When is Momentum Extragradient Optimal? A Polynomial-Based Analysis

Anonymous authors
Paper under double-blind review

Abstract

The extragradient method has recently gained increasing attention, due to its convergence behavior on smooth games. There, the eigenvalues of the Jacobian of the game vector field are distributed on the complex plane. Thus, compared to single objective minimization, games exhibit more convoluted dynamics, where –even for simple bilinear games– the extragradient method converges, while simple gradient method diverges. In this work, we study via polynomial-based analysis the momentum extragradient method, which combines appealing components of extragradient for convergence in smooth games, and momentum for accelerated convergence rates. Specifically, we show that the momentum extragradient 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. We derive the optimal hyperparameters for each case, and show that momentum extragradient achieves accelerated convergence rates.

1 Introduction

While most machine learning models are formulated as the minimization of a single loss involving data, a growing number of works adopt game formulations that involve multiple players and objectives. For example, generative adversarial networks (GANs) (Goodfellow et al., 2014), actor-critic models (Pfau & Vinyals, 2016), and sharpness aware minimization (Foret et al., 2021), can all be formulated as two-player games. Naturally, there is an increasing interest in the theoretical understanding of optimizing such differentiable games.

Due to the interaction of multiple players and objectives, there are issues in optimizing games that are not present in minimizing a single objective. Most notably, the eigenvalues of the game Jacobian are distributed on the *complex plane*, and thus exhibit much more convoluted dynamics, unlike single-objective minimization, where the eigenvalues of the Hessian lie on the *real line*. As a result, even for simple bilinear games, standard algorithms –like the gradient method– fail to converge due to the presence of rotational dynamics (Mescheder et al., 2018; Balduzzi et al., 2018; Gidel et al., 2019; Berard et al., 2020).

On the contrary, the extragradient method (EG), originally introduced by Korpelevich (1976) for saddle-point problems, converge for bilinear games (Tseng, 1995). As such, many recent works theoretically analyze EG from different perspectives, as in variational inequality (Gidel et al., 2018; Gorbunov et al., 2022), stochastic (Mishchenko et al., 2020; Li et al., 2021), and distributed (Liu et al., 2020; Beznosikov et al., 2021) settings.

Most existing works, including some of the aforementioned ones, analyze EG and relevant algorithms by assuming some structure on the objectives, such as (strong) monotonicity or Lipschitzness (Solodov & Svaiter, 1999; Tseng, 1995; Daskalakis & Panageas, 2018; Ryu et al., 2019; Azizian et al., 2020a). Such assumptions, in the context of differentiable games, confine the distribution of the eigenvalues of the game Jacobian; for instance, strong monotonicity implies a lower bound on the real part of the eigenvalues, and the Lipschitz assumption implies an upper bound on the magnitude of the eigenvalues of the Jacobian.

Azizian et al. (2020b) showed that, however, such assumption may be too crude to study the game dynamics, which is captured by the eigenvalue *distribution* on the complex plane. Indeed, they show, via polynomial-based analysis, that some first-order methods can sometimes attain faster rates using momentum, by replacing smoothness and monotonicity assumptions into more precise assumptions on the distribution of eigenvalues of the Jacobian, captured by simple shapes like ellipses or line segments.

In this work, we take a similar path, but ask the question in reverse order: for what shape of Jacobian eigenvalue distribution does the momentum extragradient (MEG) method achieve optimal performance? We focus on the MEG method, since it combines compelling components: e.g., extragradient steps improve upon simple gradient descent in bilinear games, while momentum motions is often an (almost) no-cost modification that leads to accelerated rates in minimization (Polyak, 1987). Such "reverse" analysis enables us to study the behavior of MEG in specific settings depending on the hyperparameter setup, encompassing minimization (where the eigenvalues of the Jacobian are all on the real line), regularized bilinear game (where the eigenvalues are all complex conjugates), and the intermediate case (where the eigenvalues are both on the real line and as complex conjugates), as illustrated in Figure 1. Our contributions are:

- We derive the residual polynomials of MEG, when the vector field of the game is an affine function. We then identify three different modes of convergence for MEG depending on the hyperparameter setup. Such analysis can then be applied to different eigenvalue structures of the (game) Jacobian (c.f., Theorem 3).
- For each case of the eigenvalue structure, we derive the optimal hyperparameters of MEG, as well as its (asymptotic) convergence rates. For minimization, MEG exhibits "super-acceleration," where a constant improvement upon classical lower bound rate is attained, similarly to the momentum gradient method (GDM) with cyclical step sizes (Goujaud et al., 2022). For the other two cases (involving imaginary eigenvalues), MEG exhibits accelerated convergence rates.
- For the case where the eigenvalues of the game Jacobian have a cross-shaped structure, we derive the (asymptotic) accelerated convergence rate of MEG, as well as its optimal hyperparameters (c.f., Theorem 5). This case might be closer to some observations made in Berard et al. (2020, Figure 4), where the authors empirically show that the spectrum of GANs is not contained in the imaginary axis.
- Finally, we compare the convergence rates with other first order methods including the gradient (GD), the momentum gradient (GDM), and the extragradient (EG) methods. For the classes of games considered, none of the aforementioned methods (asymptotically) converge at an accelerated rate (c.f., Corollaries 1 and 2), in contrast to MEG. We confirm the theoretical findings with numerical experiments, including the case that slightly violates our assumption, in Section 6.

2 Problem Setup and Related Work

Following Letcher et al. (2019); Balduzzi et al. (2018), we define the n-player differentiable game as a family of twice continuously differentiable losses $\ell_i : \mathbb{R}^d \to \mathbb{R}$, 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: $w^{(i)^\star} \in \arg\min_{w^{(i)} \in \mathbb{R}^{d_i}} \ell_i\big(w^{(i)}, w^{(\neg i)^\star}\big) \quad \forall i \in \{1, \ldots, n\}$, where the notation $\cdot^{(\neg i)}$ denotes all indices except for i. We also define the vector field v of the game as the concatenation of the individual gradients: $v(w) = [\nabla_{w^{(1)}} \ell_1(w) \cdots \nabla_{w^{(n)}} \ell_n(w)]^\top$, and denote its associated Jacobian with ∇v.

Unfortunately, finding Nash equilibria for general games is impractical to solve (Shoham & Leyton-Brown, 2008; Letcher et al., 2019).[1] Instead, we focus on finding a stationary point of the vector field v, since a Nash equilibrium is always a stationary point of the gradient dynamics. That is, we want to solve:

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

Notation. We denote the spectrum of a matrix M by $\mathrm{Sp}(M)$, and its spectral radius by $\rho(M) := \max\{|\lambda| : \lambda \in \mathrm{Sp}(M)\}$. $M \succ 0$ denotes that M is a positive-definite matrix. $\Re(z)$ and $\Im(z)$ respectively denote the real and the imaginary part of a complex number z.

---

[1] Finding Nash equilibira can be refomulated as a nonlinear complementarity problem, which is PPAD hard (Daskalakis et al., 2009; Letcher et al., 2019).

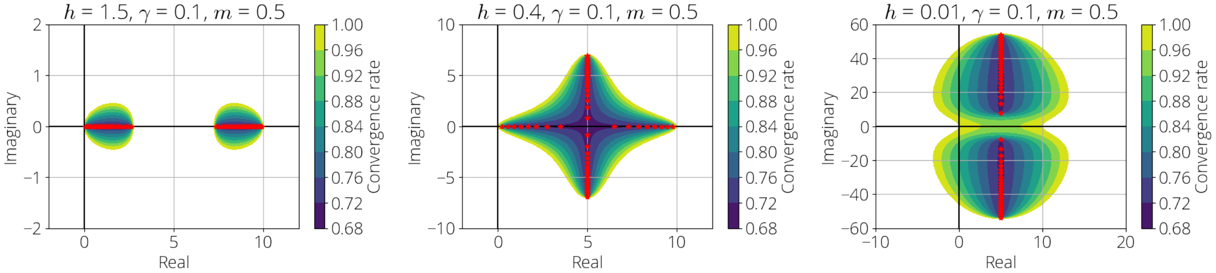

Figure 1: Red stars (⋆) represent the robust region of MEG, where it enjoys accelerated convergence rate. Convergence region is measured by $\sqrt[2t]{|P_t(\lambda)|} < 1$ for $t = 2000$, with convergence rates encoded in different colors. h, γ, m are set up based on each case of Theorem 3. **Left**: Case 1; **Middle**: Case 2; **Right**: Case 3.

2.1 Related Work

Extragradient method, originally introduced in Korpelevich (1976), is a standard algorithm for solving (unconstrained) variational inequality problems in (1) (Gidel et al., 2018). There are several works that study the convergence rate of EG for (strongly) monotone problems (Tseng, 1995; Solodov & Svaiter, 1999; Nemirovski, 2004; Monteiro & Svaiter, 2010; Mokhtari et al., 2020; Gorbunov et al., 2022). Under similar settings, stochastic variants of EG are studied in (Palaniappan & Bach, 2016; Hsieh et al., 2019; Mishchenko et al., 2020; Hsieh et al., 2020; Li et al., 2021). However, as mentioned in the introduction, assumptions like (strong) monotonicity or Lipchtizness might be too crude to capture the interaction of multiple players.

Instead, we make more fine-grained assumptions on these eigenvalues, to obtain the optimal hyperparameters and convergence rates of MEG via polynomial-based analysis. Such analysis dates back to the development of the conjugate gradient method (Hestenes & Stiefel, 1952), and is still actively used; for instance, to derive lower bounds (Arjevani & Shamir, 2016), to develop accelerated decentralized algorithms (Berthier et al., 2020), and to analyze average-case performance (Pedregosa & Scieur, 2020; Domingo-Enrich et al., 2021).

On that end, we use the following lemma Chihara (2011), which elucidates the connection between first-order methods[2] and (residual) polynomials, when the vector field v is affine:

Lemma 1 (Chihara (2011))**.** *Suppose $v(w) = Aw + b$. Then, there exists a real polynomial p_t, of degree at most t, satisfying:*

$$w_t - w^\star = p_t(A)(w_0 - w^\star)\,, \tag{2}$$

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

By taking ℓ_2-norms, (2) further implies the following worst-case convergence rate:

$$\|w_t - w^\star\| = \|p_t(A)(w_0 - w^\star)\| \leqslant \|p_t(Z\Lambda Z^{-1})\| \cdot \|w_0 - w^\star\| \leqslant \mathcal{O}\Big(\sup_{\lambda\in\mathcal{S}^\star} |p_t(\lambda)| \cdot \|w_0 - w^\star\|\Big), \tag{3}$$

where $A = Z\Lambda Z^{-1}$ is the diagonalization of A,[3] and $\mathcal{O}(\cdot)$ hides the constant $\kappa(Z) := \|Z\|\|Z^{-1}\|$, which equals 1 if A is a normal matrix. Hence, the worst-case convergence rate of a first-order method can be studied by analyzing the associated residual polynomial, p_t, evaluated at the eigenvalues of the Jacobian $\nabla v = A$, distributed over the set $S^\star$.

Azizian et al. (2020b) took a similar approach, and 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 $\mathcal{A}_{\mathcal{S}^\star}$ be the set of matrices A whose spectrum, denoted by $\mathrm{Sp}(A)$, belong to a set $\mathcal{S}^\star$ on the complex plane with positive real part:

$$\mathcal{A}_{\mathcal{S}^\star} := \{A \in \mathbb{R}^{d\times d} : \mathrm{Sp}(A) \subset \mathcal{S}^\star \subset \mathbb{C}_+\}.$$

---

[2] We consider first-order methods as those that only use first-order information in each iteration.

[3] Note that almost all matrices are diagonalizable over $\mathbb{C}$, in the sense that the set of non-diagonalizable matrices have Lebesgue measure zero (Hetzel et al., 2007).

When $\mathcal{S}^\star$ is "simple" (e.g., a real line segment or a disc on the complex plane), Azizian et al. (2020b) characterized the lower bound and the optimality of some first-order methods (Nemirovskij & Yudin, 1983; Arjevani & Shamir, 2016); see also Section 5.

Yet, any additional a priori knowledge about the exact shape of $\mathcal{S}^\star$ could potentially lead to better/more fine-grained analysis. We follow this perspective: based on the (residual) polynomials of MEG, specific spectral shapes emerge within the set $\mathcal{S}^\star$, where MEG shows the optimal performance. Similar research efforts have recently appeared in the literature: e.g., Oymak (2021); Goujaud et al. (2022) focus on the case of convex smooth minimization, and show that a priori knowledge of more information than just the largest (smoothness) and smallest (strong convexity) eigenvalues of the Jacobian lead to better convergence rate.

3 Momentum Extragradient via Chebyshev Polynomials

In this section, we study the *momentum extragradient method* (MEG) and its performance, through the lens of (residual) polynomials. MEG iterates as follows:

$$\text{(MEG)} \quad w_{t+1} = w_t - hv(w_t - \gamma v(w_t)) + m(w_t - w_{t-1}). \tag{4}$$

Here, h is the step size, γ is the extrapolation step size, and m is the momentum parameter. The extragradient (EG) method (without momentum, i.e., $m = 0$) was first proposed by Korpelevich (1976) for saddle point problems, and recently received much attention due to its convergence behavior for some class of differentiable games, such as bilinear games, for which the standard gradient method diverges (Gidel et al., 2019; Azizian et al., 2020b;a). For completeness, we remind the momentum gradient (GDM) method:

$$\text{(GDM)} \quad w_{t+1} = w_t - hv(w_t) + m(w_t - w_{t-1}), \tag{5}$$

from which the gradient method (GD) can be obtained by setting $m = 0$.

As MEG in (4) is a first-order method (Arjevani & Shamir, 2016; Azizian et al., 2020b), we can study the associated residual polynomials, in order to apply Lemma 1.

Theorem 1 (Residual polynomials of MEG). *Consider the momentum extragradient method (MEG) in (4). If $v(w) = Aw + b$, the associated residual polynomials of MEG are:*

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

Further, let $T_t(\cdot)$ and $U_t(\cdot)$ be the Chebyshev polynomials of the first and the second kind, respectively. Then, the above expressions simplify 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 MEG is *quadratic* with respect to λ. GDM admits an almost identical expression, but with a different link function. For completeness, we write below the residual polynomials of GDM, expressed in terms of the Chebyshev polynomials (Pedregosa, 2020):

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

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

Lemma 2 (Goujaud & Pedregosa (2022)). *Let z be a complex number. The sequence $\left(\left|\tfrac{2m}{1+m}T_t(z) + \tfrac{1-m}{1+m}U_t(z)\right|\right)_{t\geqslant 0}$ grows exponentially in t for $z \notin [-1, 1]$, while in that interval, the following bounds hold:*

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

Therefore, to study the optimal convergence behavior of MEG, we are interested in the case where the set of step sizes and the momentum parameters lead to $|\sigma(\lambda)| \leqslant 1$, so that we can use the bounds in (8). We will refer to those set of hyperparameters as the *robust region*, as defined below:

Definition 1 (Robust region of MEG)**.** *Consider the MEG method in* (4) *expressed via Chebyshev polynomials, as in* (6)*. We define the set of hyperparameters such that the image of the link $\sigma(\lambda)$ function lies in the interval $[-1, 1]$ as the* ***robust region****, and denote it with $\sigma^{-1}([-1, 1])$.*

Despite the polynomial-based analysis necessitates the assumption that the vector field is affine, it yet captures intuitive insights of the behavior of different algorithms in different settings, as we remark below.

Remark 1. *From the definition of $\xi(\lambda)$ in* (7)*, one can infer why negative momentum can help the convergence of GDM (Gidel et al., 2019) when $\lambda \in \mathbb{R}_+$: it forces GDM to stay within the robust region, $|\xi(\lambda)| \leqslant 1$. One can also infer about the divergence of GDM in the presence of complex eigenvalues, unless, for instance, complex momentum is used (Lorraine et al., 2022). Similarly, the residual polynomial of GD is $P_t^{GD}(\lambda) = (1 - h\lambda)^t$ (Goujaud & Pedregosa, 2022, Example 4.2), and can easily diverge in the presence of complex eigenvalues, which can potentially be alleviated by using complex step sizes. On the contrary, thanks to the quadratic link function of MEG in* (6)*, it can converge for much wider subsets of complex eigenvalues.*

Via the connection of residual polynomials and first-order methods, we can also characterize the asymptotic convergence rate of MEG for any combination of hyperparameters, as summarized in the theorem below.

Theorem 2 (Asymptotic convergence rate of MEG)**.** *Suppose $v(w) = Aw + b$. The asymptotic convergence rate of MEG in* (4) *is:*[4]

$$\limsup_{t \to \infty} \sqrt[2t]{\frac{\|x_t - x^\star\|}{\|x_0 - x^\star\|}} = \begin{cases} \sqrt[4]{m}, & \text{if } \bar{\sigma} \leqslant 1 \quad \textit{(robust region)}; \\ \sqrt[4]{m}\big(\bar{\sigma} + \sqrt{\bar{\sigma}^2 - 1}\big)^{1/2}, & \text{if } \bar{\sigma} \in \Big(1, \frac{m+1}{2\sqrt{m}}\Big); \\ \geqslant 1 \textit{ (no convergence)}, & \textit{otherwise}, \end{cases} \tag{9}$$

where $\bar{\sigma} = \sup_{\lambda \in \mathcal{S}^\star} |\sigma(\lambda; h, \gamma, m)|$, and $\sigma(\lambda; h, \gamma, m) \equiv \sigma(\lambda)$ is the link function of MEG defined in (6)*.*

3.1 Three Modes of the Momentum Extragradient

Within the robust region of MEG, we can compute its worst-case convergence rate based on (3) as follows:

$$\begin{aligned} \sup_{\lambda \in \mathcal{S}^\star} |P_t(\lambda)| &\overset{(6)}{\leqslant} m^{t/2}\left(\tfrac{2m}{1+m} \sup_{\lambda \in \mathcal{S}^\star} |T_t(\sigma(\lambda))| + \tfrac{1-m}{1+m} \sup_{\lambda \in \mathcal{S}^\star} |U_t(\sigma(\lambda))|\right) \\ &\overset{(8)}{\leqslant} m^{t/2}\left(\tfrac{2m}{1+m} + \tfrac{1-m}{1+m}(t+1)\right) \leqslant m^{t/2}(t+1). \end{aligned} \tag{10}$$

Since the Chebyshev polynomial expressions of MEG in (6) and that of GDM[5] are identical except for the link functions, the convergence rate in (10) applies to both MEG and GDM, as long as the link functions $|\sigma(\lambda)|$ and $|\xi(\lambda)|$ are bounded by 1. As a result, we see that the asymptotic convergence rate in (9) only depends on the momentum parameter m, when the hyperparameters are restricted to the robust region. This fact was utilized in tuning GDM for strongly convex quadratic minimization (Zhang & Mitliagkas, 2019).

The robust region of MEG can be described with the four extreme points below (derivation in the appendix):

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

Depending on the choice of hyperparameters, the above four extreme points (and subsequently their intermediate values) can be distributed differently in three distinct modes, as in the next theorem.

---

[4] The reason why we take the $2t$-th root is to normalize by the number of vector field computations; we compare in Section 4 the asymptotic rate of MEG in (9) with other gradient methods that use a single vector field computation in the recurrences, such as GD and GDM.

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

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

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

Remark 2. *Theorem 3 provides crude information about how to set up the hyperparameters of the MEG, depending on the spectrum of the Jacobian of the game problem at hand; for instance, if one observes only real eigenvalues (i.e., the problem is in fact minimization), the main step size* h *should be at least* $4\times$ *larger than the extrapolation step size* γ*, based on the condition* $\frac{h}{4\gamma} \geqslant (1+\sqrt{m})^2$.

We illustrate Theorem 3 in Figure 1. To represent the robust region, we discretize the interval $[-1, 1]$, and plot $\sigma^{-1}([-1, 1])$ ($\star$ - red stars). Note that the hyperparameters are set up so that each case in Theorem 3 is covered. We see that the quadratic link function induced by the MEG allows convoluted eigenvalue dynamics, such as the cross-shape observed in Case 2, to be mapped onto the $[-1, 1]$ segment.

While MEG enjoys the fastest rate within the robust region, MEG does not necessarily diverge outside of it, as in the second case of Theorem 2. We illustrate the convergence region of MEG measured by $\sqrt[2t]{|P_t(\lambda)|} < 1$ from (6) for $t = 2000$, with convergence rates encoded in different colors (slows down as moving further away from the robust region). Moreover, Figure 1 (right) shows that MEG can also converge in the absence of monotonicity (i.e., in the presence of Jacobian eigenvalues with negative real part) (Gorbunov et al., 2023).

3.2 Robust Region-Induced Problem Cases

In this subsection, we summarize the problem classes illustrated with the robust region in Figure 1. In particular, for Case 1, the problem reduces to minimization, where optimal methods already exist. Nevertheless, we show that by using MEG, one can have constant improvement over the optimal rate of convergence, by exploiting further structure of the spectrum of the Jacobian (Hessian). For Cases 2 and 3 where imaginary eigenvalues present, we show that MEG achieves accelerated convergence rates.

Case 1: The problem reduces to minimization, where eigenvalues of the Jacobian are distributed on the real line, but as a *union* of two intervals. We can model such spectrum as:

$$\mathrm{Sp}(\nabla v) \in \mathcal{S}_1^\star = [\mu_1, L_1] \cup [\mu_2, L_2] \in \mathbb{R}. \tag{12}$$

The spectrum model in (12) generalizes the familiar Hessian spectrum for minimizing μ-strongly convex and L-smooth functions, i.e., $\lambda \in [\mu, L]$, which is recovered from (12) by setting $\mu_1 = \mu$, $L_2 = L$, and $L_1 = \mu_2$. Such a spectrum is empirically observed while training DNNs, where a small number of eigenvalues of the Hessian have much larger magnitude (Papyan, 2020). In such a case, Goujaud et al. (2022) showed that, for strongly convex quadratic objectives, GDM with alternating step sizes can have a (constant factor) improvement over the traditional lower bound.

In Theorem 4 and in (18), we show that MEG enjoys similar improvement. To show that, following Goujaud et al. (2022), we define the following quantities, which elucidate the behavior of MEG applied to the minimization problem with spectrum in (12):

$$\rho := \tfrac{L_2+\mu_1}{L_2-\mu_1} = \tfrac{1+\tau}{1-\tau}, \quad \text{and} \quad R := \tfrac{\mu_2-L_1}{L_2-\mu_1} \in [0, 1). \tag{13}$$

Here, ρ is the ratio between the center of $\mathcal{S}_1^\star$ and its radius, $\tau := L_2/\mu_1$ is the inverse condition number, and R is the relative gap $\mu_2 - L_1$ and $L_2 - \mu_1$ (which becomes 0 if $\mu_2 = L_1$).

Case 2: While Case 1 reduces to minimization, for Cases 2 and 3, there are imaginary eigenvalues present. In Case 2 particularly, the eigenvalues of the Jacobian are distributed both on the real line and as complex conjugates, exhibiting a *cross-shape* spectrum. We model this spectrum as:

$$\mathrm{Sp}(\nabla v) \in \mathcal{S}_2^\star = [\mu, L] \cup \{z \in \mathbb{C} : \Re(z) = c' > 0,\ \Im(z) \in [-c, c]\}. \tag{14}$$

The first set $[\mu, L]$ denotes a segment on the real line, reminiscent of the Hessian spectrum for 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 (i.e., complex conjugates), as the Jacobian is real.

This is a strict generalization of the purely imaginary interval $\pm[ai, bi]$ commonly considered in the bilinear games literature (Liang & Stokes, 2019; Azizian et al., 2020b; Mokhtari et al., 2020). While many recent papers on bilinear games cite GANs (Goodfellow et al., 2014) as a motivation, the work in (Berard et al., 2020, Figure 4) empirically shows that the spectrum of GANs is not contained in the imaginary axis; the cross-shaped spectrum model above might be closer to some of the observed GAN spectra.

Case 3: Lastly, in this case, the eigenvalues of the Jacobian are distributed only as complex conjugates, with a fixed real component (shifted imaginary spectrum). We model this spectrum as:

$$\mathrm{Sp}(\nabla v) \in \mathcal{S}_3^\star = [c + ai, c + bi] \cup [c - ai, c - bi] \in \mathbb{C}. \tag{15}$$

Again, (15) generalizes bilinear games, where the spectrum reduces to $\pm[ai, bi]$ with $c = 0$.

Examples of Cases 2 and 3 in quadratic games. To understand these spectra better, we provide examples using quadratic games. Consider the following two player quadratic game, where $x \in \mathbb{R}^{d_1}$ and $y \in \mathbb{R}^{d_2}$ are the parameters controlled by each player, whose loss functions respectively are:

$$\ell_1(x, y) = \frac{1}{2} x^\top S_1 x + x^\top M_{12} y + x^\top b_1 \quad \text{and} \quad \ell_2(x, y) = \frac{1}{2} y^\top S_2 y + y^\top M_{21} x + y^\top b_2, \tag{16}$$

where $S_1, S_2 \succ 0$. Then, the vector field can be written as:

$$v(x, y) = \begin{bmatrix} S_1 x + M_{12} y + b_1 \\ M_{21} x + S_2 y + b_2 \end{bmatrix} = Aw + b, \text{ where } A = \begin{bmatrix} S_1 & M_{12} \\ M_{21} & S_2 \end{bmatrix}, \ w = \begin{bmatrix} x \\ y \end{bmatrix}, \text{ and } b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{17}$$

If $S_1 = S_2 = 0$ and $M_{12} = -M_{21}^\top$, the game Jacobian $\nabla v = A$ has only purely imaginary eigenvalues (Azizian et al., 2020b, Lemma 7), recovering bilinear games.

As the spectrum models in (14) and (15) generalize bilinear games, we can consider more complex quadratic games, where S_1 and S_2 does not have to be 0. Specifically, when $M_{12} = -M_{21}^\top$, and they share common bases with S_1 and S_2 as in the below proposition, then $\mathrm{Sp}(A)$ has a cross-shaped spectrum in (14) of Case 2 and a shifted imaginary spectrum in (15) of Case 3.

Proposition 1. *Let A be a matrix of the form $\begin{bmatrix} S_1 & B \\ -B^\top & S_2 \end{bmatrix}$, where $S_1, S_2 \succ 0$. Without loss of generality, assume that $dim(S_1) > dim(S_2) = d$. Then,*

- ***Case 2:*** *$\mathrm{Sp}(A)$ has a cross-shape if there exist orthonormal matrices U, V and diagonal matrices D_1, D_2 such that $S_1 = U diag(a, \dots, a, D_1) U^\top$, $S_2 = V diag(a, \dots, a) V^\top$, and $B = U D_2 V^\top$.*
- ***Case 3:*** *$\mathrm{Sp}(A)$ is shifted-imaginary if there exist orthonormal matrices U, V and diagonal matrix D_2 such that $S_1 = U diag(a, \dots, a) U^\top$, $S_2 = V diag(a, \dots, a) V^\top$, and $B = U D_2 V^\top$.*

We can interpret Case 3 as a *regularized* bilinear game, where S_1 and S_2 are digonal matrices with a constant eigenvalue. This implies that the players cannot control their parameter x and y arbitrarily, which can be seen in the loss functions in (16), where S_1 and S_2 appears in terms $x^\top S_1 x$ and $y^\top S_2 y$. Case 2 can be interpreted similarly, but player 1 (without loss of generality) has more flexibility in its parameter choice through the inclusion of an additional diagonal matrix D_1 in the eigenvalue decomposition of S_1.

4 Optimal Parameters and Convergence Rates

In this section, we obtain the optimal hyperparameters of MEG, for each of the robust region-induced problem cases. On high-level, the optimal hyperparameters are the ones that minimize the asymptotic convergence rate of MEG in (9) of Theorem 2, while satisfying the constraint for each case in Theorem 3.

4.1 Case 1: minimization

If the condition in Case 1 of Theorem 3 is satisfied (i.e., $\frac{h}{4\gamma} \geqslant (1+\sqrt{m})^2$), then $\sigma^{-1}(-1)$ and $\sigma^{-1}(1)$ (and their intermediate values) are all on real line, forming a union of two intervals, as illustrated in Figure 1 (left). We can write the robust region $\sigma^{-1}_{\text{Case}_1}([-1,1])$ as:

$$\left[\tfrac{1}{2\gamma} - \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}}, \tfrac{1}{2\gamma} - \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1+\sqrt{m})^2}{h\gamma}}\right] \bigcup \left[\tfrac{1}{2\gamma} + \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1+\sqrt{m})^2}{h\gamma}}, \tfrac{1}{2\gamma} + \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}}\right] \in \mathbb{R}.$$

The optimal hyperparameters of MEG in terms of the worst-case asymptotic convergence rate in (9) occurs when the robust region above and the spectrum model in (12) coincide, and is summarized below.

Theorem 4 (Case 1)**.** *Consider solving* (1) *for games where the Jacobian has the spectrum in* (12)*. For this problem, the optimal hyperparameters for the momentum extragradient method in* (4) *are:*

$$h = \tfrac{4(\mu_1+L_2)}{(\sqrt{\mu_2+L_1}+\sqrt{\mu_1+L_2})^2}, \; \gamma = \tfrac{1}{\mu_1+L_2} = \tfrac{1}{\mu_2+L_1}, \; m = \left(\tfrac{\sqrt{\mu_2 L_1}-\sqrt{\mu_1 L_2}}{\sqrt{\mu_2 L_1}+\sqrt{\mu_1 L_2}}\right)^2 = \left(\tfrac{\sqrt{\rho^2-R^2}-\sqrt{\rho^2-1}}{\sqrt{\rho^2-R^2}+\sqrt{\rho^2-1}}\right)^2.$$

Recalling (9), we immediately get the asymptotic convergence rate from Theorem 4. Further, this formula can be simplified in the ill-conditioned regime, where the inverse condition number $\tau := \mu_1/L_2 \to 0$:

$$\sqrt[4]{m} = \left(\tfrac{\sqrt{\rho^2-R^2}-\sqrt{\rho^2-1}}{\sqrt{\rho^2-R^2}+\sqrt{\rho^2-1}}\right)^{1/2} \underset{\tau\to 0}{=} 1 - \tfrac{2\sqrt{\tau}}{\sqrt{1-R^2}} + o(\sqrt{\tau}). \tag{18}$$

From (18), we see that MEG achieves an accelerated convergence rate $1 - O(\sqrt{\tau})$, which was known to be "optimal" for this function class, and can be asymptotically achieved by GDM (Polyak, 1987)[6] (see also Theorem 8 with $\theta = 1$). Surprisingly, the rate can be further improved by the factor $\sqrt{1-R^2}$, exhibiting "super-acceleration" phenomenon enjoyed by GDM with (optimal) cyclical step sizes (Goujaud et al., 2022).

4.2 Case 2: cross-shaped spectrum

If the condition in Case 2 of Theorem 3 is satisfied (i.e., $(1-\sqrt{m})^2 \leqslant \frac{h}{4\gamma} < (1+\sqrt{m})^2$), then the points $\sigma^{-1}(-1)$ (and the intermediate values) are complex, while $\sigma^{-1}(1)$ are real, as illustrated in Figure 1 (middle). We can write the robust region $\sigma^{-1}_{\text{Case}_2}([-1,1])$ as:

$$\underbrace{\left[\tfrac{1}{2\gamma} - \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}}, \tfrac{1}{2\gamma} + \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1-\sqrt{m})^2}{h\gamma}}\right]}_{\in\mathbb{R}} \bigcup \underbrace{\left[\tfrac{1}{2\gamma} - \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1+\sqrt{m})^2}{h\gamma}}, \tfrac{1}{2\gamma} + \sqrt{\tfrac{1}{4\gamma^2} - \tfrac{(1+\sqrt{m})^2}{h\gamma}}\right]}_{\in\mathbb{C}}.$$

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 the fixed real component: $\frac{1}{2\gamma}$. We summarize the optimal hyperparameters for this case in the next theorem below.

Theorem 5 (Case 2)**.** *Consider solving* (1) *for games where the Jacobian has a cross-shaped spectrum as in* (14)*. For this problem, the optimal hyperparameters for the momentum extragradient method in* (4) *are:*

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

We get the asymptotic rate from Theorem 5, which simplify in the ill-conditioned regime $\tau := \mu/L \to 0$ as

$$\sqrt[4]{m} = \left(\tfrac{\sqrt{4c^2+(\mu+L)^2}-\sqrt{4\mu L}}{\sqrt{4c^2+(\mu+L)^2}+\sqrt{4\mu L}}\right)^{1/2} \underset{\tau\to 0}{=} 1 - \tfrac{2\sqrt{\tau}}{\sqrt{(2c/L)^2+1}} + o(\sqrt{\tau}). \tag{19}$$

---

[6] Precisely, GDM with optimal step size and momentum asymptotically achieve $1 - 2\sqrt{\tau} + o(\sqrt{\tau})$ convergence rate, as $\tau \to 0$ (Goujaud & Pedregosa, 2022, Proposition 3.3).

We see that MEG achieves accelerated convergence rate $1 - O(\sqrt{\mu/L})$, as long as $c = O(L)$. We remark that this rate is optimal in the following sense. The lower bound for the problems with cross-shaped spectrum in (14) must be slower than the existing ones for minimizing μ-strongly convex and L-smooth functions, as the former is strictly more general. Since we reach the same asymptotic optimal rate as for acceleration methods, this must be optimal.

4.3 Case 3: shifted imaginary spectrum

Lastly, if the condition in Case 3 of Theorem 3 is satisfied (i.e., $\frac{h}{4\gamma} < (1-\sqrt{m})^2$), $\sigma^{-1}(-1)$, $\sigma^{-1}(1)$ and the intermediate values are all complex numbers, as illustrated in Figure 1 (right). We can write the robust region $\sigma^{-1}_{\text{Case}_3}([-1,1])$ as:

$$\left[\frac{1}{2\gamma} + \sqrt{\frac{1}{4\gamma^2} - \frac{(1+\sqrt{m})^2}{h\gamma}}, \frac{1}{2\gamma} + \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}}\right] \bigcup \left[\frac{1}{2\gamma} - \sqrt{\frac{1}{4\gamma^2} - \frac{(1-\sqrt{m})^2}{h\gamma}}, \frac{1}{2\gamma} - \sqrt{\frac{1}{4\gamma^2} - \frac{(1+\sqrt{m})^2}{h\gamma}}\right] \in \mathbb{C}.$$

We modeled such spectrum in (15), which generalizes bilinear games, where the spectrum reduces to $\pm[ai, bi]$ (i.e., with $c = 0$). We summarize the optimal hyperparameters for this case:

Theorem 6 (Case 3)**.** *Consider solving* (1) *for games where the Jacobian has a shifted imaginary spectrum in* (15)*. For this problem, the optimal hyperparameters for the momentum extragradient method in* (4) *are:*

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

Similarly to before, we compute the asymptotic convergence rate from Theorem 4 using (9).

$$\sqrt[4]{m} = \left(\frac{\sqrt{c^2+b^2}-\sqrt{c^2+a^2}}{\sqrt{c^2+b^2}+\sqrt{c^2+a^2}}\right)^{1/2} = \left(1 - \frac{2\sqrt{c^2+a^2}}{\sqrt{c^2+b^2}+\sqrt{c^2+a^2}}\right)^{1/2} \tag{20}$$

Note that by setting $c = 0$, the rate in (20) matches the lower bound of bilinear game: $\sqrt{\frac{b-a}{b+a}}$ (Azizian et al., 2020b, Proposition 5). Further, with $c > 0$, the convergence rate in (20) gets faster, showing the difference between classical bilinear games and the regularized one.

Remark 3. *Notice that the optimal momentum* m *in both Theorems 5 and 6 are positive. This is in contrast to Gidel et al. (2019), where the* ***gradient*** *method with negative momentum is studied. This difference elucidates distinct dynamics of how momentum interacts with the* ***gradient*** *and the* ***extragradient*** *methods.*

5 Comparison with Other Methods

We now compare the asymptotic rates of MEG we obtained in the previous section with other first-order methods, including gradient (GD), momentum gradient (GDM), and extragradient (EG) methods.

Comparison with GD and EG. In Azizian et al. (2020a), GD and EG are interpreted as fixed-point iterations, following Polyak (1987). In this framework, iterates of a method are generated by:

$$w_{t+1} = F(w_t), \quad \forall t \geqslant 0, \tag{21}$$

where $F : \mathbb{R}^d \to \mathbb{R}^d$ is an operator representing the method. Analyzing the above scheme in general settings is usually challenging, as F can be any nonlinear operator. However, if F is twice differentiable and w is near a stationary point $w^\star$, the analysis can be simplified by linearizing F at the stationary point: $F(w) \approx F(w^\star) + \nabla F(w^\star)(w - w^\star)$. Then, for w_0 in a neighborhood of $w^\star$, one can obtain an asymptotic convergence rate of (21) by $\rho(\nabla F(w^\star)) \leqslant \rho^\star < 1$, meaning that (21) converges linearly to $w^\star$ at the rate $O((\rho^\star + \varepsilon)^t)$ for $\varepsilon \geqslant 0$. Further, if F is linear, $\varepsilon = 0$.

In this scheme, the fixed point operators F_h^{GD} and F_h^{EG} of GD and EG[7] respectively are:

$$\text{(GD)} \quad w_{t+1} = w_t - hv(w_t) = F_h^{\text{GD}}(w_t), \quad \text{and} \tag{22}$$

$$\text{(EG)} \quad w_{t+1} = w_t - hv(w_t - hv(w_t)) = F_h^{\text{EG}}(w_t). \tag{23}$$

---

[7] Azizian et al. (2020a) assumes that EG uses the same step size h for both the main and the extrapolation steps.

The local convergence rate can then be obtained by bounding the spectral radius of the Jacobian of the operators under certain assumptions. We summarize the relevant results below:

Theorem 7 (Azizian et al. (2020a); Gidel et al. (2019))**.** *Let $w^\star$ be a stationary point of v. Further, assume the eigenvalues of $\nabla v(w^\star)$ all have positive real parts. Then, denoting $\mathcal{S}^\star := Sp(\nabla v(w^\star))$,*

1. *For the gradient method in (22) with step size $h = \min_{\lambda\in\mathcal{S}^\star} \Re(1/\lambda)$, it satisfies:*[8]

$$\rho(\nabla F_h^{GD}(w^\star))^2 \leqslant 1 - \min_{\lambda\in\mathcal{S}^\star} \Re\left(\tfrac{1}{\lambda}\right) \min_{\lambda\in\mathcal{S}^\star} \Re(\lambda). \tag{24}$$

2. *For the extragradient method in (23) with step size $h = (4\sup_{\lambda\in\mathcal{S}^\star} |\lambda|)^{-1}$, it satisfies:*

$$\rho(\nabla F_h^{EG}(w^\star))^2 \leqslant 1 - \tfrac{1}{4}\left(\frac{\min_{\lambda\in\mathcal{S}^\star}\Re(\lambda)}{\sup_{\lambda\in\mathcal{S}^\star}|\lambda|} + \frac{1}{16}\frac{\min_{\lambda\in\mathcal{S}^\star}|\lambda|^2}{\sup_{\lambda\in\mathcal{S}^\star}|\lambda|^2}\right). \tag{25}$$

Since all three cases of our spectrum models in (12), (14), and (15) satisfy 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. We summarize this comparison in the next corollary.

Corollary 1. *With the conditions in Theorem 7, for each case of the Jacobian spectrum $\mathcal{S}_1^\star$, $\mathcal{S}_2^\star$, and $\mathcal{S}_3^\star$, the gradient method in (22) and the extragradient method in (23) satisfy the following:*

- ***Case 1:*** *$Sp(\nabla v) \in \mathcal{S}_1^\star = [\mu_1, L_1] \cup [\mu_2, L_2] \in \mathbb{R}$:*

$$\rho(\nabla F_h^{GD}(w^\star))^2 \leqslant 1 - \tfrac{\mu_1}{L_2}, \quad \textit{and} \quad \rho(\nabla F_h^{EG}(w^\star))^2 \leqslant 1 - \tfrac{1}{4}\left(\tfrac{\mu_1}{L_2} + \tfrac{\mu_1^2}{16L_2^2}\right). \tag{26}$$

- ***Case 2:*** *$Sp(\nabla v) \in \mathcal{S}_2^\star = [\mu, L] \cup \{z \in \mathbb{C} : \Re(z) = c' > 0,\ \Im(z) \in [-c, c]\}$:*

$$\begin{aligned}
\rho(\nabla F_h^{GD}(w^\star))^2 &\leqslant \begin{cases} 1 - \frac{2\mu}{4c^2/(L-\mu)+(L-\mu)} & \textit{if } c \geqslant \sqrt{\frac{L^2-\mu^2}{4}}, \\ 1 - \frac{\mu}{L} & \textit{otherwise.} \end{cases} \\
\rho(\nabla F_h^{EG}(w^\star))^2 &\leqslant \begin{cases} 1 - \frac{1}{4}\left(\frac{\mu}{\sqrt{c^2+((L-\mu)/2)^2}} + \frac{\mu^2}{16(c^2+((L-\mu)/2)^2)}\right) & \textit{if } c \geqslant \sqrt{\frac{3L^2+2L\mu-\mu^2}{4}}, \\ 1 - \frac{1}{4}\left(\frac{\mu}{L} + \frac{\mu^2}{16L^2}\right) & \textit{otherwise.} \end{cases}
\end{aligned} \tag{27}$$

- ***Case 3:*** *$Sp(\nabla v) \in \mathcal{S}_3^\star = [c + ai, c + bi] \cup [c - ai, c - bi] \in \mathbb{C}$:*

$$\rho(\nabla F_h^{GD}(w^\star))^2 \leqslant 1 - \tfrac{c^2}{c^2+b^2}, \quad \textit{and} \quad \rho(\nabla F_h^{EG}(w^\star))^2 \leqslant 1 - \tfrac{1}{4}\left(\tfrac{c}{\sqrt{c^2+b^2}} + \tfrac{c^2+a^2}{16(c^2+b^2)}\right). \tag{28}$$

For Case 1, we see from (26) that both GD and EG have convergence rate of the form $1 - O(\mu_1/L_2) = 1 - O(\tau)$, as expected. In (18), we showed that MEG enjoys not only an accelerated convergence rate (with $\sqrt{\tau}$), but also additional constant improvement by the factor of $\sqrt{1 - R^2}$. For Case 2, we showed in (19) that MEG enjoys an accelerated convergence rate $1 - O(\sqrt{\mu/L})$ as long as $c = O(L)$. Under the same condition, both GD and EG in (27) have non-accelerated convergence rates. For Case 3, we showed in (20) that MEG again enjoys an accelerated convergence rate that matches the known lower bound for bilinear games $\sqrt{\frac{b-a}{b+a}}$ if $c = 0$, and improves for $c \neq 0$. However, as can be seen in (27), GD and EG suffer from slower rates.

Comparison with GDM. We now compare the convergence rate of MEG with that of GDM, which iterates as in (5). In Azizian et al. (2020b), it was shown that GD is optimal for the class of games where the eigenvalues of the Jacobian is included in a *disc* in the complex plane, indicating acceleration is not possible for this problem class.[9] However, it is well-known that GDM achieves the optimal convergence

---

[8] Note that the spectral radius ρ is squared, but asymptotically is almost the same as $\sqrt{1-x} \leqslant 1 - x/2$.

[9] Yet, one can consider the case, e.g., where a cross-shape is contained in a disc. Then, by knowing more fine-grained structure of the Jacobain spectrum, MEG can have faster convergence as we show in (19).

rate when the eigenvalues of the Jacobian lie on the (strictly positive) real line segment, which amounts to (strongly-convex) minimization (Polyak, 1987; Wang et al., 2022). Hence, Azizian et al. (2020b) studies the intermediate case, where the eigenvalues of the Jacobian reside in an ellipse, which can be thought of as the real segment $[\mu, L]$ perturbed with ϵ in an elliptic way. That is, they consider the following spectral shape:[10]

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

Similarly to GD and EG above, in Azizian et al. (2020b), GDM is interpreted as a fixed point iteration:[11]

$$w_{t+1} = w_t - h v(w_t) + m(w_t - w_{t-1}) = F^{\text{GDM}}(w_t, w_{t-1}). \tag{29}$$

To study the convergence rate of GDM, we use the following theorem from Azizian et al. (2020b):

Theorem 8 (Azizian et al. (2020b))**.** *Define $\epsilon(\mu, L)$ as $\epsilon(\mu, L)/L = (\mu/L)^\theta = \tau^\theta$ with $\theta > 0$ and $a \wedge b = \min(a, b)$. If $Sp(\nabla F^{GDM}(w^\star, w^\star)) \subset K_\epsilon$, and when $\tau \to 0$, it satisfies:*

$$\rho(\nabla F^{GDM}(w^\star, w^\star)) \leqslant \begin{cases} 1 - 2\sqrt{\tau} + O\left(\tau^{\theta \wedge 1}\right), & \text{if } \theta > \frac{1}{2} \\ 1 - 2(\sqrt{2}-1)\sqrt{\tau} + O\left(\tau\right), & \text{if } \theta = \frac{1}{2} \\ 1 - \tau^{1-\theta} + O\left(\tau^{1 \wedge (2-3\theta)}\right), & \text{if } \theta < \frac{1}{2}, \end{cases} \tag{30}$$

where the hyperparametes h and m are functions of μ, L, and ϵ only.

For Case 1, GDM converges at the rate $1 - 2\sqrt{\tau} + O(\tau)$ (i.e., with $\theta = 1$ from the above), which is always slower than the rate of MEG in (18) by the factor of $\sqrt{1 - R^2}$. For Case 2, we see from Theorem 8 that GDM achieves the accelerated rate, i.e., $1 - O(\sqrt{\tau})$, until $\theta = \frac{1}{2}$. In other words, the biggest elliptic perturbation ϵ where GDM permits the accelerated rate is $\epsilon = \sqrt{\mu L}$.[12] We interpret Theorem 8 for games with cross-shaped Jacobian spectrum in (14) and shifted imaginary spectrum in (15) in the following corollary:

Corollary 2. *With the conditions in Theorem 8, for games with cross-shaped Jacobian spectrum in (14) with $c = \frac{L-\mu}{2}$, GDM cannot achieve an accelerated rate when $\frac{L-\mu}{2} = c > \epsilon = \sqrt{\mu L}$. Using $L > \mu$, the above inequality implies $\frac{L}{\mu} > \sqrt{5}$. That is, when the condition number exceeds $\sqrt{5} \approx 2.236$, GDM cannot be accelerated; on the contrary, as we showed in (19), MEG converges at the accelerated rate in the ill-conditioned regime.*

Case 3 cannot be deduced from Theorem 8, as it assumes the real line segment $[\mu, L]$ with ϵ perturbation (along the imaginary axis), whereas $\mathcal{S}_3^\star$ in (15) has a fixed real component. Instead, we utilize the link function of GDM in (7) to show that it is unlikely for GDM to stay in the robust region: $\xi^{-1}([-1, 1])$.

Proposition 2. *Consider the momentum gradient method in (5), expressed via Chebyshev polynomials as in (7), with link function $\xi(\lambda)$. For any complex number $z \in \mathbb{C}$ with $\Re(z) = p$, if $\frac{2(1+m)}{h} < p$, then GDM cannot stay in the robust region.*

The condition $\frac{2(1+m)}{h} < p$ is hard to avoid even for small p, considering h is usually a small value.

6 Experiments

In this section, we perform synthetic experiments on optimizing a game where the Jacobian has a cross-shaped spectrum in (14). We chose this spectrum as it is the "hardest" case, involving both real and complex eigenvalues (c.f., Theorem 3). To test the robustness, we consider two cases where the Jacobian spectrum exactly follows a cross-shape, as well as the case when it is inexact. We illustrate them in Figure 2.

We focus on two-player quadratic games, where player 1 controls $x \in \mathbb{R}^{d_1}$ and player 2 controls $y \in \mathbb{R}^{d_2}$ with loss functions in (16). In our setting, the corresponding vector field in (17) satisfies $M_{12} = -M_{21}^\top$, but S_1

---

[10] A visual illustration of this ellipse can be found in (Azizian et al., 2020b, Figure 2).

[11] As the GDM recursion for w_{t+1} depends on w_t and w_{t-1}, Azizian et al. (2020b) uses an augmented operator; see Lemma 2 in that work for details.

[12] Observe that $(\mu/L)^{1/2} = \epsilon(\mu, L)/L \implies \epsilon(\mu, L) = \sqrt{\mu L}$.

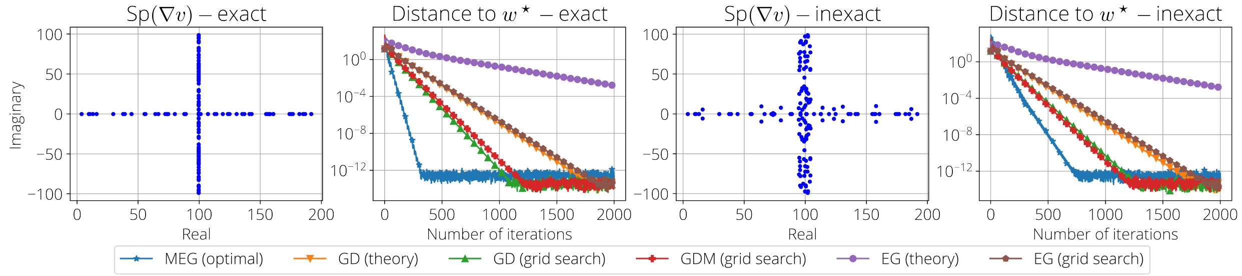

Figure 2: Comparison of the first-order methods considered in this work, for optimizing a game with a cross-shaped Jacobian spectrum in (14). All algorithms are initialized with 0, and run for 2000 iterations.

and S_2 can be nonzero symmetric matrices. Further, the Jacobian $\nabla v = A$ has the cross-shape eigenvalue structure in (14), with $c = \frac{L-\mu}{2}$. (c.f., Proposition 1, Case 2). For the problem constants, we use $\mu = 1$, and $L = 200$. The optimum $[x^\star \; y^\star]^\top = w^\star \in \mathbb{R}^{200}$ is generated using the standard normal distribution. For simplicity, we assume $b = [b_1 \; b_2]^\top = [0 \; 0]^\top$. For the algorithms, we compare GD in (22), GDM in (5), EG in (23), and MEG in (4). All algorithms are initialized with 0. We plot the experimental results in Figure 2.

For MEG (optimal), we set the hyperparameters using Theorem 5. For GD (theory) and EG (theory), we set the hyperparameters using Theorem 7, both for exact and inexact settings. For GDM (grid search), as Theorem 8 does not give a specific form for hyperparater setup, we perform grid search of h^{GDM} and m^{GDM}, and choose the best performing ones. Specifically, we consider $0.005 \leqslant h^{\text{GDM}} \leqslant 0.015$ with 10^{-3} increment, and $0.01 \leqslant m^{\text{GDM}} \leqslant 0.99$ with 10^{-2} increment. Further, as Theorem 7 might be conservative, we perform grid searches for GD and EG as well. For GD (grid search), we use the same setup as h^{GDM}. For EG (grid search), we use $0.001 \leqslant h^{\text{EG}} \leqslant 0.05$ with 10^{-4} increment.

There are several remarks to make. First, even though the bottom panel of Figure 2, where Sp(∇v) *does not* exactly follow the spectrum model in (14), MEG with the optimal hyperparameters from Theorem 5 still work well, even though MEG (optimal) took more iterations in the inexact case compared to the exact case. Comparing with other algorithms, first notice that MEG (optimal) with the hyperparameters obtained in Theorem 5 indeed exhibit orders of magnitude faster rate of convergence, even compared to other methods with grid-search hyperparameter tuning, corroborating our theoretical findings in Section 4. Second, while EG (theory) is slower than GD (theory), confirming Corrolary 1, GD (grid search) can be tuned to converge faster via grid search. Third, even though the best performance of GDM (grid search) is obtained through grid search, one can see the GD (grid search) obtains slightly faster convergence rate than GDM (grid search), confirming Corollay 2.

7 Conclusion

In this work, we focused on the momentum extragradient method for finding a stationary point of differentiable games. Via polynomial-based analysis, we showed the momentum extragradient method converges in three different modes of eigenvalue structure of the game Jacobian, depending on the hyperparameters. Then, for each case, we obtained the optimal hyperparameters for the momentum extragradient method, which enjoys accelerated asymptotic convergence rates. We compared the obtained rates of the momentum extragradient method with other first order methods, and showed that the considered methods do not achieve the accelerated convergence rate.

References

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

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*, pp. 2863–2873. PMLR, 2020a.

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*, pp. 1705–1715. PMLR, 2020b.

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*, pp. 354–363. PMLR, 2018.

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`.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Accelerated gossip in networks of given dimension using jacobi polynomial iterations. *SIAM Journal on Mathematics of Data Science*, 2(1):24–47, 2020.

Aleksandr Beznosikov, Pavel Dvurechensky, Anastasia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *arXiv preprint arXiv:2106.08315*, 2021.

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

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in neural information processing systems*, 31, 2018.

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.

Carles Domingo-Enrich, Fabian Pedregosa, and Damien Scieur. Average-case acceleration for bilinear games and normal matrices. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=H0syOoy3Ash`.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=6Tm1mposlrM`.

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.

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*, pp. 1802–1811. PMLR, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL `https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf`.

Eduard Gorbunov, Nicolas Loizou, and Gauthier Gidel. Extragradient method: O (1/k) last-iterate convergence for monotone variational inequalities and connections with cocoercivity. In *International Conference on Artificial Intelligence and Statistics*, pp. 366–402. PMLR, 2022.

Eduard Gorbunov, Adrien Taylor, Samuel Horváth, and Gauthier Gidel. Convergence of proximal point and extragradient-based methods beyond monotonicity: the case of negative comonotonicity. In *International Conference on Machine Learning*, pp. 11614–11641. PMLR, 2023.

Baptiste Goujaud and Fabian Pedregosa. Cyclical step-sizes, 2022. URL `http://fa.bianp.net/blog/2022/cyclical/`.

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*, pp. 3028–3065. PMLR, 2022.

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.

Andrew J Hetzel, Jay S Liew, and Kent E Morrison. The probability that a matrix of integers is diagonalizable. *The American Mathematical Monthly*, 114(6):491–499, 2007.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

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

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.

Chris Junchi Li, Yaodong Yu, Nicolas Loizou, Gauthier Gidel, Yi Ma, Nicolas Le Roux, and Michael I Jordan. On the convergence of stochastic extragradient for bilinear games with restarted iteration averaging. *arXiv preprint arXiv:2107.00464*, 2021.

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*, pp. 907–915. PMLR, 2019.

Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, and Payel Das. A decentralized parallel algorithm for training generative adversarial nets. *Advances in Neural Information Processing Systems*, 33:11056–11070, 2020.

Jonathan P. Lorraine, David Acuna, Paul Vicol, and David Duvenaud. Complex momentum for optimization in games. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7742–7765. PMLR, 28–30 Mar 2022.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pp. 3481–3490. PMLR, 2018.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020.

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*, pp. 1497–1507. PMLR, 2020.

Renato DC Monteiro and Benar Fux Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

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.

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

Samet Oymak. Provable super-convergence with a large cyclical learning rate. *IEEE Signal Processing Letters*, 28:1645–1649, 2021.

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

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *The Journal of Machine Learning Research*, 21(1):10197–10260, 2020.

Fabian Pedregosa. Momentum: when chebyshev meets chebyshev, 2020. URL `http://fa.bianp.net/blog/2020/momentum/`.

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

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1:32, 1987.

Ernest K Ryu, Kun Yuan, and Wotao Yin. Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems. *arXiv preprint arXiv:1905.10899*, 2019.

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

Mikhail V Solodov and Benar F Svaiter. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999.

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.

Jun-Kun Wang, Chi-Heng Lin, Andre Wibisono, and Bin Hu. Provable acceleration of heavy ball beyond quadratics for a class of polyak-lojasiewicz functions when the non-convexity is averaged-out. In *International Conference on Machine Learning*, pp. 22839–22864. PMLR, 2022.

Jian Zhang and Ioannis Mitliagkas. Yellowfin and the art of momentum tuning. *Proceedings of Machine Learning and Systems*, 1:289–308, 2019.