EXPOSITION ON THE CONNECTIONS BETWEEN DESCENT METHODS AND HAMILTONIAN MECHANICS

JAMES GILLBRAND

Abstract. Optimization techniques like gradient descent seek to find the extrema of their objective functions, just like water pools at the bottom of your hands. This expositional paper will detail the process of constructing gradient descent algorithms, starting with basic Hamiltonian Mechanics. Additionally, the paper will cover a selection of related proofs and lemmas, as well as the concept of a symplectic flow.

Contents

1. Introduction

Academic research on pursuit of faster and more efficient optimization algorithms has become a key part of modern applied math and computer science research. This focus is largely motivated by the increasing importance of these methods in the tuning of machine learning and neural network models.

One of the classic examples of an "accelerated" gradient descent algorithm is Nesterov's Accelerated Gradient Descent which has been proven to approach the minima faster than regular gradient descent [1]. It achieves this "acceleration" by incorporating a "momentum" term which preserves some information from previous steps, whereas gradient descent only considers the gradient. This mimics the movement of real physical objects and raises questions about the relationship between physical mechanics and computational optimization. This paper will focus on a similar momentum based descent method, the classical momentum method.

This paper will start with a consideration of Hamiltonian Mechanics and proceed to construct both classical momentum as well as traditional gradient descent. The first section will provide an brief introduction to Hamiltonian Mechanics and its applications for those unfamiliar. The next section will consider numerical integration methods native to Hamiltonian Mechanics and the final section will cover the motivation for the modifications to these methods which yield the methods of descent.

Date: AUGUST 14, 2024.

2. Hamiltonian Mechanics

Hamiltonian Mechanics is a method of quantifying and solving questions related to physical systems. Each object is defined by two d-dimensional vectors, p for momentum and x for position. It was preceded and built on the back of Lagrangian Mechanics, but knowledge of either is not necessary to understand this paper.

Definition 2.1. A Hamiltonian is the sum of the total energy of a system. Let m be the mass of the object. With $f(x)$ for potential function and $V = \frac{p^2}{2m}$ for kinetic function, we take

$$H \stackrel{\text{def}}{=} f(x) + V$$

Additionally, the movement of an object in this system is governed by the following set of differential equations.

Definition 2.2. Hamilton's equations dictate that,

$$\frac{dx}{dt} = \frac{\partial H}{\partial p} \quad \text{and} \quad \frac{dp}{dt} = -\frac{\partial H}{\partial x}$$

These are the two main definitions that describe Hamiltonian Mechanics and an object obeying these rules is said to follow a "Hamiltonian Flow".

Definition 2.3. Let x_0 and p_0 be the initial position and momentum vectors. Let t be time and let Hamiltonian Flow be defined as follows.

$$(x, p) \stackrel{\text{def}}{=} F_t(x_0, p_0)$$

Since the Hamiltonian is supposed to be the sum of energy in a system, it should also remain constant all along this flow, which we will prove below.

Lemma 2.4. *The Hamiltonian is a constant such that,*

$$\frac{dH}{dt} = 0$$

Proof. By the chain rule,

$$\frac{dH}{dt} = \frac{\partial H}{\partial p}\frac{dp}{dt} + \frac{\partial H}{\partial x}\frac{dx}{dt}$$

Then, applying Definition 2.2 yields the following,

$$\frac{dH}{dt} = \frac{\partial H}{\partial p} \cdot -\frac{\partial H}{\partial x} + \frac{\partial H}{\partial x} \cdot \frac{\partial H}{\partial p} = 0$$

□

With this established, it follows that that simply following a Hamiltonian flow will not usually yield a minima. Since energy is conserved, when $f(x)$ is minimized, V will be maximized. As such the object will not remain at the minima, and instead will continue to move. So, in order to find a minima, we must modify the flow to be able to descend energy levels. One way of doing this, is by breaking the flow up into steps. Then, the algorithm for following the flow, called Hamiltonian descent, looks like this.

$$(x_k, p_k) = F_t(x_{k-1}, p_{k-1}) \tag{2.5}$$

We can make this descend energy levels by scaling the momentum term or removing it entirely. Let $0 \leq \mu \leq 1$. This new equation will eliminate energy and descend.

$$(x_k, p_k) = F_t(x_{k-1}, \mu p_{k-1}) \tag{2.6}$$

Lemma 2.7. *The value of potentials yielded by Hamiltonian descent are non-increasing.*

$$f(x_k) \leq f(x_{k-1})$$

Proof. For each natural number k, by the definition of the Hamiltonian,

$$f(x_{k-1}) = H(x_{k-1}, 0)$$

Also, because energy is conserved along the Hamiltonian Flow,

$$H(x_{k-1}, 0) = H(x_k, p_k)$$

$$H(x_k, p_k) = f(x_k) + \frac{p_k^2}{2m}$$

Since m is mass, $m \geq 0$. Hence, $\frac{p_k^2}{2m} \geq 0$. Consequently,

$$f(x_k) = f(x_{k-1}) - \frac{p_k^2}{2m} \leq f(x_{k-1})$$

□

It is important to note that the step-size t is best kept small as following the flow for long enough without removing the momentum term can result in a cycle. Additionally, another issue arises when we try and implement this algorithm. In general, there is no way to find the exact flow F. As such, we will have rely on numerical approximations for this flow.

3. Numerical Integration

We will employ numerical integration to try and approximate the Hamiltonian flow of our object. Similarly to how a beginner calculus student approximates an integral with Riemann sums, these methods similarly work with the differential equations given by Hamilton's equations to try and approximate this flow.

When using a numerical integrator, it is possible to distort the space in which you are working. It is essential that the integrator that we select preserves the qualities of the Hamiltonian System like conservation of energy so that only our intentionally dampening impacts the results. This means preserving the symplectic form.

Definition 3.1. Let ξ and η be arbitrary d-dimensional vectors. The symplectic form ω is given by,

$$\omega(\xi, \eta) = \sum_{i=1}^{d} d\xi_i \wedge d\eta_i$$

This object represents a generalized concept of area for upper dimensional spaces. Additionally, there are a series of transformations which preserve this form.

Definition 3.2. Let $A \in \mathbb{R}^{2d}$ be an open set. Let J be our symplectic matrix such that,

$$J = \begin{bmatrix} 0 & I_d \\ -I_d & 0 \end{bmatrix}$$

where I_d is an identity matrix of dimension d. A differentiable transformation $g : A \to \mathbb{R}^{2d}$ is a symplectic transformation if the Jacobian matrix $g'(x,p)$ is symplectic, i.e.

$$\omega(g'(x,p)\xi, g'(x,p)\eta) = \omega(\xi,\eta)$$

or

$$g'(x,p)^T J g'(x,p) = J$$

These transformations will preserve the energy of the system and generally behave better due to this property. In fact, the very Hamiltonian Flow we are attempting to approximate is itself a symplectic transformation.

Theorem 3.3. ***(Poincaré 1899)*** *Let $H(x,p)$ be a twice continuously differentiable function on $A \subset \mathbb{R}^{2d}$. For each t, the Hamiltonian Flow F_t is a symplectic transformation.* [2]

Proof. Let $z_k = (x_k, p_k)$. For $t = 0$, F is the identity map. Hence,

$$\frac{\partial F_0}{\partial z_0} = 1$$

Thus,

$$\left(\frac{\partial F_0}{\partial z_0}\right)^T J \left(\frac{\partial F_0}{\partial z_0}\right) = J$$

Note that $\partial F_t/\partial z_0$ is a solution of the variational equation $\frac{d}{dt}\Psi = J^{-1}\nabla^2 H(x,p)\Psi$. Hence,

$$\frac{d}{dt}\frac{\partial F_t}{\partial z_0} = J^{-1}\nabla^2 H(x,p)\frac{\partial F_t}{\partial z_0}$$

Consequently, by the product rule and the previous fact,

$$\frac{d}{dt}\left(\left(\frac{\partial F_t}{\partial z_0}\right)^T J \left(\frac{\partial F_t}{\partial z_0}\right)\right) =$$

$$\left(\frac{d}{dt}\frac{\partial F_t}{\partial z_0}\right)^T J \left(\frac{\partial F_t}{\partial z_0}\right) + \left(\frac{\partial F_t}{\partial z_0}\right)^T J \left(\frac{d}{dt}\frac{\partial F_t}{\partial z_0}\right) =$$

$$\left(J^{-1}\nabla^2 H(x,p)\frac{\partial F_t}{\partial z_0}\right)^T J \left(\frac{\partial F_t}{\partial z_0}\right) + \left(\frac{\partial F_t}{\partial z_0}\right)^T J \left(J^{-1}\nabla^2 H(x,p)\frac{\partial F_t}{\partial z_0}\right)$$

Note that $J^T = -J$ by the definition of J. Hence, $J^{-T}J = -I$. Also, because H is twice continuously differentiable, for each natural i and j less than or equal to d,

$$\frac{\partial^2 H(z)}{\partial z_i \partial z_j} = \frac{\partial^2 H(z)}{\partial z_j \partial z_i}$$

It follows that $(\nabla^2 H(x,p))^T = \nabla^2 H(x,p)$. Thus,

$$\left(J^{-1}\nabla^2 H(x,p)\frac{\partial F_t}{\partial z_0}\right)^T J\left(\frac{\partial F_t}{\partial z_0}\right) + \left(\frac{\partial F_t}{\partial z_0}\right)^T J\left(J^{-1}\nabla^2 H(x,p)\frac{\partial F_t}{\partial z_0}\right) =$$
$$-\left(\frac{\partial F_t}{\partial z_0}\right)^T \nabla^2 H(x,p)\left(\frac{\partial F_t}{\partial z_0}\right) + \left(\frac{\partial F_t}{\partial z_0}\right)^T \nabla^2 H(x,p)\left(\frac{\partial F_t}{\partial z_0}\right) = 0$$

Since the time derivative of this term is 0, for any selection of t,

$$\left(\frac{\partial F_t}{\partial z_0}\right)^T J\left(\frac{\partial F_t}{\partial z_0}\right) = J$$

Thus F_t is a symplectic transformation. □

Knowing that the flow we are attempting to approximate is a symplectic transformation, it seems appropriate to use a numerical integrator that also preserves the symplectic structure of the space. We will elect to use symplectic Euler.

Definition 3.4. Let Δt be the step-size and let k be the number of iterations. Symplectic Euler is an algorithm that updates as follows.

$$x_{k+1} = x_k + \Delta t(p_{k+1}) \tag{3.5}$$

$$p_{k+1} = p_k - \Delta t\frac{\partial H}{\partial x_k} \tag{3.6}$$

4. Gradient Descent

Before proceeding, note the definitions for the two methods of descent that we wish to produce.

Definition 4.1. Let h be a non-negative real number representing the step-size. Traditional gradient descent is defined as follows.

$$x_{k+1} = x_k + h\nabla f(x_k)$$

Definition 4.2. Let h be a non-negative real number representing the step-size. Classical momentum is defined as follows.

$$x_{k+1} = x_k + p_{k+1}$$
$$p_{k+1} = \gamma p_k - h\nabla f(x_k)$$

Now that the form of integration and the concept of dampening the momentum has been established, we can modify the symplectic Euler method by adding the μ just as we did in (2.6). Further, note that $\frac{\partial H}{\partial x} = \nabla f$. Implementing these changes to symplectic Euler yields the following.

$$x_{k+1} = x_k + \Delta t(p_{k+1})$$
$$p_{k+1} = \mu p_k - \Delta t\nabla f(x_k)$$

In the case that $\mu = 0$, this formula yields exactly gradient descent.

$$x_{k+1} = x_k - \Delta t^2\nabla f(x_k) \tag{4.3}$$

Note that we must change from h to Δt to check if they are actually equivalent. However, since both are independent variables, we can replace then with the relation.

$$\Delta t^2 \stackrel{\text{def}}{=} h \tag{4.4}$$

Additionally, we can replace μ and Δt with γ as follows to make the algorithm take the form of classical momentum.

$$\mu \Delta t \stackrel{\text{def}}{=} \gamma \tag{4.5}$$

Changing the variables of the Euler method yields the following.

$$x_{k+1} = x_k + \Delta t(p_{k+1})$$
$$p_{k+1} = \frac{\gamma}{\Delta t} p_k - \Delta t \nabla f(x_k)$$

This can be rewritten as exactly classical momentum.

$$x_{k+1} = x_k + p_{k+1}$$
$$p_{k+1} = \gamma p_k - h \nabla f(x_k)$$

5. Example

To conclude this paper, we will illustrate the differences between numerical integrators and descent methods by applying them to a few physical problems. The first of which is that a spring. First we will define the potential function as follows.

$$f(x) \stackrel{\text{def}}{=} \frac{1}{2} k x^2 \tag{5.1}$$

This is the traditional notation for a spring potential. The variable k is a positive real number that depends on the strength of the spring. Given this, the symplectic Euler method, classical momentum and gradient descent can all be implemented on the system. In accordance with (4.4), Δt has been replaced with h.

Symplectic Euler

$$x_{k+1} = x_k + p_{k+1}$$
$$p_{k+1} = \sqrt{h} p_k - h(kx)$$

Classical Momentum Descent

$$x_{k+1} = x_k + p_{k+1}$$
$$p_{k+1} = \mu \sqrt{h} p_k - h(kx)$$

Gradient Descent

$$x_{k+1} = x_k + p_{k+1}$$
$$p_{k+1} = -h(kx)$$

Note that each of these algorithms are identical, excepting the coefficient on the p_k term. Further, both the symplectic Euler method and regular gradient descent are special cases of the classical momentum descent method. Intuitively, we expect that the Euler method will return an approximately harmonic oscillation while the descent methods will minimize the potential function. We can test this by writing a simple Python program to plot the progress of the two methods. First we must

choose values for the various constants.

$$k = 1$$
$$x_0 = 1$$
$$p_0 = 0$$
$$\Delta t^2 = h = .5$$
$$\gamma = .5$$

Performing the calculations yields the following plots, which illustrate how the Euler method conserves energy while the descent method seeks expressly to eliminate it in search of a minima. For classical momentum, we can calculate the μ value with the relation given by (4.5). $\mu = \frac{.5}{\sqrt{.5}} = \sqrt{.5}$

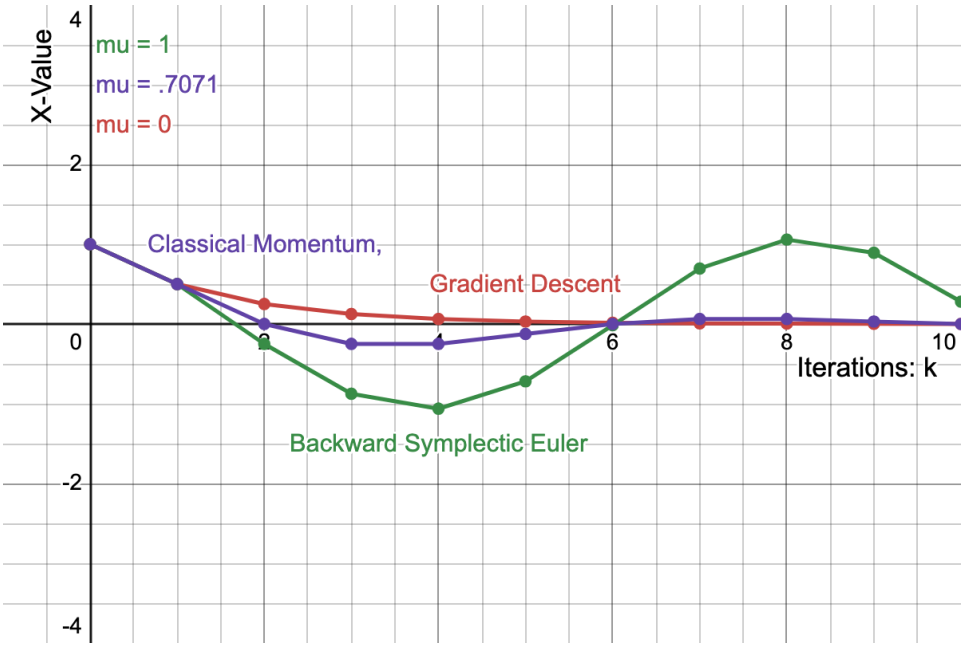

FIGURE 1. Numerical Integrators on a Spring

Since μ is the dampening coefficient, the values align with the behavior visible on the graph.

The use of integrators with values other than 1 for μ may be of use in the modeling of physical systems where natural forces like friction dampen and dissipate energy. This graph illustrates how the methods of abstract functional minimization can be traced back to methods of modeling the physical world. As such, it begs the question of which other real world phenomena might inspire innovations in the sphere of more abstract optimization. There exists research that examine the implementation of concepts of relativity into descent methods as well as examining the concept of symplectic flows in the context of descent methods. [3]. Lastly, it is worth examining other methods of Numerical Integration to see how they fit into this connection with descent methods.

6. Acknowledgements

It was a pleasure to work with my mentor, Antares Chen, throughout the writing of this paper. He proved instrumental in pointing me towards interesting topics and was always able to provide insight into the subjects I studied.

References

[1] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence O(1/k2). In Doklady AN USSR, volume 269, pages 543–547, 1983.

[2] H. Poincaré, Les Methodes Nouvelles de la Mécanique Céleste. Tome III, Gauthiers-Villars, Paris, 1899

[3] França, Guilherme, et al. "Conformal symplectic and relativistic optimization." Advances in Neural Information Processing Systems 33 (2020): 16916-16926.

[4] Ernst Hairer, Marlis Hochbruck, Arieh Iserles, Christian Lubich, Geometric Numerical Integration. Oberwolfach Rep. 3 (2006), no. 1, pp. 805–882

[5] Ahn, Kwangjun, and Suvrit Sra. "Understanding nesterov's acceleration via proximal point method." Symposium on Simplicity in Algorithms (SOSA). Society for Industrial and Applied Mathematics, 2022.