ODE ANALYSIS OF GRADIENT BASED UNCONSTRAINED OPTIMIZATION ALGORITHMS

Aditya Shirwatkar*
Indian Institute of Science (IISc)
Bangalore
adityasr@iisc.ac.in

ABSTRACT

This work focuses on analyzing well-established gradient-based unconstrained optimization algorithms like Gradient Descent (GD), Mirror Descent (MD), Nesterov's Accelerated Gradient Descent (NAGD) and Newton's Method (NM). To achieve this, we derive the continuous-time ordinary differential equations (ODEs) of these algorithms, by comparing their iterative update rule to a discretized version of the corresponding ODE. We then construct Lyapunov functions to investigate the convergence and stability properties of the equilibrium points of the ODE (which correspond to stationary points or local minima of the objective function). Additionally, we consider some special cases for which the rate of convergence for all of them can be shown as $\mathcal{O}(1/t^p)$, where $p \in \mathbb{R}_+$. Lastly, we propose a new Lyapunov function for the ODE corresponding to Newton's method and provide a new simple proof for its convergence rate.

1 Introduction

Numerous methods have been developed to solve an unconstrained optimization problem of the form $\min_X \; f(X)$, where $X \in \mathbb{R}^n$ is the decision or parameter and $X^* = 0$ is the minima. A few such popular algorithms are Gradient Descent (GD), Mirror Descent (MD) [1], Nesterov's Accelerated Gradient Descent (NAGD) [2], and Newton's Method (NM). These algorithms and their various properties, such as convergence rate and numerical stability, have been well-studied in the literature.

However, the majority of studies have focused on their discretized forms. Alternatively, some authors have tried to derive a continuous Ordinary Differential Equation (ODE) corresponding to these algorithms [3, 4, 5]. Once such an ODE is formulated, employing Lyapunov-like arguments becomes feasible to assess the stability or convergence rate concerning the ODE's equilibrium points. These equilibrium points align with the stationary points of the objective function, thus offering an avenue for analysis.

In this work, we re-establish numerous results from different studies concerning ODE analysis for earlier mentioned gradient-based unconstrained optimization algorithms. We comment on the behavior of solutions of the ODEs around the stationary points and derive their rate of convergence as $\mathcal{O}(1/t^p)$, where $p \in \mathbb{R}_+$. As a novelty, our work also proposes a new Lyapunov function for the ODE corresponding to Newton's method and a new proof for its convergence rate.

2 Preliminaries

This section outlines several key concepts essential for understanding the results presented in this work. The proofs for these statements are omitted, as they have been extensively covered in existing literature [6].

An Ordinary Differential Equation (ODE) is a system of differential equations denoted as $\dot{x}(t) = g(t, x(t)); t \geq 0$, where $x(t)$ represents the solution to the ODE with an initial condition $x(0)$, and $g(t, x(t))$ is a well-behaved nonlinear

*This work is done as a project for the course DS:211 Numerical Optimization function of the input arguments. This work assumes the existence and uniqueness of a solution to this system, indicating that $g(t, x(t))$ is Lipschitz continuous with a constant L. This assumption is justifiable as most optimization-related problems are typically well-defined. For simplicity, we use the notations $x(t) \equiv x$, $x(0) \equiv x_0$, and $\dot{x}(t) \equiv \dot{x}$.

The equilibrium points of an ODE correspond to solutions where $\dot{x} = 0$. It suffices to examine the behavior of x around these equilibrium points to assess whether the optimization algorithm converges to a stationary point. This analysis can be facilitated by utilizing the following stability definitions.

An equilibrium point $x = 0$ of an ODE $\dot{x} = g(t, x)$ is characterized as follows:

1. Stable, if $\forall\, \varepsilon > 0,\ \exists\, \delta(\varepsilon, t_0) > 0$ s.t. $||x(t_0)|| \leq \delta(\varepsilon, t_0) \implies ||x(t)|| \leq \varepsilon,\ \forall t \geq t_0 \geq 0$.
2. Unstable if it is not stable.
3. Uniformly Stable (US), if $\forall\, \varepsilon > 0, \exists\, \delta(\varepsilon) > 0$ s.t. $||x(t_0)|| \leq \delta(\varepsilon) \implies ||x(t)|| \leq \varepsilon,\ \forall t \geq t_0 \geq 0$.
4. Asymptotically Stable (AS), if it is stable and $\exists\, c(t_0) > 0$ s.t. $x(t) \to 0$ as $t \to 0,\ \forall\, ||x(t_0)|| < c(t_0)$.
5. Uniformly Asymptotically Stable (UAS), if it is US and $\exists\, c > 0$ s.t. $x(t) \to 0$ as $t \to 0,\ \forall\, ||x(t_0)|| < c$.
6. Exponentially Stable (ES), if $\exists\, c, k, \lambda > 0$ s.t. $||x(t)|| \leq k||x(t_0)||e^{-\lambda(t-t_0)},\ \forall\, ||x(t_0)|| < c$

While obtaining bounds in these stability definitions is generally straightforward, finding the solution x itself poses a non-trivial challenge. Alternatively, we can employ Lyapunov's argument, constructing a function analogous to the system's energy to comment on stability.

Theorem 1 (*Lyapunov Stability Theorem*)**.** *Let $x = 0$ be an equilibrium point for an ODE $\dot{x} = g(x)$ and $\mathbb{D}$ be a domain with $0 \in \mathbb{D}$. Let $V : \mathbb{D} \to \mathbb{R}$ (Lyapunov or Energy Function) be a C^1 function s.t.,*

1. $V(x) > 0,\ \forall x \in \mathbb{D} \setminus \{0\},\ \&\ V(0) = 0$
2. $\dot{V}(x) = \frac{\partial V}{\partial x} g(t, x) \leq 0,\ \forall\, x \in \mathbb{D}$,

then $x = 0$ is stable. Further if $V(x) < 0,\ \forall\, x \in \mathbb{D} \setminus \{0\}$, s AS.

This simplifies our analysis, as we only need to establish a candidate Lyapunov function and fulfill the two conditions. While this theorem proves highly effective for time-invariant systems, extending it to time-variant systems requires careful modification of Theorem 1. We will first introduce a useful lemma in this context to address the uniform convergence in time.

Lemma 1. *For the Ordinary Differential Equation (ODE) $\dot{x} = g(t, x)$, an equilibrium solution $x = 0$ exhibits Uniform Stability (US) if and only if there exists a class $\mathcal{K}$ function $\alpha(\cdot)$ and a constant $c > 0$ such that, for all $t \geq t_0$ and all $||x(t_0)|| < c$, the inequality $||x(t)|| \leq \alpha(||x(t_0)||)$ holds.*

Here, a continuous function $\alpha : [0, a) \to [0, \infty)$ is considered of class $\mathcal{K}$ if it is strictly increasing and $\alpha(0) = 0$. If $a = \infty$, and $\alpha(r) \to \infty$ as $r \to \infty$, then it is classified as belonging to class $\mathcal{K}_\infty$.

Using this lemma, we can now state the following theorem.

Theorem 2 (*Lyapunov Stability Theorem for Time Variant Systems*)**.** *Let $x = 0$ be an equilibrium point for an ODE $\dot{x} = g(t, x)$ and $\mathbb{D} \subseteq \mathbb{R}^n$ be a domain with $0 \in \mathbb{D}$. Let $V : [0, \infty) \times \mathbb{D} \to \mathbb{R}$ (Lyapunov-like or Energy Function) be a C^1 function s.t.,*

1. $\alpha_1(||x||) \leq V(t, x) \leq \alpha_2(||x||)$,
2. $\dot{V}(t, x) = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} g(t, x) \leq -\alpha_3(||x||)$,

$\forall t \geq 0,\ \forall x \in \mathbb{D}$, where α_1, α_2, and α_3 are class $\mathcal{K}$ functions on $[0, r)$. Then $x = 0$ is UAS.

It's worth noting that both $V(t, x)$ and $\dot{V}(t, x)$ are bounded by a Lyapunov function, which justifies the term "Lyapunov-like." When these bounding functions exhibit a particular structure involving powers of $||x||$, a stronger form of convergence can be shown. Formally, we can state it below.

Corollary 1. *If Theorem 2 satisfies $\alpha_i(||x||) = k_i||x||^a$, where $k_i, a > 0, \forall\, i \in \{1, 2, 3\}$, then $x = 0$ is ES.*

The mentioned results provide a foundation for ODE analysis in Euclidean geometry. However, additional terminologies become essential in more general scenarios, such as when dealing with the Mirror Descent algorithm or its variants.

Let $\mathbb{E}$ denote a finite-dimensional real vector space, and $\mathbb{E}^*$ represent its dual space. For a given function f, the dual function f^* is defined as follows: $f^*(z) = \sup_{x \in \mathbb{E}} \langle z, x \rangle - f(x)$. It's important to note that f^* is always convex, as it is a pointwise maximum of a linear function in z. As an example, if $f(x) = x^2/2$, then $f^*(z) = \sup_{x \in \mathbb{R}} zx - x^2/2 = z^2/2$.

Consider a continuously-differentiable, strictly convex function $\varphi : \Omega \to \mathbb{R}$ defined on a convex set Ω. For any $a, b \in \Omega$, the Bregman divergence from x to y is defined as $D_\varphi(a, b) := \varphi(a) - \varphi(b) - \langle \nabla \varphi(b), a - b \rangle$. This divergence measures the difference between two points expressed in terms of the function φ.

To illustrate, if $\varphi(\cdot) = \frac{1}{2}||\cdot||_2^2$, the induced Bregman Divergence is the Squared Euclidean Norm, given by $D_\varphi(a, b) = \frac{1}{2}||a||_2^2 - \frac{1}{2}||b||_2^2 - \langle b, a - b \rangle = \frac{1}{2}||a - b||_2^2$. This distance metric represents a statistical distance when interpreting a, b as probability distributions.

3 ODE Analysis

3.1 Gradient Descent (GD)

To better understand the convergence of the vanilla Gradient Descent (GD) algorithm, let's consider its iterative form: $X_{k+1} = X_k - \alpha_k \nabla f(X_k)$, where f is the objective function to minimize. The minimizer of f is denoted as $X^* = 0$. Viewing the search method as an Euler discretization of the Ordinary Differential Equation (ODE) $\dot{x} = -\nabla f(x), x_0 = X_0$, with α_k as the constant discretization step size. The continuous-time function x represents the trajectory of iterates. The equilibrium point $\nabla f(x) = 0$ is equivalent to $X^* = 0$, i.e., $x = 0$.

Let us construct a candidate Lyapunov function $V = \frac{1}{2}||x||^2$, then $\dot{V} = \langle -\nabla f(x), x \rangle$. By convexity of f we have, $f(0) \geq f(x) + \langle \nabla f(x), 0 - x \rangle \implies \langle -\nabla f(x), x \rangle \leq f(0) - f(x)$. Using this we get, $\dot{V} \leq f(0) - f(x) < 0$. Hence, V satisfies all the conditions of Theorem 1, implying that the $x = 0$ is AS.

While we now understand that the ODE converges to the equilibrium point, the next question is at what rate? Estimating an upper bound on $||x||$ or $f(x) - f(0)$ for the chosen Lyapunov function proves challenging without assuming the specific form of f or constructing another Lyapunov function for which upper bounds can be obtained easily. Instead, we analyze how $f\left(\frac{1}{t}\int_0^t x(\tau)d\tau\right) - f(0)$ behaves, akin to bounding the "average" of the solution.

To proceed further, we will integrate $\dot{V}(x)$ and use the Jenson's Inequality, $\frac{1}{t}\int_0^t f(x(\tau))d\tau \leq f\left(\frac{1}{t}\int_0^t x(\tau)d\tau\right)$,

$$
\begin{aligned}
\int_0^t \dot{V}(x)d\tau = V(x) - V(x_0) \leq tf(0) - \int_0^t f(x(\tau))d\tau, \\
\frac{1}{t}\int_0^t f(x(\tau))d\tau - f(0) \leq \frac{1}{t}(V(x_0) - V(x)), \\
f\left(\frac{1}{t}\int_0^t x(\tau)d\tau\right) - f(0) \leq \frac{1}{t}(V(x_0) - V(x)), \\
f\left(\frac{1}{t}\int_0^t x(\tau)d\tau\right) - f(0) \leq \frac{1}{t}V(x_0) \equiv \mathcal{O}(1/t).
\end{aligned}
$$

Thus, we have shown that for a convex objective function in Euclidean geometry, the Gradient Descent ODE converges at a rate of $\mathcal{O}(1/t)$.

3.2 Mirror Descent (MD)

In contrast to the GD analysis, which is more straightforward due to its compatibility with Euclidean geometry, the Mirror Descent (MD) algorithm demands a distinct approach to construct Lyapunov arguments. This necessity arises mainly because of dealing with dual variables and Bregman Divergence.

The following iteration gives the MD algorithm,

$$
X_{k+1} = \arg\min_X \left\{ \alpha_k \langle \nabla f(X_k), X - X_k \rangle - \frac{1}{\mu} D_\varphi(X, X_k) \right\}
$$

where, α_k is the step-size, φ is a μ-strongly convex function and D_φ is the corresponding Bregman Divergence induced by φ. To derive the ODE for this iteration, we would need a constant step size and an explicit iteration rule like GD rather than a recursive rule given by the solution of another optimization problem. Thus, we transform the above iteration rule into the following,

$$\begin{aligned} Z_{k+1} &= Z_k - \mu\alpha_k \nabla f(X_k), \\ X_{k+1} &= \nabla\varphi^*(Z_{k+1}), \end{aligned}$$

with initial $\nabla\varphi^*(Z_0) = X_0$. Here, $Z \in \mathbb{E}^*$ is the dual variable of $X \in \mathbb{E}$ (which is the primal variable). The "mirror" map $\nabla\varphi$ transports X to the dual space $\mathbb{E}^*$, and $\nabla\varphi^*$ is the dual function of $\nabla\varphi$ which maps Z back to the primal space $\mathbb{E}$. Taking $\alpha_k \to 0$, we obtain the continuous limit of above,

$$\begin{aligned} &\dot{z} = -\mu\nabla f(x), \\ &\dot{x} = \nabla\varphi^*(z), \\ &z_0 = Z_0,\ x_0 = X_0 \text{ with } \nabla\varphi^*(z_0) = x_0. \end{aligned}$$

For Lyapunov analysis, we'll construct a function on the dual variable denoted as $V(z) = D_{\varphi^*}(z, z^*)$, where z^* represents the equilibrium value of z. For simplicity, we set $x^*, z^* = 0$. Accordingly, using the definition of Bregman Divergence, $V(z) = \varphi^*(z) - \varphi^*(0) - \langle\nabla\varphi^*(0), z\rangle$, which is always Positive Definite. Taking the derivative, we obtain,

$$\begin{aligned} \dot{V}(z) &= \langle\nabla\varphi^*(z) - \nabla\varphi^*(0), \dot{z}\rangle = \langle x, \dot{z}\rangle, \\ &= -\langle x, \nabla f(x)\rangle \leq -(f(x) - f(0)) < 0. \end{aligned}$$

Applying the same reasoning as in the Gradient Descent ODE, we establish that V qualifies as a valid Lyapunov function according to Theorem 1. Consequently, $x = 0$ is AS, and $f\left(\frac{1}{t}\int_0^t x(\tau)d\tau\right) - f(0)$ converges to 0 at a rate of $\mathcal{O}(1/t)$. Lastly, it's noteworthy that the Gradient Descent ODE can be viewed as a special case of the Mirror Descent ODE by choosing $\varphi^*(z) = \frac{1}{2}||z||_2^2$. In this case, $\nabla\varphi^*$ is the identity, causing x and z to coincide.

3.3 Nesterov's Accelerated Gradient Descent (NAGD)

A family of methods utilizes accumulated gradients to improve convergence beyond the linear rates achieved by vanilla Gradient Descent (GD) or Mirror Descent (MD) algorithms. One such method of interest in this work is Nesterov's Accelerated Gradient Descent (NAGD). This algorithm exhibits faster convergence and is expressed through the following iterative steps,

$$\begin{aligned} X_k &= Y_{k-1} - \alpha_k \nabla f(Y_{k-1}), \\ Y_k &= X_k + \beta_k(X_k - X_{k-1}). \end{aligned}$$

Here, α_k denotes the stepsize, and β_k is the extrapolation parameter. In a specific scenario with a constant step size, combining the equations for $k-1$, k, and $k+1$ yields,

$$X_{k+1} - X_k = \beta_{k-1}(X_k - X_{k-1}) - \alpha\nabla f(Y_k).$$

To derive the continuous-time Ordinary Differential Equation (ODE) for $x(t)$, we must establish a connection between t and k. Setting $t = k\alpha$ initially seems intuitive, but the derivation fails. Instead, we redefine time as $t = k\sqrt{\alpha}$, leading to,

$$\begin{aligned} x(t) &= x(k\sqrt{\alpha}) = X_k + \mathcal{O}(\sqrt{\alpha}), \\ x(t + \sqrt{\alpha}) &= x((k+1)\sqrt{\alpha}) = X_{k+1} + \mathcal{O}(\sqrt{\alpha}), \\ x(t - \sqrt{\alpha}) &= x((k-1)\sqrt{\alpha}) = X_{k-1} + \mathcal{O}(\sqrt{\alpha}). \end{aligned}$$

Given that $\sqrt{\alpha} \to 0$, we can employ Taylor's series to approximate as follows,

$$\begin{aligned} x(t + \sqrt{\alpha}) &\approx x(t) + \sqrt{\alpha}\frac{\dot{x}(t)}{1!} + (\sqrt{\alpha})^2\frac{\ddot{x}(t)}{2!} + \mathcal{O}(\sqrt{\alpha}), \\ x(t - \sqrt{\alpha}) &\approx x(t) - \sqrt{\alpha}\frac{\dot{x}(t)}{1!} + (\sqrt{\alpha})^2\frac{\ddot{x}(t)}{2!} + \mathcal{O}(\sqrt{\alpha}). \end{aligned}$$

Combining these approximations, we derive the following,

$$\begin{aligned} X_{k+1} - X_k &= \sqrt{\alpha}\dot{x} + \frac{\alpha}{2}\ddot{x} + \mathcal{O}(\sqrt{\alpha}), \\ X_k - X_{k-1} &= \sqrt{\alpha}\dot{x} - \frac{\alpha}{2}\ddot{x} + \mathcal{O}(\sqrt{\alpha}), \\ \sqrt{\alpha}\dot{x} + \frac{\alpha}{2}\ddot{x} + \mathcal{O}(\sqrt{\alpha}) &= \beta_{k-1}(\sqrt{\alpha}\dot{x} - \frac{\alpha}{2}\ddot{x}) - \alpha\nabla f(Y_k). \end{aligned}$$

Assuming that in the long run $Y_k = X_k$, and we can take $y(t) = Y_k + \mathcal{O}(\sqrt{\alpha})$ and $y(t) = x(t)$. Thus we have,

$$\frac{\sqrt{\alpha}}{2}(1+\beta_{k-1})\ddot{x} + (1-\beta_{k-1})\dot{x} + \sqrt{\alpha}\nabla f(x) + \mathcal{O}(\sqrt{\alpha}) = 0.$$

For β_{k-1}, numerous choices are possible, as long as $(1-\beta_k)/\beta_k^2 \leq 1/\beta_{k-1}^2$. From the literature [7], we choose $\beta_{k-1} = 1 - 3\sqrt{\alpha}/t$. Consequently, our ODE becomes,

$$\begin{aligned}\frac{\sqrt{\alpha}}{2}\left(2 - \frac{3\sqrt{\alpha}}{t}\right)\ddot{x} + \left(\frac{3}{t}\right)\dot{x} + \sqrt{\alpha}\nabla f(x) + \mathcal{O}(\sqrt{\alpha}) = 0,\\ \frac{1}{2}\left(2 - \frac{3\sqrt{\alpha}}{t}\right)\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) + \mathcal{O}(\sqrt{\alpha}) = 0.\end{aligned}$$

Taking the limit as $\alpha \to 0$ yields a second-order ODE, $\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) = 0$. Similar to previous sections, we assume this ODE's equilibrium occurs at $x = 0$.

Given that the resulting ODE has a time-varying component, we consider a candidate Lyapunov-like function as,

$$\begin{aligned}&V(t,x) = t^2(f(x) - f(0)) + 2||x + \frac{t}{2}\dot{x}||^2,\\ &\implies \dot{V}(t,x) = 2t(f(x) - f(0)) + t^2\langle \nabla f(x), \dot{x}\rangle + 4\langle x + \frac{t}{2}\dot{x}, \frac{3}{2}\dot{x} + \frac{t}{2}\ddot{x}\rangle.\end{aligned}$$

Substituting $3\dot{x}/2 + t\ddot{x}/2$ with $-t\nabla f(x)/2$ gives us $\dot{V}(t,x) = 2t(f(x) - f(0)) + 2t\langle x, -\nabla f(x)\rangle < 0$, where the inequality follows from the strict convexity of f.

It can be shown that V is a valid Lyapunov-like function satisfying all the conditions of Theorem 2, but we omit the proof here. Now, for the order of convergence, due to the monotonicity of V and the non-negativity of $2||x + \frac{t}{2}\dot{x}||^2$, we have,

$$f(x) - f(0) \leq \frac{V(t,x)}{t^2} \leq \frac{V(0,x_0)}{t^2} \equiv \mathcal{O}(1/t^2).$$

Thus, accumulating the gradients in the iterative update of NAGD results in faster convergence.

3.4 Newton's Method (NM)

All the algorithms discussed so far have relied solely on information about the gradient of the objective function. To improve the convergence rate, a class of algorithms called Newton's and Quasi-Newton methods employs higher-order information. In this context, we will specifically focus on the classical Newton's Method (NM), which utilizes the true Hessian of the objective. The iteration rule for this method is given by

$$X_{k+1} = X_k - \alpha_k \nabla^2 f(X_k)^{-1}\nabla f(X_k).$$

Assuming that f is strictly convex, making $\nabla^2 f(\cdot)$ Positive Definite, the corresponding ODE for this update rule can be expressed as $\dot{x} = -\nabla^2 f(x)^{-1}\nabla f(x)$. Before delving further, let's derive an important property for this ODE, which will be useful later,

$$\begin{aligned}\frac{d}{dt}(\nabla f(x)) = \nabla^2 f(x)\dot{x} = -\nabla f(x),\\ \implies \nabla f(x(t)) = e^{-t}\nabla f(x_0) + \text{constant}.\end{aligned}$$

Now, consider a candidate Lyapunov function as $V(x) = \frac{1}{2}||\nabla f(x)||^2 + \frac{1}{2}\langle \nabla f(x), x\rangle$. It's easy to observe that because of the strict convexity of f, $V(x) \geq \frac{1}{2}||\nabla f(x)||^2 + \frac{1}{2}(f(x) - f(0)) \geq 0$. Taking the derivative of V, we get,

$$\begin{aligned}\dot{V}(x) &= \langle \nabla f(x), \nabla^2 f(x)\dot{x}\rangle + \langle x, \nabla^2 f(x)\dot{x}\rangle + \langle \nabla f(x), \dot{x}\rangle,\\ &= -||\nabla f(x)||^2 - \langle \nabla f(x), x\rangle - \langle \nabla f(x), \nabla^2 f(x)\nabla f(x)\rangle,\\ &\leq -||\nabla f(x)||^2 - (f(x) - f(0)) - \langle \nabla f(x), \nabla^2 f(x)\nabla f(x)\rangle < 0.\end{aligned}$$

Thus, V satisfies all the conditions of Theorem 1, making it a valid Lyapunov function, implying $x = 0$ is AS. Now, for the order of convergence,

$$
\begin{aligned}
f\Big(\frac{1}{t}\int_0^t x(\tau)d\tau\Big) - f(0) &\leq \frac{1}{t}\int_0^t f(x(\tau))d\tau - f(0), \\
&\leq \frac{1}{t}\int_0^t \Big\{ - ||\nabla f(x)||^2 - \langle \nabla f(x), \nabla^2 f(x) \nabla f(x)\rangle - \dot{V}(x(\tau))\Big\} d\tau, \\
&\leq -\frac{1}{t}\int_0^t ||\nabla f(x)||^2 d\tau + \frac{V(x_0)}{t}.
\end{aligned}
$$

In the last step, since the Hessian and V are always Positive Definite, we dropped the second term in the integral and $-V(x)/t$. Utilizing the property we derived earlier, we obtain,

$$
\begin{aligned}
f\left(\frac{1}{t}\int_0^t x(\tau)d\tau\right) - f(0) &\leq \frac{V(x_0)}{t} - \frac{||\nabla f(x_0)||^2}{t}\int_0^t e^{-2\tau}d\tau, \\
&\leq \frac{V(x_0)}{t} - \frac{||\nabla f(x_0)||^2}{t}(1 - e^{-2t}), \\
&\leq \frac{V(x_0)}{t} + \frac{||\nabla f(x_0)||^2}{t} e^{-2t}.
\end{aligned}
$$

We know that $e^t \geq 1 + t,\ \forall\ t$, thus $\frac{1}{1+2t} \geq e^{-2t}$. This implies that,

$$
f\left(\frac{1}{t}\int_0^t x(\tau)d\tau\right) - f(0) \leq \frac{V(x_0)}{t} + \frac{||\nabla f(x_0)||^2}{t + 2t^2} \equiv \mathcal{O}(1/t^2).
$$

Hence, with some mild assumptions on the Hessian of the objective, we have shown that the ODE of Newton's Method converges at the rate of $\mathcal{O}(1/t^2)$.

4 Conclusion

In conclusion, this study has delved into gradient-based unconstrained optimization algorithms, focusing mainly on formulating and analyzing the associated Ordinary Differential Equations (ODEs). While existing literature predominantly examines these algorithms in their discretized form, our work provides a self-contained re-derivation of the ODEs. This alternative viewpoint facilitates the application of Lyapunov-like arguments to evaluate the stability and convergence rates of equilibrium points corresponding to the stationary points of the optimization algorithms. Using the same Lyapunov arguments, we have also shown a new proof of convergence and its rate for the ODE of Newton's method.

Exploring ODEs offers a richer understanding of the dynamics underlying unconstrained optimization processes. By synthesizing and re-deriving critical findings from various works in this domain, our study contributes to the ongoing research on the theoretical analysis of optimization algorithms.

References

[1] Charles E. Blair. Problem complexity and method efficiency in optimization (a. s. nemirovsky and d. b. yudin). *Siam Review*, 27:264–265, 1985.

[2] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[3] Anthony M. Bloch. Hamiltonian and gradient flows, algorithms and control. 1995.

[4] A. A. Brown and Mike C. Bartholomew-Biggs. Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations. *Journal of Optimization Theory and Applications*, 62:211–224, 1989.

[5] Uwe Helmke and J. B. Moore. Optimization and dynamical systems. *Proceedings of the IEEE*, 84:907–, 1994.

[6] H.K. Khalil. *Nonlinear Systems*. Pearson Education. Prentice Hall, 2002.

[7] Weijie J. Su, Stephen P. Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17:153:1–153:43, 2014.