Abstract

Games generalize the single-objective optimization paradigm by introducing different objective functions for different players. Differentiable games often proceed by simultaneous or alternating gradient updates. In machine learning, games are gaining new importance through formulations like generative adversarial networks (GANs) and actor-critic systems. However, compared to single-objective optimization, game dynamics is more complex and less understood. In this paper, we analyze gradient-based methods with momentum on simple games. We prove that alternating updates are more stable than simultaneous updates. Next, we show both theoretically and empirically that alternating gradient updates with a negative momentum term achieves convergence in a difficult toy adversarial problem, but also on the notoriously difficult to train saturating GANs.

1 INTRODUCTION

Recent advances in machine learning are largely driven by the success of gradient-based optimization methods for the training process. A common learning paradigm is empirical risk minimization, where a (potentially non-convex) objective, that depends on the data, is minimized. However, some recently introduced approaches require the joint minimization of several objectives. For example, actor-critic methods can be written as a bi-level optimization problem (Pfau and Vinyals, 2016) and generative adversarial networks (GANs) (Goodfellow et al., 2014) use a two-player game formulation.

We prove in §5 that the alternating gradient method with negative momentum is the only setting within our study parameters (Fig. 2) that converges on a bilinear smooth game. Using a zero or positive momentum value, or doing simultaneous updates in such games fails to converge.

We show in §4 that, for general dynamics, when the eigenvalues of the Jacobian have a large imaginary part, negative momentum can improve the local convergence properties of the gradient method.

We confirm the benefits of negative momentum for training GANs with the notoriously ill-behaved saturating loss on both toy settings, and real datasets.

Outline. §2 describes the fundamentals of the analytic setup that we use. §3 provides a formulation for the optimal step-size, and discusses the constraints and intuition behind it. §4 presents our theoretical results.
and guarantees on negative momentum. §5 studies the properties of alternating and simultaneous methods with negative momentum on a bilinear smooth game. §6 contains experimental results on toy and real datasets. Finally, in §7, we review some of the existing work on smooth game optimization as well as GAN stability and convergence.

2 BACKGROUND

Notation In this paper, scalars are lower-case letters (e.g., $\lambda$), vectors are lower-case bold letters (e.g., $\mathbf{\theta}$), matrices are upper-case bold letters (e.g., $\mathbf{A}$) and operators are upper-case letters (e.g., $\mathcal{F}$). The spectrum of a squared matrix $\mathbf{A}$ is denoted by $\text{Sp}(\mathbf{A})$, and its spectral radius is defined as $\rho(\mathbf{A}) := \max\{|\lambda| \mid \lambda \in \text{Sp}(\mathbf{A})\}$. We respectively note $\sigma_{\text{min}}(\mathbf{A})$ and $\sigma_{\text{max}}(\mathbf{A})$ the smallest and the largest positive singular values of $\mathbf{A}$. The identity matrix of $\mathbb{C}^{n \times n}$ is written $\mathbf{I}_n$. We use $\mathbb{R}$ and $\mathbb{C}$ to respectively denote the real and imaginary part of a complex number. $\mathcal{O}$, $\Omega$ and $\Theta$ stand for the standard asymptotic notations. Finally, all the omitted proofs can be found in §D.

Game theory formulation of GANs Generative adversarial networks consist of a discriminator $D_{\varphi}$ and a generator $G_{\theta}$. In this game, the discriminator’s objective is to tell real from generated examples. The generator’s goal is to produce examples that are sufficiently close to real examples to confuse the discriminator.

From a game theory point of view, GAN training is a differentiable two-player game: the discriminator $D_{\varphi}$ aims at minimizing its cost function $\mathcal{L}_D$ and the generator $G_{\theta}$ aims at minimizing its own cost function $\mathcal{L}_G$. Using the same formulation as the one in Mescheder et al. (2017) and Gidel et al. (2018), the GAN objective has the following form,

$$
\begin{align}
\theta^* & \in \arg \min_{\theta \in \Theta} \mathcal{L}_G(\theta, \varphi^*) \\
\varphi^* & \in \arg \min_{\varphi \in \Phi} \mathcal{L}_D(\theta^*, \varphi).
\end{align}

(1)
$$

Given such a game setup, GAN training consists of finding a local Nash Equilibrium, which is a state $(\varphi^*, \theta^*)$ in which neither the discriminator nor the generator can improve their respective cost by a small change in their parameters. In order to analyze the dynamics of gradient-based methods near a Nash Equilibrium, we look at the gradient vector field,

$$
v(\varphi, \theta) := \left[\nabla_\varphi \mathcal{L}_D(\varphi, \theta) \quad \nabla_\theta \mathcal{L}_G(\varphi, \theta)\right]^T,
$$

(2)

and its associated Jacobian $\nabla v(\varphi, \theta)$,

$$
\left[\begin{array}{c} 
\nabla^2_{\varphi} \mathcal{L}_D(\varphi, \theta) & \nabla_\varphi \nabla_\theta \mathcal{L}_D(\varphi, \theta) \\
\nabla_\varphi \nabla_\theta \mathcal{L}_G(\varphi, \theta) & \nabla^2_\theta \mathcal{L}_G(\varphi, \theta)
\end{array} \right].
$$

(3)

Games in which $\mathcal{L}_G = -\mathcal{L}_D$ are called zero-sum games and (1) can be reformulated as a min-max problem. This is the case for the original min-max GAN formulation, but not the case for the non-saturating loss (Goodfellow et al., 2014) which is commonly used in practice.

For a zero-sum game, we note $\mathcal{L}_G = -\mathcal{L}_D = \mathcal{L}$. When the matrices $\nabla^2_{\varphi} \mathcal{L}(\varphi, \theta)$ and $\nabla^2_\theta \mathcal{L}(\varphi, \theta)$ are zero, the Jacobian is anti-symmetric and has pure imaginary eigenvalues. We call games with pure imaginary eigenvalues purely adversarial games. This is the case in a simple bilinear game $\mathcal{L}(\varphi, \theta) := \varphi^\top \mathbf{A} \theta$. This game can be formulated as a GAN where the true distribution is a Dirac on 0, the generator is a Dirac on $\theta$ and the discriminator is linear. This setup was extensively studied in 2D by Gidel et al. (2018).

Conversely, when $\nabla_\varphi \nabla_\theta \mathcal{L}(\varphi, \theta)$ is zero and the matrices $\nabla^2_{\varphi} \mathcal{L}(\varphi, \theta)$ and $-\nabla^2_\theta \mathcal{L}(\varphi, \theta)$ are symmetric and definite positive, the Jacobian is symmetric and has real positive eigenvalues. We call games with real positive eigenvalues purely cooperative games. This is the case, for example, when the objective function $\mathcal{L}$ is separable such as $\mathcal{L}(\varphi, \theta) = f(\varphi) - g(\theta)$ where $f$ and $g$ are two convex functions. Thus, the optimization can be reformulated as two separate minimization of $f$ and $g$ with respect to their respective parameters.

These notions of adversarial and cooperative games can be related to the notions of potential games (Monderer and Shapley, 1996) and Hamiltonian games recently introduced by Balduzzi et al. (2018): a game is a potential game (resp. Hamiltonian game) if its Jacobian is symmetric (resp. asymmetric). Our definition of cooperative game is a bit more general than the definition of potential game since some non-symmetric matrices may have positive eigenvalues. Similarly, the notion of

Figure 1: Decreasing trend in the value of momentum used for training GANs across time.
adversarial game generalizes the Hamiltonian games since some non-antisymmetric matrices may have pure imaginary eigenvalues, for instance,

\[
\text{Sp} \left( \begin{pmatrix} 0 & -1 \\ 2 & 3 \end{pmatrix} \right) = \{1, 2\}, \quad \text{Sp} \left( \begin{pmatrix} -1 & 1 \\ -2 & 1 \end{pmatrix} \right) = \{\pm i\}.
\]

In this work, we are interested in games in between purely adversarial games and purely cooperative ones, i.e., games which have eigenvalues with non-negative real part (cooperative component) and non-zero imaginary part (adversarial component). For \(\mathbf{A} \in \mathbb{R}^{d \times p}\), a simple class of such games is parametrized by \(\alpha \in [0, 1]\),

\[
\min_{\theta \in \mathbb{R}^p} \max_{\varphi \in \mathbb{R}^d} \alpha \|\theta\|^2 + (1 - \alpha) \theta^\top \mathbf{A} \varphi - \alpha \|\varphi\|^2,
\]

Simultaneous Gradient Method. Let us consider the dynamics of the simultaneous gradient method. It is defined as the repeated application of the operator,

\[
F_\eta(\varphi, \theta) := \left[ \varphi, \theta \right]^\top - \eta \nabla(\varphi, \theta), \quad (\varphi, \theta) \in \mathbb{R}^m,
\]

where \(\eta\) is the learning rate. Now, for brevity we write for the joint parameters \(\omega := (\varphi, \theta) \in \mathbb{R}^m\). For \(t \in \mathbb{N}\), let \(\omega_t = (\varphi_t, \theta_t)\) be the \(t\)th point of the sequence computed by the gradient method,

\[
\omega_t = F_\eta \circ \ldots \circ F_\eta(\omega_0) = F_\eta(t)(\omega_0).
\]

Then, if the gradient method converges, and its limit point \(\omega^* = (\varphi^*, \theta^*)\) is a fixed point of \(F_\eta\) such that \(\nabla \nabla(\omega^*)\) is positive-definite, then \(\omega^*\) is a local Nash equilibrium. Interestingly, some of the stable stationary points of gradient dynamics may not be Nash equilibrium (Adolphs et al., 2018). In this work, we focus on the local convergence properties near the stationary points of gradient. To the best of our knowledge, there is no first order method alleviating this issue. In the following, \(\omega^*\) is a stationary point of the gradient dynamics (i.e. a point such that \(\nabla(\omega^*) = 0\)).

### 3 TUNING THE STEP-SIZE

Under certain conditions on a fixed point operator, linear convergence is guaranteed in a neighborhood around a fixed point.

**Theorem 1** (Prop. 4.4.1 Bertsekas (1999)). If the spectral radius \(\rho_{\text{max}} := \rho(\nabla F_\eta(\omega^*)) < 1\), then, for \(\omega_0\) in a neighborhood of \(\omega^*\), the distance of \(\omega_t\) to the stationary point \(\omega^*\) converges at a linear rate of \(\mathcal{O}(\rho_{\text{max}} + \varepsilon)^t\), \(\forall \varepsilon > 0\).

From the definition in (5), we have:

\[
\nabla F_\eta(\omega^*) = I_m - \eta \nabla(\omega^*),
\]

and

\[
\text{Sp}(\nabla F_\eta(\omega^*)) = \{1 - \eta \lambda | \lambda \in \text{Sp}(\nabla(\omega^*))\}.
\]

If the eigenvalues of \(\nabla(\omega^*)\) all have a positive real-part, then for small enough \(\eta\), the eigenvalues of \(\nabla F_\eta(\omega^*)\) are inside a convergence circle of radius \(\rho_{\text{max}} < 1\), as illustrated in Fig. 3. Thm. 1 guarantees the existence of an optimal step-size \(\eta_{\text{best}}\) which yields a non-trivial convergence rate \(\rho_{\text{max}} < 1\). Thm. 2 gives analytic bounds on the optimal step-size \(\eta_{\text{best}}\), and lower-bounds the best convergence rate \(\rho_{\text{max}}(\eta_{\text{best}})\) we can expect.

**Theorem 2.** If the eigenvalues of \(\nabla(\omega^*)\) all have a positive real-part, then, the best step-size \(\eta_{\text{best}}\), which minimizes the spectral radius \(\rho_{\text{max}}(\eta)\) of \(\nabla F_\eta(\omega^*, \theta^*)\), is the solution of a (convex) quadratic by parts problem, and satisfies,

\[
\max_{1 \leq k \leq m} \sin(\psi_k)^2 \leq \rho_{\text{max}}(\eta_{\text{best}}) \leq 1 - \Re(1/\lambda_1)\delta,
\]

with

\[
\delta := \min_{1 \leq k \leq m} |\lambda_k|^2 / (2(\Re(1/\lambda_k) - \Re(1/\lambda_1)))
\]

and

\[
\Re(1/\lambda_1) \leq \eta_{\text{best}} \leq 2\Re(1/\lambda_1)
\]

where \(\lambda_k = \Re(\psi_k)^2\) are sorted such that \(0 < \Re(1/\lambda_1) \leq \cdots \leq \Re(1/\lambda_m)\). Particularly, when \(\eta_{\text{best}} = \Re(1/\lambda_1)\) we are in the case of the top plot of Fig.3 and \(\rho_{\text{max}}(\eta_{\text{best}})^2 = \sin(\psi)^2\).
Negative Momentum for Improved Game Dynamics

We illustrate and interpret these two cases in Fig. 3. In
When we can see that if the Jacobian of
imaginary eigenvalue
and we propose an illustrative example to discuss this
condition number might be meaningful for such games
dynamics of smooth games. We think that a notion of
inverse of the eigenvalues play an important role in the
condition number
depends on the extreme eigenvalues and their ratio,
thus, the convergence rate of the gradient method
may be arbitrarily close to 1. Zhang and Mitliagkas
(2017) provide an analysis of the momentum method
for quadratics, showing that momentum can actually
help to better condition the model. One interesting
point from their work is that the best conditioning is
achieved when the added momentum makes the Jaco-
bian eigenvalues turn from positive reals into complex
conjugate pairs. Our goal is to use momentum to wrangle
game dynamics into convergence manipulating the
eigenvalues of the Jacobian.

4 NEGATIVE MOMENTUM

As shown in (8), the presence of eigenvalues with
large imaginary parts can restrict us to small step-
sizes and lead to slow convergence rates. In order
to improve convergence, we add a negative mo-
tum term into the update rule. Informally, one can
think of negative momentum as friction that can
damp oscillations. The new momentum term leads
to a modification of the parameter update operator
F_\eta(\omega) of (5). We use a similar state augmentation
as Zhang and Mitliagkas (2017) to form a compound
state (\omega_t, \omega_{t-1}) := (\phi_t, \theta_t, \phi_{t-1}, \theta_{t-1}) \in \mathbb{R}^{2m}.
the update rule (5) turns into the following,

F_{\eta,\beta}(\omega_t, \omega_{t-1}) = (\omega_{t+1}, \omega_t)

(11)

where

\omega_{t+1} := \omega_t - \eta \nabla v(\omega_t) + \beta (\omega_t - \omega_{t-1}),

(12)
in which \beta \in \mathbb{R} is the momentum parameter. Therefore,
the Jacobian of F_{\eta,\beta} has the following form,

\begin{equation}
\begin{bmatrix}
I_n & 0_n \\
I_n & 0_n
\end{bmatrix} - \eta \begin{bmatrix}
\nabla v(\omega_t) & 0_n \\
0_n & 0_n
\end{bmatrix} + \beta \begin{bmatrix}
I_n & -I_n \\
I_n & I_n
\end{bmatrix}
\end{equation}

(13)

Note that for \beta = 0, we recover the gradient method.

In some situations, if \beta < 0 is adjusted properly, nega-
tive momentum can improve the convergence rate to a
local stationary point by pushing the eigenvalues of its
Jacobian towards the origin. In the following theorem,
we provide an explicit equation for the eigenvalues of
the Jacobian of F_{\eta,\beta}.

Theorem 3. The eigenvalues of \nabla F_{\eta,\beta}(\omega^*) are

\mu_{\pm}(\beta, \eta, \lambda) := (1 - \eta \lambda + \beta) \pm \Delta^{\frac{1}{2}},

(14)

where \Delta := 1 - \frac{4 \beta}{(1 - \eta \lambda + \beta)^2}, \lambda \in \text{Sp}(\nabla v(\omega^*)) and \Delta^{\frac{1}{2}} is the complex square root of \Delta with positive real part.
Moreover we have the following Taylor approximation,

\mu_{+}(\beta, \eta, \lambda) = 1 - \eta \lambda - \beta \eta \lambda (1 - \eta \lambda) + O(\beta^2),

(15)

\mu_{-}(\beta, \eta, \lambda) = \beta (1 - \eta \lambda) + O(\beta^2).

(16)

* If \Delta is a negative real number we set \Delta^{\frac{1}{2}} := i\sqrt{-\Delta}

Figure 3: Eigenvalues \lambda_i of the Jacobian \nabla v(\phi^*, \theta^*) and their trajectories 1 - \eta \lambda_i for growing step-sizes. The unit
circle is drawn in black, and the red dashed circle has
radius equal to the largest eigenvalue \lambda_{\max}, which is directly
related to the convergence rate. Therefore, smaller red
circles mean better convergence rates. Top: The red circle
is limited by the tangent trajectory line 1 - \eta \lambda_i, which
means the best convergence rate is limited only by the
eigenvalue which will pass furthest from the origin as \eta
grows, i.e., \lambda_i = \arg \min_i \Re(1/\lambda_i). Bottom: The red circle
is cut (not tangent) by the trajectories at points 1 - \eta \lambda_i
and 1 - \eta \lambda_3. The \eta is optimal because any increase in \eta
will push the eigenvalue \lambda_1 out of the red circle, while any
decrease in step-size will retract the eigenvalue \lambda_3 out of
the red circle, which will lower the convergence rate in any
case. (Figure inspired by Mescheder et al. (2017).)
As we have seen previously in Fig. 3 and Thm. 2, there are only few eigenvalues which slow down the convergence. Thm. 4 is a local result showing that a small negative momentum can improve the magnitude of the limiting eigenvalues in the following cases: when there is only one limiting eigenvalue \( \lambda_1 \) (since in that case the optimal step-size is \( \eta_{\text{best}} = \Re(1/\lambda_1) \)) or when there are several limiting eigenvalues \( \lambda_1, \ldots, \lambda_k \) and the intersection \( I(\lambda_1) \cap \cdots \cap I(\lambda_k) \) is not empty. We point out that we do not provide any guarantees on whether this intersection is empty or not but note that if the absolute value of the argument of \( \lambda_1 \) is larger than \( \pi/4 \) then by (10), our theorem provides that the optimal step-size \( \eta_{\text{best}} \) belongs to \( I(\lambda_1) \).

Since our result is local, it does not provide any guarantees on large negative values of \( \beta \). Nevertheless, we numerically optimized (17) with respect to \( \beta \) and \( \eta \) and found that for any non-imaginary fixed eigenvalue \( \lambda \), the optimal momentum is negative and the associated optimal step-size is larger than \( \eta(\lambda) \). Another interesting aspect of negative momentum is that it admits larger step-sizes (see Fig. 4 and 5).

For a game with purely imaginary eigenvalues, when \( |\eta \lambda| \ll 1 \), Thm. 3 shows that \( \mu_+(\beta, \eta, \lambda) \approx 1 - (1 + \beta) \eta \lambda \). Therefore, at the first order, \( \beta \) only has an impact on the imaginary part of \( \mu_+ \). Consequently \( \mu_+ \) cannot be pushed into the unit circle, and the convergence guarantees of Thm. 1 do not apply. In other words, the analysis above provides convergence rates for games without any pure imaginary eigenvalues. It excludes the purely adversarial bilinear example (\( \alpha = 0 \) in Eq. 4) that is discussed in the next section.

## 5 BILINEAR SMOOTH GAMES

In this section we analyze the dynamics of a purely adversarial game described by,

\[
\min_{\theta \in \mathbb{R}^d} \max_{\varphi \in \mathbb{R}^p} \theta^\top A \varphi + \theta^\top b + c^\top \varphi, \quad A \in \mathbb{R}^{d \times p}. \tag{18}
\]

The first order stationary condition for this game characterizes the solutions \((\theta^\ast, \varphi^\ast)\) as

\[
A \varphi^\ast = b \quad \text{and} \quad A^\top \theta^\ast = c. \tag{19}
\]

If \( b \) (resp. \( c \)) does not belong to the column space of \( A \) (resp. \( A^\top \)), the game (18) admits no equilibrium. In the following, we assume that an equilibrium does exist for this game. Consequently, there exist \( b' \) and \( c' \) such that \( b = Ab' \) and \( c = A^\top c' \). Using the translations \( \theta \rightarrow \theta - c' \) and \( \varphi \rightarrow \varphi - b' \), we can assume without loss of generality, that \( p \geq d \), \( b = 0 \) and \( c = 0 \). We provide upper and lower bounds on the squared distance from the known equilibrium,

\[
\Delta_1 = \|\theta_1 - \theta^\ast\|^2_2 + \|\varphi_1 - \varphi^\ast\|^2_2 \tag{20}
\]
where \((\theta^*, \varphi^*)\) is the projection of \((\theta_t, \varphi_t)\) onto the solution space. We show in §C, Lem. 2 that, for our methods of interest, this projection has a simple formulation that only depends on the initialization \((\theta_0, \varphi_0)\).

We aim to understand the difference between the dynamics of simultaneous steps and alternating steps. Practitioners have been widely using the latter instead of the former when optimizing GANs despite the rich optimization literature on simultaneous methods.

5.1 Simultaneous gradient descent

We define this class of methods with momentum using the following formulas,

\[
F_{\eta, \beta}^{\text{sim}}(\theta_t, \varphi_t, \theta_{t-1}, \varphi_{t-1}) := (\theta_{t+1}, \varphi_{t+1}, \theta_t, \varphi_t) \quad (21)
\]

where \[
\begin{align*}
\theta_{t+1} &= \theta_t - \eta_1 A \varphi_t + \beta_1 (\theta_t - \theta_{t-1}) \\
\varphi_{t+1} &= \varphi_t + \eta_2 A^\top \theta_t + \beta_2 (\varphi_t - \varphi_{t-1}).
\end{align*}
\]

In our simple setting, the operator \(F_{\eta, \beta}^{\text{sim}}\) is linear. One way to study the asymptotic properties of the sequence \((\theta_t, \varphi_t)\) is to compute the eigenvalues of \(\nabla F_{\eta, \beta}^{\text{sim}}\). The following proposition characterizes these eigenvalues.

**Proposition 1.** The eigenvalues of \(\nabla F_{\eta, \beta}^{\text{sim}}\) are the roots of the 4th order polynomials:

\[
(x-1)^2(x-\beta_1)(x-\beta_2) + \eta_1 \eta_2 \lambda x^2, \; \lambda \in \text{Sp}(A^\top A) \quad (22)
\]

Interestingly, these roots only depend on the product \(\eta_1 \eta_2\). Any re-scaling \(\eta_1 \to \gamma \eta_1\), \(\eta_2 \to \sigma \eta_2\) does not change the eigenvalues of \(\nabla F_{\eta, \beta}^{\text{sim}}\) and consequently the asymptotic dynamics of the iterates \((\theta_t, \varphi_t)\). The magnitude of the eigenvalues described in (22) characterizes the asymptotic properties for the iterates of the simultaneous method (21). We report the maximum magnitude of these roots for a given \(\lambda\) and for a grid of step-sizes and momentum values in Fig. 7. We observe that they are always larger than 1, which transcribes a diverging behavior. The following theorem provides an analytical rate of divergence.

**Theorem 5.** For any \(\eta_1, \eta_2 \geq 0\) and \(\beta_1 = \beta_2 = \beta\), the iterates of the simultaneous methods (21) diverge as,

\[
\Delta_t \in \begin{cases} 
\Omega(\Delta_0[(1 + \beta)^2 + \eta^2 \sigma_{\max}(A)^2]) \; \text{if} \; \beta \geq 0 \\
\Omega(\Delta_0(1 - \beta)^2) \; \text{if} \; -\frac{1}{10} \leq \beta < 0.
\end{cases}
\]

This theorem states that the iterates of the simultaneous method (21) diverge geometrically for \(\beta \geq -\frac{1}{10}\).

5.2 Alternating gradient descent

Alternating gradient methods take advantage of the fact that the iterates \(\theta_{t+1}\) and \(\varphi_{t+1}\) are computed sequentially, to plug the value of \(\theta_{t+1}\) (instead of \(\theta_t\) for simultaneous update rule) into the update of \(\varphi_{t+1}\),

\[
F_{\eta, \beta}^{\text{alt}}(\theta_t, \varphi_t, \theta_{t-1}, \varphi_{t-1}) := (\theta_{t+1}, \varphi_{t+1}, \theta_t, \varphi_t) \quad (23)
\]

where

\[
\begin{align*}
\theta_{t+1} &= \theta_t - \eta_1 A \varphi_t + \beta_1 (\theta_t - \theta_{t-1}) \\
\varphi_{t+1} &= \varphi_t + \eta_2 A^\top \theta_t + \beta_2 (\varphi_t - \varphi_{t-1}).
\end{align*}
\]

This slight change between (21) and (23) significantly shifts the eigenvalues of the Jacobian. We first characterize them with the following proposition.

**Proposition 2.** The eigenvalues of \(\nabla F_{\eta, \beta}^{\text{alt}}\) are the roots of the 4th order polynomials:

\[
(x-1)^2(x-\beta_1)(x-\beta_2) + \eta_1 \eta_2 \lambda x^2, \; \lambda \in \text{Sp}(A^\top A) \quad (24)
\]

The same way as in (22), these roots only depend on the product \(\eta_1 \eta_2\). The only difference is that the monomial with coefficient \(\eta_1 \eta_2 \lambda\) is of degree 2 in (22) and of degree 3 in (24). This difference is major since, for well chosen values of negative momentum, the eigenvalues described in Prop. 2 lie in the unit disk (see Fig. 7). As a consequence, the iterates of the alternating method with no momentum are bounded and do converge if we add some well chosen negative momentum:

**Theorem 6.** If we set \(\eta \leq \frac{1}{\sigma_{\text{max}}(A)}, \; \beta_1 = -\frac{1}{2}\) and \(\beta_2 = 0\) then we have

\[
\Delta_t+1 \in O\left(\max\{\frac{1}{2}, 1 - \frac{\eta \sigma_{\text{max}}(A)}{4}\}\Delta_0\right) \quad (25)
\]

If we set \(\beta_1 = 0\) and \(\beta_2 = 0\), then there exists \(M > 1\) such that for any \(\eta_1, \eta_2 \geq 0\), \(\Delta_t = \Omega(\Delta_0)\).

Our results from this section, namely Thm. 5 and Thm. 6, are summarized in Fig. 2, and demonstrate how alternating steps can improve the convergence properties of the gradient method for bilinear smooth games. Moreover, combining them with negative momentum can surprisingly lead to a linearly convergent method. The conjecture provided in Fig. 2 (divergence of the alternating method with positive momentum) is backed-up by the results provided in Fig. 5 and §A.1.

6 EXPERIMENTS AND DISCUSSION

**Min-Max Bilinear Game** [Fig. 5] In our first experiments, we showcase the effect of negative momentum in a bilinear min-max optimization setup (4) where \(\phi, \theta \in \mathbb{R}\) and \(A = 1\). We compare the effect of positive and negative momentum in both cases of alternating and simultaneous gradient steps.
Figure 5: The effect of momentum in a simple min-max bilinear game where the equilibrium is at (0, 0). (left-a) Simultaneous GD with no momentum (left-b) Alternating GD with no momentum. (left-c) Alternating GD with a momentum of +0.1. (left-d) Alternating GD with a momentum of −0.1. (right) A grid of experiments for alternating GD with different values of momentum ($\beta$) and step-sizes ($\eta$): While any positive momentum leads to divergence, small enough value of negative momentum allows for convergence with large step-sizes. The color in each cell indicates the normalized distance to the equilibrium after 500k iteration, such that 1.0 corresponds to the initial condition and values larger (smaller) than 1.0 correspond to divergence (convergence).

**Fashion MNIST and CIFAR 10 [Fig. 6]** In our third set of experiments, we use negative momentum in a GAN setup on CIFAR-10 (Krizhevsky and Hinton, 2009) and Fashion-MNIST (Xiao et al., 2017) with saturating loss and alternating steps. We use residual networks for both the generator and the discriminator with no batch-normalization. Following the same architecture as Gulrajani et al. (2017), each residual block is made of two $3 \times 3$ convolution layers with ReLU activation function. Up-sampling and down-sampling layers are respectively used in the generator and discriminator. We experiment with different values of momentum on the discriminator and a constant value of 0.5 for the momentum of the generator. We observe that using a negative value can generally result in samples with higher quality and inception scores. Intuitively, using negative momentum only on the discriminator slows down the learning process of the discriminator and allows for better flow of the gradient to the generator. Note that we provide an additional experiment on mixture of Gaussians in § A.2.

7 RELATED WORK

**Optimization** From an optimization point of view, a lot of work has been done in the context of understanding momentum and its variants (Polyak, 1964; Qian, 1999; Nesterov, 2013; Sutskever et al., 2013). Some recent studies have emphasized the importance of momentum tuning in deep learning such as Sutskever et al. (2013), Kingma and Ba (2015), and Zhang and Mitliagkas (2017), however, none of them consider using negative momentum. Among recent work, using robust control theory, Lessard et al. (2016) study optimization procedures and cover a variety of algorithms including momentum methods. Their analysis is global and they establish worst-case bounds for smooth and strongly-convex functions. Mitliagkas et al. (2016) considered negative momentum in the context of asynchronous single-objective minimization. They show that asynchronous-parallel dynamics ‘bleed’ into optimization updates introducing momentum-like behavior into SGD. They argue that algorithmic momentum and asynchrony-induced momentum add up to create an effective ‘total momentum’ value. They conclude that to attain the optimal (positive) effective momentum in an asynchronous system, one would have to reduce algorithmic momentum to small or sometimes negative values. This differs from our work where we show that for games the optimal effective momentum may be negative. Ghadimi et al. (2015) analyze momentum and provide global convergence properties for functions with Lipschitz-continuous gradients. However, all the results mentioned above are restricted to minimization problems. The purpose of our work is to try to understand how momentum influences game dynamics which is intrinsically different from minimization dynamics.

**GANs as games** A lot of recent work has attempted to make GAN training easier with new optimization methods. Daskalakis et al. (2018) extrapolate the next value of the gradient using previous history and Gidel
In this paper, we show analytically and empirically that alternating updates with negative momentum is the only method within our study parameters (Fig. 2) that converges in bilinear smooth games. We study the effects of using negative values of momentum in a GAN setup both theoretically and experimentally. We show that, for a large class of adversarial games, negative momentum may improve the convergence rate of gradient-based methods by shifting the eigenvalues of the Jacobian appropriately into a smaller convergence disk. We found that, in simple yet intuitive examples, using negative momentum makes convergence to the Nash Equilibrium easier. Our experiments support the use of negative momentum for saturating losses on mixtures of Gaussians, as well as on other tasks using CIFAR-10 and fashion MNIST. Altogether, fully stabilizing learning in GANs requires a deep understanding of the underlying highly non-linear dynamics. We believe our work is a step towards a better understanding of these dynamics. We encourage deep learning researchers and practitioners to include negative values of momentum in their hyper-parameter search.

We believe that our results explain a decreasing trend in momentum values reported in GAN literature. This research was partially supported by the Canada Excellence Research Chair in “Data Science and Focused Research Award and an IV ADO grant. Authors would like to thank NVIDIA Corporation for their generous financial support.

8 CONCLUSION
Acknowledgments

This research was partially supported by the Canada CIFAR AI Chair Program, the FRQNT nouveaux chercheurs program, 2019-NC-257943, the Canada Excellence Research Chair in “Data Science for Real-time Decision-making”, by the NSERC Discovery Grant RGPIN-2017-06936, a Google Focused Research Award and an IVADO grant. Authors would like to thank NVIDIA corporation for providing the NVIDIA DGX-1 used for this research. Authors are also grateful to Frédéric Bastien, Florian Bordes, Adam Beberg, Cam Moore and Nithya Natesan for their support.

References


A ADDITIONNAL FIGURES

A.1 Maximum magnitude of the eigenvalues gradient descent with negative momentum on a bilinear objective

In Figure 7 we numerically (using the formula provided in Proposition 1 and 2) computed the maximum magnitude of the eigenvalues gradient descent with negative momentum on a bilinear objective as a function of the step size $\eta$ and the momentum $\beta$. We can notice that on one hand, for simultaneous gradient method, no value of $\eta$ and $\beta$ provide a maximum magnitude smaller than 1, causing a divergence of the algorithm. On the other hand, for alternating gradient method there exists a sweet spot where the maximum magnitude of the eigenvalues of the operator is smaller than 1 insuring that this method does converge linearly (since the Jacobian of a bilinear minmax problem is constant).

![Contour Plot](image)

Figure 7: Contour plot of the maximum magnitude of the eigenvalues of the polynomial $(x - 1)^2(x - \beta)^2 + \eta^2 x^2$ (left, simultaneous) and $(x - 1)^2(x - \beta)^2 + \eta^2 x^2$ (right, alternated) for different values of the step-size $\eta$ and the momentum $\beta$. Note that compared to (22) and (24) we used $\beta_1 = \beta_2 = \beta$ and we defined $\eta := \sqrt{\eta_1 \eta_2 \lambda}$ without loss of generality. On the left, magnitudes are always larger than 1, and equal to 1 for $\beta = -1$. On the right, magnitudes are smaller than 1 for $\frac{\eta}{2} - 1 \leq \beta \leq 0$ and greater than 1 elsewhere.

A.2 Mixture of Gaussian

[Fig. 8] In this set of experiments we evaluate the effect of using negative momentum for a GAN with saturating loss and alternating steps. The data in this experiment comes from eight Gaussian distributions which are distributed uniformly around the unit circle. The goal is to force the generator to generate 2-D samples that are coming from all of the 8 distributions. Although this looks like a simple task, many GANs fail to generate diverse samples in this setup. This experiment shows whether the algorithm prevents mode collapse or not.
Figure 8: The effect of negative momentum for a mixture of 8 Gaussian distributions in a GAN setup. Real data and the results of using SGD with zero momentum on the Generator and using negative / zero / positive momentum ($\beta$) on the Discriminator are depicted.

We use a fully connected network with 4 hidden ReLU layers where each layer has 256 hidden units. The latent code of the generator is an 8-dimensional multivariate Gaussian. The model is trained for 100,000 iterations with a learning rate of 0.01 for stochastic gradient descent along with values of zero, −0.5 and 0.5 momentum. We observe that negative momentum considerably improves the results compared to positive or zero momentum.

B DISCUSSION ON MOMENTUM AND CONDITIONING

In this section, we analyze the effect of the conditioning of the problem on the optimal value of momentum. Consider the following formulation as an extension of the bilinear min-max game discussed in §5, Eq. 4 ($p = d = n$),

$$\min_{\theta \in \mathbb{R}^n} \max_{\varphi \in \mathbb{R}^n} \alpha \|A^{1/2} \theta\|_2^2 + (1 - \alpha) \varphi^\top A \varphi - \alpha \|D^{1/2} \varphi\|_2^2, \quad \alpha \in [0, 1], \ A \in \mathbb{R}^{n \times n},$$

(26)

where $D$ is a square diagonal positive-definite matrix,

$$D = \begin{bmatrix} d_{1,1} & 0 & 0 & \cdots & 0 \\ 0 & d_{2,2} & 0 & \cdots & 0 \\ 0 & 0 & d_{3,3} & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & d_{n,n} \end{bmatrix}$$

(27)

and its condition number is $\kappa(D) = d_{n,n}/d_{1,1}$. Thus, we can re-write the vector field and the Jacobian as a function of $\alpha$ and $D$,

$$v(\varphi, \theta, \alpha, D) = \begin{bmatrix} - (1 - \alpha) \theta + 2 \alpha D \varphi \\ 2 \alpha D \theta + (1 - \alpha) \varphi \end{bmatrix}, \quad \nabla v(\varphi, \theta, \alpha, D) = \begin{bmatrix} 2 \alpha D & (\alpha - 1) I_n \\ (1 - \alpha) I_n & 2 \alpha D \end{bmatrix}.$$  

(28)

The corresponding eigenvalues $\lambda$ of the Jacobian are,

$$\lambda = 2 \alpha \, d_{j,j} \pm (1 - \alpha) i.$$

(29)

For simplicity, in the following we will note $\nabla F_{n,\beta}$ for $\nabla F_{g,\beta}(\varphi, \theta, \alpha, D)$. 

\[ \begin{array}{c} \text{Real data} \\ \beta = 0.0 \end{array} \begin{array}{c} \beta = 0.5 \\ \beta = -0.5 \end{array} \]
Using Thm. (3), the eigenvalues of $\nabla F_{\eta,\beta}$ are,

$$
\mu_+(\beta, \eta, \lambda) = (1 - \eta \lambda + \beta) \frac{1 + \Delta^2}{2} \quad \text{and} \quad \mu_-(\beta, \eta, \lambda) = (1 - \eta \lambda + \beta) \frac{1 - \Delta^2}{2}.
$$

where $\Delta := 1 - \frac{4 \beta}{(1 - \eta \lambda + \beta)^2}$ and $\Delta^{\frac{1}{2}}$ is the complex square root of $\Delta$ with positive real part. 

Hence the spectral radius of $\nabla F_{\eta,\beta}$ can be explicitly formulated as a function of $\beta$ and $\eta$, 

$$
\rho(\nabla F_{\eta,\beta}) = \max_{\lambda \in \text{Sp}(\nabla F_{\eta,\beta})} \max \left\{ |\mu_+(\beta, \eta, \lambda)|, |\mu_-(\beta, \eta, \lambda)| \right\}
$$

In Figure 9, we numerically computed the optimal $\beta$ that minimizes $\rho_{\text{max}}(\nabla F_{\eta,\beta})$ as a function of the step-size $\eta$, for $n = 2$, $d_{1,1} = 1/\kappa$ and $d_{2,2} = 1$. To balance the game between the adversarial part and the cooperative part, we normalize the matrix $D$ such that the sum of its diagonal elements is $n$. It can be seen that there is a competition between the type of the game (adversarial and cooperative) versus the conditioning of the matrix $D$. In a more cooperative regime, increasing $\kappa$ results in more positive values of momentum which is consistent with the intuition that cooperative games are almost minimization problems where the optimum value for the momentum is known (Polyak, 1964) to be $\beta = \left(\frac{\sqrt{2} - 1}{\sqrt{2} + 1}\right)^2$. Interestingly, even if the condition number of $D$ is large, when the game is adversarial enough, the optimum value for the momentum is negative. This experimental setting seems to suggest the existence of a multidimensional condition number taking into account the difficulties introduced by the ill conditioning of $D$ as well as the adversarial component of the game.

C LEMMAS AND DEFINITIONS

Recall that the spectral radius $\rho(A)$ of a matrix $A$ is the maximum magnitude of its eigenvalues.

$$
\rho(A) := \max\{|\lambda| : \lambda \in \text{Sp}(A)\}.
$$

For a symmetric matrix, this is equal to the spectral norm, which is the operator norm induced by the vector 2-norm. However, we are dealing with general matrices, so these two values may be different. The spectral radius is always smaller than the spectral norm, but it’s not a norm itself, as illustrated by the example below:

If $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ then $\text{Sp}(A) = \{0\} \implies \rho(A) = 0$ but $A^T A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \implies \|A\|_2 = 1$ 

where we used the fact that the spectral norm is also the square root of the largest singular value.
In this section we will introduce three lemmas that we will use in the proofs of §D. The first lemma is about the determinant of a block matrix.

**Lemma 1.** Let $A, B, C, D$ four matrices such that $C$ and $D$ commute. Then

$$
\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |AD - BC|
$$

(33)

where $A$ is the determinant of $A$.

**Proof.** See (Zhang, 2006, Section 0.3).

The second lemma is about the iterates of the simultaneous and the alternating methods introduced in §5 for the bilinear game. It shows that we can pick a subspace where the iterates will remain.

**Lemma 2.** Let $(\theta_t, \varphi_t)$ the updates computed by the simultaneous (resp. alternating) gradient method with momentum (21) (resp. (23)). There exists a couple $(\theta^*, \varphi^*)$ solution of (18) only depending on $(\theta_0, \varphi_0)$ such that,

$$
\theta_t - \theta^* \in \text{span}(AA^T) \quad \text{and} \quad \varphi_t - \varphi^* \in \text{span}(A^T A), \quad \forall t \geq 0.
$$

(34)

**Proof of Lemma 2.** Let us start with the simultaneous updates (21).

Let $U^T D V = A$ the SVD of $A$ where $U$ and $V$ are orthogonal matrices and

$$
D = \begin{bmatrix}
\text{diag}(\sigma_1, \ldots, \sigma_r) & 0_{r,p-r} \\
0_{d-r,r} & 0_{d-r,p-r}
\end{bmatrix}
$$

(35)

where $r$ is the rank of $A$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are the (positive) singular values of $A$. The update rules (21) implies that,

$$
\begin{cases}
\theta_{t+1} = \theta_t - \eta_1 A \varphi_t + \beta_1 (\theta_t - \theta_{t-1}) \\
\varphi_{t+1} = \varphi_t + \eta_2 A^T \theta_t + \beta_2 (\varphi_t - \varphi_{t-1})
\end{cases} \Rightarrow
\begin{cases}
U \theta_{t+1} = U \theta_t - \eta_1 D V \varphi_t + \beta_1 U (\theta_t - \theta_{t-1}) \\
V \varphi_{t+1} = V \varphi_t + \eta_2 D^T U \theta_t + \beta_2 V (\varphi_t - \varphi_{t-1})
\end{cases}
$$

(36)

Consequently, for any $\theta_0 \in \mathbb{R}^d$ and $\varphi_0 \in \mathbb{R}^p$ we have that,

$$
A^T \begin{bmatrix}
0 \\
\vdots \\
U \theta_{0[r+1]} \\
\vdots \\
U \theta_{0[d]}
\end{bmatrix} = 0 \quad \text{and} \quad A \begin{bmatrix}
0 \\
\vdots \\
V \varphi_{0[r+1]} \\
\vdots \\
V \varphi_{0[d]}
\end{bmatrix} = 0
$$

(37)

Since the solutions $(\theta^*, \varphi^*)$ of (18) verify the following first order conditions:

$$
A^T \theta^* = 0 \quad \text{and} \quad A \varphi^* = 0
$$

(38)

One can set $(\theta^*, \varphi^*)$ as in (37) to be a couple of solution of (18) such that $U(\theta_0 - \theta^*) \in \text{span}(D)$ and $V(\varphi_0 - \varphi^*) \in \text{span}(D)$. By an immediate recurrence, using (36) we have that for any initialization $(\theta_0, \varphi_0)$ there exists a couple $(\theta^*, \varphi^*)$ such that that for any $t \geq 0$,

$$
U(\theta_t - \theta^*) \in \text{span}(D) \quad \text{and} \quad V(\varphi_t - \varphi^*) \in \text{span}(D^T)
$$

(39)

Consequently,

$$
\theta_t - \theta^* \in \text{span}(A) = \text{span}(AA^T) \quad \text{and} \quad \varphi_t - \varphi^* \in \text{span}(A^T) = \text{span}(A^T A), \quad t \geq 0
$$

(40)

The proof for the alternated updates (23) are the same since we only use the fact that the iterates stay on the span of interest.

\qed
Lemma 3. Let $M \in \mathbb{R}^{m \times m}$ and $(u_t)$ a sequence such that $u_{t+1} = Mu_t$, then we have three cases of interest for the spectral radius $\rho(M)$:

- If $\rho(M) < 1$, and $M$ is diagonalizable, then $\|u_t\|_2 \in O((\rho(M))^t\|u_0\|_2)$.
- If $\rho(M) > 1$, then there exist $u_0$ such that $\|u_t\|_2 \in \Omega(\rho(M))^t\|u_0\|_2$.
- If $|\lambda| \leq 1$, $\forall \lambda \in \text{Sp}(M)$, and $M$ is diagonalizable then $\|u_t\|_2 \in \Theta(\|u_0\|_2)$.

Proof. For that section we note $\|\cdot\|_2$ the $\ell_2$ norm of $\mathbb{C}^m$:

- If $\rho(M) < 1$:
  We have for $t \geq 0$ and any $u_0 \in \mathbb{R}^m$,
  \[
  \|u_t\|_2 = \|M^t u_0\|_2 \leq \|M^t\| \|u_0\|_2
  \]
  Then we can diagonalize $M = PDP^{-1}$ where $P$ is invertible and $D$ is a diagonal matrix. Hence using $\|\cdot\|_2$ as the norm of $\mathbb{C}^m$ (because $P$ can belong to $\mathbb{C}^{m \times m}$) we have that,
  \[
  \|u_t\|_2 \|PD^tP^{-1}\|\|u_0\|_2 \leq \|P\|\|D^t\|\|P^{-1}\|\|u_0\|_2 \leq \|P\|\|P^{-1}\|\rho(M)^t\|u_0\|_2 = O((\rho(M))^t\|u_0\|_2).
  \]

- If $\rho(M) > 1$: We have for $t \geq 0$ and any $u_0 \in \mathbb{R}^m$,
  \[
  \|u_t\|_2 = \|M^t u_0\|_2
  \]
  But we know that there exist a $u_0 \in \mathbb{R}^m$ that only depends on $M$ such that $\|M^t u_0\|_2 = \|M^t\| \|u_0\|_2$ (explicitly $u_0$ is the eigenvector associated with the largest eigenvalue of $M^T M$). But, using (Bertsekas, 1999, Proposition A.15) we know that $\rho(M) \leq \|M\|_2$. Then we have that,
  \[
  \|u_t\|_2 \geq \rho(M)^t\|u_0\|_2
  \]

- If $|\lambda| = 1$, $\forall \lambda \in \text{Sp}(M)$, we can diagonalize $M$ such that $M = PDP^{-1}$ where $P$ is invertible and $D$ is a diagonal matrix with complex values of magnitude 1.
  We have for $t \geq 0$ and any $u_0 \in \mathbb{R}^m$,
  \[
  \|u_t\|_2 = \|M^t u_0\|_2 = \|PD^tP^{-1} u_0\|_2 \leq \|P\|\|D^t\|\|P^{-1}\|\|u_0\|_2 = \|P\|\|P^{-1}\|\|u_0\|_2
  \]
  Similarly,
  \[
  \|u_0\|_2 = \|M^{-t} u_t\|_2 = \|PD^{-t} P^{-1} u_t\|_2 \leq \|P\|\|D^{-t}\|\|P^{-1}\|\|u_t\|_2 = \|P\|\|P^{-1}\|\|u_t\|_2
  \]

\[\square\]

D PROOFS OF THE THEOREMS AND PROPOSITIONS

D.1 Proof of Thm. 1

Let us recall the Theorem proposed by Bertsekas (1999, Proposition 4.4.1). We also provide a convergence rate that was not previously stated in (Bertsekas, 1999).

Theorem 1. If the spectral radius $\rho_{\text{max}} := \rho(\nabla F_\phi(\omega^*)) < 1$, then, for $\omega_0$ in a neighborhood of $\omega^*$, the distance of $\omega_t$ to the stationary point $\omega^*$ converges at a linear rate of $O((\rho_{\text{max}} + \epsilon)^t)$, $\forall \epsilon > 0$.

Proof. For brevity let us write $x_t := (\phi_t, \theta_t)$ for $t \geq 0$ and $x^* := (\phi^*, \theta^*)$. Let $\epsilon > 0$.

By Proposition A.15 (Bertsekas, 1999) there exists a norm $\|\cdot\|$ such that its induced matrix norm has the following property:

\[
\|\nabla F_\phi(x^*)\| \leq \rho(\nabla F_\phi(x^*)) + \frac{\epsilon}{2}.
\]
Then by definition of the sequence \((x_t)\) and since \(x^*\) is a fixed point of \(F_\eta\), we have that,
\[
\|x_{t+1} - x^*\| = \|F_\eta(x_t) - F_\eta(x^*)\| \tag{48}
\]
Since \(F_\eta\) is assumed to be continuously differentiable by the mean value theorem we have that
\[
F_\eta(x_t) = F_\eta(x^*) + \nabla F_\eta(\tilde{x}_t)(x_t - x^*), \tag{49}
\]
for some \(\tilde{x}_t \in [x_t, x^*]\). Then,
\[
\|x_{t+1} - x^*\| \leq \|\nabla F_\eta(\tilde{x}_t)\|\|x_t - x^*\| \tag{50}
\]
where \(\|\nabla F_\eta(\tilde{x}_t)\|\) is the induced matrix norm of \(\| \cdot \|\).

Since the induced norm of a square matrix is continuous on its elements and since we assumed that \(\nabla F_\eta\) was continuous, there exists \(\delta > 0\) such that,
\[
\|\nabla F_\eta(x) - \nabla F_\eta(x^*)\| \leq \frac{\epsilon}{2}, \quad \forall x : \|x - x^*\| \leq \delta. \tag{51}
\]

Finally, we get that if \(\|x_t - x^*\| \leq \delta\), then,
\[
\|x_{t+1} - x^*\| \leq \|\nabla F_\eta(\tilde{x}_t)\|\|x_t - x^*\| \tag{52}
\leq (\|\nabla F_\eta(x^*)\| + \|\nabla F_\eta(\tilde{x}_t) - \nabla F_\eta(x^*)\|)\|x_t - x^*\| \tag{53}
\leq \left(\rho(\nabla F_\eta(x^*)) + \frac{\epsilon}{2} + \frac{\epsilon}{2}\right)\|x_t - x^*\| \tag{54}
\]
where in the last line we used (47) and (51). Consequently, if \(\rho(\nabla F_\eta(x^*)) < 1\) and if \(\|x_0 - x^*\| \leq \delta\), we have that,
\[
\|x_t - x^*\| \leq (\rho(\nabla F_\eta(x^*)) + \epsilon)^t\|x_0 - x^*\| \leq \delta, \quad \forall \epsilon > 0. \tag{55}
\]

\[\Box\]

### D.2 Proof of Thm. 2

We are interested in the optimal step-size for the Simultaneous gradient with no momentum. Define the step-size associated to one eigenvalue \(\lambda \in \mathbb{C}\) by \(\eta(\lambda) := \frac{\Re(\lambda)}{|\lambda|}\).

**Theorem’ 2.** If the eigenvalues of \(\nabla v(\varphi^*)\) all have a positive real-part, then, the best step-size \(\eta_{\text{best}}\), which minimizes the spectral radius \(\rho_{\text{max}}(\eta)\) of \(\nabla F_\eta(\varphi^*, \theta^*)\), is the solution of a (convex) quadratic by parts problem, and satisfies,

\[
\max_{1 \leq k \leq m} \sin(\psi_k)^2 \leq \rho_{\text{max}}(\eta_{\text{best}})^2 \leq 1 - \Re(\frac{1}{\Re(\lambda_1)}), \tag{56}
\]

with \(\delta := \min_{1 \leq k \leq m} |\lambda_k|^2(2\Re(\frac{1}{\Re(\lambda_k)}) - \Re(\frac{1}{\Re(\lambda_1)})\) \tag{57}
and \(\Re(\frac{1}{\Re(\lambda_1)}) \leq \eta_{\text{best}} \leq 2\Re(\frac{1}{\Re(\lambda_1)})\) \tag{58}

where \((\lambda_k = r_k e^{i\psi_k})_{1 \leq k \leq m} = \text{Sp}(\nabla v(\varphi^*, \theta^*))\) are sorted such that \(0 < \Re(\frac{1}{\Re(\lambda_1)}) \leq \cdots \leq \Re(\frac{1}{\Re(\lambda_m)})\). Particularly, when \(\eta_{\text{best}} = \Re(\frac{1}{\Re(\lambda_1)})\) we are in the case of the top plot of Fig.3 and \(\rho_{\text{max}}(\eta_{\text{best}})^2 = \sin(\psi_1)^2\).

**Proof.** The eigenvalues of \(\nabla F_\eta\) are \(1 - \eta \lambda\), for \(\lambda \in \text{Sp}(\nabla v(\varphi, \theta))\). Our goal is to solve
\[
\rho_{\text{max}} := \min_{\eta \geq 0} \max_{1 \leq k \leq m} |1 - \eta \lambda_k|^2 \tag{59}
\]
where \(\{\lambda_1, \ldots, \lambda_m\}\) is the spectrum of \(\nabla v(\varphi^*, \theta^*)\). We can develop the magnitude to get,
\[
f_1(\eta) := |1 - \eta \lambda_1|^2 = 1 - 2\eta \Re(\lambda_1) + \eta^2 |\lambda_1|^2 \tag{60}
\]
The function $\eta \mapsto \max_{1 \leq i \leq n} f_i(\eta)$ is a convex function quadratic by part. This function goes to $+\infty$ as $\eta$ gets larger, so it reaches its minimum over $[0, \infty)$. We can notice that each function $f_i$ reaches its minimum for $\eta_i = \frac{\Re(\lambda_i)}{|\lambda_i|^2} = \Re(1/\lambda_i)$. Consequently, if we order the eigenvalues such that,

$$\eta_1 \leq \ldots \leq \eta_m$$

we have that

$$f'_i(\eta_i) \leq 0, \quad 1 \leq i \leq m \quad \text{and} \quad f_1(x) = 1, \quad \forall x \geq 2\eta_1$$

As a result,

$$\eta_1 \leq \eta_{best} \leq 2\eta_1$$

Moreover, it is easy to notice that,

$$|1 - \eta_1 \lambda_1|^2 = \min_{\eta \geq 0} |1 - \eta \lambda_1|^2 \leq \min_{\eta \geq 0} \max_{1 \leq k \leq m} |1 - \eta \lambda_k|^2$$

Then developing $|1 - \eta_1 \lambda_1|^2$, we get that,

$$|1 - \eta_1 \lambda_1|^2 = |1 - \eta \lambda_1|^2 = 1 - \frac{\Re(\lambda_1)^2}{|\lambda_1|^2} = \sin(\psi_1)^2$$

where $\lambda_1 = r_1 e^{i\psi_1}$. Moreover, we also have that

$$\rho_{\max} = \min_{\eta \geq 0} \max_{1 \leq i \leq n} |1 - \eta \lambda_i|^2 \leq \max_{1 \leq k \leq m} |1 - \eta_1 \lambda_k|^2 = 1 - \eta_1 \min_{1 \leq k \leq m} 2\Re(\lambda_k) - \eta_1 |\lambda_k|^2 = 1 - \Re(1/\lambda_1)\delta$$

This upper bound is then achieved for $\eta = \Re(1/\lambda_1)$. Moreover is $\text{Sp}(\nabla v(\varphi^*, \theta^*)) \subset [\mu, L]$ we have that, $\lambda_1 = L$ and that

$$\delta \geq \min_{\lambda \in [\mu, L]} 2\lambda - \lambda^2/L = 2\mu - \mu^2/L$$

Consequently we recover the standard upper bound $\rho_{\max}^2 \leq 1 - 2\mu^2/L + \mu^2/L = (1 - \mu/L)^2$ provided in the convex case.

\[\square\]

### D.3 Proof of Thm. 3

We are now interested in the eigenvalues of the Simultaneous Gradient Method with Momentum.

**Theorem’ 3.** The eigenvalues of $\nabla F_{\eta, \beta}(\omega^*)$ are

$$\mu_{\pm}(\beta, \eta, \lambda) := (1 - \eta \lambda + \beta) \frac{1 \pm \Delta^{\frac{1}{2}}}{2},$$

where $\Delta := 1 - \frac{4\beta}{(1 - \eta \lambda + \beta)^2}$, $\lambda \in \text{Sp}(\nabla v(\omega^*))$ and $\Delta^{\frac{1}{2}}$ is the complex square root of $\Delta$ with positive real part\(^\dagger\). Moreover we have the following Taylor approximation,

$$\mu_+(\beta, \eta, \lambda) = 1 - \eta \lambda - \beta \frac{\eta \lambda}{1 - \eta \lambda} + O(\beta^2),$$

$$\mu_-(\beta, \eta, \lambda) = \beta \frac{1}{1 - \eta \lambda} + O(\beta^2).$$

**Proof.** The Jacobian of $F_{\eta, \beta}$ is

$$M := \begin{bmatrix} I_n - \eta \nabla v(\omega^*) + \beta I_n & -\beta I_n \\ I_n & 0_n \end{bmatrix}$$

Its characteristic polynomial can be written:

$$\chi_M(X) = \det(XI_{2n} - M) = \begin{vmatrix} (X - 1 - \beta)I_n + \eta T & \beta I_n \\ -I_n & XI_n \end{vmatrix}$$

\[^\dagger\] If $\Delta$ is a negative real number we set $\Delta^{\frac{1}{2}} := i\sqrt{-\Delta}$
where $\nabla v(\omega^*) = PTP^{-1}$ and $T$ is an upper-triangular matrix. Finally by Lemma 1 we have that,

$$\chi_M(X) = \left| X((X - 1 - \beta)I_n + \eta T) + \beta I_n \right| = \prod_{i=1}^{n} (X((X - 1 - \beta) + \eta e_i) + \beta) \quad (73)$$

where

$$T = \begin{bmatrix} \lambda_1 & * & \cdots & * \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}.$$ 

Let $\lambda$ one of the $\lambda_i$ we have,

$$X((X - 1 - \beta) + \eta \lambda) + \beta = X^2 - (1 - \eta \lambda + \beta)X + \beta \quad (74)$$

The roots of this polynomial are

$$\mu_+(\lambda) = \frac{1 - \eta \lambda + \beta + \sqrt{\Delta}}{2} \quad \text{and} \quad \mu_-(\lambda) = \frac{1 - \eta \lambda + \beta - \sqrt{\Delta}}{2} \quad (75)$$

where $\Delta := (1 - \eta \lambda + \beta)^2 - 4\beta$ and $\lambda \in \text{Sp}(\nabla v(\omega^*))$. This can be rewritten as,

$$\mu_{\pm}(\beta, \eta, \lambda) := (1 - \eta \lambda + \beta) \frac{1 \pm \sqrt{\Delta}}{2} \quad (76)$$

where $\Delta := 1 - \frac{4\beta}{(1 - \eta \lambda + \beta)^2}$, $\lambda \in \text{Sp}(\nabla v(\varphi^*, \theta^*))$ and $\Delta^{\frac{1}{2}}$ is the complex square root of $\Delta$ with real positive part (if $\Delta$ is a real negative number, we set $\Delta^{\frac{1}{2}} := i\sqrt{-\Delta}$). Moreover we have the following Taylor approximation,

$$\mu_{\pm}(\beta, \eta, \lambda) = 1 - \eta \lambda - \beta \frac{\eta \lambda}{1 - \eta \lambda} + O(\beta^2) \quad \text{and} \quad \mu_{\pm}(\beta, \eta, \lambda) = \frac{\beta}{1 - \eta \lambda} + O(\beta^2). \quad (77)$$

\[\square\]

\section*{D.4 Proof of Thm. 4}

We are interested in the impact of small Momentum values on the convergence rate of Simultaneous Gradient Method.

\textbf{Theorem’ 4.} \textit{For any $\lambda \in \text{Sp}(\nabla v(\omega^*))$ s.t. $\Re(\lambda) > 0$,}

$$\rho'_{\lambda, \eta}(0) > 0 \Leftrightarrow \eta \in I(\lambda) := \left( \frac{|\lambda| - |\Re(\lambda)|}{|\lambda | - |\Re(\lambda)|}, \frac{|\lambda| + |\Re(\lambda)|}{|\lambda | - |\Re(\lambda)|} \right).$$

Particularly, we have $\rho'_{\lambda, \Re(1/\lambda)}(0) = 2\Re(\lambda)\Re(1/\lambda) > 0$ and $|\text{Arg}(\lambda)| \geq \frac{\pi}{4} \Rightarrow (\Re(1/\lambda), 2\Re(1/\lambda)) \subset I(\lambda)$.

\textit{Proof.} Recall the definitions of $\mu_+$ and $\mu_-$ from Thm. 3, and the definition of the radius:

$$\rho_{\lambda, \eta}(\beta) := \max\{||\mu_+|^2, |\mu_-|^2\} \quad (78)$$

When $\beta$ is close to 0, $\mu_-$ is close also to 0 whereas $\mu_+$ is close to $1 - \eta \lambda$. In general $1 - \eta \lambda \neq 0$, so around 0, $\rho_{\lambda, \eta}(\beta) = |\mu_+(\beta)|^2 = \mu_+(\beta)\tilde{\mu}_+(\beta)$. The special case where $1 - \eta \lambda = 0$ is excluded from this analysis because it means that the eigenvalue $\lambda$ is not one constraining the learning rate as seen in Thm. 2. Computing the derivative
of $\rho$ give us

$$
\rho'_{\lambda,\eta}(0) = (\mu_+(0)\mu'_+(0) - \eta\lambda) \left[ 1 - \eta\lambda \right]
$$

(79)

$$
= 2\Re(\mu_+(0)\mu'_+(0))
$$

(80)

$$
= 2\Re \left( (1 - \eta\lambda) \frac{-\eta\lambda}{1 - \eta\lambda} \right)
$$

(81)

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

(82)

$$
= \frac{-2\eta}{|1 - \eta\lambda|^2} \left[ \Re(\lambda) - 2\eta|\lambda|^2 + \eta^2|\lambda|^2\Re(\lambda) \right]
$$

(83)

which leads to,

$$
\rho'_{\lambda,\eta}(0) = 2\frac{\eta^2|\lambda|^2 - \eta\Re(\lambda)(1 + \eta^2|\lambda|^2)}{|1 - \eta\lambda|^2}
$$

(84)

The sign of $\rho'_{\lambda,\eta}(0)$ is determined by the sign of

$$
2\eta|\lambda|^2 - \Re(\lambda)(1 + \eta^2|\lambda|^2) = -\Re(\lambda)|\lambda|^2\eta^2 + 2|\lambda|^2\eta - \Re(\lambda)
$$

(85)

This quadratic function is strictly positive on the open interval \((|\lambda| - |\Im(\lambda)|, |\lambda| + |\Im(\lambda)|)\).

Moreover since $\Re(1/\lambda) = \frac{\Re(\lambda)}{|\lambda|^2}$, we have that $|1 - \lambda\Re(1/\lambda)|^2 = 1 - \Re(\lambda)\Re(1/\lambda)$ (see Eq. 65) and then,

$$
\rho'_{\lambda,\Re(1/\lambda)}(0) = 2\Re(\lambda)\Re(1/\lambda).
$$

(86)

Finally writing $\lambda = re^{i\psi}$ we get that,

$$
\frac{|\lambda| - |\Im(\lambda)|}{|\lambda||\Re(\lambda)|} = \frac{1 - |\sin(\psi)|}{r\cos(\psi)} = \Re(1/\lambda) \frac{1 - |\sin(\psi)|}{1 - |\sin(\psi)|^2}
$$

and

$$
\frac{|\lambda| + |\Im(\lambda)|}{|\lambda||\Re(\lambda)|} = \frac{1 + |\sin(\psi)|}{r\cos(\psi)} = \Re(1/\lambda) \frac{1 + |\sin(\psi)|}{1 - |\sin(\psi)|^2}
$$

Consequently,

$$
J(\lambda) = \left( \frac{\Re(1/\lambda)}{1 + |\sin(\psi)|}, \frac{\Re(1/\lambda)}{1 - |\sin(\psi)|} \right)
$$

(87)

and $|\arg(\lambda)| \geq \frac{\pi}{4}$ implies that \(\left(\frac{2}{3}\Re(1/\lambda), 2\Re(1/\lambda)\right)\)

\(\square\)

D.5 Proof of Thm. 5

We are now in the special case of a bilinear game. We first consider the simultaneous gradient step with momentum. The operator $F_{\eta,\beta}$ is defined as:

$$
F_{\eta,\beta}^{\text{sim}} \begin{bmatrix} \theta_t \\ \varphi_t \\ \theta_{t-1} \\ \varphi_{t-1} \end{bmatrix} := \begin{bmatrix} \theta_t - \eta_1 A\varphi_t + \beta_1 (\theta_t - \theta_{t-1}) \\ \varphi_t + \eta_2 A^\top \theta_t + \beta_2 (\varphi_t - \varphi_{t-1}) \\ \theta_t \\ \varphi_t \end{bmatrix}.
$$

(88)

**Proposition’ 1.** The eigenvalues of $\nabla F_{\eta,\beta}^{\text{sim}}$ are the roots of the $4^{th}$ order polynomials:

$$
(x - 1)^2(x - \beta_1)(x - \beta_2) + \eta_1\eta_2\lambda x^2 + \lambda \in \text{Sp}(A^\top A).
$$

(89)

**Particularly,** when $\beta_1 = \beta_2 = 0$ and $\eta_1 = \eta_2 = \eta$ we have,

$$
P_\lambda(x) = x^2(x^2 - 2x + 1 + \eta^2\lambda), \lambda \in \text{Sp}(A^\top A)
$$

(90)
Proof. \(F^{\text{sim}}_{\eta, \beta}\) is a linear operator belonging to \(\mathbb{R}^{d \times p}\), for notational compactness let us call \(m := d + p\). Let us recall that \(I_m\) and \(0_{d,p}\) are respectively the identity of \(\mathbb{R}^{m \times m}\) and the zero matrix of \(\mathbb{R}^{d \times p}\).

\[
\nabla F^{\text{sim}}_{\eta, \beta} = \begin{bmatrix} I_m & 0_m \\ I_m & 0_m \end{bmatrix} + \begin{bmatrix} 0_d & -\eta_1 A & 0_p & 0_m \\ \eta_2 A^\top & 0_p & 0_m & 0_m \end{bmatrix} + \begin{bmatrix} \beta_1 I_d & 0_{d,p} & -\beta_1 I_d & 0_{d,p} \\ 0_{p,d} & \beta_2 I_p & 0_{p,d} & -\beta_2 I_p \\ 0_m & 0_m & 0_m & 0_m \end{bmatrix}
\]

(91)

Leading to the compressed form

\[
\nabla F^{\text{sim}}_{\eta, \beta} = \begin{bmatrix} (1 + \beta) I_d & -\eta A & -\beta_1 I_d & 0_{d,p} \\ \eta A^\top & (1 + \beta) I_p & 0_{p,d} & -\beta_2 I_p \\ I_m & 0_m & \beta_1 I_d & \beta_2 I_p \\ 0_m & 0_m & 0_m & X I_m \end{bmatrix}
\]

(92)

Then the characteristic polynomial of this matrix is equal to,

\[
\chi(X) := \begin{vmatrix} (X - 1 - \beta_1) I_d & \eta A & \beta_1 I_d & 0_{d,p} \\ -\eta A^\top & (X - 1 - \beta_2) I_p & 0_{p,d} & \beta_2 I_p \\ -I_m & \beta_1 I_d & \beta_2 I_p & X I_m \end{vmatrix}
\]

(93)

Then we can use Lemma 1 to compute this determinant,

\[
\chi(X) = \det \left( X \begin{bmatrix} (X - 1 - \beta_1) I_d & \eta A \\ -\eta_2 A^\top & (X - 1 - \beta_2) I_p \end{bmatrix} + \begin{bmatrix} \beta_1 I_d & 0_{d,p} \\ 0_{p,d} & \beta_2 I_p \end{bmatrix} \right)
\]

(94)

\[
= \begin{vmatrix} (X(X - 1 - \beta_1) + \beta_1) I_d & \eta_1 X A \\ -\eta_2 X A^\top & (X(X - 1 - \beta_2) + \beta_2) I_p \end{vmatrix}
\]

(95)

\[
= \begin{vmatrix} (X - \beta_1)(X - 1) I_d + \eta_1 \eta_2 \frac{X^2}{X(X - 1 - \beta_1) + \beta_2} A^\top A & \eta_1 X A \\ 0_{d,p} & (X - \beta_2)(X - 1) I_p \end{vmatrix}
\]

(96)

Where for the last equality we added to the first block column the second one multiplied by \(\eta_2 A^\top X(X - 1 - \beta_2) + \beta_2\).

It’s now time to introduce \(r\) the rank of \(A\). We can diagonalize \(A^\top A = U^\top \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0) U\) to get the determinant of a triangular matrix,

\[
\chi(X) = ((X - \beta_1)(X - 1))^d \cdot \prod_{k=1}^{r} ((X - \beta_1)(X - 1)(X - \beta_2)(X - 1) + \eta_1 \eta_2 X^2 \lambda_k)
\]

(97)

where \(\lambda_k\) are the positive eigenvalues of \(A^\top A\). This is the characteristic polynomial we were seeking, taking into account the null singular values of \(A\).

In particular, when \(\beta_1 = \beta_2 = 0\), we get,

\[
\chi(X) = X^m(X - 1)^{m - 2r} \prod_{k=1}^{r} ((X - 1)^2 + \eta_1 \eta_2 \lambda_k)
\]

(98)

\[\square\]

**Theorem’ 5.** For any \(\eta_1, \eta_2 \geq 0\) and \(\beta_1 = \beta_2 = \beta\), the iterates of the simultaneous methods (21) diverge as,

\[\Delta_t \in \begin{cases} \Omega(\Delta_0((1 + \beta)^2 + \eta^2 \sigma_{\text{max}}^2(A))^t) & \text{if } \beta \geq 0 \\ \Omega(\Delta_0(1 - \beta)^t) & \text{if } -\frac{1}{16} \leq \beta < 0. \end{cases}\]
Proof of Thm. 5. We report the maximum magnitudes of the eigenvalues of the polynomial from Prop. 1 in Fig. 7. We observe that they are larger than 1. We now prove it in several cases. Let us start with the simpler case $\beta_1 = \beta_2 = 0$. Using Lemma 2, there exists $(\theta^*, \varphi^*)$ such that for any $t \geq 0$,
\[ \theta_t - \theta^* \in \text{span}(A) = \text{span}(AA^\top) \quad \text{and} \quad \varphi_t - \varphi^* \in \text{span}(A^\top) = \text{span}(A^\top A), \quad t \geq 0 \] (99)

Then, we have,
\[ \|\theta_{t+1} - \theta^*\|^2 = \|\theta_t - \theta^* - 2\eta A(\varphi_t - \varphi^*)\|^2 \]
(100)
\[ = \|\theta_t - \theta^*\|^2 - 2\eta \langle \theta_t - \theta^*, A(\varphi_t - \varphi^*) \rangle + \eta^2 \|A(\varphi_t - \varphi^*)\|^2 \]
(101)
\[ \geq \|\theta_t - \theta^*\|^2 - 2\eta \langle \theta_t - \theta^*, A(\varphi_t - \varphi^*) \rangle + \eta^2 \sigma_{\min}^2(A) \|\varphi_t - \varphi^*\|^2 \]
(102)

where in line 1 we used that $A\varphi^* = 0$ and in line 3 we used that $\varphi_t - \varphi^*$ is orthogonal to the null space of $A$, so that we lower bound the product by the smallest non-zero singular value $\sigma_{\min}(A)$. The same way, we get:
\[ \|\varphi_{t+1} - \varphi^*\|^2 = \|\varphi - \varphi^*\|^2 + 2\eta \langle \theta_t - \theta^*, A(\varphi_t - \varphi^*) \rangle + 2\eta^2 \|A^\top(\theta - \theta^*)\|^2 \]
(103)
\[ \geq \|\varphi - \varphi^*\|^2 + 2\eta \langle \theta_t - \theta^*, A(\varphi_t - \varphi^*) \rangle + \eta^2 \sigma_{\min}^2(A) \|\theta_t - \theta^*\|^2 \]
(104)

Summing (102) and (104), we get
\[ \Delta_{t+1} \geq (1 + \eta^2 \sigma_{\min}^2(A)) \Delta_t \]
(105)

where $\sigma_{\min}^2(A)$ is the minimal (positive) squared singular value of $A$.

Now we can try to handle the case where $\beta_1 = \beta_2 = \beta \neq 0$. To prove Thm. 5 we will prove the following Proposition

Proposition 3. Let $\delta_{t,\eta,\beta}$ the operator defined in (21).

- For $\beta \geq 0$ its radial spectrum is lower bounded by $(1 + \beta^2 + \eta_1\eta_2\sigma_{\max}^2(A))$.
- For $-1/16 \leq \beta < 0$ its radial spectrum is lower bounded by $(1 - \beta^2)$.

Proof of Proposition 3. Let us use Proposition 1 to get that the eigenvalues of our linear operator are the solutions of
\[ (x - 1)^2(x - \beta)^2 + \eta^2\lambda x^2, \quad \lambda \in S_{\text{span}}(A^\top A). \]
(106)

Let us fix $\lambda > 0$ belonging to $S_{\text{span}}(A^\top A)$. For simplicity, let us note $\alpha^2 = \eta^2\lambda$. We can then notice that this polynomial can be factorized as
\[ (x - 1)^2(x - \beta)^2 + \alpha x = ((x - 1)(x - \beta) + i\alpha x)((x - 1)(x - \beta) - i\alpha x) \]
(107)

Then the roots of these 2 quadratic polynomials are
\[ z_1 = \frac{1 + \beta + i\alpha + ((1 + \beta + i\alpha)^2 - 4\beta)^{1/2}}{2}, \quad z_2 = \frac{1 + \beta + i\alpha - ((1 + \beta + i\alpha)^2 - 4\beta)^{1/2}}{2} \]
(108)
\[ z_3 = \frac{1 + \beta - i\alpha + ((1 + \beta - i\alpha)^2 - 4\beta)^{1/2}}{2} \quad \text{and} \quad z_4 = \frac{1 + \beta - i\alpha - ((1 + \beta - i\alpha)^2 - 4\beta)^{1/2}}{2}. \]
(109)

where $z^{1/2}$ is the complex square root of $z$ with positive imaginary part. Our goal is going to be to show that $x_1$ has a magnitude larger than 1.

Let us call $z := (1 + \beta + i\alpha)^2 - 4\beta = (1 - \beta)^2 - \alpha^2 + 2i\alpha(1 + \beta)$. Since $-1 \leq \beta \leq 1$, this complex has positive real part and imaginary part. Hence its square root too. Consequently, for $\beta \geq 0$,
\[ |z_1|^2 = \Re(z_1)^2 + \Im(z_1)^2 \geq (1 + \beta)^2 + \alpha^2 > 1 \]
(110)

For $\beta < 0$ the configuration is a bit more complex since $(1 + \beta)^2 + \alpha^2$ is smaller than 1 for small enough $\alpha$. We are going to use the fact that
\[ \Re(z^{1/2}) = \sqrt{\frac{|z| + \Re(z)}{2}} \quad \text{and} \quad \Im(z^{1/2}) = \sqrt{\frac{|z| - \Re(z)}{2}} \]
(111)
Then we have that,

\[ \Re(z^{1/2}) = \sqrt{\frac{\sqrt{(1 - \beta)^2 - \alpha^2}^2 + 4\alpha^2(1 + \beta)^2 + (1 - \beta)^2 - \alpha^2}{2}} \]

\[ \Re(z^{1/2}) = \sqrt{\frac{\sqrt{(1 - \beta)^2 + \alpha^2}^2 + 16\alpha \beta - (1 - \beta)^2 + \alpha^2}{2}} \]

\[ \Re(z^{1/2}) \geq \sqrt{\frac{(1 - \beta)^2 + \alpha^2 + 16 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2}}{2}} \]

\[ \Re(z^{1/2}) \geq 1 - \beta + 8 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2} \] (116)

where for the two inequalities we used \( \sqrt{1 + x} \leq 1 + x, \forall x \leq 0 \). With the same ideas we can lower bound the Imaginary part of \( z^{1/2} \),

\[ \Im(z^{1/2}) = \sqrt{\frac{\sqrt{(1 - \beta)^2 - \alpha^2}^2 + 4\alpha^2(1 + \beta)^2 - (1 - \beta)^2 + \alpha^2}{2}} \]

\[ \Im(z^{1/2}) = \sqrt{\frac{\sqrt{(1 - \beta)^2 + \alpha^2}^2 + 16\alpha \beta - (1 - \beta)^2 + \alpha^2}{2}} \]

\[ \Im(z^{1/2}) \geq \sqrt{\frac{(1 - \beta)^2 + \alpha^2 + 16 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2}}{2}} \]

\[ \Im(z^{1/2}) \geq \alpha + 8 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2} \] (121)

Consequently we can use (116) and (121) to lower bound the magnitude of \( z_1 \) (defined in Eq. 108) as,

\[ |z_1|^2 = \Re(z_1)^2 + \Im(z_1)^2 \]

\[ |z_1|^2 \geq \left( 1 - \beta + 4 \frac{\alpha^2 \beta}{(1 - \beta)(\alpha^2 + (1 - \beta)^2)} \right)^2 + \left( \alpha + 4 \frac{\alpha \beta}{\alpha^2 + (1 - \beta)^2} \right)^2 \]

\[ |z_1|^2 \geq (1 - \beta)^2 + 8 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2} + \alpha^2 + 8 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2} \]

\[ |z_1|^2 \geq (1 - \beta)^2 + \alpha^2 + 16 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2} \] (125)

For \(-1/16 \leq \beta < 0\) we have that \( \alpha^2 + 16 \frac{\alpha^2 \beta}{\alpha^2 + (1 - \beta)^2} \geq 0 \). Hence,

\[ |z_1|^2 \geq (1 - \beta)^2 = (1 + |\beta|)^2, \quad \forall -1/16 \leq \beta < 0 \] (126)

To conclude this proof we just need to combine Proposition 3 with Lemma 3 saying that if the spectral radius is strictly larger than 1 then the iterates diverge.
Particularly for $\beta_1 = \beta_2 = 0$ and $\eta_1 \eta_2 = \eta^2$ we get
\[ P_\lambda(x) = x^2((x-1)^2 + \eta^2 \lambda x^3), \ \lambda \in \text{Sp}(A^\top A) \]

(128)

Giving the following set of eigenvalues,
\[ \{0\} \cup \left\{ 1 + \eta \lambda \pm \sqrt{\eta^2 \lambda^2 - 4\lambda} \right\} : \lambda \in \text{Sp}(A^\top A) \]

(129)

Particularly for $\beta_1 = -\frac{1}{2}$ and $\beta_2 = 0$ we get
\[ x[(x-1)^2(x+\frac{1}{2}) + \eta^2 \lambda x^2], \ \lambda \in \text{Sp}(A^\top A) \]

(130)

proof of Proposition 2. Let us recall the definition of $F_{\eta,\beta}^{\text{alt}}$, (for compactness we note $F_{\eta,\beta}^{\text{alt}} = F_{\eta_1,\eta_2,\beta_1,\beta_2}^{\text{alt}}$)

\[
F_{\eta,\beta}^{\text{alt}} = \begin{bmatrix}
\theta_t \\
\phi_t \\
\theta_{t-1} \\
\phi_{t-1}
\end{bmatrix}
= \begin{bmatrix}
\theta_t - \eta_1 A \phi_t + \beta_1 (\theta_t - \theta_{t-1}) \\
\phi_t + \eta_2 A^\top (1 + \beta_2) (\theta_t - \phi_t) + \beta_1 (\theta_t - \phi_{t-1}) \\
\theta_t - \eta_1 A \phi_t + \beta_1 (\theta_t - \theta_{t-1}) \\
\phi_t
\end{bmatrix}
\]

(131)

(132)

Hence, the matrix $F_{\eta,\beta}^{\text{alt}}$ is,
\[
F_{\eta,\beta}^{\text{alt}} = \begin{bmatrix}
(1 + \beta_1) I_d & -\eta_1 A & -\beta_1 I_d & 0_{d,p} \\
(1 + \beta_1) \eta_2 A^\top & (1 + \beta_2) I_p - \eta_1 \eta_2 A^\top A & -\beta_1 \eta_2 A^\top & -\beta_2 I_p \\
I_m & 0_{m,d} & X I_d & 0_{d,p} \\
0_{d,p} & 0_{d,p} & 0_{p,d} & X I_p
\end{bmatrix}
\]

(133)

Then the characteristic polynomial of $F_{\eta,\beta}^{\text{alt}}$ is equal to
\[
\chi(X) = \begin{vmatrix}
(X - 1 - \beta_1) I_d & \eta_1 A & \beta_1 I_d & 0_{d,p} \\
-\eta_1 A & (X - 1 - \beta_2) I_p + \eta_1 \eta_2 A^\top A & 0_{d,p} & X I_d \\
0_{d,p} & 0_{d,p} & 0_{p,d} & X I_p
\end{vmatrix}
= \begin{vmatrix}
(X - 1 - \beta_1 + \frac{\beta_1}{X}) I_d & \eta_1 A & \beta_1 I_d & 0_{d,p} \\
-\eta_1 A & (X - 1 - \beta_2 + \frac{\beta_2}{X}) I_p + \eta_1 \eta_2 A^\top A & 0_{d,p} & X I_d \\
0_{d,p} & 0_{d,p} & 0_{p,d} & X I_p
\end{vmatrix}
\]

(134)

(135)

Where in the last line we added the third block column multiplied by $\frac{1}{X}$ to the first one. Then we have
\[
\chi(X) = \begin{vmatrix}
(X - 1 - \beta_1 + \frac{\beta_1}{X}) I_d & \eta_1 A & \beta_1 I_d & 0_{d,p} \\
-\eta_2 A^\top & (X - 1 - \beta_2 + \frac{\beta_2}{X}) I_p & 0_{d,p} & X I_d \\
0_{d,p} & 0_{d,p} & 0_{p,d} & X I_p
\end{vmatrix}
\]

(136)

where we added to the second block line the first block line by $-\eta_2 A^\top$. Then our determinant is triangular by
Negative Momentum for Improved Game Dynamics

squared blocks of size \( m \times m \) and we can write,

\[
\chi(X) = \det(XI_m) \begin{vmatrix} (X - 1 - \beta_1 + \frac{\eta_1}{X})I_d & \eta_1 A \\ -X\eta_2A^T & (X - 1 - \beta_2 + \frac{\eta_2}{X})I_p \end{vmatrix}
\]  

(137)

\[
= \begin{vmatrix} (X(X - 1 - \beta_1) + \beta_1)I_d & X\eta_1 A \\ -X^2\eta_2A^T & (X(X - 1 - \beta_2) + \beta_2)I_p \end{vmatrix}
\]  

(138)

\[
= \begin{vmatrix} (X - 1)(X - \beta_1)I_d & X\eta_1 A \\ -X^2\eta_2A^T & (X - 1)(X - \beta_2)I_p \end{vmatrix}
\]  

(139)

\[
= \begin{vmatrix} (X - 1)(X - \beta_1)I_d + \eta_1\eta_2A^T \frac{X^3}{(X - 1)(X - \beta_2)} & X\eta_1 A \\ 0_{p,d} & (X - 1)(X - \beta_2)I_p \end{vmatrix}
\]  

(140)

Now we can diagonalize \( A^*A \) to get,

\[
\chi(X) = (X - \beta_1)^{d-r}(X - \beta_2)^{p-r}(X - 1)^{p+r-2r} \prod_{k=1}^{r} \left( (X - 1)^2(X - \beta_2)(X - \beta_1) + \eta_1\eta_2X^2\lambda_k \right)
\]  

(141)

where \( (\lambda_k)_{1 \leq k \leq r} \) are the positive eigenvalues of \( A^*A \) of rank \( r \). Particularly, when \( \beta_1 = \beta_2 = 0 \) we have that,

\[
\chi(X) = X^m(X - 1)^{m-2r} \prod_{k=1}^{r} \left( (X - 1)^2 + \eta_1\eta_2X\lambda_k \right)
\]  

(142)

We report the maximum magnitudes of the eigenvalues of the polynomial from Prop. 2 in Fig. 7. We observe that they are smaller than 1 for a large choice of step-size and momentum values. This is a satisfying numerical result but we want analytical convergence rates. This is what we prove in Thm. 6.

**Theorem 6.** If we set \( \eta \leq \frac{1}{\sigma_{max}(A^*)}, \beta_1 = -\frac{1}{2} \) and \( \beta_2 = 0 \) then we have

\[
\Delta_{t+1} \in O \left( \max \left\{ \frac{1}{2}, 1 - \frac{\eta\sigma_{min}(A)}{4} \right\} t \Delta_0 \right)
\]  

(143)

If we set \( \beta_1 = 0 \) and \( \beta_2 = 0 \), then there exists \( M > 1 \) such that for any \( \eta_1, \eta_2 \geq 0 \), \( \Delta_t = O(\Delta_0) \).

**Proof.** In Lemma 2 we showed of the affine transformations \( \theta_t \rightarrow U(\theta_t - \theta^*) \) and \( \phi_t \rightarrow V(\phi_t - \phi^*) \) allow us to work on the span of a diagonal matrix \( D \). Then in that case the eigenspace of \( D \) do not interact with each other. In the sense that for each coordinate of \( [U(\theta_t - \theta^*)]_i \) and \( [V(\phi_t - \phi^*)]_i \) \( (1 \leq i \leq r) \) we have from (36) that

\[
\begin{align*}
[U(\theta_{t+1} - \theta^*)]_i &= [U(\theta_t - \theta^*)]_i - \eta_1\sigma_i[V(\phi_t - \phi^*)]_i + \beta_1[U(\theta_t - \theta_{t-1})]_i \\
[V(\phi_{t+1} - \phi^*)]_i &= [V(\phi_t - \phi^*)]_i + \beta_2[V(\phi_t - \phi_{t-1})]_i
\end{align*}
\]  

(144)

Consequently we only need to study the 4 dimensional linear operators

\[
\begin{bmatrix}
(1 + \beta_1) & -\eta_1\sigma_i & -\beta_1 & 0 \\
(1 + \beta_1)\eta_2\sigma_i & (1 + \beta_2) - \eta_1\eta_2\sigma_i^2 & -\beta_1\eta_2\sigma_i & -\beta_2 \\
I_2 & 0_2
\end{bmatrix}
\]  

(145)

for \( \sigma_1 \leq \cdots \leq \sigma_r > 0 \) the positive singular values of \( A \). These equations are a particular case of (133). Using the proof of Proposition 2 the eigenvalues of these matrices are the solution of

\[
P_i(X) = (X - 1)^2(X - \beta_1)(X - \beta_2) + \eta_1\eta_2X^2\sigma_i^2, \quad 1 \leq i \leq r.
\]  

(146)

We will now consider two case:
• When, $\beta_1 = \beta_2 = 0$ we have that,

$$P_i(X) = X^2((X - 1)^2 + \eta_1 \eta_2 X \sigma_i^2), \quad 1 \leq i \leq r. \quad (147)$$

Then the roots of $P_i(X)$ are 0 and two complex conjugate value with a magnitude equal to the constant term of $(X - 1)^2 + \eta_1 \eta_2 X \sigma_i^2$ which is 1. Since these two eigenvalues are different, the matrix (145) is diagonalizable (for $\beta_1 = \beta_2 = 0$ we can remove the state augmentation to only work with these two eigenvector). Consequently our linear operator is diagonalizable and has all its eigenvalues larger than 1 in magnitude, we can then apply Lemma 3 to conclude that $\Delta_t = \Omega(\Delta_0)$.

• When, $\beta_1 = -\frac{1}{2}$ and $\beta_2 = 0$, we have that $P_i(X) = XQ_i(X)$ where

$$Q_i(X) := (X - 1)^2(X + \frac{1}{2}) + \eta_1 \eta_2 X^2 \sigma_i^2, \quad 1 \leq i \leq r. \quad (148)$$

Then $P_i(-1/2) = \frac{-\eta_1 \eta_2}{4} > 0$ and $P_i(-1) = -2 + \eta_1 \eta_2 \sigma_i^2$. If $\eta_1 \eta_2 < \frac{2}{\sigma_i^2}$ we have $P_i(-1) < 0$. Consequently, this polynomial has a negative root $\lambda_-$ such that $-1 < \lambda_- < -\frac{1}{2} < 0$. Moreover the derivative of $Q_i(X)$ is

$$Q'_i(X) = (X - 1)(2X + 1) + (X - 1)^2 + 2\eta_1 \eta_2 X \sigma_i^2 = (3X - 3 - \eta_1 \eta_2 \sigma_i^2)X. \quad (149)$$

If $\eta_1 \eta_2 < \frac{3}{\sigma_i^2}$, then $Q'_i(x) > 0, \forall x > 0$. Since $Q_i(0) = 1/2 > 0$ then $Q_i(0) > 0, \forall x \geq 0$ and consequently all the real roots of $Q_i$ are negative.

Since by the root coefficient relationship the sum of the roots of $Q_i$ has to be equal to $\frac{3}{2} - \eta_1 \eta_2 \sigma_i^2 > 0$, all the roots of $Q_i$ cannot be real (because the real roots of $Q_i$ are negative). Hence $Q_i$ has two conjugate roots $\lambda_c$ and $\bar{\lambda}_c$ and one real negative root $\lambda_r$. Moreover the roots coefficient relationship are

$$\frac{3}{2} - \alpha = 2\Re(\lambda_c) + \lambda_r \quad (150)$$

$$0 = |\lambda_c|^2 + 2\lambda_r \Re(\lambda_c) \quad (151)$$

$$-\frac{1}{2} = \lambda_r |\lambda_c|^2 \quad (152)$$

where we called $\alpha = \eta_1 \eta_2 \sigma_i^2$. Plugging (150) into (151) we get

$$0 = |\lambda_c|^2 + (\frac{3}{2} - \alpha - \lambda_r)\lambda_r \quad (153)$$

Multiplying by $\lambda_r$ and plugging (152) in we get

$$\lambda_r^2 = \frac{1}{3 - 2\alpha - 2\lambda_r} \leq \frac{1}{4 - 2\alpha} \quad (154)$$

where we used that $\lambda_r < -\frac{1}{2}$. Consequently, since in the theorem we assumed that $\eta_1 \eta_2 \leq \frac{1}{\sigma_{\min}(A)^2}$, we have that $1 \leq \alpha$, we have

$$\lambda_r^2 \leq \frac{1}{2} \quad \text{and} \quad |\lambda_c|^2 \leq \frac{-\lambda_r}{2} \leq \sqrt{\frac{4 - 2\alpha}{2}} \leq 1 - \alpha/4 \quad (155)$$

where for the last inequality we used $\sqrt{1 + x} \leq 1 + \frac{x}{2}, \forall x \in \mathbb{R}$.

One last thing to say is that the four roots of $P_i$ which are the four eigenvalues of the matrix in (145) are different and consequently this matrix is diagonalizable.

We can then apply Lemma 3 in a case of a spectral radius strictly smaller that 1 to conclude that,

$$\Delta_{t+1} \leq \max\{1/2, 1 - \eta_1 \eta_2 \sigma_{\min}(A)\} \Delta_t \quad (156)$$

where,

$$\Delta_t := \|U(\theta_{t+1} - \theta^*)\|^2 + \|U(\theta_t - \theta^*)\|^2 + \|V(\varphi_{t+1} - \varphi^*)\|^2 + \|V(\varphi_t - \varphi^*)\|^2 \quad (157)$$

$$= \|\theta_{t+1} - \theta^*\|^2 + \|\theta_t - \theta^*\|^2 + \|\varphi_{t+1} - \varphi^*\|^2 + \|\varphi_t - \varphi^*\|^2 \quad (158)$$

because $U$ and $V$ are orthogonal.