Revisiting the Role of Euler Numerical Integration on Acceleration and Stability in Convex Optimization

Peiyuan Zhang **Antonio Orvieto** **Hadi Daneshmand** **Thomas Hofmann** **Roy Smith**
ETH Zurich, Switzerland

Abstract

Viewing optimization methods as numerical integrators of ordinary differential equations (ODEs) provides a thought-provoking modern framework for studying accelerated first-order optimizers. In this literature, acceleration is often observed to be linked to the quality of the integration method (e.g., the order of accuracy, stability, structure-preservation). In this work, we propose a novel ordinary differential equation that questions this connection: both the explicit and the semi-implicit Euler discretizations on this ODE lead to an accelerated algorithm for convex programming. Although semi-implicit methods are well-known in numerical analysis to enjoy many desirable features, our findings show that these properties do not necessarily relate to acceleration.

1 Introduction

Momentum methods are the state-of-the-art choice of practitioners for the optimization of machine learning models. The simplest of such algorithms is the Heavy-ball (HB), first proposed and analyzed in the context of convex optimization by Polyak in 1964 [20]:

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

where $f : \mathbb{R}^d \to \mathbb{R}$ is the L-smooth[1] function we want to minimize, $s > 0$ is the step-size and $\beta \in [0, 1)$ a *momentum parameter*. Using a novel and quite beautiful argument on fixed point iterations, Polyak [20] proved that, if f is twice continuously differentiable and μ-strongly-convex[2], the sequence $(x_k)_{k\geq 0}$ produced by HB *locally* (i.e. if initialized close to the solution) converges to the minimizer $x^* = \arg\min_{x\in\mathbb{R}^d} f(x)$ at an *accelerated* rate. The keyword "accelerated" has a precise meaning: an algorithm for μ-strongly-convex and L-smooth problems is accelerated if and only if the convergence rate of $f(x_k)$ to $f^* := \min_{x\in\mathbb{R}^d} f(x)$ is $O((1 - \sqrt{\mu/L})^k)$. For instance, Gradient Descent (i.e. $\beta = 0$) in this setting converges linearly but with constant $1 - \mu/L$ and is therefore not accelerated[3].

Nesterov's acceleration. Supported by the lower bounds established by Nemirovski and Yudin [17], many researchers in the early 80s tried to develop an algorithm with global (i.e. iteration-independent) accelerated rate. Notably, the problem was solved by Nesterov [18] in 1983, who proposed following modification [4] of HB:

$$\begin{aligned} x_{k+1} = x_k + \beta(x_k - x_{k-1}) - s\nabla f(x_k) \\ - \beta s(\nabla f(x_k) - \nabla f(x_{k-1})). \end{aligned} \qquad \text{(NAG)}$$

The intuition behind this algorithm puzzled researchers for decades, and many papers are devoted to understanding the underlying acceleration mechanism [2, 4, 1] and the role of the small modification[5] compared to HB [5, 12, 10]. Notwithstanding the theoretical value of these contributions, they are arguably only of a descriptive nature, and leave open very fundamental questions on the reason behind acceleration.

Continuous-time models for acceleration. In 2014 a new line of research bloomed from a seminal paper by Su, Boyd and Candes [23]. This work gained a lot of attraction, as it introduces a completely novel[6] and powerful way to look at acceleration through the lens of second order ordinary differential

Correspondence to zhangp@student.ethz.ch.

[1] For all $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$, where $\|\cdot\|$ is the standard Euclidean norm.

[2] $\forall x \in \mathbb{R}^d$, $\nabla^2 f(x) - \mu I$ is positive semidefinite.

[3] If L/μ is big, $1 - \sqrt{\mu/L} \ll 1 - \mu/L$.

[4] In the original paper [18], the algorithm is presented in a more general way. Our formulation is due to [21].

[5] This is usually referred to as *gradient extrapolation*.

[6] We point out that, actually, the differential equations proposed in [23] was already written down in the original 1963 paper by Polyak [20]. Even more surprisingly, a first study of damped second order differential equations for optimization can be found already in a 1958 paper of the soviet mathematician Mark Gavurin [7].

equations (ODEs). In the strongly-convex case, this equation is

$$\ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X) = 0 \qquad \text{(NAG-ODE)}$$

and retains the essence of acceleration: namely, convergence with a rate $O(e^{-\sqrt{\mu}t})$. Analogously to the discrete-time case we just discussed, one can prove that the continuous-time model of gradient descent, i.e. the gradient flow $\dot{X} = -\nabla f(X)$, converges instead at the non-accelerated rate $O(e^{-\mu t})$.

However, as first noted by [24], while NAG-ODE is formally the continuous-time limit (for some specific choice of β) of NAG, it is also the continuous-time limit of HB. In other words, NAG-ODE does not capture the *gradient correction* (a.k.a gradient extrapolation) term $\beta s(\nabla f(x_k) - \nabla f(x_{k-1}))$, which is regarded to be a fundamental piece of the acceleration machinery in discrete-time. To solve this issue (i.e. to get a more accurate model of Nesterov's acceleration), Shi et al. [21, 22] introduced a high-resolution model of NAG:

$$\begin{aligned}\ddot{X} + (2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X))\dot{X} \\ + (1 + 2\sqrt{\mu s})\nabla f(X) = 0.\end{aligned} \qquad \text{(NAG-ODE-HR)}$$

Remarkably, here 1) the step-size s is included directly in the model, and 2) the vanishing (as $s \to 0$) term $\sqrt{s}\nabla^2 f(X)\dot{X}$ is used to include the gradient correction $\beta s(\nabla f(x_k) - \nabla f(x_{k-1}))$. The second point is referred as *Hessian damping*, and can be seen as a curvature-dependent viscosity correction. In [22] the authors show that NAG-ODE-HR also enjoys an accelerated rate, equal to the one of NAG-ODE. Crucially, the authors show that NAG is linked to the semi-implicit discretization of NAG-ODE-HR. This is then used to motivate why NAG-ODE-HR is empirically able to model NAG better than NAG-ODE (see experiments in [22]). On a parallel line, Mühlebach and Jordan [15] derived a different high resolution continuous-time model which contains terms of the form $\nabla f(X + \sqrt{s}\dot{X})$ instead of $\sqrt{s}\nabla^2 f(X)\dot{X}$, and made more evident the connection to semi-implicit integration. Finally, in the last few months, some research papers [16, 6] have been devoted to understanding the geometric properties and to Strang splitting methods for geometric integration of conformal Hamiltonian systems [14].

Our contribution. This paper focuses on the following recent result: [22] proves that semi-implicit integration of NAG-ODE-HR enjoys an accelerated rate $O(\exp(-k\sqrt{\mu/L}))$ at iteration k while the established rate upper-bound for explicit Euler discretization is not accelerated. This result has initiated a recent search for proper integration methods for momentum models. Symplectic [16], Runge-Kutta [25], and semi-implicit Euler integrators [22] are examples of sophisticated integrators used to develop an accelerated optimization from continuous-time model. In contrast, the community is also presuming that standard explicit Euler can not achieve an accelerated algorithm due to its unstable nature (see discussion in [9]). Then a question naturally arises: *among all continuous-time model of momentum methods is there one whose standard explicit Euler achieves an accelerated algorithm?* We provide a *positive* answer to this question by studying a set of momentum ODEs — indexed by three parameters. We find a set of ODEs whose explicit Euler integration simply achieves an accelerated algorithm, with no need for sophisticated integrators. Moreover, on some particular ODEs, our theoretical and empirical results show better convergence properties for algorithms obtained by explicit Euler integrators compared to those obtained by semi-implicit. This is a controversial observation, since semi-implicit integrators often are *preferred* integrators whose integration error is less than explicit Euler integrators when integrating general ODEs. Yet, we will show that on a specific set of ODEs model, which model momentum algorithms, explicit Euler integration is as good as semi-explicit Euler integration. More precisely, they both enjoy an integration error that vanishes exponentially fast with the number of integrations.

2 Summary of Results

Our work is based on the study of a novel continuous-time model for momentum algorithms, namely the following ordinary differential equation that is indexed with non-negative parameters m, n, q:

$$\begin{cases}\dot{X} = -m\nabla f(X) - nV \\ \dot{V} = \nabla f(X) - qV,\end{cases} \qquad \text{(GM-ODE)}$$

We show that the above ODE recovers a large set of momentum methods through the application of two classical numerical integrators, i.e. semi-implicit and explicit Euler. Our main idea is a reparameterization: any semi-implicit discretization of GM-ODE with parameters $\{m_{\text{SIE}}, n_{\text{SIE}}, q_{\text{SIE}}\}$ can be viewed as explicit Euler discretization of GM-ODE with different parameters $\{m_{\text{EE}}, n_{\text{EE}}, q_{\text{EE}}\}$. The exact correspondence between these two sets of parameters is characterized in Lemma 2. Such an equivalence highlights the blurred connection between discretization and optimization: the same optimizer can be interpreted as explicit Euler and, at the same time, semi-implicit Euler discretization of proper continuous-time models. This statement holds for both Heavy-ball and Nesterov's method.

We establish an accelerated convergence rate (in Thm. 3 and Thm. 6) for a set of algorithms that can be interpreted as (both two) Euler discretization of GM-ODE

with a proper parameters choice. Our convergence analysis is general enough to achieve an accelerated rate for a recently developed momentum algorithm, called quasi-hyperbolic momentum [13], whose rate in the general strongly-convex case was not yet derived (see Corollary 5) to the best of our knowledge. To better understand the difference between different integrators, we go beyond the convergence analysis and study their discretization errors. Depending on the choice of parameters, we show that the explicit Euler method enjoys the same integration error as the semi-implicit Euler method when integrating GM-ODE (see Lemma 8).

3 Continuous-time analysis

Before discussing numerical integration, we provide here a continuous-time analysis of this differential equation, in line with most related works on acceleration and numerical integration [21, 23]. The results in this section are not fundamental for the understanding of our claims on the discretization of GM-ODE. Hence, for a quick read, this section can be safely skipped.

Let us reconsider our continous-time model GM-ODE. The variable X models the position iterates of an accelerated algorithm and V can be thought of as a momentum term. This model then can be seen as as a linear combination of the gradient flow $\dot{X} = -\nabla f(X)$ (obtained for $n = 0$) and NAG-ODE (obtained for $n = 1$). Assuming the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth, one can check that the ODE admits a unique solution (see e.g. Thm. 3.2. in [11]). The model above is inspired by the quasi-hyperbolic momentum (QHM) algorithm[7] developed in [13]. We will discuss the connection to QHM later in Sec. 4.2.

Connections to existing models. GM-ODE recovers existing continuous momentum models using different choices of parameters. To see this, let us take the second derivative of X: $\ddot{X} = -m\nabla^2 f(X)\dot{X} - n\dot{V}$.

$$\ddot{X} + \big(q + m\nabla^2 f(X)\big)\dot{X} + (n + qm)\nabla f(X) = 0. \quad (1)$$

The choice[8] $m = 0$, $n = 1, q = 2\sqrt{\mu}$ recovers NAG-ODE by Polyak [20]. Moreover, the choice $m = \sqrt{s}, n = 1, q = 2\sqrt{\mu}$ recovers NAG-ODE-HR, proposed by Shi et al. [21, 22]. That is, GM-ODE includes as special cases both the *high-resolution* and *low-resolution* models of Nesterov's method. Furthermore, we note that, contrary to [21], the Hessian of f is not explicitly included in the model. Also, contrary to [15], the gradient is evaluated only at the current position X. These features give our GM-ODE higher flexibility and interpretablity than existing models – a simple linear combination of gradient and momentum can also achieve *high resolution*[9].

	m	n	q
Gradient Flow	non-zero	0	any
NAG-ODE [23]	0	1	$2\sqrt{\mu}$
NAG-ODE-HR [22]	$\sqrt{s}$	1	$2\sqrt{\mu}$

Stability and convergence rate. The equilibria of GM-ODE are easy to characterize: since m, n and q are non-negative, we have $\dot{X} = 0$ and $\dot{V} = 0$ if and only if both $\nabla f(X) = 0$ and $V = 0$. Under the assumption that f is strongly-convex, only its unique minimizer x^* is such that $\nabla f(x^*) = 0$. Therefore the point $(x^*, 0) \in \mathbb{R}^{2d}$ is the *only equilibrium* of GM-ODE. Next, we want to show that $(x^*, 0)$ is asymptotically stable and characterize the convergence rate of our model. Borrowing some inspiration from [23, 22], we propose the following Lyapunov function:

$$\mathcal{E}(X, V) = (qm + n)\big(f(X) - f(x^*)\big) + \frac{1}{4}\|q(X - x^*) - nV\|^2 + \frac{n(qm + n)}{4}\|V\|^2. \quad (2)$$

The next theorem states our result about Lyapunov stability, of which the proof is presented in the appendix.

Theorem 1 (Continuous-time stability)**.** *Suppose that f is μ-strongly-convex and L-smooth. If $n, m, q \geq 0$ then, for any value of the strong-convexity modulus $\mu \geq 0$, the point $(x^*, 0) \in \mathbb{R}^{2d}$ is globally asymptotically stable for GM-ODE, as*

$$\mathcal{E}(X(t), V(t)) \leq e^{-\gamma_1 t} \cdot \mathcal{E}(X(0), V(0)), \quad (3)$$

where $\gamma_1 := \min\left(\frac{\mu(n + qm)}{2q}, \frac{q}{2}\right)$.

Remarkably, the above stability analysis can be used to guide the analysis of different momentum methods (see Sec. 4) — obtained by the application of standard Euler integrators of our model.

How do parameter affect the ODE dynamics? One can readily check that Thm. 1 implies a linear rate in function value of the form $f(X(t)) - f(x^*) \leq O(-e^{\gamma_1 t})$. This result recovers exactly the rates in [21] as a special case. However, we note that our result is more general and leads to novel insights on the interplay between gradient amplification (controlled by

[7] QHM was introduced as weighted average of momentum and gradient descent methods. It is shown to recover both HB and NAG as special cases.

[8] Proofs for discretized GM-ODE will rely on condition $m > 0$. This discussion will be elaborated on Sec. 4.

[9] That is, a finer, compared to the original ODE in [23] approximation of Nesterov's method. For a detailed discussion on this terminology, we refer the reader to [21].

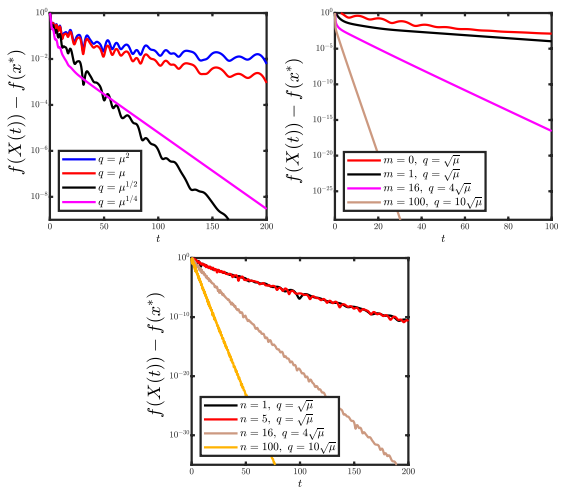

Figure 1: Role of parameters in GM-ODE. The objective function is a 10-dimensional quadratic function f with $\mu I \preceq \nabla^2 f \preceq LI$ where $\mu = 0.01$ and $L = 1$. The panels depict, from left to right, the influence of q, m and n, as suggested in above discussion. In each figure we vary the parameter we are interested in (as in the legends) and keep the others fixed. For left panel we use $m = 0.2$ and $n = 1$. For the middle panel we use $n = 0.1$. For the right panel we use $m = 0.2$. Numerical integration of GM-ODE performed using a fourth-order Runge-Kutta with step-size 10^{-4}.

n), momentum (controlled by q) and Hessian damping (controlled by m). Indeed, given the expression for γ_1 in Eq. (3), we can make the following conclusions.

- For fixed $m, n \geq 0$, the value of q which maximizes γ_1 also solves $\mu(n + qm)/q = q$, which implies $q = (m + \sqrt{4\mu n + m^2})/2$. If we restrict q to be a power of μ, set $n = 1$ and ignore the effect of m, then we get the popular choice [21, 22, 15] $q = O(\sqrt{\mu})$ (see the first panel of Fig. 1).

- For any $q \geq 0$, if $n \geq 0$ is chosen small enough such that $q^2 - \mu n \geq 0$, then by picking $m = (q^2 - \mu n)/q$ we have $\gamma_1 = q/2$. Hence, by increasing q the convergence can be sped-up *arbitrarily* (see the second panel of Fig. 1).

- If $n = q^2/\mu$, then $\gamma_1 = q/2$ for all $q \geq 0$ and any $m \geq 0$. Again, by increasing q the convergence can be sped-up *arbitrarily* (bottom panel of Fig. 1).

Remark 1. If n or m are increased, one can guarantee arbitrarily fast convergence to the minimizer. This result only holds in continuous-time, as noted also in a similar setting by [24]. Indeed, as we will see in Thm. 3, in the discrete word, *to ensure stability*, n and m have to be bounded by a constant which is inversely proportional to the discretization accuracy.

4 Discretization and acceleration

Here we show how *both* explicit and semi-implicit numerical integration, applied to GM-ODE, can yield accelerated gradient iterations.

Discretization schemes. We consider two well-understood [9] and practical first-order numerical integration schemes applied to GM-ODE with discretization step-size $\sqrt{s}$: Explicit Euler (EE) and Semi-Implicit[10] Euler (SIE).

$$\text{(EE)}: \begin{cases} x_{k+1} - x_k = -m\sqrt{s}\nabla f(x_k) - n\sqrt{s}v_k \\ v_{k+1} - v_k = \sqrt{s}\nabla f(x_k) - q\sqrt{s}v_k. \end{cases}$$

$$\text{(SIE)}: \begin{cases} x_{k+1} - x_k = -m\sqrt{s}\nabla f(x_k) - n\sqrt{s}v_k \\ v_{k+1} - v_k = \sqrt{s}\nabla f(x_{k+1}) - q\sqrt{s}v_k. \end{cases}$$

Even though the second equation in SIE is written in an *implicit* way, it can be trivially solved: indeed, one can first find x_{k+1} and then plug the solution into the second equation. Since the gradient computed at x_{k+1} can be used at the next iteration, the two algorithms have the same complexity. Indeed, for $n \neq 0$ (gradient descent is recovered for $n = 0$), by simplifying the variable v, the above schemes can be written in just one line:

$$x_{k+1} = x_k + (1 - q\sqrt{s})(x_k - x_{k-1}) - m\sqrt{s}\nabla f(x_k) + ((1 - q\sqrt{s})m\sqrt{s} - ns)\nabla f(x_{k-1}); \quad \text{(EE)}$$

$$x_{k+1} = x_k + (1 - q\sqrt{s})(x_k - x_{k-1}) - \left(m\sqrt{s} + ns\right)\nabla f(x_k) + (1 - q\sqrt{s})m\sqrt{s}\nabla f(x_{k-1}). \quad \text{(SIE)}$$

Remarkably, different choices of parameters yield a *rich set of momentum methods*, and the reader can probably already notice some rather interesting choice of parameters which can recover some very well-known algorithms (see introduction). We explore this in the next subsection.

4.1 Equivalence between SIE and EE

We show that algorithms obtained from semi-implicit discretization of an accelerated flow can be seen as explicit discretization of a different accelerated flow.

[10] Actually, there exist many semi-implicit methods that go under the name of "semi-implicit Euler". We expect many of those integrators to work equally well. For a more detailed discussion, we refer the reader to Chapter 1 of [9].

Lemma 2 (Equivalence between SIE and EE)**.** *For $n = 0$ both EE and SIE reduce to gradient descent. For $n \neq 0$, consider parameters m_{SIE}, n_{SIE}, q and set*

$$m_{EE} = m_{SIE} + \sqrt{s} n_{SIE},$$
$$n_{EE} = (1 - q\sqrt{s}) n_{SIE}.$$

EE with stepsize $\sqrt{s} > 0$ on GM-ODE with parameters m_{EE}, n_{EE}, q leads to the same exact algorithm as the one obtained using SIE with stepsize $\sqrt{s} > 0$ on GM-ODE with parameters m_{SIE}, n_{SIE}, q.

Proof. We start from the one-line representation. We get the following conditions for $n \neq 0$:

$$\begin{cases} m_{\text{SIE}}\sqrt{s} + s n_{\text{SIE}} = m_{\text{EE}}\sqrt{s} \\ (1 - q\sqrt{s}) m_{\text{SIE}}\sqrt{s} = (1 - q\sqrt{s}) m_{\text{EE}}\sqrt{s} - s n_{\text{EE}}. \end{cases}$$

We conclude by substituting the first equation into the second. □

As a crucial consequence of the last lemma, Heavy-ball and Nesterov method can be seen both as semi-implicit and explicit integrators on GM-ODE. This is illustrated in the following table.

	EE discretization	SIE discretization
HB	$q = (1-\beta)/\sqrt{s}$ $m = \sqrt{s}$ $n = \beta$	$q = (1-\beta)/\sqrt{s}$ $m = 0$ $n = 1$
NAG	$q = (1-\beta)/\sqrt{s}$ $m = (1+\beta)\sqrt{s}$ $n = \beta^2$	$q = (1-\beta)/\sqrt{s}$ $m = \sqrt{s}$ $n = \beta$

Table 1: HB and NAG with any stepsize $s > 0$ and momentum $\beta \in (0, 1)$ (see definition in the introduction) can be seen as both EE or SIE numerical integrators.

Since, as it is well known, NAG is accelerated, Lemma 2 shows that both explicit and semi-implicit Euler integrators can lead to acceleration under well-chosen parameters. In the next subsection, we elaborate more on this finding and recover parameters which lead to acceleration for both EE and SIE.

4.2 Semi-implicit Euler is accelerated

Leveraging insights the ODE stability analysis in Thm. 1 and the lessons learned from semi-implicit Lyapunov function design in recent literature [21, 22], our next result establishes a general convergence rate for the semi-implicit Euler method on GM-ODE. In the next subsection, we provide the similar result for EE, using Lemma 2.

Theorem 3 (Convergence of SIE)**.** *Assume f is L-smooth and μ-strongly-convex. Let $(x_k)_{k=1}^{\infty}$ be the sequence obtained from semi-implicit discretization of GM-ODE with step $\sqrt{s}$. Let*

$$0 < m\sqrt{s} \leq \frac{1}{2L},\ 0 < ns \leq m\sqrt{s},\ 0 < q\sqrt{s} \leq \frac{1}{2}. \quad (4)$$

There exists a constant $C > 0$ such that, for any $k \in \mathbb{N}$, it holds that

$$f(x_k) - f(x^*) \leq \left(1 + \gamma_2\sqrt{s}\right)^{-k} C,$$

where $\gamma_2 := \frac{1}{5} \min\left(\frac{n\mu}{q}, \frac{q}{1 + q^2/(nL)}\right)$.

Proof Sketch. The proof is based on the analysis of the following Lyapunov function inspired by the Lyapunov analysis for the continuous-time model (cf. Sec. 3):

$$\mathcal{E}(k) = r_1 r_2 (f(x_k) - f(x^*)) - \frac{r_1 r_2 m\sqrt{s}}{2} \|\nabla f(x_k)\|^2 + \frac{n r_1^2 r_2}{4} \|v_k\|^2 + \frac{1}{4} \|q(x_{k+1} - x^*) - n r_1 v_k\|^2,$$

where $r_1 = 1 - q\sqrt{s}$, $r_2 = n + mq$ and the last term is a vanishing (as $s \to 0$) correction that accounts for the discretization error (see also [22]). In the appendix we show that $\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\gamma_2\sqrt{s}\mathcal{E}(k+1)$, completing the proof. □

The result above is extremely general, and can be thought as a master theorem for deriving accelerated rates for many (possibly unknown) momentum methods. We illustrate this by deriving, the rate of Nesterov's method in just a few lines. We note that known results on semi-implicit integration such as the ones presented in [22] are less general since why are limited to high/low resolution or to a fixed viscosity.

From Thm. 4 to the well-known rate for NAG. By invoking Thm. 3, we can recover the acceleration of NAG since it can be regarded as SIE discretization of our GM-ODE.

Corollary 4 (NAG is accelerated)**.** *Assume that f is L-smooth and μ-strongly-convex with $L/\mu \geq 9$[11]. Consider the SIE discretization of GM-ODE with $s \leq \frac{1}{4L}$, $q = (1-\beta)/\sqrt{s}$ (with $\beta = 1 - 2\sqrt{\mu s}$) $m = \sqrt{s}$, $n = \beta$ (i.e. NAG, see Table 1). The algorithm enjoys the convergence rate $O((1 - \sqrt{\mu/L})^k)$. Namely, there exists a constant $C > 0$ such that $f(x_k) - f(x^*) \leq \left(1 + \sqrt{\mu s}/15\right)^{-k} C$.*

[11] The lower bound assumption for conditional number here and in Cor. 5 is purely technical and only serves for a

Proof. The conditions in in Eq. (4) are satisfied since $s = m\sqrt{s} \le 1/(4L)$, $n = \beta < 1 = \frac{m}{\sqrt{s}}$ and $q\sqrt{s} = 2\sqrt{\mu s} \le 2\sqrt{Ls/9} \le 1/3$. Thus, $\frac{n\mu}{5q} = \frac{(1-2\sqrt{\mu s})\sqrt{\mu}}{10} \ge \frac{(1-1/3)\sqrt{\mu}}{10} \ge \frac{\sqrt{\mu}}{15}$ and $\frac{q}{5+5q^2/(nL)} \ge \frac{2\sqrt{\mu}}{5+6\mu/L} \ge \frac{\sqrt{\mu}}{9}$. □

From Thm. 4 to a new rate for QHM. The generality of our model and discretization analysis provides an accelerated convergence rate for a broad class of momentum methods. Among these methods is Quasi-hyperbolic momentum (QHM) [13], which was recently developed and shows promises in the optimization for neural network. This method consists of the following iterative steps[12]:

$$\begin{cases} x_{k+1} = x_k - s((1-a)\nabla f(x_k) + a g_{k+1}) \\ g_{k+1} = b g_k + \nabla f(x_k), \end{cases} \tag{QHM}$$

where $a, b \in (0,1)$. For classification tasks, QHM yields accelerated rate on real-world datasets (even better than NAG) [13]. Despite empirical benefits, the convergence analysis for this algorithm is limited to quadratics [8]. Using Thm. 3, the next corollary establishes an accelerated rate for this method (proof in the appendix). To the best of our knowledge, this rate is novel and provides a generalization to the result of [8].

Corollary 5 (Convergence of QHM)**.** *Assume f is L-smooth and μ-strongly-convex with $L/\mu \ge 9$. The iterates of QHM enjoy a linear convergence rate for $s \le \frac{1}{4L}$ and $a \le 1/2$. In particular, QHM also enjoys convergence rate $O((1-\sqrt{\mu/L})^k)$ for $b = 1 - 2\sqrt{\mu s}$. Namely, there exists a constant $C > 0$ such that $f(x_k) - f(x^*) \le \left(1 + a\sqrt{\mu s}/10\right)^{-k} C$.*

Fig. 2 shows the accelerated rate established in the last lemma and its dependency on the parameter a.

Thm. 3 fails to prove accelerated rate for HB. An interesting question may arise as a consequence of our results: since HB can be recast as semi-implicit discretization of GM-ODE, then does invoking Thm. 3 produce a global acceleration proof for HB? The answer is no, since the convergence result in Thm. 3 is conditioned on $m > 0$; while one needs to set $m = 0$ to obtain HB by SIE integrator.

The trade-off speed-stability. As noted in Remark 1, in continuous time one can increase either m or n to infinity and get an arbitrarily fast rate. Thm. 3 shows why a similar phenomenon is not possible in discrete time (would violate the lower bound in [17]): for a specific discretization stepsize $\sqrt{s}$, Eq. (4) gives us a bound on the maximum m and n we can choose to have guaranteed stability. In other words, if we choose a large value for either m and n to get a faster rate, we would end up with a slow algorithm since numerical stability would require a very small integration stepsize. Hence, as expected by the classic theory of convex optimization [17], there is a sweet spot which yields $\gamma_2 = O(\sqrt{\mu/L})$ — a.k.a acceleration.

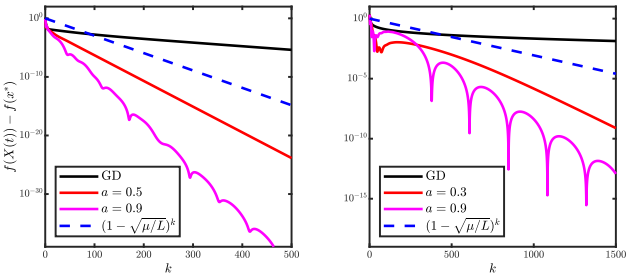

Figure 2: Convergence of QHM. The left plot shows the convergence of 10-dimensional quadratic with $\mu = 0.01$ and $L = 1$; the right plot reports 10-dimensional regularized logistic regression (random data and labels) with regularization weight $l = 10^{-4}$. We specifically used $s = 0.5$, $\beta = 1 - 2\sqrt{\mu s}$ and $s = 0.5$, $\beta = 1 - 2\sqrt{\mu l}$, respectively, in the two experiments.

4.3 Explicit Euler is *also* accelerated!

In the last subsection, we provided a convergence rate for semi-implicit discretization of GM-ODE and showed how this general result can be applied to derive (old and new) convergence rates for momentum methods. However, as already noted a few times, Lemma 2 implies that an equivalent theorem can be written for the explicit Euler method.

Corollary 6 (Convergence of EE)**.** *Assume f is L-smooth and μ-strongly-convex. Let $(x_k)_{k=1}^{\infty}$ be the sequence obtained from semi-implicit discretization of GM-ODE with step $\sqrt{s}$. Let*

$$\begin{aligned} &0 < m\sqrt{s} - ns/(1-q\sqrt{s}) \le \frac{1}{2L}, \\ &0 < ns \le \frac{1-q\sqrt{s}}{2} m\sqrt{s}, \quad 0 < q\sqrt{s} \le \frac{1}{2}. \end{aligned} \tag{5}$$

There exists a constant $C > 0$ such that, for any $k \in \mathbb{N}$, it holds that

$$f(x_k) - f(x^*) \le \left(1 + \gamma_3\sqrt{s}\right)^{-k} C,$$

where $\gamma_3 := \frac{1}{5} \min\left(\frac{n\mu}{q(1-\mu\sqrt{s})}, \frac{q}{1 + q^2/(nL)} \right)$.

simple illustration of these corollaries.

[12] In fact, in [13] presented a normalized second iteration, i.e. $g_{k+1} = bg_k + (1-b)\nabla f(x_k)$, which is generally equivalent to the one we derive here by factor rescaling.

Proof. Consider an explicit method with parameters $(m_{\text{EE}}, n_{\text{EE}}, q)$ and a semi-implicit method with parameters $(m_{\text{SIE}}, n_{\text{SIE}}, q)$. Thm. 3 holds if $0 < m_{\text{SIE}}\sqrt{s} \leq 1/(2L), 0 < sn_{\text{SIE}} \leq m_{\text{SIE}}\sqrt{s}$ and $q\sqrt{s} \leq 1/2$, then it is convergent. By Lemma 2, we can recover the parameter of an equivalent explicit method by setting $n_{\text{EE}} = (1 - q\sqrt{s})n_{\text{SIE}}$ and $m_{\text{EE}} = m_{\text{SIE}} + \sqrt{s}n_{\text{SIE}}$. Combining these conditions with the theorem requirements on n_{SIE}, we get:

$$0 < \frac{sn_{\text{EE}}}{1 - q\sqrt{s}} \leq m_{\text{SIE}}\sqrt{s} = m_{\text{EE}}\sqrt{s} - \frac{sn_{\text{EE}}}{1 - q\sqrt{s}},$$

which implies the condition on n_{EE}. For the condition on m_{EE}, just note that the condition on m_{SIE} from Thm. 3 implies

$$\sqrt{s}m_{\text{SIE}} = \sqrt{s}m_{\text{EE}} - \frac{s}{1 - q\sqrt{s}}n_{\text{EE}} \leq \frac{1}{2L}.$$

□

EE vs. SIE. Comparing conditions on parameters in Thm. 3 with those in Cor. 6 implies that explicit Euler discretization obtains a convergent algorithm for a wider range of m if n is considerably small. Vice versa, semi-implicit Euler converges for a wider range of n for a fixed m. Experiments presented in Fig. 3 substantiates this theoretical finding. According to this experimental result, EE discretization enjoys better stability properties compared to SIE for particular configurations of parameters in GM-ODE. This is in a clear contrast to conventional presumption that advanced integrators are needed for momentum ODE to archive an accelerated algorithm, which is stated repeatedly in the recent literature (see for example [22, 16]).

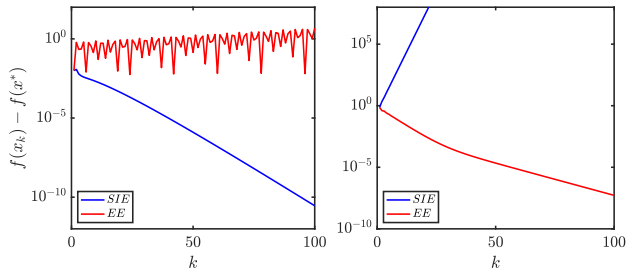

Figure 3: EE vs. SIE. To show that SIE and EE are neither superior nor inferior to each other, in each subplot, we use same parameter $\{m, n, q\}$ for both SIE and EE discretization, while we observe different convergence behaviours. This suggests the stability and convergence is determined by the joint choice of parameters and numerical integrator together. Specifically, the objective function, used for this experiment, is a 2-dimensional quadratic with $\mu = 0.01, L = 1$ and the step-size $s = 1$. In the left plot we use $m = \sqrt{s}, n = 1$ and $q = 2\sqrt{\mu}$ and in the right plot we use $m = 2\sqrt{s}, n = 0.5$ and $q = 2\sqrt{\mu}$.

5 Behaviour of the discretization error

In the last sections, we studied the properties of explicit and semi-implicit integration of GM-ODE and showed that both can lead to acceleration. Yet, most recent literature [24, 22, 15, 16] claims that semi-implicit integration *is somehow more natural* for the approximation of partitioned dissipative systems such as GM-ODE. Indeed, [6, 16] recently showed that the geometric properties of semi-implicit methods combined with backward error analysis [9] can be used to successfully prove the preservation continuous-time rates of convergence up to a controlled error. Instead, our results in Thm. 6, shows that explicit Euler –of a proper ODE– also leads to an accelerated method (see also Table 1). To elaborate on this, we compare semi-implicit and explicit Euler in terms of their approximation error specifically for the integration of GM-ODE. For this particular ODE, EE suffers from a worse local discretization error compared to SIE for the general choice of parameters. Under particular choices of parameters, EE and SIE yield contractive algorithms. In this case, the error of the both discretization schemes decays exponentially fast.

A trap: analysis for the general case. Consider the following discretization errors:

$$\Delta_k^{(\text{EE})} := \|X(k\sqrt{s}) - x_k^{(\text{EE})}\|, \tag{6}$$

$$\Delta_k^{(\text{SIE})} := \|X(k\sqrt{s}) - x_{k+1}^{(\text{SIE})}\|. \tag{7}$$

We compare the above errors for $k = 1$ (for one step). Proof/details are provided in the appendix.

Lemma 7. *Let f be L-smooth and of class C^2. If $m = O(s^{1/2})$, then $\Delta_1^{(SIE)} = O(s^{3/2})$ and $\Delta_1^{(EE)} = O(s)$.*

The above lemma holds for any finite choice of the parameters, and shows that SIE provides a better one-step integration error in the position variable[13]. This result may lead to a wrong conclusion: semi-implicit integration leads to faster algorithm when discretizing GM-ODE. However, this analysis does not provide us a complete picture. Indeed, as we proved in the last section, explicit discretization *can also lead to acceleration* — in particular, it can recover Nesterov's method. To provide some intuition on why a local error analysis leads to misleading conclusions, we provide a tighter analysis of the integration error for a narrowed set of parameters in GM-ODE.

Analysis for contractive cases. A line of recent works around the connection between acceleration and numerical integration [19, 16, 6] studied the behavior of the discretization error of NAG-ODE as $k \to \infty$,

[13] It is well known [9] that these methods actually have the same order, since they are $O(s)$ in the velocity.

showing interesting *shadowing*[14] properties. The main idea behind shadowing is studying the discretization error when the choice of parameters leads to a contractive algorithm. In this case, one can provide a tighter analysis for the discretization error. Particularly, next lemma proves that the integration error of semi-implicit and explicit Euler discretization of GM-ODE decays exponentially fast if one choose parameters n, m, q, and s properly.

Lemma 8. *Suppose f is μ-strongly-convex and L-smooth. For EE discretization of GM-ODE obeying Eq.* (5)*, the discretization error decays as $\Delta_k^{(EE)} = O((1+\gamma_3\sqrt{s})^{-k})$ where γ_3 is defined in Thm. 6. Furthermore, SIE also enjoys $\Delta_k^{(SIE)} = O((1+\gamma_2\sqrt{s})^{-k})$ where γ_2 is defined in Thm. 3 as long as conditions in Eq.* (4) *are satisfied.*

The proof of the last lemma is postponed to the appendix. According to this result, SIE and EE discretization have the same asymptotic integration error properties — under particular choice of parameters. This similarity is also reflected in the convergence rate of obtained algorithms established in Thm. 3 and 6.

6 Conclusion

In this paper, we proposed a general ODE model of momentum-based methods for optimizing smooth strongly-convex functions. The generality of our model allows to view different old and new momentum methods as semi-implicit or explicit Euler integrators and to establish novel accelerated convergence rates for both integrators. In particular, our new findings overturn the old notion: explicit Euler is inferior than semi-implicit due to its unstable nature. Here, we show that the stability and the convergence of discretization is tied to the integrated continuous-time ODE. Therefore, at a deeper level, our methodology provides new challenging insights on the link between accelerated optimization, and numerical integration.

References

[1] Kwangjun Ahn. From proximal point method to Nesterov's acceleration. *arXiv preprint arXiv:2005.08304*, 2020.

[2] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[3] Shui-Nee Chow and Erik S Van Vleck. A shadowing lemma approach to global error analysis for initial value odes. *SIAM Journal on Scientific Computing*, 15(4):959–976, 1994.

[4] Aaron Defazio. On the curved geometry of accelerated optimization. In *Advances in Neural Information Processing Systems*, pages 1764–1773, 2019.

[5] Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695, 2015.

[6] Guilherme França, Michael I Jordan, and René Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *arXiv preprint arXiv:2004.06840*, 2020.

[7] Mark Konstantinovich Gavurin. Nonlinear functional equations and continuous analogues of iteration methods. *Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika*, pages 18–31, 1958.

[8] Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 9630–9640, 2019.

[9] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

[10] Bin Hu and Laurent Lessard. Dissipativity theory for Nesterov's accelerated method. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1549–1557. JMLR. org, 2017.

[11] Hassan K Khalil and Jessy W Grizzle. *Nonlinear systems*, volume 3. Prentice Hall Upper Saddle River, NJ, 2002.

[12] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[13] Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and adam for deep learning. *arXiv preprint arXiv:1810.06801*, 2018.

[14] Robert McLachlan and Matthew Perlmutter. Conformal Hamiltonian systems. *Journal of Geometry and Physics*, 39(4):276–300, 2001.

[14] That is, the discretization bound does *not* explode exponentially due to error accumulation [3] if the objective is convex, due to the contraction provided by the landscape.

[15] Michael Muehlebach and Michael I Jordan. A dynamical systems perspective on Nesterov acceleration. *arXiv preprint arXiv:1905.07436*, 2019.

[16] Michael Muehlebach and Michael I Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *arXiv preprint arXiv:2002.12493*, 2020.

[17] Arkadi Semenovich Nemirovsky and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley, 1983.

[18] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[19] Antonio Orvieto and Aurelien Lucchi. Shadowing properties of optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 12671–12682, 2019.

[20] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[21] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

[22] Bin Shi, Simon S Du, Weijie J Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. *arXiv preprint arXiv:1902.03694*, 2019.

[23] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[24] Ashia C Wilson, Benjamin Recht, and Michael I Jordan. A Lyapunov analysis of momentum methods in optimization. *arXiv preprint arXiv:1611.02635*, 2016.

[25] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. In *Advances in neural information processing systems*, pages 3900–3909, 2018.

Appendix: Proofs and Supplementaries

A Proof for Theorem 1

For convenience of the reader, we report here our generalized model for momentum methods (GM-ODE), motivated in the main paper.

$$\begin{cases} \dot{X} = -m\nabla f(X) - nV \\ \dot{V} = \nabla f(X) - qV. \end{cases} \tag{GM-ODE}$$

Theorem 1 (Continuous-time stability)**.** *Suppose that f is μ-strongly-convex and L-smooth. If $n, m, q \geq 0$ then, for any value of the strong-convexity modulus $\mu \geq 0$, the point $(x^*, 0) \in \mathbb{R}^{2d}$ is globally asymptotically stable for GM-ODE, as*

$$\mathcal{E}(X(t), V(t)) \leq e^{-\gamma_1 t} \cdot \mathcal{E}(X(0), V(0)), \tag{3}$$

where $\gamma_1 := \min\left(\frac{\mu(n+qm)}{2q}, \frac{q}{2}\right)$.

Proof. We propose the Lyapunov function

$$\mathcal{E}(t) = \underbrace{(qm+n)}_{c_1}\big(f(X(t)) - f(x^*)\big) + \underbrace{\frac{n(qm+n)}{4}}_{c_2}\|V(t)\|^2 + \underbrace{\frac{1}{4}}_{c_3}\|q(X(t)-x^*) - nV(t)\|^2, \tag{8}$$

consisting of quadratic and mixing parts

$$\mathcal{E}_1(t) = f(X(t)) - f(x^*), \quad \mathcal{E}_2(t) = \|V(t)\|^2, \quad \mathcal{E}_3(t) = \| - nV(t) + q(X(t)-x^*)\|^2. \tag{9}$$

The derivatives of each quadratic part are

$$\frac{d}{dt}\mathcal{E}_1(t) = -m\|\nabla f(X(t)\|^2 - n\langle \nabla f(X(t)), V(t)\rangle \tag{10}$$

and

$$\frac{d}{dt}\mathcal{E}_2(t) = -2q\|V(t)\|^2 + 2\langle \nabla f(X(t)), V(t)\rangle, \tag{11}$$

along with that of the mixing term:

$$\begin{aligned} \frac{d}{dt}\mathcal{E}_3(t) =& 2\langle -n\dot{V}(t) + q\dot{X}(t), -nV(t) + q(X(t)-x^*)\rangle \\ =& -2(qm+n)\langle \nabla f(X(t)), -nV(t) + q(X(t)-x^*)\rangle \\ =& -2q(qm+n)\langle \nabla f(X(t)), X(t)-x^*\rangle + 2n(qm+n)\langle \nabla f(X(t)), V(t)\rangle \\ \leq& -2q(qm+n)\Big(f(X(t)) - f(x^*)\Big) - \mu q(qm+n)\|X(t)-x^*\|^2 \\ & + 2n(qm+n)\langle \nabla f(X(t)), V(t)\rangle, \end{aligned} \tag{12}$$

where last inequality is due to the strong convexity. Plugging the value of c_1, c_2 and c_3, we have

$$\frac{d}{dt}\mathcal{E}(t) \leq -\frac{q(n+qm)}{2}\Big(\big(f(X(t)) - f(x^*)\big) + \frac{\mu}{2}\|X(t)-x^*\|^2 + n\|V(t)\|^2\Big). \tag{13}$$

Besides, the mixing term can be upper-bounded by

$$\mathcal{E}_3(t) \leq 2q^2\|X(t)-x^*\|^2 + 2n^2\|V(t)\|^2. \tag{14}$$

Therefore we have $\mathcal{E}(t)$ satisfying

$$\mathcal{E}(t) \leq (qm+n)\big(f(X(t)) - f(x^*)\big) + q^2\|X(t) - x^*\|^2/2 + \Big(n^2/2 + \frac{n(n+qm)}{4}\Big)\|V(t)\|^2, \tag{15}$$

which implies

$$\frac{d}{dt}\mathcal{E}(t) \leq -\min\left\{\frac{\mu(n+qm)}{2q}, \frac{q}{2}\right\} \cdot \mathcal{E}(t). \tag{16}$$

We then conclude using Gronwall's lemma [11]. □

B Proof for Theorem 3

For convenience of the reader, we repeat here the semi-implicit integrator of GM-ODE we seek to study:

$$\text{(SIE)}: \quad \begin{cases} x_{k+1} - x_k = -m\sqrt{s}\nabla f(x_k) - n\sqrt{s}v_k \\ v_{k+1} - v_k = \sqrt{s}\nabla f(x_{k+1}) - q\sqrt{s}v_k. \end{cases}$$

In compact notation, the second iteration can be written as

$$r_1(v_{k+1} - v_k) = \sqrt{s}\nabla f(x_{k+1}) - q\sqrt{s}v_{k+1} \tag{17}$$

or

$$r_1 v_k = v_{k+1} - \sqrt{s}\nabla f(x_{k+1}), \tag{18}$$

where $r_1 = 1 - q\sqrt{s}$.

Theorem 3 (Convergence of SIE). *Assume f is L-smooth and μ-strongly-convex. Let $(x_k)_{k=1}^{\infty}$ be the sequence obtained from semi-implicit discretization of GM-ODE with step $\sqrt{s}$. Let*

$$0 < m\sqrt{s} \leq \frac{1}{2L}, \ 0 < ns \leq m\sqrt{s}, \ 0 < q\sqrt{s} \leq \frac{1}{2}. \tag{4}$$

There exists a constant $C > 0$ such that, for any $k \in \mathbb{N}$, it holds that

$$f(x_k) - f(x^*) \leq \big(1 + \gamma_2\sqrt{s}\big)^{-k} C,$$

where $\gamma_2 := \frac{1}{5}\min\left(\frac{n\mu}{q}, \frac{q}{1+q^2/(nL)}\right)$.

Proof. We propose the discrete Lyapunov function defined as

$$\mathcal{E}(k) = r_1 r_2(f(x_k) - f(x^*)) + \frac{1}{4}\|q(x_{k+1} - x^*) - nr_1 v_k\|^2 + \frac{nr_1^2 r_2}{4}\|v_k\|^2 - \frac{r_1 r_2 m\sqrt{s}}{2}\|\nabla f(x_k)\|^2. \tag{19}$$

We use colors for different parts to keep track of related terms in the derivation. As the first step, thanks to L-Lipshitz smoothness, we have

$$\begin{aligned} f(x_{k+1}) - f(x_k) \leq& \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle - \frac{1}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\ =& -m\sqrt{s}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle - n\sqrt{s}\langle v_k, \nabla f(x_{k+1})\rangle \\ & - \frac{1}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2. \end{aligned} \tag{20}$$

We proceed by computing the difference in $\mathcal{E}$ in two subsequent iterations. Denote $r_2 = n + mq$, we have

$$\mathcal{E}(k+1) - \mathcal{E}(k) \overset{(A)}{\leq} -r_1 r_2 m\sqrt{s}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle - r_1 r_2 n\sqrt{s}\langle v_k, \nabla f(x_{k+1})\rangle$$

$$
\begin{aligned}
&- \frac{r_1 r_2}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + \frac{1}{4}\|q(x_{k+2} - x_{k+1}) - nr_1(v_{k+1} - v_k)\|^2 \\
&+ \frac{1}{2}\langle q(x_{k+2} - x_{k+1}) - nr_1(v_{k+1} - v_k), q(x_{k+1} - x^*) - nv_{k+1} + n\sqrt{s}\nabla f(x_{k+1})\rangle \\
&+ \frac{nr_1^2 r_2}{4}\|v_{k+1}\|^2 - \frac{nr_2}{4}\|v_{k+1} - \sqrt{s}\nabla f(x_{k+1})\|^2 \\
&- \frac{r_1 r_2 m\sqrt{s}}{2}\Big(\|\nabla f(x_{k+1})\|^2 - \|\nabla f(x_k)\|^2\Big) \\
\overset{(B)}{=} &-r_1 r_2 m\sqrt{s}\langle \nabla f(x_k), \nabla f(x_{k+1})\rangle - r_1 r_2 n\sqrt{s}\langle v_k, \nabla f(x_{k+1})\rangle \\
&- \frac{r_1 r_2}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \frac{r_2(2n - r_2)}{4}s\|\nabla f(x_{k+1})\|^2 \\
&- \frac{r_2}{2}\sqrt{s}\langle \nabla f(x_{k+1}), q(x_{k+1} - x^*) - nv_{k+1}\rangle \\
&- \frac{nr_2(1 - r_1^2)}{4}\|v_{k+1}\|^2 - \frac{nr_2}{4}s\|\nabla f(x_{k+1})\|^2 + \frac{nr_2}{2}\sqrt{s}\langle \nabla f(x_{k+1}), v_{k+1}\rangle \\
&- \frac{r_1 r_2 m\sqrt{s}}{2}\Big(\|\nabla f(x_{k+1})\|^2 - \|\nabla f(x_k)\|^2\Big) \\
\overset{(C)}{=} &nr_2\sqrt{s}\langle \nabla f(x_{k+1}), v_{k+1}/2 + v_{k+1}/2 - r_1 v_k\rangle \\
&+ \frac{r_1 r_2}{2}m\sqrt{s}\Big(\|\nabla f(x_{k+1})\|^2 - 2\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle + \|\nabla f(x_k)\|^2\Big) \\
&- \Big(\frac{r_2(2n - r_2)}{4}s + \frac{nr_2}{4}s + r_1 r_2 m\sqrt{s}\Big)\|\nabla f(x_{k+1})\|^2 - \frac{nr_2(1 - r_1^2)}{4}\|v_{k+1}\|^2 \\
&- \frac{r_1 r_2}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \frac{r_2}{2}q\sqrt{s}\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle.
\end{aligned} \tag{21}
$$

In step (A), we use smoothness of f as stated in Eq. (20) for the blue term. Also, we used the inequality $\|a\|^2 - \|b\|^2 = \|a - b\|^2 + 2\langle a - b, b\rangle$ where $a = q(x_{k+2} - x^*) - nr_1 v_k$ and $b = q(x_{k+1} - x^*) - nr_1 v_k$ to obtain the red term. In particular,

$$
\begin{aligned}
a - b &= q(x_{k+2} - x_{k+1}) - nr_1(v_{k+1} - v_k) \\
&= -mq\sqrt{s}\nabla f(x_{k+1}) - nq\sqrt{s}v_{k+1} - n\sqrt{s}\nabla f(x_{k+1}) + nq\sqrt{s}v_{k+1} \\
&= -r_2\sqrt{s}\nabla f(x_{k+1}).
\end{aligned} \tag{22}
$$

In step (B), we incorporate the recurrence of SIE. Step (C) is a simple re-arrangement of terms.

We can easily verify the following identities:

$$
\sqrt{s}\langle \nabla f(x_{k+1}), v_{k+1} - r_1 v_k\rangle = s\|\nabla f(x_{k+1})\|^2 \tag{23}
$$

and

$$
\|\nabla f(x_{k+1})\|^2 - 2\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle + \|\nabla f(x_k)\|^2 = \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2. \tag{24}
$$

We have

$$
\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) \leq &nr_2 s\|\nabla f(x_{k+1})\|^2 + \frac{r_1 r_2}{2}m\sqrt{s}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\
&- r_2 s\Big(\frac{2n - r_2}{4} + \frac{n}{4} + \frac{r_1 m}{\sqrt{s}}\Big)\|\nabla f(x_{k+1})\|^2 - \frac{nr_2(1 - r_1^2)}{4}\|v_{k+1}\|^2 \\
&- \frac{r_1 r_2}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \frac{r_2}{2}q\sqrt{s}\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle.
\end{aligned} \tag{25}
$$

We leverage μ-strong convexity of f to get

$$
\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle \geq f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x_{k+1} - x^*\|^2. \tag{26}
$$

Applying the above inequality to the last term of Eq. (25), we obtain

$$
\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\frac{r_2}{2}q\sqrt{s}(f(x_{k+1}) - f(x^*)) - \frac{r_2\mu}{4}q\sqrt{s}\|x_{k+1} - x^*\|^2
$$

$$- \frac{r_1 r_2}{2}(1/L - m\sqrt{s})\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 - \frac{n r_2(1-r_1^2)}{4}\|v_{k+1}\|^2$$
$$- r_2 s\Big(\frac{2n-r_2}{4} + \frac{n}{4} + \frac{r_1 m}{\sqrt{s}} - n\Big)\|\nabla f(x_{k+1})\|^2. \tag{27}$$

Now we plug in the the value of r_1, r_2 and calculate

$$1 - r_1^2 = 1 - (1 - q\sqrt{s})^2 = q\sqrt{s}(2 - q\sqrt{s}) \geq q\sqrt{s}, \tag{28}$$

where we used the condition $q\sqrt{s} \leq 1/2$. Next, since $m\sqrt{s} \leq 1/(2L)$, $n \leq m/\sqrt{s}$ and $r_1 = 1 - q\sqrt{s} \geq 1/2$, it holds that

$$\frac{2n - r_2}{4} + \frac{r_1 m}{\sqrt{s}} - \frac{3n}{4} = \frac{n - mq}{4} + \frac{r_1 m}{\sqrt{s}} - \frac{3n}{4} = \frac{r_1 m}{\sqrt{s}} - \frac{n}{2} - \frac{mq}{4} \geq -\frac{mq}{4}. \tag{29}$$

Hence, the difference between two iterations can be upper-bounded as follows:

$$\begin{aligned}
\mathcal{E}(k+1) - \mathcal{E}(k) \leq & - \frac{r_2 q\sqrt{s}}{2}\Big(f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x_{k+1} - x^*\|^2 + n\|v_{k+1}\|^2/2 - \frac{m\sqrt{s}}{2}\|\nabla f(x_{k+1})\|^2\Big) \\
= & - \frac{r_2 q\sqrt{s}}{2}\Big((1-r_3)[f(x_{k+1}) - f(x^*)] + \frac{\mu}{2}\|x_{k+1} - x^*\|^2 \\
& + n\|v_{k+1}\|^2/2 + r_3[f(x_{k+1}) - f(x^*) - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2]\Big),
\end{aligned} \tag{30}$$

where $r_3 = Lm\sqrt{s} \leq 1/2$ and the bound remains legal since $1 - r_3 \geq 1/2$.

On the other hand, our candidate Lyapunov function at iteration k itself can be upper-bounded as

$$\begin{aligned}
\mathcal{E}(k) &= r_1 r_2(f(x_k) - f(x^*) + \frac{1}{4}\|q(x_{k+1} - x^*) - n r_1 v_k\|^2 + \frac{n r_1^2 r_2}{4}\|v_k\|^2 - \frac{r_1 r_2 m\sqrt{s}}{2}\|\nabla f(x_k)\|^2 \\
&\overset{(A)}{=} r_1 r_2(f(x_k) - f(x^*)) + \frac{1}{4}\|q(x_k - x^*) - n v_k - mq\sqrt{s}\nabla f(x_k)\|^2 + \frac{n r_1^2 r_2}{4}\|v_k\|^2 \\
&\quad - r_1 r_2 m\sqrt{s}\|\nabla f(x_k)\|^2/2 \\
&\overset{(B)}{\leq} r_1 r_2(f(x_k) - f(x^*)) + q^2\|x_k - x^*\|^2 + n^2\|v_k\|^2 + \frac{q^2 m^2 s}{2}\|\nabla f(x_k)\|^2 + \frac{n r_1^2 r_2}{4}\|v_k\|^2 \\
&\quad - r_1 r_2 m\sqrt{s}\|\nabla f(x_k)\|^2/2 \\
&= r_1 r_2(1 - r_3 + r_4)(f(x_k) - f(x^*)) + q^2\|x_k - x^*\|^2 + (n^2 + n r_1^2 r_2/4)\|v_k\|^2 \\
&\quad + r_1 r_2(r_3 - r_4)[f(x_k) - f(x^*) - \frac{1}{2L}\|\nabla f(x_k)\|^2],
\end{aligned} \tag{31}$$

with $r_4 = Lq^2 m^2 s/(r_1 r_2)$. Precisely, step (A) is obtained by replacing SIE update for the term x_{k+1}. (B) is obtained by repeatedly using the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Finally, noting that $f(x_k) - f(x^*) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2$, we have

$$\begin{aligned}
\mathcal{E}(k) \leq r_2\Big(& r_1(1 - r_3 + r_4)[f(x_k) - f(x^*)] + \frac{q^2}{n}\|x_k - x^*\|^2 + 5n\|v_k\|^2/4 \\
& + r_1(r_3 - r_4)[f(x_k) - f(x^*) - \frac{1}{2L}\|\nabla f(x_k)\|^2]\Big),
\end{aligned} \tag{32}$$

since $r_2 = n + mq \geq n$. It is reckoned that $\mathcal{E}(k+1) - \mathcal{E}(k)$ and $\mathcal{E}(k)$ share identical parts except for different coefficients. Now we aim at obtaining following inequality

$$\mathcal{E}(k+1) - \mathcal{E}(k) \leq -\gamma_2\sqrt{s}\mathcal{E}(k+1). \tag{33}$$

To achieve this, γ_2 should be the minimal ratio for coefficients of each parts of $\mathcal{E}(k+1) - \mathcal{E}(k)$ to those of $\mathcal{E}(k)$. It is easy then to notice that γ_2 should be smaller than $q/5$ and $n\mu/(4q)$. Besides it should also hold that

$$\frac{r_2 q}{2 r_1 r_2}\frac{r_3}{r_3 - r_4} \geq \frac{q}{2}\frac{1 - r_3}{1 - (r_3 - r_4)} = \frac{q}{2}\frac{1 - r_3}{1 - r_3(1 - r_4/r_3)} \geq \frac{q}{2}\frac{1 - 1/2}{1 - 1/2(1 - \frac{q^2}{nL})} \geq \frac{q}{2}\frac{1}{1 + \frac{q^2}{nL}} \geq \gamma_2, \tag{34}$$

due to the fact $\frac{r_4}{r_3} = \frac{q^2 m\sqrt{s}}{r_1 r_2} \leq \frac{q^2}{nL}$ and $r_3 \leq 1/2$. Therefore $\gamma_2 = \frac{1}{5}\min\{\frac{q}{1 + \frac{q^2}{nL}}, \frac{n\mu}{q}\}$ satisfies the above inequality and completes the proof. □

We now use the above result to prove the convergence of QHM iterations (see Section 4).

Corollary 5 (Convergence of QHM)**.** *Assume f is L-smooth and μ-strongly-convex with $L/\mu \geq 9$. The iterates of QHM enjoy a linear convergence rate for $s \leq \frac{1}{4L}$ and $a \leq 1/2$. In particular, QHM also enjoys convergence rate $O((1-\sqrt{\mu/L})^k)$ for $b = 1 - 2\sqrt{\mu s}$. Namely, there exists a constant $C > 0$ such that $f(x_k) - f(x^*) \leq \left(1 + a\sqrt{\mu s}/10\right)^{-k} C$.*

Proof. First, we show how one can alternatively write QHM as one-line scheme. The original QHM algorithm is reported here for convenience of the reader

$$\begin{cases} x_{k+1} = x_k - s((1-a)\nabla f(x_k) + a g_{k+1}) \\ g_{k+1} = b g_k + \nabla f(x_k). \end{cases} \tag{QHM}$$

We replace the second line of QHM into the first one :

$$x_{k+1} = x_k - s(1-a)\nabla f(x_k) - s \cdot b \cdot a \cdot g_k - as\nabla f(x_k). \tag{35}$$

Using the first iterate we get:

$$-(x_k - x_{k-1})/s - (1-\alpha)\nabla f(x_{k-1}) = a g_k. \tag{36}$$

Replacing this into the result of first equation, we get:

$$x_{k+1} = x_k - s(1-a)\nabla f(x_k) + b((x_k - x_{k-1}) + s(1-a)\nabla f(x_{k-1})) - as\nabla f(x_k). \tag{37}$$

By rearrangment, we finally obtain

$$x_{k+1} = x_k + b(x_k - x_{k-1}) - s\nabla f(x_k) + sb(1-a)\nabla f(x_{k-1}). \tag{38}$$

The above iterates can be viewed as SIE discretization of GM-ODE with the following specific choice of parameters (see the single sequence of iterates of SIE in the last section):

$$m = (1-a)\sqrt{s}, \quad n = a, \quad q = \frac{1-b}{\sqrt{s}}. \tag{39}$$

Invoking Thm. 3, we get the convergence rate for QHM. More precisely, choosing $b = 1 - 2\sqrt{\mu s}$ we obtain

$$q = 2\sqrt{\mu}. \tag{40}$$

The above choice of parameters obeys the constraints in Thm. 3:

$$m\sqrt{s} = (1-a)s < s \leq \frac{1}{4L}, \quad n = a \leq (1-a) = \frac{m}{\sqrt{s}}, \tag{41}$$

and

$$q\sqrt{s} = 2\sqrt{\mu s} \leq 2\sqrt{sL/9} \leq 1/3, \tag{42}$$

since we assumed $s \leq 1/(4L)$, $a \leq 1/2$ and $L/\mu \geq 9$. The rate — thanks to Thm. 3 — is determined by $\gamma_2 = \frac{1}{5}\min\{\frac{n\mu}{q}, \frac{q}{1+q^2/(nL)})\}$. We conclude the proof by showing that $\gamma_2 = a\sqrt{\mu}/8$ in the case of QHM. First, one can readily check that $(n\mu)/(5q) = a\mu/(10\sqrt{\mu})$ holds due to the choice of parameters. Second, with some patience, one can check that the following chain of inequality holds:

$$\frac{1}{5} \cdot \frac{q}{1 + \frac{q^2}{nL}} = \frac{2\sqrt{\mu}}{5 + \frac{20\mu}{aL}} \geq \frac{2a\sqrt{\mu}}{5a + 20/9} \geq \frac{a\sqrt{\mu}}{10}.$$

□

C Proofs for Section 5

As stated in the main paper, we consider the following discretization errors:

$$\Delta_k^{(\mathrm{EE})} := \|X(k\sqrt{s}) - x_k\|, \quad x_k \text{ obtained by EE}$$
$$\Delta_k^{(\mathrm{SIE})} := \|X(k\sqrt{s}) - x_{k+1}\|, \quad x_k \text{ obtained by SIE.}$$

We define $w_k := x_k$ for EE and $w_k := x_{k+1}$ for SIE. We compare the error $\Delta_k = \|X(k\sqrt{s}) - w_k\|$ for $k = 1$ in the next lemma, assuming $\Delta_0 = 0$ and $v_0 = V(0)$. This is also called local (or one-step) integration error.

Lemma 7. *Let f be L-smooth and of class C^2. If $m = O(s^{1/2})$, then $\Delta_1^{(SIE)} = O(s^{3/2})$ and $\Delta_1^{(EE)} = O(s)$.*

Proof. We introduce the notation $X_k := X(k\sqrt{s})$, $V_k := V(k\sqrt{s})$. Our problem setting requires $w_0 = X_0$ and $v_0 = V_0$. For SIE, $w_k = x_{k+1}$ and we begin from Taylor expansion of X as

$$X_1 - X_0 = \sqrt{s}\dot{X}_0 + s\ddot{X}_0 + O(s^{3/2}), \tag{43}$$

and therefore

$$\begin{aligned}
X_1 - w_1 =& X_1 - X_0 - (w_1 - w_0) + X_0 - w_0 \\
=& X_1 - X_0 - (x_2 - x_1) + X_0 - x_1 \\
=& \sqrt{s}\dot{X}_0 + s\ddot{X}_0 + m\sqrt{s}\nabla f(x_1) + n\sqrt{s}v_1 + O(s^{3/2}) \\
=& \sqrt{s}\Big(- m\nabla f(X_0) - nV_0\Big) + s\frac{d}{dt}\Big(- m\nabla f(X_0) - nV_0\Big) \\
&+ m\sqrt{s}\nabla f(x_1) + n\sqrt{s}\Big(\sqrt{s}\nabla f(x_1) + (1 - q\sqrt{s})v_0\Big) + O(s^{3/2}).
\end{aligned} \tag{44}$$

where in the third equality we used the fact that, by hypothesis, $X_0 - x_1 = 0$. And in particular, since $\frac{d\nabla f(X)}{dt} = \nabla^2 f(X)\dot{X}$,

$$\begin{aligned}
s\frac{d}{dt}\Big(- m\nabla f(X_0) - nV_0\Big) =& - sm\nabla^2 f(X_0)\dot{X}_0 - sn\dot{V}_0 \\
=& - sn\nabla f(X_0) + snqV_0 + sm^2\nabla^2 f(X_0)\nabla f(X_0) + smn\nabla^2 f(X_0)V_0.
\end{aligned} \tag{45}$$

Then it holds that

$$X_1 - w_1 = - (m\sqrt{s} + ns)\Big(\nabla f(X_0) - \nabla f(x_1)\Big) - n\sqrt{s}(V_0 - v_0) + snq(V_0 - v_0) + O(s^{3/2}) \le O(s^{3/2}) \tag{46}$$

and $\Delta_1^{(\mathrm{SIE})} \le O(s^{3/2})$.

We proceed with the EE iterations (remember: $w_k = x_k$). We expand $\Delta_1^{(\mathrm{EE})}$ as

$$\begin{aligned}
X_1 - w_1 =& X_1 - X_0 - (w_1 - w_0) + X_0 - x_0 + O(s^{3/2}) \\
=& X_1 - X_0 - (x_1 - x_0) + X_0 - x_0 + O(s^{3/2}) \\
=& \sqrt{s}\dot{X}_0 + s\ddot{X}_0 + m\sqrt{s}\nabla f(x_0) + n\sqrt{s}v_0 + O(s^{3/2}) \\
=& \sqrt{s}\Big(- m\nabla f(X_0) - nV_k\Big) + m\sqrt{s}\nabla f(x_0) + n\sqrt{s}v_0 \\
&+ sm\Big(m\nabla^2 f(X_0)\nabla f(X_0) + n\nabla^2 f(X_k)V_0\Big) \\
&- sn\Big(\nabla f(X_0) - qV_0\Big) + O(s^{3/2}) \\
=& - m\sqrt{s}\Big(\nabla f(X_0) - \nabla f(x_0)\Big) - n\sqrt{s}\Big(V_0 - v_0\Big) + O(s).
\end{aligned} \tag{47}$$

Therefore, we conclude that $\Delta_1^{(\mathrm{EE})} \le O(s)$. □

Lemma 8. *Suppose f is μ-strongly-convex and L-smooth. For EE discretization of GM-ODE obeying Eq. (5), the discretization error decays as $\Delta_k^{(EE)} = O((1 + \gamma_3\sqrt{s})^{-k})$ where γ_3 is defined in Thm. 6. Furthermore, SIE also enjoys $\Delta_k^{(SIE)} = O((1 + \gamma_2\sqrt{s})^{-k})$ where γ_2 is defined in Thm. 3 as long as conditions in Eq. (4) are satisfied.*

Proof. The proof is based on the following consequence of strong convexity

$$\mu\|x - x^*\|^2/2 \leq f(x) - f(x^*). \tag{48}$$

Using the above inequality together with a straightforward application of triangular inequality we complete the proof:

$$\begin{aligned}\|X(k\sqrt{s}) - x_k\| &= \|X(k\sqrt{s}) - x^* + x^* - x_k\| \\ &\leq \|X(k\sqrt{s}) - x^*\| + \|x_k - x^*\| \\ &\leq \sqrt{2}\mu^{-1/2}\left(\left(f(X(k\sqrt{s})) - f(x^*)\right)^{1/2} + \left(f(x_k) - f(x^*)\right)^{1/2}\right).\end{aligned} \tag{49}$$

Replacing the convergence results in Thm. 1, 3, and 6 into the the above bound concludes the proof. □