Stochastic Hamiltonian Gradient Methods for Smooth Games

Nicolas Loizou [1] **Hugo Berard** [1 2] **Alexia Jolicoeur-Martineau** [1] **Pascal Vincent**† [1 2] **Simon Lacoste-Julien**† [1] **Ioannis Mitliagkas**† [1]

Abstract

The success of adversarial formulations in machine learning has brought renewed motivation for smooth games. In this work, we focus on the class of stochastic Hamiltonian methods and provide the first convergence guarantees for certain classes of stochastic smooth games. We propose a novel unbiased estimator for the stochastic Hamiltonian gradient descent (SHGD) and highlight its benefits. Using tools from the optimization literature we show that SHGD converges linearly to the neighbourhood of a stationary point. To guarantee convergence to the exact solution, we analyze SHGD with a decreasing step-size and we also present the first stochastic variance reduced Hamiltonian method. Our results provide the first global non-asymptotic last-iterate convergence guarantees for the class of stochastic unconstrained bilinear games and for the more general class of stochastic games that satisfy a "sufficiently bilinear" condition, notably including some non-convex non-concave problems. We supplement our analysis with experiments on stochastic bilinear and sufficiently bilinear games, where our theory is shown to be tight, and on simple adversarial machine learning formulations.

1. Introduction

We consider the min-max optimization problem

$$\min_{x_1 \in \mathbb{R}^{d_1}} \max_{x_2 \in \mathbb{R}^{d_2}} g(x_1, x_2) \tag{1}$$

where $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is a smooth objective. Our goal is to find $x^* = (x_1^*, x_2^*)^\top \in \mathbb{R}^d$ where $d = d_1 + d_2$ such that

$$g(x_1^*, x_2) \leq g(x_1^*, x_2^*) \leq g(x_1, x_2^*), \tag{2}$$

for every $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$. We call point, x^*, a *saddle point*, *min-max solution* or *Nash equilibrium* of (1). In its general form, this problem is hard. In this work we focus on the simplest family of problems where some important questions are still open: the case where all stationary points are global min-max solutions.

Motivated by recent applications in machine learning, we are particularly interested in cases where the objective, g, is naturally expressed as a finite sum

$$\min_{x_1 \in \mathbb{R}^{d_1}} \max_{x_2 \in \mathbb{R}^{d_2}} g(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} g_i(x_1, x_2)\,, \tag{3}$$

where each component function $g_i : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is assumed to be smooth. Indeed, in problems like domain generalization (Albuquerque et al., 2019), generative adversarial networks (Goodfellow et al., 2014), and some formulations in reinforcement learning (Pfau & Vinyals, 2016), empirical risk minimization yields finite sums of the form of (3). We refer to this formulation as a ***stochastic smooth game***.[1] We call problem (1) a *deterministic game*.

The deterministic version of the problem has been studied in a number of classic (Korpelevich, 1976; Nemirovski, 2004) and recent results (Mescheder et al., 2017; Ibrahim et al., 2019; Gidel et al., 2018; Daskalakis et al., 2018; Gidel et al., 2019; Mokhtari et al., 2020; Azizian et al., 2020a;b) in various settings. Importantly, the majority of these results provide **last-iterate convergence** guarantees. In contrast, for the stochastic setting, guarantees on the classic extragradient method and its variants rely on iterate averaging over compact domains (Nemirovski, 2004). However, Chavdarova et al. (2019) highlighted a possibility of pathological behavior where the iterates *diverge towards* and then rotate near the boundary of the domain, far from the solution, while their average is shown to converge to the solution (by convexity).[2] This behavior is also problematic in the context of applying the method on non-convex problems, where averaging do not necessarily yield a solution (Daskalakis et al.,

[1]Mila, Université de Montréal † Canada CIFAR AI Chair [2]Facebook AI Research. Correspondence to: Nicolas Loizou <loizouni@mila.quebec>.

Proceedings of the 37^th International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]We note that all of our results except the one on variance reduction do not require the finite-sum assumption and can be easily adapted to the stochastic setting (see Appendix C.3).

[2]This is qualitatively very different to stochastic minimization where the iterates converge towards a neighborhood of the solution and averaging is only used to stabilize the method.

2018; Abernethy et al., 2019). It is only very recently that last-iterate convergence guarantees over a **non-compact domain** appeared in literature for the stochastic problem (Palaniappan & Bach, 2016; Chavdarova et al., 2019; Hsieh et al., 2019; Mishchenko et al., 2020) under the assumption of strong monotonicity. Strong monotonicity, a generalization of strong convexity for general operators, seems to be an essential condition for fast convergence in optimization. Here, we make **no strong monotonicity assumption**.

The algorithms we consider belong to a recently introduced family of computationally-light second order methods which in each step require the computation of a Jacobian-vector product. Methods that belong to this family are the consensus optimization (CO) method (Mescheder et al., 2017) and Hamiltonian gradient descent (Balduzzi et al., 2018; Abernethy et al., 2019). Even though some convergence results for these methods are known for the deterministic problem, there is no available analysis for the stochastic problem. We close this gap. We study *stochastic Hamiltonian gradient descent* (SHGD), and propose the first stochastic variance reduced Hamiltonian method, named L-SVRHG. Our contributions are summarized as follows:

- Our results provide the first set of global non-asymptotic last-iterate convergence guarantees for a stochastic game over a non-compact domain, in the absence of strong monotonicity assumptions.
- The proposed stochastic Hamiltonian methods use *novel unbiased estimators* of the gradient of the Hamiltonian function. This is an essential point for providing convergence guarantees. Existing practical variants of SHGD use biased estimators (Mescheder et al., 2017).
- We provide the first efficient convergence analysis of stochastic Hamiltonian methods. In particular, we focus on solving two classes of stochastic smooth games:
  - *Stochastic Bilinear Games*.
  - Stochastic games satisfying the "sufficiently bilinear" condition or simply *Stochastic Sufficiently Bilinear Games*. The deterministic variant of this class of games was firstly introduced by Abernethy et al. (2019) to study the deterministic problem and notably includes some **non-monotone problems**.
- For the above two classes of games, we provide convergence guarantees for SHGD with a constant step-size (linear convergence to a neighborhood of stationary point), SHGD with a variable step-size (sub-linear convergence to a stationary point) and L-SVRHG. For the latter, we guarantee a linear rate.
- We show the benefits of the proposed methods by performing numerical experiments on simple stochastic bilinear and sufficiently bilinear problems, as well as toy GAN problems for which the optimal solution is known. Our numerical findings corroborate our theoretical results.

2. Further Related work

In recent years, several second-order methods have been proposed for solving the min-max optimization problem (1). Some of them require the computation or inversion of a Jacobian which can be an inefficient operation (Wang et al., 2019; Mazumdar et al., 2019).[3] In contrast, second-order methods like the ones presented in Mescheder et al. (2017); Balduzzi et al. (2018); Abernethy et al. (2019) and in this work are more efficient as they only rely on the computation of a Jacobian-vector product in each step.

Abernethy et al. (2019) provide the first last-iterate convergence rates for the deterministic Hamiltonian gradient descent (HGD) for several classes of games including games satisfying the sufficiently bilinear condition. The authors briefly touch upon the stochastic setting and by using the convergence results of Karimi et al. (2016), explain how a stochastic variant of HGD with decreasing stepsize behaves. Their approach was purely theoretical and they did not provide an efficient way of selecting the unbiased estimators of the gradient of the Hamiltonian. In addition, they assumed bounded gradient of the Hamiltonian function which is restrictive for functions satisfying the Polyak-Lojasiewicz (PL) condition (Gower et al., 2020). In this work we provide the first efficient variants and analysis of SHGD. We did that by choosing practical unbiased estimator of the full gradient and by using the recently proposed assumptions of expected smoothness (Gower et al., 2019) and expected residual (Gower et al., 2020) in our analysis. The proposed theory of SHGD allow us to obtain as a corollary tight convergence guarantees for the deterministic HGD recovering the result of Abernethy et al. (2019).

In another line of work, Carmon et al. (2019) analyze variance reduction methods for constrained finite-sum problems and Ryu et al. (2019) provide an ODE-based analysis and guarantees in the monotone but potentially non-smooth case. Chavdarova et al. (2019) show that both alternate stochastic descent-ascent and stochastic extragradient diverge on an unconstrained stochastic bilinear problem. In the same paper, Chavdarova et al. (2019) propose the stochastic variance reduced extragradient (SVRE) algorithm with restart, which empirically achieves last-iterate convergence on this problem. However, it came with no theoretical guarantees. In Section 7, we observe in our experiments that SVRE is slower than the proposed L-SVRHG for both the stochastic bilinear and sufficiency bilinear games that we tested.

In concurrent work, Yang et al. (2020) provide global convergence guarantees for stochastic alternate gradient descent-ascent (and its variance reduction variant) for a subclass of nonconvex-nonconcave objectives satisfying a so-called

[3] See also Schäfer & Anandkumar (2019) for an efficient approximation of the inverse through a conjugate gradient approach.

two-sided Polyak-Lojasiewicz inequality, but this does not include the stochastic bilinear problem that we cover.

3. Technical Preliminaries

In this section, we present the necessary background and the basic notation used in the paper. We also describe the update rule of the deterministic Hamiltonian method.

3.1. Optimization Background: Basic Definitions

We start by presenting some definitions that we will later use in the analysis of the proposed methods.

Definition 3.1. Function $f : \mathbb{R}^d \to \mathbb{R}$ is μ–quasi-strongly convex if there exists a constant $\mu > 0$ such that $\forall x \in \mathbb{R}^d$: $f^* \geq f(x) + \langle \nabla f(x), x^* - x\rangle + \frac{\mu}{2} \|x^* - x\|^2$, where f^* is the minimum value of f and x^* is the projection of x onto the solution set $\mathcal{X}^*$ minimizing f.

Definition 3.2. We say that a function satisfies the Polyak-Lojasiewicz (PL) condition if there exists $\mu > 0$ such that

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left[f(x) - f^*\right] \quad \forall x \in \mathbb{R}^d\,, \qquad (4)$$

where f^* is the minimum value of f.

An analysis of several stochastic optimization methods under the assumption of PL condition (Polyak, 1987) was recently proposed in Karimi et al. (2016). A function can satisfy the PL condition and not be strongly convex, or even convex. However, if the function is $\mu-$quasi strongly convex then it satisfies the PL condition with the same μ (Karimi et al., 2016).

Definition 3.3. Function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if there exists $L > 0$ such that:
$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d$.

If $f = \frac{1}{n}\sum_{i=1}^n f_i(x)$, then a more refined analysis of stochastic gradient methods has been proposed under new notions of smoothness. In particular, the notions of *expected smoothness (ES)* and *expected residual (ER)* have been introduced and used in the analysis of SGD in Gower et al. (2019) and Gower et al. (2020) respectively. ES and ER are generic and remarkably weak assumptions. In Section 6 and Appendix B.2, we provide more details on their generality. We state their definitions below.

Definition 3.4 (Expected smoothness, (Gower et al., 2019))**.** We say that the function $f = \frac{1}{n}\sum_{i=1}^n f_i(x)$ satisfies the *expected smoothness* condition if there exists $\mathcal{L} > 0$ such that for all $x \in \mathbb{R}^d$,

$$\mathbb{E}_i\left[\|\nabla f_i(x) - \nabla f_i(x^*)\|^2\right] \leq 2\mathcal{L}(f(x) - f(x^*))\,. \quad (5)$$

Definition 3.5 (Expected residual, (Gower et al., 2020))**.** We say that the function $f = \frac{1}{n}\sum_{i=1}^n f_i(x)$ satisfies the *expected residual* condition if $\exists \rho > 0$ such that $\forall x \in \mathbb{R}^d$,

$$\mathbb{E}_i\left[\|\nabla f_i(x) - \nabla f_i(x^*) - (\nabla f(x) - \nabla f(x^*))\|^2\right] \leq 2\rho\left(f(x) - f(x^*)\right). \quad (6)$$

3.2. Smooth Min-Max Optimization

We use standard notation used previously in Mescheder et al. (2017); Balduzzi et al. (2018); Abernethy et al. (2019); Letcher et al. (2019).

Let $x = (x_1, x_2)^\top \in \mathbb{R}^d$ be the column vector obtained by stacking x_1 and x_2 one on top of the other. With $\xi(x) := (\nabla_{x_1} g, -\nabla_{x_2} g)^\top$, we denote the signed vector of partial derivatives evaluated at point x. Thus, $\xi(x) : \mathbb{R}^d \to \mathbb{R}^d$ is a vector function. We use

$$\mathbf{J} = \nabla\xi = \begin{pmatrix} \nabla^2_{x_1,x_1} g & \nabla^2_{x_1,x_2} g \\ -\nabla^2_{x_2,x_1} g & -\nabla^2_{x_2,x_2} g \end{pmatrix} \in \mathbb{R}^{d\times d}$$

to denote the Jacobian of the vector function ξ. Note that using the above notation, the simultaneous gradient descent/ascent (SGDA) update can be written simply as: $x^{k+1} = x^k - \eta_k \xi(x_k)$.

Definition 3.6. The objective function g of problem (1) is L_g-smooth if there exist $L_g > 0$ such that:
$\|\xi(x) - \xi(y)\| \leq L_g\|x - y\| \quad \forall x, y \in \mathbb{R}^d$.

We also say that g is L-smooth in x_1 (in x_2) if $\|\nabla_{x_1} g(x_1, x_2) - \nabla_{x_1} g(x_1', x_2)\| \leq L\|x_1 - x_1'\|$ (if $\|\nabla_{x_2} g(x_1, x_2) - \nabla_{x_2} g(x_1, x_2')\| \leq L\|x_2 - x_2'\|$) for all $x_1, x_1' \in \mathbb{R}^{d_1}$ (for all $x_2, x_2' \in \mathbb{R}^{d_2}$).

Definition 3.7. A stationary point of function $f : \mathbb{R}^d \to \mathbb{R}$ is a point $x^* \in \mathbb{R}^d$ such that $\nabla f(x^*) = 0$. Using the above notation, in min-max problem (1), point $x^* \in \mathbb{R}^d$ is a stationary point when $\xi(x^*) = 0$.

As mentioned in the introduction, in this work we focus on smooth games satisfying the following assumption.

Assumption 3.8. The objective function g of problem (3) has at least one stationary point and all of its stationary points are global min-max solutions.

With Assumption 3.8, we can guarantee convergence to a min-max solution of problem (3) by proving convergence to a stationary point. This assumption is true for several classes of games including strongly convex-strongly concave and convex-concave games. However, it can also be true for some classes of non-convex non-concave games (Abernethy et al., 2019). In Section 4, we describe in more details the two classes of games that we study. Both satisfy this assumption.

3.3. Deterministic Hamiltonian Gradient Descent

Hamiltonian gradient descent (HGD) has been proposed as an efficient method for solving min-max problems in Balduzzi et al. (2018). To the best of our knowledge, the first convergence analysis of the method is presented in Abernethy et al. (2019) where the authors prove non-asymptotic linear last-iterate convergence rates for several classes of games.

In particular, HGD converges to saddle points of problem (1) by performing gradient descent on a particular objective function $\mathcal{H}$, which is called the Hamiltonian function (Balduzzi et al., 2018), and has the following form:

$$\min_x \quad \mathcal{H}(x) = \frac{1}{2}\|\xi(x)\|^2. \tag{7}$$

That is, HGD is a gradient descent method that minimizes the square norm of the gradient $\xi(x)$. Note that under Assumption 3.8, solving problem (7) is equivalent to solving problem (1). The equivalence comes from the fact that $\mathcal{H}$ only achieves its minimum at stationary points. The update rule of HGD can be expressed using a Jacobian-vector product (Balduzzi et al., 2018; Abernethy et al., 2019):

$$x^{k+1} = x^k - \eta_k \nabla\mathcal{H}(x) = x^k - \eta_k \left[\mathbf{J}^\top \xi\right], \tag{8}$$

making HGD a second-order method. However, as discussed in Balduzzi et al. (2018), the Jacobian-vector product can be efficiently evaluated in tasks like training neural networks and the computation time of the gradient and the Jacobian-vector product is comparable (Pearlmutter, 1994).

4. Stochastic Smooth Games and Stochastic Hamiltonian Function

In this section, we provide the two classes of stochastic games that we study. We define the stochastic counterpart to the Hamiltonian function as a step towards solving problem (3) and present its main properties.

Let us start by presenting the basic notation for the stochastic setting. Let $\xi(x) = \frac{1}{n}\sum_{i=1}^n \xi_i(x)$, where $\xi_i(x) := \left(\nabla_{x_1} g_i, -\nabla_{x_2} g_i\right)^\top$, for all $i \in [n]$ and let

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

Using the above notation, the stochastic variant of SGDA can be written as $x^{k+1} = x^k - \eta_k \xi_i(x_k)$ where $\mathbb{E}_i[\xi_i(x_k)] = \xi(x_k)$.[4]

In this work, we focus on stochastic smooth games of the form (3) that satisfy the following assumption.

[4] Here the expectation is over the uniform distribution. That is, $\mathbb{E}_i[\xi_i(x)] = \frac{1}{n}\sum_{i=1}^n \xi_i(x)$.

Assumption 4.1. Functions $g_i : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ of problem (3) are twice differentiable, L_i-smooth with S_i-Lipschitz Jacobian. That is, for each $i \in [n]$ there are constants $L_i > 0$ and $S_i > 0$ such that $\|\xi_i(x) - \xi_i(y)\| \leq L_i\|x - y\|$ and $\|\mathbf{J}_i(x) - \mathbf{J}_i(y)\| \leq S_i\|x - y\|$ for all $x, y \in R^d$.

4.1. Classes of Stochastic Games

Here we formalize the two families of stochastic smooth games under study: (i) stochastic bilinear, and (ii) stochastic sufficiently bilinear. Both families satisfy Assumption 3.8. Interestingly, the latter family includes some non-convex non-concave games, i.e. non-monotone problems.

Stochastic Bilinear Games. A stochastic bilinear game is the stochastic smooth game (3) in which function g has the following structure:

$$g(x_1, x_2) = \frac{1}{n}\sum_{i=1}^n \left(x_1^\top b_i + x_1^\top \mathbf{A}_i x_2 + c_i^\top x_2\right). \tag{9}$$

While this game appears simple, standard methods diverge on it (Chavdarova et al., 2019) and L-SVRHG gives the first stochastic method with last-iterate convergence guarantees.

Stochastic sufficiently bilinear games. A game of the form (3) is called *stochastic sufficiently bilinear* if it satisfies the following definition.

Definition 4.2. Let Assumption 4.1 be satisfied and let the objective function g of problem (3) be L-smooth in x_1 and L-smooth in x_2. Assume that a constant $C > 0$ exists, such that $\mathbb{E}_i\|\xi_i(x)\| < C$. Assume the cross derivative $\nabla^2_{x_1,x_2} g$ be full rank with $0 < \delta \leq \sigma_i\left(\nabla^2_{x_1,x_2} g\right) \leq \Delta$ for all $x \in \mathbb{R}^d$ and for all singular values σ_i. Let $\rho^2 = \min_{x_1,x_2} \lambda_{\min}\left[\nabla^2_{x_1,x_1} g(x_1,x_2)\right]^2$ and $\beta^2 = \min_{x_1,x_2} \lambda_{\min}\left[\nabla^2_{x_2,x_2} g(x_1,x_2)\right]^2$. Finally let the following condition to be true:

$$(\delta^2 + \rho^2)(\delta^2 + \beta^2) - 4L^2\Delta^2 > 0. \tag{10}$$

Note that the definition of the stochastic sufficiently bilinear game has no restriction on the convexity of functions $g_i(x)$ and $g(x)$. The most important condition that needs to be satisfied is the expression in equation (10). Following the terminology of Abernethy et al. (2019), we call the condition (10): "*sufficiently bilinear*" condition. Later in our numerical evaluation, we present stochastic non convex-non concave min-max problems that satisfy condition (10).

We highlight that the deterministic counterpart of the above game was first proposed in Abernethy et al. (2019). The deterministic variant of Abernethy et al. (2019) can be obtained as special case of the above class of games when $n = 1$ in problem (3).

4.2. Stochastic Hamiltonian Function

Having presented the two main classes of stochastic smooth games, in this section we focus on the structure of the stochastic Hamiltonian function and highlight some of its properties.

Finite-Sum Structure Hamiltonian Function. Having the objective function g of problem (3) to be stochastic and in particular to be a finite-sum function, leads to the following expression for the Hamiltonian function:

$$\mathcal{H}(x) = \frac{1}{n^2}\sum_{i,j=1}^{n} \underbrace{\frac{1}{2}\langle \xi_i(x), \xi_j(x)\rangle}_{\mathcal{H}_{i,j}(x)} . \tag{11}$$

That is, the Hamiltonian function $\mathcal{H}(x)$ can be expressed as a finite-sum with n^2 components.

Properties of the Hamiltonian Function. As we will see in the following sections, the finite-sum structure of the stochastic Hamiltonian function (11) allows us to use popular stochastic optimization problems for solving problem (7). However in order to be able to provide convergence guarantees of the proposed stochastic Hamiltonian methods, we need to show that the stochastic Hamiltonian function (11) satisfies specific properties for the two classes of games we study. This is what we do in the following two propositions.

Proposition 4.3. For stochastic bilinear games of the form (9), the stochastic Hamiltonian function (11) is a smooth quadratic $\mu_{\mathcal{H}}$–quasi-strongly convex function with constants $L_{\mathcal{H}} = \sigma^2_{\max}(\mathbf{A})$ and $\mu_{\mathcal{H}} = \sigma^2_{\min}(\mathbf{A})$ where $\mathbf{A} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_i$ and $\sigma_{\max}$ and $\sigma_{\min}$ are the maximum and minimum non-zero singular values of $\mathbf{A}$.

Proposition 4.4. For stochastic sufficiently bilinear games, the stochastic Hamiltonian function (11) is a $L_{\mathcal{H}} = \bar{S}C + \bar{L}^2$ smooth function and satisfies the PL condition (4) with $\mu_{\mathcal{H}} = \frac{(\delta^2+\rho^2)(\delta^2+\beta^2)-4L^2\Delta^2}{2\delta^2+\rho^2+\beta^2}$. Here $\bar{S} = \mathbb{E}_i[S_i]$ and $\bar{L} = \mathbb{E}_i[L_i]$.

5. Stochastic Hamiltonian Gradient Methods

In this section we present the proposed stochastic Hamiltonian methods for solving the stochastic min-max problem (3). Our methods could be seen as extensions of popular stochastic optimization methods into the Hamiltonian setting. In particular, the two algorithms that we build upon are the popular stochastic gradient descent (SGD) and the recently introduced loopless stochastic variance reduced gradient (L-SVRG). For completeness, we present their form for solving finite-sum optimization problems in Appendix A.

Algorithm 1 Stochastic Hamiltonian Gradient Descent (SHGD)

Input: Starting stepsize $\gamma^0 > 0$. Choose initial points $x^0 \in \mathbb{R}^d$. Distribution $\mathcal{D}$ of samples.
for $k = 0, 1, 2, \cdots, K$ **do**
  Generate fresh samples $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ and evaluate $\nabla\mathcal{H}_{i,j}(x^k)$.
  Set step-size γ^k following one of the selected choices (constant, decreasing)
  Set $x^{k+1} = x^k - \gamma^k \nabla\mathcal{H}_{i,j}(x^k)$
end for

5.1. Unbiased Estimator

One of the most important elements of stochastic gradient-based optimization algorithms for solving finite-sum problems of the form (11) is the selection of unbiased estimators of the full gradient $\nabla\mathcal{H}(x)$ in each step. In our proposed optimization algorithms for solving (11), at each step we use the gradient of only one component function $\mathcal{H}_{i,j}(x)$:

$$\nabla\mathcal{H}_{i,j}(x) = \frac{1}{2}\left[\mathbf{J}_i^\top \xi_j + \mathbf{J}_j^\top \xi_i\right]. \tag{12}$$

It can easily be shown that this selection is an unbiased estimator of $\nabla\mathcal{H}(x)$. That is, $\mathbb{E}_{i,j}\left[\nabla\mathcal{H}_{i,j}(x)\right] = \nabla\mathcal{H}(x)$.

5.2. Stochastic Hamiltonian Gradient Descent (SHGD)

Stochastic gradient descent (SGD) (Robbins & Monro, 1951; Nemirovski & Yudin, 1978; 1983; Nemirovski et al., 2009; Hardt et al., 2016; Gower et al., 2019; 2020; Loizou et al., 2020) is the workhorse for training supervised machine learning problems. In Algorithm 1, we apply SGD to (11), yielding stochastic Hamiltonian gradient descent (SHGD) for solving problem (3). Note that at each step, $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ are sampled from a given well-defined distribution $\mathcal{D}$ and then are used to evaluate $\nabla\mathcal{H}_{i,j}(x^k)$ (unbiased estimator of the full gradient). In our analysis, we provide rates for two selections of step-sizes for SHGD. These are the constant step-size $\gamma^k = \gamma$ and the decreasing step-size (switching rule which describe when one should switch from a constant to a decreasing stepsize regime).

5.3. Loopless Stochastic Variance Reduced Hamiltonian Gradient (L-SVRHG)

One of the main disadvantage of Algorithm 1 with constant step-size selection is that it guarantees linear convergence only to a neighborhood of the min-max solution x^*. As we will
allow us to obtain exact convergence to the min-max but at the expense of slower rate (sublinear).

One of the most remarkable algorithmic breakthroughs in recent years was the development of variance-reduced stochas-

Algorithm 2 Loopless Stochastic Variance Reduced Hamiltonian Gradient (L-SVRHG)

Input: Starting stepsize $\gamma > 0$. Choose initial points $x^0 = w^0 \in \mathbb{R}^d$. Distribution $\mathcal{D}$ of samples. Probability $p \in (0, 1]$
for $k = 0, 1, 2, \cdots, K-1$ **do**
Generate fresh samples $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ and evaluate $\nabla \mathcal{H}_{i,j}(x^k)$.
Evaluate $g^k = \nabla \mathcal{H}_{i,j}(x^k) - \nabla \mathcal{H}_{i,j}(w^k) + \nabla \mathcal{H}(w^k)$.
Set $x^{k+1} = x^k - \gamma g^k$
Set

$$w^{k+1} = \begin{cases} x^k & \text{with probability} \quad p \\ w^k & \text{with probability} \quad 1-p \end{cases}$$

end for
Output:
Option I: The last iterate $x = x^k$.
Option II: x is chosen uniformly at random from $\{x^i\}_{i=0}^K$.

tic gradient algorithms for solving finite-sum optimization problems. These algorithms, by reducing the variance of the stochastic gradients, are able to guarantee convergence to the exact solution of the optimization problem with faster convergence than classical SGD. For example, for smooth strongly convex functions, variance reduced methods can guarantee linear convergence to the optimum. This is a vast improvement on the sub-linear convergence of SGD with decreasing step-size. In the past several years, many efficient variance-reduced methods have been proposed. Some popular examples of variance reduced algorithms are SAG (Schmidt et al., 2017), SAGA (Defazio et al., 2014), SVRG (Johnson & Zhang, 2013) and SARAH (Nguyen et al., 2017). For more examples of variance reduced methods in different settings, see Defazio (2016); Konečný et al. (2016); Gower et al. (2018); Sebbouh et al. (2019).

In our second method Algorithm 2, we propose a variance reduced Hamiltonian method for solving (3). Our method is inspired by the recently introduced and well behaved variance reduced algorithm, Loopless-SVRG (L-SVRG) first proposed in Hofmann et al. (2015); Kovalev et al. (2020) and further analyzed under different settings in Qian et al. (2019); Gorbunov et al. (2020); Khaled et al. (2020). We name our method loopless stochastic variance reduced Hamiltonian gradient (L-SVRHG). The method works by selecting at each step the unbiased estimator $g^k = \nabla \mathcal{H}_{i,j}(x^k) - \nabla \mathcal{H}_{i,j}(w^k) + \nabla \mathcal{H}(w^k)$ of the full gradient. As we will prove in the next section, this method guarantees linear convergence to the min-max solution of the stochastic bilinear game (9).

To get a linearly convergent algorithm in the more general setup of sufficiently bilinear games 4.2, we had to propose a restarted variant of Alg. 2, presented in Alg. 3, which calls at each step Alg. 2 with the second option of output, that is L-SVRHG$_{II}$. Using the property from Proposition 4.4 that the Hamiltonian function (11) satisfy the PL condition 3.2, we show that Alg. 3 converges linearly to the solution of the sufficiently bilinear game (Theorem 6.8).

Algorithm 3 L-SVRHG (with Restart)

Input: Starting stepsize $\gamma > 0$. Choose initial points $x^0 = w^0 \in \mathbb{R}^d$. Distribution $\mathcal{D}$ of samples. Probability $p \in (0, 1]$, T
for $t = 0, 1, 2, \cdots, T$ **do**
Set x^{t+1} = L-SVRHG$_{II}(x^t, K, \gamma, p \in (0, 1])$
end for
Output: The last iterate x^T.

6. Convergence Analysis

We provide theorems giving the performance of the previously described stochastic Hamiltonian methods for solving the two classes of stochastic smooth games: stochastic bilinear and stochastic sufficiently bilinear. In particular, we present three main theorems for each one of these classes describing the convergence rates for (i) SHGD with constant step-size, (ii) SHGD with decreasing step-size and (iii) L-SVRHG and its restart variant (Algorithm 3).

The proposed results depend on the two main parameters $\mu_{\mathcal{H}}$, $L_{\mathcal{H}}$ evaluated in Propositions 4.3 and 4.4. In addition, the theorems related to the bilinear games (the Hamiltonian function is quasi-strongly convex) use the expected smoothness constant $\mathcal{L}$ (5), while the theorems related to the sufficiently bilinear games (the Hamiltonian function satisfied the PL condition) use the expected residual constant ρ (6). We note that the expected smoothness and expected residual constants can take several values according to the well-defined distributions $\mathcal{D}$ selected in our algorithms and the proposed theory will still hold (Gower et al., 2019; 2020).

As a concrete example, in the case of τ-minibatch sampling,[5] the expected smoothness and expected residual parameters take the following values:

$$\mathcal{L}(\tau) = \frac{n^2(\tau-1)}{\tau(n^2-1)} L_{\mathcal{H}} + \frac{n^2-\tau}{\tau(n^2-1)} L_{\max} \tag{13}$$

$$\rho(\tau) = L_{\max} \frac{n^2-\tau}{(n^2-1)\tau} \tag{14}$$

where $L_{\max} = \max_{\{1,\ldots,n^2\}}\{L_{\mathcal{H}_{i,j}}\}$ is the maximum smoothness constant of the functions $\mathcal{H}_{i,j}$. By using the expressions (13) and (14), it is easy to see that for single element sampling where $\tau = 1$ (the one we use in our ex-

[5] In each step we draw uniformly at random τ components of the n^2 possible choices of the stochastic Hamiltonian function (11). For more details on the τ-minibatch sampling see Appendix B.2.

periments) $\mathcal{L} = \rho = L_{\max}$. On the other limit case where a full-batch is used ($\tau = n^2$), that is we run the deterministic Hamiltonian gradient descent, these values become $\mathcal{L} = L_{\mathcal{H}}$ and $\rho = 0$ and as we explain below, the proposed theorems include the convergence of the deterministic method as special case.

6.1. Stochastic Bilinear Games

We start by presenting the convergence of SHGD with constant step-size and explain how we can also obtain an analysis of the HGD (8) as special case. Then we move to the convergence of SHGD with decreasing step-size and the L-SVRHG where we are able to guarantee convergence to a min-max solution x^*. In the results related to SHGD we use $\sigma^2 := \mathbb{E}_{i,j}[\|\nabla \mathcal{H}_{i,j}(x^*)\|^2]$ to denote the finite gradient noise at the solution.

Theorem 6.1 (Constant stepsize)**.** Let us have the stochastic bilinear game (9). Then iterates of SHGD with constant step-size $\gamma^k = \gamma \in (0, \frac{1}{2\mathcal{L}}]$ satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1-\gamma\mu_{\mathcal{H}})^k \|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}. \quad (15)$$

That is, Theorem 6.1 shows linear convergence to a neighborhood of the min-max solution. Using Theorem 6.1 and following the approach of Gower et al. (2019), we can obtain the following corollary on the convergence of deterministic Hamiltonian gradient descent (HGD) (8). Note that for the deterministic case $\sigma = 0$ and $\mathcal{L} = L$ (13).

Corollary 6.2. Let us have a deterministic bilinear game. Then the iterates of HGD with step-size $\gamma = \frac{1}{2L}$ satisfy:

$$\|x^k - x^*\|^2 \leq (1-\gamma\mu_{\mathcal{H}})^k \|x^0 - x^*\|^2 . \quad (16)$$

To the best of our knowledge, Corollary 6.2 provides the first linear convergence guarantees for HGD in terms of $\|x^k - x^*\|^2$ (Abernethy et al. (2019) gave guarantees only on $\mathcal{H}(x^k)$). Let us now select a decreasing step-size rule (switching strategy) that guarantees a sublinear convergence to the exact min-max solution for the SHGD.

Theorem 6.3 (Decreasing stepsizes/switching strategy)**.** Let us have the stochastic bilinear game (9). Let $\mathcal{K} := \mathcal{L}/\mu_{\mathcal{H}}$. Let

$$\gamma^k = \begin{cases} \frac{1}{2\mathcal{L}} & \text{for} \quad k \leq 4\lceil \mathcal{K} \rceil \\ \frac{2k+1}{(k+1)^2\mu_{\mathcal{H}}} & \text{for} \quad k > 4\lceil \mathcal{K} \rceil . \end{cases} \quad (17)$$

If $k \geq 4\lceil \mathcal{K} \rceil$, then SHGD given in Algorithm 1 satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq \frac{\sigma^2}{\mu_{\mathcal{H}}^2}\frac{8}{k} + \frac{16\lceil \mathcal{K} \rceil^2}{e^2k^2}\|x^0 - x^*\|^2. \quad (18)$$

Lastly, in the following theorem, we show under what selection of step-size L-SVRHG convergences linearly to a min-max solution.

Theorem 6.4 (L-SVRHG)**.** Let us have the stochastic bilinear game (9). Let step-size $\gamma = 1/6L_{\mathcal{H}}$ and $p \in (0, 1]$. Then L-SVRHG with Option I for output as given in Algorithm 2 convergences linearly to the min-max solution x^* and satisfies:

$$\mathbb{E}[\Phi^k] \leq \max\left\{1 - \frac{\mu}{6L_{\mathcal{H}}}, 1 - \frac{p}{2}\right\}^k \Phi^0$$

where $\Phi^k := \|x^k - x^*\|^2 + \frac{4\gamma^2}{pn^2}\sum_{i,j=1}^n \|\nabla \mathcal{H}_{i,j}(w^k) - \nabla \mathcal{H}_{i,j}(x^*)\|^2$.

6.2. Stochastic Sufficiently-Bilinear Games

As in the previous section, we start by presenting the convergence of SHGD with constant step-size and explain how we can obtain an analysis of the HGD (8) as special case. Then we move to the convergence of SHGD with decreasing step-size and the L-SVRHG (with restart) where we are able to guarantee linear convergence to a min-max solution x^*. In contrast to the results on bilinear games, the convergence guarantees of the following theorems are given in terms of the Hamiltonian function $\mathbb{E}[\mathcal{H}(x^k)]$. In all theorems we call “sufficiently-bilinear game” the game described in Definition 4.2. With $\sigma^2 := \mathbb{E}_{i,j}[\|\nabla \mathcal{H}_{i,j}(x^*)\|^2]$, we denote the finite gradient noise at the solution.

Theorem 6.5. Let us have a stochastic sufficiently-bilinear game. Then the iterates of SHGD with constant steps-size $\gamma^k = \gamma \leq \frac{\mu}{L(\mu+2\rho)}$ satisfy:

$$\mathbb{E}[\mathcal{H}(x^k)] \leq (1-\gamma\mu_{\mathcal{H}})^k [\mathcal{H}(x^0)] + \frac{L_{\mathcal{H}}\gamma\sigma^2}{\mu_{\mathcal{H}}}. \quad (19)$$

Using the above Theorem and by following the approach of Gower et al. (2020), we can obtain the following corollary on the convergence of deterministic Hamiltonian gradient descent (HGD) (8). It shows linear convergence of HGD to the min-max solution. Note that for the deterministic case $\sigma = 0$ and $\rho = 0$ (14).

Corollary 6.6. Let us have a deterministic sufficiently-bilinear game. Then the iterates of HGD with step-size $\gamma = \frac{1}{L_{\mathcal{H}}}$ satisfy:

$$\mathcal{H}(x^k) \leq (1-\gamma\mu_{\mathcal{H}})^k \mathcal{H}(x^0) . \quad (20)$$

The result of Corollary 6.6 is equivalent to the convergence of HGD as proposed in Abernethy et al. (2019).

Let us now show that with decreasing step-size (switching strategy), SHGD can converge (with sub-linear rate) to the min-max solution.

Theorem 6.7 (Decreasing stepsizes/switching strategy)**.** Let us have a stochastic sufficiently-bilinear game. Let $k^* := 2\frac{L}{\mu}\left(1 + 2\frac{\rho}{\mu}\right)$ and

$$\gamma^k = \begin{cases} \frac{\mu_{\mathcal{H}}}{L_{\mathcal{H}}(\mu_{\mathcal{H}}+2\rho)} & \text{for} \quad k \leq \lceil k^* \rceil \\ \frac{2k+1}{(k+1)^2\mu_{\mathcal{H}}} & \text{for} \quad k > \lceil k^* \rceil. \end{cases} \tag{21}$$

If $k \geq \lceil k^* \rceil$, then SHGD given in Algorithm 1 satisfy:

$$\mathbb{E}[\mathcal{H}(x^k)] \leq \frac{4L_{\mathcal{H}}\sigma^2}{\mu_{\mathcal{H}}^2}\frac{1}{k} + \frac{(k^*)^2}{k^2e^2}[\mathcal{H}(x^0)].$$

In the next Theorem we show how the updates of L-SVRHG with Restart (Algorithm 3) converges linearly to the min-max solution. We highlight that each step t of Alg. 3 requires $K = \frac{4}{\mu_{\mathcal{H}}\gamma}$ updates of the L-SVRHG.

Theorem 6.8 (L-SVRHG with Restart)**.** Let us have a stochastic sufficiently-bilinear game. Let $p \in (0, 1]$ and $\gamma \leq \min\left\{\frac{1}{4L_{\mathcal{H}}}, \frac{p^{2/3}}{36^{1/3}(L_{\mathcal{H}}\rho)^{1/3}}, \frac{\sqrt{p}}{\sqrt{6}\rho}\right\}$ and let $K = \frac{4}{\mu_{\mathcal{H}}\gamma}$. Then the iterates of L-SVRHG (with Restart) given in Algorithm 3 satisfies

$$\mathbb{E}[\mathcal{H}(x^t)] \leq (1/2)^t\,[\mathcal{H}(x^0)].$$

7. Numerical Evaluation

In this section, we compare the algorithms proposed in this paper to existing methods in the literature. Our goal is to illustrate the good convergence properties of the proposed algorithms as well as to explore how these algorithms behave in settings not covered by the theory. We propose to compare the following algorithms: **SHGD** with constant step-size and decreasing step-size, a biased version of SHGD (Mescheder et al., 2017), **L-SVRHG** with and without restart, consensus optimization (**CO**)[6] (Mescheder et al., 2017), the stochastic variant of **SGDA**, and finally the stochastic variance-reduced extragradient with restart **SVRE** proposed in Chavdarova et al. (2019). For all our experiments, we ran the different algorithms with 10 different seeds and plot the mean and 95% confidence intervals. We provide further details about the experiments and choice of hyperparameters for the different methods in Appendix F.

7.1. Bilinear Games

First we compare the different methods on the stochastic bilinear problem (9). Similarly to Chavdarova et al. (2019), we choose $n = d_1 = d_2 = 100$, $[\mathbf{A}_i]_{kl} = 1$ if $i = k = l$ and 0 otherwise, and $[b_i]_k, [c_i]_k \sim \mathcal{N}(0, 1/n)$.

[6]**CO** is a mix between SGDA and SHGD, with the following update rule $x^{k+1} = x^k - \eta_k(\xi_i(x^k) + \lambda\nabla\mathcal{H}_{i,j}(x^k))$ (See Appendix F.5)

We show the convergence of the different algorithms in Fig. 1a. As predicted by theory, **SHGD** with decreasing step-size converges at a sublinear rate while **L-SVRHG** converges at a linear rate. Among all the methods we compared to, **L-SVRHG** is the fastest to converge; however, the speed of convergence depends a lot on parameter p. We observe that setting $p = 1/n$ yields the best performance.

To further illustrate the behavior of the Hamiltonian methods, we look at the trajectory of the methods on a simple 2D version of the bilinear game, where we choose x_1 and x_2 to be scalars. We observe that while previously proposed methods such as **SGDA** and **SVRE** suffer from rotations which slow down their convergence and can even make them diverge, the Hamiltonian methods converge much faster by removing rotation and converging “straight” to the solution.

7.2. Sufficiently-Bilinear Games

In section 6.2, we showed that Hamiltonian methods are also guaranteed to converge when the problem is non-convex non-concave but satisfies the sufficiently-bilinear condition (10). To illustrate these results, we propose to look at the following game inspired by Abernethy et al. (2019):

$$\min_{x_1\in\mathbb{R}^d}\max_{x_2\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}\big(F(x_1) + \delta\; x_1^\top\mathbf{A}_i x_2 + b_i^\top x_1 + c_i^\top x_2 - F(x_2)\big), \tag{22}$$

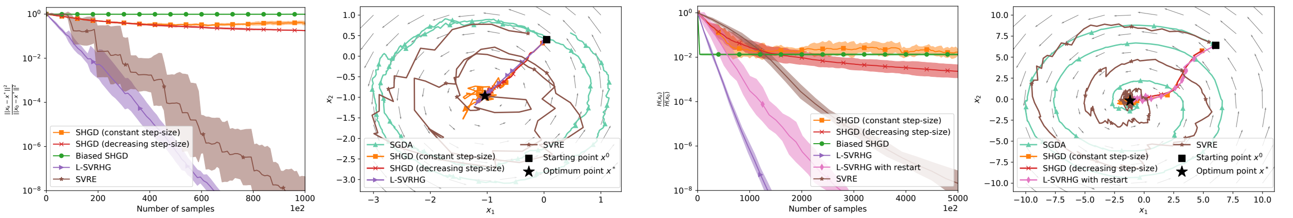
where $F(x)$ is a non-linear function (see details in Appendix F.2). This game is non-convex non-concave and satisfies the sufficiently-bilinear condition if $\delta > 2L$, where L is the smoothness of $F(x)$. Thus, the results and theorems from Section 6.2 hold.

Results are shown in Fig.1b. Similarly to the bilinear case, the methods follow very closely the theory. We highlight that while the proposed theory for this setting only guarantees convergence for **L-SVRHG** with restart, in practice using restart is not strictly necessary: **L-SVRHG** with the correct choice of stepsize also converges in our experiment. Finally we show the trajectories of the different methods on a 2D version of the problem. We observe that contrary to the bilinear case, stochastic SGDA converges but still suffers from rotation compared to Hamiltonian methods.

7.3. GANs

In previous experiments, we verify the proposed theory for the stochastic bilinear and sufficiently-bilinear games. Although we do not have theoretical results for more complex games, we wanted to test our algorithms on a simple GAN setting, which we call *GaussianGAN*.

In *GaussianGAN*, we have a dataset of real data x_{real} and latent variable z from a normal distribution with mean

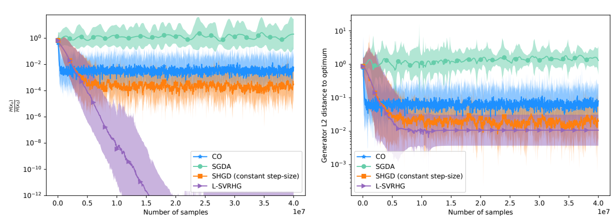


(a) Bilinear game (b) Sufficiently-bilinear game

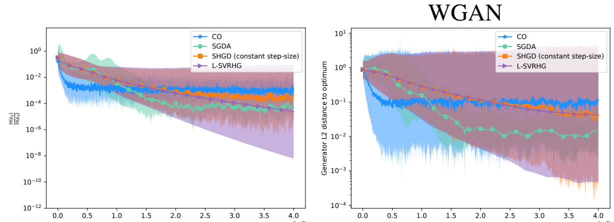
Figure 1. **a)** Comparison of different methods on the stochastic bilinear game (9). Left: Distance to optimality $\frac{||x_k - x^*||^2}{||x_0 - x^*||^2}$ as a function of the number of samples seen during training. Right: The trajectory of the different methods on a 2D version of the problem.
b) Comparison of different methods on the sufficiently bilinear games (22). Left: The Hamiltonian $\frac{H(x_k)}{H(x_0)}$ as a function of the number of samples seen during training. Right: The trajectory of the different methods on a 2D version of the problem.

0 and standard deviation 1. The generator is defined as $G(z) = \mu + \sigma z$ and the discriminator as $D(x_{data}) = \phi_0 + \phi_1 x_{data} + \phi_2 x_{data}^2$, where x_{data} is either real data (x_{real}) or fake generated data ($G(z)$). In this setting, the parameters are $x = (x_1, x_2) = ([\mu, \sigma], [\phi_0, \phi_1, \phi_2])$. In GaussianGAN, we can directly measure the L^2 distance between the generator's parameters and the true optimal parameters: $||\hat{\mu} - \mu|| + ||\hat{\sigma} - \sigma||$, where $\hat{\mu}$ and $\hat{\sigma}$ are the sample's mean and standard deviation.

We consider three possible minmax games: Wasserstein GAN (WGAN) (Arjovsky et al., 2017), saturating GAN (satGAN) (Goodfellow et al., 2014), and non-saturating GAN (nsGAN) (Goodfellow et al., 2014). We present the results for WGAN and satGAN in Figure 2. We provide the nsGAN results in Appendix G.2 and details for the different experiments in Appendix F.3.



(a) Hamiltonian for WGAN (b) Distance to optimum for WGAN



(c) Hamiltonian for satGAN (d) Distance to optimum for satGAN

Figure 2. The Hamiltonian $\frac{H(x_k)}{H(x_0)}$ (**left**) and the distance to the optimal generator (**right**) as a function of the number of samples seen during training for WGAN and satGAN. The distance to the optimal generator corresponds to $\frac{||\hat{\mu} - \mu_k|| + ||\hat{\sigma} - \sigma_k||}{||\hat{\mu} - \mu_0|| + ||\hat{\sigma} - \sigma_0||}$.

For WGAN, we see that stochastic **SGDA** fails to converge and that **L-SVRHG** is the only method to converge linearly on the Hamiltonian. For satGAN, **SGDA** seems to perform best. Algorithms that take into account the Hamiltonian have high variance. We looked at individual runs and found that, in 3 out of 10 runs, the algorithms other than stochastic **SGDA** fail to converge, and the Hamiltonian does not significantly decrease over time. While WGAN is guaranteed to have a unique critical point, which is the solution of the game, this is not the case for satGAN and nsGAN due to the non-linear component. Thus, as expected, Assumption 3.8 is very important in order for the proposed stochastic Hamiltonian methods to perform well.

8. Conclusion and Extensions

We introduce new variants of SHGD (through novel unbiased estimator and step-size selection) and present the first variance reduced Hamiltonian method L-SVRHG. Using tools from optimization literature, we provide convergence guarantees for the two methods and we show how they can efficiently solve stochastic unconstrained bilinear games and the more general class of games that satisfy the "sufficiently bilinear" condition. An important result of our analysis is the first set of global non-asymptotic last-iterate convergence guarantees for a stochastic game over a non-compact domain, in the absence of strong monotonicity assumptions. We believe that our results and the Hamiltonian viewpoint could work as a first step in closing the gap between the stochastic optimization algorithms and methods for solving stochastic games and can open up many avenues for further development and research in both areas. A natural extension of our results will be the proposal of accelerated Hamiltonian methods that use momentum (Loizou & Richtárik, 2017; Assran & Rabbat, 2020) on top of the Hamiltonian gradient update. We speculate that similar ideas to the ones presented in this work can be used for the development of efficient decentralized methods (Assran et al., 2019; Koloskova et al., 2020) for solving problem (3).

Acknowledgements

The authors would like to thank Reyhane Askari, Gauthier Gidel and Lewis Liu for useful discussions and feedback.

Nicolas Loizou acknowledges support by the IVADO postdoctoral funding program. This work was partially supported by the FRQNT new researcher program (2019-NC-257943), the NSERC Discovery grants (RGPIN-2017-06936 and RGPIN-2019-06512) and the Canada CIFAR AI chairs program. Ioannis Mitliagkas acknowledges support by an IVADO startup grant and a Microsoft Research collaborative grant. Simon Lacoste-Julien acknowledges support by a Google Focused Research award. Simon Lacoste-Julien and Pascal Vincent are CIFAR Associate Fellows in the Learning in Machines & Brains program.

References

Abernethy, J., Lai, K. A., and Wibisono, A. Last-iterate convergence rates for min-max optimization. *arXiv preprint arXiv:1906.02027*, 2019.

Albuquerque, I., Monteiro, J., Falk, T. H., and Mitliagkas, I. Adversarial target-invariant representation learning for domain generalization. *arXiv preprint arXiv:1911.00804*, 2019.

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

Assran, M. and Rabbat, M. On the convergence of nesterov's accelerated gradient method in stochastic settings. *arXiv preprint arXiv:2002.12414*, 2020.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. *ICML*, 2019.

Azizian, W., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, 2020a.

Azizian, W., Scieur, D., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. Accelerating smooth games by manipulating spectral shapes. *AISTATS*, 2020b.

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

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Variance reduction for matrix games. In *NeurIPS*, 2019.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. Reducing noise in gan training with variance reduced extragradient. In *NeurIPS*, 2019.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. In *ICLR*, 2018.

Defazio, A. A simple practical accelerated method for finite sums. In *NeurIPS*, 2016.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NeurIPS*, 2014.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2018.

Gidel, G., Hemmat, R. A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. In *AISTATS*, 2019.

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

Gorbunov, E., Hanzely, F., and Richtárik, P. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *AISTATS*, 2020.

Gower, R. M., Richtárik, P., and Bach, F. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *arxiv:1805.02632*, 2018.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. In *ICML*, 2019.

Gower, R. M., Sebbouh, O., and Loizou, N. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. *arXiv preprint arXiv:2006.10311*, 2020.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: stability of stochastic gradient descent. In *ICML*, 2016.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. Variance reduced stochastic gradient descent with neighbors. In *NeurIPS*, 2015.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, 2019.

Ibrahim, A., Azizian, W., Gidel, G., and Mitliagkas, I. Linear lower bounds and conditioning of differentiable games. *arXiv preprint arXiv:1906.07300*, 2019.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NeurIPS*, 2013.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *ECML-PKDD*, 2016.

Khaled, A., Sebbouh, O., Loizou, N., Gower, R. M., and Richtárik, P. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *arXiv preprint arXiv:2006.11573*, 2020.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. A unified theory of decentralized SGD with changing topology and local updates. *arXiv preprint arXiv:2003.10422*, 2020.

Konečný, J., Liu, J., Richtárik, P., and Takáč, M. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10 (2):242–255, 2016.

Korpelevich, G. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Kovalev, D., Horváth, S., and Richtárik, P. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory*, 2020.

Letcher, A., Balduzzi, D., Racanière, S., Martens, J., Foerster, J. N., Tuyls, K., and Graepel, T. Differentiable game mechanics. *Journal of Machine Learning Research*, 20 (84):1–40, 2019.

Loizou, N. and Richtárik, P. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

Loizou, N., Vaswani, S., Laradji, I., and Lacoste-Julien, S. Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence. *arXiv preprint arXiv:2002.10542*, 2020.

Mazumdar, E. V., Jordan, M. I., and Sastry, S. S. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of gans. In *NeurIPS*, 2017.

Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. Revisiting stochastic extragradient. In *AISTATS*, 2020.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, 2020.

Necoara, I., Nesterov, Y., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *Math. Program.*, pp. 1–39, 2018.

Nemirovski, A. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004.

Nemirovski, A. and Yudin, D. B. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. *Soviet Mathetmatics Doklady*, 19, 1978.

Nemirovski, A. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley Interscience, 1983.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. SGD and hogwild! Convergence without the bounded gradients assumption. In *ICML*, 2018.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. Sarah: a novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

Palaniappan, B. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *NeurIPS*, 2016.

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

Pfau, D. and Vinyals, O. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

Polyak, B. Introduction to optimization. translations series in mathematics and engineering. *Optimization Software*, 1987.

Qian, X., Qu, Z., and Richtárik, P. L-SVRG and L-Katyusha with arbitrary sampling. *arXiv preprint arXiv:1906.01481*, 2019.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

Ryu, E. K., Yuan, K., and Yin, W. ODE analysis of stochastic gradient methods with optimism and anchoring for minimax problems and GANs. *arXiv preprint arXiv:1905.10899*, 2019.

Schäfer, F. and Anandkumar, A. Competitive gradient descent. In *NeurIPS*, 2019.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

Sebbouh, O., Gazagnadou, N., Jelassi, S., Bach, F., and Gower, R. Towards closing the gap between the theory and practice of SVRG. In *NeurIPS*, 2019.

Wang, Y., Zhang, G., and Ba, J. On solving minimax optimization locally: A follow-the-ridge approach. In *ICLR*, 2019.

Yang, J., Kiyavash, N., and He, N. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.