Mathematical Programming
https://doi.org/10.1007/s10107-021-01681-8

FULL LENGTH PAPER

Series A

Understanding the acceleration phenomenon via high-resolution differential equations

Bin Shi[1] · **Simon S. Du**[2] · **Michael I. Jordan**[3] · **Weijie J. Su**[4]

Received: 27 October 2018 / Accepted: 6 June 2021

Abstract
Gradient-based optimization algorithms can be studied from the perspective of limiting ordinary differential equations (ODEs). Motivated by the fact that existing ODEs do not distinguish between two fundamentally different algorithms—Nesterov's accelerated gradient method for strongly convex functions (NAG-SC) and Polyak's heavy-ball method—we study an alternative limiting process that yields *high-resolution ODEs*. We show that these ODEs permit a general Lyapunov function framework for the analysis of convergence in both continuous and discrete time. We also show that these ODEs are more accurate surrogates for the underlying algorithms; in particular, they not only distinguish between NAG-SC and Polyak's heavy-ball method, but they allow the identification of a term that we refer to as "gradient correction" that is present in NAG-SC but not in the heavy-ball method and is responsible for the qualitative difference in convergence of the two methods. We also use the high-resolution ODE framework to study Nesterov's accelerated gradient method for (non-strongly) convex functions, uncovering a hitherto unknown result—that NAG-C minimizes the squared gradient norm at an inverse cubic rate. Finally, by modifying the high-resolution ODE of NAG-C, we obtain a family of new optimization methods that are shown to maintain the accelerated convergence rates of NAG-C for smooth convex functions.

Keywords Convex optimization · First-order method · Polyak's heavy ball method · Nesterov's accelerated gradient methods · Ordinary differential equation · Lyapunov function · Gradient minimization

Mathematics Subject Classification 34E10 · 65L20 · 65P10 · 90C25 · 90C30 · 93D05

This work was supported in part by the NSF via Grant CCF-1763314 and CAREER Award DMS-1847415, and Army Research Office via Grant W911NF-17-1-0304.

Extended author information available on the last page of the article

Published online: 06 July 2021

1 Introduction

Machine learning has become one of the major application areas for optimization algorithms during the past decade. While there have been many kinds of applications, to a wide variety of problems, the most prominent applications have involved large-scale problems in which the objective function is the sum over terms associated with individual data, such that stochastic gradients can be computed cheaply, while gradients are much more expensive and the computation (and/or storage) of Hessians is often infeasible. In this setting, simple first-order gradient descent algorithms have become dominant, and the effort to make these algorithms applicable to a broad range of machine learning problems has triggered a flurry of new research in optimization, both methodological and theoretical.

We will be considering unconstrained minimization problems,

$$\min_{x\in\mathbb{R}^n} \quad f(x), \tag{1.1}$$

where f is a smooth convex function. Perhaps the simplest first-order method for solving this problem is gradient descent. Taking a fixed step size s, gradient descent is implemented as the recursive rule

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

given an initial point x_0.

As has been known at least since the advent of conjugate gradient algorithms, improvements to gradient descent can be obtained within a first-order framework by using the history of past gradients. Modern research on such extended first-order methods arguably dates to Polyak [39], whose *heavy-ball method* incorporates a momentum term into the gradient step. This approach allows past gradients to influence the current step, while avoiding the complexities of conjugate gradients and permitting a stronger theoretical analysis. Explicitly, starting from an initial point $x_0,\ x_1 \in \mathbb{R}^n$, the heavy-ball method updates the iterates according to

$$x_{k+1} = x_k + \alpha\,(x_k - x_{k-1}) - s\nabla f(x_k), \tag{1.2}$$

where $\alpha > 0$ is the momentum coefficient. While the heavy-ball method provably attains a faster rate of *local* convergence than gradient descent near a minimum of f, it does not come with *global* guarantees. Indeed, [31] demonstrate that even for strongly convex functions the method can fail to converge for some choices of the step size.[1]

The next major development in first-order methodology was due to Nesterov, who discovered a class of *accelerated gradient methods* that have a faster global convergence rate than gradient descent [34,36]. For a μ-strongly convex objective f with

[1] Note that [39] considers $s = 4/(\sqrt{L} + \sqrt{\mu})^2$ and $\alpha = (1 - \sqrt{\mu s})^2$. This momentum coefficient is basically the same as the choice of $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ (adopted starting from Sect. 1.1) if s is small.

L-Lipschitz gradients, Nesterov's accelerated gradient method (NAG-SC) involves the following pair of update equations:

$$\begin{aligned} y_{k+1} &= x_k - s\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\,(y_{k+1} - y_k)\,, \end{aligned} \tag{1.3}$$

given an initial point $x_0 = y_0 \in \mathbb{R}^n$. Equivalently, NAG-SC can be written in a single-variable form that is similar to the heavy-ball method:

$$\begin{aligned} x_{k+1} = x_k + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\,(x_k - x_{k-1}) - s\nabla f(x_k) - \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \\ \cdot s\,(\nabla f(x_k) - \nabla f(x_{k-1}))\,, \end{aligned} \tag{1.4}$$

starting from x_0 and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$. It is worthwhile mentioning that the Ravine method of Gelfand and Tsetlin is in a similar form [22]. Like the heavy-ball method, NAG-SC blends gradient and momentum contributions into its update direction, but defines a specific momentum coefficient $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$. Nesterov also developed the estimate sequence technique to prove that NAG-SC achieves an accelerated linear convergence rate:

$$f(x_k) - f(x^\star) \le O\left(\left(1-\sqrt{s\mu}\right)^k\right),$$

if the step size satisfies $0 < s \le 1/L$. Moreover, for a (weakly) convex objective f with L-Lipschitz gradients, Nesterov defined a related accelerated gradient method (NAG-C) that takes the following form:

$$\begin{aligned} y_{k+1} &= x_k - s\nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k), \end{aligned} \tag{1.5}$$

with $x_0 = y_0 \in \mathbb{R}^n$. The choice of momentum coefficient $\frac{k}{k+3}$, which tends to one, is fundamental to the estimate-sequence-based argument used by Nesterov to establish the following inverse quadratic convergence rate:

$$f(x_k) - f(x^\star) \le O\left(\frac{1}{sk^2}\right), \tag{1.6}$$

for any step size $s \le 1/L$. Under an oracle model of optimization complexity, the convergence rates achieved by NAG-SC and NAG-C are *optimal* for smooth strongly convex functions and smooth convex functions, respectively [33]. For completeness, we remark that the convergence results for these accelerated methods can be carried over to the iterates y_k's when the objective is smooth [36].

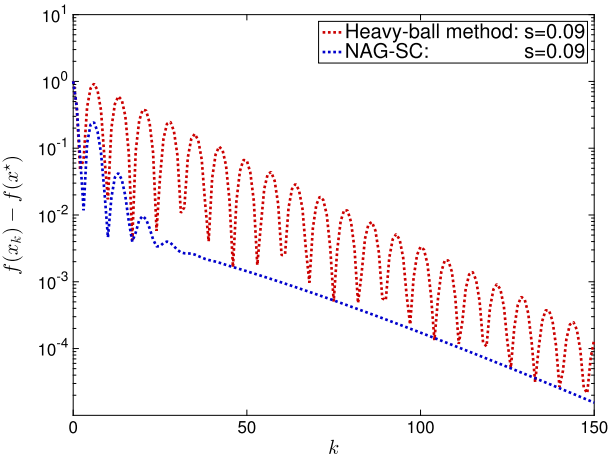

Fig. 1 A numerical comparison between NAG-SC and heavy-ball method. The objective function (ill-conditioned $\mu/L \ll 1$) is $f(x_1, x_2) = 5 \times 10^{-3} x_1^2 + x_2^2$, with the initial iterate (1, 1)

1.1 Gradient correction: small but essential

Throughout the present paper, we set $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$ to define a specific implementation of the heavy-ball method in (1.2). This choice of the momentum coefficient and the second initial point renders the heavy-ball method and NAG-SC identical except for the last (small) term in (1.4). Note that this choice of α can take any value between 0 and 1. Despite their close resemblance, however, the two methods are in fact fundamentally different, with contrasting convergence results (see, for example, [14]). Notably, the former algorithm in general only achieves *local* acceleration, while the latter achieves acceleration method for all initial values of the iterate [31]. As a numerical illustration, Fig. 1 presents the trajectories that arise from the two methods when minimizing an ill-conditioned convex quadratic function. We see that the heavy-ball method exhibits pronounced oscillations throughout the iterations, whereas NAG-SC is monotone in the function value once the iteration counter exceeds 50.

This striking difference between the two methods can *only* be attributed to the last term in (1.4):

$$\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \cdot s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right), \tag{1.7}$$

which we refer to henceforth as the *gradient correction*.[2] This term corrects the update direction in NAG-SC by contrasting the gradients at consecutive iterates. Although an essential ingredient in NAG-SC, the effect of the gradient correction is unclear from the vantage point of the estimate-sequence technique used in Nesterov's proof.

[2] The gradient correction for NAG-C is $\frac{k}{k+3} \cdot s(\nabla f(x_k) - \nabla f(x_{k-1}))$, as seen from the single-variable form of NAG-C: $x_{k+1} = x_k + \frac{k}{k+3}(x_k - x_{k-1}) - s\nabla f(x_k) - \frac{k}{k+3} \cdot s(\nabla f(x_k) - \nabla f(x_{k-1}))$.

Accordingly, while the estimate-sequence technique delivers a proof of acceleration for NAG-SC, it does not explain why the absence of the gradient correction prevents the heavy-ball method from achieving acceleration for strongly convex functions.

A recent line of research has taken a different point of view on the theoretical analysis of acceleration, formulating the problem in continuous time and obtaining algorithms via discretization This can be done by taking continuous-time limits of existing algorithms to obtain ordinary differential equations (ODEs) that can be analyzed using the rich toolbox associated with ODEs, including Lyapunov functions[3]. For instance, [41] shows that

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X(t)) = 0, \tag{1.8}$$

with initial conditions $X(0) = x_0$ and $\dot{X}(0) = 0$, is the exact limit of NAG-C (1.5) by taking the step size $s \to 0$. Alternatively, the starting point may be a Lagrangian or Hamiltonian framework [43]. In either case, the continuous-time perspective not only provides analytical power and intuition, but it also provides design tools for new accelerated algorithms.

Unfortunately, existing continuous-time formulations of acceleration stop short of differentiating between the heavy-ball method and NAG-SC. In particular, these two methods have the *same* limiting ODE (see, for example, [44]):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \nabla f(X(t)) = 0, \tag{1.9}$$

and, as a consequence, this ODE does not provide any insight into the stronger convergence results for NAG-SC as compared to the heavy-ball method. As will be shown in Sect. 2, this is because the gradient correction, $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) = O(s^{1.5})$, is an order of magnitude smaller than the other terms in (1.4) if $s = o(1)$. Consequently, the gradient correction is *not* reflected in the *low-resolution* ODE (1.9) associated with NAG-SC, which is derived by simply taking $s \to 0$ in both (1.2) and (1.4).

1.2 Overview of contributions

Just as there is not a single preferred way to discretize a differential equation, there is not a single preferred way to take a continuous-time limit of a difference equation. Inspired by dimensional-analysis strategies widely used in fluid mechanics in which physical phenomena are investigated at multiple scales via the inclusion of various orders of perturbations [38], we propose to incorporate $O(\sqrt{s})$ terms into the limiting process for obtaining an ODE, including the (Hessian-driven) gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in (1.7). This will yield *high-resolution ODEs* that differentiate between the NAG methods and the heavy-ball method.

[3] One can think of the Lyapunov function as a generalization of the idea of the energy of a system. Then the method studies stability by looking at the rate of change of this measure of energy.

We list the high-resolution ODEs that we derive in the paper here:[4]

(a) The high-resolution ODE for the heavy-ball method (1.2):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + (1+\sqrt{\mu s})\nabla f(X(t)) = 0, \tag{1.10}$$

with $X(0) = x_0$ and $\dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$.

(b) The high-resolution ODE for NAG-SC (1.3):

$$\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + \left(1+\sqrt{\mu s}\right)\nabla f(X(t)) = 0, \tag{1.11}$$

with $X(0) = x_0$ and $\dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$.

(c) The high-resolution ODE for NAG-C (1.5):

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + \left(1+\frac{3\sqrt{s}}{2t}\right)\nabla f(X(t)) = 0, \tag{1.12}$$

for $t \geq 3\sqrt{s}/2$, with $X(3\sqrt{s}/2) = x_0$ and $\dot{X}(3\sqrt{s}/2) = -\sqrt{s}\nabla f(x_0)$.

High-resolution ODEs are more accurate continuous-time counterparts for the corresponding discrete algorithms than low-resolution ODEs, thus allowing for a better characterization of the accelerated methods. This is illustrated in Fig. 2, which presents trajectories and convergence of the discrete methods, and the low- and high-resolution ODEs. For both NAGs, the high-resolution ODEs are in much better agreement with the discrete methods than the low-resolution ODEs.[5] Moreover, for NAG-SC, its high-resolution ODE captures the non-oscillatory pattern while the low-resolution ODE does not.

The three new ODEs include $O(\sqrt{s})$ terms that are not present in the corresponding low-resolution ODEs (compare, for example, (1.12) and (1.8)). Note also that if we let $s \to 0$, each high-resolution ODE reduces to its low-resolution counterpart. Thus, the difference between the heavy-ball method and NAG-SC is reflected only in their high-resolution ODEs—the gradient correction (1.7) of NAG-SC is preserved only in its high-resolution ODE in the form $\sqrt{s}\nabla^2 f(X(t))\dot{X}(t)$. This term, which we refer to as the (Hessian-driven) gradient correction, is connected with the discrete gradient correction by the approximate identity:

$$\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \cdot s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) \approx s\nabla^2 f(x_k)(x_k - x_{k-1}) \approx s^{\frac{3}{2}}\nabla^2 f(X(t))\dot{X}(t),$$

[4] We note that the form of the initial conditions is fixed for each ODE throughout the paper. For example, while x_0 is arbitrary, $X(0)$ and $\dot{X}(0)$ must always be equal to x_0 and $-2\sqrt{s}f(x_0)/(1+\sqrt{\mu s})$, respectively, in the high-resolution ODE of the heavy-ball method. This is in accordance with the choice of $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$.

[5] Note that for the heavy-ball method, the trajectories of the high-resolution ODE and the low-resolution ODE are almost identical.

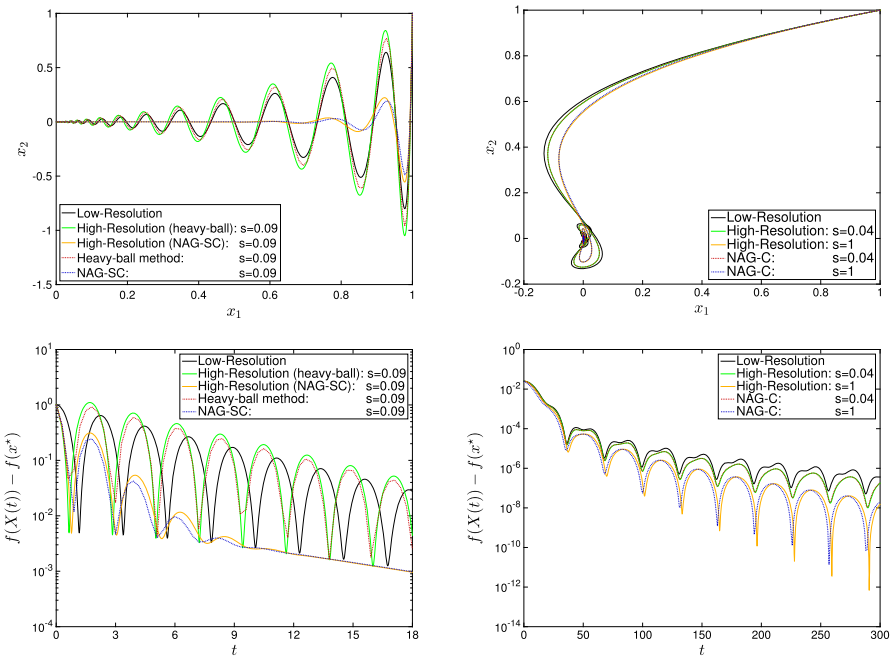

Fig. 2 Top left and bottom left: trajectories and errors of NAG-SC and the heavy-ball method for minimizing $f(x_1, x_2) = 5 \times 10^{-3} x_1^2 + x_2^2$, from the initial value (1, 1), the same setting as Fig. 1. Top right and bottom right: trajectories and errors of NAG-C for minimizing $f(x_1, x_2) = 2 \times 10^{-2} x_1^2 + 5 \times 10^{-3} x_2^2$, from the initial value (1, 1). For the two bottom plots, we use the identification $t = k\sqrt{s}$ between time and iterations for the x-axis

for small s, with the identification $t = k\sqrt{s}$. The gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in NAG-C arises in the same fashion.[6] Interestingly, although both NAGs are first-order methods, their gradient corrections brings in second-order information from the objective function.

Despite being small, the gradient correction has a fundamental effect on the behavior of both NAGs, and this effect is revealed by inspection of the high-resolution ODEs. We provide two illustrations of this.

– *Effect of the gradient correction in acceleration* Viewing the coefficient of $\dot{X}$ as a damping ratio, the ratio $2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X)$ of $\dot{X}$ in the high-resolution ODE (1.11) of NAG-SC is *adaptive* to the position X, in contrast to the *fixed* damping ratio $2\sqrt{\mu}$ in the ODE (1.10) for the heavy-ball method. To appreciate the effect of this adaptivity, imagine that the velocity $\dot{X}$ is highly correlated with an eigenvector of $\nabla^2 f(X)$ with a large eigenvalue, such that the large friction $(2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X))\dot{X}$ effectively "decelerates" along the trajectory of the ODE (1.11) of NAG-SC. This feature of NAG-SC is appealing as taking a cautious step in the presence of high curvature generally helps avoid oscillations. Figure 1 and the left plot of Fig. 2 confirm the superiority of NAG-SC over the heavy-ball method in this respect.

[6] Henceforth, the dependence of X on t is suppressed when clear from the context.

If we can translate this argument to the discrete case we can understand why NAG-SC achieves acceleration globally for strongly convex functions but the heavy-ball method does not. We will be able to make this translation by leveraging the high-resolution ODEs to construct discrete-time Lyapunov functions that allow maximal step sizes to be characterized for the NAG-SC and the heavy-ball method. The detailed analysis is given in Sect. 3.

– *Effect of gradient correction in gradient norm minimization* We will also show how to exploit the high-resolution ODE of NAG-C to construct a continuous-time Lyapunov function to analyze convergence in the setting of a smooth convex objective with L-Lipschitz gradients. Interestingly, the time derivative of the Lyapunov function is not only negative, but it is smaller than $-O(\sqrt{s}t^2\|\nabla f(X)\|^2)$. This bound arises from the gradient correction and, indeed, it cannot be obtained from the Lyapunov function studied in the low-resolution case by [41]. This finer characterization in the high-resolution case allows us to establish a new phenomenon:

$$\min_{0\le i\le k} \|\nabla f(x_i)\|^2 \le O\left(\frac{L^2}{k^3}\right).$$

That is, we discover that NAG-C achieves an inverse *cubic* rate for minimizing the squared gradient norm. By comparison, from (1.6) and the L-Lipschitz continuity of ∇f we can only show that $\|\nabla f(x_k)\|^2 \le O\left(L^2/k^2\right)$. See Sect. 4 for further elaboration on this cubic rate for NAG-C.

1.3 Related work

There is a long history of using ODEs to analyze optimization methods. Recently, the work of [41] has sparked a renewed interest in leveraging continuous dynamical systems to understand and design first-order methods and to provide more intuitive proofs for the discrete methods. Below is a rather incomplete review of recent work that uses continuous-time dynamical systems to study accelerated methods.

In the work of [13,43,44], Lagrangian and Hamiltonian frameworks are used to generate a large class of continuous-time ODEs for a unified treatment of accelerated gradient-based methods. Indeed, [43] extends NAG-C to non-Euclidean settings, mirror descent and accelerated higher-order gradient methods, all from a single "Bregman Lagrangian." In [44], the connection between ODEs and discrete algorithms is further strengthened by establishing an equivalence between the estimate sequence technique and Lyapunov function techniques, allowing for a principled analysis of the discretization of continuous-time ODEs. Recent papers have considered symplectic [13] and Runge–Kutta [45] schemes for discretization of the low-resolution ODEs. Notably, there is a venerable line of work that studies inertial dynamics with a Hessian-driven term [1,7,9,10]. In particular, [2] relates these ODEs to the analysis of associated optimization methods in both convex and non-convex settings, and [9] analyzes forward-backward methods using inertial dynamics with Hessian-driven terms, where the viscous damping coefficient is fixed. While the continuous-time limits considered in these works resemble closely with our ODEs, it is important to note that the

Hessian-driven terms therein result from the second-order information of Newton's method [9], and in contrast, the gradient correction entirely relies on the first-order information of Nesterov's accelerated gradient method.

An ODE-based analysis of mirror descent has been pursued in another line of work by [28–30], delivering new connections between acceleration and constrained optimization, averaging and stochastic mirror descent.

In addition to the perspective of continuous-time dynamical systems, there has also been work on the acceleration from a control-theoretic point of view [11,24,25,31] and from a geometric point of view [15]. See also [18,19,21,23,32,37] for a number of other recent contributions to the study of the acceleration phenomenon.

1.4 Organization and notation

The remainder of the paper is organized as follows. In Sect. 2, we briefly introduce our high-resolution-ODE-based analysis framework. This framework is used in Sect. 3 to study the heavy-ball method and NAG-SC for smooth strongly convex functions. In Sect. 4, we turn our focus to NAG-C for a general smooth convex objective. In Sect. 5 we derive some extensions of NAG-C. We conclude the paper in Sect. 6 with a list of future research directions. Most technical proofs are deferred to the "Appendix".

We mostly follow the notation of [36], with slight modifications tailored to the present paper. Let $\mathcal{F}_L^1(\mathbb{R}^n)$ be 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 $x, y \in \mathbb{R}^n$ and its gradient is L-Lipschitz continuous in the sense that $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$, where $\|\cdot\|$ denotes the standard Euclidean norm and $L > 0$ is the Lipschitz constant. (Note that this implies that ∇f is also L'-Lipschitz for any $L' \geq L$.) The function class $\mathcal{F}_L^2(\mathbb{R}^n)$ is the subclass of $\mathcal{F}_L^1(\mathbb{R}^n)$ such that each f has a Lipschitz-continuous Hessian. 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 μ-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 $x, y \in \mathbb{R}^n$. Note that this is equivalent to the convexity of $f(x) - \frac{\mu}{2}\|x - x^\star\|^2$, where $x^\star$ denotes a minimizer of the objective f.

2 The high-resolution ODE framework

This section introduces a high-resolution ODE framework for analyzing gradient-based methods, with NAG-SC being a guiding example. Given a (discrete) optimization algorithm, the first step in this framework is to derive a high-resolution ODE using dimensional analysis, the next step is to construct a continuous-time Lyapunov function to analyze properties of the ODE, the third step is to derive a discrete-time Lyapunov function from its continuous counterpart and the last step is to translate properties of the ODE into that of the original algorithm. The overall framework is illustrated in Fig. 3.

Step 1: Deriving high-resolution ODEs Our focus is on the single-variable form (1.4) of NAG-SC. For any nonnegative integer k, let $t_k = k\sqrt{s}$ and take the

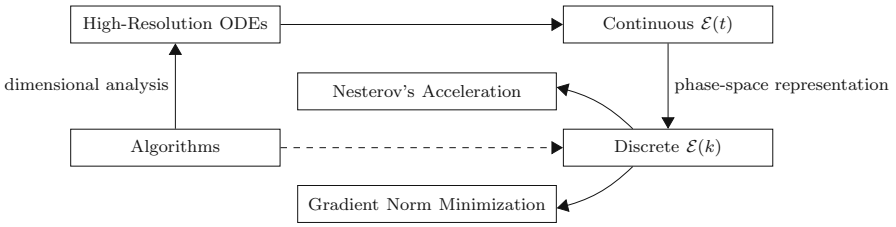

Fig. 3 An illustration of our high-resolution ODE framework. The three solid straight lines represent Steps 1, 2 and 3, and the two curved lines denote Step 4. The dashed line is used to emphasize that it is difficult, if not impractical, to construct discrete Lyapunov functions directly from the algorithms

ansatz that $x_k = X(t_k)$ for some sufficiently smooth curve $X(t)$. Performing a Taylor expansion in powers of $\sqrt{s}$, we get

$$\begin{aligned} x_{k+1} = X(t_{k+1}) &= X(t_k) + \dot{X}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}(t_k)\left(\sqrt{s}\right)^2 + \frac{1}{6}\dddot{X}(t_k)\left(\sqrt{s}\right)^3 + O\left(\left(\sqrt{s}\right)^4\right) \\ x_{k-1} = X(t_{k-1}) &= X(t_k) - \dot{X}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}(t_k)\left(\sqrt{s}\right)^2 - \frac{1}{6}\dddot{X}(t_k)\left(\sqrt{s}\right)^3 + O\left(\left(\sqrt{s}\right)^4\right). \end{aligned} \tag{2.13}$$

We now use a Taylor expansion for the gradient correction, which gives

$$\nabla f(x_k) - \nabla f(x_{k-1}) = \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right). \tag{2.14}$$

Multiplying both sides of (1.4) by $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{s}$ and rearranging the equality, we can rewrite NAG-SC as

$$\begin{aligned} &\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{x_{k+1} - x_k}{s} + \nabla f(x_k) - \nabla f(x_{k-1}) \\ &\quad + \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\nabla f(x_k) = 0. \end{aligned} \tag{2.15}$$

Next, plugging (2.13) and (2.14) into (2.15), we have[7]

$$\begin{aligned} &\ddot{X}(t_k) + O\left(\left(\sqrt{s}\right)^2\right) + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}\left[\dot{X}(t_k) + \frac{1}{2}\ddot{X}(t_k)\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right)\right] \\ &\quad + \nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s} + O\left(\left(\sqrt{s}\right)^2\right) + \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\nabla f(X(t_k)) = 0, \end{aligned}$$

[7] Note that we use the approximation $\frac{x_{k+1}+x_{k-1}-2x_k}{s} = \ddot{X}(t_k) + O(s)$, whereas [41] relies on the low-accuracy Taylor expansion $\frac{x_{k+1}+x_{k-1}-2x_k}{s} = \ddot{X}(t_k) + o(1)$ in the derivation of the low-resolution ODE of NAG-C.

which can be rewritten as

$$\frac{\ddot{X}(t_k)}{1-\sqrt{\mu s}}+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}\dot{X}(t_k)+\sqrt{s}\nabla^2 f(X(t_k))\dot{X}(t_k)+\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\nabla f(X(t_k))+O(s)=0.$$

Multiplying both sides of the last display by $1-\sqrt{\mu s}$, we obtain the following high-resolution ODE of NAG-SC:

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

where we ignore any $O(s)$ terms but retain the $O(\sqrt{s})$ terms (note that $(1-\sqrt{\mu s})\sqrt{s}=\sqrt{s}+O(s)$).

Our analysis is inspired by dimensional analysis [38], a strategy widely used in physics to construct a series of differential equations that involve increasingly high-order terms corresponding to small perturbations. In more detail, taking a small s, one first derives a differential equation that consists only of $O(1)$ terms, then derives a differential equation consisting of both $O(1)$ and $O(\sqrt{s})$, and next, one proceeds to obtain a differential equation consisting of $O(1)$, $O(\sqrt{s})$ and $O(s)$ terms. High-order terms in powers of $\sqrt{s}$ are introduced sequentially until the main characteristics of the original algorithms have been extracted from the resulting approximating differential equation. Thus, we aim to understand Nesterov acceleration by incorporating $O(\sqrt{s})$ terms into the ODE, including the (Hessian-driven) gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ which results from the (discrete) gradient correction (1.7) in the single-variable form (1.4) of NAG-SC. We also show (see "Appendix A.1" for the detailed derivation) that this $O(\sqrt{s})$ term appears in the high-resolution ODE of NAG-C, but is not found in the high-resolution ODE of the heavy-ball method.

As shown below, each ODE admits a unique global solution under mild conditions on the objective, and this holds for an arbitrary step size $s>0$. The solution is accurate in approximating its associated optimization method if s is small. To state the result, we use $C^2(I;\mathbb{R}^n)$ to denote the class of twice continuously differentiable maps from I to $\mathbb{R}^n$ for $I=[0,\infty)$ (the heavy-ball method and NAG-SC) and $I=[1.5\sqrt{s},\infty)$ (NAG-C).

Proposition 1 *For any $f\in\mathcal{S}^2_\mu(\mathbb{R}^n):=\cup_{L\geq\mu}\mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$, each of the ODEs (1.10) and (1.11) with the specified initial conditions has a unique global solution $X\in C^2([0,\infty);\mathbb{R}^n)$. Moreover, the two methods converge to their high-resolution ODEs, respectively, in the sense that*

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

for any fixed $T>0$.

In fact, Proposition 1 holds for $T=\infty$ because both the discrete iterates and the ODE trajectories converge to the unique minimizer when the objective is stongly convex.

Proposition 2 *For any $f \in \mathcal{F}^2(\mathbb{R}^n) := \cup_{L>0}\mathcal{F}_L^2(\mathbb{R}^n)$, the ODE (1.12) with the specified initial conditions has a unique global solution $X \in C^2([1.5\sqrt{s}, \infty); \mathbb{R}^n)$. Moreover, NAG-C converges to its high-resolution ODE in the sense that*

$$\limsup_{s\to 0} \max_{0\le k\le \frac{T}{\sqrt{s}}} \left\| x_k - X(k\sqrt{s} + 1.5\sqrt{s}) \right\| = 0,$$

for any fixed $T > 0$.

The proofs of the two propositions are standard in the theory of ordinary differential equations (see, e.g., the proofs of Theorems 1 and 2 in [41]) and thus are omitted.

Step 2: Analyzing ODEs using Lyapunov functions With these high-resolution ODEs in place, the next step is to construct Lyapunov functions for analyzing the dynamics of the corresponding ODEs, as is done in previous work [31,41,44]. For NAG-SC, we consider the Lyapunov function

$$\begin{aligned}\mathcal{E}(t) = (1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right) + \frac{1}{4}\|\dot{X}\|^2 \\ + \frac{1}{4}\|\dot{X} + 2\sqrt{\mu}(X - x^\star) + \sqrt{s}\nabla f(X)\|^2.\end{aligned} \tag{2.16}$$

The first and second terms $(1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right)$ and $\frac{1}{4}\|\dot{X}\|^2$ can be regarded, respectively, as the potential energy and kinetic energy, and the last term is a mix. For the mixed term, it is interesting to note that the time derivative of $\dot{X} + 2\sqrt{\mu}(X - x^\star) + \sqrt{s}\nabla f(X)$ equals $-(1 + \sqrt{\mu s})\nabla f(X)$.

From a dimensional analysis viewpoint, the step size s has dimension $[\mathrm{T}^{-2}]$, where T denotes the time unit. Consequently, both μ and L have the same dimension $[\mathrm{T}^2]$. Recognizing the assumptions imposed on the objective, which in particular give rise to $\frac{\mu}{2}\|X - x^\star\|^2 \le f(X) - f(x^\star) \le \frac{L}{2}\|X - x^\star\|^2$, one can readily show that every term in this Lyapunov function, such as $(1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right)$, $\frac{1}{4}\|\dot{X}\|^2$, $\frac{1}{4}\|\sqrt{s}\nabla f(X)\|^2$ and any cross terms in the mixed energy, have dimension $[\mathrm{T}^2\mathrm{L}^2]$, where the length unit L is the dimension of X. Indeed, this dimensional analysis viewpoint in part formalizes the intuition for the construction of all Lyapunov functions in this paper.

The differentiability of $\mathcal{E}(t)$ will allow us to investigate properties of the ODE (1.11) in a principled manner. For example, we will show that $\mathcal{E}(t)$ decreases exponentially along the trajectories of (1.11), recovering the accelerated linear convergence rate of NAG-SC. Furthermore, a comparison between the Lyapunov function of NAG-SC and that of the heavy-ball method will explain why the gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ yields acceleration in the former case. This is discussed in Sect. 3.1.

Step 3: Constructing discrete Lyapunov functions Our framework make it possible to translate continuous Lyapunov functions into discrete Lyapunov functions via a phase-space representation (see, for example, [3]). We illustrate the procedure in the case of NAG-SC. The first step is formulate explicit position and velocity updates:

$$\begin{aligned} x_k - x_{k-1} &= \sqrt{s} v_{k-1} \\ v_k - v_{k-1} &= -\frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} v_k - \sqrt{s}(\nabla f(x_k) - \nabla f(x_{k-1})) - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \sqrt{s} \nabla f(x_k), \end{aligned} \tag{2.17}$$

where the velocity variable v_k is defined as $v_k = \frac{x_{k+1}-x_k}{\sqrt{s}}$. The initial velocity is $v_0 = -\frac{2\sqrt{s}}{1+\sqrt{\mu s}} \nabla f(x_0)$. Interestingly, this phase-space representation has the flavor of symplectic discretization, in the sense that the update for $x_k - x_{k-1}$ is explicit (it only depends on the last iterate v_{k-1}) while the update for $v_k - v_{k-1}$ is implicit (it depends on the current iterates x_k and v_k, see [40]).

The representation (2.17) suggests translating the continuous-time Lyapunov function (2.16) into a discrete-time Lyapunov function of the following form:

$$\begin{aligned} \mathcal{E}(k) = & \underbrace{\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(f(x_k) - f(x^\star) \right)}_{\mathbf{I}} \\ & + \underbrace{\frac{1}{4} \|v_k\|^2}_{\mathbf{II}} + \underbrace{\frac{1}{4} \left\| v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}} (x_{k+1} - x^\star) + \sqrt{s} \nabla f(x_k) \right\|^2}_{\mathbf{III}} \\ & \underbrace{- \frac{s \|\nabla f(x_k)\|^2}{2(1-\sqrt{\mu s})}}_{\text{a negative term}}, \end{aligned} \tag{2.18}$$

by replacing continuous terms (e.g., $\dot{X}$) by their discrete counterparts (e.g., v_k). Akin to the continuous (2.16), here **I**, **II**, and **III** correspond to potential energy, kinetic energy, and mixed energy, respectively. To better appreciate this translation, note that the factor $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}$ in **I** results from the term $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \sqrt{s} \nabla f(x_k)$ in (2.17). Likewise, $\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}$ in **III** is from the term $\frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} v_k$ in (2.17). We use x_{k+1} in lieu of x_k as a reflection of the fact that the x variable takes a forward step in the phase-space representation of NAG-SC. The need for the final (small) negative term is technical; we discuss it in Sect. 3.2.

Step 4: Analyzing algorithms using discrete Lyapunov functions The last step is to map properties of high-resolution ODEs to corresponding properties of optimization methods. This step closely mimics Step 2 except that now the object is a discrete algorithm and the tool is a discrete Lyapunov function such as (2.18). Given that Step 2 has been performed, this translation is conceptually straightforward, albeit often calculation-intensive. For example, using the discrete Lyapunov function (2.18), we will recover the optimal linear rate of NAG-SC and gain insights into the fundamental effect of the gradient correction in accelerating NAG-SC. In addition, NAG-C is shown to minimize the squared gradient norm at an inverse cubic rate by a simple analysis of the decreasing rate of its discrete Lyapunov function.

3 Gradient correction for acceleration

In this section, we use our high-resolution ODE framework to analyze NAG-SC and the heavy-ball method. Section 3.1 focuses on the ODEs with an objective function $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$, and in Sect. 3.2 we extend the results to the discrete case for $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. Throughout this section, the strategy is to analyze the two methods in parallel, thereby highlighting the differences between the two methods. In particular, the comparison will demonstrate the vital role of the gradient correction, namely $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \cdot s\left(\nabla f(x_k) - \nabla f(x_{k-1})\right)$ in the discrete case and $\sqrt{s}\nabla^2 f(X)\dot{X}$ in the ODE case, in making NAG-SC an accelerated method.

3.1 The ODE case

The following theorem characterizes the convergence rate of the high-resolution ODE corresponding to NAG-SC.

Theorem 1 (Convergence of NAG-*SC* ODE) *Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $0 < s \le 1/L$, the solution $X = X(t)$ of the high-resolution ODE (1.11) satisfies*

$$f(X(t)) - f(x^\star) \le \frac{2\left\|x_0 - x^\star\right\|^2}{s} \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}.$$

The theorem states that the functional value $f(X)$ tends to the minimum $f(x^\star)$ at a linear rate. By setting $s = 1/L$, we obtain $f(X) - f(x^\star) \le 2L\left\|x_0 - x^\star\right\|^2 \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}$.

The proof of Theorem 1 is based on analyzing the Lyapunov function $\mathcal{E}(t)$ for the high-resolution ODE of NAG-SC. Recall that $\mathcal{E}(t)$ defined in (2.16) is

$$\begin{aligned}\mathcal{E}(t) = (1+\sqrt{\mu s})\left(f(X) - f(x^\star)\right) + \frac{1}{4}\|\dot{X}\|^2 + \frac{1}{4}\|\dot{X} + 2\sqrt{\mu}(X - x^\star) \\ + \sqrt{s}\nabla f(X)\|^2.\end{aligned}$$

The next lemma states the key property we need from this Lyapunov function.

Lemma 1 (Lyapunov function for NAG-*SC* ODE) *Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $s > 0$, and with $X = X(t)$ being the solution to the high-resolution ODE (1.11), the Lyapunov function (2.16) satisfies*

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \le -\frac{\sqrt{\mu}}{4}\mathcal{E}(t). \tag{3.19}$$

The proof of this lemma is placed at the end of this subsection. In particular, the proof reveals that (3.19) can be strengthened to

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \le -\frac{\sqrt{\mu}}{4}\mathcal{E}(t) - \frac{\sqrt{s}}{2}\left[\|\nabla f(X(t))\|^2 + \dot{X}(t)^\top \nabla^2 f(X(t))\dot{X}(t)\right].$$

The term $\frac{\sqrt{s}}{2}(\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X}) \geq 0$ plays no role at the moment, but Sect. 3.2 will shed light on its profound effect in the discretization of the high-resolution ODE of NAG-SC.

Proof of Theorem 1 Lemma 1 implies $\dot{\mathcal{E}}(t) \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}(t)$, which amounts to $\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathcal{E}(t)\mathrm{e}^{\frac{\sqrt{\mu}t}{4}}\right) \leq 0$. By integrating out t, we get

$$\mathcal{E}(t) \leq \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}\mathcal{E}(0). \tag{3.20}$$

Recognizing the initial conditions $X(0) = x_0$ and $\dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$, we write (3.20) as

$$\begin{aligned} f(X) - f(x^\star) \leq \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}} \Bigg[& f(x_0) - f(x^\star) + \frac{s}{\left(1+\sqrt{\mu s}\right)^3}\|\nabla f(x_0)\|^2 \\ & + \frac{1}{4(1+\sqrt{\mu s})}\left\|2\sqrt{\mu}(x_0 - x^\star) - \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot\sqrt{s}\nabla f(x_0)\right\|^2\Bigg]. \end{aligned}$$

Since $f \in \mathcal{S}^2_{\mu,L}$, we have that $\|\nabla f(x_0)\| \leq L\|x_0 - x^\star\|$ and $f(x_0) - f(x^\star) \leq L\|x_0 - x^\star\|^2/2$. Together with the Cauchy–Schwarz inequality, the two inequalities yield

$$\begin{aligned} f(X) - f(x^\star) &\leq \left[f(x_0) - f(x^\star) + \frac{2 + (1-\sqrt{\mu s})^2}{2(1+\sqrt{\mu s})^3}\cdot s\,\|\nabla f(x_0)\|^2 \right. \\ &\quad \left. + \frac{2\mu}{1+\sqrt{\mu s}}\left\|x_0 - x^\star\right\|^2\right]\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}} \\ &\leq \left[\frac{L}{2} + \frac{3 - 2\sqrt{\mu s} + \mu s}{2(1+\sqrt{\mu s})^3}\cdot sL^2 + \frac{2\mu}{1+\sqrt{\mu s}}\right]\left\|x_0 - x^\star\right\|^2\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}, \end{aligned}$$

which is valid for all $s > 0$. To simplify the coefficient of $\|x_0 - x^\star\|^2\,\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}$, note that L can be replaced by $1/s$ in the analysis since $s \leq 1/L$. It follows that

$$f(X(t)) - f(x^\star) \leq \left[\frac{1}{2} + \frac{3 - 2\sqrt{\mu s} + \mu s}{2(1+\sqrt{\mu s})^3} + \frac{2\mu s}{1+\sqrt{\mu s}}\right]\frac{\|x_0 - x^\star\|^2\,\mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}}{s}.$$

Furthermore, a bit of analysis reveals that

$$\frac{1}{2} + \frac{3 - 2\sqrt{\mu s} + \mu s}{2(1+\sqrt{\mu s})^3} + \frac{2\mu s}{1+\sqrt{\mu s}} < 2,$$

since $\mu s \leq \mu/L \leq 1$, and this step completes the proof of Theorem 1. □

We now consider the heavy-ball method (1.2). Recall that the momentum coefficient α is set to $\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$. The following theorem characterizes the rate of convergence of this method.

Theorem 2 (Convergence of heavy-ball ODE) *Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $0 < s \le 1/L$, the solution $X = X(t)$ of the high-resolution ODE (1.10) satisfies*

$$f(X(t)) - f(x^\star) \le \frac{7\,\|x_0 - x^\star\|^2}{2s} \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}.$$

As in the case of NAG-SC, the proof of Theorem 2 is based on a Lyapunov function:

$$\mathcal{E}(t) = (1 + \sqrt{\mu s})\left(f(X) - f(x^\star)\right) + \frac{1}{4}\|\dot{X}\|^2 + \frac{1}{4}\|\dot{X} + 2\sqrt{\mu}(X - x^\star)\|^2, \quad (3.21)$$

which is the same as the Lyapunov function (2.16) for NAG-SC except for the lack of the $\sqrt{s}\nabla f(X)$ term. In particular, (2.16) and (3.21) are identical if $s = 0$. The following lemma considers the decay rate of (3.21).

Lemma 2 (Lyapunov function for the heavy-ball ODE) *Let $f \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$. For any step size $s > 0$, the Lyapunov function (3.21) for the high-resolution ODE (1.10) satisfies*

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \le -\frac{\sqrt{\mu}}{4}\mathcal{E}(t).$$

The proof of Theorem 2 follows the same strategy as the proof of Theorem 1. In brief, Lemma 2 gives $\mathcal{E}(t) \le \mathrm{e}^{-\sqrt{\mu}t/4}\mathcal{E}(0)$ by integrating over the time parameter t. Recognizing the initial conditions

$$X(0) = x_0, \quad \dot{X}(0) = -\frac{2\sqrt{s}\nabla f(x_0)}{1 + \sqrt{\mu s}}$$

in the high-resolution ODE of the heavy-ball method and using the L-smoothness of ∇f, Lemma 2 yields

$$f(X) - f(x^\star) \le \left[\frac{1}{2} + \frac{3}{(1+\sqrt{\mu s})^3} + \frac{2(\mu s)}{1+\sqrt{\mu s}}\right] \frac{\|x_0 - x^\star\|^2\, \mathrm{e}^{-\frac{\sqrt{\mu}t}{4}}}{s},$$

if the step size $s \le 1/L$. Finally, since $0 < \mu s \le \mu/L \le 1$, the coefficient satisfies $\frac{1}{2} + \frac{3}{(1+\sqrt{\mu s})^3} + \frac{2\mu s}{1+\sqrt{\mu s}} < \frac{7}{2}$.

The proofs of Lemmas 1 and 2 share similar ideas. In view of this, we present only the proof of the former here, deferring the proof of Lemma 2 to "Appendix B.2".

Proof of Lemma 1 Along the trajectories of (1.11), the Lyapunov function (2.16) satisfies

$$\begin{aligned}\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} &= (1+\sqrt{\mu s})\langle \nabla f(X), \dot{X}\rangle + \frac{1}{2}\left\langle \dot{X}, -2\sqrt{\mu}\dot{X} - \sqrt{s}\nabla^2 f(X)\dot{X} - (1+\sqrt{\mu s})\nabla f(X)\right\rangle \\ &\quad + \frac{1}{2}\left\langle \dot{X} + 2\sqrt{\mu}\left(X - x^\star\right) + \sqrt{s}\nabla f(X), -(1+\sqrt{\mu s})\nabla f(X)\right\rangle \\ &= -\sqrt{\mu}\left(\|\dot{X}\|^2 + (1+\sqrt{\mu s})\left\langle \nabla f(X), X - x^\star\right\rangle + \frac{s}{2}\|\nabla f(X)\|^2\right) \\ &\quad - \frac{\sqrt{s}}{2}\left[\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X}\right] \\ &\le -\sqrt{\mu}\left(\|\dot{X}\|^2 + (1+\sqrt{\mu s})\left\langle \nabla f(X), X - x^\star\right\rangle + \frac{s}{2}\|\nabla f(X)\|^2\right).\end{aligned} \tag{3.22}$$

Furthermore, $\langle \nabla f(X), X - x^\star\rangle$ is greater than or equal to both $f(X) - f(x^\star) + \frac{\mu}{2}\|X - x^\star\|^2$ and $\mu\|X - x^\star\|^2$ due to the μ-strong convexity of f. This yields

$$\begin{aligned}&(1+\sqrt{\mu s})\left\langle \nabla f(X), X - x^\star\right\rangle \\ &\quad \ge \frac{1+\sqrt{\mu s}}{2}\left\langle \nabla f(X), X - x^\star\right\rangle + \frac{1}{2}\left\langle \nabla f(X), X - x^\star\right\rangle \\ &\quad \ge \frac{1+\sqrt{\mu s}}{2}\left[f(X) - f(x^\star) + \frac{\mu}{2}\|X - x^\star\|^2\right] + \frac{\mu}{2}\|X - x^\star\|^2 \\ &\quad \ge \frac{1+\sqrt{\mu s}}{2}(f(X) - f(x^\star)) + \frac{3\mu}{4}\|X - x^\star\|^2,\end{aligned}$$

which together with (3.22) suggests that the time derivative of this Lyapunov function can be bounded as

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} \le -\sqrt{\mu}\left(\frac{1+\sqrt{\mu s}}{2}(f(X) - f(x^\star)) + \|\dot{X}\|^2 + \frac{3\mu}{4}\left\|X - x^\star\right\|^2 + \frac{s}{2}\|\nabla f(X)\|^2\right). \tag{3.23}$$

Next, the Cauchy–Schwarz inequality yields

$$\left\|2\sqrt{\mu}(X - x^\star) + \dot{X} + \sqrt{s}\nabla f(X)\right\|^2 \le 3\left(4\mu\left\|X - x^\star\right\|^2 + \|\dot{X}\|^2 + s\|\nabla f(X)\|^2\right),$$

from which it follows that

$$\mathcal{E}(t) \le \left(1+\sqrt{\mu s}\right)\left(f(X) - f(x^\star)\right) + \|\dot{X}\|^2 + 3\mu\left\|X - x^\star\right\|^2 + \frac{3s}{4}\|\nabla f(X)\|^2. \tag{3.24}$$

Combining (3.23) and (3.24) completes the proof of the theorem. □

Remark 1 The only inequality in (3.22) is due to the term $\frac{\sqrt{s}}{2}(\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X})$, which is discussed right after the statement of Lemma 1. This term

results from the gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in the NAG-SC ODE. For comparison, this term does not appear in Lemma 2 in the case of the heavy-ball method as its ODE does not include the gradient correction and, accordingly, its Lyapunov function (3.21) is free of the $\sqrt{s}\nabla f(X)$ term.

3.2 The discrete case

This section carries over the results in Sect. 3.1 to the two discrete algorithms, namely NAG-SC and the heavy-ball method. Here we consider an objective $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$ since second-order differentiability of f is not required in the two discrete methods. Recall that both methods start with an arbitrary x_0 and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$.

Theorem 3 (Convergence of NAG-*SC*) *Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. If the step size is set to $s = 1/(4L)$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by NAG-SC (1.3) satisfy*

$$f(x_k) - f(x^\star) \le \frac{5L\,\|x_0 - x^\star\|^2}{\left(1 + \frac{1}{12}\sqrt{\mu/L}\right)^k},$$

for all $k \ge 0$.

In brief, the theorem states that $\log(f(x_k) - f(x^\star)) \le -O(k\sqrt{\mu/L})$, which matches the optimal rate for minimizing smooth strongly convex functions using only first-order information [36]. More precisely, [36] shows that $f(x_k) - f(x^\star) = O((1-\sqrt{\mu/L})^k)$ by taking $s = 1/L$ in NAG-SC. Although this optimal rate of NAG-SC is well known in the literature, this is the first Lyapunov-function-based proof of this result.

As indicated in Sect. 2, the proof of Theorem 3 rests on the Lyapunov function $\mathcal{E}(k)$ from (2.18):

$$\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star)\right) + \frac{1}{4}\|v_k\|^2 + \frac{1}{4}\left\|v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star) + \sqrt{s}\nabla f(x_k)\right\|^2 - \frac{s\,\|\nabla f(x_k)\|^2}{2(1-\sqrt{\mu s})}.$$

Recall that this functional is derived by writing NAG-SC in the phase-space representation (2.17). Analogous to Lemma 1, the following lemma gives an upper bound on the difference $\mathcal{E}(k+1) - \mathcal{E}(k)$.

Lemma 3 (Lyapunov function for NAG-*SC*) *Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. Taking any step size $0 < s \le 1/(4L)$, the discrete Lyapunov function (2.18) with $\{x_k\}_{k=0}^{\infty}$ generated by NAG-SC satisfies*

$$\mathcal{E}(k+1) - \mathcal{E}(k) \le -\frac{\sqrt{\mu s}}{6}\mathcal{E}(k+1).$$

The form of the inequality ensured by Lemma 3 is consistent with that of Lemma 1. Alternatively, it can be written as $\mathcal{E}(k+1) \leq \frac{1}{1+\frac{\sqrt{\mu s}}{6}}\mathcal{E}(k)$. With Lemma 3 in place, we give the proof of Theorem 3.

Proof of Theorem 3 Given $s = 1/(4L)$, we have

$$f(x_k) - f(x^\star) \leq \frac{4(1-\sqrt{\mu/(4L)})}{3+4\sqrt{\mu/(4L)}}\mathcal{E}(k). \tag{3.25}$$

To see this, first note that

$$\begin{aligned}\mathcal{E}(k) &\geq \frac{1+\sqrt{\mu/(4L)}}{1-\sqrt{\mu/(4L)}}\left(f(x_k)-f(x^\star)\right) - \frac{\|\nabla f(x_k)\|^2}{8L(1-\sqrt{\mu/(4L)})},\\ \frac{1}{2L}\|\nabla f(x_k)\|^2 &\leq f(x_k)-f(x^\star).\end{aligned}$$

Combining these two inequalities, we get

$$\begin{aligned}\mathcal{E}(k) &\geq \frac{1+\sqrt{\mu/(4L)}}{1-\sqrt{\mu/(4L)}}\left(f(x_k)-f(x^\star)\right) - \frac{f(x_k)-f(x^\star)}{4(1-\sqrt{\mu/(4L)})}\\ &= \frac{3+4\sqrt{\mu/(4L)}}{4(1-\sqrt{\mu/(4L)})}(f(x_k)-f(x^\star)),\end{aligned}$$

which gives (3.25).

Next, we inductively apply Lemma 3, yielding

$$\begin{aligned}\mathcal{E}(k) &\leq \frac{\mathcal{E}(0)}{\left(1+\frac{\sqrt{\mu s}}{6}\right)^k}\\ &= \frac{\mathcal{E}(0)}{\left(1+\frac{1}{12}\sqrt{\mu/L}\right)^k}.\end{aligned} \tag{3.26}$$

Recognizing the initial velocity $v_0 = -\frac{2\sqrt{s}\nabla f(x_0)}{1+\sqrt{\mu s}}$ in NAG-SC, one can show that

$$\begin{aligned}\mathcal{E}(0) &\leq \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_0)-f(x^\star)\right) + \frac{s}{(1+\sqrt{\mu s})^2}\|\nabla f(x_0)\|^2\\ &\quad + \frac{1}{4}\left\|\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_0-x^\star) - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\sqrt{s}\nabla f(x_0)\right\|^2\\ &\leq \left[\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right) + \frac{Ls}{(1+\sqrt{\mu s})^2} + \frac{2\mu/L}{(1-\sqrt{\mu s})^2}\right.\\ &\quad \left. + \frac{Ls}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\right]\cdot L\left\|x_0-x^\star\right\|^2.\end{aligned} \tag{3.27}$$

Taking $s = 1/(4L)$ in (3.27), it follows from (3.25) and (3.26) that

$$f(x_k) - f(x^\star) \le \frac{C_{\mu/L}\, L \|x_0 - x^\star\|^2}{\left(1 + \frac{1}{12}\sqrt{\mu/L}\right)^k}.$$

Here the constant factor $C_{\mu/L}$ is a short-hand for

$$\frac{4\left(1-\sqrt{\mu/(4L)}\right)}{3+4\sqrt{\mu/(4L)}} \cdot \left[\frac{1+\sqrt{\mu/(4L)}}{2-2\sqrt{\mu/(4L)}} + \frac{1}{4(1+\sqrt{\mu/(4L)})^2} + \frac{2\mu/L}{(1-\sqrt{\mu/(4L)})^2} + \frac{1}{8}\left(\frac{1+\sqrt{\mu/(4L)}}{1-\sqrt{\mu/(4L)}}\right)^2\right],$$

which is less than five by making use of the fact that $\mu/L \le 1$. This completes the proof. □

We now turn to the heavy-ball method (1.2). Recall that $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$ and $x_1 = x_0 - \frac{2s\nabla f(x_0)}{1+\sqrt{\mu s}}$.

Theorem 4 (Convergence of heavy-ball method) *Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. If the step size is set to $s = \mu/(16L^2)$, the iterates $\{x_k\}_{k=0}^\infty$ generated by the heavy-ball method satisfy*

$$f(x_k) - f(x_0) \le \frac{5L\|x_0 - x^\star\|^2}{\left(1 + \frac{\mu}{16L}\right)^k},$$

for all $k \ge 0$.

The heavy-ball method minimizes the objective at the rate $\log(f(x_k) - f(x^\star)) \le -O(k\mu/L)$, as opposed to the optimal rate $-O(k\sqrt{\mu/L})$ obtained by NAG-SC. Thus, the acceleration phenomenon is not observed in the heavy-ball method for minimizing functions in the class $\mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. This difference is, on the surface, attributed to the much smaller step size $s = \mu/(16L^2)$ in Theorem 4 as compared to the $s = 1/(4L)$ step size in Theorem 3. Further discussion of this difference is given after Lemma 4.

In addition to allowing us to complete the proof of Theorem 4, Lemma 4 will shed light on why the heavy-ball method needs a more conservative step size. To state this lemma, we consider the discrete Lyapunov function defined as

$$\mathcal{E}(k) = \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star)\right) + \frac{1}{4}\|v_k\|^2 + \frac{1}{4}\left\|v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star)\right\|^2, \tag{3.28}$$

which is derived by discretizing the continuous Lyapunov function (3.21) using the phase-space representation of the heavy-ball method:

$$\begin{aligned} x_k - x_{k-1} &= \sqrt{s} v_{k-1} \\ v_k - v_{k-1} &= -\frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} v_k - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \sqrt{s} \nabla f(x_k). \end{aligned} \tag{3.29}$$

Lemma 4 (Lyapunov function for the heavy-ball method) *Let $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$. For any step size $s > 0$, the discrete Lyapunov function (3.28) with $\{x_k\}_{k=0}^{\infty}$ generated by the heavy-ball method satisfies*

$$\begin{aligned} \mathcal{E}(k+1) - \mathcal{E}(k) \leq & -\sqrt{\mu s} \min\left\{ \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}, \frac{1}{4} \right\} \mathcal{E}(k+1) \\ & - \left[\frac{3\sqrt{\mu s}}{4} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right) \left(f(x_{k+1}) - f(x^\star) \right) \right. \\ & \left. - \frac{s}{2} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \right]. \end{aligned} \tag{3.30}$$

The proof of Lemma 4 can be found in "Appendix B.3". To apply this lemma to prove Theorem 4, we need to ensure

$$\frac{3\sqrt{\mu s}}{4} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right) \left(f(x_{k+1}) - f(x^\star) \right) - \frac{s}{2} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \geq 0. \tag{3.31}$$

A sufficient and necessary condition for (3.31) is

$$\frac{3\sqrt{\mu s}}{4} \left(f(x_{k+1}) - f(x^\star) \right) - \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right) sL \left(f(x_{k+1}) - f(x^\star) \right) \geq 0. \tag{3.32}$$

This is because $\|\nabla f(x_{k+1})\|^2 \leq 2L \left(f(x_{k+1}) - f(x^\star) \right)$, which can be further reduced to an equality (for example, $f(x) = \frac{L}{2}\|x\|^2$). Thus, the step size s must obey $s = O\left(\frac{\mu}{L^2}\right)$. In particular, the choice of $s = \frac{\mu}{16L^2}$ fulfills (3.32) and, as a consequence, Lemma 4 implies $\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\frac{\mu}{16L}\mathcal{E}(k+1)$. The remainder of the proof of Theorem 4 is similar to that of Theorem 3 and is therefore omitted. As an aside, [39] uses $s = 4/(\sqrt{L} + \sqrt{\mu})^2$ for *local* accelerated convergence of the heavy-ball method. This choice of step size is larger than our step size $s = \frac{\mu}{16L^2}$, which yields a non-accelerated but global convergence rate.

The term $\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 \|\nabla f(x_{k+1})\|^2$ in (3.30) that arises from finite differencing of (3.28) is a (small) term of order $O(s)$ and, as a consequence, this term is not reflected in Lemma 2. In relating to the case of NAG-SC, one would be tempted to ask why this term does not appear in Lemma 3. In fact, a similar term can be found in $\mathcal{E}(k+1) - \mathcal{E}(k)$ by taking a closer look at the proof of Lemma 3. However, this term is canceled out by

the discrete version of the quadratic term $\frac{\sqrt{s}}{2}(\|\nabla f(X)\|^2 + \dot{X}^\top \nabla^2 f(X)\dot{X})$ in Lemma 1 and is, therefore, not present in the statement of Lemma 3. Note that this quadratic term results from the gradient correction (see Remark 1). In light of the above, the gradient correction is the key ingredient that allows for a larger step size in NAG-SC, which is necessary for achieving acceleration.

Before finishing Sect. 3.2 by proving Lemma 3, we briefly remark on the proof strategies for Lemmas 3 and 4. One can rewrite the second lines of the phase-space representations (2.17) and (3.29) as

$$\begin{aligned}
&(1+\sqrt{\mu s})\left(v_k + \sqrt{s}\nabla f(x_k)\right) - (1-\sqrt{\mu s}) \\
&\quad \left(v_{k-1} + \sqrt{s}\nabla f(x_{k-1})\right) = -(1-\sqrt{\mu s})\cdot\sqrt{s}\nabla f(x_k) \\
&(1+\sqrt{\mu s})v_k - (1-\sqrt{\mu s})v_{k-1} = -(1+\sqrt{\mu s})\cdot\sqrt{s}\nabla f(x_k),
\end{aligned}$$

respectively, from which it is straightforward to obtain (3.35) below in the proof and (3.30). Notably, the derivation of (3.35) additionally relies on the fact that the dimension of the first term is the same as the Lyapunov function plus the gradient term.

Proof of Lemma 3 Using the Cauchy–Schwarz inequality, we have (see the definition of **III** in (2.18))

$$\begin{aligned}
\mathbf{III} =& \frac{1}{4}\left\|\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_k - x^\star) + \sqrt{s}\nabla f(x_k)\right\|^2 \\
\leq& \frac{3}{4}\left[\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|v_k\|^2 + \frac{4\mu}{(1-\sqrt{\mu s})^2}\left\|x_k - x^\star\right\|^2 + s\left\|\nabla f(x_k)\right\|^2\right],
\end{aligned}$$

which, together with the inequality

$$\begin{aligned}
\frac{3s}{4}\left\|\nabla f(x_k)\right\|^2 - \frac{s\left\|\nabla f(x_k)\right\|^2}{2(1-\sqrt{\mu s})} &= \frac{s}{4}\left\|\nabla f(x_k)\right\|^2 + \frac{s}{2}\left\|\nabla f(x_k)\right\|^2 - \frac{s\left\|\nabla f(x_k)\right\|^2}{2(1-\sqrt{\mu s})} \\
&\leq \frac{Ls}{2}\left(f(x_k) - f(x^\star)\right) - \frac{s\sqrt{\mu s}\left\|\nabla f(x_k)\right\|^2}{2(1-\sqrt{\mu s})},
\end{aligned}$$

for $f \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$, shows that the Lyapunov function (2.18) satisfies

$$\begin{aligned}
\mathcal{E}(k) \leq& \left(\frac{1}{1-\sqrt{\mu s}} + \frac{Ls}{2}\right)\left(f(x_k) - f(x^\star)\right) + \frac{1+\sqrt{\mu s}+\mu s}{(1-\sqrt{\mu s})^2}\|v_k\|^2 \\
&+ \frac{3\mu}{(1-\sqrt{\mu s})^2}\left\|x_k - x^\star\right\|^2 + \frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star) - \frac{s}{2}\left\|\nabla f(x_k)\right\|^2\right).
\end{aligned} \tag{3.33}$$

Take the following inequality as given for the moment:

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \leq -\sqrt{\mu s}\Bigg[&\frac{1-2Ls}{\left(1-\sqrt{\mu s}\right)^2}\left(f(x_{k+1})-f(x^\star)\right)+\frac{1}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2\\&+\frac{\mu}{2(1-\sqrt{\mu s})^2}\left\|x_{k+1}-x^\star\right\|^2\\&+\frac{\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\left(f(x_{k+1})-f(x^\star)-\frac{s}{2}\left\|\nabla f(x_{k+1})\right\|^2\right)\Bigg],\end{aligned}\tag{3.34}$$

which holds for $s \leq 1/(2L)$. Comparing the coefficients of the same terms in (3.33) for $\mathcal{E}(k+1)$ and (3.34), we conclude that the first difference of the discrete Lyapunov function (2.18) must satisfy

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) &\leq -\sqrt{\mu s}\min\left\{\frac{1-2Ls}{1-\sqrt{\mu s}+\frac{Ls}{2}\left(1-\sqrt{\mu s}\right)^2},\right.\\&\left.\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}+\mu s},\frac{1}{6},\frac{1}{1-\sqrt{\mu s}}\right\}\mathcal{E}(k+1)\\&\leq -\sqrt{\mu s}\min\left\{\frac{1-2Ls}{1+\frac{Ls}{2}},\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}+\mu s},\frac{1}{6},\frac{1}{1-\sqrt{\mu s}}\right\}\mathcal{E}(k+1)\\&=-\frac{\sqrt{\mu s}}{6}\mathcal{E}(k+1),\end{aligned}$$

since $s \leq 1/(4L)$, as desired.

To complete the proof of this lemma, we now verify (3.34) below. First, we point out that

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \leq -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\Bigg[&\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle\right.\\&\left.-s\left\|\nabla f(x_{k+1})\right\|^2\right)+\left\|v_{k+1}\right\|^2\Bigg]\\&-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}+\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\left(\frac{1}{L}-s\right)\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2\end{aligned}\tag{3.35}$$

implies (3.34) for $s \leq 1/L$. With (3.35) in place, recognizing that

$$\begin{cases}f(x^\star) \geq f(x_{k+1})+\left\langle\nabla f(x_{k+1}),x^\star-x_{k+1}\right\rangle+\dfrac{1}{2L}\left\|\nabla f(x_{k+1})\right\|_2^2\\ f(x^\star) \geq f(x_{k+1})+\left\langle\nabla f(x_{k+1}),x^\star-x_{k+1}\right\rangle+\dfrac{\mu}{2}\left\|x_{k+1}-x^\star\right\|_2^2,\end{cases}$$

when the step size satisfies $s \leq 1/(2L) \leq 1/L$, we have

$$\mathcal{E}(k+1)-\mathcal{E}(k) \leq -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left[\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)(f(x_{k+1})\right.$$

$$
\begin{aligned}
&-f(x^\star)\big)+\frac{1}{2L}\left(\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\|\nabla f(x_{k+1})\|^2\\
&+\frac{\mu}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)\left\|x_{k+1}-x^\star\right\|^2-\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)s\,\|\nabla f(x_{k+1})\|^2+\|v_{k+1}\|^2\Bigg]\\
\le&-\sqrt{\mu s}\Bigg[\left(\frac{1}{1-\sqrt{\mu s}}\right)^2\left(f(x_{k+1})-f(x^\star)-s\,\|\nabla f(x_{k+1})\|^2\right)\\
&+\frac{\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\left(f(x_{k+1})-f(x^\star)-\frac{s}{2}\|\nabla f(x_{k+1})\|^2\right)\\
&+\frac{\mu}{2(1-\sqrt{\mu s})^2}\left\|x_{k+1}-x^\star\right\|^2+\frac{1}{1-\sqrt{\mu s}}\|v_{k+1}\|^2\Bigg]\\
\le&-\sqrt{\mu s}\Bigg[\frac{1-2Ls}{\left(1-\sqrt{\mu s}\right)^2}\left(f(x_{k+1})-f(x^\star)\right)+\frac{1}{1-\sqrt{\mu s}}\|v_{k+1}\|^2\\
&+\frac{\mu}{2(1-\sqrt{\mu s})^2}\left\|x_{k+1}-x^\star\right\|^2+\frac{\sqrt{\mu s}}{(1-\sqrt{\mu s})^2}\\
&\left(f(x_{k+1})-f(x^\star)-\frac{s}{2}\|\nabla f(x_{k+1})\|^2\right)\Bigg].
\end{aligned}
$$

Now, we conclude this section by deriving (3.35). Recall the discrete Lyapunov function (2.18),

$$
\begin{aligned}
\mathcal{E}(k)=&\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_k)-f(x^\star)\right)}_{\mathbf{I}}+\underbrace{\frac{1}{4}\|v_k\|^2}_{\mathbf{II}}+\\
&\underbrace{\frac{1}{4}\left\|v_k+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1}-x^\star)+\sqrt{s}\nabla f(x_k)\right\|^2}_{\mathbf{III}}\\
&\underbrace{-\frac{s}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)\|\nabla f(x_k)\|^2}_{\textbf{additional term}}.
\end{aligned}
$$

Next, we evaluate the difference between $\mathcal{E}(k)$ and $\mathcal{E}(k+1)$ by the three parts, **I**, **II** and **III** respectively.

– For part **I** (potential), using the convexity of the objective, we have

$$
\begin{aligned}
&\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1})-f(x^\star)\right)-\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_k)-f(x^\star)\right)\\
&\quad\le\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left[\langle\nabla f(x_{k+1}),x_{k+1}-x_k\rangle-\frac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\right]
\end{aligned}
$$

$$\leq \underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\,\langle \nabla f(x_{k+1}), v_k\rangle}_{\mathbf{I}_1} \underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2}_{\mathbf{I}_2}.$$

– For part **II** (kinetic energy), using the phase representation of NAG-SC (2.17), we see that $\frac{1}{4}\|v_{k+1}\|^2-\frac{1}{4}\|v_k\|^2 \equiv \frac{1}{2}\langle v_{k+1}-v_k, v_{k+1}\rangle - \frac{1}{4}\|v_{k+1}-v_k\|^2$ equals

$$\begin{aligned}
&-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\|v_{k+1}\|^2-\frac{\sqrt{s}}{2}\langle \nabla f(x_{k+1})-\nabla f(x_k), v_{k+1}\rangle\\
&\quad-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{s}}{2}\langle \nabla f(x_{k+1}), v_{k+1}\rangle-\frac{1}{4}\|v_{k+1}-v_k\|^2\\
&=\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\|v_{k+1}\|^2}_{\mathbf{II}_1}\\
&\quad\underbrace{-\frac{\sqrt{s}}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\langle \nabla f(x_{k+1})-\nabla f(x_k), v_k\rangle}_{\mathbf{II}_2}\\
&\quad\underbrace{+\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot\frac{s}{2}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2}_{\mathbf{II}_3}\\
&\quad\underbrace{+\frac{s}{2}\langle \nabla f(x_{k+1})-\nabla f(x_k), \nabla f(x_{k+1})\rangle}_{\mathbf{II}_4}\\
&\quad\underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{s}}{2}\langle \nabla f(x_{k+1}), v_{k+1}\rangle}_{\mathbf{II}_5}\underbrace{-\frac{1}{4}\|v_{k+1}-v_k\|^2}_{\mathbf{II}_6}.
\end{aligned}$$

– For part **III** (mixed energy), using the phase representation of NAG-SC (2.17), we have

$$\begin{aligned}
&\frac{1}{4}\left\|v_{k+1}+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2}-x^\star)+\sqrt{s}\nabla f(x_{k+1})\right\|^2\\
&\quad-\frac{1}{4}\left\|v_k+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1}-x^\star)+\sqrt{s}\nabla f(x_k)\right\|^2\\
&=\frac{1}{2}\left\langle-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\sqrt{s}\nabla f(x_{k+1}), \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}v_{k+1}\right.\\
&\quad\left.+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1}-x^\star)+\sqrt{s}\nabla f(x_{k+1})\right\rangle\\
&\quad-\frac{1}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\|\nabla f(x_{k+1})\|^2
\end{aligned}$$

$$= \underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle}_{\mathbf{III}_1}$$
$$\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle}_{\mathbf{III}_2}$$
$$\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_3}$$
$$\underbrace{-\frac{1}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_4}.$$

Now, we evaluate the difference of the discrete Lyapunov function (2.18) at $k+1$ and k:

$$\begin{aligned}
\mathcal{E}(k+1)-\mathcal{E}(k) \le\ & \underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle \nabla f(x_{k+1}), v_k\right\rangle}_{\mathbf{I}_1}\underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{I}_2}\\
& \underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\|v_{k+1}\right\|^2}_{\mathbf{II}_1}\underbrace{-\frac{\sqrt{s}}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\left\langle \nabla f(x_{k+1})-\nabla f(x_k), v_k\right\rangle}_{\mathbf{II}_2}\\
& +\underbrace{\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot\frac{s}{2}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2}_{\mathbf{II}_3}+\underbrace{\frac{s}{2}\left\langle \nabla f(x_{k+1})-\nabla f(x_k), \nabla f(x_{k+1})\right\rangle}_{\mathbf{II}_4}\\
& \underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{s}}{2}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle}_{\mathbf{II}_5}\underbrace{-\frac{1}{4}\left\|v_{k+1}-v_k\right\|^2}_{\mathbf{II}_6}\\
& \underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle}_{\mathbf{III}_1}\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle}_{\mathbf{III}_2}\\
& \underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_3}\underbrace{-\frac{1}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_4}\\
& \underbrace{-\frac{s}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)\left(\left\|\nabla f(x_{k+1})\right\|^2-\left\|\nabla f(x_k)\right\|^2\right)}_{\textbf{additional term}}\\
\le\ & \underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle \nabla f(x_{k+1}), x_{k+1}-x^\star\right\rangle+\left\|v_{k+1}\right\|^2\right)}_{\mathbf{II}_1+\mathbf{III}_1}
\end{aligned}$$

$$
\begin{aligned}
&\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left[\sqrt{s}\left\langle\nabla f(x_{k+1}),\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\right\rangle+s\,\|\nabla f(x_{k+1})\|^2\right]}_{\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2+\mathbf{III}_3}\\
&\underbrace{-\frac{\sqrt{s}}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\langle\nabla f(x_{k+1})-\nabla f(x_k),v_k\rangle}_{\mathbf{II}_2}+\underbrace{\frac{s}{2}\langle\nabla f(x_{k+1})-\nabla f(x_k),\nabla f(x_{k+1})\rangle}_{\mathbf{II}_4}\\
&\underbrace{-\frac{1}{4}\left[\|v_{k+1}-v_k\|^2+2\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\,\langle\nabla f(x_{k+1}),v_{k+1}-v_k\rangle+\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\,\|\nabla f(x_{k+1})\|^2\right]}_{\frac{1}{2}\mathbf{I}_1+\mathbf{II}_5+\mathbf{II}_6+\mathbf{III}_4}\\
&\underbrace{-\frac{1}{2}\left[\frac{1}{L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)-s\left(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\right]\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2}_{\mathbf{I}_2+\mathbf{II}_3}\\
&\underbrace{-\frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right)s\left(\|\nabla f(x_{k+1})\|^2-\|\nabla f(x_k)\|^2\right)}_{\textbf{additional term}}.
\end{aligned}
$$

The term $(1/2)\mathbf{I}_1+\mathbf{II}_5+\mathbf{II}_6+\mathbf{III}_4$ is identical to

$$
\begin{aligned}
&-\frac{1}{4}\left[\|v_{k+1}-v_k\|^2+2\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\,\langle\nabla f(x_{k+1}),v_{k+1}-v_k\rangle\right.\\
&\left.+\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2 s\,\|\nabla f(x_{k+1})\|^2\right]\\
&=-\frac{1}{4}\left\|v_{k+1}-v_k+\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\nabla f(x_k)\right\|^2\le 0.
\end{aligned}
$$

Using the phase representation of NAG-SC (2.17), we have

$$
\begin{aligned}
&\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2+\mathbf{III}_3\\
&\quad=-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left[\sqrt{s}\left\langle\nabla f(x_{k+1}),\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\right\rangle+s\,\|\nabla f(x_{k+1})\|^2\right]\\
&\quad=\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)s\left(\langle\nabla f(x_{k+1})-\nabla f(x_k),\nabla f(x_{k+1})\rangle\right.\\
&\quad\left.+\frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}}\|\nabla f(x_{k+1})\|^2\right)\\
&=\underbrace{\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\cdot s\cdot\langle\nabla f(x_{k+1})-\nabla f(x_k),\nabla f(x_{k+1})\rangle}_{\mathbf{IV}_1}\\
&\quad+\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\cdot\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot s\,\|\nabla f(x_{k+1})\|^2}_{\mathbf{IV}_2}.
\end{aligned}
$$

Note that $\mathbf{IV} = (1/2)\mathbf{I}_1 + \mathbf{III}_2 + \mathbf{III}_3$. Then, using the phase representation of NAG-SC (2.17), we have

$$
\begin{aligned}
&\mathcal{E}(k+1) - \mathcal{E}(k) \\
&\leq \underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle - s\,\|\nabla f(x_{k+1})\|^2\right) + \|v_{k+1}\|^2\right)}_{\mathbf{II}_1+\mathbf{III}_1+\mathbf{IV}_2} \\
&\quad \underbrace{-\frac{1}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\,\langle \nabla f(x_{k+1}) - \nabla f(x_k), x_{k+1} - x_k\rangle}_{\mathbf{II}_2} \\
&\quad \underbrace{+\left(\frac{1}{1-\sqrt{\mu s}}\right) s\,\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nabla f(x_{k+1})\rangle}_{\mathbf{II}_4+\mathbf{IV}_1} \\
&\quad \underbrace{-\frac{1}{2}\left[\frac{1}{L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right) - s\left(\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\right]\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2}_{\mathbf{I}_2+\mathbf{II}_3} \\
&\quad \underbrace{-\frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right) s\left(\|\nabla f(x_{k+1})\|^2 - \|\nabla f(x_k)\|^2\right)}_{\textbf{additional term}}.
\end{aligned}
$$

To proceed, note that $\mathbf{II}_4 + \mathbf{IV}_1 +$ **additional term** is a perfect square as this term is identical to

$$
\begin{aligned}
&\left(\frac{1}{1-\sqrt{\mu s}}\right) s\,\langle \nabla f(x_{k+1}) - \nabla f(x_k), \nabla f(x_{k+1})\rangle \\
&\quad - \frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right) s\left(\|\nabla f(x_{k+1})\|^2 - \|\nabla f(x_k)\|^2\right) \\
&\quad = \frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}}\right) s\,\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2.
\end{aligned}
$$

Combining $\mathbf{II}_4 + \mathbf{IV}_1 +$ **additional term**, $\mathbf{I}_2 + \mathbf{II}_3$, we see that ($\mathbf{II}_4 + \mathbf{IV}_1 +$ **additional term**) + ($\mathbf{I}_2 + \mathbf{II}_3$) equals

$$
\begin{aligned}
&\frac{1}{2}\left(\frac{1}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{1}{Ls}\right) s\,\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\
&\leq \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{1}{Ls}\right) s\,\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2.
\end{aligned}
$$

Now, we obtain that the difference of Lyapunov function (2.18) obeys

$$
\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right.
$$

$$
\begin{aligned}
&\left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle - s\,\|\nabla f(x_{k+1})\|^2\right) + \|v_{k+1}\|^2\Big) \\
&- \frac{1}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\,\langle \nabla f(x_{k+1}) - \nabla f(x_k), x_{k+1} - x_k\rangle \\
&+ \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right. \\
&\left. - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{1}{Ls}\right) s\,\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 .
\end{aligned}
$$

Because $\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \le L\,\langle \nabla f(x_{k+1}) - \nabla f(x_k), x_{k+1} - x_k\rangle$ for any $f(x) \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$, we have

$$
\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) \le & -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}} \\
&\left[\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle - s\,\|\nabla f(x_{k+1})\|^2\right) + \|v_{k+1}\|^2\right] \\
&- \frac{1}{2}\cdot\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\cdot\frac{1}{L}\cdot\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\
&+ \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} - \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{1}{Ls}\right) s\,\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\
\le & -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle\right.\right. \\
&\left.\left. - s\,\|\nabla f(x_{k+1})\|^2\right) + \|v_{k+1}\|^2\right) \\
&- \frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}\right)\left(\frac{1}{L} - s\right)\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 .
\end{aligned}
$$

This completes the proof. □

4 Gradient correction for gradient norm minimization

In this section, we extend the use of the high-resolution ODE framework to NAG-C (1.5) in the setting of minimizing an L-smooth convex function f. The main result is an improved rate of NAG-SC for minimizing the squared gradient norm. Indeed, we show that NAG-C achieves the $O(L^2/k^3)$ rate of convergence for minimizing $\|\nabla f(x_k)\|^2$. To the best of our knowledge, this is the *sharpest* known bound for this problem using NAG-C *without* any modification. Moreover, we will show that the gradient correction in NAG-C is responsible for this rate and, as it is therefore unsurprising that this inverse cubic rate was not perceived within the low-resolution ODE frameworks such as that of [41].

4.1 The ODE case

We begin by studying the high-resolution ODE (1.12) corresponding to NAG-C with an objective $f \in \mathcal{F}_L^2(\mathbb{R}^n)$ and an arbitrary step size $s > 0$. For convenience, let $t_0 = 1.5\sqrt{s}$.

Theorem 5 *Assume $f \in \mathcal{F}_L^2(\mathbb{R}^n)$ and let $X = X(t)$ be the solution to the ODE (1.12). The squared gradient norm satisfies*

$$\inf_{t_0 \le u \le t} \|\nabla f(X(u))\|^2 \le \frac{(12 + 9sL)\|x_0 - x^\star\|^2}{2\sqrt{s}(t^3 - t_0^3)},$$

for all $t > t_0$.

By taking the step size $s = 1/L$, this theorem shows that $\inf_{t_0 \le u \le t} \|\nabla f(X(u))\|^2 = O(\sqrt{L}/t^3)$, where the infimum operator is necessary as the squared gradient norm is generally not decreasing in t. In contrast, directly combining the convergence rate of the function value (see Corollary 1) and inequality $\|\nabla f(X)\|^2 \le 2L(f(X) - f(x^\star))$ only gives a $O(L/t^2)$ rate for squared gradient norm minimization. We remark that this inverse cubic rate is also found in an ODE for modeling Newton's method [10].

The proof of the theorem is based on the continuous Lyapunov function

$$\mathcal{E}(t) = t\left(t + \frac{\sqrt{s}}{2}\right)\big(f(X) - f(x^\star)\big) + \frac{1}{2}\|t\dot{X} + 2(X - x^\star) + t\sqrt{s}\nabla f(X)\|^2, \tag{4.36}$$

which reduces to the continuous Lyapunov function in [41] when setting $s = 0$.

Lemma 5 *Let $f \in \mathcal{F}_L^2(\mathbb{R}^n)$. The Lyapunov function defined in (4.36) with $X = X(t)$ being the solution to the ODE (1.12) satisfies*

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t} \le -\left[\sqrt{s}t^2 + \left(\frac{1}{L} + \frac{s}{2}\right)t + \frac{\sqrt{s}}{2L}\right]\|\nabla f(X)\|^2, \tag{4.37}$$

for all $t \ge t_0$.

The decreasing rate of $\mathcal{E}(t)$ as specified in the lemma is sufficient for the proof of Theorem 5. First, note that Lemma 5 readily gives

$$\begin{aligned}&\int_{t_0}^{t}\left[\sqrt{s}u^2 + \left(\frac{1}{L} + \frac{s}{2}\right)u + \frac{\sqrt{s}}{2L}\right]\|\nabla f(X(u))\|^2\,\mathrm{d}u\\ &\quad\le -\int_{t_0}^{t}\frac{\mathrm{d}\mathcal{E}(u)}{\mathrm{d}u}\mathrm{d}u = \mathcal{E}(t_0) - \mathcal{E}(t) \le \mathcal{E}(t_0),\end{aligned}$$

where the last step is due to the fact $\mathcal{E}(t) \geq 0$. Thus, it follows that

$$\begin{aligned}\inf_{t_0\leq u\leq t} \|\nabla f(X(u))\|^2 &\leq \frac{\int_{t_0}^{t}\left[\sqrt{s}u^2+\left(\frac{1}{L}+\frac{s}{2}\right)u+\frac{\sqrt{s}}{2L}\right]\|\nabla f(X(u))\|^2\,\mathrm{d}u}{\int_{t_0}^{t}\sqrt{s}u^2+\left(\frac{1}{L}+\frac{s}{2}\right)u+\frac{\sqrt{s}}{2L}\mathrm{d}u} \\ &\leq \frac{\mathcal{E}(t_0)}{\sqrt{s}(t^3-t_0^3)/3+\left(\frac{1}{L}+\frac{s}{2}\right)(t^2-t_0^2)/2+\frac{\sqrt{s}}{2L}(t-t_0)}.\end{aligned} \tag{4.38}$$

Recognizing the initial conditions of the ODE (1.12), we get

$$\begin{aligned}\mathcal{E}(t_0) &= t_0(t_0+\sqrt{s}/2)(f(x_0)-f(x^\star)) \\ &\quad+\frac{1}{2}\left\|-t_0\sqrt{s}\nabla f(x_0)+2(x_0-x^\star)+t_0\sqrt{s}\nabla f(x_0)\right\|^2 \\ &\leq 3s\cdot\frac{L}{2}\|x_0-x^\star\|^2+2\left\|x_0-x^\star\right\|^2,\end{aligned}$$

which together with (4.38) gives

$$\inf_{t_0\leq u\leq t} \|\nabla f(X(u))\|^2 \leq \frac{(2+1.5sL)\|x_0-x^\star\|^2}{\sqrt{s}(t^3-t_0^3)/3+\left(\frac{1}{L}+\frac{s}{2}\right)(t^2-t_0^2)/2+\frac{\sqrt{s}}{2L}(t-t_0)}. \tag{4.39}$$

This bound reduces to the one claimed by Theorem 5 by only keeping the first term $\sqrt{s}(t^3-t_0^3)/3$ in the denominator.

The gradient correction $\sqrt{s}\nabla^2 f(X)\dot{X}$ in the high-resolution ODE (1.12) plays a pivotal role in Lemma 5 and is, thus, key to Theorem 5. As will be seen in the proof of the lemma, the factor $\|\nabla f(X)\|^2$ in (4.37) results from the term $t\sqrt{s}\nabla f(X)$ in the Lyapunov function (4.36), which arises from the gradient correction in the ODE (1.12). In light of this, the low-resolution ODE (1.8) of NAG-C cannot yield a result similar to Lemma 5; furthermore, we conjecture that the $O(\sqrt{L}/t^3)$ rate does applies to this ODE. Sect. 4.2 will discuss this point further in the discrete case.

In passing, it is worth pointing out that the analysis above applies to the case of $s=0$. In this case, we have $t_0=0$, and (4.39) turns out to be $\inf_{0\leq u\leq t}\|\nabla f(X(u))\|^2 \leq \frac{4L\|x_0-x^\star\|^2}{t^2}$. This result is similar to that of the low-resolution ODE in [41].[8]

This section is concluded with the proof of Lemma 5.

Proof of Lemma 5 The time derivative of the Lyapunov function (4.36) obeys

$$\frac{\mathrm{d}\mathcal{E}(t)}{\mathrm{d}t}=\left(2t+\frac{\sqrt{s}}{2}\right)\left(f(X)-f(x^\star)\right)+t\left(t+\frac{\sqrt{s}}{2}\right)\left\langle\nabla f(X),\dot{X}\right\rangle$$

[8] To see this, recall that [41] shows that $f(X(t))-f(x^\star)\leq\frac{2\|x_0-x^\star\|^2}{t^2}$, where $X=X(t)$ is the solution to (4.39) with $s=0$. Using the L-smoothness of f, we get $\|\nabla f(X(t))\|^2\leq 2L(f(X(t))-f(x^\star))\leq\frac{4L\|x_0-x^\star\|^2}{t^2}$.

$$
\begin{aligned}
&+\left\langle t\dot{X}+2(X-x^\star)+t\sqrt{s}\nabla f(X), -\left(\frac{\sqrt{s}}{2}+t\right)\nabla f(X)\right\rangle \\
&=\left(2t+\frac{\sqrt{s}}{2}\right)\left(f(X)-f(x^\star)\right)-(\sqrt{s}+2t)\left\langle X-x^\star, \nabla f(X)\right\rangle \\
&\quad -\sqrt{s}t\left(t+\frac{\sqrt{s}}{2}\right)\|\nabla f(X)\|^2 .
\end{aligned}
$$

Making use of the basic inequality $f(x^\star) \geq f(X) + \langle \nabla f(X), x^\star - X\rangle + \frac{1}{2L}\|\nabla f(X)\|^2$ for L-smooth f, the expression for $\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}$ above satisfies

$$
\begin{aligned}
\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} &\leq -\frac{\sqrt{s}}{2}\left(f(X)-f(x^\star)\right)-\left(\sqrt{s}t+\frac{1}{L}\right)\left(t+\frac{\sqrt{s}}{2}\right)\|\nabla f(X)\|^2 \\
&\leq -\left(\sqrt{s}t+\frac{1}{L}\right)\left(t+\frac{\sqrt{s}}{2}\right)\|\nabla f(X)\|^2 \\
&= -\left[\sqrt{s}t^2+\left(\frac{1}{L}+\frac{s}{2}\right)t+\frac{\sqrt{s}}{2L}\right]\|\nabla f(X)\|^2 .
\end{aligned}
$$

□

Noting that Lemma 5 shows $\mathcal{E}(t)$ is a decreasing function, we obtain:

$$
f(X)-f(x^\star) \leq \frac{\mathcal{E}(t_0)}{t\left(t+\frac{\sqrt{s}}{2}\right)} = \frac{3s(f(x_0)-f(x^\star))+2\|x_0-x^\star\|^2}{t\left(t+\frac{\sqrt{s}}{2}\right)},
$$

by recognizing the initial conditions of the high-resolution ODE (1.12). This gives the following corollary.

Corollary 1 *Under the same assumptions as in Theorem 5, for any $t > t_0$, we have*

$$
f(X(t))-f(x^\star) \leq \frac{(4+3sL)\|x_0-x^\star\|^2}{t\left(2t+\sqrt{s}\right)}.
$$

4.2 The discrete case

We now turn to the discrete NAG-C (1.5) for minimizing an objective $f \in \mathcal{F}_L^1(\mathbb{R}^n)$. Recall that this algorithm starts from any x_0 and $y_0 = x_0$. The discrete counterpart of Theorem 5 is as follows.

Theorem 6 *Let $f \in \mathcal{F}_L^1(\mathbb{R}^n)$. For any step size $0 < s \leq 1/(3L)$, the iterates $\{x_k\}_{k=0}^\infty$ generated by NAG-C obey*

$$
\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \leq \frac{8568\|x_0-x^\star\|^2}{s^2(k+1)^3},
$$

for all $k \geq 0$. In additional, we have $f(x_k)-f(x^\star) \leq \frac{119\|x_0-x^\star\|^2}{s(k+1)^2}$ for all $k \geq 0$.

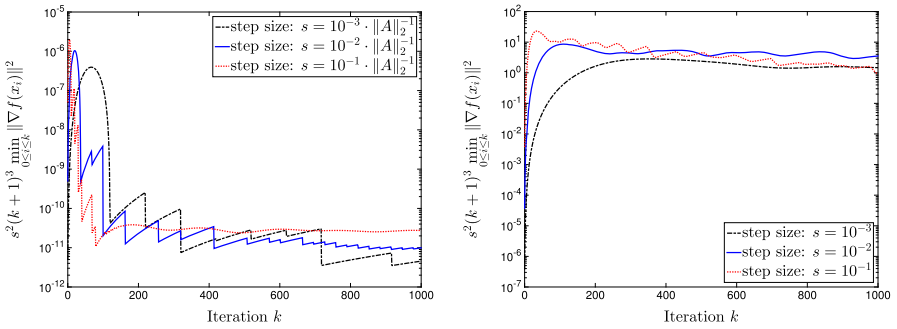

Fig. 4 Scaled squared gradient norm $s^2(k+1)^3 \min_{0\le i\le k} \|\nabla f(x_i)\|^2$ of NAG-C. In both plots, the scaled squared gradient norm stays bounded as $k \to \infty$. Left: $f(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle$, where $A = T'T$ is a 500×500 positive semidefinite matrix and b is 1×500. All entries of b, $T \in \mathbb{R}^{500\times 500}$ are i.i.d. uniform random variables on $(0, 1)$, and $\|\cdot\|_2$ denotes the matrix spectral norm. Right: $f(x) = \rho \log\left\{\sum_{i=1}^{200} \exp\left[(\langle a_i, x\rangle - b_i)/\rho\right]\right\}$, where $A = [a_1, \ldots, a_{200}]'$ is a 200×50 matrix and b is a 200×1 column vector. All entries of A and b are sampled i.i.d. from $\mathcal{N}(0, 1)$ with $\rho = 20$

Remark 2 The convergence result of this theorem carries over effortlessly to the iterate sequence $\{y_k\}_{k=0}^{\infty}$, since the smoothness of the objective ensures that the two iterates are sufficiently close due to the smoothness of the objective. It is important to note, however, that this equivalence is in general not true when applying proximal gradient methods to nonsmooth objectives [12]. While it is beyond the scope of this paper, we refer interested readers to [4,8] for extensions to nonsmooth objectives.

Taking $s = 1/(3L)$, Theorem 6 shows that NAG-C minimizes the squared gradient norm at the rate $O(L^2/k^3)$. This theoretical prediction is in agreement with two numerical examples illustrated in Fig. 4. To our knowledge, the bound $O(L^2/k^3)$ is sharper than any existing bounds in the literature for NAG-C for squared gradient norm minimization. In fact, the convergence result $f(x_k) - f(x^\star) = O(L/k^2)$ for NAG-C and the L-smoothness of the objective immediately give $\|\nabla f(x_k)\|^2 \le O(L^2/k^2)$. This well-known but loose bound can be improved by using a recent result from [8], which shows that a slightly modified version of NAG-C satisfies $f(x_k) - f(x^\star) = o(L/k^2)$ (see Sect. 5.2 for more discussion of this improved rate). This reveals $\|\nabla f(x_k)\|^2 \le o\left(\frac{L^2}{k^2}\right)$, which, however, remains looser than the bound of Theorem 6. In addition, the rate $o(L^2/k^2)$ is not valid for $k \le n/2$ and, as such, the bound $o(L^2/k^2)$ on the squared gradient norm is *dimension-dependent* [8]. For completeness, the rate $O(L^2/k^3)$ can be achieved by introducing an additional sequence of iterates and a more aggressive step-size policy in a variant of NAG-C [23]. In stark contrast, our result shows that no adjustments are needed for NAG-C to yield an accelerated convergence rate for minimizing the gradient norm.

An $\Omega(L^2/k^4)$ lower bound has been established by [35] as the optimal convergence rate for minimizing $\|\nabla f\|^2$ with access to only first-order information. (For completeness, "Appendix C" presents an exposition of this fundamental barrier.) In the same paper, a regularization technique is used in conjunction with NAG-SC to

obtain a matching upper bound (up to a logarithmic factor). This method, however, takes as input the distance between the initial point and the minimizer, which is not practical in general [27].

Returning to Theorem 6, we present a proof of this theorem using a Lyapunov function argument. By way of comparison, we remark that Nesterov's estimate sequence technique is unlikely to be useful for characterizing the convergence of the gradient norm as this technique is essentially based on local quadratic approximations. The phase-space representation of NAG-C (1.5) takes the following form:

$$\begin{aligned} x_k - x_{k-1} &= \sqrt{s} v_{k-1} \\ v_k - v_{k-1} &= -\frac{3}{k} v_k - \sqrt{s}(\nabla f(x_k) - \nabla f(x_{k-1})) - \left(1 + \frac{3}{k}\right) \sqrt{s} \nabla f(x_k), \end{aligned} \tag{4.40}$$

for any initial position x_0 and the initial velocity $v_0 = -\sqrt{s}\nabla f(x_0)$. This representation allows us to discretize the continuous Lyapunov function (4.36) into

$$\begin{aligned} \mathcal{E}(k) = s(k+3)(k+1)\left(f(x_k) - f(x^\star)\right) \\ + \frac{1}{2}\left\|(k+1)\sqrt{s} v_k + 2(x_{k+1} - x^\star) + (k+1)s\nabla f(x_k)\right\|^2. \end{aligned} \tag{4.41}$$

The following lemma characterizes the dynamics of this Lyapunov function. Its proof is relegated to "Appendix C".

Lemma 6 *Under the assumptions of Theorem 6, for all $k \geq 0$ we have*

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\frac{s^2\left((k+3)(k-1) - Ls(k+3)(k+1)\right)}{2} \left\|\nabla f(x_{k+1})\right\|^2.$$

Next, we provide the proof of Theorem 6.

Proof of Theorem 6 We start with the fact that

$$(k+3)(k-1) - Ls(k+3)(k+1) \geq 0, \tag{4.42}$$

for $k \geq 2$. To show this, note that it suffices to guarantee

$$s \leq \frac{1}{L} \cdot \frac{k-1}{k+1}, \tag{4.43}$$

which is self-evident since $s \leq 1/(3L)$ by assumption.

Next, by a telescoping-sum argument, Lemma 6 leads to the following inequalities for $k \geq 4$:

$$\begin{aligned}\mathcal{E}(k) - \mathcal{E}(3) &= \sum_{i=3}^{k-1} (\mathcal{E}(i+1) - \mathcal{E}(i)) \\ &\leq \sum_{i=3}^{k-1} -\frac{s^2}{2} \left[(i+3)(i-1) - Ls(i+3)(i+1)\right] \|\nabla f(x_{i+1})\|^2 \\ &\leq -\frac{s^2}{2} \min_{4\leq i\leq k} \|\nabla f(x_i)\|^2 \sum_{i=3}^{k-1} \left[(i+3)(i-1) - Ls(i+3)(i+1)\right] \\ &\leq -\frac{s^2}{2} \min_{4\leq i\leq k} \|\nabla f(x_i)\|^2 \sum_{i=3}^{k-1} \left[(i+3)(i-1) - \frac{1}{3}(i+3)(i+1)\right], \end{aligned} \tag{4.44}$$

where the second inequality is due to (4.42). To further simplify the bound, observe that

$$\sum_{i=3}^{k-1} \left[(i+3)(i-1) - \frac{1}{3}(i+3)(i+1)\right] = \frac{2k^3 - 38k + 60}{9} \geq \frac{(k+1)^3}{36},$$

for $k \geq 4$. Plugging this inequality into (4.44) yields

$$\mathcal{E}(k) - \mathcal{E}(3) \leq -\frac{s^2(k+1)^3}{72} \min_{4\leq i\leq k} \|\nabla f(x_i)\|^2,$$

which gives

$$\min_{4\leq i\leq k} \|\nabla f(x_i)\|^2 \leq \frac{72(\mathcal{E}(3) - \mathcal{E}(k))}{s^2(k+1)^3} \leq \frac{72\mathcal{E}(3)}{s^2(k+1)^3}. \tag{4.45}$$

It is shown in "Appendix C.4" that $\mathcal{E}(3) \leq \mathcal{E}(2) \leq 119 \|x_0 - x^\star\|^2$, for $s \leq 1/(3L)$. As a consequence of this, (4.45) gives

$$\min_{4\leq i\leq k} \|\nabla f(x_i)\|^2 \leq \frac{8568 \|x_0 - x^\star\|^2}{s^2(k+1)^3}. \tag{4.46}$$

For completeness, "Appendix C.4" proves, via a brute-force calculation, that $\|\nabla f(x_0)\|^2$, $\|\nabla f(x_1)\|^2$, $\|\nabla f(x_2)\|^2$, and $\|\nabla f(x_3)\|^2$ are all bounded above by the right-hand side of (4.46). This completes the proof of the first inequality claimed by Theorem 6.

For the second claim in Theorem 6, the definition of the Lyapunov function and its decreasing property ensured by (4.42) implies

$$f(x_k) - f(x^\star) \le \frac{\mathcal{E}(k)}{s(k+3)(k+1)} \le \frac{\mathcal{E}(2)}{s(k+3)(k+1)} \le \frac{119\,\|x_0 - x^\star\|^2}{s(k+1)^2}, \tag{4.47}$$

for all $k \ge 2$. "Appendix C.4" establishes that $f(x_0) - f(x^\star)$ and $f(x_1) - f(x^\star)$ are bounded by the right-hand side of (4.47). This completes the proof. □

In passing, we remark that the gradient correction sheds light on the superiority of the high-resolution ODE over its low-resolution counterpart, just as in Sect. 3. Indeed, the absence of the gradient correction in the low-resolution ODE leads to the lack of the term $(k+1)s\nabla f(x_k)$ in the Lyapunov function (see Section 4 of [41]), as opposed to the high-resolution Lyapunov function (4.41). Accordingly, it is unlikely to carry over the bound $\mathcal{E}(k+1) - \mathcal{E}(k) \le -O(s^2k^2\|\nabla f(x_{k+1})\|^2)$ of Lemma 6 to the low-resolution case and, consequently, the low-resolution ODE approach pioneered by [41] is insufficient to obtain the $O(L^2/k^3)$ rate for squared gradient norm minimization.

5 Extensions

Motivated by the high-resolution ODE (1.12) of NAG-C, this section considers a family of generalized high-resolution ODEs that take the form

$$\ddot{X} + \frac{\alpha}{t}\dot{X} + \beta\sqrt{s}\nabla^2 f(X)\dot{X} + \left(1 + \frac{\alpha\sqrt{s}}{2t}\right)\nabla f(X) = 0, \tag{5.48}$$

for $t \ge \alpha\sqrt{s}/2$, with initial conditions $X(\alpha\sqrt{s}/2) = x_0$ and $\dot{X}(\alpha\sqrt{s}/2) = -\sqrt{s}\nabla f(x_0)$. As demonstrated in [6,41,42], the low-resolution counterpart (that is, set $s = 0$) of (5.48) achieves acceleration if and only if $\alpha \ge 3$. Accordingly, we focus on the case where the friction parameter $\alpha \ge 3$ and the gradient correction parameter $\beta > 0$. An investigation of the case of $\alpha < 3$ is left for future work.

By discretizing the ODE (5.48), we obtain a family of new accelerated methods for minimizing smooth convex functions:

$$\begin{aligned} y_{k+1} &= x_k - \beta s\nabla f(x_k) \\ x_{k+1} &= x_k - s\nabla f(x_k) + \frac{k}{k+\alpha}(y_{k+1} - y_k), \end{aligned} \tag{5.49}$$

starting with $x_0 = y_0$. The second line of the iteration is equivalent to

$$x_{k+1} = \left(1 - \frac{1}{\beta}\right)x_k + \frac{1}{\beta}y_{k+1} + \frac{k}{k+\alpha}(y_{k+1} - y_k).$$

In Sect. 5.1, we study the convergence rates of this family of generalized NAC-C algorithms along the lines of Sect. 4. To further our understanding of (5.49), Sect.

5.2 shows that this method in the super-critical regime (that is, $\alpha > 3$) converges to the optimum faster than $O(1/(sk^2))$. As earlier, the proofs of all the results follow the high-resolution ODE framework introduced in Sect. 2. Proofs are deferred to "Appendix D". Finally, we note that Sect. 6 briefly sketches the extensions along this direction for NAG-SC.

5.1 Convergence rates

The theorem below characterizes the convergence rates of the generalized NAG-C (5.49).

Theorem 7 *Let $f \in \mathcal{F}_L^1(\mathbb{R}^n)$, $\alpha \geq 3$, and $\beta > \frac{1}{2}$. There exists $c_{\alpha,\beta} > 0$ such that, taking any step size $0 < s \leq c_{\alpha,\beta}/L$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the generalized NAG-C (5.49) obey*

$$\min_{0\leq i\leq k} \|\nabla f(x_i)\|^2 \leq \frac{C_{\alpha,\beta}\|x_0 - x^\star\|^2}{s^2(k+1)^3}, \tag{5.50}$$

for all $k \geq 0$. In addition, we have

$$f(x_k) - f(x^\star) \leq \frac{C_{\alpha,\beta}\|x_0 - x^\star\|^2}{s(k+1)^2},$$

for all $k \geq 0$. The constants $c_{\alpha,\beta}$ and $C_{\alpha,\beta}$ only depend on α and β.

The proof of Theorem 7 is given in "Appendix D.1" for $\alpha = 3$ and "Appendix D.1" for $\alpha > 3$. This theorem shows that the generalized NAG-C achieves the same rates as the original NAG-C in both squared gradient norm and function value minimization. The constraint $\beta > \frac{1}{2}$ reveals that further leveraging of the gradient correction does not hurt acceleration, but perhaps not the other way around (note that NAG-C in its original form corresponds to $\beta = 1$). It is an open question whether this constraint is a technical artifact or is fundamental to acceleration.

5.2 Faster convergence in the super-critical regime

We turn to the case in which $\alpha > 3$, where we show that the generalized NAG-C in this regime attains a faster rate for minimizing the function value. The following proposition provides a technical inequality that motivates the derivation of the improved rate.

Proposition 3 *Let $f \in \mathcal{F}_L^1(\mathbb{R}^n)$, $\alpha > 3$, and $\beta > \frac{1}{2}$. There exists $c'_{\alpha,\beta} > 0$ such that, taking any step size $0 < s \leq c'_{\alpha,\beta}/L$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the generalized NAG-C (5.49) obey*

$$\sum_{k=0}^{\infty}\left[(k+1)\left(f(x_k) - f(x^\star)\right) + s(k+1)^2\|\nabla f(x_k)\|^2\right] \leq \frac{C'_{\alpha,\beta}\|x_0 - x^\star\|^2}{s},$$

where the constants $c'_{\alpha,\beta}$ and $C'_{\alpha,\beta}$ only depend on α and β.

In relating to Theorem 7, one can show that Proposition 3 in fact implies (5.50) in Theorem 7. To see this, note that for $k \geq 1$, one has

$$\min_{0\leq i\leq k} \|\nabla f(x_i)\|^2 \leq \frac{\sum_{i=0}^{k} s(i+1)^2 \|\nabla f(x_i)\|^2}{\sum_{i=0}^{k} s(i+1)^2} \leq \frac{\frac{C'_{\alpha,\beta}\|x_0-x^\star\|^2}{s}}{\frac{s}{6}(k+1)(k+2)(2k+1)} = O\left(\frac{\|x_0-x^\star\|^2}{s^2k^3}\right),$$

where the second inequality follows from Proposition 3.

Proposition 3 can be thought of as a generalization of Theorem 6 of [41]. In particular, this result implies an intriguing and important message. To see this, first note that, by taking $s = O(1/L)$, Proposition 3 gives

$$\sum_{k=0}^{\infty}(k+1)\left(f(x_k)-f(x^\star)\right) = O(L\left\|x_0-x^\star\right\|^2), \tag{5.51}$$

which would not be valid if $f(x_k)-f(x^\star) \geq cL\|x_0-x^\star\|^2/k^2$ for a constant $c>0$. Thus, it is tempting to suggest that there might exist a faster convergence rate in the sense that

$$f(x_k)-f(x^\star) \leq o\left(\frac{L\|x_0-x^\star\|^2}{k^2}\right). \tag{5.52}$$

This faster rate is indeed achievable as we show next, though there are examples where (5.51) and $f(x_k)-f(x^\star) = O(L\|x_0-x^\star\|^2/k^2)$ are both satisfied but (5.52) does not hold (a counterexample is given in "Appendx D.1").

Theorem 8 *Under the same assumptions as in Proposition 3, taking the step size $s = c'_{\alpha,\beta}/L$, the iterates $\{x_k\}_{k=0}^{\infty}$ generated by the generalized NAG-C (5.49) starting from any $x_0 \neq x^\star$ satisfy*

$$\lim_{k\to\infty} \frac{k^2(f(x_k)-f(x^\star))}{L\|x_0-x^\star\|^2} = 0.$$

Figures 5 and 6 present several numerical studies concerning the prediction of Theorem 8. For a fixed dimension n, the convergence in Theorem 8 is uniform over functions in $\mathcal{F}^1 = \cup_{L>0}\mathcal{F}^1_L$ and, consequently, is independent of the Lipschitz constant L and the initial point x_0. In addition to following the high-resolution ODE framework, the proof of this theorem reposes on the finiteness of the series in Proposition 3. See "Appendices D.1" and "D.2" for the full proofs of the proposition and the theorem, respectively.

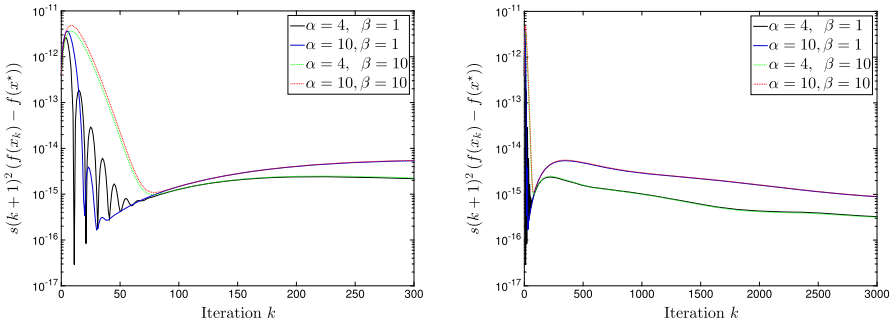

Fig. 5 Scaled error $s(k+1)^2(f(x_k)-f(x^\star))$ of the generalized NAG-C (5.49) with various (α,β). The setting is the same as the left plot of Fig. 4, with the objective $f(x)=\frac{1}{2}\langle Ax,x\rangle+\langle b,x\rangle$. The step size is $s=10^{-1}\|A\|_2^{-1}$. The left shows the short-time behaviors of the methods, while the right focuses on the long-time behaviors. The scaled error curves with the same β are very close to each other in the short-time regime, but in the long-time regime, the scaled error curves with the same α almost overlap. The four scaled error curves slowly tend to zero

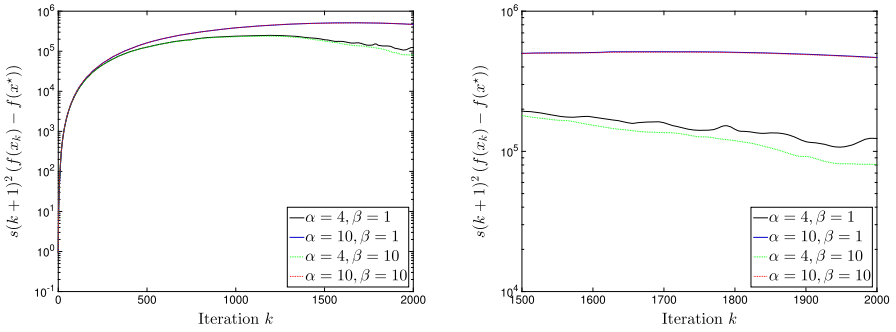

Fig. 6 Scaled error $s(k+1)^2(f(x_k)-f(x^\star))$ of the generalized NAG-C (5.49) with various (α,β). The setting is the same as the right plot of Fig. 4, with the objective $f(x)=\rho\log\left\{\sum_{i=1}^{200}\exp\left[(\langle a_i,x\rangle-b_i)/\rho\right]\right\}$. The step size is $s=0.1$. This set of simulation studies implies that the convergence in Theorem 8 is slow for some problems

In the literature, [5,8] use low-resolution ODEs to establish the faster rate $o(1/k^2)$ for the generalized NAG-C (5.49) in the special case of $\beta=1$. In contrast, our proof of Theorem 8 is more general and applies to a broader class of methods. Notably, [5,8] show in addition that the iterates of NAG-C converge in this regime (see also [17]).

In passing, we make the observation that Proposition 3 reveals that $\sum_{k=1}^{\infty}sk^2\|\nabla f(x_k)\|^2\le\frac{C'_{\alpha,\beta}\|x_0-x^\star\|^2}{s}$, which would not hold if $\min_{0\le i\le k}\|\nabla f(x_i)\|^2\ge c\|x_0-x^\star\|^2/(s^2k^3)$ for all k and a constant $c>0$. In view of the above, it might be true that the rate of the generalized NAG-C for minimizing the squared gradient norm can be improved to $\min_{0\le i\le k}\|\nabla f(x_i)\|^2=o\left(\frac{\|x_0-x^\star\|^2}{s^2k^3}\right)$. We leave the proof or disproof of this asymptotic result for future research.

6 Discussion

In this paper, we have proposed high-resolution ODEs for modeling three first-order optimization methods—the heavy-ball method, NAG-SC, and NAG-C. These new ODEs are more faithful surrogates for the corresponding discrete optimization methods than existing ODEs in the literature, thus serving as a more effective tool for understanding, analyzing, and generalizing first-order methods. Using this tool, we identified a term that we refer to as "gradient correction" in NAG-SC and in its high-resolution ODE, and we demonstrate its critical effect in making NAG-SC an accelerated method, as compared to the heavy-ball method. We also showed via the high-resolution ODE of NAG-C that this method minimizes the squared norm of the gradient at a faster rate than expected for smooth convex functions, and again the gradient correction is the key to this rate. Finally, the analysis of this tool suggested a new family of accelerated methods with the same optimal convergence rates as NAG-C.

The aforementioned results are obtained using the high-resolution ODEs in conjunction with a new framework for translating findings concerning the amenable ODEs into those of the less "user-friendly" discrete methods. This framework encodes an optimization property under investigation into a continuous-time Lyapunov function for an ODE and a discrete-time Lyapunov function for the discrete method. As an appealing feature of this framework, the transformation from the continuous Lyapunov function to its discrete version is through a phase-space representation. This representation links continuous objects such as position and velocity variables to their discrete counterparts in a faithful manner, permitting a transparent analysis of the three discrete methods that we studied.

There are a number of avenues open for future research using the high-resolution ODE framework. First, the discussion of Sect. 5 can carry over to the heavy-ball method and NAG-SC, which correspond to the high-resolution ODE, $\ddot{X}(t) + 2\sqrt{\mu}\dot{X}(t) + \beta\sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + \left(1 + \sqrt{\mu s}\right)\nabla f(X(t)) = 0$, with $\beta = 0$ and $\beta = 1$, respectively. This ODE with a general $0 < \beta < 1$ corresponds to a new algorithm that can be thought of as an interpolation between the two methods. It is of interest to investigate the convergence properties of this class of algorithms. Second, we recognize that new optimization algorithms are obtained in [43] by using different discretization schemes on low-resolution ODE. Hence, a direction of interest is to apply the techniques therein to our high-resolution ODEs and to explore possible appealing properties of the new methods. Third, the technique of dimensional analysis, which we have used to derive high-resolution ODEs, can be further used to incorporate even higher-order powers of $\sqrt{s}$ into the derivation of ODEs. This might lead to further fine-grained findings concerning the discrete methods. Last, the powerful toolbox developed for inertial dynamics might provide further insight in the analysis of our high-resolution ODEs [1,2,7,9,10].

More broadly, we wish to remark on possible extensions of the high-resolution ODE framework beyond smooth convex optimization in the Euclidean setting. In the non-Euclidean case, it would be interesting to derive a high-resolution ODE for mirror descent [43]. This framework might also admit extensions to non-smooth optimization with proximal methods and stochastic optimization, where the ODEs are replaced, respectively, by differential inclusions and stochastic differential equations. Finally,

recognizing that the high-resolution ODEs are well-defined for nonconvex functions, we believe that this framework will provide more accurate characterization of local behaviors of first-order algorithms near saddle points [20,26]. On a related note, given the centrality of the problem of finding an approximate stationary point in the nonconvex setting [16], it is worth using the high-resolution ODE framework to explore possible applications of the faster rate for minimizing the squared gradient norm that we have uncovered.

Acknowledgements Bin Shi is indebted to Xiaoping Yuan for advising him the modern theory of ordinary differential equations and would like to thank Rui Xin Huang for teaching him how to leverage intuitions from physics to understand differential equations. We are grateful to the associate editor and the reviewers for many constructive comments that significantly improved the presentation of this work.

Appendix

A Technical details in Sect. 2

A.1 Derivation of high-resolution ODEs

In this section, we formally derive the high-resolution ODEs of the heavy-ball method and NAG-C. Let $t_k = k\sqrt{s}$. For the moment, let $X(t)$ be a sufficiently smooth map from $[0, \infty)$ (the heavy-ball method) or $[1.5\sqrt{s}, \infty)$ (NAG-C) to $\mathbb{R}^n$, with the correspondence $X(t_k) = X(k\sqrt{s}) = x_k$, where $\{x_k\}_{k=0}^{\infty}$ is the sequence of iterates generated by the heavy-ball method or NAG-C, depending on the context.

The heavy-ball method. For any function $f(x) \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$, setting $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$, multiplying both sides of (1.2) by $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \frac{1}{s}$ and rearranging the equality, we obtain

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu s}}{1-\sqrt{\mu s}} \frac{x_{k+1} - x_k}{s} + \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \nabla f(x_k) = 0. \tag{A.53}$$

Plugging (2.13) into (A.53), we have

$$\ddot{X}(t_k) + O\left(\sqrt{s}\right) + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}} \left[\dot{X}(t_k) + \frac{1}{2}\sqrt{s}\ddot{X}(t_k) + O\left(\left(\sqrt{s}\right)^2\right)\right]$$

$$+\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\nabla f(X(t_k))=0.$$

Ignoring the $O(s)$ term, we obtain the high-resolution ODE (1.10) for the heavy-ball method $\ddot{X}+2\sqrt{\mu}\dot{X}+\left(1+\sqrt{\mu s}\right)\nabla f(X)=0$.

NAG-C For any function $f(x)\in\mathcal{F}_L^2(\mathbb{R}^n)$, multiplying both sides of (1.5) by $\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{1}{s}$ and rearranging the equality, we get

$$\frac{x_{k+1}+x_{k-1}-2x_k}{s}+\frac{3}{k}\cdot\frac{x_{k+1}-x_k}{s}+(\nabla f(x_k)-\nabla f(x_{k-1}))+\left(1+\frac{3}{k}\right)\nabla f(x_k)=0. \tag{A.54}$$

For convenience, we slightly change the definition $t_k=k\sqrt{s}+(3/2)\sqrt{s}$ instead of $t_k=k\sqrt{s}$. Plugging (2.13) into (A.54), we have

$$\ddot{X}(t_k)+O\left(\left(\sqrt{s}\right)^2\right)+\frac{3}{t_k-(3/2)\sqrt{s}}\left[\dot{X}(t_k)+\frac{1}{2}\sqrt{s}\ddot{X}(t_k)+O\left(\left(\sqrt{s}\right)^2\right)\right]$$
$$+\nabla^2 f(X(t_k))\dot{X}(t_k)\sqrt{s}+O\left(\left(\sqrt{s}\right)^2\right)+\frac{t_k+(3/2)\sqrt{s}}{t_k-(3/2)\sqrt{s}}\nabla f(X(t_k))=0.$$

Ignoring any $O(s)$ terms, we obtain the high-resolution ODE (1.12) for NAG-C $\ddot{X}+\frac{3}{t}\dot{X}+\sqrt{s}\nabla^2 f(X)\dot{X}+\left(1+\frac{3\sqrt{s}}{2t}\right)\nabla f(X)=0$.

B Technical details in Sect. 3

B.2 Proof of Lemma 2

Using the Cauchy–Schwarz inequality $\|\dot{X}+2\sqrt{\mu}(X-x^\star)\|^2\le 2\Big(\|\dot{X}\|^2+4\mu\|X-x^\star\|_2^2\Big)$, the Lyapunov function (3.21) can be estimated as

$$\mathcal{E}\le(1+\sqrt{\mu s})\left(f(X)-f(x^\star)\right)+\frac{3}{4}\|\dot{X}\|^2+2\mu\left\|X-x^\star\right\|^2. \tag{B.55}$$

Along the solution to the high-resolution ODE (1.10), the time derivative of the Lyapunov function (3.21) is

$$\begin{aligned}\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}&=(1+\sqrt{\mu s})\left\langle\nabla f(X),\dot{X}\right\rangle+\frac{1}{2}\left\langle\dot{X},-2\sqrt{\mu}\dot{X}-(1+\sqrt{\mu s})\nabla f(X)\right\rangle\\&\quad+\frac{1}{2}\left\langle\dot{X}+2\sqrt{\mu}\left(X-x^\star\right),-(1+\sqrt{\mu s})\nabla f(X)\right\rangle\\&=-\sqrt{\mu}\left[\|\dot{X}\|_2^2+(1+\sqrt{\mu s})\left\langle\nabla f(X),X-x^\star\right\rangle\right].\end{aligned}$$

With (B.55) and the inequality for any function $f(x) \in \mathcal{S}^2_{\mu,L}(\mathbb{R}^n)$, $f(x^\star) \geq f(X) + \langle \nabla f(X), x^\star - X\rangle + \frac{\mu}{2}\|X - x^\star\|_2^2$, the time derivative of the Lyapunov function can be estimated as

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t} \leq -\sqrt{\mu}\left[(1+\sqrt{\mu s})(f(X) - f(x^\star)) + \|\dot{X}\|_2^2 + \frac{\mu}{2}\left\|X - x^\star\right\|_2^2\right] \leq -\frac{\sqrt{\mu}}{4}\mathcal{E}$$

Hence, the proof is complete.

B.3 Proof of Lemma 4

With the phase representation of the heavy-ball method (3.29) and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left\|v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star)\right\|_2^2 &= \left\|\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_k - x^\star)\right\|_2^2 \\ &\leq 2\left[\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|v_k\|_2^2 + \frac{4\mu}{(1-\sqrt{\mu s})^2}\left\|x_k - x^\star\right\|_2^2\right].\end{aligned}$$

The discrete Lyapunov function (3.28) can be estimated as

$$\mathcal{E}(k) \leq \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star)\right) + \frac{1+\mu s}{(1-\sqrt{\mu s})^2}\|v_k\|_2^2 + \frac{2\mu}{(1-\sqrt{\mu s})^2}\left\|x_k - x^\star\right\|_2^2. \tag{B.56}$$

For convenience, we also split the discrete Lyapunov function (3.28) into three parts and mark them as below

$$\mathcal{E}(k) = \underbrace{\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(f(x_k) - f(x^\star)\right)}_{\mathbf{I}} + \underbrace{\frac{1}{4}\|v_k\|^2}_{\mathbf{II}} + \underbrace{\frac{1}{4}\left\|v_k + \frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1} - x^\star)\right\|^2}_{\mathbf{III}},$$

where the three parts **I**, **II** and **III** are corresponding to potential, kinetic energy and mixed energy in classical mechanics, respectively.

– For the part **I**, potential, with the basic convex of $f(x) \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$

$$f(x_k) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle + \frac{1}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|_2^2,$$

we have

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

$$\leq \underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\,\langle \nabla f(x_{k+1}), v_k\rangle}_{\mathbf{I}_1} \underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2}_{\mathbf{I}_2}.$$

– For the part **II**, kinetic energy, with the phase representation of the heavy-ball method (3.29), we have

$$\begin{aligned}
&\frac{1}{4}\|v_{k+1}\|^2-\frac{1}{4}\|v_k\|^2=\frac{1}{2}\langle v_{k+1}-v_k, v_{k+1}\rangle-\frac{1}{4}\|v_{k+1}-v_k\|^2\\
&\quad=\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\|v_{k+1}\|^2}_{\mathbf{II}_1}\underbrace{-\frac{1}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\sqrt{s}\,\langle\nabla f(x_{k+1}), v_{k+1}\rangle}_{\mathbf{II}_2}\\
&\quad\underbrace{-\frac{1}{4}\|v_{k+1}-v_k\|^2}_{\mathbf{II}_3}
\end{aligned}$$

– For the part **III**, mixed energy, with the phase representation of the heavy-ball method (3.29), we have

$$\begin{aligned}
&\frac{1}{4}\left\|v_{k+1}+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2}-x^\star)\right\|^2-\frac{1}{4}\left\|v_k+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+1}-x^\star)\right\|^2\\
&\quad=\frac{1}{4}\left\langle v_{k+1}-v_k+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2}-x_{k+1}), v_{k+1}\right.\\
&\quad\left.+v_k+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2}+x_{k+1}-2x^\star)\right\rangle=-\frac{1}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\sqrt{s}\,\langle\nabla f(x_{k+1}),\\
&\quad v_{k+1}+\frac{2\sqrt{\mu}}{1-\sqrt{\mu s}}(x_{k+2}-x^\star)\rangle-\frac{s}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2\\
&\quad=\underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\langle\nabla f(x_{k+1}), x_{k+1}-x^\star\rangle}_{\mathbf{III}_1}\\
&\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\,\langle\nabla f(x_{k+1}), v_{k+1}\rangle}_{\mathbf{III}_2}\underbrace{-\frac{s}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2}_{\mathbf{III}_3}.
\end{aligned}$$

Next, we calculate the difference of discrete Lyapunov function (2.18) at the k-th iteration by the simple operation as

$$\mathcal{E}(k+1)-\mathcal{E}(k)\leq\underbrace{\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\,\langle\nabla f(x_{k+1}), v_k\rangle}_{\mathbf{I}_1}$$

$$
\begin{aligned}
&\underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2}_{\mathbf{I}_2}\\
&\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\|v_{k+1}\|^2}_{\mathbf{II}_1}\underbrace{-\frac{1}{2}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\sqrt{s}\,\langle\nabla f(x_{k+1}),v_{k+1}\rangle}_{\mathbf{II}_2}\\
&\underbrace{-\frac{1}{4}\|v_{k+1}-v_k\|^2}_{\mathbf{II}_3}\underbrace{-\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\cdot\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle}_{\mathbf{III}_1}\\
&\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\sqrt{s}\,\langle\nabla f(x_{k+1}),v_{k+1}\rangle}_{\mathbf{III}_2}\\
&\underbrace{-\frac{s}{4}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2}_{\mathbf{III}_3}\\
&=\underbrace{-\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle+\|v_{k+1}\|^2\right)}_{\mathbf{II}_1+\mathbf{III}_1}\\
&\underbrace{-\frac{1}{2L}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2}_{\mathbf{I}_2}\\
&\underbrace{-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\left\langle\nabla f(x_{k+1}),\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\right\rangle}_{\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2}\\
&\underbrace{-\frac{1}{4}\left(\|v_{k+1}-v_k\|^2+2\sqrt{s}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\,\langle\nabla f(x_{k+1}),v_{k+1}-v_k\rangle+s\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2\right)}_{\frac{1}{2}\mathbf{I}_1+\mathbf{II}_2+\mathbf{II}_3+\mathbf{III}_3}.
\end{aligned}
$$

With the phase representation of the heavy-ball method (3.29), we have

$$
\begin{aligned}
\frac{1}{2}\mathbf{I}_1+\mathbf{III}_2=&-\frac{1}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\sqrt{s}\,\langle\nabla f(x_{k+1}),\\
&\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)v_{k+1}-v_k\Big\rangle=\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2,
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{1}{2}\mathbf{I}_1+\mathbf{II}_2+\mathbf{II}_3+\mathbf{III}_3=-\frac{1}{4}\Bigg[&\|v_{k+1}-v_k\|^2+2\sqrt{s}\cdot\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\,\langle\nabla f(x_{k+1}),\\
&v_{k+1}-v_k\rangle+s\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2\Bigg]
\end{aligned}
$$

$$= -\frac{1}{4}\left\| v_{k+1} - v_k + \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \cdot \sqrt{s}\nabla f(x_{k+1}) \right\|^2 \leq 0.$$

Now, the difference of discrete Lyapunov function (3.28) can be rewritten as

$$\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) \leq & -\frac{\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right. \\
& \left. \left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star \right\rangle + \|v_{k+1}\|^2 \right) \\
& - \frac{1}{2L} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right) \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\
& + \frac{s}{2} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 .
\end{aligned}$$

With the inequality for any function $f(x) \in \mathcal{S}^1_{\mu,L}(\mathbb{R}^n)$ $f(x^\star) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1} \rangle + \frac{\mu}{2} \|x_{k+1} - x^\star\|^2$, we have

$$\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) \leq & -\sqrt{\mu s} \left[\frac{1+\sqrt{\mu s}}{(1-\sqrt{\mu s})^2} \left(f(x_{k+1}) - f(x^\star) \right) \right. \\
& \left. + \frac{\mu}{2} \cdot \frac{1+\sqrt{\mu s}}{(1-\sqrt{\mu s})^2} \left\| x_{k+1} - x^\star \right\|^2 + \frac{1}{1-\sqrt{\mu s}} \|v_{k+1}\|^2 \right] \\
& + \frac{s}{2} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \\
\leq & -\sqrt{\mu s} \left[\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(f(x_{k+1}) - f(x^\star) \right) \right. \\
& \left. + \frac{\mu}{2} \cdot \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left\| x_{k+1} - x^\star \right\|^2 + \frac{1}{1-\sqrt{\mu s}} \|v_{k+1}\|^2 \right] \\
& + \frac{s}{2} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \\
\leq & -\sqrt{\mu s} \left[\frac{1}{4} \cdot \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left(f(x_{k+1}) - f(x^\star) \right) \right. \\
& \left. + \frac{1}{1-\sqrt{\mu s}} \|v_{k+1}\|^2 + \frac{\mu}{2} \cdot \frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \left\| x_{k+1} - x^\star \right\|^2 \right] \\
& - \left[\frac{3}{4}\sqrt{\mu s} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right) \left(f(x_{k+1}) - f(x^\star) \right) \right. \\
& \left. - \frac{s}{2} \left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}} \right)^2 \|\nabla f(x_{k+1})\|^2 \right].
\end{aligned}$$

Comparing the coefficient of the estimate of Lyapunov function (B.56), we have

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \leq & -\sqrt{\mu s}\min\left\{\frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}, \frac{1}{4}\right\}\mathcal{E}(k+1)\\ & -\left[\frac{3}{4}\sqrt{\mu s}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)\left(f(x_{k+1})-f(x^\star)\right)\right.\\ & \left.-\frac{s}{2}\left(\frac{1+\sqrt{\mu s}}{1-\sqrt{\mu s}}\right)^2\|\nabla f(x_{k+1})\|^2\right].\end{aligned}$$

The proof is complete.

C Technical details in Sect. 4

C.4 Technical details in Proof of Theorem 6

C.4.1 Iterates (x_k, y_k) at $k = 1, 2, 3$

The iterate (x_k, y_k) at $k = 1$ is

$$x_1 = y_1 = x_0 - s\nabla f(x_0). \tag{C.57}$$

When $k = 2$, the iterate (x_k, y_k) is

$$\begin{cases} y_2 = x_0 - s\nabla f(x_0) - s\nabla f(x_0 - s\nabla f(x_0))\\ x_2 = x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0)).\end{cases} \tag{C.58}$$

When $k = 3$, the iterate (x_k, y_k) is

$$\begin{cases} y_3 = x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0)) - s\nabla f\left(x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0))\right)\\ x_3 = x_0 - s\nabla f(x_0) - \dfrac{27}{20}s\nabla f(x_0 - s\nabla f(x_0)) - \dfrac{7}{5}s\nabla f\left(x_0 - s\nabla f(x_0) - \dfrac{5}{4}s\nabla f(x_0 - s\nabla f(x_0))\right).\end{cases} \tag{C.59}$$

C.4.2 Estimate for $\|\nabla f(x_k)\|^2$ at $k = 0, 1, 2, 3$

From (C.57), we have

$$\begin{aligned}\|\nabla f(x_1)\|^2 &= \|\nabla f(x_0 - s\nabla f(x_0))\|^2 \leq L^2\left\|x_0 - x^\star - s\nabla f(x_0)\right\|^2\\ &\leq 2L^2(1+L^2s^2)\left\|x_0 - x^\star\right\|^2.\end{aligned} \tag{C.60}$$

According to (C.59), we have

$$\begin{aligned}\|\nabla f(x_2)\|^2 &= \left\|\nabla f\left(x_0 - s\nabla f(x_0) - \frac{5}{4}s\nabla f\left(x_0 - s\nabla f(x_0)\right)\right)\right\|^2 \\ &\leq L^2\left\|x_0 - x^\star - s\nabla f(x_0) - \frac{5}{4}s\nabla f\left(x_0 - s\nabla f(x_0)\right)\right\|^2 \\ &\leq 3L^2\left(1 + \frac{33}{8}L^2s^2 + \frac{25}{8}L^4s^4\right)\left\|x_0 - x^\star\right\|^2.\end{aligned}$$

From (C.57) and-(C.59), we have

$$\begin{aligned}\|\nabla f(x_3)\|^2 &\leq L^2\left\|x_3 - x^\star\right\|^2 \\ &\leq L^2\left\|x_0 - x^\star - s\nabla f(x_0) - \frac{27}{20}s\nabla f(x_1) - \frac{7}{5}s\nabla f\left(x_2\right)\right\|^2 \\ &= \frac{L^2(40 + 381L^2s^2 + 1156L^4s^4 + 735L^6s^6)}{10}\left\|x_0 - x^\star\right\|^2 \quad \text{(C.61)}\end{aligned}$$

Taking $s \leq 1/(3L)$ and using (C.60), (C.61) and (C.61), we have

$$\|\nabla f(x_0)\|^2 \leq \frac{\|x_0 - x^\star\|^2}{9s^2}, \qquad \|\nabla f(x_1)\|^2 \leq \frac{20\,\|x_0 - x^\star\|^2}{81s^2},$$
$$\|\nabla f(x_2)\|^2 \leq \frac{485\,\|x_0 - x^\star\|^2}{972s^2}, \qquad \|\nabla f(x_3)\|^2 \leq \frac{2372\,\|x_0 - x^\star\|^2}{2187s^2}.$$

C.4.3 Estimate for $f(x_k) - f(x^\star)$ at $k = 0, 1$

According to (C.57), we have

$$\begin{aligned}f(x_1) - f(x^\star) &\leq \frac{L}{2}\left\|x_1 - x^\star\right\|^2 \leq \frac{L}{2}\left\|x_0 - s\nabla f(x_0) - x^\star\right\|^2 \\ &\leq L(1 + L^2s^2)\left\|x_0 - x^\star\right\|^2. \quad \text{(C.62)}\end{aligned}$$

Taking $s \leq 1/(3L)$, (C.62) tells us that

$$f(x_0) - f(x^\star) \leq \frac{\|x_0 - x^\star\|^2}{6s}, \qquad f(x_1) - f(x^\star) \leq \frac{10\,\|x_0 - x^\star\|^2}{27s}.$$

C.4.4 Estimate for Lyapunov function $\mathcal{E}(2)$ and $\mathcal{E}(3)$

With the phase-space representation form (4.40), we have

$$v_2 = \frac{x_3 - x_2}{\sqrt{s}} = \frac{1}{10}\nabla f(x_1) + \frac{7}{5}\nabla f(x_2). \quad \text{(C.63)}$$

According to (4.41), the Lyapunov function $\mathcal{E}(2)$ can be written as

$$\mathcal{E}(2) = 15s\left(f(x_2) - f(x^\star)\right) + \frac{1}{2}\left\|2(x_2 - x^\star) + 5\sqrt{s}v_2 + 3s\nabla f(x_2)\right\|.^2$$

With (C.63) and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathcal{E}(2) &\le \frac{15Ls}{2}\left\|x_2 - x^\star\right\|^2 + \frac{3}{2}\left(4\left\|x_2 - x^\star\right\|^2 + 25s\left\|v_2\right\|^2 + 9s^2\left\|\nabla f(x_2)\right\|^2\right)\\ &= \left(\frac{15Ls}{2} + 6\right)\left\|x_2 - x^\star\right\|^2 + \frac{321}{2}s^2\left\|\nabla f(x_2)\right\|^2 + \frac{3}{4}s^2\left\|\nabla f(x_1)\right\|^2.\end{aligned}$$

Furthermore, with (C.58), we have

$$\begin{aligned}\mathcal{E}(2) \le \left(\frac{15Ls}{2} + 6\right)\left\|x_0 - x^\star - s\nabla f(x_0) - \frac{5}{4}s\nabla f(x_0 - s\nabla f(x_0))\right\|^2 + \frac{321}{2}s^2\|\\ \nabla f(x_2)\|^2 + \frac{3}{4}s^2\left\|\nabla f(x_1)\right\|^2.\end{aligned}$$

Finally, with (C.60) and-(C.61), the Cauchy-Schwarz inequality tells

$$\begin{aligned}\mathcal{E}(2) &\le \left\{\left[\frac{3}{16}(12 + 15Ls) + \frac{963}{16}L^2s^2\right]\left(8 + 33L^2s^2 + 25L^4s^4\right) + \frac{3}{2}L^2s^2(1 + L^2s^2)\right\}\cdot\\ &\quad\left\|x_0 - x^\star\right\|^2\\ &= \frac{288 + 360Ls + 8916L^2s^2 + 1485L^3s^3 + 32703L^4s^4 + 1125L^5s^5 + 24075L^6s^6}{16}\\ &\quad\cdot\left\|x_0 - x^\star\right\|^2. \qquad \text{(C.64)}\end{aligned}$$

By Lemma 6, when the step size $s \le 1/(3L)$, (C.64) tells us $\mathcal{E}(3) \le \mathcal{E}(2) \le 119\left\|x_0 - x^\star\right\|^2$.

Proof of Lemma 6 The difference of the Lyapunov function (4.41) satisfies

$$\begin{aligned}\mathcal{E}(k+1) - \mathcal{E}(k) &= s(k+3)(k+1)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right)\\ &\quad + \big\langle 2(x_{k+2} - x_{k+1}) + \sqrt{s}(k+2)(v_{k+1} + \sqrt{s}\nabla f(x_{k+1}))\\ &\quad -\sqrt{s}(k+1)(v_k + \sqrt{s}\nabla f(x_k)),\\ &\quad 2(x_{k+2} - x^\star) + (k+2)\sqrt{s}(v_{k+1} + \sqrt{s}\nabla f(x_{k+1}))\big\rangle\\ &\quad - \frac{1}{2}\left\|2(x_{k+2} - x_{k+1}) + \sqrt{s}(k+2)(v_{k+1} + \sqrt{s}\nabla f(x_{k+1}))\right.\\ &\quad \left.-(k+1)\sqrt{s}(v_k + \sqrt{s}\nabla f(x_k))\right\|^2\\ &= s(k+3)(k+1)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right)\\ &\quad + \left\langle -s(k+3)\nabla f(x_{k+1}), 2(x_{k+2} - x^\star) + \sqrt{s}(k+2)(v_{k+1} + \sqrt{s}\nabla f(x_{k+1}))\right\rangle\\ &\quad - \frac{1}{2}\left\|s(k+3)\nabla f(x_{k+1})\right\|^2\end{aligned}$$

$$
\begin{aligned}
&= s(k+3)(k+1)\left(f(x_{k+1}) - f(x_k)\right) + s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right) \\
&\quad - s^{\frac{3}{2}}(k+3)(k+4)\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle - 2s(k+3)\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle \\
&\quad - s^2(k+3)(k+2)\left\|\nabla f(x_{k+1})\right\|^2 - \frac{s^2}{2}(k+3)^2\left\|\nabla f(x_{k+1})\right\|^2,
\end{aligned}
$$

where the last two equalities are due to

$$
(k+3)\left(v_k + \sqrt{s}\nabla f(x_k)\right) - k\left(v_{k-1} + \sqrt{s}\nabla f(x_{k-1})\right) = -k\sqrt{s}\nabla f(x_k), \tag{C.65}
$$

which follows from the phase-space representation (4.40). Rearranging the identity for $\mathcal{E}(k+1) - \mathcal{E}(k)$, we get

$$
\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) &= s(k+3)(k+1)\left(f(x_{k+1}) - f(x_k)\right) \\
&\quad - s^{\frac{3}{2}}(k+3)(k+4)\left\langle \nabla f(x_{k+1}), v_{k+1}\right\rangle \\
&\quad + s(2k+5)\left(f(x_{k+1}) - f(x^\star)\right) \\
&\quad - s(2k+6)\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle \\
&\quad - \frac{s^2(k+3)(3k+7)}{2}\left\|\nabla f(x_{k+1})\right\|^2.
\end{aligned} \tag{C.66}
$$

The next step is to recognize that the convexity and the L-smoothness of f gives

$$
\begin{aligned}
f(x_{k+1}) - f(x_k) &\le \left\langle \nabla f(x_{k+1}), x_{k+1} - x_k\right\rangle - \frac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
f(x_{k+1}) - f(x^\star) &\le \left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle.
\end{aligned}
$$

Plugging these two inequalities into (C.66), we have

$$
\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) &\le -s^{\frac{3}{2}}(k+3)\left\langle \nabla f(x_{k+1}), (k+4)v_{k+1} - (k+1)v_k\right\rangle \\
&\quad - \frac{s}{2L}(k+3)(k+1)\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&\quad - s\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle \\
&\quad - \frac{s^2(k+3)(3k+7)}{2}\left\|\nabla f(x_{k+1})\right\|^2 \\
&\le -s^{\frac{3}{2}}(k+3)\left\langle \nabla f(x_{k+1}), (k+4)v_{k+1} - (k+1)v_k\right\rangle \\
&\quad - \frac{s}{2L}(k+3)(k+1)\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\
&\quad - \frac{s^2(k+3)(3k+7)}{2}\left\|\nabla f(x_{k+1})\right\|^2,
\end{aligned}
$$

where the second inequality uses the fact that $\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle \ge 0$.

To further bound $\mathcal{E}(k+1)-\mathcal{E}(k)$, making use of (C.65) with $k+1$ in place of k, we get

$$\begin{aligned}
\mathcal{E}(k+1)-\mathcal{E}(k) &\le s^2(k+3)(k+1)\,\langle \nabla f(x_{k+1}), \nabla f(x_{k+1})-\nabla f(x_k)\rangle \\
&\quad -\frac{s}{2L}(k+3)(k+1)\,\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\
&\quad -s^2\left(\frac{(k+3)(3k+7)}{2}-(k+3)(k+4)\right)\|\nabla f(x_{k+1})\|^2 \\
&= \frac{Ls^3(k+3)(k+1)}{2}\|\nabla f(x_{k+1})\|^2-\frac{s(k+3)(k+1)}{2L} \\
&\quad \|(1-Ls)\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\
&\quad -\frac{s^2(k+3)(k-1)}{2}\|\nabla f(x_{k+1})\|^2 \\
&\le -\frac{s^2}{2}\left[(k+3)(k-1)-Ls(k+3)(k+1)\right]\|\nabla f(x_{k+1})\|^2 .
\end{aligned}$$

This completes the proof. □

C Nesterov's lower bound

Recall [36, Theorem 2.1.7], for any k, $1\le k\le (1/2)(n-1)$, and any $x_0\in\mathbb{R}^n$, there exists a function $f\in\mathcal{F}_L^1(\mathbb{R}^n)$ such that any first-order method obeys

$$f(x_k)-f(x^\star)\ge \frac{3L\,\|x_0-x^\star\|^2}{32(k+1)^2}.$$

Using the basic inequality for $f(x)\in\mathcal{F}_L^1(\mathbb{R}^n)$,

$$\|\nabla f(x_k)\|\,\|x_k-x^\star\|\ge \langle \nabla f(x_k), x_k-x^\star\rangle \ge f(x_k)-f(x^\star),$$

we have

$$\|\nabla f(x_k)\|\ge \frac{3L\,\|x_0-x^\star\|^2}{32(k+1)^2\max\limits_{1\le k\le\frac{n-1}{2}}\|x_k-x^\star\|},$$

for $1\le k\le (1/2)(n-1)$.

D Technical details in Sect. 5

D.1 Proof of Theorem 7: Case $\alpha = 3$

Before starting to prove Theorem 7, we first look back our high-resolution ODE framework in Sect. 2.

- *Step* 1, the generalized high-resolution ODE has been given in (5.48).
- *Step* 2, the continuous Lyapunov function is constructed as

$$\begin{aligned}\mathcal{E}(t) = t\left[t + \left(\frac{3}{2} - \beta\right)\sqrt{s}\right]\left(f(X(t)) - f(x^\star)\right) \\ + \frac{1}{2}\left\|2(X(t) - x^\star) + t\left(\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right)\right\|^2.\end{aligned} \tag{D.67}$$

  Following this Lyapunov function (D.67), we can definitely obtain similar results as Theorem 5 and Corollary 1. The detailed calculation, about the estimate of the optimal constant β and how the constant β influence the initial point, is left for readers.
- *Step* 3, before constructing discrete Lyapunov functions, we show the phase-space representation (5.49) as

$$\begin{aligned}x_k - x_{k-1} &= \sqrt{s}v_{k-1} \\ v_k - v_{k-1} &= -\frac{\alpha}{k}v_k - \beta\sqrt{s}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) \\ &\quad - \left(1 + \frac{\alpha}{k}\right)\sqrt{s}\nabla f(x_k).\end{aligned} \tag{D.68}$$

Now, we show how to construct the discrete Lyapunov function and analyze the algorithms (5.49) with $\alpha = 3$ in order to prove Theorem 7.

D.1.1 Case: $\beta < 1$

When $\beta < 1$, we know that the function $g(k) = \frac{k+3}{k+3-\beta}$ decreases monotonically. Hence we can construct the discrete Lyapunov function as

$$\begin{aligned}\mathcal{E}(k) \quad &= s(k+4)(k+1)\left(f(x_k) - f(x^\star)\right) \\ &+ \tfrac{k+3}{2(k+3-\beta)}\left\|2(x_{k+1} - x^\star) + \sqrt{s}(k+1)\right. \\ &\qquad \left.\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\right\|^2,\end{aligned} \tag{D.69}$$

which is slightly different from the discrete Lyapunov function (4.41) for NAG-C. When $\beta \to 1$, the discrete Lyapunov function (D.69) approximate to (4.41) as $k \to \infty$.

With the phase-space representation (D.68) for $\alpha = 3$, we can obtain

$$(k+3)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right) - k\left(v_{k-1} + \beta\sqrt{s}\nabla f(x_{k-1})\right)$$

$$= -\sqrt{s}\,(k+3-3\beta)\,\nabla f(x_k). \tag{D.70}$$

The difference of the discrete Lyapunov function (D.69) of the k-th iteration is

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \le\ & s\,(k+4)\,(k+1)\,(f(x_{k+1})-f(x_k))\\ &+s(2k+6)\left(f(x_{k+1})-f(x^\star)\right)\\ &-\left\langle s(k+4)\nabla f(x_{k+1}),\,2(x_{k+2}-x^\star)\right.\\ &\left.+\sqrt{s}(k+2)\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\right\rangle\\ &-\frac{1}{2}s^2(k+4)\,(k+4-\beta)\,\|\nabla f(x_{k+1})\|^2.\end{aligned}$$

With the basic inequality of any function $f(x)\in\mathcal{F}_L^1(\mathbb{R}^n)$

$$\begin{cases} f(x_k)\ge f(x_{k+1})+\langle\nabla f(x_{k+1}),x_k-x_{k+1}\rangle+\dfrac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\\ f(x^\star)\ge f(x_{k+1})+\left\langle\nabla f(x_{k+1}),x^\star-x_{k+1}\right\rangle,\end{cases}$$

and the phase-space representation (D.68) $x_{k+2}=x_{k+1}+\sqrt{s}v_{k+1}$, the difference of the discrete Lyapunov function (D.69) can be estimated as

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \le\ & -s^{\frac{3}{2}}(k+4)\,\langle\nabla f(x_{k+1}),\,(k+4)v_{k+1}-(k+1)v_k\rangle\\ &-\frac{s(k+4)(k+1)}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\\ &-2s\left(f(x_{k+1})-f(x^\star)\right)-s^2\left[\beta(k+4)(k+2)\right.\\ &\left.+\frac{1}{2}(k+4)\,(k+4-\beta)\right]\|\nabla f(x_{k+1})\|^2.\end{aligned}$$

Utilizing the phase-space representation (D.68) again, we calculate the difference of the discrete Lyapunov function (D.69) as

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \le -\Bigg[&\beta(k+2)-\frac{1}{2}\,(k+4+\beta)\\ &-\frac{L\beta^2 s}{2}(k+1)\Bigg](k+4)s^2\,\|\nabla f(x_{k+1})\|^2.\end{aligned}$$

To guarantee that the Lyapunov function $\mathcal{E}(k)$ is decreasing, a sufficient condition is

$$\beta(k+2)-\frac{1}{2}\,(k+4+\beta)-\frac{L\beta^2 s}{2}(k+1)\ge 0. \tag{D.71}$$

Simple calculation tells us that (D.71) can be rewritten as

$$s\le\frac{(2\beta-1)k+3\beta-4}{(k+1)L\beta^2}=\frac{1}{L\beta^2}\left(2\beta-1+\frac{\beta-3}{k+1}\right). \tag{D.72}$$

Apparently, when $\beta \to 1$, the step size satisfies $0 < s \leq \frac{k-1}{k+1} \cdot \frac{1}{L}$, which is consistent with (4.43). Now, we turn to discuss the parameter $0 \leq \beta < 1$ case by case.

- When the parameter $\beta \leq 1/2$, the sufficient condition (D.71) for the Lyapunov function $\mathcal{E}(k)$ decreasing cannot be satisfied for sufficiently large k.
- When the parameter $1/2 < \beta < 1$, since the function $h(k) = \frac{1}{L\beta^2}\left(2\beta - 1 + \frac{\beta-3}{k+1}\right)$ increases monotonically for $k \geq 0$, there exists $k_{3,\beta} = \left\lfloor \frac{4-3\beta}{2\beta-1} \right\rfloor + 1$ such that the step size $s \leq \frac{(2\beta-1)k_{3,\beta}+3\beta-4}{(k_{3,\beta}+1)L\beta^2}$ works for any $k \geq k_{3,\beta}$ ($k_{3,\beta} \to 2$ with $\beta \to 1$). Then, the difference of the discrete Lyapunov function (D.69) can be estimated as

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -s^2\left(\frac{2\beta - 1 - L\beta^2 s}{2}\right)(k - k_{3,\beta})^2 \left\|\nabla f(x_{k+1})\right\|^2.$$

  Here, the proof is actually complete. Without loss of generality, we briefly show the expression is consistent with Theorem 7 and omit the proofs for the following facts. When $k \geq k_{3,\beta} + 1$, there exists some constant $\mathfrak{C}^0_{3,\beta} > 0$ such that

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -s^2\mathfrak{C}^0_{3,\beta}(k+1)^2 \left\|\nabla f(x_{k+1})\right\|^2.$$

  For $k \leq k_{3,\beta}$, using mathematic induction, there also exists some constant $\mathfrak{C}^1_{3,\beta} > 0$ such that for $s = O(1/L)$, we have

$$\begin{aligned}\left\|\nabla f(x_{k+1})\right\|^2 &\leq \frac{\mathfrak{C}^1_{3,\beta}\left\|x_0 - x^\star\right\|^2}{s^2} \quad \text{and} \quad f(x_k) - f(x^\star)\\ &\leq \frac{\mathcal{E}(k)}{4s} \leq \frac{\mathfrak{C}^1_{3,\beta}\left\|x_0 - x^\star\right\|^2}{s}.\end{aligned}$$

D.1.2 Case: $\beta \geq 1$

When $\beta \geq 1$, we know that the function $g(k) = \frac{k+2}{k+3-\beta}$ decreases monotonically. Hence we can construct the discrete Lyapunov function as

$$\begin{aligned}\mathcal{E}(k) = s(k+3)(k+1)\left(f(x_k) - f(x^\star)\right) + \frac{k+2}{2(k+3-\beta)}\Big\|2(x_{k+1} - x^\star)\\ + \sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\Big\|^2,\end{aligned} \tag{D.73}$$

which for $\beta = 1$ is consistent with the discrete Lyapunov function (4.41) for NAG-C.

With the expression (D.70)

$$\begin{aligned}&(k+3)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right) - k\left(v_{k-1} + \beta\sqrt{s}\nabla f(x_{k-1})\right)\\ &= -\sqrt{s}\,(k + 3 - 3\beta)\,\nabla f(x_k),\end{aligned}$$

the difference of the discrete Lyapunov function (D.73) of the k-th iteration is

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \leq\ & s\,(k+3)\,(k+1)\,(f(x_{k+1})-f(x_k))+s(2k+5)\\ & \left(f(x_{k+1})-f(x^\star)\right)-\frac{1}{2}s^2(k+3)\,(k+4-\beta)\\ & \|\nabla f(x_{k+1})\|^2\\ & -\left\langle s(k+3)\nabla f(x_{k+1}),\,2(x_{k+2}-x^\star)\right.\\ & \left.+\sqrt{s}(k+2)\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\right\rangle.\end{aligned}$$

With the basic inequality of any function $f(x)\in\mathcal{F}_L^1(\mathbb{R}^n)$

$$\begin{cases} f(x_k)\geq f(x_{k+1})+\langle\nabla f(x_{k+1}),x_k-x_{k+1}\rangle+\dfrac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\\ f(x^\star)\geq f(x_{k+1})+\left\langle\nabla f(x_{k+1}),x^\star-x_{k+1}\right\rangle,\end{cases}$$

and the phase-space representation (D.68) $x_{k+2}=x_{k+1}+\sqrt{s}v_{k+1}$, the difference of the discrete Lyapunov function (D.73) can be estimated as

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k)\leq\ & -s^{\frac{3}{2}}(k+3)\,\langle\nabla f(x_{k+1}),\,(k+4)v_{k+1}-(k+1)v_k\rangle\\ & -\frac{s(k+3)(k+1)}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\\ & -2s\left(f(x_{k+1})-f(x^\star)\right)\\ & -s^2\left[\beta(k+3)(k+2)+\frac{1}{2}(k+3)\,(k+4-\beta)\right]\\ & \|\nabla f(x_{k+1})\|^2.\end{aligned}$$

Utilize the phase-space representation (D.68) again, we calculate the difference of the discrete Lyapunov function (D.73) as

$$\begin{aligned}&\mathcal{E}(k+1)-\mathcal{E}(k)\\ &\leq-\left[\beta(k+2)-\frac{1}{2}\,(k+4+\beta)-\frac{L\beta^2 s}{2}(k+1)\right](k+3)s^2\,\|\nabla f(x_{k+1})\|^2.\end{aligned}$$

Consistently, we can obtain the sufficient condition for the Lyapunov function $\mathcal{E}(k)$ decreasing (D.71) and the sufficient condition for step size (D.72).

Now, we turn to discuss the parameter $\beta\geq 1$ case by case.

– When the parameter $\beta\geq 3$, since the function $h(k)=\frac{1}{L\beta^2}\left(2\beta-1+\frac{\beta-3}{k+1}\right)$ decreases monotonically for $k\geq 0$, then the condition of the step size $s\leq\frac{2\beta-1}{(1+\epsilon)L\beta^2}<\frac{2\beta-1}{L\beta^2}$ holds for (D.71), where $\epsilon>0$ is a real number. Hence, when $k\geq k_{3,\beta}+1$, where $k_{3,\beta}=\max\left\{0,\lfloor\beta-3\rfloor+1,\left\lfloor\frac{4-3\beta+L\beta^2 s}{2\beta-1-L\beta^2 s}\right\rfloor+1\right\}$, the differ-

ence of the discrete Lyapunov function (D.73) can be estimated as

$$\mathcal{E}(k+1)-\mathcal{E}(k)\leq -s^2\left(\frac{2\beta-1-L\beta^2 s}{2}\right)(k-k_{3,\beta})^2\left\|\nabla f(x_{k+1})\right\|^2.$$

– When the parameter $1\leq\beta<3$, since the function $h(k)=\frac{1}{L\beta^2}\left(2\beta-1+\frac{\beta-3}{k+1}\right)$ increases monotonically for $k\geq 0$, there exists $k_{3,\beta}=\max\left\{0,\lfloor\beta-3\rfloor+1,\left\lfloor\frac{4-3\beta}{2\beta-1}\right\rfloor+1\right\}$ such that the step size $s\leq\frac{(2\beta-1)k_{3,\beta}+3\beta-4}{(k_{3,\beta}+1)L\beta^2}$ works for any $k\geq k_{3,\beta}$. When $\beta=1$, the step size satisfies $0<s\leq\frac{k-1}{k+1}\cdot\frac{1}{L}$ which is consistent with (4.43) and $k_{3,\beta}=2$. Then, the difference of the discrete Lyapunov function (D.69) can be estimated as

$$\mathcal{E}(k+1)-\mathcal{E}(k)\leq -s^2\left(\frac{2\beta-1-L\beta^2 s}{2}\right)(k-k_{3,\beta})^2\left\|\nabla f(x_{k+1})\right\|^2.$$

for all $k\geq k_{3,\beta}+1$.

By simple calculation, we complete the proof.

D Proof of Theorem 7: Case $\alpha > 3$

Before starting to prove Theorem 7: Case $\alpha>3$, we first also look back our high-resolution ODE framework in Sect. 2.

– *Step* 1, the generalized high-resolution ODE has been given in (5.48).
– *Step* 2, the continuous Lyapunov function is constructed as

$$\begin{aligned}\mathcal{E}(t)=&\,t\left[t+\left(\frac{\alpha}{2}-\beta\right)\sqrt{s}\right]\left(f(X(t))-f(x^\star)\right)\\&+\frac{1}{2}\left\|(\alpha-1)(X(t)-x^\star)+t\left(\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right)\right\|^2,\end{aligned}\tag{D.74}$$

which is consistent with (D.74) for $\alpha\to 3$. Following this Lyapunov function (D.74), we can obtain

$$\begin{aligned}&f(X(t))-f(x^\star)\leq O\left(\frac{\|X(t_0)-x^\star\|^2}{(t-t_0)^2}\right)\\&\int_{t_0}^{t}u\left(f(X(u))-f(x^\star)\right)+\sqrt{s}u^2\left\|\nabla f(X(u))\right\|^2du\leq O\left(\|X(t_0)-x^\star\|^2\right)\end{aligned}\tag{D.75}$$

for any $t>t_0=\max\left\{\sqrt{s}(\alpha/2-\beta)(\alpha-2)/(\alpha-3),\sqrt{s}(\alpha/2)\right\}$. The two inequalities of (D.75) for the convergence rate of function value is stronger than Corollary 1. The detailed calculation, about the estimate of the optimal constant β and how the constant β influences the initial point, is left for readers.

– *Step* 3, before constructing discrete Lyapunov functions, we look back the phase-space representation (D.68)

$$\begin{aligned} x_k - x_{k-1} &= \sqrt{s} v_{k-1} \\ v_k - v_{k-1} &= -\frac{\alpha}{k} v_k - \beta\sqrt{s}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \left(1 + \frac{\alpha}{k}\right)\sqrt{s}\nabla f(x_k). \end{aligned}$$

The discrete functional is constructed as

$$\begin{aligned} \mathcal{E}(k) = & s(k+1)(k+\alpha-\beta+1)\left(f(x_k) - f(x^\star)\right) \\ & + \frac{1}{2}\left\|(\alpha-1)(x_{k+1} - x^\star) + \sqrt{s}(k+1)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right)\right\|^2. \end{aligned} \tag{D.76}$$

When $\beta = 1$, with $\alpha \to 3$, the discrete Lyapunov function $\mathcal{E}(k)$ degenerates to (4.41).

Now, we procced to *Step* 4 to analyze the algorithms (5.49) with $\alpha > 3$ in order to prove Theorem 3. The simple transformation of (D.68) for $\alpha > 3$ is

$$\begin{aligned} & (k+\alpha)\left(v_k + \beta\sqrt{s}\nabla f(x_k)\right) - k\left(v_{k-1} + \beta\sqrt{s}\nabla f(x_{k-1})\right) \\ & \quad = -\sqrt{s}\left(k + \gamma - \gamma\beta\right)\nabla f(x_k). \end{aligned} \tag{D.77}$$

Thus, the difference of the Lyapunov function (D.76) on the k-th iteration is

$$\begin{aligned} \mathcal{E}(k+1) - \mathcal{E}(k) = & s(k+1)\left(k+\alpha-\beta+1\right)\left(f(x_{k+1}) - f(x_k)\right) \\ & + s\left(2k+\alpha-\beta+3\right)\left(f(x_{k+1}) - f(x^\star)\right) \\ & - \Big\langle s\left(k+\alpha-\beta+1\right)\nabla f(x_{k+1}), (\alpha-1)(x_{k+1} - x^\star) \\ & + \sqrt{s}(k+\alpha+1)v_{k+1} + \beta s(k+2)\nabla f(x_{k+1})\Big\rangle \\ & - \frac{1}{2}s^2(k+\alpha-\beta+1)^2\left\|\nabla f(x_{k+1})\right\|^2. \end{aligned}$$

With the basic inequality of convex function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$,

$$\begin{cases} f(x_k) \geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle + \dfrac{1}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|^2 \\ f(x^\star) \geq f(x_{k+1}) + \left\langle \nabla f(x_{k+1}), x^\star - x_{k+1}\right\rangle \end{cases}$$

and the phase-space representation (D.68) $x_{k+2} = x_{k+1} + \sqrt{s}v_{k+1}$, the difference of the discrete Lyapunov function (D.76) can be estimated as

$$\begin{aligned} \mathcal{E}(k+1) - \mathcal{E}(k) \leq & \left(k+\alpha-\beta+1\right)\Big[-s^{\frac{3}{2}}\langle \nabla f(x_{k+1}), (k+\alpha+1)v_{k+1} \\ & -(k+1)v_k\rangle - \frac{s(k+1)}{2L}\left\|\nabla f(x_{k+1}) - \nabla f(x_k)\right\|_2^2\Big] \\ & - s\left[(\alpha-3)k + (\alpha-2)\left(\alpha-\beta+1\right) - 2\right]\left(f(x_{k+1}) - f(x^\star)\right) \end{aligned}$$

$$-\frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta+1)k+\alpha+3\beta+1\right]\\ \|\nabla f(x_{k+1})\|^2.$$

Utilizing the phase-space representation (D.68) again, we calculate the difference of the discrete Lyapunov function (D.76) as

$$\mathcal{E}(k+1)-\mathcal{E}(k)\le -s\left[(\alpha-3)k+(\alpha-2)(\alpha-\beta+1)-2\right]\left(f(x_{k+1})-f(x^\star)\right)\\ -\frac{1}{2}s^2(k+\alpha-\beta+1)\left[(2\beta-1)k-\alpha+3\beta-1-L\beta^2s(k+1)\right]\\ \|\nabla f(x_{k+1})\|^2.$$

To guarantee the Lyapunov function $\mathcal{E}(k)$ is decreasing, a sufficient condition is

$$(2\beta-1)k-\alpha+3\beta-1-L\beta^2s(k+1)\ge 0. \tag{D.78}$$

With the inequality (D.78), the step size can be estimated as $s\le \frac{2\beta-1}{L\beta^2}-\frac{\alpha-\beta}{(k+1)L\beta^2}$.

– When the parameter $\beta>1/2$ and $\alpha<\beta$, since the function $h(k)=\frac{2\beta-1}{L\beta^2}-\frac{\alpha-\beta}{(k+1)L\beta^2}$ decreases monotonically for $k\ge 0$, thus the step size $s\le\frac{2\beta-1}{(1+\epsilon)L\beta^2}<\frac{2\beta-1}{L\beta^2}$ holds for (D.78), where $\epsilon>0$ is a real number. Hence, when $k\ge k_{\alpha,\beta}+1$, where

$$k_{\alpha,\beta}=\max\left\{0,\left\lfloor\frac{2-(\alpha-2)(\alpha-\beta+1)}{\alpha-3}\right\rfloor+1,\left\lfloor\frac{4-3\beta+L\beta^2s}{-1+2\beta-L\beta^2s}\right\rfloor\right.\\ \left.+1,\lfloor\beta-\alpha-1\rfloor+1\right\},$$

the difference of the discrete Lyapunov function (D.76) can be estimated as

$$\mathcal{E}(k+1)-\mathcal{E}(k)\le -s(\alpha-3)\left(k-k_{\alpha,\beta}\right)\left(f(x_{k+1})-f(x^\star)\right)\\ -s^2\left(\frac{2\beta-1-L\beta^2s}{2}\right)\left(k-k_{\alpha,\beta}\right)^2\|\nabla f(x_{k+1})\|^2.$$

– When the parameter $\beta>1/2$ and $\alpha\ge\beta$, since the function $h(k)=\frac{2\beta-1}{L\beta^2}-\frac{\alpha-\beta}{(k+1)L\beta^2}$ increases monotonically for $k\ge 0$, there exists

$$k_{\alpha,\beta}=\max\left\{0,\left\lfloor\frac{2-(\alpha-2)(\alpha-\beta+1)}{\alpha-3}\right\rfloor+1,\lfloor\beta-\alpha-1\rfloor+1,\left\lfloor\frac{1+\alpha-3\beta}{2\beta-1}\right\rfloor+1\right\}$$

such that the step size satisfies $s\le\frac{(2\beta-1)k_{\alpha,\beta}-\alpha+3\beta-1}{L\beta^2(k_{\alpha,\beta}+1)}$. When $\beta=1$, the step size satisfies

$$s\le\frac{1}{L}\cdot\frac{k_{\alpha,\beta}-\alpha+2}{(k_{\alpha,\beta}+1)}\to\frac{1}{L}\cdot\frac{k_{\alpha,\beta}-1}{k_{\alpha,\beta}+1}\quad\text{as}\quad\alpha\to 3,$$

which is consistent with (4.43). Then, the difference of the discrete Lyapunov function (D.76) can be estimated as

$$\begin{aligned}\mathcal{E}(k+1)-\mathcal{E}(k) \leq & -s(\alpha-3)\left(k-k_{\alpha,\beta}\right)\left(f(x_{k+1})-f(x^\star)\right)\\ & -s^2\left(\frac{2\beta-1-L\beta^2 s}{2}\right)\left(k-k_{\alpha,\beta}\right)^2\left\|\nabla f(x_{k+1})\right\|^2.\end{aligned}$$

D.1 A simple counterexample

The simple counterexample is constructed as

$$f(x_k)-f(x^\star)=\begin{cases}\dfrac{L\left\|x_0-x^\star\right\|^2}{(k+1)^2}, & k=j^2\\ 0, & k\neq j^2\end{cases}$$

where $j \in \mathbb{N}$. Plugging it into (5.51), we have $\sum_{k=0}^{\infty}(k+1)\left(f(x_k)-f(x^\star)\right) = L\left\|x_0-x^\star\right\|^2\cdot\sum_{j=0}^{\infty}\left(\frac{1}{j^2+1}\right)<\infty$. Hence, Proposition 3 cannot guarantee the faster convergence rate.

D.2 Super-critical regime: sharper convergence rate $o(1/t^2)$ and $o(L/k^2)$

D.2.1 The ODE case

Here, we still turn back to our high-resolution ODE framework in Sect. 2. The generalized high-resolution ODE has been still shown in (5.48). A more general Lyapunov function is constructed as

$$\begin{aligned}\mathcal{E}_\nu(t)= & t\left[t+\left(\frac{\alpha}{2}-\beta\right)\sqrt{s}+(\alpha-\nu-1)\beta\sqrt{s}\right]\left(f(X(t))-f(x^\star)\right)\\ & +\frac{\nu(\alpha-\nu-1)}{2}\left\|X(t)-x^\star\right\|^2+\frac{1}{2}\left\|\nu(X(t)-x^\star)+t\left(\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right)\right\|^2,\end{aligned}\tag{D.79}$$

where $2<\nu\leq\alpha-1$. When $\nu=\alpha-1$, the Lyapunov function (D.79) degenerates to (D.74). Furthermore, when $\nu=\alpha-1\to 2$, the Lyapunov function (D.79) degenerates to (D.67). Finally, when $2=\nu=\alpha-1$ and $\beta=1$, the Lyapunov function (D.79) is consistent with (4.36). We assume that initial time is $t_{\alpha,\beta,\nu}=\max\left\{\sqrt{s}\left(\beta-\frac{\alpha}{2}\right),\sqrt{s}\left(\frac{\beta(\alpha-2)}{\nu-2}-\frac{\alpha(\nu-1)}{2(\nu-2)}\right),\frac{\sqrt{s}\alpha}{2}\right\}$. Based on the Lyapunov function (D.79), we have the following results.

Theorem 9 *Let $f(x) \in \mathcal{F}_L^2(\mathbb{R}^n)$ and $X = X(t)$ be the solution of the ODE (5.48) with $\alpha > 3$ and $\beta > 0$. Then, there exists $t_{\alpha,\beta,v} > 0$ such that*

$$\begin{cases} \lim\limits_{t\to\infty} t^2 \left(\left(f(X(t)) - f(x^\star) \right) + \left\| \dot{X}(t) + \beta\sqrt{s}\nabla f(X(t)) \right\|^2 \right) = \mathfrak{C}^2_{\alpha,\beta,v} \left\| x_0 - x^\star \right\|^2 \\ \int_{t_0}^{t} \left[u \left(f(X(u)) - f(x^\star) \right) + u \left\| \dot{X}(u) + \beta\sqrt{s}\nabla f(X(u)) \right\|^2 \right] \mathrm{d}u < \infty, \end{cases} \tag{D.80}$$

for all $t \geq t_{\alpha,\beta,v}$, where the positive constant $\mathfrak{C}^2_{\alpha,\beta,v}$ and the integer $t_{\alpha,\beta,v}$ depend only on α, β and v. In other words, the equivalent expression of (D.80) is $f(X(t)) - f(x^\star) + \left\| \dot{X}(t) + \beta\sqrt{s}\nabla f(X(t)) \right\|^2 \leq o\left(\frac{\|x_0 - x^\star\|^2}{t^2} \right)$.

Now, we start to show the proof. Since $X = X(t)$ is the solution of the ODE (5.48) with $\alpha > 3$ and $\beta > 0$, when $t > t_{\alpha,\beta,v}$, the time derivative of Lyapunov function (D.79) is

$$\begin{aligned} \frac{\mathrm{d}\mathcal{E}_v(t)}{\mathrm{d}t} = & \left[2t + \left(\frac{\alpha}{2} - \beta \right) \sqrt{s} + (\alpha - v - 1)\beta\sqrt{s} \right] \\ & \left(f(X(t)) - f(x^\star) \right) - (\alpha - 1 - v)t \left\| \dot{X}(t) \right\|^2 \\ & - v \left[t + \left(\frac{\alpha}{2} - \beta \right) \sqrt{s} \right] \left\langle \nabla f(X(t)), X(t) - x^\star \right\rangle \\ & - \beta t\sqrt{s} \left[t + \left(\frac{\alpha}{2} - \beta \right) \sqrt{s} \right] \left\| \nabla f(X(t)) \right\|^2. \end{aligned} \tag{D.81}$$

With the basic inequality for any $f(x) \in \mathcal{F}_L^2(\mathbb{R}^n)$ $f(x^\star) \geq f(X(t)) + \langle \nabla f(X(t)), x^\star - X(t) \rangle$, the time derivative of Lyapunov function (D.81) can be estimated as

$$\begin{aligned} \frac{\mathrm{d}\mathcal{E}_v(t)}{\mathrm{d}t} \leq & - \left\{ (v-2)t + \sqrt{s} \left[\frac{\alpha(v-1)}{2} - (\alpha - 2)\beta \right] \right\} \left(f(X(t)) - f(x^\star) \right) \\ & -(\alpha - 1 - v)t \left\| \dot{X}(t) \right\|^2 - \beta t\sqrt{s} \left[t + \left(\frac{\alpha}{2} - \beta \right) \sqrt{s} \right] \left\| \nabla f(X(t)) \right\|^2. \end{aligned}$$

With the Lyapunov function $\mathcal{E}_v(t) \geq 0$, for any $t > t_0$ we have

$$\begin{aligned} \int_{t_0}^{t} u(f(X(u)) - f(x^\star))du \leq & \int_{t_0}^{t_0+\delta} u(f(X(u)) - f(x^\star))du + \left(1 + \frac{t_0}{\delta} \right) \\ & \int_{t_0+\delta}^{t} (u - t_0)(f(X(u)) - f(x^\star))du, \end{aligned}$$

where $\delta < t - t_0$. Thus, we can obtain the following lemma.

Lemma 7 *Under the same assumption of Theorem 9, the following limits exist*

$$\lim_{t\to\infty} \mathcal{E}_v(t), \lim_{t\to\infty} \int_{t_0}^{t} u(f(X(u))$$

$$-f(x^\star))\mathrm{d}u,\ \lim_{t\to\infty}\int_{t_0}^{t} u\left\|\dot{X}(u)\right\|^2\mathrm{d}u,\ \lim_{t\to\infty}\int_{t_0}^{t} u^2\left\|\nabla f(X(u))\right\|^2 du.$$

With (D.81) and Lemma 7, the following lemma holds.

Lemma 8 *Under the same assumption of Theorem 9, the following limit exists* $\lim_{t\to\infty}\int_{t_0}^{t} u\left\langle\nabla f(X(u)), X(u)-x^\star\right\rangle du$.

Lemma 9 *Under the same assumption of Theorem 9, the following limits exist*

$$\lim_{t\to\infty}\left\|X(t)-x^\star\right\| \quad \textit{and}$$
$$\lim_{t\to\infty} t\left\langle X(t)-x^\star, \dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\rangle.$$

Proof of Lemma 9 Taking $\nu\neq\nu'\in[2,\gamma-1]$, we have

$$\begin{aligned}\mathcal{E}_\nu(t)-\mathcal{E}_{\nu'}(t)=(\nu-\nu')\Big[&-\beta\sqrt{s}t\left(f(X(t))-f(x^\star)\right)\\&+t\left\langle X(t)-x^\star,\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\rangle+\frac{\alpha-1}{2}\left\|X(t)-x^\star\right\|^2\Big].\end{aligned}$$

With Lemma 7 and (D.75), the following limit exists

$$\lim_{t\to\infty}\left[t\left\langle X(t)-x^\star,\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\rangle+\frac{\alpha-1}{2}\left\|X(t)-x^\star\right\|^2\right]. \tag{D.82}$$

Define a new function about time variable t: $\pi(t) := \frac{1}{2}\|X(t) - x^\star\|^2 + \beta\sqrt{s}\int_{t_0}^{t} \langle \nabla f(X(u)), X(u) - x^\star\rangle du$. If we can prove the existence of the limit $\pi(t)$ with $t \to \infty$, we can guarantee $\lim_{t\to\infty}\|X(t) - x^\star\|$ exists via Lemma 8. We observe the following equality

$$\begin{aligned} t\dot{\pi}(t) + (\alpha-1)\pi(t) = \beta(\alpha-1)\sqrt{s}\int_{t_0}^{t} & \\ \left\langle \nabla f(X(u)), X(u) - x^\star\right\rangle du & \\ + t\left\langle X(t) - x^\star, \dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right\rangle & \\ + \frac{\alpha-1}{2}\left\|X(t) - x^\star\right\|^2. & \end{aligned}$$

With (D.82) and Lemma 8, we obtain that the following limit exists: $\lim_{t\to\infty}\left[t\dot{\pi}(t) + (\alpha-1)\pi(t)\right]$; that is, there exists some constant $\mathfrak{C}^3$ such that the following equality holds: $\lim_{t\to\infty}\frac{\mathrm{d}(t^{\alpha-1}\pi(t))}{t^{\alpha-2}} = \lim_{t\to\infty}\left[t\dot{\pi}(t) + (\alpha-1)\pi(t)\right] = \mathfrak{C}^3$. For any $\epsilon > 0$, there exists $t_0 > 0$ such that when $t \geq t_0$, we have

$$t^{\alpha-1}\left(\pi(t) - \frac{\mathfrak{C}^3}{\alpha-1}\right) - t_0^{\alpha-1}\left(\pi(t_0) - \frac{\mathfrak{C}^3}{\alpha-1}\right) \leq \frac{\epsilon}{\alpha-1}\cdot\left(t^{\alpha-1} - t_0^{\alpha-1}\right);$$

that is,

$$\left|\pi(t) - \frac{\mathfrak{C}^3}{\alpha-1}\right| \leq \left|\pi(t_0) - \frac{\mathfrak{C}^3}{\alpha-1}\right|\left(\frac{t_0}{t}\right)^{\alpha-1} + \frac{\epsilon}{\alpha-1}.$$

The proof is complete. □

Finally, we finish the proof for Theorem 9.

Proof of Theorem 9 When $t > t_{\alpha,\beta,\nu}$, we expand the Lyapunov function (D.79) as

$$\begin{aligned} \mathcal{E}_\nu(t) = t\Big[t + \Big(\frac{\alpha}{2} - \beta\Big) & \\ \sqrt{s} + (\alpha - \nu - 1)\beta\sqrt{s}\Big] & \\ \big(f(X(t)) - f(x^\star)\big) & \\ + \frac{\nu(\alpha-1)}{2}\left\|X(t) - x^\star\right\|^2 & \\ + \frac{t^2}{2}\left\|\dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right\|^2 & \\ + t\left\langle X(t) - x^\star, \dot{X}(t) + \beta\sqrt{s}\nabla f(X(t))\right\rangle. & \end{aligned}$$

With Lemmas 7 and 9, we obtain the first equation of (D.80). Furthermore, the Cauchy-Schwarz inequality gives

$$
\begin{aligned}
&\left[t+\left(\frac{\alpha}{2}-\beta\right)\sqrt{s}+(\alpha-\nu-1)\beta\sqrt{s}\right]\left(f(X(t))-f(x^\star)\right)\\
&+\frac{t}{2}\left\|\dot{X}(t)+\beta\sqrt{s}\nabla f(X(t))\right\|^2\\
&\le\left[t+\left(\frac{\alpha}{2}-\beta\right)\sqrt{s}\right.\\
&\left.+(\alpha-\nu-1)\beta\sqrt{s}\right]\left(f(X(t))-f(x^\star)\right)+t\left\|\dot{X}(t)\right\|^2\\
&+\beta^2 st\left\|\nabla f(X(t))\right\|^2.
\end{aligned}
$$

With Lemma 7, we obtain the second equation of (D.80). After a straightforward calculation, we complete the proof. □

D.2.2 Proof of Theorem 8

Similarly, under the assumption of Theorem 8, we can show a discrete version of (D.80); that is, there exists some constant $\mathfrak{C}^4_{\alpha,\beta,\nu}>0$ and $\mathfrak{c}_{\alpha,\beta,\nu}>0$ such that when the step size satisfies $0<s\le \mathfrak{c}_{\alpha,\beta,\nu}/L$, the following relationship holds:

$$
\begin{cases}
\lim\limits_{k\to\infty}(k+1)^2\left(f(x_k)-f(x^\star)+\left\|v_k+\beta\sqrt{s}\nabla f(x_k)\right\|^2\right)=\dfrac{\mathfrak{C}^4_{\alpha,\beta,\nu}\|x_0-x^\star\|^2}{s}\\
\sum\limits_{k=0}^{\infty}(k+1)\left(\left(f(x_k)-f(x^\star)\right)+\left\|v_k+\beta\sqrt{s}\nabla f(x_k)\right\|^2\right)<\infty.
\end{cases}
\tag{D.83}
$$

Thus, we obtain the sharper convergence rate as $f(x_k)-f(x^\star)+\left\|v_k+\beta\sqrt{s}\nabla f(x_k)\right\|^2\le o\left(\frac{\|x_0-x^\star\|^2}{sk^2}\right)$.

Now we show the derivation of the inequality (D.83). The discrete Lyapunov function is constructed as

$$
\begin{aligned}
\mathcal{E}(k)=&\underbrace{s(k+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left(f(x_k)-f(x^\star)\right)}_{\mathbf{I}}\\
&+\underbrace{\frac{\nu(\alpha-\nu-1)}{2}\left\|x_{k+1}-x^\star\right\|^2}_{\mathbf{II}}\\
&+\underbrace{\frac{1}{2}\left\|\nu(x_{k+1}-x^\star)+(k+1)\sqrt{s}\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right)\right\|^2}_{\mathbf{III}},
\end{aligned}
\tag{D.84}
$$

where $2\le\nu<\alpha-1$ and parts **I**, **II** and **III** are the potential, Euclidean distance and mixed energy respectively. Apparently, when $\nu=\alpha-1$, the discrete Lyapunov function (D.84) is consistent with (D.76). When $\beta=1$ and $\nu=\alpha-1\to 2$, the discrete Lyapunov function (D.84) degenerates to (4.41),

Now, we turn to estimate the difference of Lyapunov function (D.84).

– For the part **I**, the potential, we have

$$
\begin{aligned}
&s(k+2)\left[k+\alpha+2-\beta+\frac{(k+3)(\alpha-1-\nu)\beta}{k+\alpha+2}\right]\left(f(x_{k+1})-f(x^\star)\right)\\
&-s(k+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left(f(x_k)-f(x^\star)\right)\\
&\leq \underbrace{s(k+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right](f(x_{k+1})-f(x_k))}_{\mathbf{I_1}}\\
&\underbrace{+s\left[2k+\alpha+3+(2\alpha-3-2\nu)\beta\right]\left(f(x_{k+1})-f(x^\star)\right)}_{\mathbf{I_2}},
\end{aligned}
$$

where the last inequality follows from $k+\alpha+2>k+\alpha+1>k+2$.

– For the part **II**, the Euclidean distance, we have

$$
\begin{aligned}
&\frac{\nu(\alpha-\nu-1)}{2}\left\|x_{k+2}-x^\star\right\|^2-\frac{\nu(\alpha-\nu-1)}{2}\left\|x_{k+1}-x^\star\right\|^2\\
&=\underbrace{\nu(\alpha-\nu-1)\left\langle x_{k+2}-x_{k+1},x_{k+2}-x^\star\right\rangle}_{\mathbf{II}_1}\\
&\underbrace{-\frac{\nu(\alpha-\nu-1)}{2}\left\|x_{k+2}-x_{k+1}\right\|^2}_{\mathbf{II}_2}.
\end{aligned}
$$

– For the part **III**, the mixed energy, with the simple transformation (D.77) for $\alpha>3$

$$
\begin{aligned}
&(k+\alpha)\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right)-k\left(v_{k-1}+\beta\sqrt{s}\nabla f(x_{k-1})\right)\\
&=-\sqrt{s}\,(k+\gamma-\gamma\beta)\,\nabla f(x_k),
\end{aligned}
$$

we have

$$
\begin{aligned}
&\frac{1}{2}\left\|\nu(x_{k+2}-x^\star)\right.\\
&\left.+(k+2)\sqrt{s}\left(v_{k+1}+\beta\sqrt{s}\nabla f(x_{k+1})\right)\right\|^2\\
&-\frac{1}{2}\left\|\nu(x_{k+1}-x^\star)+(k+1)\sqrt{s}\left(v_k+\beta\sqrt{s}\nabla f(x_k)\right)\right\|^2\\
&=\underbrace{-\nu(\alpha-\nu-1)\left\langle x_{k+2}-x_{k+1},x_{k+2}-x^\star\right\rangle}_{\mathbf{III}_1}\\
&\underbrace{-\frac{(2k+\alpha+3-\nu)(\alpha-\nu-1)}{2}\left\|x_{k+2}-x_{k+1}\right\|^2}_{\mathbf{III}_2}\\
&\underbrace{-s(k+\alpha+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left\langle\nabla f(x_{k+1}),x_{k+2}-x_{k+1}\right\rangle}_{\mathbf{III}_3}\\
&\underbrace{-s\nu\,(k+\alpha+1-\beta)\left\langle\nabla f(x_{k+1}),x_{k+1}-x^\star\right\rangle}
\end{aligned}
$$

$$\underbrace{\mathbf{III}_4 -\frac{1}{2}s^2\left[k+\alpha+1-\beta+2(k+2)\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2}_{\mathbf{III}_5}.$$

Clearly, we see that $\mathbf{II}_1+\mathbf{III}_1=0$, and $\mathbf{II}_2+\mathbf{III}_2=-\frac{s(2k+\alpha+3)(\alpha-\nu-1)}{2}\left\|v_{k+1}\right\|^2$. Using the basic inequality for $f(x)\in\mathcal{F}_L^1(\mathbb{R}^n)$, $f(x_k)\geq f(x_{k+1})+\langle\nabla f(x_{k+1}),x_k-x_{k+1}\rangle+\frac{1}{2L}\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2$, we have

$$\begin{aligned}\mathbf{I}_1+\mathbf{III}_3+\mathbf{III}_5\leq&-s^{\frac{3}{2}}\left[k+\alpha+1-\beta\right.\\&\left.+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\langle\nabla f(x_{k+1}),(k+\alpha+1)v_{k+1}-(k+1)v_k\rangle\\&-\frac{s(k+1)}{2L}\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\\&\left\|\nabla f(x_{k+1})-\nabla f(x_k)\right\|^2\\&-\frac{1}{2}s^2\left[k+\alpha+1-\beta+2(k+2)\beta\right]\\&(k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2.\end{aligned}$$

Utilizing (D.77) again, we have

$$\begin{aligned}\mathbf{I}_1+\mathbf{III}_3+\mathbf{III}_5\leq&s^2\left[\frac{L\beta^2s}{2}(k+1)+(k+\alpha+1)\right]\\&\left[k+\alpha+1-\beta+(\alpha-1-\nu)\beta\right]\left\|\nabla f(x_{k+1})\right\|^2\\&-\frac{1}{2}s^2\left[(2\beta+1)k+\alpha+1+3\beta\right](k+\alpha+1-\beta)\left\|\nabla f(x_{k+1})\right\|^2.\end{aligned}$$

Since $\beta>1/2$, let $n\in\mathbb{N}^+$ satisfy $n=\left\lfloor\frac{2}{2\beta-1}\right\rfloor+1$. When $k\geq n(\alpha-1-\nu)\beta-(\alpha+1-\beta)$, we have

$$\begin{aligned}\mathbf{I}_1+\mathbf{III}_3+\mathbf{III}_5\leq&s^2\left[\frac{L\beta^2s}{2}(k+1)+(k+\alpha+1)\right]\\&\left[k+\alpha+1-\beta+(\alpha-1-\nu)\beta\right]\\&\left\|\nabla f(x_{k+1})\right\|^2\\&-\frac{s^2n}{2(n+1)}\cdot\left[(2\beta+1)k+\alpha+1+3\beta\right]\\&\left[k+\alpha+1-\beta+(\alpha-1-\nu)\beta\right]\left\|\nabla f(x_{k+1})\right\|^2.\end{aligned}$$

Given the monotonicity of the following function in k:

$$h(k)=\frac{\left(\frac{n(2\beta+1)}{2(n+1)}-1\right)k+\frac{n}{2(n+1)}\cdot(\alpha+1+3\beta)-\alpha-1}{\frac{L\beta^2(k+1)}{2}}$$

$$= \frac{(2\beta n - n - 2)(k+1) + (\beta - \alpha)n - 2\alpha}{L\beta^2(n+1)(k+1)},$$

we know there exist constants $\mathfrak{c}_{\alpha,\beta,\nu}$ and $k_{1,\alpha,\beta,\nu}$ such that the step size satisfies $0 < s \le \mathfrak{c}_{\alpha,\beta,\nu}/L$. When $k \ge k_{1,\alpha,\beta,\nu}$, the following inequality holds:

$$\mathbf{I}_1 + \mathbf{III}_3 + \mathbf{III}_5 \le -\frac{s^2}{2}\left(\frac{2\beta n}{n+1} - \frac{n+2}{n+1} - L\beta^2 s\right)(k - k_{1,\alpha,\beta,\nu})^2 \left\|\nabla f(x_{k+1})\right\|^2.$$

With the basic inequality for $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$, $f(x^\star) \ge f(x_{k+1}) + \langle \nabla f(x_{k+1}), x^\star - x_{k+1}\rangle$, we know that there exists $k_{2,\alpha,\beta,\nu}$ such that when $k \ge k_{2,\alpha,\beta,\nu}$, $\mathbf{I}_2 + \mathbf{III}_4 \le -s(\nu-2)(k - k_{2,\alpha,\beta,\nu})\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle$.

Let $k_{\alpha,\beta,\nu} = \max\{k_{1,\alpha,\beta,\nu}, k_{2,\alpha,\beta,\nu}\} + 1$. Summing up all the estimates above, when $\beta > 1/2$, the difference of discrete Lyapunov function satisfies, for any $k \ge k_{\alpha,\beta,\nu}$,

$$\begin{aligned}\mathcal{E}(k+1) - \mathcal{E}(k) \le &-\frac{s^2}{2}\left(\frac{2\beta n}{n+1} - \frac{n+2}{n+1} - L\beta^2 s\right)\\ &(k - k_{\alpha,\beta,\nu})^2 \left\|\nabla f(x_{k+1})\right\|^2\\ &- s(\nu-2)(k - k_{\alpha,\beta,\nu})\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle\\ &- \frac{s(2k+\alpha+3)(\alpha-\nu-1)}{2}\left\|v_{k+1}\right\|^2.\end{aligned}$$

Using the basic inequality for any function $f(x) \in \mathcal{F}_L^1(\mathbb{R}^n)$ $\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\rangle \ge f(x_{k+1}) - f(x^\star)$, we can obtain the following lemma.

Lemma 10 *Under the same assumption of Theorem 8, the limit* $\lim_{k\to\infty}\mathcal{E}(k)$ *exists, and the summation of the following series exist:*

$$\sum_{k=0}^{\infty}(k+1)^2\left\|\nabla f(x_{k+1})\right\|^2, \qquad \sum_{k=0}^{\infty}(k+1)\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle,$$
$$\sum_{k=0}^{\infty}(k+1)(f(x_{k+1}) - f(x^\star)), \qquad \sum_{k=0}^{\infty}(k+1)\left\|v_{k+1}\right\|^2.$$

Lemma 11 *Under the same assumption of Theorem 8, the following limits exist:*

$$\lim_{k\to\infty}\left\|x_k - x^\star\right\| \quad \textbf{and} \quad \lim_{k\to\infty}(k+1)\left\langle x_{k+1} - x^\star, v_k + \beta\sqrt{s}\nabla f(x_k)\right\rangle.$$

Proof of Lemma 11 Taking $\nu \ne \nu' \in (2, \gamma - 1]$, we have

$$\mathcal{E}_\nu(k) - \mathcal{E}_{\nu'}(k) = (\nu - \nu')\left[-s\beta\cdot\frac{(k+1)(k+2)}{k+\alpha+1}\left(f(x_k) - f(x^\star)\right)\right.$$

$$+(k+1)\sqrt{s}\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle + \frac{(\alpha-1)}{2}\left\|x_{k+1}-x^\star\right\|^2\Bigg].$$

Given Lemma 10, the following limit exists:

$$\lim_{k\to\infty}\left[(k+1)\sqrt{s}\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle + \frac{\alpha-1}{2}\left\|x_{k+1}-x^\star\right\|^2\right] \tag{D.85}$$

Define a function $\pi(k) := \frac{1}{2}\|x_k - x^\star\|^2 + \beta s\sum_{i=k_0}^{k-1}\langle\nabla f(x_i), x_{i+1}-x^\star\rangle$. If we can show the existence of the limit $\pi(k)$ with $k\to\infty$, we can guarantee $\lim_{k\to\infty}\|x_{k+1}-x^\star\|$ exists using Lemma 10. We observe the following equality:

$$\begin{aligned}
&(k+1)(\pi(k+1)-\pi(k)) + (\alpha-1)\pi(k+1) - s(\alpha-1)\beta\sum_{i=0}^{k}\left\langle\nabla f(x_i), x_{i+1}-x^\star\right\rangle\\
&= (k+1)\sqrt{s}\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle - \frac{(k+1)s}{2}\|v_k\|^2\\
&\quad + \frac{\alpha-1}{2}\left\|x_{k+1}-x^\star\right\|^2.
\end{aligned}$$

Lemma 10 and (D.85) tell us there exists some constant $\mathfrak{C}^5$ such that $\lim_{k\to\infty}[(k+\alpha)\pi(k+1)-(k+1)\pi(k)] = \mathfrak{C}^5$, that is, taking a simple translation $\pi'(k) = \pi(k) - \mathfrak{C}^5/(\gamma-1)$, we have $\lim_{k\to\infty}\left[(k+\alpha)\pi'(k+1)-(k+1)\pi'(k)\right] = 0$. Since $\mathcal{E}(k)$ decreases for $k\geq k_{\alpha,\beta,v}$, we have that $\|x_k-x^\star\|^2$ is bounded. Using Lemma 10, we obtain that $\pi(k)$ is bounded; that is, $\pi'(k)$ is bounded. Thus we have $\lim_{k\to\infty}\frac{(k+2)^{\alpha-1}\pi'(k+1)-(k+1)^{\alpha-1}\pi'(k)}{(k+1)^{\alpha-2}} = 0$; that is, for any $\epsilon>0$, there exists $k_0'>0$ such that $\left|\pi'(k)\right| \leq \left(\frac{k_0'+1}{k+1}\right)^{\alpha-1}\left|\pi'(k_0')\right| + \frac{\epsilon\sum_{i=k_0'}^{k-1}(i+1)^{\alpha-2}}{(k+1)^{\alpha-1}}$. With arbitrary $\epsilon>0$, we complete the proof of Lemma 11. □

Proof of (D.83) When $k\geq k_{\alpha,\beta,v}$, we expand the discrete Lyapunov function (D.84) as

$$\begin{aligned}
\mathcal{E}(k) = {}& s(k+1)\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-v)\beta}{k+\alpha+1}\right]\\
&\left(f(x_k)-f(x^\star)\right)\\
&+\sqrt{s}(k+1)v\left\langle x_{k+1}-x^\star, v_k+\beta\sqrt{s}\nabla f(x_k)\right\rangle\\
&+\frac{v(\alpha-1)}{2}\left\|x_{k+1}-x^\star\right\|^2 + \frac{s(k+1)^2}{2}\left\|v_k+\beta\sqrt{s}\nabla f(x_k)\right\|^2.
\end{aligned}$$

Using Lemma 10 and Lemma 11, we obtain the first equation of (D.83). Additionally, we have

$$\begin{aligned} & s\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left(f(x_k)-f(x^\star)\right) \\ & \quad+\frac{(k+1)s}{2}\left\|v_k+\beta\sqrt{s}\nabla f(x_k)\right\|^2 \\ & \leq s\left[k+\alpha+1-\beta+\frac{(k+2)(\alpha-1-\nu)\beta}{k+\alpha+1}\right]\left(f(x_k)-f(x^\star)\right) \\ & \quad+(k+1)s\left\|v_k\right\|^2+(k+1)\beta^2 s^2\left\|\nabla f(x_k)\right\|^2. \end{aligned}$$

Using Lemma 10, we obtain the second equation of (D.83). □

References

1. Alvarez, F.: On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM J. Control Optim. **38**(4), 1102–1119 (2000)
2. Alvarez, F., Attouch, H., Bolte, J., Redont, P.: A second-order gradient-like dissipative dynamical system with Hessian-driven damping: application to optimization and mechanics. J. Math. Pures Appl. **81**(8), 747–779 (2002)
3. Arnold, V.I.: Mathematical Methods of Classical Mechanics, vol. 60. Springer Science and Business Media, Berlin (2013)
4. Attouch, H., Cabot, A.: Convergence rates of inertial forward-backward algorithms. SIAM J. Optim. **28**(1), 849–874 (2018)
5. Attouch, H., Chbani, Z., Peypouquet, J., Redont, P.: Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. Math. Progr. **168**(1–2), 123–175 (2018)
6. Attouch, H., Chbani, Z., Riahi, H.: Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. arXiv preprint arXiv:1706.05671 (2017)
7. Attouch, H., Maingé, P.-E., Redont, P.: A second-order differential system with Hessian-driven damping; application to non-elastic shock laws. Differ. Equ. Appl. **4**(1), 27–65 (2012)
8. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. SIAM J. Optim. **26**(3), 1824–1834 (2016)
9. Attouch, H., Peypouquet, J., Redont, P.: A dynamical approach to an inertial forward-backward algorithm for convex minimization. SIAM J. Optim. **24**(1), 232–256 (2014)
10. Attouch, H., Peypouquet, J., Redont, P.: Fast convex optimization via inertial dynamics with Hessian driven damping. J. Differ. Equ. **261**(10), 5734–5783 (2016)
11. Badithela, A., Seiler P.: Analysis of the heavy-ball algorithm using integral quadratic constraints. In: 2019 American Control Conference (ACC), pp. 4081–4085, IEEE, (2019)
12. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)
13. Betancourt, M., Jordan, M.I., Wilson, A.C.: On symplectic optimization. arXiv preprint arXiv:1802.03653, (2018)
14. Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends Mach. Learn. **8**(3–4), 231–357 (2015)
15. Bubeck, S., Lee, Y.T., Singh, M.: A geometric alternative to Nesterov's accelerated gradient descent. arXiv preprint arXiv:1506.08187, (2015)
16. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Lower bounds for finding stationary points I. arXiv preprint arXiv:1710.11606, (2017)
17. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". J. Optim. Theory Appl. **166**(3), 968–982 (2015)
18. Diakonikolas, J., Orecchia, L.: The approximate duality gap technique: a unified theory of first-order methods. arXiv preprint arXiv:1712.02485, (2017)

19. Drusvyatskiy, D., Fazel, M., Roy, S.: An optimal first order method based on optimal quadratic averaging. SIAM J. Optim. **28**(1), 251–271 (2018)
20. Du, S.S., Jin, C., Lee, J.D., Jordan, M.I., Singh, A., Poczos, B.: Gradient descent can take exponential time to escape saddle points. In: Advances in Neural Information Processing Systems, pp. 1067–1077, (2017)
21. Flammarion, N., Bach, F.: From averaging to acceleration, there is only a step-size. In: Conference on Learning Theory, pp. 658–695, (2015)
22. Gelfand, I.M., Tsetlin, M.L.: The method of" ravines" (Russian). Uspekhi Matematicheskikh Nauk (Progres in Mathematics) **17**, 103–131 (1962)
23. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Progr. **156**(1–2), 59–99 (2016)
24. Hu, B., Lessard, L.: Control interpretations for first-order optimization methods. In: 2017 American Control Conference (ACC), pp. 3114–3119. IEEE, (2017)
25. Hu, B., Lessard, L.: Dissipativity theory for Nesterov's accelerated method. arXiv preprint arXiv:1706.04381, (2017)
26. Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., Jordan, M.I.: How to escape saddle points efficiently. In: International Conference on Machine Learning (ICML), New York, ACM Press, (2017)
27. Kim, D., Fessler, J.A.: Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. arXiv preprint arXiv:1803.06600, (2018)
28. Krichene, W., Bartlett, P.L.: Acceleration and averaging in stochastic descent dynamics. In: Advances in Neural Information Processing Systems, pp. 6796–6806, (2017)
29. Krichene, W., Bayen, A., Bartlett, P.L.: Accelerated mirror descent in continuous and discrete time. In: Advances in Neural Information Processing Systems, pp. 2845–2853, (2015)
30. Krichene, W., Bayen, A., Bartlett, P.L.: Adaptive averaging in accelerated descent dynamics. In: Advances in Neural Information Processing Systems, pp. 2991–2999, (2016)
31. Lessard, L., Recht, B., Packard, A.: Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim. **26**(1), 57–95 (2016)
32. Lin, H., Mairal, J., Harchaoui, Z.: Catalyst acceleration for first-order convex optimization: from theory to practice. J. Mach. Learn. Res. **18**(212), 1–54 (2018)
33. Nemirovsky, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization. SIAM Rev. **27**(2), 264–265 (1983)
34. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Sov. Math. Doklady **27**(2), 372–376 (1983)
35. Nesterov, Y.: How to make the gradients small. Optima **88**, 10–11 (2012)
36. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Springer Science and Business Media, Berlin (2013)
37. O'Donoghue, B., Candès, E.J.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**(3), 715–732 (2015)
38. Pedlosky, J.: Geophysical Fluid Dynamics. Springer Science and Business Media, Berlin (2013)
39. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
40. Shi, B., Du, S.S., Su, W.J., Jordanm M.I.: Acceleration via symplectic discretization of high-resolution differential equations. In: Advances in Neural Information Processing Systems, pp. 5745–5753, (2019)
41. Su, W., Boyd, S., Candès, E.L.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**(153), 1–43 (2016)
42. Vassilis, A., Jean-François, A., Charles, D.: The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b<3$. SIAM J. Optim. **28**(1), 551–574 (2018)
43. Wibisono, A., Wilson, A.C., Jordan, M.I.: A variational perspective on accelerated methods in optimization. Proc. Natl. Acad. Sci. **113**(47), E7351–E7358 (2016)
44. Wilson, A.C., Recht, B., Jordan, M.I.: A Lyapunov analysis of momentum methods in optimization. J. Mach. Learn. Res. **22**, 1–34 (2021)
45. Zhang, J., Mokhtari, A., Sra, S., Jadbabaie, A.: Direct Runge–Kutta discretization achieves acceleration. arXiv preprint arXiv:1805.00521, (2018)

Affiliations

Bin Shi[1] · Simon S. Du[2] · Michael I. Jordan[3] · Weijie J. Su[4]

✉ Bin Shi
shibin@lsec.cc.ac.cn

Simon S. Du
ssdu@cs.washington.edu

Michael I. Jordan
jordan@cs.berkeley.edu

Weijie J. Su
suw@wharton.upenn.edu

1 State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

2 University of Washington, Seattle, USA

3 University of California, Berkeley, USA

4 University of Pennsylvania, Philadelphia, USA