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

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

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

$$g(x_1^*, x_2) \le g(x_1^*, x_2^*) \le g(x_1, x_2^*), \qquad (2)$$

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

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

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 extra-gradient 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]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 is a highly inefficient operation (Wang et al., 2019; Mazumdar et al., 2019). 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) for the sufficiently bilinear games.

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 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

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

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 there exists $\rho > 0$ such that for all $x \in \mathbb{R}^d$,

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

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:

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

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):

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

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 $\mathbb{E}_i[\xi_i(x_k)] = \xi(x_k)$.[3]

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

---

[3]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:

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

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:

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

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:

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

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 $L_{\mathcal{H}} = \sigma_{\max}^2(\mathbf{A})$ and $\mu_{\mathcal{H}} = \sigma_{\min}^2(\mathbf{A})$ 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 $\mathcal{H}_{i,j}(x)$:

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

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

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 present in Section 6, the decreasing step-size selection 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 $\{x^i\}_{i=0}^K$.

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.

---

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).

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,[4] the expected smoothness and expected residual parameters take the following values:

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

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

where $L_{\max} = \max_{\{1,...,n^2\}} \{L_{\mathcal{H}_{i,j}}\}$ 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-

---

[4]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:

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

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:

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

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

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

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

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

Lastly, in the following theorem, we show under what se-

lection 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:

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

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 $\gamma^k = \gamma \le \frac{\mu}{L(\mu+2\rho)}$ satisfy:

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

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:

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

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

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

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

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

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**)[5] (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 $[b_i]_k, [c_i]_k \sim \mathcal{N}(0, 1/n)$.

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):

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

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
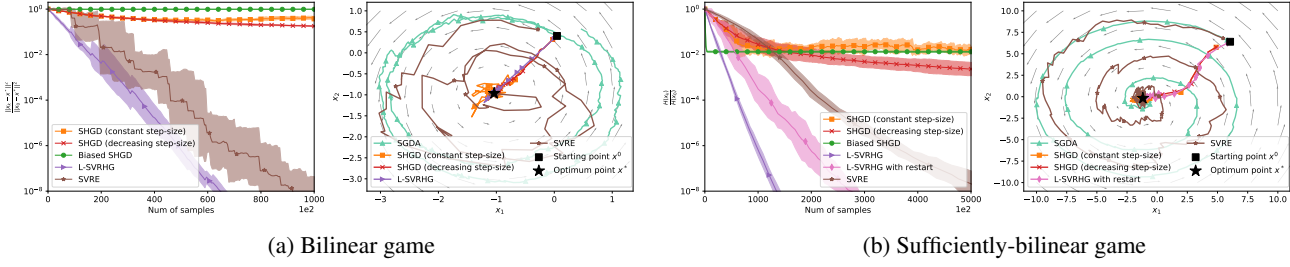
---

[5]**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)

(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.
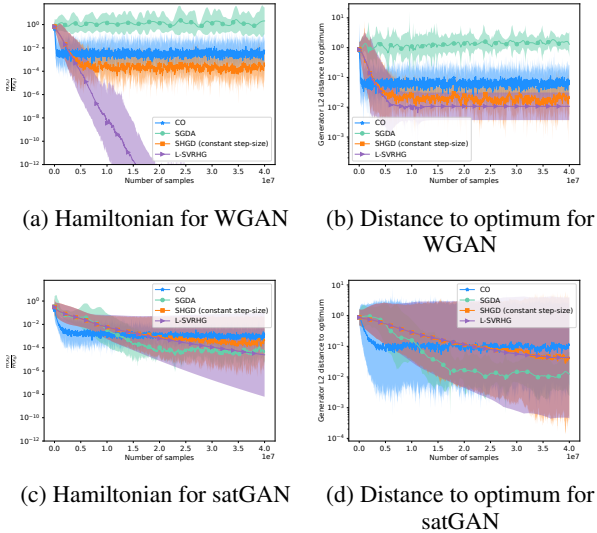
For WGAN, we see that stochastic **SGDA** fails to converge and that **L-SVRHG** is the only method to converge linearly on the Hamiltonian. For satGAN, **SGDA** seems to perform best. Algorithms that take into account the Hamiltonian have high variance. We looked at individual runs and found that, in 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.

(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 $\frac{||\hat{\mu} - \mu_k|| + ||\hat{\sigma} - \sigma_k||}{||\hat{\mu} - \mu_0|| + ||\hat{\sigma} - \sigma_0||}$.

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

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 extragradient 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 Richtrik, 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. Dont 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.

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.

Supplementary Material
Stochastic Hamiltonian Gradient Methods for Smooth Games

In the Appendix we present the proofs of the main Propositions and Theorems proposed in the main paper together with additional experiments on different Bilinear and sufficiently bilinear games.

In particular in Section A, we start by presenting the pseudo-codes of the stochastic optimization algorithms SGD and L-SVRG based on which we build our stochastic Hamiltonian methods. In Section B we provide more details on the assumptions and definitions used in the main paper. In Section D we present the proofs of the two main propositions and in Section E we explain how these propositions can be combined with existing convergence results in order to obtain the Theorems of Section 6. Finally in Sections F and G we present the experimental details and provide additional experiments.

A. Stochastic Optimization Algorithms

In this section we present the pseudocodes of SGD and L-SVRG for solving the finite-sum optimization problem:

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right]. \tag{23}$$

Algorithm 4 Stochastic Gradient Descent (SGD)

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

Algorithm 5 Loopless Stochastic Variance Reduced Gradient (L-SVRG)

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 $k = 0, 2, \cdots, K$ **do**
   Generate fresh sample $i \sim \mathcal{D}$ evaluate $\nabla f_i(x^k)$.
   Evaluate $g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$.
   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 $\{x^i\}_{i=0}^K$.

Algorithm 6 L-SVRG (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-SVRG$_{II}(x^t, K, \gamma, p \in (0, 1])$
end for

Output: The last iterate x^T.

B. Connections of Main Assumptions and Definitions

As we mentioned above SGD (Algorithm 4) and L-SVRG (Algorithm 5) are popular methods for solving the stochastic optimization problem (23). Several convergence analyses of the two algorithms have been proposed under different assumptions on the functions f and f_i. In this section we describe in more details the assumptions used in the analysis of the stochastic Hamiltonian methods in the main paper.

B.1. On Quasi-strong convexity and PL condition

In Section 3.1 we present the definitions of quasi-strong convexity and the PL condition and later in Section 4.2 we explain that for the two classes of games (Bilinear and Sufficiently bilinear) the stochastic Hamiltonian function (11) satisfies one of these conditions. Here using Karimi et al. (2016) we explain the connection between these conditions and the more well-known definition of strong convexity.

Definition B.1 (Strong Convexity). A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, is μ-strongly convex, if there exists a constant $\mu > 0$ such that $\forall x, y \in \mathbb{R}^n$:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \tag{24}$$

In particular the following connection hold:

$$SC \subseteq QSC \subseteq PL, \tag{25}$$

where SC denotes the class of strongly convex functions, QSC the class of quasi-strongly convex (Definition 3.1) and PL the class of functions satisfy the PL condition (Definition 3.2). For more details on the connections of the μ parameter between these methods we refer the reader to Karimi et al. (2016) and Necoara et al. (2018).

B.2. On Smoothness and Expected Smoothness / Expected Residual

In Section 3.1 we present the definitions of Expected Smoothness (ES) and Expected Residual (ER). In the main theoretical results of Section 6 we also use the expected smoothness parameter $\mathcal{L}$ and the expected residual parameter ρ to provide the convergence guarantees of SHGD and L-SVRHG. In this section we provide more details on these assumptions as presented in Gower et al. (2019; 2020).

As explained in Gower et al. (2019; 2020) expected smoothness and expected residual are assumptions that combine both the properties of the distribution $\mathcal{D}$ of drawing samples and the smoothness properties of function f. In particular, ES and ER can be seen as two different ways to measure how far the gradient estimate $\nabla f_i(x)$ is from the true gradient $\nabla f(x)$ where $i \sim \mathcal{D}$.

ES was first used for the analysis of SGD in Gower et al. (2019) for solving stochastic optimization problems of the form (23) where the objective function f is assumed to be μ–quasi-strongly convex (see Definition 3.1). Later in Gower et al. (2020) a similar analysis for SGD has been proposed for functions satisfying the PL condition. As explained in Gower et al. (2020), assuming ES in the analysis of SGD for functions satisfying the PL condition is not ideal as it does not allow the recovery of the the best known dependence on the condition number for the deterministic Gradient Descent (full batch). For this reason Gower et al. (2020) used the notion of Expected residual (ER) in the proposed analysis and explained its benefits.

In both, Gower et al. (2019) and (Gower et al., 2020), the ES and ER assumptions have been used in combination with the arbitrary sampling paradigm. That is, the proposed theorems of Gower et al. (2019; 2020) that describe the convergence of SGD include an infinite array of variants of SGD as special cases. Each one of these variants is associated with a specific probability law governing the data selection rule used to form minibatches.

B.2.1. FORMAL DEFINITIONS

Let us present the definitions of ES and ER as presented in Gower et al. (2019) and Gower et al. (2020). In Gower et al. (2019; 2020) to allow for any form of minibatching the *arbitrary sampling* notation was used. That is,

$$\nabla f_v(x) := \frac{1}{n} \sum_{i=1}^{n} v_i \nabla f_i(x), \tag{26}$$

where $v \in \mathbb{R}^n_+$ is a random *sampling vector* such that $\mathbb{E}[v_i] = 1$, for $i = 1, \ldots, n$ and $f_v(x) := \frac{1}{n} \sum_{i=1}^{n} v_i f_i(x)$. Note that it follows immediately from this definition of sampling vector that $\mathbb{E}[\nabla f_v(x)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[v_i] \nabla f_i(x) = \nabla f(x)$.

In addition note that using the notion of *arbitrary sampling* the update rule of SGD is simply: $x^{k+1} = x^k - \gamma^k \nabla f_v(x^k)$.

Under the notion of *arbitrary sampling* the expected smoothness assumption (Gower et al., 2019) and the expected residual assumption (Gower et al., 2020) take the following form (generalization of the definitions presented in Section 3.1).

Assumption B.2 (Expected Smoothness (ES))**.** We say that f is $\mathcal{L}$–smooth in expectation with respect to a distribution $\mathcal{D}$ if there exists $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$ such that

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x) - \nabla f_v(x^*)\|^2\right] \le 2\mathcal{L}(f(x) - f(x^*)), \tag{27}$$

for all $x \in \mathbb{R}^d$. For simplicity, we will write $(f, \mathcal{D}) \sim ES(\mathcal{L})$ to say that expected smoothness holds.

Assumption B.3 (Expected Residual (ER))**.** We say that f satisfied the expected residual assumption if there exists $\rho = \rho(f, \mathcal{D}) > 0$ such that

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x) - \nabla f_v(x^*) - (\nabla f(x) - \nabla f(x^*))\|^2\right] \le 2\rho(f(x) - f(x^*)). \tag{28}$$

for all $x \in \mathbb{R}^d$. For simplicity, we will write $(f, \mathcal{D}) \sim ER(\rho)$ to say that expected residual holds.

As we explain in Section 6, in this work we focus on τ-minibatch sampling, where in each step we select uniformly at random a minibatch of size $\tau \in [n^2]$ (recall that the Hamiltonian function (11) has n^2 components). However we highlight that the proposed analysis of the stochastic Hamiltonian methods holds for any form of sampling vector following the results presented in Gower et al. (2019; 2020) for the case of SGD and Qian et al. (2019) for the case of L-SVRG methods, including importance sampling variants.

Let us provide a formal definition of the τ-minibatch sampling when $\tau \in [n]$.

Definition B.4 (τ-Minibatch sampling)**.** Let $\tau \in [n]$. We say that $v \in \mathbb{R}^n$ is a τ–minibatch sampling if for every subset $S \in [n]$ with $|S| = \tau$ we have that $\mathbb{P}\left[v = \frac{n}{\tau} \sum_{i \in S} e_i\right] = 1/\binom{n}{\tau} := \tau!(n-\tau)!/n!$

It is easy to verify by using a double counting argument that if v is a τ–minibatch sampling, it is also a valid sampling vector ($\mathbb{E}[v_i] = 1$) (Gower et al., 2019).

Let $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ with functions f_i be L_i–smooth and function f be L-smooth and let $L_{\max} = \max_{\{1,\ldots,n\}}\{L_i\}$. In this setting as it was shown in Gower et al. (2019; 2020) for the case of τ-minibatch sampling ($\tau \in [n]$), the expected smoothness and expected residual parameters and the finite gradient noise σ^2 take the following form:

$$\mathcal{L}(\tau) = \frac{n(\tau-1)}{\tau(n-1)}L + \frac{n-\tau}{\tau(n-1)}L_{\max} \tag{29}$$

$$\rho(\tau) = L_{\max} \frac{n-\tau}{(n-1)\tau} \tag{30}$$

$$\sigma^2(\tau) := \mathbb{E}_{\mathcal{D}}[\|\nabla f_v(x^*)\|^2] = \frac{1}{\tau} \frac{n-\tau}{n-1} \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2. \tag{31}$$

Using the above expressions (29), (30) and (31) it is easy to see that for single element sampling where $\tau = 1$ it holds that $\mathcal{L} = \rho = L_{\max}$ and that $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2$. On the other limit case where a full-batch is used ($\tau = n$), these values

become $\mathcal{L} = L$ and $\rho = \sigma^2 = 0$. Note that these are exactly the values for $\mathcal{L}$, ρ and σ^2 we use in Section 6 with the only difference that $\tau \in [n^2]$ because the stochastic Hamiltonian function (11) has n^2 components $\mathcal{H}_{i,j}$.

In particular, as we explained in Section 6, for the Theorems related to SHGD we use $\sigma^2 := \mathbb{E}_{i,j}[\|\nabla\mathcal{H}_{i,j}(x^*)\|^2]$. From the above expression and for the case of τ-minibatch sampling with $\tau \in [n^2]$ this is equivalent to:

$$\sigma^2 := \mathbb{E}_{i,j}[\|\nabla\mathcal{H}_{i,j}(x^*)\|^2] = \frac{1}{\tau}\frac{n^2 - \tau}{n^2 - 1}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\nabla\mathcal{H}_{i,j}(x^*)\|^2 .$$

Connection between τ-minibatch sampling and sampling step of main algorithms. Note that one of the main steps of Algorithms 1 and 5 is the generation of fresh samples $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ and the evaluation of $\nabla\mathcal{H}_{i,j}(x^k)$. In the case of uniform single element sampling, the samples i and j are selected with probability $p_i = 1/n$ and $p_j = 1/n$ respectively. This is equivalent on selecting samples $\{i,j\}$ uniformly at random from the n^2 components of the Hamiltonian function. In both cases the probability of selecting the component $\mathcal{H}_{i,j}$ is equal to $p_{\mathcal{H}_{i,j}} = 1/n^2$.

In other words, for the case of 1-minibatch sampling (uniform single element sampling), one can simply substitute the sampling step of SHGD and L-SVRHG: "*Generate fresh samples $i \sim \mathcal{D}$ and $j \sim \mathcal{D}$ and evaluate $\nabla\mathcal{H}_{i,j}(x^k)$.*" with the "*Sample uniformly at random the component $H_{i,j}$ and evaluate $\nabla\mathcal{H}_{i,j}(x^k)$.*"

Trivially, using the definition (B.4) and the above notion of *sampling vector*, this connection can be extended to capture the more general τ-minibatch sampling where $\tau \in [n^2]$. In this case, we will have

$$\nabla\mathcal{H}_v(x) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}v_{i,j}\nabla\mathcal{H}_{i,j}(x),$$

where $v \in \mathbb{R}_+^{n^2}$ is a random *sampling vector* such that $\mathbb{E}_{i,j}[v_{i,j}] = 1$, for $i = 1, \ldots, n$, and $j = 1, \ldots, n$ and $\mathcal{H}_v(x) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}v_{i,j}\mathcal{H}_{i,j}(x)$. Note that it follows immediately from this definition of sampling vector that $\mathbb{E}[\nabla\mathcal{H}_v(x)] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_{i,j}[v_{i,j}]\nabla\mathcal{H}_{i,j}(x) = \nabla\mathcal{H}(x)$.

In this case the update rule of SHGD (Algorithm 1) will simply be: $x^{k+1} = x^k - \gamma^k\nabla\mathcal{H}_v(x)$ and the proposed theoretical results will still hold.

B.2.2. SUFFICIENT CONDITIONS AND CONNECTIONS BETWEEN NOTIONS OF SMOOTHNESS.

In Gower et al. (2019) it was proved that convexity and L_i–smoothness of f_i in problem (23) implies expected smoothness of function f. However, the opposite implication does not hold. The expected smoothness assumption can hold even when the f_i's and f are not convex. See (Gower et al., 2019) for more details.

Similar results have been shown in Gower et al. (2020) for the case of expected residual. More specifically, it was shown that if the functions f_i of problem (23) are L_i–smooth and also x^*-convex for $x^* \in \mathcal{X}^*$ (where $\mathcal{X}^*$ is solution set of f) then function f satisfies the expected residual conditions, that is $(f, \mathcal{D}) \sim ER(\rho)$ and the expected residual parameter ρ has a meaningful expression.

Another interesting connections between the smoothness parameters is the following Gower et al. (2019). If we assume that function f of problem (23) is Lsmooth and that each f_i function is $L_i$smooth then the expected smoothness $\mathcal{L}$ constant is bounded as follows:

$$L \leq \mathcal{L} \leq L_{\max},$$

where $L_{\max} = \max L_i{}_{i=1}^n$.

Let us also present the following lemma as proved in (Gower et al., 2020) that connects the ES and ER assumptions.

Lemma B.5. (Expected smoothness implies Expected Residual, from (Gower et al., 2020).) If function f of problem (23) satisfies the expected smoothness $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then it satisfies the expected residual $(f, \mathcal{D}) \sim ER(\rho)$ with $\rho = \mathcal{L}$. If in addition the function satisfied the PL condition that satisfied the expected residual with $\rho = \mathcal{L} - \mu$.

B.2.3. BOUNDS ON THE STOCHASTIC GRADIENT

A common assumption used to prove the convergence of SGD is uniform boundedness of the stochastic gradients: there exist $0 < c < \infty$ such that $\mathbb{E}\|\nabla f_v(x)\|^2 \leq c$ for all x. However, this assumption often does not hold, such as in the case

when f is strongly convex (Nguyen et al., 2018). Recall that the class of μ-strongly convex functions is a special case of both the μ-quasi strongly convex and functions satisfying the PL condition (see (25)).

Using ES and ER in the proposed theorems we do not need to assume such a bound. Instead, we use the following direct consequence of expected smoothness and expected residual to bound the expected norm of the stochastic gradients.

Lemma B.6. (Gower et al., 2019) If $(f, \mathcal{D}) \sim ES(\mathcal{L})$, then

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x)\|^2\right] \leq 4\mathcal{L}(f(x) - f(x^*)) + 2\sigma^2, \tag{32}$$

where $\sigma^2 := \mathbb{E}_{\mathcal{D}}[\|\nabla f_v(x^*)\|^2]$.

Similar upper bound on the stochastic gradients can be obtained if one assumed expected residual:

Lemma B.7. (Gower et al., 2020) If $(f, \mathcal{D}) \sim ER(\rho)$ then

$$\mathbb{E}_{\mathcal{D}}\left[\|\nabla f_v(x)\|^2\right] \leq 4\rho(f(x) - f^*) + \|\nabla f(x)\|^2 + 2\sigma^2. \tag{33}$$

where $\sigma^2 := \mathbb{E}_{\mathcal{D}}[\|\nabla f_v(x^*)\|^2]$.

C. On Stochastic Hamiltonian Function and Unbiased Estimator of the Gradient

C.1. Finite-Sum Structure of Hamiltonian Methods

Having g to be a finite sum function leads to the following derivations on the Hamiltonian functions and gradients:

$$
\begin{aligned}
\mathcal{H}(x) &= \frac{1}{2}\|\xi(x)\|^2 = \frac{1}{2}\langle\xi(x), \xi(x)\rangle = \frac{1}{2}\langle\frac{1}{n}\sum_{i=1}^{n}\xi_i(x), \frac{1}{n}\sum_{j=1}^{n}\xi_j(x)\rangle \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\underbrace{\frac{1}{2}\langle\xi_i(x), \xi_j(x)\rangle}_{\mathcal{H}_{i,j}(x)} = \frac{1}{n^2}\sum_{i,j=1}^{n}\underbrace{\frac{1}{2}\langle\xi_i(x), \xi_j(x)\rangle}_{\mathcal{H}_{i,j}(x)}
\end{aligned}
\tag{34}
$$

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

C.2. Unbiased Estimator of the Full Gradient

The gradient of $\mathcal{H}_{i,j}(x)$ has the following form:

$$\nabla\mathcal{H}_{i,j}(x) = \frac{1}{2}\nabla\langle\xi_i(x), \xi_j(x)\rangle = \frac{1}{2}\left[\langle\nabla\xi_i(x), \xi_j(x)\rangle + \langle\xi_i(x), \nabla\xi_j(x)\rangle\right] = \frac{1}{2}\left[\mathbf{J}_i^\top\xi_j + \mathbf{J}_j^\top\xi_i\right], \tag{35}$$

and it is an unbiased estimator of the full gradient. That is, $\nabla\mathcal{H}(x) = \mathbb{E}_{i,j}[\nabla\mathcal{H}_{i,j}(x)]$.

$$
\begin{aligned}
\nabla\mathcal{H}(x) &= \nabla\frac{1}{2}\|\xi(x)\|^2 = \nabla\frac{1}{2}\langle\xi(x), \xi(x)\rangle = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\nabla\langle\xi_i(x), \xi_j(x)\rangle \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\left[\langle\nabla\xi_i(x), \xi_j(x)\rangle + \langle\xi_i(x), \nabla\xi_j(x)\rangle\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\underbrace{\frac{1}{2}\left[\mathbf{J}_i^\top\xi_j + \mathbf{J}_j^\top\xi_i\right]}_{\nabla\mathcal{H}_{i,j}(x)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{n}\nabla\mathcal{H}_{i,j}(x) \\
&= \mathbb{E}_i\mathbb{E}_j[\nabla\mathcal{H}_{i,j}(x)] = \mathbb{E}_{i,j}[\nabla\mathcal{H}_{i,j}(x)]
\end{aligned}
\tag{36}
$$

C.3. Beyond Finite-Sum

All results presented in the main paper related to SHGD can be trivially extended beyond finite sum problems (with exactly the same rates). The finite-sum structure is required only for the variance reduced method L-SVRHG. In the stochastic case, problem (3) will be

$$\min_{x_1 \in \mathbb{R}^{d_1}} \max_{x_2 \in \mathbb{R}^{d_2}} g(x_1, x_2) = E_\zeta[g(x, \zeta)]$$

where ζ is a random variable obeying some distribution. Then $\xi(x) = E_\zeta[\xi(x, \zeta)]$, $\mathbf{J} = E_\zeta[\mathbf{J}(x, \zeta)]$ and the stochastic Hamiltonian function will become

$$H(x) = E_{\zeta_i} E_{\zeta_j} \underbrace{\frac{1}{2} \langle \xi(x, \zeta_i), \xi(x, \zeta_j) \rangle}_{\mathcal{H}_{i,j}(x)}.$$

In this case $\nabla \mathcal{H}_{i,j}(x) = \frac{1}{2} \left[\mathbf{J}(x, \zeta_i)^\top \xi(x, \zeta_j) + \mathbf{J}(x, \zeta_j)^\top \xi(x, \zeta_i) \right]$ and $\nabla \mathcal{H}(x) = E_{\zeta_i} E_{\zeta_j}[\nabla \mathcal{H}_{i,j}(x)]$.

In this case SHGD will execute the following updates in each step $k \in \{0, 1, 2, \cdots, K\}$:

1. Generate i.i.d random variables ζ_i and ζ_j and evaluate $\nabla \mathcal{H}_{i,j}(x^k)$.

2. Set step-size γ^k following one of the selected choices (constant, decreasing)

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

D. Proofs of Main Propositions

Let us first present the main notation used for eigenvalues and singular values (similar to the main paper).

Eigenvalues, singular values Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. We denote with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ its eigenvalues. Let $\lambda_{\min} = \lambda_1$ be the smallest non-zero eigenvalue, and $\lambda_{\max} = \lambda_n$ be the largest eigenvalue. With $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$ we denote its singular values. With $\sigma_{\max}$ and $\sigma_{\min}$ we denote the maximum singular value and the minimum non-zero singular value of matrix $\mathbf{A}$.

D.1. Proof of Proposition 4.3

In our proof we use the following result proved in Necoara et al. (2018).

Lemma D.1 (Necoara et al. (2018)). *Let function $z : \mathbb{R}^m \to \mathbb{R}$ be μ_z-strongly convex with L_z-Lipschitz continuous gradient and $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a nonzero matrix. Then, the convex function $f(x) = z(Ax)$ is a smooth μ–quasi-strongly convex function with constants $L = L_z \|\mathbf{A}\|^2$ and $\mu = \mu_z \sigma_{\min}^2(\mathbf{A})$ where $\sigma_{\min}(\mathbf{A})$ is the smallest nonzero singular value of matrix $\mathbf{A}$ and $\|\mathbf{A}\|$ is the spectral norm.*

Proof. Recall that in the stochastic bilinear game we have that:

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

By evaluating the partial derivatives we obtain:

- $\nabla_{x_1} g_i(x_1, x_2) = \mathbf{A}_i x_2 + b_i \in \mathbb{R}^{d_1 \times 1}, \forall i \in [n]$

- $\nabla_{x_2} g_i(x_1, x_2) = \mathbf{A}_i^\top x_1 + c_i \in \mathbb{R}^{d_2 \times 1}, \forall i \in [n]$

Thus, from definition of $\xi_i(x)$ we get:

$$\xi_i(x) = (\nabla_{x_1} g_i, -\nabla_{x_2} g_i) = \left(\mathbf{A}_i x_2 + b_i, -[\mathbf{A}_i^\top x_1 + c_i] \right) \quad \forall i \in [n]$$

and as a result by simple computations:

$$
\begin{aligned}
\mathcal{H}_{i,j}(x) &= \frac{1}{2}\langle \xi_i(x), \xi_j(x)\rangle \\
&= \frac{1}{2}\left\langle \left(\mathbf{A}_i x_2 + b_i, -[\mathbf{A}_i^\top x_1 + c_i]\right), \left(\mathbf{A}_j x_2 + b_j, -[\mathbf{A}_j^\top x_1 + c_j]\right)\right\rangle \\
&= \frac{1}{2}(x_1, x_2)^\top \begin{pmatrix} \mathbf{A}_i \mathbf{A}_j^\top & 0 \\ 0 & \mathbf{A}_i^\top \mathbf{A}_j \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2}\left(c_j^\top \mathbf{A}_i^\top + c_i^\top \mathbf{A}_j^\top, b_j^\top \mathbf{A}_i + b_i^\top \mathbf{A}_j\right)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2}c_i^\top c_j + \frac{1}{2}b_i^\top b_j \\
&= \frac{1}{2}x^\top \mathbf{Q}_{i,j} x + q_{i,j}^\top x + \ell_{i,j},
\end{aligned}
\tag{37}
$$

where $\mathbf{Q}_{i,j} = \begin{pmatrix} \mathbf{A}_i \mathbf{A}_j^\top & 0 \\ 0 & \mathbf{A}_i^\top \mathbf{A}_j \end{pmatrix}$ and

$q_{i,j}^\top = \frac{1}{2}\left(c_j^\top \mathbf{A}_i^\top + c_i^\top \mathbf{A}_j^\top, b_j^\top \mathbf{A}_i + b_i^\top \mathbf{A}_j\right)$ and $\ell_{i,j} = \frac{1}{2}c_i^\top c_j + \frac{1}{2}b_i^\top b_j$.

Using the finite-sum structure of the Hamiltonian function (11) the stochastic Hamiltonian function takes the following form:

$$
\begin{aligned}
\mathcal{H}(x) &= \frac{1}{n^2}\sum_{i,j=1}^n \mathcal{H}_{i,j}(x) \\
&= \frac{1}{n^2}\sum_{i,j=1}^n \frac{1}{2}x^\top \mathbf{Q}_{i,j} x + q_{i,j}^\top x + \ell_{i,j} \\
&= \frac{1}{2}x^\top \mathbf{Q} x + q^\top x + \ell
\end{aligned}
\tag{38}
$$

where $\mathbf{Q} = \left[\frac{1}{n^2}\sum_{i,j=1}^n \mathbf{Q}_{i,j}\right] = \begin{pmatrix} \mathbf{A}\mathbf{A}^\top & 0 \\ 0 & \mathbf{A}^\top \mathbf{A} \end{pmatrix}$ with $\mathbf{A} = \frac{1}{n}\sum_{i=1}^n \mathbf{A}_i$ and

$q^\top = \left[\frac{1}{n^2}\sum_{i,j=1}^n q_{i,j}^\top\right]$ and $\ell = \left[\frac{1}{n^2}\sum_{i,j=1}^n \ell_{i,j}\right]$.

Thus, the stochastic Hamiltonian function

$$
\boxed{\mathcal{H}(x) = \frac{1}{2}x^\top \mathbf{Q} x + q^\top x + \ell}
$$

is a quadratic smooth function (not necessarily strongly convex).

Note also that throughout the paper we assume that a stationary point of function g exist. This in combination with Assumption 3.8 guarantees that the Hamiltonian function $\mathcal{H}(x)$ has at least on minimum point x^*. That is, there exist x^* such that $\nabla \mathcal{H}(x^*) = \mathbf{Q}x^* + q = 0$.

Here matrix $\mathbf{Q}$ is symmetric and $\mathbf{Q} \succeq 0$. Thus, let $\mathbf{Q} = \mathbf{L}_\mathbf{Q}^\top \mathbf{L}_\mathbf{Q}$ be the Cholesky decomposition of matrix $\mathbf{Q}$. In addition, since we already assume that the Hamiltonian function $\mathcal{H}(x)$ has at least on minimum point x^* we have that $q = -\mathbf{Q}x^* = -\mathbf{L}_\mathbf{Q}^\top \mathbf{L}_\mathbf{Q} x^*$.

Using this note that:

$$
\mathcal{H}(x) = \phi(\mathbf{L}_\mathbf{Q} x)
$$

where function $\phi(y) = \frac{1}{2}\|y\|^2 - (\mathbf{L}_\mathbf{Q} x^*)^\top y + \ell$. In addition, note that function ϕ is 1-strongly convex with 1-Lipschitz continuous gradient.

Thus, using Lemma D.1 we have that the the Hamiltonian function is a $L_\mathcal{H}$−smooth, $\mu_\mathcal{H}$−quasi-strongly convex function with constants $L_\mathcal{H} = \|\mathbf{L}_\mathbf{Q}\|^2 = \lambda_{\max}(\mathbf{L}_\mathbf{Q}^\top \mathbf{L}_\mathbf{Q}) = \lambda_{\max}(\mathbf{Q}) = \sigma_{\max}^2(\mathbf{A})$ and $\mu_\mathcal{H} = \sigma_{\min}^2(\mathbf{L}_\mathbf{Q}) = \lambda_{\min}^+(\mathbf{Q}) = \sigma_{\min}^2(\mathbf{A})$.

This completes the proof. $\qquad\square$

D.2. Proof of Proposition 4.4

Proof. Our proof follows closely the approach of Abernethy et al. (2019). The main difference is that our game by its definition is stochastic and as a result quantities like $\bar{S} = \mathbb{E}_i[S_i]$ and $\bar{L} = \mathbb{E}_i[L_i]$ appear in the expression of $L_{\mathcal{H}}$.

We divide the proof into two parts. In the first one we show that the Hamiltonian function is smooth and in the next one that satisfies the PL condition.

Smoothness. Note that:

$$\nabla \mathcal{H}(x) \overset{(36)}{=} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \underbrace{\frac{1}{2} \left[\mathbf{J}_i^\top \xi_j + \mathbf{J}_j^\top \xi_i \right]}_{\nabla \mathcal{H}_{i,j}(x)} = \frac{1}{n^2} \left[\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{J}_i^\top \xi_j + \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \mathbf{J}_j^\top \xi_i \right] = \frac{1}{n^2} \left[\sum_{i=1}^n \sum_{j=1}^n \mathbf{J}_i^\top \xi_j \right] \quad (39)$$

Thus,

$$\begin{aligned}
\|\nabla \mathcal{H}(x) - \nabla \mathcal{H}(y)\| &\overset{(39)}{=} \left\| \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{J}_i^\top(x)\xi_j(x) - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{J}_i^\top(y)\xi_j(y) \right\| \\
&= \left\| \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left[\mathbf{J}_i^\top(x)\xi_j(x) - \mathbf{J}_i^\top(y)\xi_j(y) \right] \right\| \\
&\overset{Jensen}{\leq} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left\| \mathbf{J}_i^\top(x)\xi_j(x) - \mathbf{J}_i^\top(y)\xi_j(y) \right\| \\
&\overset{(*)}{\leq} \mathbb{E}_i \mathbb{E}_j \left\| \mathbf{J}_i^\top(x)\xi_j(x) - \mathbf{J}_i^\top(y)\xi_j(y) \right\| \\
&= \mathbb{E}_i \mathbb{E}_j \left\| \mathbf{J}_i^\top(x)\xi_j(x) + \mathbf{J}_i^\top(y)\xi_j(x) - \mathbf{J}_i^\top(y)\xi_j(x) - \mathbf{J}_i^\top(y)\xi_j(y) \right\| \\
&= \mathbb{E}_i \mathbb{E}_j \left\| \left[\mathbf{J}_i^\top(x) - \mathbf{J}_i^\top(y) \right] \xi_j(x) + \mathbf{J}_i^\top(y) \left[\xi_j(x) - \xi_j(y) \right] \right\| \\
&\leq \mathbb{E}_i \mathbb{E}_j \left\| \left[\mathbf{J}_i^\top(x) - \mathbf{J}_i^\top(y) \right] \xi_j(x) \right\| + \mathbb{E}_i \mathbb{E}_j \left\| \mathbf{J}_i^\top(y) \left[\xi_j(x) - \xi_j(y) \right] \right\| \\
&= \mathbb{E}_i \left\| \mathbf{J}_i^\top(x) - \mathbf{J}_i^\top(y) \right\| \mathbb{E}_j \|\xi_j(x)\| + \mathbb{E}_i \left\| \mathbf{J}_i^\top(y) \right\| \mathbb{E}_j \|\xi_j(x) - \xi_j(y)\| \\
&\overset{Assumption\ 4.1}{\leq} \mathbb{E}_i \left[S_i \|x - y\| \right] \mathbb{E}_j \|\xi_j(x)\| + \mathbb{E}_i[L_i] \mathbb{E}_j \left[L_j \|x - y\| \right] \\
&\overset{\mathbb{E}_j \|\xi_j(x)\| < C}{\leq} C\bar{S} \|x - y\| + \bar{L}^2 \|x - y\| \\
&= \left(\bar{S}C + \bar{L}^2 \right) \|x - y\| \qquad (40)
\end{aligned}$$

where in $(*)$ we use that i and j are sampled from a uniform distribution.

PL Condition. To show that the Hamiltonian function satisfies the PL condition (3.2) we use a linear algebra lemma from (Abernethy et al., 2019).

Lemma D.2. (Lemma H.2 in (Abernethy et al., 2019))
Let matrix $\mathbf{M} = \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ -\mathbf{C}^\top & -\mathbf{B} \end{pmatrix}$ where matrix $\mathbf{C}$ is a square full rank matrix. Let $c = \left(\sigma_{\min}^2(C) + \lambda_{\min}(\mathbf{A}^2) \right) \left(\lambda_{\min}(\mathbf{B}^2) + \sigma_{\min}^2(\mathbf{C}) \right) - \sigma_{\max}^2(\mathbf{C}) \left(\|\mathbf{A}\| + \|B\| \right)^2$ and let assume that $c > 0$. Here $\lambda_{\min}$ denotes the smaller eigenvalue and $\sigma_{\min}$ and $\sigma_{\max}$ the smallest and largest singular values respectively. Then if λ is an eigenvalue of $\mathbf{M}\mathbf{M}^\top$ it holds that:

$$\lambda > \frac{\left(\sigma_{\min}^2(C) + \lambda_{\min}(\mathbf{A}^2) \right) \left(\lambda_{\min}(\mathbf{B}^2) + \sigma_{\min}^2(\mathbf{C}) \right) - \sigma_{\max}^2(\mathbf{C}) \left(\|\mathbf{A}\| + \|B\| \right)^2}{\left(2\sigma_{\min}^2(\mathbf{C}) + \lambda(\mathbf{A}^2) + \lambda_{\min}(\mathbf{B}^2) \right)^2}$$

In addition note that if there exist $\mu > 0$ such that $\mathbf{J}(x)\mathbf{J}^\top(x) \succeq \mu\mathbf{I}$ then the Hamiltonian function satisfies the PL condition with parameter μ.

Lemma D.3. Let $g(x)$ of min-max problem 3 be twice differentiable function. If there exist $\mu > 0$ such that $\mathbf{J}(x)\mathbf{J}^\top(x) \succeq \mu\mathbf{I}$ for all $x \in \mathbb{R}^d$ then the Hamiltonian function $\mathcal{H}(x)$ (11) satisfies the PL condition (3.2) with parameter μ.

Proof.

$$\frac{1}{2}\|\nabla\mathcal{H}(x)\|^2 = \frac{1}{2}\|\mathbf{J}^\top(x)\xi(x)\|^2 \qquad = \qquad \frac{1}{2}\xi(x)^\top\mathbf{J}(x)\mathbf{J}^\top(x)\xi(x)$$

$$\overset{\mathbf{J}(x)\mathbf{J}^\top(x)\succeq\mu\mathbf{I}}{\geq} \quad \frac{\mu}{2}\xi(x)^\top\xi(x)$$

$$= \qquad \mu\frac{1}{2}\|\xi(x)\|^2$$

$$= \qquad \mu\left[\mathcal{H}(x)\right]$$

$$\overset{\mathcal{H}(x^*)=0}{=} \quad \mu\left[\mathcal{H}(x) - \mathcal{H}(x^*)\right] \tag{41}$$

$\square$

Combining the above two lemmas we can now show that for the sufficiently bilinear games that Hamiltonian function satisfies the PL condition with parameter $\mu_\mathcal{H} = \frac{(\delta^2+\rho^2)(\delta^2+\beta^2)-4L^2\Delta^2}{2\delta^2+\rho^2+\beta^2}$.

Recall that $\mathbf{J} = \nabla\xi = \begin{pmatrix} \nabla^2_{x_1,x_1}g & \nabla^2_{x_1,x_2}g \\ -\nabla^2_{x_2,x_1}g & -\nabla^2_{x_2,x_2}g \end{pmatrix} \in \mathbb{R}^{d\times d}$. Now let $\mathbf{C}(x) = \nabla^2_{x_1,x_2}g(x)$ and note that this is a square full rank matrix. In particular, by the assumption of the sufficiently bilinear games we have that the cross derivative $\nabla^2_{x_1,x_2}g$ is full rank matrix 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. In addition since we assume that the sufficiently bilinear condition (10) holds we can apply Lemma D.2 with $\mathbf{M} = \mathbf{J}$. Since function g is smooth in x_1 and x_2 and using the bounds on the singular values of matrix $\mathbf{C}(x)$ we have that:

$$\mathbf{J}\mathbf{J}^\top \succeq \left[\frac{(\delta^2 + \rho^2)(\delta^2 + \beta^2) - 4L^2\Delta^2}{2\delta^2 + \rho^2 + \beta^2}\right]\mathbf{I},$$

where $\rho^2 = \min_{x_1,x_2}\lambda_{\min}\left[\nabla^2_{x_1,x_1}g(x_1,x_2)\right]^2$ and $\beta^2 = \min_{x_1,x_2}\lambda_{\min}\left[\nabla^2_{x_2,x_2}g(x_1,x_2)\right]^2$. Using Lemma D.3 it is clear that the Hamiltonian function of the sufficiently bilinear games satisfies the PL condition with $\mu_\mathcal{H} = \frac{(\delta^2+\rho^2)(\delta^2+\beta^2)-4L^2\Delta^2}{2\delta^2+\rho^2+\beta^2}$.

This completes the proof. $\square$

E. Proofs of Main Theorems

In this section we present the proofs of the convergence analysis Theorems presented in Section 6 for the convergence of SHGD (constant and decreasing step-size) and L-SVRHG (and its variant for PL functions) for solving the bilinear games and sufficiently bilinear games.

E.1. Derivation of Convergence Results

The Theorems of the papers can be obtained by combining existing and new optimization convergence results with the two main proposition proved in the previous section (Propositions 4.3 and 4.4).

In particular we use the following combination of results:

- **For the Bilinear Games:**
  - Convergence of SHGD with constant step-size (Theorem 6.1): Combination of constant step-size SGD theorem from Gower et al. (2019) and Proposition 4.3.
  - Convergence of SHGD with decreasing step-size (Theorem 6.3): Combination of decreasing step-size SGD theorem from Gower et al. (2019) and Proposition 4.3.
  - Convergence of L-SVRHG (Theorem 6.4): Combination of the convergence of L-SVRG from Kovalev et al. (2020); Gorbunov et al. (2020) and Proposition 4.3.

- **For the Sufficiently Bilinear Games:**
  - Convergence of SHGD with constant step-size (Theorem 6.5): Combination of constant step-size SGD theorem from Gower et al. (2020) and Proposition 4.4.
  - Convergence of SHGD with decreasing step-size (Theorem 6.7): Combination of decreasing step-size SGD theorem from Gower et al. (2020) and Proposition 4.4.
  - Convergence of L-SVRHG with Restarts (Theorem 6.8): Combination of Theorem E.8 describing the convergence of Algorithm 6 in the optimization setting (extension of the convergence results from Qian et al. (2019)) and Proposition 4.4.

In the rest of this section we present the Theorems of the convergence of SGD (Algorithm 4) and L-SVRG (Algorithm 5) for solving the finite sum problem (23) as presented in the above papers with some brief comments on their convergence. As we explain above combining these results with the Propositions 4.3 and 4.4 yield the Theorems presented in Section 6.

E.2. Convergence of Stochastic Optimization Methods for μ–Quasi-strongly Convex Functions

In this subsection we present the main convergence results as presented in Gower et al. (2019) for SGD and in Kovalev et al. (2020); Gorbunov et al. (2020) for L-SVRG. The main assumption of these Theorems is the function f of problem (23) to be μ–quasi-strongly convex function and that the expected smoothness is satisfied. Note that no assumption on convexity of f_i is made.

E.2.1. CONVERGENCE OF SHGD

Two Theorems have been presented in Gower et al. (2019) for the convergence of SGD one for constant step-size and one with a decreasing step-size. In particular the second theorem provide insightful stepsize-switching rules which describe when one should switch from a constant to a decreasing stepsize regime. As expected for the choice of constant step-size, SGD converges with linear rate to a neighborhood of the solution while for the decreasing step-size converges to the exact optimum but with a slower sublinear rate.

For the case of our stochastic Hamiltonian methods this is exactly the behavior of the SHGD where for the constant step-size the method convergence to a neighborhood of the min-max solution while for the case of decreasing step-size the method converges with a slower rate to the min-max solution of problem (3).

Theorem E.1 (Constant Stepsize). Assume f is μ-quasi-strongly convex and that $(f, \mathcal{D}) \sim ES(\mathcal{L})$. Choose $\gamma^k = \gamma \in (0, \frac{1}{2\mathcal{L}}]$ for all k. Then iterates of Stochastic Gradient Descent (SGD) given in Algorithm 4 satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \gamma\mu)^k \|x^0 - x^*\|^2 + \frac{2\gamma\sigma^2}{\mu}. \tag{42}$$

Theorem E.2 (Decreasing stepsizes/switching strategy). Assume f is μ-quasi-strongly convex and that $(f, \mathcal{D}) \sim ES(\mathcal{L})$. Let $\mathcal{K} := \mathcal{L}/\mu$ and

$$\gamma^k = \begin{cases} \frac{1}{2\mathcal{L}} & \text{for} \quad k \leq 4\lceil \mathcal{K} \rceil \\ \frac{2k+1}{(k+1)^2\mu} & \text{for} \quad k > 4\lceil \mathcal{K} \rceil. \end{cases} \tag{43}$$

If $k \geq 4\lceil \mathcal{K} \rceil$, then Stochastic Gradient Descent (SGD) given in Algorithm 4 satisfy:

$$\mathbb{E}\|x^k - x^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{k} + \frac{16\lceil \mathcal{K} \rceil^2}{e^2 k^2}\|x^0 - x^*\|^2. \tag{44}$$

E.2.2. CONVERGENCE OF L-SVRG

As we explained in the main paper L-SVRG is a variance reduced method which means that allow us to obtain convergence to the exact solution of the problem. The methods is analyzed in Kovalev et al. (2020) for the case of strongly convex functions and extended to the class of μ-quasi strongly convex in Gorbunov et al. (2020). Following the theorem proposed in Gorbunov et al. (2020) it can be shown that the method converges to the solution x^* with linear rate as follows:

Theorem E.3. Assume f is μ-quasi-strongly convex. Let step-size $\gamma = 1/6L$ and let $p \in (0,1]$ and $\mathcal{D}$ be the uniform distribution. Then L-SVRG$_I$ given in Algorithm 5 convergences to the optimum and satisfies:

$$\mathbb{E}[\Phi^k] \leq \max\{1 - \mu/6L, 1 - p/2\}^k \Phi^0$$

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

Note that in statement of Theorem 6.4 we replace n in the above expression with n^2 because the Hamiltonian function has finite-sum structure with n^2 components. As we explain before for obtaining Theorem 6.4 of the main paper one can simply combine the above Theorem with Proposition 4.3.

We highlight that in (Qian et al., 2019) a convergence Theorem of L-SVRG for smooth strongly convex functions was presented under the arbitrary sampling paradigm (well defined distribution $\mathcal{D}$) . This result can be trivially extended to capture the class of smooth quasi-strongly convex functions and as such it can be also used in combination with 4.3. In this case the step-size will become $\gamma = 1/6\mathcal{L}$ where $\mathcal{L}$ is the expected smoothness parameter. Using this, one can guarantee linear convergence of L-SVRG, and as a result of L-SVRHG, with more general distribution $\mathcal{D}$ (beyond uniform sampling). For other well defined choices of distribution $\mathcal{D}$ we refer the interested reader to (Qian et al., 2019).

E.3. Convergence of Stochastic Optimization Methods for Functions Satisfying the PL Condition

As we have already mentioned the Theorems of the convergence of the stochastic Hamiltonian methods for solving the sufficiently bilinear games can be obtain by combining Proposition 4.4 with existing results on the analysis of SGD and L-SVRG for functions satisfying the PL condition.

In particular, in this subsection we present the main convergence Theorems as presented in Gower et al. (2020) for the analysis of SGD for functions satisfying the PL condition and we explain how we can extend the results of Qian et al. (2019) in order to provide an analysis of L-SVRG with restart.

The main assumption of these Theorems is that function f of problem (23) satisfies the PL condition and that the expected residual is satisfied. Note that again no assumption on convexity of f_i is made.

An important remark that we need to highlight is that all convergence result are presented in terms of function suboptimality $\mathbb{E}[f(x^k) - f(x^*)]$. When these results are used for the Hamiltonian method that we know that $\mathcal{H}(x^*) = 0$ they can be written as $\mathbb{E}[\mathcal{H}(x^k)]$. This is exactly the quantity for which we show convergence in Theorems 6.5, 6.7 and 6.8.

E.3.1. CONVERGENCE OF SHGD

In Gower et al. (2020) several convergence theorems describing the performance of SGD were presented for two large classes of structured nonconvex functions: (i) the Quasar (Strongly) Convex functions and (ii) the functions satisfying the Polyak-Lojasiewicz (PL) condition. The proposed analysis of Gower et al. (2020) relied on the Expected Residual assumption 6. The authors proved the weakness of this assumption compared to previous one used for the analysis of SGD and highlight the benefits of using it when the function satisfying the PL condition.

In particular, one of the main contributions of this work is a novel analysis of minibatch SGD for PL functions which recovers the best known dependence on the condition number for Gradient Descent (GD) while also matching the current state-of-the-art rate derived in for SGD. Recall that this is what we used to obtain the convergence of deterministic Hamiltonian method (equivalent to GD) in Corollary 6.6.

In Gower et al. (2020) two theorems have been proposed for PL functions for the convergence of SGD, one for constant step-size and one with a decreasing step-size. In particular the second theorem provide insightful stepsize-switching rules which describe when one should switch from a constant to a decreasing stepsize regime. For the case of constant step-size, SGD converges with linear rate to a neighborhood of the solution while for the decreasing step-size converges to the exact optimum but with a slower sublinear rate.

Theorem E.4. Let f be L-smooth. Assume expected residual and that $f(x)$ satisfies the PL condition (4). Set $\sigma^2 =$

$\mathbb{E}(\|\nabla f_i(x^*)\|^2)$, where $x^* = \arg\min_x f(x)$. Let $\gamma_k = \gamma \leq \frac{\mu}{L(\mu+2\rho)}$, for all k. Then the iterates of SGD satisfy:

$$\mathbb{E}[f(x^k) - f^*] \leq (1 - \gamma\mu)^k [f(x^0) - f^*] + \frac{L\gamma\sigma^2}{\mu}. \tag{45}$$

Theorem E.5 (Decreasing step sizes/switching strategy)**.** Let f be an L-smooth. Assume expected residual and that $f(x)$ satisfies the PL condition (4). Let $k^* := 2\frac{L}{\mu}\left(1 + 2\frac{\rho}{\mu}\right)$ and

$$\gamma^k = \begin{cases} \dfrac{\mu}{L(\mu + 2\rho)}, & \text{for} \quad k \leq \lceil k^* \rceil \\ \dfrac{2k + 1}{(k+1)^2 \mu} & \text{for} \quad k > \lceil k^* \rceil \end{cases} \tag{46}$$

If $k \geq \lceil k^* \rceil$, then SGD given in Algorithm 4 satisfies:

$$\mathbb{E}[f(x^k) - f^*] \leq \frac{4L\sigma^2}{\mu^2}\frac{1}{k} + \frac{(k^*)^2}{k^2 e^2}[f(x^0) - f^*]. \tag{47}$$

E.3.2. CONVERGENCE OF L-SVRG (WITH RESTART)

In Qian et al. (2019) an analysis of L-SVRG (Algorithm 5) was provided for non-convex and smooth functions. In particular Lemma E.7 has been proved under the Expected residual Assumption E.6.

Assumption E.6. There is a constant $\rho_{nc} > 0$ such that

$$\mathbb{E}\left[\|\nabla f_i(x) - \nabla f_i(y) - [\nabla f(x) - \nabla f(y)]\|^2\right] \leq \rho_{nc}\|x - y\|^2 \tag{48}$$

Assumption E.6 is similar to the expected residual condition presented in the main paper. For the case of τ-minibatch sampling, in Qian et al. (2019) it was shown that parameter ρ_{nc} of the assumption can be upper bounded by $\rho_{nc} \leq \frac{n^2-\tau}{(n^2-1)\tau}\frac{1}{n}\sum L_i^2$ where L_i is the smoothness parameter of function f_i.

Under the Expected residual Assumption E.6 the following lemma was proven in Qian et al. (2019).

Lemma E.7 (Theorem 5.1 in Qian et al. (2019))**.** Let f be nonconvex and smooth function. Let Assumption E.6 be satisfied and let $p \in (0, 1]$. Consider the Lyapunov function $\Psi^k = f(x^k) + \alpha\|x^k - w^k\|^2$ where $\alpha = 3\gamma^2 L\rho_{nc}/p$. If stepsize γ satisfies:

$$\gamma \leq \min\left\{\frac{1}{4L}, \frac{p^{2/3}}{36^{1/3}(L\rho_{nc})^{1/3}}, \frac{\sqrt{p}}{\sqrt{6}\rho_{nc}}\right\} \tag{49}$$

then the update of L-SVRG (Algorithm 5) satisfies

$$\mathbb{E}_i[\Psi^{k+1}] \leq \Psi^k - \frac{\gamma}{4}\|\nabla f(x^k)\|^2.$$

Having the result of Lemma E.7 let us now present the main Theorem describing the convergence of L-SVRG with restart presented in Algorithm 6. Let us, run L-SVRG with step-size γ that satisfies (49) and select the output x^u of the method to be its Option II. That is x^u is chosen uniformly at random from $\{x^i\}_{i=0}^K$. In this case we name the method L-SVRG$_{II}(x^0 = w^0, K, \gamma, p \in (0, 1])$.

Theorem E.8 (Convergence of Algorithm 6)**.** Let f be $L-$smooth function that satisfies the PL condition (4) with parameter μ. Let Assumption E.6 be satisfied and let $p \in (0, 1]$. If stepsize γ satisfies:

$$\gamma \leq \min\left\{\frac{1}{4L}, \frac{p^{2/3}}{36^{1/3}(L\rho_{nc})^{1/3}}, \frac{\sqrt{p}}{\sqrt{6}\rho_{nc}}\right\}$$

and $K = \frac{4}{\mu \gamma}$ then the update of Algorithm 6 satisfies

$$\mathbb{E}[f(x^t) - f(x^*)] \leq \left(\frac{1}{2}\right)^t [f(x^0) - f(x^*)], \tag{50}$$

and

$$\mathbb{E}\|\nabla f(x^t)\|^2 \leq \left(\frac{1}{2}\right)^t \|\nabla f(x^0)\|^2. \tag{51}$$

Proof. Using Lemma E.7 we obtain:

$$\mathbb{E}_i[\Psi^{k+1}] \leq \Psi^k - \frac{\gamma}{4}\|\nabla f(x^k)\|^2.$$

By taking expectation again and by rearranging:

$$\mathbb{E}\|\nabla f(x^k)\|^2 \leq \frac{4}{\gamma}\left[\mathbb{E}\left[\Psi^k\right] - \mathbb{E}[\Psi^{k+1}]\right]$$

By letting x^u to be chosen uniformly at random from $\{x^i\}_{i=0}^K$ we obtain:

$$
\begin{aligned}
\mathbb{E}\|\nabla f(x^u)\|^2 &\leq \frac{1}{K}\sum_{i=0}^{K-1}\mathbb{E}\|\nabla f(x^i)\|^2 \\
&\leq \frac{1}{K}\frac{4}{\gamma}\sum_{i=0}^{K-1}\left[\mathbb{E}\left[\Psi^k\right] - \mathbb{E}[\Psi^k]\right] \\
&= \frac{1}{K}\frac{4}{\gamma}\left(\Psi^0 - \mathbb{E}[\Psi^K]\right) \\
&= \frac{1}{K}\frac{4}{\gamma}\left(f(x^0) - \mathbb{E}[f(x^k)] - \alpha\mathbb{E}[\|x^k - w^k\|^2]\right) \\
&\leq \frac{1}{K}\frac{4}{\gamma}\left[f(x^0) - f(x^*)\right] \tag{52}
\end{aligned}
$$

Convergence on function values.   The above derivation, (52), shows that the iterates of Algorithm 6 satisfy:

$$\mathbb{E}\|\nabla f(x^t)\|^2 \leq \frac{4}{\gamma K}\mathbb{E}\left[f(x^{t-1}) - f(x^*)\right]$$

Substitute the specified value of $K = \frac{4}{\gamma\mu}$ in the above inequality, we have

$$\mathbb{E}\|\nabla f(x^t)\|^2 \leq \mu\mathbb{E}\left[f(x^{t-1}) - f(x^*)\right]$$

and since the function satisfies the PL condition we have $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu\left[f(x) - f(x^*)\right]$ which means that:

$$2\mu\mathbb{E}\left[f(x^t) - f(x^*)\right] \leq \mathbb{E}\|\nabla f(x^t)\|^2 \leq \mu\mathbb{E}\left[f(x^{t-1}) - f(x^*)\right]$$

Thus,

$$\mathbb{E}\left[f(x^t) - f(x^*)\right] \leq \left(\frac{1}{2}\right)\mathbb{E}\left[f(x^{t-1}) - f(x^*)\right]$$

by unrolling the recurrence we obtain (50).

Convergence on norm of the gradient.   Similar to the previous case, using (52), the iterates of Algorithm 6 satisfy:

$$
\begin{aligned}
\mathbb{E}\|\nabla f(x^t)\|^2 &\leq \frac{4}{\gamma K}\left[f(x^{t-1}) - f(x^*)\right] \\
&\overset{(4)}{\leq} \frac{4}{\gamma K}\frac{1}{2\mu}\|\nabla f(x^{t-1})\|^2 \\
&= \frac{2}{\gamma\mu K}\|\nabla f(x^{t-1})\|^2 \tag{53}
\end{aligned}
$$

Using the specified value $K = \frac{4}{\gamma\mu}$ in the above inequality, we have:

$$\mathbb{E}\|\nabla f(x^t)\|^2 \leq \left(\frac{1}{2}\right)\|\nabla f(x^{t-1})\|^2 \tag{54}$$

and by unrolling the recurrence we obtain (51). □

F. Experimental Details

In the experimental section we compare several different algorithms, we provide a short explanation of the different algorithms here:

- **SHGD** with constant and decreasing step-size: This is the Alg. 1 proposed in the paper.

- **Biased SHGD**: This is a biased version of Alg. 1 that was proposed by Mescheder et al. (2017), where $\nabla\mathcal{H}_{i,j}(x) = \frac{1}{2}\nabla\langle\xi_i(x), \xi_j(x)\rangle$ is replaced by $\nabla\hat{\mathcal{H}}_{i,j}(x) = \frac{1}{2}\nabla\|\xi_i(x) + \xi_j(x)\|^2$, note that this a biased estimator of $\nabla\mathcal{H}(x)$.

- **L-SVRHG** with or without restart: This is the Alg.2 proposed in the paper, with Option II for the restart and Option I for the version without restart. Restart is not used unless specified.

- **CO**: This is the Consensus Optimization algorithm proposed in Mescheder et al. (2017). We provide more details in App. F.5.

- **SGDA**: This is the stochastic version of Simultaneous Gradient Descent/Ascent algorithm, which uses the following update $x^{k+1} = x^k - \eta_k\xi_i(x_k)$.

- **SVRE** with restart: This is the Alg. 3 described in Chavdarova et al. (2019).

In the following sections we provide the details for the different hyper-parameters used in our different experiments.

F.1. Bilinear Game

We first provide the details about the bilinear experiments presented in section 7.1:

$$\min_{x_1\in\mathbb{R}^d}\max_{x_2\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}x_1^\top\mathbf{A}_i x_2 + b_i^\top x_1 + c_i^\top x_2 \tag{55}$$

where:

$$n = d = 100$$
$$\mathbf{A}_i \in R^{d\times d}, \quad [\mathbf{A}_i]_{kl} = \begin{cases} 1 & \text{if } i = k = l \\ 0 & \text{otherwise} \end{cases}$$
$$b_i, c_i \in R^d, \quad [b_i]_k, [c_i]_k \sim \mathcal{N}(0, 1/d)$$

The hyper-parameters used for the different algorithms are described in Table 1:

Table 1. Hyper-parameters used for the different algorithms in the Bilinear Experiments (section 7.1).

ALGORITHMS	STEP-SIZE γ^k	PROBABILITY p	RESTART
SHGD WITH CONSTANT STEP-SIZE	0.5	N/A	N/A
SHGD WITH DECREASING STEP-SIZE	$\begin{cases} 0.5 & \text{FOR } k \leq 10,000 \\ \frac{2k+1}{(k+1)^2}\frac{1}{2500} & \text{FOR } k > 10,000. \end{cases}$	N/A	N/A
BIASED SHGD	0.5	N/A	N/A
L-SVRHG	10	$\frac{1}{n} = 0.01$	N/A
SVRE	0.3	$\frac{1}{n} = 0.01$	RESTART WITH PROBABILITY 0.1

The optimal constant step-size suggested by the theory for **SHGD** is $\gamma = \frac{1}{2\mathcal{L}}$. In this experiment we have that $\mathcal{L} = 1$, thus the optimal step-size is 0.5, this is also what we observed in practice. However we observe that while the theory recommends to decrease the step-size after $4\lceil \mathcal{K} \rceil = 40,000$ we observe in this experiment that it actually converges faster if we decrease the step-size a bit earlier after only $10,000$ iterations.

F.2. Sufficiently-Bilinear Games

In this section we provide more details about the sufficiently-bilinear experiments of section 7.2:

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

where:

$$n = d = 100 \quad \text{and} \quad \delta = 7$$

$$\mathbf{A}_i \in R^{d \times d}, \quad [\mathbf{A}_i]_{kl} = \begin{cases} 1 & \text{if } i = k = l \\ 0 & \text{otherwise} \end{cases}$$

$$b_i, c_i \in R^d, \quad [b_i]_k, [c_i]_k \sim \mathcal{N}(0, 1/d)$$

$$F(x) = \frac{1}{d} \sum_{k=1}^{d} f(x_k), a \quad f(x) = \begin{cases} -3\left(x + \frac{\pi}{2}\right) & \text{for } x \leq -\frac{\pi}{2} \\ -3\cos x & \text{for } -\frac{\pi}{2} < x \leq \frac{\pi}{2} \\ -\cos x + 2x - \pi & \text{for } x > \frac{\pi}{2} \end{cases} \tag{56}$$

Note that this game satisfies the sufficiently-bilinear condition as long as $\delta > 2L$, where L is the smoothness of $F(x)$, in our case $L = 3$. Thus we choose $\delta = 7$ in order for the sufficiently-bilinear condition to be satisfied.

The hyper-parameters used for the different algorithms are described in Table 2:

Table 2. Hyper-parameters used for the different algorithms in the sufficiently-bilinear experiments (section 7.2).

ALGORITHMS	STEP-SIZE γ^k	PROBABILITY p	RESTART
SHGD WITH CONSTANT STEP-SIZE	0.02	N/A	N/A
SHGD WITH DECREASING STEP-SIZE	$\begin{cases} 0.02 & \text{FOR} \quad k \leq 10,000 \\ \frac{2k+1}{(k+1)^2 \frac{1}{2500}} & \text{FOR} \quad k > 10,000. \end{cases}$	N/A	N/A
BIASED SHGD	0.01	N/A	N/A
L-SVRHG	0.1	$\frac{1}{n} = 0.01$	N/A
L-SVRHG WITH RESTART	0.1	$\frac{1}{n} = 0.01$	RESTART EVERY $1,000$ ITERATIONS
SVRE	0.05	0.1	RESTART WITH PROBABILITY 0.1

F.3. GANs

In this section we present the details for the GANs experiments. We first present the different problem we try to solve.

satGAN solve the following problem:

$$\min_{\mu, \sigma} \max_{\phi_0, \phi_1, \phi_2} \frac{1}{n} \sum_{i=1}^{n} \log(\text{sigmoid}(\phi_0 + \phi_1 y_i + \phi_2 y_i^2)) + \log(1 - \text{sigmoid}(\phi_0 + \phi_1(\mu + \sigma z_i) + \phi_2(\mu + \sigma z_i)^2))$$

nsGAN solve the following problem:

$$\max_{\phi_0,\phi_1,\phi_2} \frac{1}{n} \sum_{i=1}^{n} \log(\text{sigmoid}(\phi_0 + \phi_1 y_i + \phi_2 y_i^2)) + \log(1 - \text{sigmoid}(\phi_0 + \phi_1(\mu + \sigma z_i) + \phi_2(\mu + \sigma z_i)^2))$$

$$\max_{\mu,\sigma} \frac{1}{n} \sum_{i=1}^{n} \log(\text{sigmoid}(\phi_0 + \phi_1(\mu + \sigma z_i) + \phi_2(\mu + \sigma z_i)^2)) + \log(1 - \text{sigmoid}(\phi_0 + \phi_1 y_i + \phi_2 y_i^2))$$

WGAN solve the following problem:

$$\min_{\mu,\sigma} \max_{\phi_1,\phi_2} \frac{1}{n} \sum_{i=1}^{n} (\phi_1 y_i + \phi_2 y_i^2) - (\phi_1(\mu + \sigma z_i) + \phi_2(\mu + \sigma z_i)^2)$$

All Discriminator and Generator parameters are initialized randomly with $U(-1, 1)$ prior. The data is set as $y_i \sim N(0, 1)$, $z_i \sim N(0, 1)$. We run all experiments 10 times (with seed 1, 2, ..., 10).

The hyper-parameters used for the different algorithms are described in Table 3:

Table 3. Hyper-parameters used for the different algorithms in the GAN Experiments (section 7.3).

ALGORITHMS	STEP-SIZE γ^k	PROBABILITY p	SAMPLE SIZE	MINI-BATCH SIZE
CO	.02	N/A	10K	100
SGDA	.02	N/A	10K	100
SHGD	.02	N/A	10K	100
L-SVRHG	.02	$\frac{1}{n} = 0.01$	10K	100

F.4. Implementation details for L-SVRHG

L-SVRHG requires the computation of the gradient of the full Hamiltonian with probability p. As a reminder the Hamiltonian can be written as the sum of n^2 terms (see eq 11), a naive implementations would thus requires n^2 operation to compute each of the $\nabla\mathcal{H}_{i,j}(x)$ to get the full gradient. However a more efficient alternative is to notice that the Hamiltonian can be written as $\mathcal{H}(x) = \frac{1}{2}\|\xi(x)\|^2$, by first computing $\xi(x)$ and then using back-propagation to compute the gradient of $\mathcal{H}(x)$, we can reduce the cost of computing the full gradient to $2n$ instead of n^2.

F.5. Details for Consensus Optimization (CO)

Consensus optimization can be formulated as solving the following problem using SGDA:

$$\min_{x_1 \in \mathbb{R}^{d_1}} \max_{x_2 \in \mathbb{R}^{d_2}} g(x_1, x_2) + \lambda\mathcal{H}((x1, x2)) \tag{57}$$

Using SGDA to solve this problem is equivalent to do the following update:

$$x^{k+1} = x^k - \eta_k(\xi(x_k) + \lambda\nabla\mathcal{H}_{i,j}(x^k)) \tag{58}$$

As per Mescheder et al. (2017), we used $\lambda = 10$ in all the experiments and a biased estimator of the Hamiltonian $\nabla\hat{\mathcal{H}}_{i,j}(x) = \frac{1}{2}\nabla\|\xi_i(x) + \xi_j(x)\|^2$. Note that we also tried to use the unbiased estimator proposed in section 5.1 but found no significant difference in our results, and thus only included the results for the original algorithm proposed by Mescheder et al. (2017) that uses the biased estimator.

F.6. Cost per iteration

In the experimental section, we compare the different algorithms as a function of the number of gradient computations. In Table 4, we give the number of gradient computations per iteration for all the different algorithms compared in the paper.

We also give a brief explanation on the cost per iteration of each methods:

- **SHGD**: We can write $\mathcal{H}_{i,j}(x) = \frac{1}{2} < \xi_i(x), \xi_j(x) >$, thus at every iteration we need to compute two gradients $\xi_i(x)$ and $\xi_j(x)$, which leads to a cost of 2 per iteration.

- **Biased SHGD**: The biased estimate is based on $\mathcal{H}_i(x) = \frac{1}{2}\|\xi_i(x) + \xi_j(x)\|^2$, which requires the computation of two gradients $\xi_i(x)$ and thus also has a cost of 2 per iteration.

- **L-SVRHG**: At every iteration we need to compute two Hamiltonian updates which cost 2 each, and with probability p we need to compute the full Hamiltonian which cost $2n$ (see App. F.4), which leads to a cost of $4 + p \cdot 2n$

- **SVRE**: At each iteration **SVRE** need to do an extrapolation step and an update step, both the extrapolation step and the update step requires to evaluate two gradients, and with probability p we need to compute the full gradient which cost n, which leads to a total cost of $4 + p \cdot n$.

Table 4. Number of the gradient computations per iteration for the different algorithms compared in the paper.

ALGORITHM	COST PER ITERATION
SGDA	1
SHGD	2
BIASED SHGD	2
L-SVRHG	$4 + p \cdot 2n$
SVRE	$4 + p \cdot n$

G. Additional Experiments

In this section we provide additional experiments that we couldn't include in the main paper. Those experiments provide further observations on the behavior of our proposed methods in different settings.

G.1. Bilinear and Sufficiently-Bilinear Games

G.1.1. SYMMETRIC POSITIVE DEFINITE MATRIX

In all the experiments presented in the paper the matrix $\mathbf{A}_i$ have a particular structure, they are very sparse, and the matrix $\mathbf{A} = 1/n \sum_{i=1}^{n} \mathbf{A}_i$ is the identity. We thus propose here to compare the methods on the bilinear game (9) and the sufficiently-bilinear game from Section 7.2 but with different matrices $\mathbf{A}_i$. We choose $\mathbf{A}_i$ to be random symmetric positive definite matrices. For the sufficiently bilinear experiments we choose δ, such that the sufficiently-bilinear condition is satisfied. We show the results in Fig. 3, we observe results very similar to the results observed in section 7.1 and section 7.2, the experiments again shows that our proposed methods follow closely the theory and that **L-SVRHG** is the fastest method to converge.
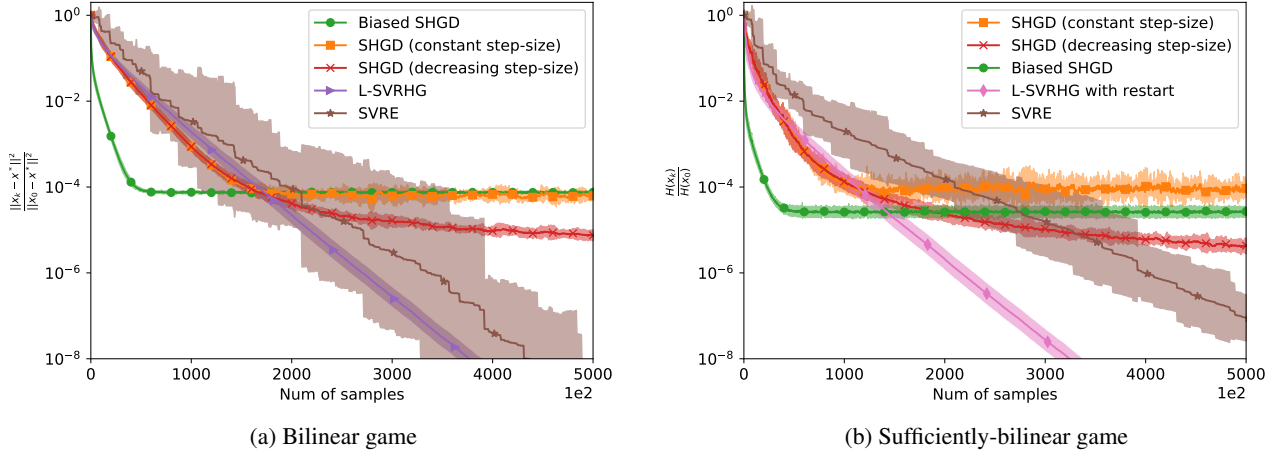
(a) Bilinear game
(b) Sufficiently-bilinear game

Figure 3. Results for the bilinear game and sufficiently-bilinear game with symmetric positive definite matrices. We observe results very similar to the results observed in section 7.1 and section 7.2, the experiments again shows that our proposed methods follow closely the theory and that **L-SVRHG** is the fastest method to converge.

G.1.2. INTERPOLATED GAMES

In this section we present a particular class of games that we call interpolated games.

Definition G.1 (Interpolated Games). If a game is such that at the equilibrium x^*, we have $\forall i \; \xi_i(x^*) = 0$, then we say that the game satisfies the interpolation condition.

If a game satisfies the interpolation condition, then **SHGD** with constant step-size converges linearly to the solution.

In the bilinear game (9) and the sufficiently-bilinear game from Section 7.2, if we choose to set $\forall i \; b_i = c_i = 0$, then both problems satisfies the interpolation condition. We provide additional experiments in this particular setting where we compare **SHGD** with constant step-size, **Biased SHGD**, and **L-SVRHG**. We show the results in Fig. 4. We observe that all methods converge linearly to the solution, surprisingly in this setting **Biased SHGD** converges much faster than all other methods.

We argue that this is due to the fact that **Biased SHGD** is optimizing an upper-bound on the Hamiltonian. Indeed we can show using Jensen's inequality, that:

$$\mathcal{H}(x) = \frac{1}{2}\|\xi(x)\|^2 = \frac{1}{2}\|\frac{1}{n}\sum_{i=1}^n \xi_i(x)\|^2 \overset{Jensen}{\leq} \frac{1}{2n}\sum_{i=1}^n \|\xi_i(x)\|^2 = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\|\xi_i(x)\|^2 = \frac{1}{n}\sum_{i=1}^n \mathcal{H}_i(x) \quad (59)$$

If the interpolation condition is satisfied, then we have that at the optimum x^*, the inequality becomes an equality:

$$\mathcal{H}(x^*) = \frac{1}{n}\sum_{i=1}^n \mathcal{H}_i(x^*) = 0 \quad (60)$$

Thus in this particular setting **Biased SHGD** also converges to the solution. Furthermore we can notice that because the $\mathbf{A}_i$ are very sparse, $\forall i \neq j \; \nabla\mathcal{H}_{i,j}(x) = 0$. Thus most of the time **SHGD** will not update the current iterate, which is not the case of **Biased SHGD** which only considers the $\nabla\mathcal{H}_{i,i}(x) = 0$ to do its update and thus always has signal. The convergence of **SHGD** could thus be improved by using non-uniform sampling. We leave this for future work.
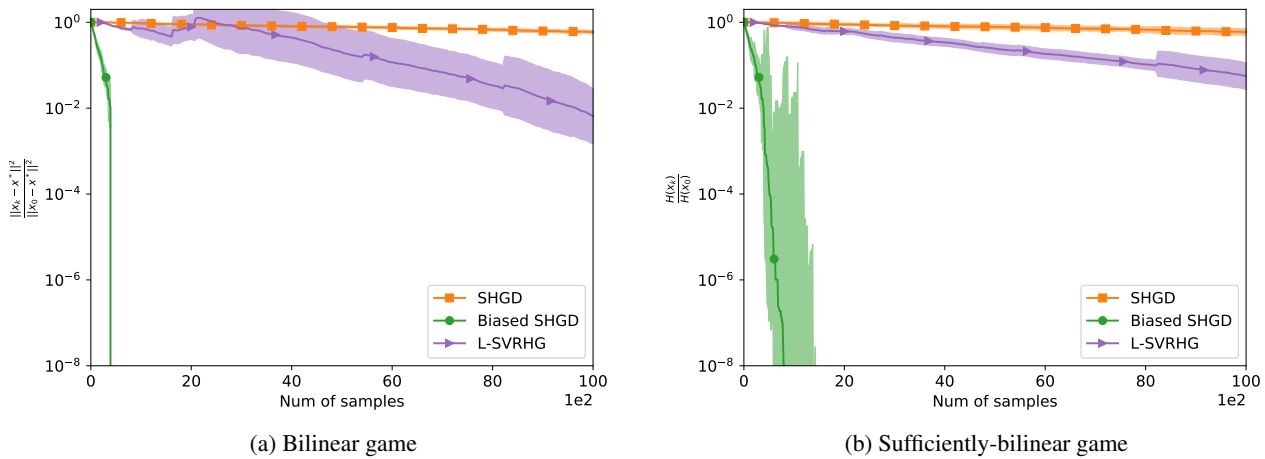
(a) Bilinear game

(b) Sufficiently-bilinear game

Figure 4. Results for the bilinear game and sufficiently-bilinear game when $\forall i\ b_i = c_i = 0$. We observe that all the methods converge linearly in this setting. Surprisingly in this setting **Biased SHGD** is the fastest method to converge. We give a brief informal explanation on why this is the case above.

G.2. GANs

We present the missing experiments for satGAN (with batch size 100) in Figure 5 As can be observed, results for nsGAN are very similar to results for satGAN (see Figure 2).
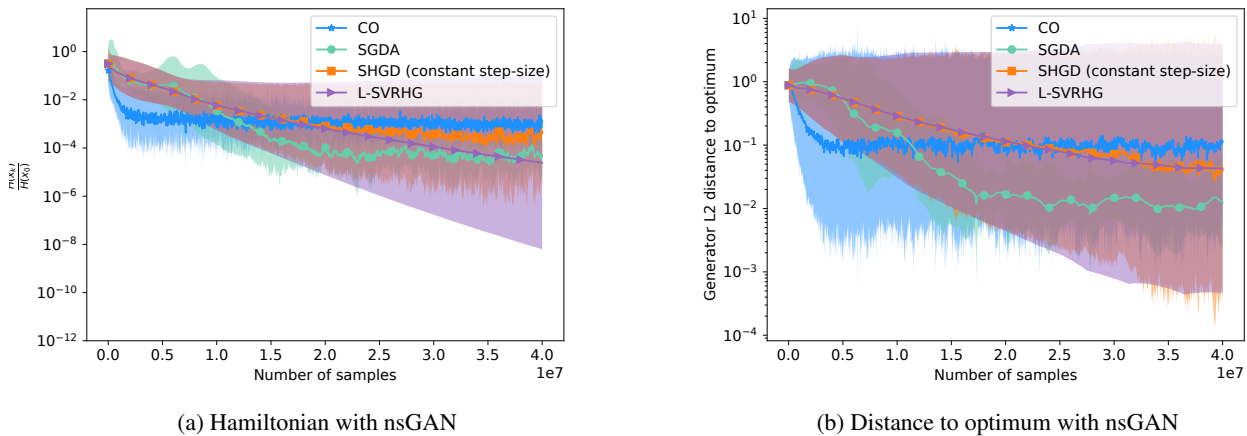

(a) Hamiltonian with nsGAN

(b) Distance to optimum with nsGAN

Figure 5. nsGAN with batch size of 100