Abstract. In this paper, we study the convergence properties of an algorithm that can be viewed as an interpolation between two gradient based optimization methods, Nesterov’s acceleration method for strongly convex functions (NAG-SC) and Polyak’s heavy ball method. Recent Progress [11] has been made on using High-Resolution ordinary differential equations (ODEs) to distinguish these two fundamentally different methods. The key difference between them can be attributed to the gradient correction term, which is reflected by the Hessian term in the High-Resolution ODE. Our goal is to understand how this term can affect the convergence rate and the choice of our step size. To achieve this goal, we introduce the notion of $\beta$-High Resolution ODE, $0 \leq \beta \leq 1$ and prove that within certain range of step size, there is a phase transition happening at $\beta_c$. When $\beta_c \leq \beta \leq 1$, the algorithm associated with $\beta$-High Resolution ODE have the same convergence rate as NAG-SC. When $0 \leq \beta \leq \beta_c$, this algorithm will have the slower convergence rate than NAG-SC.

Contents

1. Introduction 1
1.1. Overview 1
1.2. Notation and Basic Setup 3
1.3. Statement of the Main Result 4
2. Derivation of $\beta$-High Resolution ODE 4
3. Global Existence and Uniqueness of ODE 5
4. Convergence Rate of Continuous ODE 8
5. Convergence Rate of discrete method 10
5.1. Discrete Energy Functional 10
5.2. Lemmata 11
5.3. Proof of Main Results 17
References 18

1. Introduction

1.1. Overview. In modern machine learning and (convex) optimization, we are interested in efficiently finding the minimizer of a smooth convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, i.e.

$$\min_{x \in \mathbb{R}^n} f(x)$$

There are several ways of solving this unconstrained optimization problem, among which the simplest and most straightforward method is gradient descent. For any

Key words and phrases. Optimization.
initial point $x_0 \in \mathbb{R}^n$, we update our $x_k$ by the following recursive rule,

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

(1.2)

where $s > 0$ is a fixed step size. Significant amount of work has been devoted to improve \[ \text{(1.2)} \] afterwards. Polyak in \[ \text{[2, 3]} \] introduced the following heavy-ball method. For any two initial points $x_0, x_1 \in \mathbb{R}^n$, we iteratively update our $x_k$ by

$$x_{k+1} = x_k + \alpha (x_k - x_{k-1}) - s\nabla f(x_k)$$

(1.3)

where $s > 0$ is the step size, $\alpha > 0$ is called the momentum coefficient. Heuristically, at each step, we accelerate the minimizing process by giving a momentum from the previous two steps. The main advantage of this method is the faster local convergence rate near the minimum of $f$.

It turns out that we can do better. Nesterov discovered the accelerated gradient method, see \[ \text{[4, 5]} \] for details. For (weakly) convex function $f$ (called NAG-C), NAG-C takes the form

$$y_{k+1} = x_k - s\nabla f(x_k)$$

$$x_{k+1} = y_{k+1} + \frac{k}{k+3} (y_{k+1} - y_k)$$

(1.4)

with $x_0 = y_0 \in \mathbb{R}^n$. For $\mu$-strongly convex and $L$-Lipschitz function $f$ (called NAG-SC), NAG-SC takes the following form

$$y_{k+1} = x_k - s\nabla f(x_k)$$

$$x_{k+1} = y_{k+1} + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} (y_{k+1} - y_k)$$

(1.5)

with $x_0 = y_0 \in \mathbb{R}^n$ as initial data points. (all the terms above will be defined in the next section) Plugging the $y_k$ and $y_{k+1}$ into the second line and we get

$$x_{k+1} = x_k + \left( \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \right) (x_k - x_{k-1}) - s\nabla f(x_k) - \left( \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \right) s(\nabla f(x_k) - \nabla f(x_{k-1}))$$

(1.6)

with $x_0$ and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1 + \sqrt{\mu s}}$. If we compare \[ \text{(1.6)} \] with \[ \text{(1.3)} \], \[ \text{(1.6)} \] is just the \[ \text{(1.3)} \] with momentum coefficient $\alpha = \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}}$ and an additional term

$$\left( \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \right) s(\nabla f(x_k) - \nabla f(x_{k-1}))$$

(1.7)

This term is called the gradient correction term. Mathematically, we want to understand why this term \[ \text{(1.7)} \] gives a faster convergence rate.

Recently, the work of B.Shi, S.Du, M.Jordan and W.Su \[ \text{[1]} \] provides an High-Resolution ODE approach to unravel the mystery of the gradient correction term. The crucial point in their approach is that when deriving the ODE, we take the step size $s$ small but non-vanishing. Here, we recall that High-Resolution ODE of heavy-ball method is

$$\dot{X}(t) + 2\sqrt{\mu} \dot{X}(t) + (1 + \sqrt{\mu s}) \nabla f(X(t)) = 0$$

(1.8)

and the High-Resolution ODE of NAG-SC is

$$\dot{X}(t) + 2\sqrt{\mu} \dot{X}(t) + \sqrt{s} \nabla^2 f(X(t)) \dot{X}(t) + (1 + \sqrt{\mu s}) \nabla f(X(t)) = 0$$

(1.9)
If we simply take the step size \( s \to 0 \), then both heavy ball method and NAG-SC will have the same limiting ODE (see [6] and [1] for a more detailed discussion)

\[
(1.10) \quad X(t) + 2\sqrt{\mu}X(t) + \nabla f(X(t)) = 0
\]

We can see that the only difference between (1.8) and (1.9) is the \( \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) \). In order to better understand how this term would make a difference on convergence rate and step size, we consider the so-called \( \beta \) High-Resolution ODE,

\[
(1.11) \quad \dot{X}(t) + 2\sqrt{\mu}X(t) + \beta \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s})\nabla f(X(t)) = 0, \quad 0 \leq \beta \leq 1
\]

Its corresponding discrete counterpart

\[
\begin{align*}
y_{k+1} &= x_k - s\nabla f(x_k) \\
y_{k+1}^\beta &= x_k - \beta s\nabla f(x_k) \\
x_{k+1} &= y_{k+1} + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} (y_{k+1}^\beta - y_k)
\end{align*}
\]

can be viewed as an interpolation between NAG-SC and heavy ball method. (see Section 2 for a detailed derivation)

The main objective of this paper is to understand the "cutoff" point of the convergence rate of this generalized class of algorithm when \( \beta \) continuously vary from 1 to 0. Suppose \( \beta \) is negligible, the Hessian term only contributes a little "acceleration". Hence it cannot achieve the same convergence rate as NAG-SC. Similarly, suppose \( \beta \) is very close to 1, it is essentially NAG-SC, which should give us a faster convergence rate than heavy ball method. To start, we first introduce some basic definitions.

1.2. Notation and Basic Setup. Let \( \mathcal{F}_L^1(\mathbb{R}^n) \) denote the class of L-smooth convex functions defined on \( \mathbb{R}^n \), that is, \( f \in \mathcal{F}_L^1 \) if \( f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle \) for all \( x, y \in \mathbb{R}^n \). Its gradient is L-Lipschitz continuous in the sense that

\[
\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|
\]

where \( \| \cdot \| \) denotes standard Euclidean norm and \( L > 0 \) is the Lipschitz constant. The function class \( \mathcal{F}_L^2(\mathbb{R}^n) \) denotes the subclass of \( \mathcal{F}_L^1(\mathbb{R}^n) \) such that each \( f \) has a Lipschitz continuous Hessian in the sense that

\[
\left\| \nabla^2 f(x) - \nabla^2 f(y) \right\|_F \leq L'\|x - y\|
\]

where \( \| \cdot \|_F \) denotes the Frobenius norm and \( L' > 0 \) is an arbitrary constant. For \( p = 1, 2 \), let \( \mathcal{S}_{\mu L}^p(\mathbb{R}^n) \) denote the subclass of \( \mathcal{F}_L^p(\mathbb{R}^n) \) such that each member \( f \) is \( \mu \)-strongly convex for some \( 0 < \mu \leq L \). That is, \( f \in \mathcal{S}_{\mu L}^p(\mathbb{R}^n) \) if \( f \in \mathcal{F}_L^p(\mathbb{R}^n) \) and

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

for all \( x, y \in \mathbb{R}^n \). This is equivalent to the convexity of \( f(x) - \frac{\mu}{2} \|x - x^*\|^2 \), where \( x^* \) is the minimizer of the objective \( f \). Now, we are ready to state the main result.
1.3. Statement of the Main Result.

**Theorem 1.1.** Let $f \in S^1_{\mu, L}([R^n])$. If the step size $s$ satisfies \( \frac{25\mu}{(12L-\mu)^2} \leq s = \frac{1}{cL} \leq \frac{1}{12L} \) (equivalently, $4 \leq c \leq \frac{(12L-\mu)^2}{25\mu}$, $c$ may possibly depend on $\mu, L$), then there exists a $\beta_c = \beta_c(\mu, L, s) \in (0, 1)$ such that when $0 \leq \beta \leq \beta_c$,

\[
\begin{align*}
\beta_c 	ext{ is computed explicitly in Remark 5.7.}
\end{align*}
\]

**Remark 1.2** (Comparison with the known results). In [1], Theorem 3, when $s = \frac{1}{12L}$, NAG-SC ($\beta = 1$) gives us a monotone convergence rate of

\[
\begin{align*}
f(x_k) - f(x^*) \leq O \left( \frac{L \cdot \|x_0 - x^*\|^2}{1 + \frac{\beta}{\sqrt{\mu/L}}} \right)
\end{align*}
\]

In our $\beta$-High Resolution Approach, assume $s \propto \frac{1}{L}$, we can see that as $\beta$ decreases from 1 to 0, after passing the critical value $\beta_c$, the convergence rate cannot match the (1.14) anymore (It slows down). Instead, the denominator is a rational function of $\sqrt{\mu/L}$ as in (1.12).

2. Derivation of $\beta$-High Resolution ODE

For variable $\beta \in [0, 1]$, define the $\beta$ generalized NAG-SC method to be

\[
\begin{align*}
\mathbf{y}_{k+1} &= x_k - s \nabla f(x_k) \\
\mathbf{y}^\beta_{k+1} &= x_k - \beta s \nabla f(x_k) \\
x_{k+1} &= \mathbf{y}_{k+1} + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} (\mathbf{y}^\beta_{k+1} - \mathbf{y}_{k})
\end{align*}
\]
with initial condition $x_0 \in \mathbb{R}^n$ and $y_0^\beta = \frac{(1-\sqrt{\mu s})x_0 - s\nabla f(x_0)[1-\sqrt{\mu s}]\beta + \sqrt{\mu s} - 1}{1-\sqrt{\mu s}}$. This is equivalent to
\begin{equation}
2.2 \quad x_{k+1} = x_k + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}}(x_k - x_{k-1}) - s\nabla f(x_k) - \beta \cdot \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \cdot s(\nabla f(x_k) - \nabla f(x_{k-1}))
\end{equation}
with initial condition $x_0$ and $x_1 = x_0 - \frac{2s \nabla f(x_0)}{1 + \sqrt{\mu s}}$. Fix a nonnegative integer $k$ and let $t_k = k\sqrt{s}$ and $x_k = X(t_k)$ for some $C^\infty$ curve. Using Taylor expansion with respect to $\sqrt{s}$, we get
\begin{equation}
2.3 \quad x_{k+1} = X(t_{k+1}) = X(t_k) + \dot{X}(t_k)\sqrt{s} + \frac{1}{2} \ddot{X}(t_k)(\sqrt{s})^2 + \frac{1}{6} \dot{X}(t_k)(\sqrt{s})^3 + O((\sqrt{s})^4)
\end{equation}
\begin{equation}
2.4 \quad x_{k-1} = X(t_{k-1}) = X(t_k) - \dot{X}(t_k)\sqrt{s} + \frac{1}{2} \ddot{X}(t_k)(\sqrt{s})^2 - \frac{1}{6} \dot{X}(t_k)(\sqrt{s})^3 + O((\sqrt{s})^4)
\end{equation}
Applying Taylor expansion again to the gradient correction gives us
\begin{equation}
2.5 \quad \nabla f(x_k) - \nabla f(x_{k-1}) = \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O((\sqrt{s})^2)
\end{equation}
Multiplying both sides of (2.2) by $\frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \cdot \frac{1}{\sqrt{s}}$ and rearranging the terms,
\begin{equation}
2.6 \quad \frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu s}}{1 - \sqrt{\mu s}} \frac{x_{k+1} - x_k}{s} + \beta (\nabla f(x_k) - \nabla f(x_{k-1})) + \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \nabla f(x_k) = 0
\end{equation}
Plugging (2.3), (2.4) and (2.5) into (2.6), we have
\begin{equation}
\ddot{X}(t_k) + O((\sqrt{s})^2) + \frac{2\sqrt{\mu}}{1 - \sqrt{\mu s}} \left[ X(t_k) + \frac{1}{2} \dot{X}(t_k)\sqrt{s} + O((\sqrt{s})^2) \right] \\
+ \beta \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O((\sqrt{s})^2) + \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \nabla f(X(t_k)) = 0
\end{equation}
After rearranging,
\begin{equation}
\ddot{X}(t_k) + \frac{2\sqrt{\mu}}{1 - \sqrt{\mu s}} \ddot{X}(t_k) + \beta \sqrt{s} \nabla^2 f(X(t_k))\dot{X}(t_k) + \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \nabla f(X(t_k)) + O(s) = 0
\end{equation}
Multiplying both sides by $1 - \sqrt{\mu s}$ and by ignoring any $O(s)$ terms but keep $O(\sqrt{s})$ terms, we finally get the $\beta$-High Resolution ODE,
\begin{equation}
2.7 \quad \ddot{X}(t) + 2\sqrt{\mu} \dot{X}(t) + \beta \sqrt{s} \nabla^2 f(X(t))\dot{X}(t) + (1 + \sqrt{\mu s}) \nabla f(X(t)) = 0
\end{equation}
with $0 \leq \beta \leq 1$. The initial conditions of (2.7) throughout this paper are assumed to be $X(0) = x_0$ and $X(0) = \frac{2\sqrt{\mu s} f(x_0)}{1 + \sqrt{\mu s}}$.

3. Global Existence and Uniqueness of ODE
Suppose $X_\beta(t)$ is the solution of (2.7), then by the following Lyapunov function
\begin{equation}
3.1 \quad \mathcal{E}(t) = (1 + \sqrt{\mu s}) (f(X_\beta) - f(x^*)) + \frac{1}{2} \| \dot{X}_\beta \|^2
\end{equation}
we can deduce that there exists some $C_1 > 0$ such that
\[
\sup_{0 \leq t < \infty} \|X_s(t)\| \leq C_1
\]
Now, we investigate the global existence and uniqueness of the $\beta$-High Resolution ODE (2.7). Recall that the initial value problem (IVP) for first-order ODE system in $\mathbb{R}^m$ is
\[
\dot{x} = b(x), \quad x(0) = x_0
\]
and the following theorem deals with the global existence and uniqueness of (3.2)

**Theorem 3.1** (Chillingworth [7], Chapter 3.1, Theorem 4). Let $M \in \mathbb{R}^m$ be a compact manifold and $b \in C^1(M)$. If the vector fields $b$ satisfies the global Lipschitz condition
\[
\|b(x) - b(y)\| \leq \mathcal{L}\|x - y\| \quad \text{for all } x, y \in M.
\]
Then for any $x_0 \in M$, the IVP (3.2) has a unique solution $x(t)$ defined for all $t \in \mathbb{R}$.

**Theorem 3.2.** For any $f \in S^2_\mu(\mathbb{R}^n) := \bigcup_{L \geq \mu} S^2_{\mu L}(\mathbb{R}^n)$, the $\beta$-High Resolution ODE (2.7) with the specified initial conditions has a unique global solution $X \in C^2(I; \mathbb{R}^n)$.

**Proof.** Notice that
\[
M_{\epsilon_1} := \left\{ (X_s, \dot{X}_s) \in \mathbb{R}^{2n} : \|X_s\| \leq \epsilon_1 \right\}
\]
is a compact manifold. The phase-space representation for (2.7) is
\[
\frac{d}{dt} \begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} = \begin{pmatrix} -2\sqrt{\mu}X_s - \beta \sqrt{\frac{\mu}{s}} \nabla^2 f(X_s) \dot{X}_s - (1 + \sqrt{\mu s}) \nabla f(X_s) \\ \dot{X}_s - \dot{Y}_s \end{pmatrix}
\]
Now, for any $(X_s, \dot{X}_s)^T, (Y_s, \dot{Y}_s)^T \in M_{\epsilon_1},$
\[
\left\| \frac{d}{dt} \begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} - \frac{d}{dt} \begin{pmatrix} Y_s \\ \dot{Y}_s \end{pmatrix} \right\| \leq \left\| \begin{pmatrix} -2\sqrt{\mu}X_s - \beta \sqrt{\frac{\mu}{s}} \nabla^2 f(X_s) \dot{X}_s - (1 + \sqrt{\mu s}) \nabla f(X_s) \\ \dot{X}_s - \dot{Y}_s \end{pmatrix} \right\|
\]
\[
+ \beta \sqrt{s} \left\| \begin{pmatrix} 0 \\ (\nabla^2 f(X_s) - \nabla^2 f(Y_s)) \dot{Y}_s \end{pmatrix} \right\|
\]
\[
+ (1 + \sqrt{s}) \left\| \begin{pmatrix} 0 \\ \nabla f(X_s) - \nabla f(Y_s) \end{pmatrix} \right\|
\]
\[
\leq \sqrt{1 + 8\mu + 2\beta^2 s L^2} \|X_s - Y_s\| + \left[ \beta \sqrt{s} \epsilon_1 + (1 + \sqrt{s}) L \right] \|X_s - Y_s\|
\]
\[
\leq 2 \max \left\{ \sqrt{1 + 8\mu + 2\beta^2 s L^2}, \beta \sqrt{s} \epsilon_1 + (1 + \sqrt{s}) L \right\} \left\| \begin{pmatrix} X_s \\ \dot{X}_s \end{pmatrix} - \begin{pmatrix} Y_s \\ \dot{Y}_s \end{pmatrix} \right\|
\]
Hence, based on the above calculation and the the phase space representation (3.3), we get the desired results. \qed

Here we quickly remark that the low resolution counterparts of this $\beta$-High Resolution ODE is the same as both of the heavy-ball method and NAG-SC, which is
\[
\ddot{X}(t) + 2\sqrt{\mu} \dot{X}(t) + \nabla f(X(t)) = 0
\]
Based on the Lyapunov function (3.1), the gradient norm is also bounded, i.e.
\[
\sup_{0 \leq t < \infty} \|\nabla f(X_s(t))\| \leq C_2
\]
For the low resolution ODE \((3.4)\), it has phase representation
\[
\frac{d}{dt} \left( \begin{array}{c} X \\ \dot{X} \end{array} \right) = \left( \begin{array}{c} \dot{X} \\ -2\sqrt{\beta}X - \nabla f(X) \end{array} \right)
\]
and again by Lyapunov function, the solution \(X = X(t)\) of \((3.4)\) is bounded, i.e.
\[
\sup_{0 \leq t < \infty} \|\dot{X}(t)\| \leq \mathcal{C}_3
\]
It is easy to see that we can find a constant \(\mathcal{L}_1\) such that
\[
\| \left( \begin{array}{c} \dot{X} \\ -2\sqrt{\beta}X - \nabla f(X) \end{array} \right) - \left( \begin{array}{c} \dot{Y} \\ -2\sqrt{\beta}Y - \nabla f(Y) \end{array} \right) \| \leq \mathcal{L}_1 \| X - Y \|
\]
Now, we study the approximation. We first introduce several lemmas.

**Lemma 3.3** (Gronwall’s Lemma). Let \(m(t), t \in [0, T]\), be a nonnegative function with the following relation,
\[
m(t) \leq C + \alpha \int_0^t m(s) ds
\]
with \(C, \alpha > 0\). Then we have
\[
m(t) \leq Ce^{\alpha t}
\]
**Proof.** Trivially by calculus. \(\square\)

**Lemma 3.4.** Let \(X_s(t)\) and \(X(t)\) be the solutions of \(\beta\)-High Resolution ODE \((2.7)\) and Low Resolution Counterpart \((3.4)\), respectively. Then
\[
\lim_{s \to 0} \max_{0 \leq t \leq T} \|X_s(t) - X(t)\| = 0
\]
**Proof.** By \((3.3)\) and \((3.5)\),
\[
\frac{d}{dt} \left( \begin{array}{c} X_s - X \\ \dot{X}_s - \dot{X} \end{array} \right) = \left( \begin{array}{c} \dot{X}_s - \dot{X} \\ -2\sqrt{\beta}(X_s - X) - \nabla f(X_s) - \nabla f(X) \end{array} \right) - \sqrt{s} \left( \frac{\beta}{2}(X_s)X_s + \sqrt{\beta}f(X_s) \right)
\]
Then, we have
\[
\|X_s(t) - X(t)\|^2 + \|\dot{X}_s(t) - \dot{X}(t)\|^2
\]
\[
= 2 \int_0^t \left( X_s(u) - X(u) \right) \cdot \left( X_s(u) - X(u) \right) du + \|X_s(0) - X(0)\|^2 + \|\dot{X}_s(0) - \dot{X}(0)\|^2
\]
\[
\leq 2\mathcal{L}_1 \int_0^t \|X_s(u) - X(u)\|^2 + \|\dot{X}_s(u) - \dot{X}(u)\|^2 du
\]
\[
+ \left( \mathcal{C}_1 + \mathcal{C}_3 \right) \left( \beta L \mathcal{C}_1 + \mathcal{C}_2 \sqrt{\beta} \right) + \frac{4\sqrt{s}}{\left( 1 + \sqrt{\beta s} \right)^2} \left\| \nabla f(x_0) \right\|^2 \sqrt{s}
\]
\[
\leq 2\mathcal{L}_1 \int_0^t \|X_s(u) - X(u)\|^2 + \|\dot{X}_s(u) - \dot{X}(u)\|^2 du + \mathcal{C}_5 \sqrt{s}
\]
By Lemma \((3.3)\), we have that
\[
\|X_s(t) - X(t)\|^2 + \|\dot{X}_s(t) - \dot{X}(t)\|^2 \leq \mathcal{C}_5 \sqrt{s} \exp(2\mathcal{L}_1) t
\]
This completes the proof. \(\square\)
Lemma 3.5. The discrete method of $\beta$-High Resolution ODE converges to their low-resolution ODE in the sense that

$$\lim_{s \to 0} \max_{0 \leq k \leq \frac{T}{s}} \|x_k - X(k\sqrt{s})\| = 0$$

Proof. The proof of this Lemma follows closely from the method used in [8] and [6]. Here we do not go into any details. 

Proposition 3.6. For any $f \in S^{2}_{\mu_L}(\mathbb{R}^n) := \cup_{L \geq \mu} S^{2}_{\mu_L}(\mathbb{R}^n)$, the $\beta$-High Resolution ODE (2.7) with the specified initial conditions has a unique global solution $X \in C^{2}([0, \infty); \mathbb{R}^n)$. Moreover, the discretized method converges to the $\beta$-High Resolution ODE in the sense that

$$\limsup_{s \to 0} \max_{0 \leq k \leq \frac{T}{\sqrt{s}}} \|x_k - X(k\sqrt{s})\| = 0$$

for any fixed $T > 0$.

Proof. This result follows from the Lemma 3.3, Lemma 3.4 and Lemma 3.5.

4. Convergence Rate of Continuous ODE

In this section, we prove the following theorem.

Theorem 4.1. Let $f \in S^{2}_{\mu_L}(\mathbb{R}^n)$. Then for any step size $0 \leq s \leq 1/L$, the solution $X = X(t)$ of the $\beta$-High Resolution ODE (2.7) satisfies

$$f(X(t)) - f(x^*) \leq \frac{3 + (2 - \beta)^2}{2s} \|x_0 - x^*\|^2 \cdot e^{-\sqrt{\mu}t}$$

We first define the Energy Functional $E_{\beta}(t)$ of $\beta$-High Resolution ODE as the following:

(4.1)

$$E_{\beta}(t) := (1 + \sqrt{\mu s})(f(X) - f(x^*)) + \frac{1}{4} \|\dot{X}\|^2 + \frac{1}{4} \|X + 2\sqrt{\mu}(X - x^*) + \beta \sqrt{s} \nabla f(X)\|^2$$

The next lemma is of key importance to us.

Lemma 4.2. For any step size $s > 0$, the energy functional (4.1) with $X = X(t)$ being the our solution to the $\beta$-High Resolution ODE satisfies

(4.2)

$$\frac{dE_{\beta}(t)}{dt} \leq -\frac{\sqrt{\mu}}{4} E_{\beta}(t) - \frac{1}{4} \left( 8\beta s \sqrt{\mu - 3s} \beta^2 \sqrt{\mu} \right) \|\nabla f(X)\|^2 + 2\sqrt{\mu} \|\dot{X}\|^2 + (\sqrt{\mu + \mu \sqrt{s}})(f(X) - f(x^*))$$

In particular,

(4.3)

$$\frac{dE_{\beta}(t)}{dt} \leq -\frac{\sqrt{\mu}}{4} E_{\beta}(t)$$
Proof. The energy functional $[4.1]$ together with $[2.7]$ give us

\[
\frac{d\mathcal{E}_\beta(t)}{dt} = (1 + \sqrt{\mu s}) \langle \nabla f(X), \dot{X} \rangle + \frac{1}{2} \left\langle \dot{X}, -2\sqrt{\mu} \dot{X} - \beta \sqrt{s} \nabla^2 f(X) \dot{X} - (1 + \sqrt{\mu s}) \nabla f(X) \right\rangle \\
+ \frac{1}{2} \left\langle \dot{X} + 2\sqrt{\mu} (X - x^*) + \beta \sqrt{s} \nabla f(X), -(1 + \sqrt{\mu s}) \nabla f(X) \right\rangle \\
= -\sqrt{\mu} \left( \|\dot{X}\|^2 + (1 + \sqrt{\mu s}) \langle \nabla f(X), X - x^* \rangle + \frac{\beta s}{2} \|\nabla f(X)\|^2 \right) \\
- \frac{\beta \sqrt{s}}{2} \left( \|\nabla f(X)\|^2 + \dot{X}^T \nabla^2 f(X) \dot{X} \right) \\
\leq -\sqrt{\mu} \left( \|\dot{X}\|^2 + (1 + \sqrt{\mu s}) \langle \nabla f(X), X - x^* \rangle + \frac{\beta s}{2} \|\nabla f(X)\|^2 \right)
\]

Also, by $\mu$-strong convexity of $f$,

\[
\langle \nabla f(X), X - x^* \rangle \geq \begin{cases} 
    f(X) - f(x^*) + \frac{\mu}{2} \|X - x^*\|^2 \\
    \mu \|X - x^*\|^2 
\end{cases}
\]

This gives us

\[
(1 + \sqrt{\mu s}) \langle \nabla f(X), X - x^* \rangle \geq \frac{1 + \sqrt{\mu s}}{2} \langle \nabla f(X), X - x^* \rangle + \frac{1}{2} \langle \nabla f(X), X - x^* \rangle \\
\geq \frac{1 + \sqrt{\mu s}}{2} \left( f(X) - f(x^*) + \frac{\mu}{2} \|X - x^*\|^2 \right) + \frac{\mu}{2} \|X - x^*\|^2 \\
\geq \frac{1 + \sqrt{\mu s}}{2} \left( f(X) - f(x^*) \right) + \frac{3\mu}{4} \|X - x^*\|^2
\]

Hence, the derivative of Energy Functional can be bounded by (4.4)

\[
\frac{d\mathcal{E}_\beta(t)}{dt} \leq -\sqrt{\mu} \left( \frac{1 + \sqrt{\mu s}}{2} (f(X) - f(x^*)) + |\dot{X}|^2 + \frac{3\mu}{4} \|X - x^*\|^2 + \frac{\beta s}{2} \|\nabla f(X)\|^2 \right)
\]

Next, by Cauchy-Schwarz inequality,

\[
\|2\sqrt{\mu} (X - x^*) + \dot{X} + \beta \sqrt{s} \nabla f(X)\|^2 \leq 3 \left( 4\mu \|X - x^*\|^2 + |\dot{X}|^2 + \beta^2 s \|\nabla f(X)\|^2 \right)
\]

from which we can deduce that

\[
4.5 \quad \mathcal{E}_\beta(t) \leq (1 + \sqrt{\mu s}) (f(X) - f(x^*)) + |\dot{X}|^2 + 3\mu \|X - x^*\|^2 + \frac{3\beta^2}{4} \|\nabla f(X)\|^2
\]

Finally, combining (4.4) and (4.5) and we get the (4.2). The (4.3) holds since $\Delta_\beta \geq 0$. (Notice that $0 \leq \beta \leq 1$ and $x^*$ is the minimizer) 

Proof of Theorem 4.1. By previous lemma,

\[
\dot{\mathcal{E}}_\beta(t) \leq -\frac{\sqrt{\mu}}{4} \mathcal{E}_\beta(t) \implies \frac{d}{dt} \left( \mathcal{E}_\beta(t) e^{\frac{\sqrt{\mu}}{4} t} \right) \leq 0 \implies \mathcal{E}_\beta(t) \leq e^{-\frac{\sqrt{\mu}}{4} t} \mathcal{E}_\beta(0)
\]
Noticing the initial condition \( X(0) = x_0 \) and \( \dot{X}(0) = -\frac{2\sqrt{sv(x_0)}}{1+\sqrt{\mu s}} \), we get

\[
\begin{align*}
f(X) - f(x^*) &\leq e^{-\frac{\mu t}{4}} \left[ f(x_0) - f(x^*) + \frac{s}{(1 + \sqrt{\mu s})^3} \|\nabla f(x_0)\|^2 \right. \\
&\quad + \left. \frac{1}{4(1 + \sqrt{\mu s})} \right] \frac{2\sqrt{\mu s}}{1 + \sqrt{\mu s}} \cdot \frac{2 - \beta - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \cdot \sqrt{s \nabla f(x_0)} \|\|^{2}\right] \\
\end{align*}
\]

Since \( f \in S^2_{\mu, L} \),

\[
\|\nabla f(x_0)\| \leq L \|x_0 - x^*\| \quad \text{and} \quad f(x_0) - f(x^*) \leq \frac{L}{2} \|x_0 - x^*\|^2
\]

Together with Cauchy-Schwartz inequality,

\[
\begin{align*}
f(X) - f(x^*) &\leq \left[ f(x_0) - f(x^*) + \frac{2 + (2 - \beta - \sqrt{\mu s})^2}{2(1 + \sqrt{\mu s})^3} \cdot s \|\nabla f(x_0)\|^2 + \frac{2\mu}{1 + \sqrt{\mu s}} \|x_0 - x^*\|^2 \right] e^{-\frac{\mu t}{4}} \\
&\leq \left[ \frac{L}{2} + \frac{2 + (2 - \beta - \sqrt{\mu s})^2}{2(1 + \sqrt{\mu s})^3} \cdot s \right] L^2 + \frac{2\mu}{1 + \sqrt{\mu s}} \|x_0 - x^*\|^2 e^{-\frac{\mu t}{4}} \\
&\leq \left[ \frac{1}{2} + \frac{2 + (2 - \beta - \sqrt{\mu s})^2}{2(1 + \sqrt{\mu s})^3} + \frac{2\mu s}{1 + \sqrt{\mu s}} \right] \cdot \frac{1}{s} \|x_0 - x^*\|^2 e^{-\frac{\mu t}{4}}
\end{align*}
\]

Now, by a little bit of analysis, under the assumption \( \mu s \leq \mu / L \leq 1 \),

\[
\frac{1}{2} + \frac{2 + (2 - \beta - \sqrt{\mu s})^2}{2(1 + \sqrt{\mu s})^3} + \frac{2\mu s}{1 + \sqrt{\mu s}} \leq \frac{3 + (2 - \beta)^2}{2}
\]

This completes the proof of the Theorem. \( \square \)

5. Convergence Rate of discrete method

5.1. Discrete Energy Functional. We first write the (2.2) as

\[
\begin{align*}
x_k - x_{k-1} &= \sqrt{sv_{k-1}} \\
(5.1) \quad v_k - v_{k-1} &= -\frac{2\sqrt{\mu s}}{1 - \sqrt{\mu s}} v_k - \beta \sqrt{s} \left( \nabla f(x_k) - \nabla f(x_{k-1}) \right) - \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \cdot \sqrt{s} \nabla f(x_k)
\end{align*}
\]

in the position variable \( x_k \) and the velocity variable \( v_k \) that is defined as

\[
v_k = \frac{x_{k+1} - x_k}{\sqrt{s}}
\]

The initial velocity is

\[
v_0 = -\frac{2\sqrt{s}}{1 + \sqrt{\mu s}} \nabla f(x_0)
\]
Next, we construct the $\beta$ discrete-time energy functional

$$
\mathcal{E}_B(k) = \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_k) - f(x^*)) + \frac{1}{4} \|v_k\|^2 + \frac{1}{4} \left\| v_k + \frac{2 \sqrt{\mu} s}{1 - \sqrt{\mu s}} (x_k - x^*) + \beta \sqrt{s} \nabla f(x_k) \right\|^2
$$

(5.2)

- $\beta s \|\nabla f(x_k)\|^2$
- $\frac{3}{2(1 - \sqrt{\mu s})}$

negative term

5.2. Lemmata.

Lemma 5.1. For $f \in S^1_{\mu,L}(\mathbb{R}^n)$,

$$
\mathcal{E}_B(k) \leq \left( \frac{1}{1 - \sqrt{\mu s}} + \frac{\beta^2 L s}{2} \right) (f(x_k) - f(x^*)) + \frac{1 + \sqrt{\mu s} + \mu s}{(1 - \sqrt{\mu s})^2} \|v_k\|^2
$$

$$
+ \frac{3\mu}{(1 - \sqrt{\mu s})^2} \|x_k - x^*\|^2 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \left[ f(x_k) - f(x^*) - \left( \frac{\beta s \sqrt{\mu s} - (\beta^2 - \beta) s}{2 \sqrt{\mu s}} \right) \|\nabla f(x_k)\|^2 \right]
$$

Proof. In the definition of $\beta$ discrete-time energy functional (5.2), by the Cauchy-Schwarz inequality, we have

$$
\text{III} = \frac{1}{4} \left\| v_k + \frac{2 \sqrt{\mu} s}{1 - \sqrt{\mu s}} (x_k - x^*) + \beta \sqrt{s} \nabla f(x_k) \right\|^2
$$

$$
\leq \frac{3}{4} \left[ \left( \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right)^2 \|v_k\|^2 + \frac{4\mu}{(1 - \sqrt{\mu s})^2} \|x_k - x^*\|^2 + \beta^2 s \|\nabla f(x_k)\|^2 \right]
$$

Notice that $\|\nabla f(x_k)\|^2 \leq 2L (f(x_k) - f(x^*))$,

$$
\frac{3\beta^2 s}{4} \|\nabla f(x_k)\|^2 - \frac{\beta s |\nabla f(x_k)|^2}{2(1 - \sqrt{\mu s})} = \frac{\beta^2 s}{4} \|\nabla f(x_k)\|^2 + \frac{\beta^2 s}{2} \|\nabla f(x_k)\|^2 - \frac{\beta s |\nabla f(x_k)|^2}{2(1 - \sqrt{\mu s})}
$$

$$
\leq \frac{\beta^2 L s}{2} (f(x_k) - f(x^*)) - \frac{\beta^2 s \sqrt{\mu s} - (\beta^2 - \beta) s}{2 (1 - \sqrt{\mu s})} \cdot \|\nabla f(x_k)\|^2
$$

for $f \in S^1_{\mu,L}(\mathbb{R}^n)$, which gives us the following estimate,

$$
\mathcal{E}_B(k) \leq \left( \frac{1}{1 - \sqrt{\mu s}} + \frac{\beta^2 L s}{2} \right) (f(x_k) - f(x^*)) + \frac{1 + \sqrt{\mu s} + \mu s}{(1 - \sqrt{\mu s})^2} \|v_k\|^2
$$

$$
+ \frac{3\mu}{(1 - \sqrt{\mu s})^2} \|x_k - x^*\|^2 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \left[ f(x_k) - f(x^*) - \left( \frac{\beta s \sqrt{\mu s} - (\beta^2 - \beta) s}{2 \sqrt{\mu s}} \right) \|\nabla f(x_k)\|^2 \right]
$$

$\square$
Lemma 5.2. For $f \in \mathcal{S}^1_{\mu, L} (\mathbb{R}^n)$,
\[
\mathcal{E}_\beta (k+1) - \mathcal{E}_\beta (k) \leq - \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \left( 1 + \sqrt{\mu s} \right) \left( \nabla f(x_{k+1})^T (x_{k+1} - x^*) + \| v_{k+1} \|^2 \right) \\
+ \frac{1}{2} \left( 1 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) \cdot s \cdot \frac{2 \mu s - (1 + \beta)\sqrt{s} + (1 - \beta)}{1 - \sqrt{\mu s}} \cdot \| \nabla f(x_{k+1}) \|^2 \\
- \frac{1}{2L} \left( \beta \sqrt{s} + 1 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) \| \nabla f(x_{k+1}) - \nabla f(x_k) \|^2 \\
+ \frac{\beta s}{2} \left( 1 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) \| \nabla f(x_{k+1}) - \nabla f(x_k) \|^2
\]

Proof. The proof of this Lemma is only a slight variation of argument in [1], Appendix B.2.2 so here we only give the first several steps in order to illustrate the difference. Recall the $\beta$ discrete time energy functional (5.2).

\[
\mathcal{E}_\beta (k) = \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_k) - f(x^*)) + \frac{1}{4} \| v_k \|^2 + \frac{1}{4} \left\| v_{k+1} + \frac{2 \sqrt{\mu s}}{1 - \sqrt{\mu s}} (x_k - x^*) + \beta \sqrt{s} \nabla f(x_k) \right\|^2 \\
- \frac{\beta s \| \nabla f(x_k) \|^2}{2(1 - \sqrt{\mu s})}
\]

Let $\Delta_I, \Delta_{II}$ and $\Delta_{III}$ be the difference between I, II and III respectively. For the first part, same as in [1], Appendix B.2.2

\[
\Delta_I = \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_{k+1}) - f(x^*)) - \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_k) - f(x^*)) \\
\leq \left( \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) \sqrt{s} (\nabla f(x_{k+1}), v_k) + \frac{1}{2} \left( \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) \| \nabla f(x_{k+1}) - \nabla f(x_k) \|^2
\]

For the second part, by using (5.1),

\[
\Delta_{II} = \left( \frac{1}{4} \| v_{k+1} \|^2 - \frac{1}{4} \| v_k \|^2 \right) \\
= \frac{1}{2} \langle v_{k+1} - v_k, v_{k+1} \rangle - \frac{1}{4} \| v_{k+1} - v_k \|^2 \\
= - \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \| v_{k+1} \|^2 - \frac{\beta \sqrt{s}}{2} \langle \nabla f(x_{k+1}) - \nabla f(x_k), v_{k+1} \rangle \\
- \frac{1 + \sqrt{\mu s}}{2 \sqrt{s}} \langle \nabla f(x_{k+1}), v_{k+1} \rangle - \frac{1}{4} \| v_{k+1} - v_k \|^2 \\
= - \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \| v_{k+1} \|^2 - \frac{\beta \sqrt{s}}{2} \cdot \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \langle \nabla f(x_{k+1}) - \nabla f(x_k), v_{k+1} \rangle \\
+ \frac{\beta \sqrt{s}}{2} \| \nabla f(x_{k+1}) - \nabla f(x_k) \|^2 + \frac{\beta \sqrt{s}}{2} \langle \nabla f(x_{k+1}) - \nabla f(x_k), \nabla f(x_{k+1}) \rangle \\
- \frac{1 + \sqrt{\mu s}}{2 \sqrt{s}} \langle \nabla f(x_{k+1}), v_{k+1} \rangle - \frac{1}{4} \| v_{k+1} - v_k \|^2
\]
For the third part,

\[
\Delta_{\text{III}} \\
= \frac{1}{4} \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 - \frac{1}{4} \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2 \\
= \frac{1}{2} \left( 1 + \frac{2\sqrt{s}}{1 - \sqrt{\mu}} \right) \left( \beta - \frac{1}{2} \right) \left( \frac{1 + \sqrt{\mu}}{1 - \sqrt{\mu}} \right) \left( \frac{1 + \sqrt{\mu}}{1 - \sqrt{\mu}} \right) \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2
\]

The rest of the argument on estimating the difference \(E_{\beta}(k + 1) - E_{\beta}(k)\) follows the same method as in [1] so here we do not go into further details.

**Remark 5.3.** Notice that for the last two terms above,

\[
\left[ \frac{\sqrt{s}}{2} \left( \frac{1 + \sqrt{\mu}}{1 - \sqrt{\mu}} \right) + \frac{1}{2} \frac{\beta s}{L} \left( \frac{1 + \sqrt{\mu}}{1 - \sqrt{\mu}} \right) \right] \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2
\]

Therefore, under the assumption that \(s \leq \frac{1}{t}\),

\[
E_{\beta}(k + 1) - E_{\beta}(k) \leq \frac{1}{2} \left( \frac{1 + \sqrt{\mu}}{1 - \sqrt{\mu}} \right) \left( \beta - \frac{1}{2} \right) \left( 1 + \sqrt{\mu} \right) \left( 1 + \sqrt{\mu} \right) \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2
\]

**Corollary 5.4.** If \(s \leq \frac{1}{2t} \leq \frac{1}{t}\) and \(f \in \mathcal{B}_{\mu,1}(\mathbb{R}^n)\), then we have

\[
E_{\beta}(k + 1) - E_{\beta}(k) \leq \frac{1}{2} \left( \frac{1 + \sqrt{\mu}}{1 - \sqrt{\mu}} \right) \left( \beta - \frac{1}{2} \right) \left( 1 + \sqrt{\mu} \right) \left( 1 + \sqrt{\mu} \right) \left\| \nabla f(x_{k+1}) - \nabla f(x_k) \right\|^2
\]
Lemma 5.5. Let \( f \in S^1_{\mu,E}(\mathbb{R}^n) \), together with the inequality

\[
\begin{align*}
    f(x^*) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^* - x_{k+1} \rangle + \frac{1}{1/L} \| \nabla f(x_{k+1}) \|^2 \\
    1/L &\geq 1/2L 
\end{align*}
\]

We have that

\[
\begin{align*}
    \varepsilon_{\beta}(k+1) - \varepsilon_{\beta}(k) &\leq -\frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \left[ f(x_{k+1}) - f(x^*) + \frac{1}{2L} \frac{1}{1 - \sqrt{\mu s}} \| \nabla f(x_{k+1}) \|^2 \right] \\
    &\quad - \sqrt{\mu s} \left[ \frac{1}{(1 - \sqrt{\mu s})^2} \| x_{k+1} - x^* \|^2 + \frac{1}{2(1 - \sqrt{\mu s})} \| \nu_{k+1} \|^2 \right] \\
    &\quad - \sqrt{\mu s} \left[ \frac{1}{(1 - \sqrt{\mu s})^2} \left( 1 - 2Ls \cdot \frac{(\beta - \beta^2)\mu s + (3 + \beta^2 - 2\beta)\sqrt{\mu s} + 2 - 2\beta}{2 \sqrt{\mu s}} \right) \right] \\
    &\quad - \sqrt{\mu s} \left[ \frac{1}{(1 - \sqrt{\mu s})^2} \| x_{k+1} - x^* \|^2 + \frac{1}{2(1 - \sqrt{\mu s})} \| \nu_{k+1} \|^2 \right] \\
\end{align*}
\]

\( \blacksquare \)

Lemma 5.5. Let \( f \in S^1_{\mu,E}(\mathbb{R}^n) \), \( \mu \leq L \). Taking any step size \( 0 < s \leq \frac{1}{4L} \), the discrete-time energy functional with \( \{ x_k \}_{k=0}^\infty \) generated by the discrete method satisfies

\[
\varepsilon_{\beta}(k+1) - \varepsilon_{\beta}(k) \leq -\sqrt{\mu s} \min \left\{ \frac{1}{B_{\beta}}, \frac{A_{\beta}}{B_{\beta}} \right\} \varepsilon_{\beta}(k+1)
\]

where

\[
\begin{align*}
    A_{\beta} &= \frac{1}{(1 - \sqrt{\mu s})^2} \left[ 1 - 2Ls \cdot \frac{(\beta - \beta^2)\mu s + (3 + \beta^2 - 2\beta)\sqrt{\mu s} + 2 - 2\beta}{2 \sqrt{\mu s}} \right] \\
    B_{\beta} &= \frac{1}{1 - \sqrt{\mu s}} + \frac{\beta^2 L s}{2}
\end{align*}
\]
\(\beta\)-HIGH RESOLUTION ODE AND PHASE TRANSITION BETWEEN NAG-SC AND HEAVY BALL METHOD

**Proof.** Notice that by the previous lemma,

\[
\begin{align*}
\mathcal{E}(k+1) - \mathcal{E}(k) &\leq -\mu s \left\{ \frac{1}{1 - \sqrt{\mu s}} \left[ 1 - 2Ls \cdot \frac{(\beta - \beta^2)\mu s + (3 + \beta^2 - 2\beta)\sqrt{\mu s} + 2 - 2\beta}{2\sqrt{\mu s}} \right] (f(x_{k+1}) - f(x^*)) \right. \\
&\left. - \sqrt{\mu s} \left\{ \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \left[ f(x_{k+1}) - f(x^*) - \left( \frac{\beta^2 s \sqrt{\mu s} - (\beta^2 - \beta)s}{2\sqrt{\mu s}} \right) \|\nabla f(x_{k+1})\|^2 \right] \right. \\
&\left. - \sqrt{\mu s} \left\{ \frac{\mu s}{2(1 - \sqrt{\mu s})^2} \|x_{k+1} - x^*\|^2 + \frac{1}{1 - \sqrt{\mu s}} \|v_{k+1}\|^2 \right) \right. \\
&\left. := A_{\beta} \right. \end{align*}
\]

On the other hand, we have

\[
\begin{align*}
\mathcal{E}(k) \leq &\left( \frac{1}{1 - \sqrt{\mu s}} + \frac{\beta^2 Ls}{2} \right) (f(x_k) - f(x^*)) + \frac{1 + \sqrt{\mu s} + \mu s}{(1 - \sqrt{\mu s})^2} \|v_k\|^2 \\
&+ \frac{3\mu s}{(1 - \sqrt{\mu s})^2} \|x_k - x^*\|^2 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \left[ f(x_k) - f(x^*) - \left( \frac{\beta^2 s \sqrt{\mu s} - (\beta^2 - \beta)s}{2\sqrt{\mu s}} \right) \|\nabla f(x_k)\|^2 \right] 
\end{align*}
\]

By comparing the coefficients,

\[
\begin{align*}
\mathcal{E}(k+1) - \mathcal{E}(k) &\leq -\mu s \min \left\{ \frac{1}{1 - \sqrt{\mu s}}, \frac{1}{6}, \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s} + \mu s}, A_{\beta} \right\} \mathcal{E}(k+1) \\
&= -\mu s \min \left\{ \frac{1}{6}, \frac{A_{\beta}}{B_{\beta}} \right\} \mathcal{E}(k+1) 
\end{align*}
\]

since \(\frac{1}{1 - \sqrt{\mu s}} > 1 \geq \frac{1}{6}\) and \(\frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s} + \mu s} > \frac{2}{7} > \frac{1}{6}\). \(\square\)

**Lemma 5.6.** When \(\frac{25\mu}{(12L - \mu)^2} \leq s \leq \frac{1}{4L}\), there exists a \(\beta_c \in [0, 1]\) depending on \(\mu, s, L\) such that

\[
\begin{align*}
\frac{A_{\beta}}{B_{\beta}} &\leq \frac{1}{6} \quad \text{when } 0 \leq \beta \leq \beta_c \\
\frac{A_{\beta}}{B_{\beta}} &> \frac{1}{6} \quad \text{when } \beta_c < \beta \leq 1
\end{align*}
\]

**Proof.** For general \(\beta \in [0, 1]\),

\[
\begin{align*}
\frac{A_{\beta}}{B_{\beta}} &= \frac{1 - Ls \cdot (\beta - \beta^2)\mu s + (3 + \beta^2 - 2\beta)\sqrt{\mu s} + 2 - 2\beta}{(1 - \sqrt{\mu s})^2 \left( \frac{1}{1 - \sqrt{\mu s}} + \frac{\beta^2 Ls}{2} \right)} \\
&= \frac{(L\mu s^2 - Ls\sqrt{\mu s})\beta^2 + (2Ls\sqrt{\mu s} - Ls\mu s^2 + 2Ls)\beta + (\sqrt{\mu s} - 3Ls\sqrt{\mu s} - 2Ls)}{\frac{Ls}{2}\sqrt{\mu s}(1 - \sqrt{\mu s})^2 \beta^2 + \sqrt{\mu s} - \mu s}
\end{align*}
\]
To compare \( \frac{\Delta_{\beta}}{p_{\beta}} \) with \( \alpha \), we only need to compare the function \( h(\beta) \) with 0 where
\[
h(\beta) = (L \mu s^2 - L s \sqrt{\mu s}) \beta^2 + (2Ls \sqrt{\mu s} - L \mu s^2 + 2Ls) \beta + (\sqrt{\mu s} - 3Ls \sqrt{\mu s} - 2Ls)
- \frac{1}{6} \left\{ \frac{Ls}{2} \sqrt{\mu s} (1 - \sqrt{\mu s})^2 \beta^2 + \sqrt{\mu s} - \mu s \right\}.
\]
First, it is easy to see that
\[
h(0) = \frac{5}{6} \sqrt{\mu s} - 3Ls \sqrt{\mu s} + \frac{1}{6} \mu s - 2Ls \leq 0
\]
and
\[
h(1) = \frac{1 - 2Ls}{1 - \sqrt{\mu s} + \frac{Ls}{2} (1 - \sqrt{\mu s})^2} \geq 0
\]
when \( \frac{25\mu}{(12L - \mu)^2} \leq s \leq \frac{1}{4L} \). Secondly,
\[
h'(\beta) = Ls \sqrt{\mu s} \left[ 2(\sqrt{\mu s} - 1) - \frac{1}{6} (1 - \sqrt{\mu s})^2 \right] \beta + 2Ls \sqrt{\mu s} - L \mu s^2 + 2Ls
\]
which is a monotone decreasing function on \([0, 1]\). Hence,
\[
h'(\beta) \geq h'(1) = L \mu s^2 + 2Ls - \frac{1}{6} Ls \sqrt{\mu s} (1 - \sqrt{\mu s})^2 \geq L \mu s^2 + 2Ls - \frac{1}{12} Ls \geq 0
\]
Therefore, \( h'(\beta) \geq 0 \) for all \( 0 \leq \beta \leq 1 \). This completes the proof. \( \square \)

**Remark 5.7.** The \( \beta_c \) in the Lemma above is computable,
\[
\beta_c = \frac{-b - \sqrt{b^2 - 4ac}}{2a}
\]
where
\[
a = Ls \sqrt{\mu s} (\sqrt{\mu s} - 1) \left[ 1 - \frac{1}{12} (\sqrt{\mu s} - 1) \right]
\]
\[
b = Ls (2 \sqrt{\mu s} - \mu s + 2)
\]
\[
c = \frac{5}{6} \sqrt{\mu s} - 3Ls \sqrt{\mu s} - 2Ls + \frac{1}{6} \mu s
\]

**Corollary 5.8.** Suppose \( \frac{25\mu}{(12L - \mu)^2} \leq s \leq \frac{1}{4L} \). When \( 0 \leq \beta \leq \beta_c \),
\[
\mathcal{E}_\beta(k + 1) - \mathcal{E}_\beta(k) \leq -\frac{\sqrt{\mu s} - Ls \left[ (\beta - \beta^2) \mu s + (3 + \beta^2 - 2\beta) \sqrt{\mu s} + 2 - 2\beta \right]}{\sqrt{\mu s} \left( 1 - \sqrt{\mu s} + \frac{\beta^2 Ls}{2} (1 - \sqrt{\mu s})^2 \right)} \mathcal{E}_\beta(k)
\]
When \( \beta_c \leq \beta \leq 1 \),
\[
\mathcal{E}_\beta(k + 1) - \mathcal{E}_\beta(k) \leq -\frac{\sqrt{\mu s}}{6} \mathcal{E}_\beta(k)
\]
(2)

**Proof.** Trivially from Lemma 5.6. \( \square \)
5.3. **Proof of Main Results.**

**Proof of Theorem 1.1.** Notice that

\[
\mathcal{E}_\beta(k) \geq \frac{1 + \sqrt{\mu/(cL)}}{1 - \sqrt{\mu/(cL)}} (f(x_k) - f(x^*)) - \frac{\beta \| \nabla f(x_k) \|^2}{(2cL)(1 - \sqrt{\mu/(cL)})}
\]

Together with

\[
f(x_k) - f(x^*) \geq \frac{1}{2L} \| \nabla f(x_k) \|^2
\]

we get

\[
\mathcal{E}_\beta(k) \geq \frac{1 + \sqrt{\mu/(cL)}}{1 - \sqrt{\mu/(cL)}} (f(x_k) - f(x^*)) - \frac{\beta (f(x_k) - f(x^*))}{c(1 - \sqrt{\mu/(cL)})}
\]

Equivalently,

\[
f(x_k) - f(x^*) \leq \frac{c + c \sqrt{\mu/(cL)} - \beta}{c \left( 1 - \sqrt{\mu/(cL)} \right)} \cdot \mathcal{E}_\beta(k)
\]

Applying Corollary 5.8 inductively and plugging in \( s = \frac{1}{ct} \) gives us

\[
\mathcal{E}_\beta(k) \leq \frac{\mathcal{E}_\beta(0)}{\left\{ 1 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} \left[ \left( \beta - \sqrt{\mu s} \right)^2 + 3 + 2 \sqrt{\mu s} - 2 \beta \sqrt{\mu s} \right] \right\}^k}
\]

Recall that the initial velocity \( v_0 = -\frac{2 \sqrt{\mu s} f(x_0)}{1 + \sqrt{\mu s}} \), hence

\[
\mathcal{E}_\beta(0) \leq 1 + \frac{\sqrt{\mu s}}{1 - \sqrt{\mu s}} (f(x_0) - f(x^*)) + \frac{s}{(1 + \sqrt{\mu s})^2} \| \nabla f(x_0) \|^2
\]

\[
+ \frac{1}{4} \left\| \frac{2 \sqrt{\mu s}}{1 - \sqrt{\mu s}} (x_0 - x^*) - \left( \frac{2 - \beta - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \right) \cdot \sqrt{\mu} \cdot \nabla f(x_0) \right\|^2
\]

\[
\leq \left[ \frac{1}{2} \left( \frac{1 + \sqrt{\mu s}}{1 - \sqrt{\mu s}} \right) + \frac{\sqrt{\mu s}}{(1 + \sqrt{\mu s})^2} + \frac{2 \mu / L}{(1 - \sqrt{\mu s})^2} + \frac{L s}{2} \left( \frac{2 - \beta - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \right)^2 \right] L \| x_0 - x^* \|^2
\]

\[
= C_{\beta,\mu,L} \cdot L \cdot \| x_0 - x^* \|^2
\]

where

\[
C_{\beta,\mu,L} = \left[ \frac{1 + \sqrt{\mu/(sL)}}{2 - 2 \sqrt{\mu/(sL)}} + \frac{4(1 + \sqrt{\mu/(sL)})^2}{(1 - \sqrt{\mu/(sL))^2} + \frac{2 \mu / L}{(1 - \sqrt{\mu/(sL))^2} + \frac{1}{2c} \left( \frac{2 - \beta - \sqrt{\mu s}}{1 + \sqrt{\mu s}} \right)^2} \right]
\]

since we write \( s = \frac{1}{ct} \). Let

\[
C'_{\beta,\mu,L} = \frac{c + c \sqrt{\mu/(cL)} - \beta}{c \left( 1 - \sqrt{\mu/(cL)} \right)} \cdot C_{\beta,\mu,L}
\]
and we conclude that
\[
f(x_k) - f(x^*) \leq \frac{C'_{\beta, \mu, L} \cdot L \cdot \|x_0 - x^*\|^2}{1 + \frac{\sqrt{\frac{\mu}{c}} \cdot \beta}{c} \left[ (\beta - \beta^2) \frac{1}{cL} + (3 + \beta^2 - 2\beta) \sqrt{\frac{\mu}{cL}} + 2 - 2\beta \right]}^k
\]

This completes the proof of the subcritical regime. The supercritical regime follows directly from Corollary 5.8 and [1] Theorem 3. □

Remark 5.9. As \( \beta \) reaches to 0, since \( \frac{A_0}{B_0} \leq \frac{1}{6} \) if \( \frac{25\mu}{(12L - \mu)^2} \leq s \leq \frac{1}{4L} \), we have to choose a step size smaller than \( \frac{25\mu}{(12L - \mu)^2} \) in order to let \( \frac{A_0}{B_0} > \frac{1}{6} \). For instance, \( \mu = \frac{1}{10L^2} \) works here. This matches the (1.15) and gives a different reasoning than the one stated in [1] of why we need a more conservative step size on Heavy ball method.

References

[1] B. Shi, S. Du, M. Jordan, W. Su Understanding the Acceleration Phenomenon Via High-Resolution Differential Equations arXiv:1810.08907v3
[2] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1-17,1964
[3] Boris T Polyak. Introduction to Optimization. Optimization Software, Inc, New York, 1987
[4] Yurii Nesterov. A method of solving a convex programming problem with convergence rate \( O(1/k^2) \). Soviet Mathematics Doklady, 27(2):371-376, 1983
[5] Yurii Nesterov. Introductory Lectures on Convex Optimization: A Basic Course, volume 87 Springer Science and Business Media, 2013
[6] Weijie Su, Stephen Boyd and Emmanuel J Candès. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. Journal of Machine Learning Research, 17(153):1-43,2016
[7] Lawrence Perko. Differential equation and dynamical systems, volume 7. Springer Science and Business Media, 2013.
[8] Ashia C Wilson, Benjamin Recht and Michael I Jordan. A Lyapunov Analysis of momentum methods in optimization. Proceedings of National Academy of Sciences, 113(47): E7351-E7358, 2016

University of Pennsylvania, Department of Mathematics, David Rittenhouse Lab, 209 South 33rd Street, Philadelphia, PA 19104-6395
E-mail address: dawu@math.upenn.edu