A Stochastic Extension of Hamiltonian Descent Methods

Jaivardhan Kapoor
Roll Number - 150300
Dept of EE, IIT Kanpur
jkapoor@iitk.ac.in

Harshvardhan
Roll Number - 150283
Dept of EE, IIT Kanpur
harshv@iitk.ac.in

***Abstract*—We analyse the recently released work in [1], which employs Hamiltonian dynamics to allow faster convergence rate on a wider class of objectices. The proposed framework allows for linear rates of convergence on certain classes of non-strongly convex functions and generalizes the momentum method to non-classical kinetic energies. We propose a stochastic variant of one such method, and sketch its convergence proof. We also implement the deterministic methods as well as the stochastic counterpart and compare various facets of the methods with baselines such as Gradient Descent and Momentum.**

I. Introduction

In recent years, there has been a growing interest in the study of continuous dynamics to motivate faster and more robust accelerated methods in various sub-fields of optimization and inference [2] [1] [3]. Continuous dynamics allow us to generalize various classes of optimization and simulation algorithms by converting the discretization to a continuous version. The methods then allow for analysis in this continuous domain.

The recently released work in [1], employs this methodology to analyse classical accelerated Gradient methods that use first-order gradient information. This results in a generalization of Polyak's Heavy Ball Method. The new suite of algorithms have fast convergence on a wider range of convex objectives than preciousl possible, hile allowing for many dgrees of freedom in their tailoring. We analyse their framework, and also propose our stochastic extension to their algorithm. We then simulate their algorithms on functions showing pathological behaviour, and report our results and interpretations.

A. Hamiltonians

In mechanics, a Hamiltonian system is used to model the continuous time dynamics of a system acted upon by a field ∇f. The total energy of the system is given by –

$$\mathcal{H}\left(\mathbf{x},\mathbf{p}\right) = k(\mathbf{p}) + f(\mathbf{x}) - f(\mathbf{x}^{*}) \tag{1}$$

Here, $\mathbf{x}, \mathbf{p}, k, f$ and $\mathbf{x}^{*}$ represent the position,momentum,kinetic energy, potential energy and potential energy minima position respectively. The update equations for a Hamiltonian ensure that the energy is always constant. Hamiltonians have existed in mechanics for decades but are now being incorporated into machine learning algorithms with appropriate discretizations with the most popular application being Hamiltonian MCMC [4].

For this paper, a special variant, namely, conformal Hamiltonian is used to model the system. The continuous time update equations for such a system are –

$$\begin{aligned} \mathbf{x}_t^{'} &= \mathbf{p}_t \\ \mathbf{p}_t^{'} &= -\nabla f\left(\mathbf{x}_t\right) - \gamma \mathbf{p}_t \end{aligned} \tag{2}$$

This differs from the standard Hamiltonian in only the $(-\gamma \mathbf{p}_t)$ term. This acts like a frictional force which dissipates energy from the Hamiltonian.

$$\begin{aligned} \mathcal{H}_t^{'} &= \left\langle \nabla k(\mathbf{p}_t), \mathbf{p}_t^{'} \right\rangle + \left\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t^{'} \right\rangle \\ &= -\gamma \left\langle \nabla k(\mathbf{p}_t), \mathbf{p}_t \right\rangle \leq -\gamma k(p) \leq 0 \end{aligned} \tag{3}$$

Note that this convergence requires $\gamma > 0$,k to be convex and have minima as $k(0) = 0$. Also, this formulation bears resemblance to Polyak's heavy ball method [5] whose discretization gives us the momentum schemes of gradient descent. The conformal Hamiltonian continuously loses energy and converges to the lowest energy state $(\mathbf{x}^{*}, 0)$ under certain assumptions. For specific design choices of k and f, this convergence has an exponential rate in continuous time. Note that any other dissipation field $D(p_t)$ satisfying $\langle \nabla k(\mathbf{p}_t), D(\mathbf{p}_t) \rangle \leq 0$ will also lead to a descending Hamiltonian but linear convergence rates and convergence to minima has not been investigated.

B. Defining $\mathcal{H}$

Before we state the lemma for the continuous time convergence rate, we will investigate the formulations of k and f for our case. Allowing f to be the convex objective function is an obvious choice. The standard definition for k has been $\frac{\langle \mathbf{p}, \mathbf{p} \rangle}{2m}$, where m is the mass. For our case, the design of k turns out to be–

$$k(\mathbf{p}) = \frac{1}{2}\left(f_c^*\left(\mathbf{p}\right) + f_c^*\left(\mathbf{p}\right)\right) \tag{4}$$

$$\text{where } f_c(\mathbf{x}) = f(\mathbf{x} + \mathbf{x}^*) - f(\mathbf{x}^*) \tag{5}$$

Here f_c^* denotes the conjugate function of f_c. This design choice can nice properties which allow us to guarantee linear convergence in continuous and discrete time. For a convex f, k is convex and $k(0) = 0$. An interesting point to note is that considering the unconstrained optimization of f as the primal problem, one form of the dual turns out to be $\frac{1}{2}\left(f^*(\mathbf{p}) + f^*(-\mathbf{p})\right)$ and the duality gap turns out to be exactly equal to the Hamiltonian total energy. Thus, in some way minimizing the Hamiltonian leads to reducing the duality gap. This property can be of paramount importance when we try to extend this to constrained settings.

We now present the convergence rate for the continuous time Hamiltonian system with our given definition of k. Note that the assumptions to guarantee convergence are stated in the next section where the discrete time algorithms have been described.

Theorem I.1. *Given $f, k, \gamma, \alpha, C_{\alpha,\gamma}$ satisfying assumptions A. Let $(\mathbf{x}_t, \mathbf{p}_t)$ be a solution to 2 with initial states $(\mathbf{x}_0, \mathbf{p}_0) = (\mathbf{x}, 0)$. Let $\alpha^* = \alpha\left(3\mathcal{H}_0\right)$, $\lambda = \frac{(1-\gamma)C_{\alpha,\gamma}}{4}$ and $\mathcal{W} : [0, \infty) \to [0, \infty)$ be the solution of*

$$\mathcal{W}_t^{'} = -\lambda \cdot \alpha\left(2\mathcal{W}_t\right)\mathcal{W}_t,$$

with $\mathcal{W}_0 := \mathcal{H}_0 = f(\mathbf{x}_0) - f(\mathbf{x}^*)$. Then, for every $t \in [0, \infty)$, we have

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq & 2\exp\left(-\lambda \int_0^t \alpha\left(2\mathcal{W}_t\right)\right) \\ \leq & 2\mathcal{H}_0 \exp\left(-\lambda\alpha^* t\right) \end{aligned} \tag{6}$$

We do not provide a proof for this theorem as our main focus is on convergence of the discretized algorithms. The proof proceeds by defining the Lyapunov function as $\mathcal{V}_t\left(\mathbf{x}, \mathbf{p}\right) = \mathcal{H}\left(\mathbf{x}, \mathbf{p}\right) + \beta\left\langle \mathbf{x} - \mathbf{x}^*, \mathbf{p}\right\rangle$ and then establishing $\mathcal{V}_t^{'} \leq -\mu\mathcal{V}_t$ for some constants μ, β. One of the initial lemmas used for the proof which relates $\mathcal{V}$ and $\mathcal{H}$ will be described later as it is used in the proofs of the discretized algorithms.

II. Methods

This section describes the stochastic and non-stochastic discretized versions of this continuous time Hamitonian, its convergence rates, their proofs and the assumption. The stochastic version is our novel contribution and involves very little modification in the non-stochastic version. We cover both these cases side by side.

A. Algorithms

The original paper proposes 3 discretization schemes – 1 implicit and 2 explicit. We cover the 2 explicit schemes as their stochastic analogues were easier to prove.

Algorithm 1 First explicit Method

```
Require: $f, k, \epsilon, \gamma, \mathbf{x}_0, \mathbf{p}_0, \delta = (1 + \gamma\epsilon)^{-1}$
    $\mathbf{p}_{t+1} = \delta\mathbf{p}_t - \epsilon\nabla f(\mathbf{x}_t)$
    $\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon\nabla k(\mathbf{p}_{t+1})$
```

Algorithm 2 Stochastic First explicit Method

```
Require: $f, k, \epsilon, \gamma, \mathbf{x}_0, \mathbf{p}_0, \delta = (1 + \gamma\epsilon)^{-1}$
    Sample random variables $a, b$
    $\mathbf{p}_{t+1} = \delta\mathbf{p}_t - \epsilon\nabla f_a(\mathbf{x}_t)$
    $\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon\nabla k_b(\mathbf{p}_{t+1})$
```

Algorithm 3 Second explicit Method

```
Require: $f, k, \epsilon, \gamma, \mathbf{x}_0, \mathbf{p}_0$
    $\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon\nabla k(\mathbf{p}_t)$
    $\mathbf{p}_{t+1} = (1 - \epsilon\gamma)\,\mathbf{p}_t - \epsilon\nabla f(\mathbf{x}_{t+1})$
```

Algorithm 4 Stochastic Second explicit Method

```
Require: $f, k, \epsilon, \gamma, \mathbf{x}_0, \mathbf{p}_0$
    Sample random variables $a, b$
    $\mathbf{x}_{t+1} = \mathbf{x}_t + \epsilon\nabla k_b(\mathbf{p}_t)$
    $\mathbf{p}_{t+1} = (1 - \epsilon\gamma)\,\mathbf{p}_t - \epsilon\nabla f_a(\mathbf{x}_{t+1})$
```

The two explicit methods are discretize the system 2 on points $(\mathbf{x}_t, \mathbf{p}_{t+1})$ and $(\mathbf{x}_{t+1}, \mathbf{p}_t)$ respectively. We will see later that this small change in discretization allows us to expand the linear rate of convergence to even non-smooth and non-strongly convex functions. The stochastic versions of these algorithms assume that you have a random realization of the actual gradient.

B. Assumptions

Here $\mathcal{S}$: denotes the stochastic version of the same constraint.

Assumptions A –

1) $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable and convex with unique minimum at $\mathbf{x}^*$. $\mathcal{S}$: This property holds for each f_a and $\mathbb{E}[f_a] = f$ has minima at $\mathbf{x}^*$.
2) $k : \mathbb{R}^d \rightarrow \mathbb{R}$ differentiable and strictly convex with minima $k(0) = 0.\mathcal{S}$: Assumption holds for each k_b and $\mathbb{E}[k_b] = k$ has minima $k(0) = 0$.
3) $\gamma \in (0, 1)$
4) There exists some differentiable non-increasing convex function $\alpha : [0, \infty) \rightarrow (0, 1]$ and constant $C_{\alpha,\gamma} \in (0, \gamma]$ such that for every $\mathbf{p} \in \mathbb{R}^d$

$$k(\mathbf{p}) \geq \alpha\left(k(\mathbf{p})\right) \max\left(f_c^*(\mathbf{p}), f_c^*(-\mathbf{p})\right) \tag{7}$$

and that for every $y \in [0, \infty)$

$$-C_{\alpha,\gamma}\alpha^{'}(y)y < \alpha(y) \tag{8}$$

If $k(\mathbf{p}) \geq \alpha^* \max\left(f_c^*(\mathbf{p}), f_c^*(-\mathbf{p})\right)$, for a constant $\alpha^* \in (0, 1]$, then $\alpha(y) = \alpha^*$ is a valid choice. $\mathcal{S}$: For the stochastic version, we use the choice of a constant α as it made subsequent results easier to prove. The exact condition required is

$$\mathbb{E}\left[k_a(\mathbf{p})\right] \geq \alpha^* \max\left(\mathbb{E}\left[f_{c,a}^*(\mathbf{p})\right], \mathbb{E}\left[f_{c,a}^*(-\mathbf{p})\right]\right) \tag{9}$$

5) Additionally, for the stochastic case, $\mathbb{E}[f_a] = f, \mathbb{E}[\nabla f_a] = \nabla f, \mathbb{E}[k_b] = k, \mathbb{E}[\nabla k_b] = k$

Assumption B – There exists $C_{f,k} \in (0, \infty)$ such that $\forall \mathbf{x}, \mathbf{p} \in \mathbb{R}^d$,

$$|\langle \nabla f(\mathbf{x}), \nabla k(\mathbf{p}) \rangle| \leq C_{f,k}\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right) \tag{10}$$

$\mathcal{S}$: The same condition holds but for functions f_a, k_b on expectation,i.e.,

$$\mathbb{E}\left[|\langle \nabla f_a(\mathbf{x}), \nabla k_b(\mathbf{p}) \rangle|\right] \leq C_{f,k}\mathbb{E}\left[\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right)\right] \tag{11}$$

Assumptions C –

1) There exists $C_k \in (0, \infty)$ such that for every $\mathbf{p} \in \mathbb{R}^d$,

$$\langle \nabla k(\mathbf{p}), \mathbf{p} \rangle \leq C_k k(\mathbf{p}) \tag{12}$$

$\mathcal{S}$:

$$\mathbb{E}\left[\langle \nabla k_b(\mathbf{p}), \mathbf{p} \rangle\right] \leq C_k \mathbb{E}\left[k_b(\mathbf{p})\right] \tag{13}$$

2) f is twice continuously differentiable for every $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{x}^*\}$ $\mathcal{S}$: Each f_a is twice continuously differentiable as well.
3) There exists $D_{f,k} \in (0, \infty)$ such that for every $\mathbf{p} \in \mathbb{R}^d, \mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{x}^*\}$,

$$\begin{aligned}&\left\langle \nabla k(\mathbf{p}), \nabla^2 f(\mathbf{x}) \nabla k(\mathbf{p}) \right\rangle \\ &\leq D_{f,k}\alpha(3\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right))\,\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right)\end{aligned} \tag{14}$$

$\mathcal{S}$:

$$\begin{aligned}&\mathbb{E}\left[\left\langle \nabla k(\mathbf{p}), \nabla^2 f(\mathbf{x}) \nabla k(\mathbf{p}) \right\rangle\right] \\ &\leq D_{f,k}\mathbb{E}\left[\alpha(3\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right))\,\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right)\right]\end{aligned} \tag{15}$$

Assumptions D –

1) k is twice continuously differentiable for every $\mathbf{p} \in \mathbb{R}^d \setminus \{0\}$
2) Exactly similar to Assumption 13
3) There exists $D_k \in (0, \infty)$ such that for every $\mathbf{p} \in \mathbb{R}^d \setminus \{0\}$,

$$\left\langle \mathbf{p}, \nabla^2 k(\mathbf{p}) \right\rangle \leq D_k k(\mathbf{p}) \tag{16}$$

4) There exists $E_k, F_k \in (0, \infty)$ such that for every $\mathbf{p}, \mathbf{q} \in \mathbb{R}^d$,

$$k(\mathbf{p}) - k(\mathbf{q}) \leq E_k k(\mathbf{p}) + F_k \langle \nabla k(\mathbf{p}) - \nabla k(\mathbf{q}), \mathbf{p} - \mathbf{q} \rangle \tag{17}$$

5) There exists $D_{f,k} \in (0, \infty)$ such that for every $\mathbf{x} \in \mathbb{R}^d, \mathbf{p} \in \mathbb{R}^d \setminus \{0\}$,

$$\begin{aligned}&\left\langle \nabla f(\mathbf{x}), \nabla^2 k(\mathbf{p}) \nabla f(\mathbf{x}) \right\rangle \\ &\leq D_{f,k}\alpha(3\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right))\,\mathcal{H}\left(\mathbf{x}, \mathbf{p}\right)\end{aligned} \tag{18}$$

For convergence of the continuous time formulation Assumption A needs to be satisfied, for implicit method A and B, for first explicit method A, B and C and for second explicit method A, B, D. We will prove convergence for first explicit method and its stochastic variant only so the stochastic versions for Assumption D have been omitted. The assumptions for first explicit method can be easily satisfied for a quadratic strongly smooth function and its quadratic version of kinetic energy defined in previous section. The roles of f and k are interchanged in the two explicit methods. The first explicit method requires PSD hessian of f while the second necessitates the same for k. Since, the objective function f is more or less fixed but the kinetic energy is somewhat under our control, we are able to achieve linear convergence through the second explicit method for f having unbounded or zero eigenvalues in its Hessian. These assumptions appear to be too restrictive, but the authors have demonstrated that when f and k are suitably chosen power functions according to the definitions in previous section all these assumptions are satisfied. More specifically, power functions of the form $f(x) = \frac{|x|^d}{d}, k(p) = \frac{|p|^c}{c}$ satisfy these assumptions when $\frac{1}{c} + \frac{1}{d} = 1$. In the experiments section, we choose functions of this form which satisfy all our assumptions.

C. Convergence Analysis

We provide convergence analysis for first explicit method and its stochastic variant. The proof for the second explicit method is on similar lines interchanging the roles of f and k with a few additional steps. The stochastic version for the first algorithm is also proved simultaneously.

The proof strategy will involve the following steps –

1) First, we find the relation between $\mathcal{H}$ and $\mathcal{V}$. This is Lemma 2.3 from [1]

Theorem II.1. *Let* $\mathbf{x} \in \mathbb{R}^d, f_a : \mathbb{R}^d \to \mathbb{R}$ *convex with unique minima at* $\mathbf{x}^*$*,* $k_b : \mathbb{R}^d \to \mathbb{R}$ *strictly convex with minima* $k(0) = 0$*,* $\alpha \in (0, 1]$ *and* $\beta \in (0, \alpha]$*. If* $\mathbf{p} \in \mathbb{R}^d$ *is such that* $\mathbb{E}\left[k_b(\mathbf{p})\right] \geq \alpha \mathbb{E}\left[f^*_{c,b}(-\mathbf{p})\right]$*, then*

$$\begin{aligned}\mathbb{E}\left[\langle \mathbf{x} - \mathbf{x}^*, p\rangle\right] \geq & -\mathbb{E}\left[\left(\frac{k_b(\mathbf{p})}{\alpha} + f_a(\mathbf{x}) - f_a(\mathbf{x}^*)\right)\right] \\ & \geq -\frac{\mathbb{E}\left[\mathcal{H}(\mathbf{x}, \mathbf{p})\right]}{\alpha}\end{aligned} \tag{19}$$

$$\frac{\alpha - \beta}{\alpha}\mathbb{E}\left[\mathcal{H}(\mathbf{x}, \mathbf{p})\right] \leq \mathbb{E}\left[\mathcal{V}(\mathbf{x}, \mathbf{p})\right] \tag{20}$$

If $\mathbf{p} \in \mathbb{R}^d$ *is such that* $\mathbb{E}\left[k_b(\mathbf{p})\right] \geq \alpha \mathbb{E}\left[f^*_{c,b}(+\mathbf{p})\right]$*, then*

$$\begin{aligned}\mathbb{E}\left[\langle \mathbf{x} - \mathbf{x}^*, p\rangle\right] \leq & \mathbb{E}\left[\left(\frac{k_b(\mathbf{p})}{\alpha} + f_a(\mathbf{x}) - f_a(\mathbf{x}^*)\right)\right] \\ & \leq \frac{\mathbb{E}\left[\mathcal{H}(\mathbf{x}, \mathbf{p})\right]}{\alpha}\end{aligned} \tag{21}$$

$$\frac{\alpha + \beta}{\alpha}\mathbb{E}\left[\mathcal{H}(\mathbf{x}, \mathbf{p})\right] \geq \mathbb{E}\left[\mathcal{V}(\mathbf{x}, \mathbf{p})\right] \tag{22}$$

The role of Assumptions A is in proving this inequality. This is further used to port results from $\mathcal{H}$ to $\mathcal{V}$ and vice-versa. The non-stochastic variants don't have expectations keeping the rest of the result unchanged. The proof can be obtained by expanding $\frac{\mathcal{H}}{\alpha}$.

2) We now try to establish an equation of the form $\mathcal{V}_t^{'} \leq -\lambda \mathcal{V}_t$. In discrete terms, such an equation would look like

$$\mathbb{E}\left[\mathcal{V}_{t+1} - \mathcal{V}_t\right] \leq -\epsilon\beta\left[1 - \gamma\epsilon C_2\right]\mathbb{E}\left[\mathcal{V}_{t+1}\right] \tag{23}$$

Note that this equation can be found in Lemma C.2 in [1]. The terms inside the square brackets are positive for some constant C_2. Such an equation would admit a linear convergence rate. Also, for the stochastic case, a constant α admits a constant $\beta = \frac{C_{\alpha,\gamma}}{2}\alpha(2\mathcal{V}_i)$ and admits a slightly easier proof. However, expanding the LHS of above equation, we obtain –

$$\begin{aligned}\mathbb{E}\left[\mathcal{V}_{t+1} - \mathcal{V}_t\right] = & \mathbb{E}\left[\mathcal{H}_{t+1} - \mathcal{H}_t\right] \\ & + \beta\mathbb{E}\left[\langle \mathbf{x}_{t+1} - \mathbf{x}^*, \mathbf{p}_{t+1}\rangle\right] \\ & - \beta\mathbb{E}\left[\langle \mathbf{x}_t - \mathbf{x}^*, \mathbf{p}_t\rangle\right]\end{aligned} \tag{24}$$

The next two steps will involve bounding these two difference terms individually in terms of a constant multiple of $\mathbb{E}\left[\mathcal{H}\right]$.

3) Bounding $\mathbb{E}\left[\mathcal{H}_{t+1} - \mathcal{H}_t\right]$. Using convexity and assumption B, this can be simplified to –

$$\begin{aligned}& \mathbb{E}\left[\mathcal{H}_{t+1} - \mathcal{H}_t\right] \\ & \leq -\gamma\epsilon\mathbb{E}\left[\langle \nabla k_b(\mathbf{p}_{t+1}), \mathbf{p}_{t+1}\rangle\right] \\ & + \mathbb{E}\left[\langle \nabla f_a(\mathbf{x}_{t+1}) - \nabla f_a(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle\right]\end{aligned} \tag{25}$$

This part can be found in Proposition 3.3 in [1]. The second term on RHS is bounded separately.

4) Bound on $\mathbb{E}\left[\langle \nabla f_a(\mathbf{x}_{t+1}) - \nabla f_a(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle\right]$ In our case, the bound on this term comes out to be –

$$\begin{aligned}& \mathbb{E}\left[\langle \nabla f_a(\mathbf{x}_{t+1}) - \nabla f_a(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle\right] \\ & \leq 3\epsilon^2 D_{f,k}\alpha\mathbb{E}\left[\mathcal{H}_{t+1}\right]\end{aligned} \tag{26}$$

The non-stochastic version has the same structure, but the non-constant α complicates one of the constant terms of the RHS. The proof of this term is where we actually switch from continuous to discrete settings. We define $\mathbf{x}^{(i)}_{t+1} := \mathbf{x}_t - i\epsilon\nabla k_b\left(\mathbf{p}_{t+1}\right)$. Here i is a continuous variable. We similarly define $\mathcal{H}^i_{t+1} := \mathcal{H}\left(\mathbf{x}^i_{t+1}, \mathbf{p}_{t+1}\right)$. We bound the LHS by integrating i from 0 to 1. We interchange expectations and integrals to obtain our stochastic form. The bound we obtain is in terms of $\mathcal{H}^i_{t+1}$. Then, we bound $\mathcal{H}^i_{t+1}$ in terms of $\mathcal{H}_{t+1}$. This is where we use Assumption B and Assumption A.

5) Bound on $\mathbb{E}\left[\langle \mathbf{x}_{t+1} - \mathbf{x}^*, \mathbf{p}_{t+1}\rangle - \langle \mathbf{x}_t - \mathbf{x}^*, \mathbf{p}_t\rangle\right]$. Using convexity and Assumption B, we first obtain the following result-

$$\mathbb{E}\left[f_a(\mathbf{x}_t)\right] \leq -f(\mathbf{x}_t) + \epsilon C_{f,k}\mathcal{H}_{t+1} \tag{27}$$

This result, along with the convexity of k_b, f_a and expanding the update steps, we get the following bound–

$$\begin{aligned}& \mathbb{E}\left[\langle \mathbf{x}_{t+1} - \mathbf{x}^*, \mathbf{p}_{t+1}\rangle - \langle \mathbf{x}_t - \mathbf{x}^*, \mathbf{p}_t\rangle\right] \\ & \leq \left(\gamma + \gamma\epsilon^2\right)\mathbb{E}\left[\langle \nabla k_b(\mathbf{p}_{t+1}), \mathbf{p}_{t+1}\rangle\right] + \epsilon C_{f,k}\mathcal{V}_{t+1} \\ & - \epsilon\mathbb{E}\left[f_a(\mathbf{x}_{t+1}) - f_a(\mathbf{x}^*)\right] - \gamma\epsilon\mathbb{E}\left[\langle \mathbf{x}_{t+1}, \mathbf{p}_{t+1}\rangle\right]\end{aligned} \tag{28}$$

6) The remaining terms are components of $\mathcal{V}$ but they have different coefficients. Thus, we set bounds on

the constants $\alpha, \beta,$, so that it can be upper bounded in terms of $\mathcal{V}$.

The organization of the original paper and the sequence in which results are presented are different from what we have done, but, we believe, this gives a better understanding.

1) Remarks on Stochastic variant: We were able to apply stochasticity to the analysis of this algorithm due to the convexity and the assumptions in the original analysis. We believe that the paper has really strong assumptions on the behaviour of the functions and these are essential in guaranteeing convergence. The paper proposes guidelines on making kinetic energies suitable to given objective functions, however, due to lack of time, we cannot comment on the same for the stochastic case. These methods have not been tested on real world datasets as the loss functions then might not obey the strict assumptions and fail to converge. This posed a problem as minibatching would have been a very easy way to obtain stochastic gradients. Thus, to experimentally check our method, we have used a simple noise model similar to 1001[2]. Further, we could work on an appropriate noise model to meet our strict stochastic assumptions.

III. Simulations and Experiments

In this section we use our own implementation[1] of the algorithms mentioned above with baselines, and plotting routines to investigate certain aspects of the methods. Concretely, our simulations focus on the following questions:

1) What is the dependence of convergence of the first Explicit Method on γ?
2) How do the methods converge with increasing dimensionality, and starting points for functions with different power behavior near zero and far from it?
3) Finally, how does the non-stochastic First Explicit Method compare to its stochastic counterpart?

A. Convergence of First Explicit method with varying γ

The function used in this case is $f(\mathbf{x}) = [x^{(1)} + x^{(2)}]^4 + [x^{(1)}/2 - x^{(2)}/2]^4$, with its corresponding kinetic energy as $k(\mathbf{p}) = \frac{3}{4}[(p^{(1)})^{4/3} + (p^{(2)})^{4/3}]$. Comparisons with Momentum and Gradient Descent have already been performed in [1]. We present the plot in Figure 1. We have set the learning rate to $\epsilon = 0.007$.

From the figure, we see that as γ increases, the convergence is faster. Interestingly, although the convergence proof requires that γ be less that 1, we found that larger values of gamma here allowed for even faster rates of convergence. This is one such case where the convergence is allowed in spite of this condition.

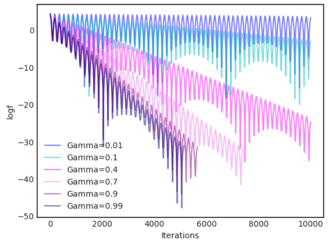


Fig. 1: Convergence of $logf$ with increasing values of γ.

B. Comparing non-stochastic methods

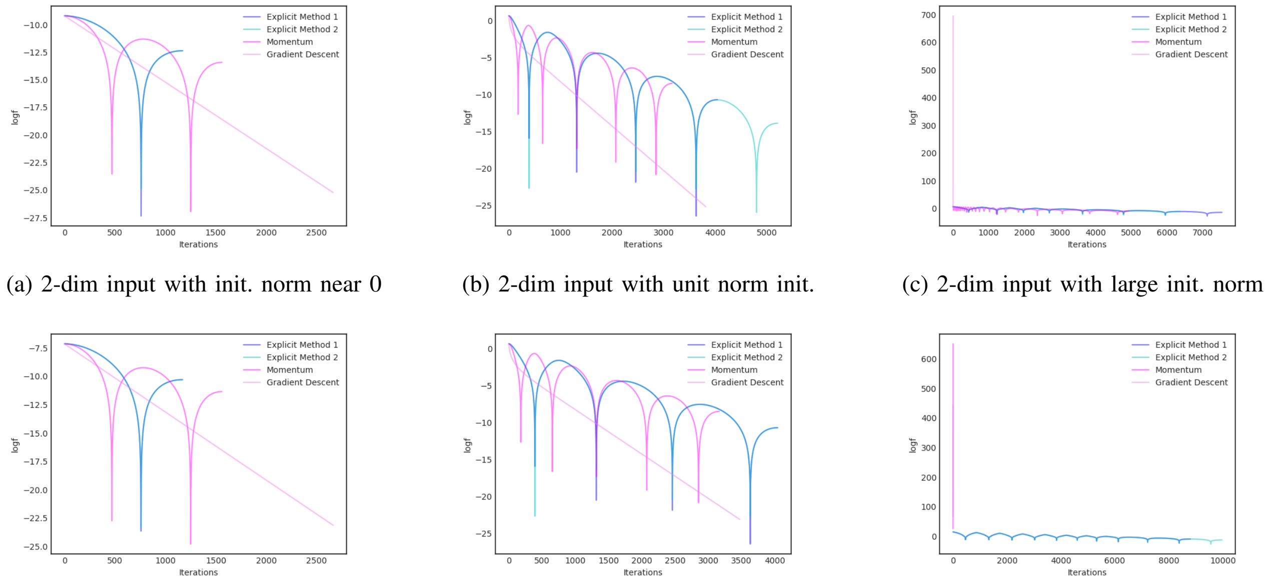
Below we provide an empirical comparison between First Explicit Method, Second Explicit Method, Momentum and Gradient Descent. Out primary variables of change here would be the dimensionality of $\mathbf{x}$, and the starting point of the optimizer. To look at varying power behaviours using a single objective, we consider the function $f(\mathbf{x}) = \psi_b^B(||\mathbf{x}||) = \frac{1}{B}(||\mathbf{x}||^b + 1)^{\frac{B}{b}} - \frac{1}{B}$. This function exhibits power behaviour of $||\mathbf{x}||^b$ near 0 and $||\mathbf{x}||^B$ far from it. The kinetic function map of such a function is $k(\mathbf{p}) = \psi_a^A(||\mathbf{p}||_*)$, where $A = \frac{B}{B-1}$ and $a = \frac{b}{b-1}$, with $||\cdot||_*$ being the dual norm. In this and the next subsection, we fix $B = 8$ and $b = 2$. Also, for this subsection, $\epsilon = 0.003, \gamma = 0.9$.

Figure 2 shows us the comparison. The first row is dimensionality 2, while the lower row of plots is dimensionality 16. We see here that for low dimensionality, the Explicit Methods and Momentum are able to converge, with similar rates. However Gradient Descent diverges with a starting point that is far from 0. This may be explained by overshooting the function due to large gradient far from 0, while the other 2 methods are able to mitigate this overshooting eventually using the auxilliary variable $\mathbf{p}$. With large dimensions, however, (Figure 2(d)-(f)) we see that even Momentum fails to converge in the case of large norm starting point. Only the kinetic maps prescribed by the Explicit methods are able to converge in high dimensions. We also tested this with 256 dimensions, and found the observation to be exaggerated from the previous case.

C. Comparing Stochastic and Non-Stochastic variants of First Explicit Method

In this subsection we analyse how the stochastic variant of the First Explicit method performs compared to its deterministic counterpart. The objective used here is the

[1] Code is available at https://github.com/jaivardhankapoor/ee609



(a) 2-dim input with init. norm near 0

(b) 2-dim input with unit norm init.

(c) 2-dim input with large init. norm

(d) 16-dim input with init norm near 0

(e) 16-dim input with unit norm init.

(f) 16-dim input with large init. norm

Fig. 2: Convergence plots of $logf$ with different dimensions and starting points

same as in Subsection III-A. We specify the noise to be added to the gradients as $\nabla f(\mathbf{x}) \cdot s$ and $\nabla k(\mathbf{p}) \cdot s$, where s is sampled from a Gaussian distribution with mean 0 and variance σ^2. We test the convergence on a grid of variance learning rate values, and plot the cases most archetypal of the behaviour of these methods.

Figure 3 shows the convergence plots of the 2 methods. We infer that for small σ^2, the stochastic method with the above noise formulation does not diverge too much from the deterministic one. In fact, in a few of the cases we saw, it even beat the deterministic method by a small margin. However, for large variance, we see that the bottom row of the figure clearly shows small to moderate divergence for small ϵ and large divergence for large ϵ. We hypothesize this is due to a very simplistic noise formulation, which we do not know if it satisfies the assumptions for the stochastic case described previously. A more hand-crafted noise function might lead to similar convergence behaviour as the deterministic one.

IV. Conclusion

We have analysed both theoretically and experimentally the facets of the Hamiltonian Gradient Descent. We proposed a stochastic variant to the original formulation and reproduced the results as well as studying other aspects like the effect of γ. Due to space constraint, we had to skip certain details. The effectiveness of this method motivates us to further investigate ODE literature and continuous time methods.

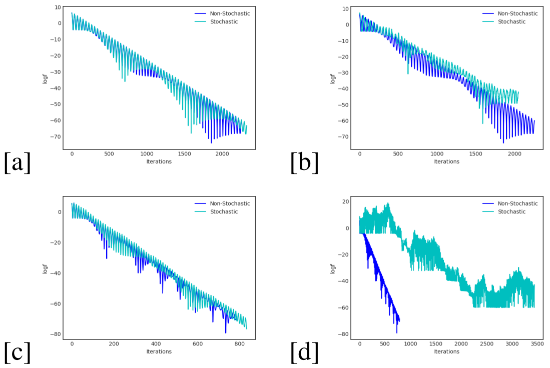


Fig. 3: Convergence comparison of variants of First Explicit method with different variance and learning rates. (a):$(\sigma^2 = 0.1, \epsilon = 0.003)$, (b):$(\sigma^2 = 1, \epsilon = 0.003)$, (c):$(\sigma^2 = 1, \epsilon = 0.01)$, (d):$(\sigma^2 = 1, \epsilon = 0.01)$

References

[1] C. J. Maddison, D. Paulin, Y. Whye Teh, B. O'Donoghue, and A. Doucet, "Hamiltonian Descent Methods," *arXiv e-prints*, p. arXiv:1809.05042, Sep 2018.

[2] T. Chen, E. B. Fox, and C. Guestrin, "Stochastic gradient hamiltonian monte carlo," in *Proceedings of the 31st International*

Conference on International Conference on Machine Learning - Volume 32, ser. ICML'14. JMLR.org, 2014, pp. II–1683–II–1691. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3045080

[3] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, 2018, pp. 6571–6583.

[4] R. M. Neal, "MCMC using Hamiltonian dynamics," *arXiv e-prints*, p. arXiv:1206.1901, Jun 2012.

[5] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *Ussr Computational Mathematics and Mathematical Physics*, vol. 4, pp. 1–17, 12 1964.