Last-iterate convergence rates for min-max optimization

Jacob Abernethy, Kevin A. Lai, Andre Wibisono
Georgia Institute of Technology

PROBLEM SETTING: MIN-MAX OPTIMIZATION

An unconstrained min-max optimization problem is written as:

$$\min_{x_1 \in \mathbb{R}^d} \max_{x_2 \in \mathbb{R}^d} g(x_1, x_2)$$

where $g: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a smooth function.

Goal: Find $(x_1^*, x_2^*) \in \mathbb{R}^d \times \mathbb{R}^d$ such that $\forall x_1 \in \mathbb{R}^d$ and $\forall x_2 \in \mathbb{R}^d$:

$$g(x_1^*, x_2) \leq g(x_1^*, x_2^*) \leq g(x_1, x_2^*).$$

- Classic result: *Average iterate* of no-regret algorithms like Simultaneous Gradient Descent/Ascent (SGDA) converges to a min-max in *convex-concave* problems.
- Modern applications such as GANs involve *non-convex* min-max problems, for which averaging no longer gives the same guarantees.
- On the other hand, *last-iterate* guarantees transfer more readily to the non-convex setting.

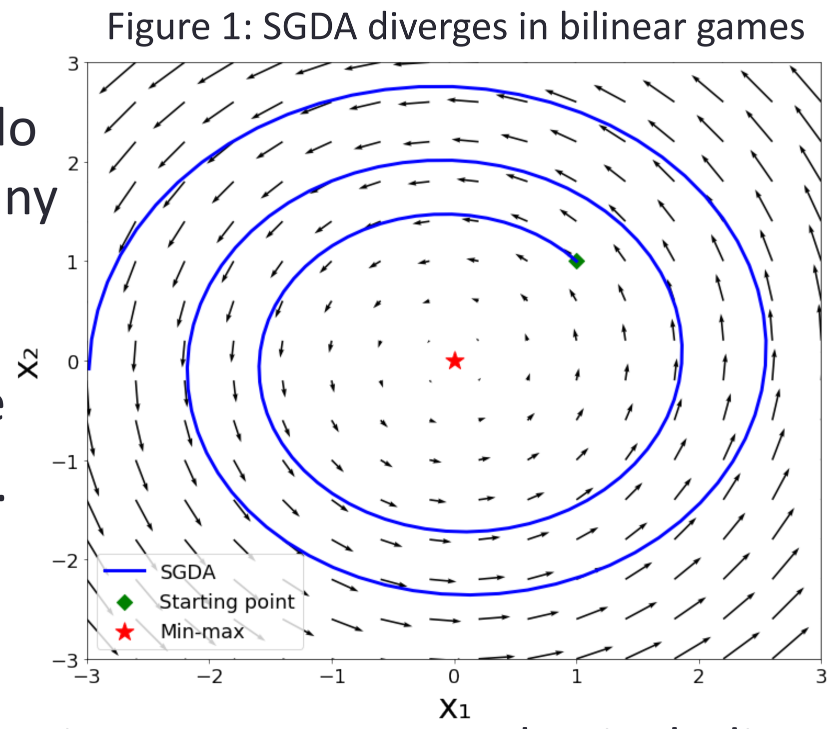
Question: What last-iterate convergence rates are possible for the convex-concave setting?

LAST-ITERATE CONVERGENCE

Last-iterate convergence is tricky!

- [MPP18] All FTRL algorithms provably do not have last-iterate convergence in many cases ⇒ standard algorithms like SGDA can't be used.
- SGDA diverges even in the bilinear case where $g(x_1, x_2) = x_1^\top x_2$ (see Figure 1).

Figure 1: SGDA diverges in bilinear games



Existing results are limited

- Many recent works give local or asymptotic convergence results, including in nonconvex-nonconcave settings.
- Global convergence rates have only been proven in very limited settings:
  - [LS18] show convergence in the bilinear case for various algorithms as well as convergence for SGDA in the strongly convex-strongly concave case.
  - [DH19] show convergence for SGDA in a specific case where g is strongly convex in x_1 and concave in x_2.

Prior to our work, no global last-iterate convergence rates existed beyond bilinear or strongly convex/PL settings!

Note: Concurrent work by [AMLJG19] shows global linear convergence rates for various algorithms in a very similar setting to ours.

CONVERGENCE DEFINITION AND ASSUMPTIONS

For $x = (x_1, x_2)$, let $\xi(x) \coloneqq \left(\nabla_{x_1} g(x_1, x_2), -\nabla_{x_2} g(x_1, x_2)\right)^\top$.

Assumption 1: $\nabla^2 g$ is bounded and Lipschitz (i.e. g is sufficiently smooth).

Assumption 2: All critical points are min-maxes (true for convex-concave g).

Definition of Convergence: We measure convergence rates in terms of $\|\xi\|$.

HAMILTONIAN GRADIENT DESCENT

As in [BRMFTG18], we define the Hamiltonian $\mathcal{H}(x) \coloneqq \frac{1}{2}\|\xi(x)\|^2$.

Our main algorithm is **Hamiltonian Gradient Descent** (HGD), defined as:

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

- Note that $\nabla \mathcal{H} = \nabla \xi^\top \xi$ and that under Assumption 1, $\mathcal{H}$ is smooth over the algorithm's iterates. Let $L_{\mathcal{H}}$ be the smoothness constant of $\mathcal{H}$.
- HGD is a second-order algorithm, but can be implemented with Hessian-vector products, which are as fast as gradients for neural networks.

MAIN RESULT

We show that HGD achieves a linear rate under a novel "sufficiently bilinear" condition. This demonstrates a new setting where linear rates are possible.

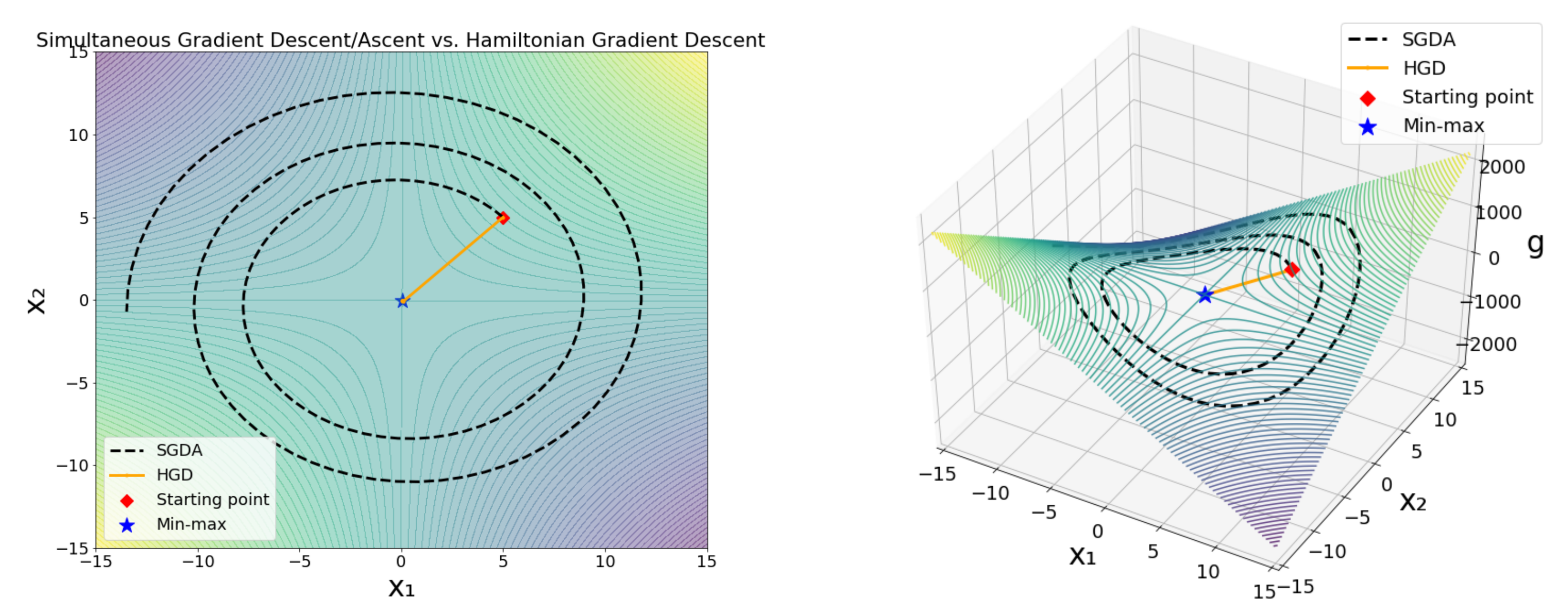


Figure 1: SGDA vs. HGD for $g(x_1, x_2) = f(x_1) + 10 x_1^\top x_2 - f(x_2)$ where $f(x) = \log(1 + e^x)$. SGDA slowly circles away from the min-max, while HGD goes directly to the min-max.

Main Theorem: For all $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$, let

- $\lambda\left(\left(\nabla^2_{x_1 x_1} g\right)^2\right) \in [\rho^2, L^2]$, $\lambda\left(\left(\nabla^2_{x_2 x_2} g\right)^2\right) \in [\mu^2, L^2]$
- $\sigma\left(\nabla^2_{x_1 x_2} g(x_1, x_2)\right) \in [\gamma, \Gamma]$ for $\gamma > 0$.

Assume the following "sufficiently bilinear" condition holds:

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

Then HGD with $\eta = 1/L_{\mathcal{H}}$ has the following convergence rate:

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

We also show results for a stochastic variant of HGD provided that the stochastic gradient is bounded over all iterates.

SUFFICIENTLY BILINEAR CONDITION

The "sufficiently bilinear" condition is

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

- This is a property of $\nabla^2 g$ and is satisfied if $\gamma = \Gamma$ and $\gamma \geq 4L$.
- Note that in the bilinear case $g(x_1, x_2) = x_1^\top x_2$, so $L = 0$.
- Satisfied for functions $g(x_1, x_2) = f(x_1) - 3L x_1^\top x_2 - h(x_2)$ for L-smooth convex functions f and h with Lipschitz Hessian.

CONSENSUS OPTIMIZATION

Our results imply a linear convergence rate for some parameter regimes of the Consensus Optimization (CO) algorithm of [MNG17], defined as:

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

- [MNG17] show CO can train GANs effectively in practice for $\gamma = 10$.
- We show that CO with large enough γ converges at the same rate as HGD (up to constants) and in the same settings.

ANALYSIS

We show that the Hamiltonian satisfies the Polyak-Łojasiewicz (PL) condition, which implies linear convergence of gradient descent.

Lemma: If $\nabla \xi \nabla \xi^\top \succcurlyeq \alpha I$, then $\mathcal{H}$ satisfies the PL condition with parameter α.

References:
- [AMLJG19] Azizian, Mitliagkas, Lacoste-Julien, and Gidel. A Tight and Unified Analysis of Extragradient for a Whole Spectrum of Differentiable Games. Preprint. 2019.
- [BRMFTG18] Balduzzi, Racaniere, Martens, Foerster, Tuyls, and Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning (ICML)*, 2018.
- [DH19] Du and Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. *AISTATS* 2019.
- [LS19] Liang and Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *AISTATS* 2019.
- [MPP18] Mertikopoulos, Papadimitriou, and Piliouras. Cycles in adversarial regularized learning. *SODA* 2018.
- [MNG17] Mescheder, Nowozin, and Geiger. The numerics of GANs. *NeurIPS* 2017.