational inequality problems [9, 29]. For minimization, the vector field v and the Jacobian ∇v are, respectively, the gradient and the Hessian of the objective function. The Hessian is a symmetric matrix, so its eigenvalues lie on the real line; the ratio of the largest and the smallest eigenvalues is the condition number. It is now a common wisdom in numerical linear algebra and optimization theory how the notion of condition number characterizes the difficulty of a (strongly convex) minimization problem [23]. In conjunction with provable lower bounds within various function classes [22], an algorithm, whose attainable upper bound on the convergence rate matches the lower bound of its function class, is considered "opimal" for that function class [23, 24].

Differentiable games, on the other hand, have much more convoluted dynamics, as the eigenvalues of the Jacobian at the solution are distributed over the complex plane [4, 19]. To address this, recent line of work proposes new algorithms with better upper bounds on the convergence rate [8, 10, 11, 19–21, 26]. Vast majority of the aforementioned literature proves convergence based on two approaches: either (i) bounding the real part of the eigenvalues of the (submatrices of) Jacobian, or (ii) the magnitude. However, such approaches do not fully exploit the distribution of eigenvalues of the Jacobian, resulting in loose bounds [3]. Simply put, the notion of lower bounds, and subsequently that of optimal algorithms, is not well established for optimizing games.

To understand the difficulty of a differentiable game, [3] proposed a geometric interpretation of the conditioning of a game via the notion of *spectral shape*, defined as the set containing all eigenvalues of the Jacobian. Specifically, let $\mathcal{M}_\mathcal{K}$ be the set of matrices A whose spectrum, denoted by $\text{Sp}(A)$, belong to a set $\mathcal{K}$ on the complex plane with positive real part:

$$\mathcal{M}_\mathcal{K} := \{A \in \mathbb{R}^{d\times d} : \text{Sp}(A) \subset \mathcal{K} \subset \mathbb{C}_+\}.$$

When $\mathcal{K}$ is "simple" (e.g., a real line segment or a disc on the complex plane), [3] characterizes the lower bound and the optimality of some first-order methods [1].

This work. We consider a more specific class of games, where the distribution of the eigenvalues of the Jacobian can be modeled via (shifted) *cross-shape* on the complex plane. That is, we consider the following spectrum model:

$$\text{Sp}(\nabla v(w)) \in \mathcal{S}^\star = [\mu, L] \cup \{a + bi \in \mathbb{C} : a = c' > 0,\ b \in [-c, c]\}. \tag{2}$$

In (2), the first set can be thought of as a segment on the real line, similarly to the case of minimizing μ-strongly convex and L-smooth functions. The second set has a fixed real component ($c' > 0$), along with imaginary components symmetric across the real line, as the Jacobian is real.

This is a strict generalization of the purely imaginary interval $[-bi, bi]$ commonly considered in the bilinear games literature [3, 18, 20]. While many recent papers on bilinear games cite generative adversarial networks (GANs) [12] as a motivation, the work in [5, Figure 4] empirically shows that the spectrum of GANs is not contained in the imaginary axis. Instead, we note that the cross shape spectrum assumption above would include some of the observed GAN spectra.

In this work, we focus on solving (1) for games with Jacobian eigenvalue structure in (2). We use the extragradient with momentum, which we expound in the next section.

Notation. We denote the spectrum of a matrix A by $\text{Sp}(A)$, and its spectral radius by $\rho(A) := \max\{|\lambda| : \lambda \in \text{Sp}(A)\}$. $\mathfrak{R}(z)$ and $\mathfrak{I}(z)$ respectively denote the real and the imaginary part of a complex number z.

2. Algorithm

Extragradient with momentum via Chebyshev polynomials. To solve (1) for games with Jacobian eigenvalue structure in (2), we focus on the the *extragradient method with momentum* (EGM):

$$\text{(EGM)}\quad w_{t+1} = w_t - hv(w_t - \gamma v(w_t)) + m(w_t - w_{t-1}), \tag{3}$$

where h is the step size, γ is the extrapolation step size, and m is the momentum parameter. The extragradient (EG) method (without momentum) was first proposed in [16] for saddle point problems, and recently got popularized due to its convergence behavior for some class of differentiable games such as bilinear games, for which the standard gradient method diverges [2, 3, 11]. For completeness, we remind the gradient method (GD) and the gradient method with momentum (GDM):

$$\text{(GD)}\quad w_{t+1} = w_t - hv(w_t), \quad \text{and} \quad \text{(GDM)}\quad w_{t+1} = w_t - hv(w_t) + m(w_t - w_{t-1}). \tag{4}$$

We take a polynomial based approach to analyze EGM. Such analysis was used in [3, 14, 15, 27]. On that end, we use the following lemma [3, 6], which connects first-order methods with polynomials, when the vector field satisfies $v(w) = Aw + b$:

Lemma 1 *There exists a real polynomial p_t of degree at most t satisfying*

$$w_t - w^\star = p_t(A)(w_0 - w^\star), \tag{5}$$

where $p_t(0) = 1$, and $v(w^\star) = Aw^\star + b = 0$.

By taking norms, (5) further implies $\|w_t - w^\star\|_2 = \|p_t(A)(w_0 - w^\star)\|_2$, from which a convergence rate can be obtained. As EGM in (3) is a first-order method [1, 3], we can study the associated residual polynomial. We summarize this in the following theorem:

Theorem 2 (Residual polynomials of EGM) *Consider the extragradient with momentum in (3). Its associated residual polynomial is:*

$$\tilde{P}_0(\lambda) = 1, \;\; \tilde{P}_1(\lambda) = 1 - \tfrac{h\lambda(1-\gamma\lambda)}{1+m}, \;\; \textit{and} \;\; \tilde{P}_{t+1}(\lambda) = (1 + m - h\lambda(1 - \gamma\lambda))\tilde{P}_t(\lambda) - m\tilde{P}_{t-1}(\lambda).$$

Further, let $T_t(\cdot)$ and $U_t(\cdot)$ be the Chebyshev polynomials of the first and the second kind respectively. Then, the above expression simplifies to:

$$P_t(\lambda) = m^{t/2}\left(\tfrac{2m}{1+m}T_t(\sigma(\lambda)) + \tfrac{1-m}{1+m}U_t(\sigma(\lambda))\right) \textit{ with } \sigma(\lambda) = \tfrac{1+m-h\lambda(1-\gamma\lambda)}{2\sqrt{m}}, \tag{6}$$

where we refer to the term $\sigma(\lambda)$ as the link function.

As can be seen in (6), the link function for EGM is *quadratic* with respect to λ, which is different from the link function of GDM. We write below the residual polynomial for GDM, expressed in terms of the Chebyshev polynomials:

$$P_t^{\text{GDM}}(\lambda) = m^{t/2}\left(\tfrac{2m}{1+m}T_t(\xi(\lambda)) + \tfrac{1-m}{1+m}U_t(\xi(\lambda))\right) \text{ with } \xi(\lambda) = \tfrac{1+m-h\lambda}{2\sqrt{m}}. \tag{7}$$

Notice that the Chebyshev polynomial-based expressions of EGD in (6) and that of GDM in (7) are identical except for the link functions $\sigma(\lambda)$ and $\xi(\lambda)$, which enter $T_t(\cdot)$ and $U_t(\cdot)$ as arguments. This difference in link functions is crucial, as the Chebyshev polynomials behave very differently based on the domain, as shown in the following lemma:

Lemma 3 ([13]) *Let z be a complex number. The sequence $\left(\left|\frac{2m}{1+m}T_t(z) + \frac{1-m}{1+m}U_t(z)\right|\right)_{t\geqslant 0}$ grows exponentially in t for $z \notin [-1, 1]$, while in that interval they are bounded as the following:*

$$|T_t(z)| \leqslant 1 \quad \textit{and} \quad |U_t(z)| \leqslant t+1. \tag{8}$$

Therefore, we are interested in the case where the set of step sizes and momentum parameters lead to $|\sigma(\lambda)| \leqslant 1$, so that we can use the bound in (8). We call the set of hyperparameters such that the image of the link function is in $[-1, 1]$ as the *robust region*.

Three modes of the extragradient with momentum. Within the robust region, we can compute the worst-case convergence rate as follows:

$$\begin{aligned} r_t := \max_{\lambda\in\mathcal{S}^\star} |P_t(\lambda)| &\leqslant m^{t/2}\left(\tfrac{2m}{1+m}\max_{\lambda\in\mathcal{S}^\star}|T_t(\sigma(\lambda))| + \tfrac{1-m}{1+m}\max_{\lambda\in\mathcal{S}^\star}|U_t(\sigma(\lambda))|\right) \\ &\overset{(8)}{\leqslant} m^{t/2}\left(\tfrac{2m}{1+m} + \tfrac{1-m}{1+m}(t+1)\right) \leqslant m^{t/2}(t+2), \end{aligned} \tag{9}$$

where the last inequality uses $0 < m < 1$. Note that by definition r_t denotes the worst-case convergence rate. To get the asymptotic rate, we take the limit supremum of the $2t$-th root[2] of r_t:

$$\limsup_{t\to\infty} \sqrt[2t]{r_t} = \limsup_{t\to\infty}\left(m^{t/2}(t+2)\right)^{\frac{1}{2t}} = \sqrt[4]{m}. \tag{10}$$

Recall that the Chebyshev polynomial expressions of EGD in (6) and that of GDM[3] are identical except for the link functions. Hence, the convergence rate in (9) applies to both EGM and GDM, as long as the link functions $|\sigma(\lambda)|$ and $|\xi(\lambda)|$ are bounded by 1.

Yet, due to the difference in link functions, the robust region of EGM and GDM are drastically different. Specifically, due to the *quadratic* link function of EGM in (6), the robust region becomes more flexible; in particular, this link function can still satisfy the upper bounds (8) in Lemma 3 for a subset of complex values, resulting in the convergence rate of (9). Hence, compared to GDM whose link function is linear in λ and so can only satisfy (8) for a subset of real values, EGM enjoys the convergence rate in (9) for a much wider set of λ values.

The robust region of EGM can be described with the four extreme points below:

$$\sigma^{-1}(-1) = \tfrac{1}{2\gamma} \pm \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1+\sqrt{m})^2}{h\gamma}} \quad \text{and} \quad \sigma^{-1}(1) = \tfrac{1}{2\gamma} \pm \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}}. \tag{11}$$

Depending on the choice of hyperparameters, the above four extreme points can be distributed differently on three different modes, which we summarize in the following theorem:

[2] The reason why we take the $2t$-th root of r_t is to normalize by the number of vector field computations.

[3] Asymptotically, GDM enjoys $\sqrt{m}$ convergence rate instead of the $\sqrt[4]{m}$ of EGM, as it uses a single vector field computation per iteration instead of the two.

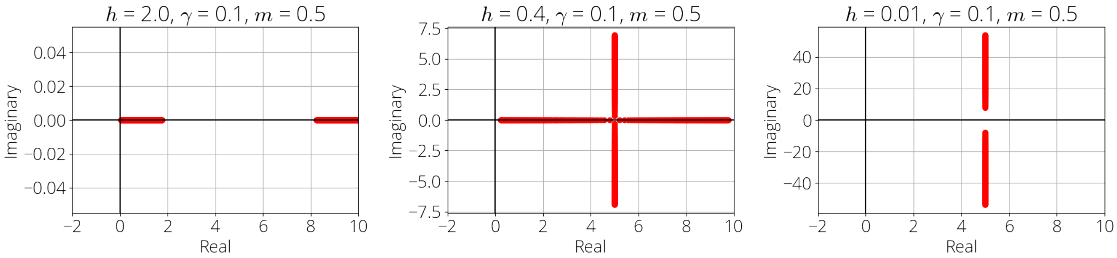

Figure 1: Blue points represent input arguments to the link function in (6) that outputs the range in $[-1, 1]$. **Left**: Case 1; **Middle**: Case 2; **Right**: Case 3. Red solid color is the real and the imaginary axes.

Theorem 4 *Consider the extragradient with momentum in* (3)*, expressed with the Chebyshev polynomials as in* (6)*. Then, the robust region summarized in* (11) *have the following three modes:*

1. *If* $\frac{h}{4\gamma} \geqslant (1+\sqrt{m})^2$*, then* $\sigma^{-1}(-1)$ *and* $\sigma^{-1}(1)$ *are all real numbers;*
2. *If* $(1-\sqrt{m})^2 \leqslant \frac{h}{4\gamma} < (1+\sqrt{m})^2$*, then* $\sigma^{-1}(-1)$ *are complex, and* $\sigma^{-1}(1)$ *are real;*
3. *If* $(1-\sqrt{m})^2 > \frac{h}{4\gamma}$*, then* $\sigma^{-1}(-1)$ *and* $\sigma^{-1}(1)$ *are all complex numbers.*

We empirically verify Theorem 4 in Figure 1. Note that the hyperparameters are set up so that each case in Theorem 4 is covered.

3. Optimal parameters and convergence rates

Optimal parameters for cross-shaped game problem. We focus on solving (1) for games with *cross-shaped* Jacobian eigenvalue structure as in (2). Therefore, in the remainder of the paper, we focus on the second case of Theorem 4, which is illustrated in the middle pannel of Figure 1.

We now connect the cross-shaped spectrum model in (2) and the robust region of EGM described with four extra points in (11). As we focus on the second case of Theorem 4, the hyperparameters satisfy the condition $(1-\sqrt{m})^2 \leqslant \frac{h}{4\gamma} < (1+\sqrt{m})^2$. In this case, we can write the robust region of EGM as follows:

$$\begin{aligned}\sigma^{-1}([-1,1]) = &\left[\tfrac{1}{2\gamma} - \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}},\ \tfrac{1}{2\gamma} + \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}}\right] \bigcup \\ &\left[\tfrac{1}{2\gamma} - \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1+\sqrt{m})^2}{h\gamma}},\ \tfrac{1}{2\gamma} + \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1+\sqrt{m})^2}{h\gamma}}\right].\end{aligned} \tag{12}$$

Here, the first interval lies on $\mathbb{R}$, as the square root term is real; conversely, in the second interval, the square root term is imaginary, with fixed real component: $\frac{1}{2\gamma}$.

The optimal parameters of EGM in terms of the worst-case convergence rate for this problem occurs when the robust region in (12) and the spectrum model in (2) coincide. This condition can be summarized to the following three equalities:

$$\tfrac{1}{2\gamma} - \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}} = \mu, \quad \tfrac{1}{2\gamma} + \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}} = L, \text{ and } \sqrt{\tfrac{(1+\sqrt{m})^2}{h\gamma} - \tfrac{1}{4\gamma^2}} = c. \tag{13}$$

Based on (13), we can compute the optimal hyperparameters, summarized in the next theorem:

Theorem 5 (Optimal hyperparameters for EGM) *Consider solving* (1) *for games where the Jacobian has cross-shaped spectrum as in* (2)*. For this problem, the optimal hyperparameters for the extragradient with momentum in* (3) *can be set as follows:*

$$h = \frac{16(\mu+L)}{(\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L})^2}, \quad \gamma = \frac{1}{\mu+L}, \quad \textit{and} \quad m = \left(\frac{\sqrt{4c^2+(\mu+L)^2}-\sqrt{4\mu L}}{\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L}}\right)^2.$$

Recalling (10), we immediately get the asymptotic convergence rate from Theorem 5. Further, this formula can be simplified the ill-conditioned regime, where $\tau := \frac{\mu}{L}$ approaches 0:

$$\sqrt[4]{m} = \left(\frac{\sqrt{4c^2+(\mu+L)^2}-\sqrt{4\mu L}}{\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L}}\right)^{\frac{1}{2}} = \left(\frac{\sqrt{(2c/L)^2+(1+\tau)^2}-2\sqrt{\tau}}{\sqrt{(2c/L)^2+(1+\tau)^2}+2\sqrt{\tau}}\right)^{\frac{1}{2}} \underset{\tau\to 0}{=} 1 - \frac{2\sqrt{\tau}}{\sqrt{(2c/L)^2+1}} + o(\sqrt{\tau}). \tag{14}$$

We see that EGM achieves accelerated convergence rate: $1 - O(\sqrt{\mu/L})$ as long as $c = O(L)$.

A special case of the spectrum model in (2) is when the length of the first set and the second equals; i.e., when $c = \frac{L-\mu}{2}$. In this setting, the optimal hyperparameters of EGM in Theorem 5 further simplify, as we summarize in the next corollary:

Corollary 6 *Consider the same setting as Theorem 5. Further, assume that* $c = \frac{L-\mu}{2}$*. Then, the optimal hyperparameters for the extragradient with momentum in* (3) *can be set as follows:*

$$h = \frac{8(\mu+L)}{(\sqrt{\mu^2+L^2}+\sqrt{2\mu L})^2}, \quad \gamma = \frac{1}{\mu+L}, \quad \textit{and} \quad m = \left(\frac{\sqrt{\mu^2+L^2}-\sqrt{2\mu L}}{\sqrt{\mu^2+L^2}+\sqrt{2\mu L}}\right)^2.$$

Again, from Corollary 6, we can immediately get the asymptotic convergence rate via (10), which can be interpreted further in the ill-conditioned regime where $\tau \to 0$:

$$\sqrt[4]{m} = \left(\frac{\sqrt{\mu^2+L^2}-\sqrt{2\mu L}}{\sqrt{\mu^2+L^2}+\sqrt{2\mu L}}\right)^{\frac{1}{2}} \underset{\tau\to 0}{=} 1 - \sqrt{2}\sqrt{\tau} + o(\sqrt{\tau}) \approx 1 - \sqrt{2}\sqrt{\frac{\mu}{L}}. \tag{15}$$

Comparison with other methods. In [2], GD and EG[4] are interpreted as the fixed-point iterations of operators F_h^{GD} and F_h^{EG} respectively representing each method, and the local convergence rate is obtained by bounding the spectral radius of the Jacobian of the operators under certain assumptions. We make the comparison for the simplified setting when $c = \frac{L-\mu}{2}$ in (2); we leave the general case with any c for future work. We summarize the relevant theorems from [2, 11] below:

Theorem 7 ([2, 11]) *Let* $w^\star$ *be a stationary point of* v*. Denote by* $\sigma^\star$ *the spectrum of* $\nabla v(w^\star)$ *for notational brevity. Further, assume the eigenvalues of* $\nabla v(w^\star)$ *all have positive real parts. Then,*

1. *For the gradient method with step size* $h = \min_{\lambda\in\sigma^\star}$*,*

$$\textit{(GD)} \quad \rho(\nabla F_h^{GD}(w^\star))^2 \leqslant 1 - \min_{\lambda\in\sigma^\star} \Re(1/\lambda) \min_{\lambda\in\sigma^\star} \Re(\lambda). \tag{16}$$

2. *For the extragradient method with step size* $h = (4\max_{\lambda\in\sigma^\star}|\lambda|)^{-1}$*, it satisfies:*

$$\textit{(EG)} \quad \rho(\nabla F_h^{EG}(w^\star))^2 \leqslant 1 - \frac{1}{4}\left(\frac{\min_{\lambda\in\sigma^\star}\Re(\lambda)}{\max_{\lambda\in\sigma^\star}|\lambda|} + \frac{\min_{\lambda\in\sigma^\star}|\lambda|^2}{\max_{\lambda\in\sigma^\star}|\lambda|^2}\right). \tag{17}$$

Since our spectrum model in (2) satisfies the condition that the eigenvalues of $\nabla v(w^\star)$ all have positive real parts, we can obtain the convergence rate of GD and EG, based on Theorem 7. With our (simplified) spectrum model with $c = \frac{L-\mu}{2}$, it follows that $\min_{\lambda\in\sigma^\star}\Re(\lambda) = \mu$, and $\min_{\lambda\in\sigma^\star}\Re(1/\lambda) =$

[4] In this work, it is assumed that same step size h is used for both the main and the extrapolation step sizes.

$1/L$. Similarly, for EG, it follows that $\min_{\lambda\in\sigma^\star}\Re(\lambda)=\mu$, $\max_{\lambda\in\sigma^\star}|\lambda|=L$, and $\max_{\lambda\in\sigma^\star}|\lambda|^2=L^2$. Furthermore, $\min_{\lambda\in\sigma^\star}|\lambda|^2=\mu^2$ if $L\geqslant(\sqrt{2}+1)\mu$, and $\min_{\lambda\in\sigma^\star}|\lambda|^2=(L-\mu)^2/2$ otherwise. Thus, the convergence rates of GD and EG in (16) and (17) respectively satisfy:

$$\rho(\nabla F_h^{\mathrm{GD}}(w^\star))^2\leqslant 1-\tfrac{\mu}{L},\ \text{ and }\ \rho(\nabla F_h^{\mathrm{EG}}(w^\star))^2\leqslant\begin{cases}1-\frac{1}{4}\left(\frac{\mu}{L}+\frac{\mu^2}{16L^2}\right) & \text{if } L\geqslant(\sqrt{2}+1)\mu\\ 1-\frac{1}{4}\left(\frac{\mu}{L}+\frac{(L-\mu)^2}{16L^2}\right) & \text{otherwise.}\end{cases}$$

We see that both GD and EG have non-accelerated convergence rate, unlike (14) and (15).

4. Conclusion

In this work, we focused on finding a stationary point of games. Via polynomial-based analysis, we showed the extragradient with momentum converges in three different modes of eigenvalue structure of the game Jacobian, depending on the hyperparameters. Then, we focused on a more specific problem class, where the eigenvalues of the game Jacobian are distributed in the complex plane with a cross-shape. For this problem class, we obtained the optimal hyperparameters for the momentum extragradient method, which enjoys an accelerated linear asymptotic convergence rate.

Acknowledgements

This work is supported by NSF FET no. 1907936, NSF/Intel MLWiNS no. 2003137, NSF CMMI no. 2037545, and NSF CAREER no. 2145629.

References

[1] Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In *International Conference on Machine Learning*, pages 908–916. PMLR, 2016.

[2] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873. PMLR, 2020.

[3] Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *International Conference on Artificial Intelligence and Statistics*, pages 1705–1715. PMLR, 2020.

[4] 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 354–363. PMLR, 2018.

[5] Hugo Berard, Gauthier Gidel, Amjad Almahairi, Pascal Vincent, and Simon Lacoste-Julien. A closer look at the optimization landscapes of generative adversarial networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJeVnCEKwH.

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

[7] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[8] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[9] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

[10] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[11] 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*, pages 1802–1811. PMLR, 2019.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[13] Baptiste Goujaud and Fabian Pedregosa. Cyclical step-sizes, 2022. URL http://fabianp.net/blog/2022/cyclical/.

[14] Baptiste Goujaud, Damien Scieur, Aymeric Dieuleveut, Adrien B Taylor, and Fabian Pedregosa. Super-acceleration with cyclical step-sizes. In *International Conference on Artificial Intelligence and Statistics*, pages 3028–3065. PMLR, 2022.

[15] Magnus R Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving. *Journal of research of the National Bureau of Standards*, 49(6):409, 1952.

[16] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[17] Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *The Journal of Machine Learning Research*, 20(1):3032–3071, 2019.

[18] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.

[19] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. *Advances in neural information processing systems*, 30, 2017.

[20] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[21] 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*, 15(1):229–251, 2004.

[22] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[23] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[24] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $o(k^2)$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[25] Roy A Nicolaides. Existence, uniqueness and approximation for generalized saddle point problems. *SIAM Journal on Numerical Analysis*, 19(2):349–357, 1982.

[26] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

[27] Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7553–7562. PMLR, November 2020. URL https://proceedings.mlr.press/v119/pedregosa20a.html. ISSN: 2640-3498.

[28] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[29] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

Appendix A. Proofs

A.1. Proof for Theorem 2

DERIVATION OF THE FIRST PART

Proof We want to find the residual polynomial $\tilde{P}_t(A)$ of the extragradient with momentum (EGM) in (3). That is, we want to find

$$w_t - w^\star = \tilde{P}_t(A)(w_0 - w^\star), \tag{18}$$

where $\{w_t\}_{t=0}$ is the iterates generated by EGM, which is possible by Lemma 1, as EGM is a first-order method [3]. We now prove this is by induction.
Base case. For $t = 0$, $\tilde{P}_0(A)$ is a degree-zero polynomial, and hence equals $\mathbb{I}$, which denotes the identity matrix. Thus, $w_0 - w^\star = \mathbb{I}(w_0 - w^\star)$ holds true.
Induction step. As the induction hypothesis, assume $\tilde{P}_t$ satisfies (18). Also, since we are looking for a stationary point, it holds that $v(w^\star) = 0$. Further, as v is linear by the assumption of Lemma 1, it holds that $v(w) = A(w - w^\star)$. Applying this to EGM, we have:

$$\begin{aligned}
w_{t+1} &= w_t - hv(w_t - \gamma v(w_t)) + m(w_t - w_{t-1}) \\
&= w_t - hv(w_t - \gamma A(w_t - w^\star)) + m(w_t - w_{t-1}) \\
&= w_t - hA(w_t - \gamma A(w_t - w^\star) - w^\star) + m(w_t - w_{t-1}) \\
&= w_t - hA(w_t - w^\star) + h\gamma A^2(w_t - w^\star) + m(w_t - w_{t-1}) \\
&= w_t - hA(\mathbb{I} - \gamma A)(w_t - w^\star) + m(w_t - w_{t-1}).
\end{aligned}$$

Subtracting $w^\star$ on both sides, we have:

$$\begin{aligned}
w_{t+1} - w^\star &= w_t - w^\star - hA(\mathbb{I} - \gamma A)(w_t - w^\star) + m(w_t - w_{t-1}) \\
&= (\mathbb{I} - hA(\mathbb{I} - \gamma A))\,(w_t - w^\star) + m(w_t - w^\star - (w_{t-1} - w^\star)) \\
&= (\mathbb{I} - hA(\mathbb{I} - \gamma A))\,\tilde{P}_t(A)(w_0 - w^\star) + m(\tilde{P}_t(A)(w_0 - w^\star) - \tilde{P}_{t-1}(A)(w_0 - w^\star)) \\
&= (\mathbb{I} + m\mathbb{I} - hA(\mathbb{I} - \gamma A))\,\tilde{P}_t(A)(w_0 - w^\star) - m\tilde{P}_{t-1}(A)(w_0 - w^\star) \\
&= \tilde{P}_{t+1}(A)(w_0 - w^\star),
\end{aligned}$$

where in the third equality, we used the induction hypothesis in (18).

■

DERIVATION OF THE SECOND PART IN (6)

Proof We show $P_t = \tilde{P}_t$ for all t via induction.

Base case.

$$\begin{aligned}
P_1(\lambda) &= m^{t/2}\left(\frac{2m}{1+m}T_1(\sigma(\lambda)) + \frac{1-m}{1+m}U_1(\sigma(\lambda))\right)\\
&= m^{t/2}\left(\frac{2m}{1+m}\,\sigma(\lambda) + \frac{1-m}{1+m}\cdot 2\cdot\sigma(\lambda)\right)\\
&= m^{t/2}\left(\frac{2\sigma(\lambda)}{1+m}\right)\\
&= 1 - \frac{h\lambda(1-\gamma\lambda)}{1+m} = \tilde{P}_1(\lambda).
\end{aligned}$$

Induction step. As the induction hypothesis, assume that $P_t = \tilde{P}_t$ for t. We want to show this holds for $t+1$.

$$\begin{aligned}
P_{t+1} &= m^{(t+1)/2}\left[\frac{2m}{1+m}T_{t+1}(\sigma(\lambda)) + \frac{1-m}{1+m}U_{t+1}(\sigma(\lambda))\right]\\
&= m^{(t+1)/2}\left[\frac{2m}{1+m}\Big(2\sigma(\lambda)T_t(\sigma(\lambda)) - T_{t-1}(\sigma(\lambda))\Big)\right.\\
&\qquad\left. + \frac{1-m}{1+m}\Big(2\sigma(\lambda)U_t(\sigma(\lambda) - U_{t-1}(\sigma(\lambda)))\Big)\right]\\
&= 2\sigma(\lambda)\cdot m^{1/2}\cdot \underbrace{m^{t/2}\left(\frac{2m}{1+m}T_t(\sigma(\lambda)) + \frac{1-m}{1+m}U_t(\sigma(\lambda))\right)}_{P_t(\lambda)}\\
&\qquad - m\cdot \underbrace{m^{(t-1)/2}\left(\frac{2m}{1+m}T_{t-1}(\sigma(\lambda)) + \frac{1-m}{1+m}U_{t-1}(\sigma(\lambda))\right)}_{P_{t-1}(\lambda)}\\
&= 2\sigma(\lambda)\cdot\sqrt{m}\cdot\tilde{P}_t(\lambda) - m\cdot\tilde{P}_{t-1}(\lambda)\\
&= (1+m-h\lambda(1-\gamma\lambda))\tilde{P}_t(\lambda) - m\tilde{P}_{t-1}(\lambda),
\end{aligned}$$

where in the second to last equality we use the induction hypothesis. ■

A.2. Proof of Lemma 3

Proof for Lemma 3 can be found in [13].

A.3. Proof of Theorem 4

Proof We analyze each case separately.

Case 1: There are two square roots: $\sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}}$ and $\sqrt{\frac{1}{4\gamma^2} - \frac{(1+\sqrt{m})^2}{h\gamma}}$. The second one is real if:

$$\frac{1}{4\gamma^2} \geqslant \frac{(1+\sqrt{m})^2}{h\gamma} \implies \frac{h\gamma}{4\gamma^2} = \frac{h}{4\gamma} \geqslant (1+\sqrt{m})^2,$$

which implies the first is real, as $(1+\sqrt{m})^2 \geqslant (1-\sqrt{m})^2$.

Case 3: There are two square roots: $\sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}}$ and $\sqrt{\frac{1}{4\gamma^2} - \frac{(1+\sqrt{m})^2}{h\gamma}}$. The first one is complex if:

$$\frac{1}{4\gamma^2} < \frac{(1-\sqrt{m})^2}{h\gamma} \implies \frac{h\gamma}{4\gamma^2} = \frac{h}{4\gamma} < (1-\sqrt{m})^2,$$

which implies the second is complex, as $(1+\sqrt{m})^2 \geqslant (1-\sqrt{m})^2$.

Case 2: This case follows automatically from the above two cases.

■

A.4. Proof of Theorem 5

Proof We recall the conditions in (13) below:

$$\frac{1}{2\gamma} - \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}} = \mu, \tag{19}$$

$$\frac{1}{2\gamma} + \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}} = L, \quad \text{and} \tag{20}$$

$$\sqrt{\frac{(1+\sqrt{m})^2}{h\gamma} - \frac{1}{4\gamma^2}} = c. \tag{21}$$

First, by adding (19) and (20), we get:

$$\frac{1}{\gamma} = \mu + L \implies \gamma = \frac{1}{\mu + L}. \tag{22}$$

Plugging (22) back into (19), we have:

$$\begin{aligned}
\frac{1}{2\gamma} - \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}} &= \mu \\
\frac{\mu+L}{2} - \mu &= \sqrt{\left(\frac{\mu+L}{2}\right)^2 - \frac{(1-\sqrt{m})^2(\mu+L)}{h}} \\
\left(\frac{L-\mu}{2}\right)^2 &= \left(\frac{\mu+L}{2}\right)^2 - \frac{(1-\sqrt{m})^2(\mu+L)}{h} \\
\frac{(1-\sqrt{m})^2(\mu+L)}{h} &= \left(\frac{\mu+L}{2}\right)^2 - \left(\frac{L-\mu}{2}\right)^2 = \mu L \\
h &= \frac{(1-\sqrt{m})^2(\mu+L)}{\mu L}. \qquad (23)
\end{aligned}$$

Plugging (22) and (23) into (21), we have:

$$\begin{aligned}
\sqrt{\frac{(1+\sqrt{m})^2}{h\gamma} - \frac{1}{4\gamma^2}} &= c \\
\sqrt{\frac{(1+\sqrt{m})^2 \cdot \mu L}{(1-\sqrt{m})^2} - \left(\frac{\mu+L}{2}\right)^2} &= c \\
\frac{(1+\sqrt{m})^2 \cdot \mu L}{(1-\sqrt{m})^2} &= c^2 + \left(\frac{\mu+L}{2}\right)^2 = \frac{4c^2+(\mu+L)^2}{4} \\
\frac{(1+\sqrt{m})^2}{(1-\sqrt{m})^2} &= \frac{4c^2+(\mu+L)^2}{4\mu L} \\
(1+\sqrt{m})\sqrt{4\mu L} &= (1-\sqrt{m})\sqrt{4c^2+(\mu+L)^2} \\
\sqrt{m}(\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L}) &= \sqrt{4c^2+(\mu+L)^2}-\sqrt{4\mu L} \\
\sqrt{m} &= \frac{\sqrt{4c^2+(\mu+L)^2}-\sqrt{4\mu L}}{\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L}}. \qquad (24)
\end{aligned}$$

Finally, to simplify (23) further, from (24), we have:

$$1-\sqrt{m} = \frac{4\sqrt{\mu L}}{\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L}}.$$

Hence, from (23),

$$h = \frac{(\mu+L)(1-\sqrt{m})^2}{\mu L} = \frac{\frac{16\mu L(\mu+L)}{(\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L})^2}}{\mu L} = \frac{16(\mu+L)}{(\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L})^2}. \qquad (25)$$

■

A.5. Proof of Corollary 6

Proof This result can be obtained either by plugging in $c = \frac{L-\mu}{2}$ to (24) and (25), or by using the same technique as the proof of Theorem 5. ■