Last-iterate convergence rates for min-max optimization

Jacob Abernethy[*]
Georgia Institute of Technology
prof@gatech.edu

Kevin A. Lai[*]
Georgia Institute of Technology
kevinlai@gatech.edu

Andre Wibisono[*]
Georgia Institute of Technology
wibisono@gatech.edu

Abstract

While classic work in convex-concave min-max optimization relies on average-iterate convergence results, the emergence of nonconvex applications such as training Generative Adversarial Networks has led to renewed interest in last-iterate convergence guarantees. Proving last-iterate convergence is challenging because many natural algorithms, such as Simultaneous Gradient Descent/Ascent, provably diverge or cycle even in simple convex-concave min-max settings, and previous work on global last-iterate convergence rates has been limited to the bilinear and convex-strongly concave settings. In this work, we show that the HAMILTONIAN GRADIENT DESCENT (HGD) algorithm achieves linear convergence in a variety of more general settings, including convex-concave problems that satisfy a "sufficiently bilinear" condition. We also prove similar convergence rates for some parameter settings of the Consensus Optimization (CO) algorithm of [MNG17].

1 Introduction

In this paper we consider methods to solve smooth unconstrained min-max optimization problems. In the most classical setting, a min-max objective has the form

$$\min_{x_1} \max_{x_2} g(x_1, x_2)$$

where $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a smooth objective function with two inputs. The usual goal in such problems is to find a saddle point, also known as a *min-max solution*, which is a pair $(x_1^*, x_2^*) \in \mathbb{R}^d \times \mathbb{R}^d$ that satisfies $g(x_1^*, x_2) \le g(x_1^*, x_2^*) \le g(x_1, x_2^*)$ for every $x_1 \in \mathbb{R}^d$ and $x_2 \in \mathbb{R}^d$. Min-max problems have a long history, going back at least as far as [Neu28], which formed the basis of much of modern game theory, and including a great deal of work in the 1950s when algorithms such as *fictitious play* were explored [Bro51, Rob51].

The *convex-concave* setting, where we assume g is convex in x_1 and concave in x_2, is a classic min-max problem that has a number of different applications, such as solving constrained convex optimization problems. One popular approach to solving such problems is to use *no-regret* learning algorithms [CBL06, Haz16]. If no-regret learning algorithms are used to iteratively update the inputs x_1 and x_2, one can prove that the average-iterates $(\bar{x}_1, \bar{x}_2)$ converge to a min-max solution [Han57, FS99].

[*] Author order is alphabetical and all authors contributed equally.

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

Recently, interest in min-max optimization has surged due to the enormous popularity of Generative Adversarial Networks (GANs), whose training involves solving a nonconvex min-max problem where x_1 and x_2 correspond to the parameters of two different neural nets [GPAM$^+$14]. The fundamentally nonconvex nature of this problem changes two things. First, it is infeasible to find a "global" solution of the min-max objective, and one instead seeks a suitably-defined local version of a min-max. Second, iterate averaging lacks the theoretical guarantees present in the convex-concave setting. This has motivated research on *last-iterate* convergence guarantees, which are appealing because they more easily carry over from convex to nonconvex settings.

Last-iterate convergence guarantees for min-max problems have been challenging to prove since standard analysis of no-regret algorithms says essentially nothing about last-iterate convergence. Widely used no-regret algorithms, such as Simultaneous Gradient Descent/Ascent (SGDA), fail to converge even in the simple *bilinear* setting where $g(x_1, x_2) = x_1^\top C x_2$ for some arbitrary matrix C. SGDA provably cycles in continuous time and diverges in discrete time (see for example [DISZ18, MGN18]). In fact, the full range of Follow-The-Regularized-Leader (FTRL) algorithms provably do not converge in zero-sum games with interior equilibria [MPP18]. This occurs because the iterates of the FTRL algorithms exhibit cyclic behavior, a phenomenon commonly observed when training GANs in practice as well.

Much of the recent research on last-iterate convergence in min-max problems has focused on *asymptotic* or *local* convergence [MLZ$^+$19, MNG17, DP18, BRM$^+$18, LFB$^+$19, MJS19]. While these results are certainly useful, one would ideally like to prove *global non-asymptotic* last-iterate convergence rates. Provable global convergence rates allow for quantitative comparison of different algorithms and can aid in choosing learning rates and architectures to ensure fast convergence in practice. Yet despite the extensive amount of literature on convergence rates for convex optimization, very few global last-iterate convergence rates have been proved for min-max problems. Existing work on global last-iterate convergence rates has been limited to the *bilinear* or *convex-strongly concave* settings [Tse95, LS19, DH19, MOP19]. In particular, the following basic question is still open:

"What global last-iterate convergence rates are achievable for convex-concave min-max problems?"

Our Contribution We give a partial answer for this question by proving linear last-iterate convergence rates for an algorithm called HAMILTONIAN GRADIENT DESCENT (HGD) under weaker assumptions compared to previous results. HGD is gradient descent on the squared norm of the gradient, and it has been mentioned in [MNG17, BRM$^+$18]. Our results are the first to show non-asymptotic convergence of an efficient algorithm in settings that not linear or strongly convex in either input. In particular, we introduce a novel "sufficiently bilinear" condition on the second-order derivatives of the objective g and show that this condition is sufficient for HGD to achieve linear convergence in convex-concave settings. The "sufficiently bilinear" condition appears to be a new sufficient condition for linear convergence rates that is distinct from previously known conditions such as the Polyak-Łojasiewicz (PL) condition or pure bilinearity. Our analysis relies on showing that the squared norm of the gradient satisfies the PL condition in various settings. As a corollary of this result, we can leverage [KNS16] to show that a stochastic version of HGD will have a last-iterate convergence rate of $O(1/\sqrt{k})$ in the "sufficiently bilinear" setting. We also show convergence rates for some parameter settings of the Consensus Optimization algorithm of [MNG17], which has been shown to have good practical performance in training GANs.

1.1 Preliminaries

Since g is a function of $x_1 \in \mathbb{R}^d$ and $x_2 \in \mathbb{R}^d$, we will often consider x_1 and x_2 to be components of one vector $x = (x_1 \ , \ x_2)$. We will use superscripts to denote iterate indices. Following [BRM$^+$18], we use $\xi = (\frac{\partial g}{\partial x_1}, -\frac{\partial g}{\partial x_2})$ to denote the signed vector of partial derivatives. We will use J to denote the Jacobian of ξ, i.e. $J \equiv \nabla \xi$. Note that unlike the Hessian in standard optimization, J is not symmetric, due to the negative sign in ξ. When clear from the context, we often omit dependence on x when writing $\xi, J, g, \mathcal{H}$, and other functions. Note that ξ, J, and $\mathcal{H}$ are defined for a given objective g – we omit this dependence as well for notational clarity. We will always assume g is sufficiently differentiable whenever we take derivatives.

Our main proofs will rely on the well-known Polyak-Łojasiewicz (PL) condition, defined as follows:

Definition 1.1 (Polyak-Łojasiewicz (PL) condition [Pol63, Loj63]). *A function $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfies the PL condition with parameter $\mu > 0$ if for all $x \in \mathbb{R}^d$, $\frac{1}{2} \left\|\nabla f(x)\right\|^2 \geq \mu(f(x) - \min_{x^* \in \mathbb{R}^d} f(x^*))$.*

If a smooth function f satisfies the PL condition, then gradient descent on f converges at a linear rate, as shown in the following classic theorem:

Theorem 1.2 (Linear rate under PL [Pol63, Loj63])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and let $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x)$. Suppose f satisfies the PL condition with parameter μ. Then if we run gradient descent from $x^{(0)} \in \mathbb{R}^d$ with step-size $\frac{1}{L}$, we have: $f(x^{(k)}) - f(x^*) \leq (1 - \frac{\mu}{L})^k (f(x^{(0)}) - f(x^*))$.*

Notions of convergence in min-max problems The convergence rates in this paper will apply to min-max problems where g satisfies the following:

Assumption 1.3. *All critical points of the objective g are global min-maxes.*

In other words, we prove convergence rates to min-maxes in settings where convergence to critical points is necessary and sufficient for convergence to min-maxes. Assumption 1.3 holds for convex-concave settings, but also holds for some nonconvex-nonconcave settings. This assumption allows us to measure convergence of our algorithms to *ϵ-approximate critical points*, defined as follows:

Definition 1.4. *Let $\epsilon \geq 0$. A point $x \in \mathbb{R}^d \times \mathbb{R}^d$ is an ϵ-approximate critical point if $||\xi(x)|| \leq \epsilon$.*

Convergence to approximate critical points is a necessary condition for convergence in min-max optimization, and it is a natural measure of convergence since the value of g at a given point gives no information about how close we are to a min-max.

2 Hamiltonian Gradient Descent

Our main algorithm for finding saddle points of $g(x_1, x_2)$ is called HAMILTONIAN GRADIENT DESCENT (HGD), which consists of performing gradient descent on a particular objective function $\mathcal{H}$ that we refer to as the *Hamiltonian*,[2] defined as follows:

$$\mathcal{H}(x) := \tfrac{1}{2}\|\xi(x)\|^2 = \tfrac{1}{2}\left(\|\tfrac{\partial g}{\partial x_1}(x)\|^2 + \|\tfrac{\partial g}{\partial x_2}(x)\|^2\right).$$

Since a critical point occurs when $\xi(x) = 0$, we can find a (approximate) critical point by finding a (approximate) minimizer of $\mathcal{H}$. Moreover, under Assumption 1.3, finding a critical point is equivalent to finding a saddle point. This motivates the HGD update procedure on $x^{(k)} = (x_1^{(k)}, x_2^{(k)})$ with step-size $\eta > 0$:

$$x^{(k+1)} = x^{(k)} - \eta \nabla \mathcal{H}(x^{(k)}), \tag{1}$$

HGD has been mentioned in [MNG17, BRM$^+$18], and it strongly resembles the Consensus Optimization (CO) approach of [MNG17]. The HGD update requires a Hessian-vector product because $\nabla \mathcal{H} = \xi^\top J$, making HGD a second-order iterative scheme. However, Hessian-vector products are cheap to compute when the objective is defined by a neural net, taking only two gradient oracle calls [Pea94]. This makes the Hessian-vector product oracle a theoretically appealing primitive, and it has been used widely in nonconvex min-max optimization [MNG17, BRM$^+$18, ADLH19, LFB$^+$19, MJS19].

3 Results

We now state our main theorems for this paper, which show convergence to critical points. When Assumption 1.3 holds, we get convergence to min-maxes. All of our main results will use the following multi-part assumption:

Assumption 3.1. *Let $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Assume a critical point for g exists. Moreover, assume that for all $x \in \mathbb{R}^d$, $||\xi(x)|| \leq L_1$ and $||\nabla\xi(x)|| \leq L_2$, and for all $x, y \in \mathbb{R}^d$, $||\nabla\xi(x) - \nabla\xi(y)|| \leq L_3 ||x - y||$. Let $L_{\mathcal{H}} = L_1 L_3 + L_2$.*

Our main results are Theorems 3.2 to 3.4, which show that HGD converges when 1) g is strongly convex-strongly concave 2) g is linear in one input and the cross-derivative is full-rank 3) g is smooth in both inputs and has a large, well-conditioned cross-derivative.

[2] Although we call $\mathcal{H}$ the Hamiltonian to match the terminology of [BRM$^+$18], we note that $\mathcal{H}$ is not the Hamiltonian as in the sense of classical physics, as we do not use the symplectic structure in our analysis. Rather we only perform gradient descent on $\mathcal{H}$.

Theorem 3.2. *Let Assumption 3.1 hold and let $g(x_1, x_2)$ be α-strongly convex in x_1 and α-strongly concave in x_2. Then HGD with $\eta = 1/L_{\mathcal{H}}$ starting from some $x^{(0)} \in \mathbb{R}^d \times \mathbb{R}^d$ will have the following convergence rate:* $\left|\left|\xi(x^{(k)})\right|\right| \leq \left(1 - \frac{\alpha^2}{L_{\mathcal{H}}}\right)^{k/2} \left|\left|\xi(x^{(0)})\right|\right|.$

Theorem 3.3. *Let Assumption 3.1 hold and let $g(x_1, x_2)$ be L-smooth in x_1 and linear in x_2, and assume the cross derivative $\nabla^2_{x_1,x_2} g$ is full rank with all singular values at least $\gamma > 0$ for all $x \in \mathbb{R}^d \times \mathbb{R}^d$. Then HGD with $\eta = 1/L_{\mathcal{H}}$ starting from some $x^{(0)} \in \mathbb{R}^d \times \mathbb{R}^d$ will have the following convergence rate:* $\left|\left|\xi(x^{(k)})\right|\right| \leq \left(1 - \frac{\gamma^4}{(2\gamma^2+L^2)L_{\mathcal{H}}}\right)^{k/2} \left|\left|\xi(x^{(0)})\right|\right|.$

Theorem 3.4. *Let Assumption 3.1 hold and let g be L-smooth in x_1 and L-smooth in x_2. Let $\mu^2 = \min_{x_1,x_2} \lambda_{\min}((\nabla^2_{x_2x_2} g(x_1, x_2))^2)$ and $\rho^2 = \min_{x_1,x_2} \lambda_{\min}((\nabla^2_{x_1x_1} g(x_1, x_2))^2)$, and assume the cross derivative $\nabla^2_{x_1x_2} g$ is full rank with all singular values lower bounded by $\gamma > 0$ and upper bounded by Γ for all $x \in \mathbb{R}^d \times \mathbb{R}^d$. Moreover, let the following "sufficiently bilinear" condition hold:*

$$(\gamma^2 + \rho^2)(\mu^2 + \gamma^2) - 4L^2\Gamma^2 > 0. \tag{2}$$

Then HGD with $\eta = 1/L_{\mathcal{H}}$ starting from some $x^{(0)} \in \mathbb{R}^d \times \mathbb{R}^d$ will satisfy

$$\left|\left|\xi(x^{(k)})\right|\right| \leq \left(1 - \frac{(\gamma^2+\rho^2)(\gamma^2+\mu^2)-4L^2\Gamma^2}{(2\gamma^2+\rho^2+\mu^2)L_{\mathcal{H}}}\right)^{k/2} \left|\left|\xi(x^{(0)})\right|\right|. \tag{3}$$

As discussed above, Theorem 3.4 provides the first last-iterate convergence rate for min-max problems that does not require strong convexity or linearity in either input. For example, the objective $g(x_1, x_2) = f(x_1) + 3Lx_1{}^\top x_2 - h(x_2)$, where f and h are L-smooth convex functions, satisfies the assumptions of Theorem 3.4 and is not strongly concave or linear in either input. One can also construct simple examples that are not convex-concave. Note that the "sufficiently bilinear" condition (2) is crucial for the linear rate, as linear convergence is impossible for general convex-concave problems due to lower bounds for convex optimization. In simple experiments for HGD on convex-concave and nonconvex-nonconcave objectives, the convergence rate speeds up when there is a larger bilinear component, as expected from our theoretical results.

Proof sketch for Theorems 3.2 to 3.4. The proofs for Theorems 3.2 to 3.4 all involve showing that $\mathcal{H}$ satisfies the Polyak-Łojasiewicz (PL) condition with some parameter c that depends on the setting. Since HGD is gradient descent on $\mathcal{H}$, this implies that HGD converges to a critical point at a linear rate. To show that $\mathcal{H}$ satisfies the PL condition, we show that it suffices to lower bound the eigenvalues of $JJ^\top$. The rest of the proofs consist of bounding the eigenvalues of $JJ^\top$ in different settings. □

Stochastic HGD Since we show that $\mathcal{H}$ satisfies the PL condition with parameter c in different settings, we can use Theorem 4 in [KNS16] to show that stochastic HGD converges at a $O(1/\sqrt{k})$ rate in the settings of Theorems 3.2 to 3.4, including the "sufficiently bilinear" setting:

Theorem 3.5. *Let Assumption 3.1 hold and suppose $\mathcal{H}$ satisfies the PL condition with parameter c. Suppose we use the update $x^{(k+1)} = x^{(k)} - \eta_k v(x^{(k)})$, where v is a stochastic estimate of $\nabla\mathcal{H}$ such that $\mathrm{E}[v] = \nabla\mathcal{H}$ and $\mathrm{E}[\|v(x^{(k)})\|^2] \leq C^2$ for all $x^{(k)}$. Then if we use $\eta_k = \frac{2k+1}{2c(k+1)^2}$, we have the following convergence rate:* $\mathrm{E}[\|x^{(k)}\|] \leq \sqrt{\frac{L_{\mathcal{H}}C^2}{kc^2}}$.

3.1 Consensus Optimization

We can also show convergence rates for the Consensus Optimization (CO) algorithm of [MNG17]:

$$x^{(k+1)} = x^{(k)} - \eta(\xi(x^{(k)}) + \gamma\nabla\mathcal{H}(x^{(k)})) \tag{4}$$

where $\gamma > 0$. [MNG17] show that CO can effectively train GANs in a variety of settings, including on CIFAR-10 and celebA. When γ is large, the CO update is close to that of HGD, which allows us to prove convergence rates for the large γ parameter regime of CO:

Theorem 3.6. *Let Assumption 3.1 hold. Let g be L_g smooth and suppose $\mathcal{H}$ satisfies the PL condition with parameter c. Then the CO algorithm (4) starting at some $x^{(0)} \in \mathbb{R}^d \times \mathbb{R}^d$ with step-size $\eta = \frac{c}{4L_{\mathcal{H}}L_g}$ and $\gamma = \frac{4L_g}{c}$ has the following convergence rate:* $\left|\left|\xi(x^{(k)})\right|\right| \leq \left(1 - \frac{c}{4L_{\mathcal{H}}}\right)^{k} \left|\left|\xi(x^{(0)})\right|\right|.$

Previously, CO only had provable convergence rates in the bilinear setting [LS19], so our result greatly expands the settings where CO has provable non-asymptotic convergence.

References

[ADLH19] Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. In *Artificial Intelligence and Statistics (AISTATS)*, 2019.

[BRM$^+$18] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning (ICML)*, 2018.

[Bro51] George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[DH19] Simon S Du and Wei Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *Artificial Intelligence and Statistics (AISTATS)*, 2019.

[DISZ18] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations (ICLR)*, 2018.

[DP18] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9255–9265, 2018.

[FS99] Yoav Freund and Robert E. Schapire. Adaptive Game Playing Using Multiplicative Weights. *Games and Economic Behavior*, 29(1-2):79–103, October 1999.

[GPAM$^+$14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[Han57] James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

[Haz16] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[LFB$^+$19] Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*, 2019.

[Loj63] Lojasiewicz. A topological property of real analytic subsets (in french). *Coll. du CNRS, Les équations aux dérivées partielles*, page 87–89, 1963.

[LS19] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *Artificial Intelligence and Statistics (AISTATS)*, 2019.

[MGN18] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, pages 3478–3487, 2018.

[MJS19] Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

[MLZ$^+$19] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *International Conference on Learning Representations (ICLR)*, 2019.

[MNG17] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1825–1835, 2017.

[MOP19] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extragradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

[MPP18] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2703–2717. SIAM, 2018.

[Neu28] J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[Pea94] Barak A Pearlmutter. Fast exact multiplication by the hessian. *Neural computation*, 6(1):147–160, 1994.

[Pol63] B. T. Polyak. Gradient methods for minimizing functionals (in russian). *Zh. Vychisl. Mat. Mat. Fiz.*, page 643–653, 1963.

[Rob51] Julia Robinson. An iterative method of solving a game. *Annals of mathematics*, pages 296–301, 1951.

[Tse95] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.