Stochastic Hamiltonian Gradient Methods for Smooth Games

Nicolas Loizou

joint work with Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent†, Simon Lacoste-Julien†, Ioannis Mitliagkas†.

— : — : — : —

ICML 2020

— : — : — : —

† Canada CIFAR AI Chair

Overview

The Min-Max Optimization Problem

Problem: Stochastic Smooth Game.

$$\min_{x_1 \in \mathbb{R}^{d_1}} \max_{x_2 \in \mathbb{R}^{d_2}} g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} g_i(x_1, x_2) \tag{1}$$

where $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is a smooth objective.

Goal: Find Min-max solution / Nash Equilibrium.

Find $x^* = (x_1^*, x_2^*) \in \mathbb{R}^d$ such that, for every $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$,

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

Appears in many applications:

- Domain Generalization (Albuquerque et al., 2019)
- Generative Adversarial Networks (GANs) (Goodfellow et al., 2014)
- Formulations in Reinforcement Learning (Pfau, Vinyals, 2016)

Related Work

- Deterministic Games:
  Last-iterate convergence guarantees. Classic results (Korpelevich, 1976; Nemirovski, 2004) and recent results (Mescheder et al., 2017; Daskalakis et al., 2017; Gidel et al., 2018b; Azizian et al., 2019).

Related Work

- <u>Deterministic Games:</u>
  Last-iterate convergence guarantees. Classic results (Korpelevich, 1976; Nemirovski, 2004) and recent results (Mescheder et al., 2017; Daskalakis et al., 2017; Gidel et al., 2018b; Azizian et al., 2019).
- <u>Stochastic Games:</u>
  Convergent methods rely on **iterate averaging over compact domains** (Nemirovski, 2004).
  Palaniappan & Bach, 2016 and Chavdarova et al., 2019 proposed methods with last-iterate convergence guarantees over a non-compact domain but under **strong monotonicity assumption.**

Related Work

- Deterministic Games:
  Last-iterate convergence guarantees. Classic results (Korpelevich, 1976; Nemirovski, 2004) and recent results (Mescheder et al., 2017; Daskalakis et al., 2017; Gidel et al., 2018b; Azizian et al., 2019).
- Stochastic Games:
  Convergent methods rely on **iterate averaging over compact domains** (Nemirovski, 2004).
  Palaniappan & Bach, 2016 and Chavdarova et al., 2019 proposed methods with last-iterate convergence guarantees over a non-compact domain but under **strong monotonicity assumption.**
- Second-Order Methods:
  Consensus optimization method (Mescheder et al., 2017) and Hamiltonian gradient descent (Balduzzi et al., 2018; Abernethy et al., 2019). **No available analysis for the stochastic problem.**

Main Contributions

1. **First global non-asymptotic last-iterate convergence guarantees** in the stochastic setting (without assuming strong monotonicity or bounded domain) including a class of non-convex non-concave games.

Main Contributions

1. **First global non-asymptotic last-iterate convergence guarantees** in the stochastic setting (without assuming strong monotonicity or bounded domain) including a class of non-convex non-concave games.
2. **First convergence analysis of stochastic Hamiltonian methods** for solving min-max problems. Existing papers on these methods are empirical (Mescheder et al. 2017, Balduzzi et al. 2018).

Main Contributions

1. **First global non-asymptotic last-iterate convergence guarantees** in the stochastic setting (without assuming strong monotonicity or bounded domain) including a class of non-convex non-concave games.
2. **First convergence analysis of stochastic Hamiltonian methods** for solving min-max problems. Existing papers on these methods are empirical (Mescheder et al. 2017, Balduzzi et al. 2018).
3. A **novel unbiased estimator** of the Hamiltonian gradient. Crucial point for proving convergence for the proposed methods (existing methods use biased estimators).

Main Contributions

1. **First global non-asymptotic last-iterate convergence guarantees** in the stochastic setting (without assuming strong monotonicity or bounded domain) including a class of non-convex non-concave games.
2. **First convergence analysis of stochastic Hamiltonian methods** for solving min-max problems. Existing papers on these methods are empirical (Mescheder et al. 2017, Balduzzi et al. 2018).
3. A **novel unbiased estimator** of the Hamiltonian gradient. Crucial point for proving convergence for the proposed methods (existing methods use biased estimators).
4. First stochastic **Hamiltonian variance reduced method** (linear convergence guarantees).

Main Contributions

1. **First global non-asymptotic last-iterate convergence guarantees** in the stochastic setting (without assuming strong monotonicity or bounded domain) including a class of non-convex non-concave games.
2. **First convergence analysis of stochastic Hamiltonian methods** for solving min-max problems. Existing papers on these methods are empirical (Mescheder et al. 2017, Balduzzi et al. 2018).
3. A **novel unbiased estimator** of the Hamiltonian gradient. Crucial point for proving convergence for the proposed methods (existing methods use biased estimators).
4. First stochastic **Hamiltonian variance reduced method** (linear convergence guarantees).

Hamiltonian Perspective: Popular stochastic optimization algorithms can be used as methods for solving stochastic min-max problems.

Smooth Games and Hamiltonian Gradient Descent

$$\min_{x_1 \in \mathbb{R}^{d_1}} \max_{x_2 \in \mathbb{R}^{d_2}} g(x_1, x_2) \tag{2}$$

$$x = (x_1, x_2)^\top \in \mathbb{R}^d \quad \xi(x) = \begin{pmatrix} \nabla_{x_1} g \\ -\nabla_{x_2} g \end{pmatrix} \quad \mathbf{J} = \nabla \xi = \begin{pmatrix} \nabla^2_{x_1,x_1} g & \nabla^2_{x_1,x_2} g \\ -\nabla^2_{x_2,x_1} g & -\nabla^2_{x_2,x_2} g \end{pmatrix}$$

Vector $x^* \in \mathbb{R}^d$ is a **stationary point** when $\xi(x^*) = 0$.

Key Assumption:

All stationary points of the objective g are global min-max solutions.

Hamiltonian Gradient Descent (HGD) (Balduzzi et al., 2018)

$$\min_x \quad \mathcal{H}(x) = \frac{1}{2}\|\xi(x)\|^2. \tag{3}$$

HGD can be expressed using a Jacobian-vector product:

$$x^{k+1} = x^k - \eta_k \nabla \mathcal{H}(x) = x^k - \eta_k \left[J^\top \xi \right]$$

Stochastic Hamiltonian Function

$$\min_{x_1 \in \mathbb{R}^{d_1}} \max_{x_2 \in \mathbb{R}^{d_2}} g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} g_i(x_1, x_2) \tag{4}$$

$$\xi_i(x) = \begin{pmatrix} \nabla_{x_1} g_i \\ -\nabla_{x_2} g_i \end{pmatrix} \quad \mathbf{J} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{J}_i, \quad \text{where } \mathbf{J}_i = \begin{pmatrix} \nabla^2_{x_1,x_1} g_i & \nabla^2_{x_1,x_2} g_i \\ -\nabla^2_{x_2,x_1} g_i & -\nabla^2_{x_2,x_2} g_i \end{pmatrix}.$$

Finite-Sum Structure Hamiltonian Function

$$\mathcal{H}(x) = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathcal{H}_{i,j}(x) \quad \text{where} \quad \mathcal{H}_{i,j}(x) = \frac{1}{2} \langle \xi_i(x), \xi_j(x) \rangle \tag{5}$$

Algorithms use gradient of only one component function $\mathcal{H}_{i,j}(x)$:

$$\nabla \mathcal{H}_{i,j}(x) = \frac{1}{2} \left[\mathbf{J}_i^\top \xi_j + \mathbf{J}_j^\top \xi_i \right]. \tag{6}$$

Unbiased estimator of the $\nabla \mathcal{H}(x)$. That is, $\mathbb{E}_{i,j} \left[\nabla \mathcal{H}_{i,j}(x) \right] = \nabla \mathcal{H}(x)$.

Stochastic Bilinear Games.

$$g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} x_1^\top b_i + x_1^\top \mathbf{A}_i x_2 + c_i^\top x_2 \tag{7}$$

Stochastic sufficiently bilinear games. (Abernethy et al., 2019)
Games where the following condition is true:

$$(\delta^2 + \rho^2)(\delta^2 + \beta^2) - 4L^2\Delta^2 > 0, \tag{8}$$

where $0 < \delta \leq \sigma_i \left(\nabla^2_{x_1,x_2} g\right) \leq \Delta$, $\rho^2 = \min_{x_1,x_2} \lambda_{\min} \left[\nabla^2_{x_1,x_1} g(x_1, x_2)\right]^2$ and $\beta^2 = \min_{x_1,x_2} \lambda_{\min} \left[\nabla^2_{x_2,x_2} g(x_1, x_2)\right]^2$.

Classes of Stochastic Smooth Games

Stochastic Bilinear Games.

$$g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} x_1^\top b_i + x_1^\top \mathbf{A}_i x_2 + c_i^\top x_2 \tag{7}$$

Proposition: Stochastic bilinear game (7) $\Rightarrow$ Stochastic Hamiltonian function (5) is a smooth quadratic quasi-strongly convex function.

Stochastic sufficiently bilinear games. (Abernethy et al., 2019)
Games where the following condition is true:

$$(\delta^2 + \rho^2)(\delta^2 + \beta^2) - 4L^2\Delta^2 > 0, \tag{8}$$

where $0 < \delta \leq \sigma_i \left(\nabla^2_{x_1,x_2} g\right) \leq \Delta$, $\rho^2 = \min_{x_1,x_2} \lambda_{\min} \left[\nabla^2_{x_1,x_1} g(x_1,x_2)\right]^2$ and $\beta^2 = \min_{x_1,x_2} \lambda_{\min} \left[\nabla^2_{x_2,x_2} g(x_1,x_2)\right]^2$.

Classes of Stochastic Smooth Games

Stochastic Bilinear Games.

$$g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} x_1^\top b_i + x_1^\top \mathbf{A}_i x_2 + c_i^\top x_2 \quad (7)$$

Proposition: Stochastic bilinear game (7) $\Rightarrow$ Stochastic Hamiltonian function (5) is a smooth quadratic quasi-strongly convex function.

Stochastic sufficiently bilinear games. (Abernethy et al., 2019)
Games where the following condition is true:

$$(\delta^2 + \rho^2)(\delta^2 + \beta^2) - 4L^2\Delta^2 > 0, \quad (8)$$

where $0 < \delta \leq \sigma_i \left(\nabla^2_{x_1, x_2} g\right) \leq \Delta$, $\rho^2 = \min_{x_1, x_2} \lambda_{\min} \left[\nabla^2_{x_1, x_1} g(x_1, x_2)\right]^2$ and $\beta^2 = \min_{x_1, x_2} \lambda_{\min} \left[\nabla^2_{x_2, x_2} g(x_1, x_2)\right]^2$.

Proposition: Stochastic sufficiently bilinear game $\Rightarrow$ Stochastic Hamiltonian function (5) is smooth and satisfies the PL condition.

Stochastic Hamiltonian Gradient Methods

Stochastic Hamiltonian Gradient Descent (SHGD)

1. Generate fresh samples $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ and evaluate $\nabla \mathcal{H}_{i,j}(x^k)$.
2. Set step-size γ^k (constant, decreasing)
3. Set
$$x^{k+1} = x^k - \gamma^k \nabla \mathcal{H}_{i,j}(x^k)$$

Stochastic Hamiltonian Gradient Methods

Stochastic Hamiltonian Gradient Descent (SHGD)

1. Generate fresh samples $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ and evaluate $\nabla \mathcal{H}_{i,j}(x^k)$.
2. Set step-size γ^k (constant, decreasing)
3. Set

$$x^{k+1} = x^k - \gamma^k \nabla \mathcal{H}_{i,j}(x^k)$$

Loopless Stochastic Variance Reduced Hamiltonian Gradient (L-SVRHG)

Input: Choose initial points $x^0 = w^0 \in \mathbb{R}^d$ and probability $p \in (0,1]$.

1. Generate fresh samples $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ and evaluate $\nabla \mathcal{H}_{i,j}(x^k)$.
2. Evaluate $\boxed{g^k = \nabla \mathcal{H}_{i,j}(x^k) - \nabla \mathcal{H}_{i,j}(w^k) + \nabla \mathcal{H}(w^k)}$.
3. Set $x^{k+1} = x^k - \gamma g^k$
4. Set $w^{k+1} = \begin{cases} x^k & \text{with probability} \quad p \\ w^k & \text{with probability} \quad 1-p \end{cases}$

Convergence Guarantees

Algorithm	**Stochastic Bilinear Game** $\mathbb{E}\left[\|x^k - x^*\|^2\right]$	**Stochastic Sufficiently Bilinear Game** $\mathbb{E}\left[\mathcal{H}(x)\right]$	Remarks on Rates (all: global, non-asymptotic)
SHGD Constant step-size	Linear	Linear	last-iterate convergence to neighborhood
SHGD Decreasing step-size	sublinear: $\mathcal{O}(1/k)$	sublinear: $\mathcal{O}(1/k)$	last-iterate convergence to min-max solution
L-SVRHG with/without restarts	Linear	Linear	last-iterate convergence to min-max solution

Table: Summary of Convergence Analysis Results

Remark: In our results we do not assume bounded gradient or bounded variance. We use the recently introduced weak assumptions of *Expected smoothness* and *Expected Residual.* (Gower et al., 2019, 2020)

Numerical Evaluation

- Stochastic Bilinear Games
- Stochastic Sufficiently Bilinear Games
- GANs

Stochastic Bilinear Game

$$g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} x_1^\top b_i + x_1^\top \mathbf{A}_i x_2 + c_i^\top x_2$$

$n = d_1 = d_2 = 100$, $[b_i]_k, [c_i]_k \sim \mathcal{N}(0, 1/n)$ and $[\mathbf{A}_i]_{kl} = 1$ if $i = k = l$.

Stochastic Bilinear Game

$$g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} x_1^\top b_i + x_1^\top \mathbf{A}_i x_2 + c_i^\top x_2$$

$n = d_1 = d_2 = 100$, $[b_i]_k, [c_i]_k \sim \mathcal{N}(0, 1/n)$ and $[\mathbf{A}_i]_{kl} = 1$ if $i = k = l$.

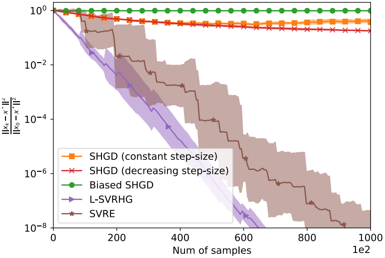

Figure: Distance to optimality $||x_k - x^*||^2/||x_0 - x^*||^2$

Stochastic Bilinear Game

$$g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} x_1^\top b_i + x_1^\top \mathbf{A}_i x_2 + c_i^\top x_2$$

$n = d_1 = d_2 = 100$, $[b_i]_k, [c_i]_k \sim \mathcal{N}(0, 1/n)$ and $[\mathbf{A}_i]_{kl} = 1$ if $i = k = l$.

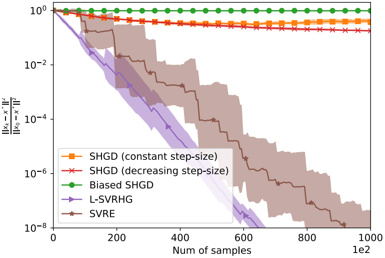

Figure: Distance to optimality $||x_k - x^*||^2 / ||x_0 - x^*||^2$

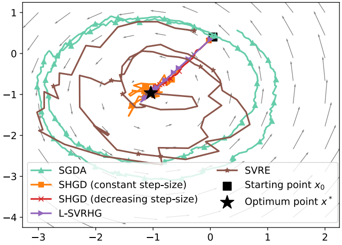

Figure: Gradient Vector Field and Trajectory. (x_1 and x_2 are scalars)

Take-Away Message

1. First set of global non-asymptotic last-iterate convergence guarantees for stochastic smooth games over a non-compact domain, in the absence of strong monotonicity assumptions.
2. Present the first variance reduced Hamiltonian method (linear convergence).
3. Hamiltonian Perspective: Popular stochastic optimization algorithms can be used as methods for solving stochastic min-max problems.

Future Extensions

- Hamiltonian-type methods for solving more classes of games.
- Development of efficient accelerated, distributed / decentralized Hamiltonian methods.

Thank You!

(for questions welcome to our virtual poster)