Negative Momentum for Improved Game Dynamics

Gauthier Gidel* **Reyhane Askari Hemmat*** **Mohammad Pezeshki** **Gabriel Huang**
Rémi Lepriol **Simon Lacoste-Julien†** **Ioannis Mitliagkas**
Mila & DIRO, Université de Montréal

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 are more complex and less understood. In this paper, we analyze gradient-based methods with momentum on simple games. Next, we show empirically that alternating gradient updates with a negative momentum term achieves convergence on the notoriously difficult to train saturating GANs.

1 Introduction and Background

Recent advances in machine learning are largely driven by the success of gradient-based optimization methods for the training process. Games generalize the standard optimization framework by introducing different objective functions for different optimizing agents, known as *players*. For example generative adversarial networks (GANs) [Goodfellow et al., 2014] use a two-player game formulation. We are commonly interested in finding a local *Nash equilibrium*: a set of parameters from which no player can (locally and unilaterally) improve its objective function. Games with differentiable objectives often proceed by simultaneous or alternating gradient steps on the players' objectives.

In this work we are interested in studying the effect of two particular algorithmic choices: (i) the choice between simultaneous and alternating updates, and (ii) the choice of step size and momentum value. We summarize our main contributions as follows: We show that the alternating gradient method with negative momentum is the only setting within our study parameters (Tab. 1) 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 also show in §3 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. Finally, we confirm the benefits of negative momentum for training GANs with the notoriously ill-behaved saturating loss on CIFAR10 and FASHION-MNIST datasets.

Related work A lot of work has been done in the context of understanding momentum and its variants but all are restricted to minimization problems [Polyak, 1964, Qian, 1999, Nesterov, 2013, Sutskever et al., 2013]. Balduzzi et al. [2018] develop new methods to understand the dynamics of general games and propose Symplectic Gradient Adjustment which helps in discovering stable fixed points in general games. Mescheder et al. [2017] show how presence of eigenvalues with zero real-part and large imaginary-part result in oscillatory behavior. Nagarajan and Kolter [2017] analyze the local stability of GANs and show that during training of a GAN, the eigenvalues of the Jacobian of the vector field are pushed away from one along the real axis.

*Equal contribution, correspondence to <gauthier.gidel,reyhane.askari.hemmat>@umontreal.ca
†CIFAR fellow

Extended abstract. Smooth Games Optimization and Machine Learning Workshop (NIPS 2018), Montréal, Canada. Longer version available: https://arxiv.org/abs/1807.04740

Method	β	Bounded	Converges
Simultaneous	$\beta \in \mathbb{R}$	✗	✗
Alternated	>0	✗	✗
	0	✓	✗
	<0	✓	✓

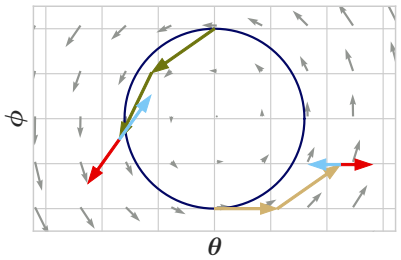

Figure 1: Effect of Gradient Methods on an unconstrained bilinear example: $\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} \ \boldsymbol{\theta}^\top \boldsymbol{A} \boldsymbol{\varphi}$.

Game theory formulation of GANs From a game theory point of view, GAN training is a differentiable two-player game: the discriminator $D_{\boldsymbol{\varphi}}$ aims at minimizing its cost function $\mathcal{L}^{(\boldsymbol{\varphi})}$ and the generator $G_{\boldsymbol{\theta}}$ aims at minimizing its own cost function $\mathcal{L}^{(\boldsymbol{\theta})}$,

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\theta}, \boldsymbol{\varphi}^*) \quad \text{and} \quad \boldsymbol{\varphi}^* \in \arg\min_{\boldsymbol{\varphi} \in \boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\theta}^*, \boldsymbol{\varphi}) \,. \tag{1}$$

State $(\boldsymbol{\varphi}^*, \boldsymbol{\theta}^*)$. In order to analyze the dynamics of gradient-based methods near a Nash Equilibrium, we look at the *gradient vector field* and its associated *Jacobian* $\nabla \boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta})$,

$$\boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta}) := \begin{bmatrix} \nabla_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \\ \nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \end{bmatrix} \quad \nabla \boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta}) := \begin{bmatrix} \nabla^2_{\boldsymbol{\varphi}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) & \nabla_{\boldsymbol{\varphi}} \nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\varphi})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \\ \nabla_{\boldsymbol{\varphi}} \nabla_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\varphi}, \boldsymbol{\theta})^T & \nabla^2_{\boldsymbol{\theta}} \mathcal{L}^{(\boldsymbol{\theta})}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \end{bmatrix} . \tag{2}$$

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

$$F_\eta(\boldsymbol{\varphi}, \boldsymbol{\theta}) := \begin{bmatrix} \boldsymbol{\varphi} & \boldsymbol{\theta} \end{bmatrix}^\top - \eta \, \boldsymbol{v}(\boldsymbol{\varphi}, \boldsymbol{\theta}) \,, \quad (\boldsymbol{\varphi}, \boldsymbol{\theta}) \in \mathbb{R}^m \,, \tag{3}$$

where η is the learning rate. Now, for brevity we write the joint parameters $\boldsymbol{\omega} := (\boldsymbol{\varphi}, \boldsymbol{\theta}) \in \mathbb{R}^m$. For $t \in \mathbb{N}$, let $\boldsymbol{\omega}_t = (\boldsymbol{\varphi}_t, \boldsymbol{\theta}_t)$ be the t^{th} point of the sequence computed by the gradient method. If the gradient method converges, its limit point $\boldsymbol{\omega}^* = (\boldsymbol{\varphi}^*, \boldsymbol{\theta}^*)$ is a *fixed point* of F_η. In addition, if $\nabla v(\boldsymbol{\omega}^*)$ is positive-definite, then $\boldsymbol{\omega}^*$ is a local Nash equilibrium.

2 Tuning the Step Size

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

Theorem 1 (Prop. 4.4.1 Bertsekas [1999])**.** *If the spectral radius $\rho_{\max} := \rho(\nabla F_\eta(\boldsymbol{\omega}^*)) < 1$, then, for $\boldsymbol{\omega}_0$ in a neighborhood of $\boldsymbol{\omega}^*$, the distance of $\boldsymbol{\omega}_t$ to the stationary point $\boldsymbol{\omega}^*$ converges at a linear rate of $\mathcal{O}\big((\rho_{\max} + \epsilon)^t\big)$, $\forall \epsilon > 0$.*

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

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

$$\max_{1 \le k \le m} \sin(\psi_k)^2 \le \rho_{\max}(\eta_{best})^2 \le 1 - \Re(1/\lambda_1)\delta \quad \textit{and} \quad \Re(1/\lambda_1) \le \eta_{best} \le 2\Re(1/\lambda_1) \,, \tag{4}$$

where $\delta := \min_{1 \le k \le m} |\lambda_k|^2 (2\Re(1/\lambda_k) - \Re(1/\lambda_1))$ and $(\lambda_k = r_k e^{i\psi_k})_{1 \le k \le m} = \mathrm{Sp}(\nabla \boldsymbol{v}(\boldsymbol{\varphi}^*, \boldsymbol{\theta}^*))$ are sorted such that $0 < \Re(1/\lambda_1) \le \cdots \le \Re(1/\lambda_m)$. Particularly, when $\eta_{best} = \Re(1/\lambda_1)$ we are in the case of the top plot of Fig.2 and $\rho_{\max}(\eta_{best})^2 = \sin(\psi_1)^2$.

When $\nabla \boldsymbol{v}$ is positive-definite, the η_{best} is achieved because of either one or several limiting eigenvalues, we illustrate and interpret these two cases in Fig. 2. In (4) we can see that if the Jacobian of $\boldsymbol{v}$ has an almost purely imaginary eigenvalue then the convergence rate of the gradient method can be arbitrarily close to 1. Our goal is to use momentum to wrangle game dynamics into convergence.

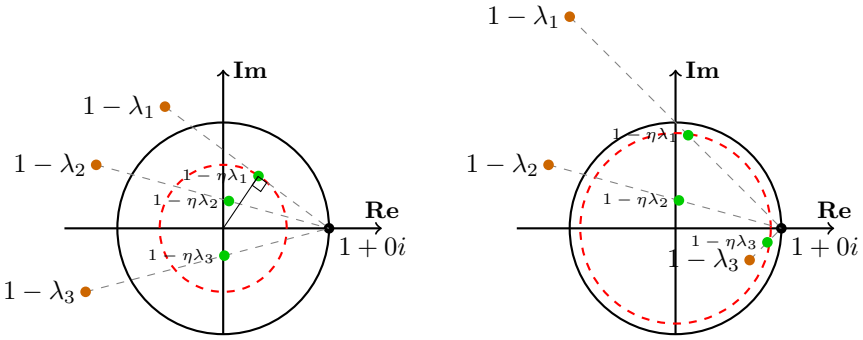

Figure 2: Trajectories of $1-\eta\lambda_i$ for growing step sizes, and the optimal step-size for $\lambda_i \in \mathrm{Sp}(\nabla \boldsymbol{v}(\boldsymbol{\phi}^*, \boldsymbol{\theta}^*))$. The unit circle is drawn in **black**, and the **red** dashed circle has radius equal to the largest eigenvalue $\mu_{\max}$, which is directly related to the convergence rate. Therefore, smaller red circles mean better convergence rates. **Left:** The red circle is limited by the tangent trajectory line $1-\eta\lambda_1$, which means the best convergence rate is limited only by the eigenvalue which will pass furthest from the origin as η grows, i.e., $\lambda_1 = \arg\min \frac{|\Re(\lambda_i)|}{|\lambda_i|^2}$. **Right:** The red circle is cut (not tangent) by the trajectories at points $1-\eta\lambda_1$ and $1-\eta\lambda_3$. The η is optimal because any increase in η will push the eigenvalue λ_1 out of the red circle, while any decrease in step-size will retract the eigenvalue λ_3 out of the red circle, which will lower the convergence rate in any case.

3 Negative Momentum

As shown in (4), the presence of eigenvalues with large imaginary parts can restrict us to using small step sizes and lead to slow convergence rates. In order to improve convergence, we add a *negative* momentum term into the update rule. The new momentum term leads to a modification of the *parameter update operator* $F_\eta(\boldsymbol{\omega})$ of (3). It requires to augment the state $\boldsymbol{\omega}_t$ with the previous iterate $\boldsymbol{\omega}_{t-1}$ (similar to Zhang and Mitliagkas [2017]) to form a compound state $(\boldsymbol{\omega}_t, \boldsymbol{\omega}_{t-1}) := (\boldsymbol{\varphi}_t, \boldsymbol{\theta}_t, \boldsymbol{\varphi}_{t-1}, \boldsymbol{\theta}_{t-1}) \in \mathbb{R}^{2m}$. The update rule (3) turns into the following,

$$F_{\eta,\beta}(\boldsymbol{\omega}_t, \boldsymbol{\omega}_{t-1}) = (\boldsymbol{\omega}_{t+1}, \boldsymbol{\omega}_t) \quad \text{where} \quad \boldsymbol{\omega}_{t+1} := \boldsymbol{\omega}_t - \eta \boldsymbol{v}(\boldsymbol{\omega}_t) + \beta(\boldsymbol{\omega}_t - \boldsymbol{\omega}_{t-1})\,, \tag{5}$$

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

$$\nabla F_{\eta,\beta}(\boldsymbol{\omega}_t, \boldsymbol{\omega}_{t-1}) = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0}_n \\ \boldsymbol{I}_n & \boldsymbol{0}_n \end{bmatrix} - \eta \begin{bmatrix} \nabla \boldsymbol{v}(\boldsymbol{\omega}_t) & \boldsymbol{0}_n \\ \boldsymbol{0}_n & \boldsymbol{0}_n \end{bmatrix} + \beta \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{I}_n \\ \boldsymbol{0}_n & \boldsymbol{0}_n \end{bmatrix} \tag{6}$$

In the following theorem, we provide an explicit equation for the eigenvalues of the Jacobian of $F_{\eta,\beta}$.

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

$$\mu_\pm(\beta, \eta, \lambda) := (1 - \eta\lambda + \beta)\frac{1 \pm \Delta^{\frac{1}{2}}}{2}\,, \quad \lambda \in \mathrm{Sp}(\nabla \boldsymbol{v}(\boldsymbol{\omega}^*))\,, \tag{7}$$

where $\Delta := 1 - \frac{4\beta}{(1-\eta\lambda+\beta)^2}$ *and* $\Delta^{\frac{1}{2}}$ *is the* complex *square root of* Δ *with positive real part*[3].

When β is small enough, Δ is a complex number close to 1. Consequently, μ_+ is close to the original eigenvalue for gradient dynamics $1-\eta\lambda$, and μ_-, the eigenvalue introduced by the state augmentation, is close to 0. Since our goal is to minimize the largest magnitude of the eigenvalues of $F_{\eta,\beta}$ computed in Thm. 3, we want to understand the effect of β on these eigenvalues with potential large magnitude. Let $\lambda \in \mathrm{Sp}(\nabla \boldsymbol{v}(\boldsymbol{\omega}^*))$, we define the squared magnitude $\rho_{\lambda,\eta}(\beta)$ we want to optimize,

$$\rho_{\lambda,\eta}(\beta) := \max\left\{|\mu_+(\beta, \eta, \lambda)|^2, |\mu_-(\beta, \eta, \lambda)|^2\right\}. \tag{8}$$

We study the local behavior of $\rho_{\lambda,\eta}$ for small values of β. The following theorem shows that a well suited β decreases $\rho_{\lambda,\eta}$, which corresponds to faster convergence.

Theorem 4. *For any* $\lambda \in \mathrm{Sp}(\nabla \boldsymbol{v}(\boldsymbol{\omega}^*))$ *s.t.* $\Re(\lambda) > 0$,

$$\rho'_{\lambda,\eta}(0) > 0 \Leftrightarrow \eta \in I(\lambda) := \left(\frac{|\lambda| - \Im(\lambda)}{|\lambda|\Re(\lambda)}, \frac{|\lambda| + \Im(\lambda)}{|\lambda|\Re(\lambda)}\right). \tag{9}$$

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

[3] If Δ is a negative real number we set $\Delta^{\frac{1}{2}} := i\sqrt{-\Delta}$

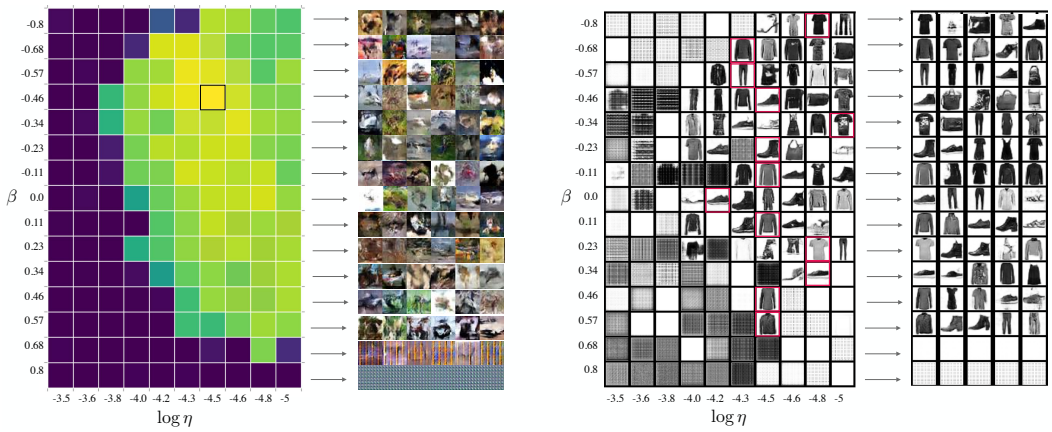

Figure 3: Comparison between negative and positive momentum on GANs with saturating loss on CIFAR-10 (left) and on Fashion MNIST (right) using a residual network. For each dataset, a grid of different values of momentum (β) and step sizes (η) is provided which describes the discriminator's settings while a constant momentum of 0.5 and step size of 10^{-4} is used for the generator. Each cell in CIFAR-10 (or Fashion MNIST) grid contains a single configuration in which its color (or its content) indicates the inception score (or a single sample) of the model. For CIFAR-10 experiments, yellow is higher while blue is the lower inception score. Along each row, the best configuration is chosen and more samples from that configuration are presented on the right side of each grid.

As we have seen previously in Fig. 2 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 λ_1 (since in that case the optimal step-size is $\eta_{best} = \Re(1/\lambda_1) \in I(\lambda_1)$) or when there are several limiting eigenvalues $\lambda_1, \ldots, \lambda_k$ and the intersection $I(\lambda_1) \cap \ldots \cap I(\lambda_k)$ 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 λ_1 is larger than $\pi/4$ then by (4), our theorem provides that the optimal step-size η_{best} belongs to $I(\lambda_1)$. Nevertheless, we numerically optimized (8) with respect to β and η and found that for any non-imaginary fixed eigenvalue λ, the optimal β is negative and the associated optimal step size is larger than $\Re(1/\lambda)$.

4 Experiments and Discussion

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. Fig. 3 shows the results. We use residual networks for both the generator and the discriminator with no batch-normalization. Each residual block is made of two 3×3 convolution layers with *ReLU* activation function. Up-sampling and down-sampling layers are respectively used in the generator and discriminator. We observe that using a negative value generally result in samples with higher quality and inception scores. We use negative momentum only on the discriminator because intuitively, negative momentum slows down the learning process of the discriminator and allows for better gradient flow to the generator.

5 Conclusion

We theoretically show that alternating updates with negative momentum is the only method within our study parameters (Tab.1) that converges in bilinear smooth games. Next, we study the effect of using negative values of momentum in a GAN setup both theoretically and experimentally and show that negative momentum can improve the convergence rate of gradient-based methods by shifting the eigenvalues of the Jacobian appropriately into a smaller convergence disk. We test out theory on CIFAR-10 and fashion MNIST datasets. We believe that our results explain a decreasing trend in momentum values reported in GAN literature in the past few years. Some of the most successful papers use zero momentum [Arjovsky et al., 2017, Gulrajani et al., 2017] for architectures that would otherwise call for high momentum values in a non-adversarial setting.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The mechanics of n-player differentiable games. In *ICML*, 2018.

D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein GANs. In *NIPS*, 2017.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

L. Mescheder, S. Nowozin, and A. Geiger. The numerics of GANs. In *NIPS*, 2017.

V. Nagarajan and J. Z. Kolter. Gradient descent GAN optimization is locally stable. In *NIPS*, 2017.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 1964.

N. Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 1999.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

J. Zhang and I. Mitliagkas. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471*, 2017.