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

Peiyuan Zhang* ETH Zürich
Antonio Orvieto ETH Zürich
Hadi Daneshmand INRIA Paris, ETH Zürich
Thomas Hofmann ETH Zürich
Roy Smith ETH Zürich

Abstract

Viewing optimization methods as numerical integrators for ordinary differential equations (ODEs) provides a thought-provoking modern framework for studying accelerated first-order optimizers. In this literature, acceleration is often supposed to be linked to the quality of the integrator (accuracy, energy preservation, symplecticity). In this work, we propose a novel ordinary differential equation that questions this connection: both the explicit and the semi-implicit (a.k.a symplectic) 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 for the integration of physical systems, 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 (1964):

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

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 $\beta \in [0, 1)$ the momentum parameter. Using a novel and beautiful argument on fixed point iterations, Polyak (1964) 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 Nemirovsky and Yudin (1983), many researchers in the early 80s tried to develop an algorithm with a *global* accelerated convergence rate. The problem was solved by Nesterov (1983), who proposed the 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 articles are devoted to understanding the underlying mechanism (Allen-Zhu and Orecchia, 2014; Defazio, 2019; Ahn, 2020) and the role of the small yet crucial modification[5] compared to HB (Flammarion and Bach, 2015; Lessard et al., 2016; Hu and Lessard, 2017). Notwithstanding the theoretical value of these contributions, they are arguably only of a descriptive nature and leave open more fundamental questions on the reason behind acceleration.

Continuous-time models for acceleration. A new line of research bloomed from a seminal paper by Su et al. (2014). This work gained a lot of attraction,

*Correspondence to talantyeri@gmail.com.

[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 large, $1 - \sqrt{\mu/L} \ll 1 - \mu/L$.

[4] In the original paper Nesterov (1983), the algorithm is presented in a more general way. Our formulation is similar to Shi et al. (2018).

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

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

as it introduces[6] a powerful way to look at acceleration through the lens of second order ordinary differential equations (ODEs). In the μ-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 $O(e^{-\mu t})$. Other interesting properties of damped gradient systems such as NAG-ODE can be found in the (stochastic) optimization literature (Krichene et al., 2015; Xu et al., 2018; Cabot et al., 2009; Orvieto et al., 2020; Orvieto and Lucchi, 2019; Diakonikolas and Jordan, 2019; Alecsa et al., 2019; Alimisis et al., 2020), and in the applied mathematics literature (Sanz-Serna and Zygalakis, 2020; Attouch et al., 2000; Attouch and Alvarez, 2000; Alvarez, 2000; Bégout et al., 2015).

High-resolution ODEs. As first noted by Wilson et al. (2016), 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 vanishing 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. (2018) 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 capture the gradient correction $\beta s(\nabla f(x_k) - \nabla f(x_{k-1}))$. The term $\sqrt{s}\nabla^2 f(X)\dot{X}$ is referred to as Hessian damping, and can be seen as a curvature-dependent viscosity correction. As a validation for their new ODE, Shi et al. (2018) showed that NAG-ODE-HR enjoys the same accelerated rate of NAG-ODE — but it is empirically more faithful to NAG compared to NAG-ODE, for finite values of s.

Connection to numerical integration. In a second article, Shi et al. (2019) showed that NAG can be approximately recovered[7] through a semi-implicit (a.k.a. symplectic) Euler discretization of NAG-ODE-HR. The authors also claim that if the same system is integrated with the explicit Euler method, the resulting optimizer *might*[8] not be accelerated because it is only found stable for small values of s. Semi-implicit methods are well-known to perform remarkably well for integrating second-order ODEs in physics (Hairer et al., 2006) and chemistry (Lubich, 2008); namely, one can use big step-sizes while preserving the geometry of the original flow. *Shi et al. suggested that the essence of acceleration can be explained by the same phenomenon*, which is mathematically well understood in the Hamiltonian (i.e. energy-conserving) setting thanks to the theory of backward error analysis (Hairer, 1994; Benettin and Giorgilli, 1994).

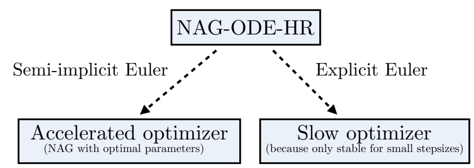

Figure 1: Sketch of the storyline of Shi et al. (2019): while semi-implicit discretization of NAG-ODE-HR yields an accelerated method, explicit discretization results in a method not known in the literature, which is claimed to be stable only for very small step-sizes ($s \leq O(\mu/L^2)$, compared to $s \leq O(1/L)$ of the semi-implicit method). This is used by the authors to advocate that the structure provided by semi-implicit integration is somethow critical for the construction of accelerated methods. This storyline (and the associated conclusion) is much different from ours, sketched in Fig. 2. Superiority of semi-implicit methods is also claimed/hinted in several works (Shi et al., 2018; Bravetti et al., 2019; Betancourt et al., 2018; França et al., 2020a; Muehlebach and Jordan, 2019).

On a parallel line, Muehlebach and Jordan (2019) derived a different continuous-time model that contains terms of the form $\nabla f(X + \sqrt{s}\dot{X})$ instead of $\sqrt{s}\nabla^2 f(X)\dot{X}$. This ODE can also relate to Nesterov's method through semi-implicit integration. Moreover, inspired by the variational perspective presented in Wibisono et al. (2016), many research papers (Betancourt et al., 2018; Muehlebach and Jordan, 2020; França et al., 2020a,b; Alecsa, 2020; Bravetti et al., 2019) have been devoted to understanding the geometric properties of Nesterov's method, seen as either (1)

[6]We point out that, actually, the differential equations proposed in Su et al. (2014) was already written down and partly analyzed in the original 1963 paper by Polyak (1964). 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 Gavurin (1958).

[7]See approximations in Sec. 2.2 of Shi et al. (2019).

[8]We point out here a potential problem in the main claim of Shi et al. (2019): the authors show that the explicit integrator of NAG-ODE-HR is stable only for small step-sizes by finding necessary conditions for the steady decrease of a particular energy function. While this fact surely hints at potential instabilities of the associated algorithm, it does not *per se* provide a sufficient condition for slow convergence. In other words, one could in principle find a different result by choosing a different Lyapunov function.

a (Strang/Lie-Trotter) splitting scheme for structure-preserving integration of conformal Hamiltonian systems (McLachlan and Perlmutter, 2001; McLachlan and Quispel, 2002) or (2) the composition of a map derived from a contact Hamiltonian (de León and Lainz Valcázar, 2019; Bravetti et al., 2017) and a gradient descent step. Finally, the application of Runge-Kutta schemes was explored (Zhang et al., 2018, 2019; Sanz Serna and Zygalakis, 2020); in particular, Zhang et al. (2018) first showed that fast rates can be also achieved via high-order explicit methods.

To sum it up, to the best of our knowledge, *all* recent convex optimization literature advocates that, in order to achieve acceleration from an ODE model, one needs to use either structure-preserving integrators (Bravetti et al., 2019; Shi et al., 2018; França et al., 2020a), high-order explicit methods (Zhang et al., 2018, 2019), or implicit methods (Shi et al., 2019; Wilson et al., 2016; Diakonikolas and Orecchia, 2017).

Our contribution. We show that, contrary to what is often claimed (or hinted at) in recent literature (see paragraph above) acceleration can also be achieved by means of simple low-order explicit numerical integrators — such as the explicit Euler method. While explicit Euler is well-known to be provably suboptimal for accurate integration of Hamiltonian systems (Hairer et al., 2006; Hairer, 1994; Benettin and Giorgilli, 1994), we show that this does not necessarily imply slow convergence of the resulting optimizer. In particular, our work suggests that the structure-preserving properties of semi-implicit (symplectic) methods are not a necessary component of accelerated algorithms.

We start by introducing a generalized momentum ODE (GM-ODE), dependent on three parameters, which recovers both NAG-ODE and NAG-ODE-HR as special cases. In Sec. 3 we study the convergence of this ODE. Next, in Sec. 4 we show that both the explicit and the semi-implicit Euler methods, when applied for numerical integration of GM-ODE, can achieve an accelerated rate. Finally, in Sec. 5, we go one step further and show that there exist damped gradient systems for which the semi-implicit Euler method is unstable, while the explicit Euler method (with the same step-size) is stable. Of course, for other ODE systems, we observe the opposite behavior. This showcases that the stability of the integrator depends on the underlying accelerated ODE.

At its core, our work showcases some unintuitive aspects of the connection between the fields of numerical integration and optimization. Namely, while for accurate integration of physical systems symplectic integrators are provably superior to explicit methods (Hairer et al., 2006; Benettin and Giorgilli, 1994), we show that the same ranking might not hold when seeking fast optimizers through ODE discretization. We think that clarifying this critical point makes the (already vast) literature on this topic richer, motivating future research on the connections between the fields of optimization and numerical analysis.

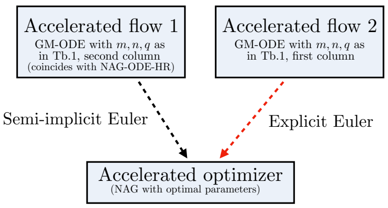

Figure 2: The main conceptual finding of our paper: both the semi-implicit and the explicit Euler integrators are able to recover Nesterov's method, when used to discretize different accelerated flows (ODEs are given in Sec. 4).

2 Summary of the 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 by non-negative parameters m, n, q:

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

We show that the above ODE includes both NAG-ODE and NAG-ODE-HR as special cases and recovers a large set of momentum methods through the application of two classical numerical integrators, i.e. semi-implicit and explicit Euler. In Lemma 2 we show that these integrators are equivalent and can both lead to acceleration (see also Fig. 2). Equivalence is shown by 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 $(m_{\text{EE}}, n_{\text{EE}}, q_{\text{EE}})$. These parameters can be computed in closed form starting from $(m_{\text{SIE}}, n_{\text{SIE}}, q_{\text{SIE}})$ (see Lemma 2 for precise formulas). Such an equivalence suggests that the energy-preservation properties of semi-implicit integration might no be strictly necessary to achieve acceleration, as instead hinted in recent works (Bravetti et al., 2017; McLachlan and Perlmutter, 2001; Shi et al., 2019).

To make our analysis complete, we establish an accelerated convergence rate (in Thm. 3 and Cor. 6) for a set of algorithms that can be interpreted as (both the two) Euler discretizations of GM-ODE with a proper parameters choice. As a side product of our novel analysis of semi-implicit and explicit methods, we derive a novel accelerated convergence rate for the quasi-hyperbolic momentum (QHM) method introduced by Ma and Yarats (2018). Indeed, along with HB and NAG, the

QHM method can also be seen as a numerical integrator on GM-ODE. QHM was shown to be very competitive in deep learning tasks (Choi et al., 2019) as well as in the strongly-convex setting (see Appendix J in (Ma and Yarats, 2018)). However, to the best of our knowledge, QHM has only been studied in the quadratic case (Gitman et al., 2019) (hence the novelty of our rate). We like to point out that this is not the main contribution of our paper, but is presented here nonetheless to showcases the flexibility of our novel ODE and of the numerical integration approach.

Finally, we go beyond convergence analysis and study the discretization errors in Sec. 5. Under some conditions 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 GM-ODE, in line with most related works on acceleration and numerical integration (Shi et al., 2018; Su et al., 2014). 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.

GM-ODE 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 is L-smooth, one can check that GM-ODE admits a unique solution (follows directly from Thm. 3.2 in Khalil and Grizzle (2002)). The model above is inspired by the quasi-hyperbolic momentum (QHM) algorithm[9] developed in Ma and Yarats (2018). We discuss the connection to QHM later in Sec. 4.

Connections to existing ODE models. GM-ODE recovers existing continuous momentum models under 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[10] $m = 0, n = 1, q = 2\sqrt{\mu}$ recovers NAG-ODE by Polyak (1964). Moreover, the choice $m = \sqrt{s}, n = 1, q = 2\sqrt{\mu}$ recovers NAG-ODE-HR, proposed by Shi et al. (2018, 2019). That is, GM-ODE includes as special cases both the high-resolution and low-resolution models of Nesterov's method (see discussion in the introduction). We note that, contrary to Shi et al. (2018), the Hessian of f is not explicitly included in the model. Also, contrary to Muehlebach and Jordan (2019), the gradient is evaluated only at the current position X. This feature arguably gives GM-ODE higher interpretability than existing models – a simple linear combination of gradient and momentum can also achieve high resolution[11].

	m	n	q
Gradient Flow	1	0	any
NAG-ODE (Su et al., 2014)	0	1	$2\sqrt{\mu}$
NAG-ODE-HR (Shi et al., 2019)	$\sqrt{s}$	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 Su et al. (2014); Shi et al. (2019), we propose the following energy function:

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

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

Theorem 1 (Continuous-time stability)**.** *Let f be 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*

$$\mathcal{E}(X(t), V(t)) \leq e^{-\gamma_1 t} \cdot \mathcal{E}(X(0), V(0)), \quad (2)$$

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

Remarkably, the stability analysis in the proof 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.

Remark 1. The rate in Thm. 1 is not affected by the gradient Lipshitz constant L. This might look strange

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

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

[11] That is, a finer, compared to the original ODE in Su et al. (2014) approximation of Nesterov's method. For a detailed discussion on this terminology, we refer the reader to Shi et al. (2018).

at first for a reader familiar with the optimization literature. However, we point to the fact that this is a feature of most continuous-time models (see e.g. rates in Shi et al. (2018)). The Lipschitz constant comes back into the rate after discretization, since one has to introduce a bound on the maximum integrator step-size, usually proportional to $1/L$ (see Eq. 3 and 4).

How do m, n, q 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 Shi et al. (2018) 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 n), momentum (controlled by q) and Hessian damping (controlled by m). Indeed, given the expression for γ_1 in Eq. 2, 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 (Shi et al., 2018, 2019; Muehlebach and Jordan, 2019) $q = O(\sqrt{\mu})$ (see the first panel of Fig. 3). Sanz-Serna and Zygalakis (2020) recently showed that this choice is optimal using the linear matrix inequalities framework (Lessard et al., 2016; Fazlyab et al., 2018).
- 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 (and adapting m accordingly) the convergence in continuous-time can be sped-up arbitrarily (see the second panel of Fig. 3).
- 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. 3).

Remark 2. If n or m are increased, one can guarantee arbitrarily fast convergence to the minimizer. This result only holds true in continuous-time, as noted also in a similar setting by Wilson et al. (2016). 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

We now jump to the discrete world and 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 (Hairer et al., 2006) and practical first-order

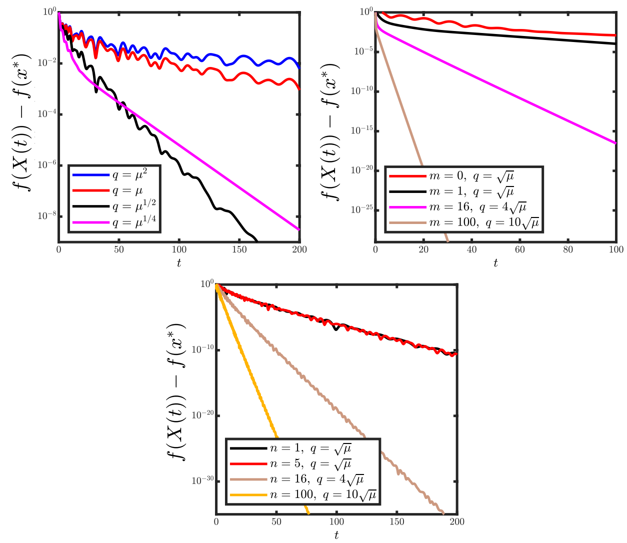

Figure 3: Role of parameters in GM-ODE. The objective function is a 10-dimensional quadratic function f with $\mu I \preceq \nabla^2 f \preceq LI$ where $\mu = 0.01$ and $L = 1$. The panels depict, from left to right, the influence of q, m and n, as suggested in above discussion. In each figure we vary 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}.

numerical integration schemes applied to GM-ODE with discretization step-size $\sqrt{s}$ (see discussion in Su et al. (2014); Shi et al. (2019)): Explicit Euler (EE) and Semi-Implicit[12] 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 shall 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, both schemes can be written in one line:

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

[12] 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 the sake of our discussion on equivalence. For a more detailed discussion, we refer the reader to Chapter 1 of Hairer et al. (2006).

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

Remarkably, different choices of parameters yield a rich set of momentum methods, and the reader can probably already notice some configurations which recover well-known optimizers (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 also be seen as explicit discretization of a different accelerated flow.

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 $n \neq 0$:

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

We substitute 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 Tb. 1. 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 EE and SIE.

An ODE which gives NAG under the explicit Euler method. From Tb. 1 and Eq. 1, we get that

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

leads to NAG through EE (choosing $q = 2\sqrt{\mu}$), while

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

recovers NAG through SIE discretization. These parameter choices lead to acceleration (see Cor. 4). Note

	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.

that the last equation is equivalent[13] to NAG-ODE-HR, while the first is not known in the literature. However, Thm. 1 ensures that both ODEs are accelerated. This is enough to show that the sketch in Fig. 2 is correct.

4.2 Semi-implicit Euler is accelerated

Leveraging insights from the ODE stability analysis in Thm. 1 and the lessons learned from semi-implicit Lyapunov function design in recent literature Shi et al. (2018, 2019), our next result establishes a general convergence rate for the semi-implicit Euler method on GM-ODE. In the next subsection, we also provide a similar result for EE, using Lemma 2.

Theorem 3 (Convergence of SIE)**.** *Assume f 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*

$$0 < m\sqrt{s} \leq \frac{1}{2L}, \ 0 < ns \leq m\sqrt{s}, \ 0 < q\sqrt{s} \leq \frac{1}{2}. \tag{3}$$

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 following energy function inspired by the ODE 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

[13] The careful reader might notice a factor $1 + \sqrt{\mu s}$ in front of the gradient for the ODE in Shi et al. (2019). This small difference is only due to the particular definition of semi-implicit integration. If one replaces $q\sqrt{s}v_k$ in the RHS of SIE with $q\sqrt{s}v_{k+1}$, then we have complete equivalence.

the discretization error (cf. Shi et al. (2019)). We show $\mathcal{E}(k+1)-\mathcal{E}(k) \leq -\gamma_2\sqrt{s}\mathcal{E}(k+1)$ in App. B, completing the proof. □

The generality of the convergence result allows us to derive accelerated rates for different momentum methods whose convergence rates may even be unknown. We illustrate this by deriving the well-known 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 Shi et al. (2019) are less general since are limited to high/low resolution or to a fixed viscosity $O(\sqrt{\mu})$.

From Thm. 3 to the well-known rate for NAG. By invoking Thm. 3, we can recover acceleration of NAG since it can be written as SIE discretization of GM-ODE (see Tb. 1).

Corollary 4 (NAG is accelerated)**.** *Let f be L-smooth and μ-strongly-convex with large*[14] *condition number $L/\mu \geq 9$. 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 Tb. 1). The algorithm enjoys the accelerated convergence rate $O((1-\sqrt{\mu/L})^k)$. Namely, $\exists C > 0$ such that*

$$f(x_k) - f(x^*) \leq \left(1+\sqrt{\mu s}/15\right)^{-k} C.$$

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

From Thm. 4 to a new rate for QHM. The generality of our model and our discretization analysis provides an accelerated convergence rate for a broad class of momentum methods. Among these methods is quasi-hyperbolic momentum (Ma and Yarats, 2018), which[15] shows promises in optimization for neural nets (Choi et al., 2019).

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

where $a, b \in (0,1)$. For classification tasks, QHM yields an accelerated rate on real-world datasets (even better than NAG) (Ma and Yarats, 2018). Despite empirical benefits, the convergence analysis for this algorithm is limited to quadratics (Gitman et al., 2019). Using

[14]The lower bound assumption for conditional number here and in Cor. 5 is purely technical and only serves for a simple illustration of these corollaries.

[15]Ma and Yarats (2018) 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 present here by factor rescaling.

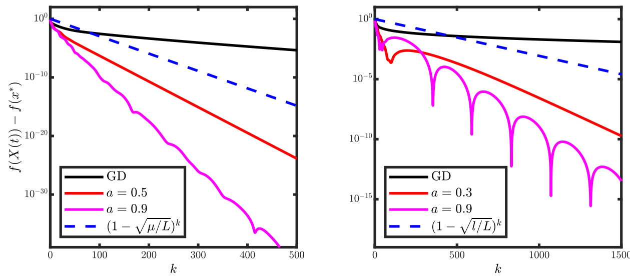

Figure 4: Convergence of QHM. The left plot shows the convergence of 10-dimensional quadratic with $\mu = 0.01$ and $L = 1$; the right plot reports 10-dimensional regularized logistic regression (random data and labels) with regularization weight $l = 10^{-4}$. We specifically used $s = 0.5$, $\beta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

Thm. 3, the next corollary establishes an accelerated rate for QHM (proof in the appendix).

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

$$f(x_k) - f(x^*) \leq \left(1 + a\sqrt{\mu s}/10\right)^{-k} C.$$

Fig. 4 shows the accelerated rate established in the corollary, and its dependency on the parameter a. We leave the extension to the stochastic case (possible with the methodology in Assran and Rabbat (2020)) to future work, for the sake of continuing our discussion on numerical integration.

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 integration. This is not surprising since the Lyapunov functions used in the literature to prove acceleration for NAG often differ from the one used for convergence of HB (see Eq. 3.3 in Shi et al. (2018)). Nonetheless, It is possible to construct an analogue of Thm. 3, using a different Lyapunov function, to derive (non accelerated) convergence for an HB-like method.

The trade-off speed-stability. As noted in Remark 2, 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 Nemirovsky and Yudin (1983)): for a specific discretization step-size $\sqrt{s}$, Eq. 3 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 step-size. Hence, as expected by the classic theory of convex optimization (Nemirovsky and Yudin, 1983), there is a sweet spot which yields $\gamma_2 = O(\sqrt{\mu/L})$ — a.k.a acceleration.

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 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*

$$\begin{aligned} &0 < m\sqrt{s} - ns/(1-q\sqrt{s}) \leq \frac{1}{2L}, \\ &0 < ns \leq \frac{1-q\sqrt{s}}{2} m\sqrt{s}, \quad 0 < q\sqrt{s} \leq \frac{1}{2}. \end{aligned} \tag{4}$$

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

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

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

Proof. Consider an explicit method with parameters $(m_{\mathrm{EE}}, n_{\mathrm{EE}}, q)$ and a semi-implicit method with parameters $(m_{\mathrm{SIE}}, n_{\mathrm{SIE}}, q)$. Thm. 3 holds if $0 < m_{\mathrm{SIE}}\sqrt{s} \leq 1/(2L), 0 < sn_{\mathrm{SIE}} \leq m_{\mathrm{SIE}}\sqrt{s}$ 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 $n_{\mathrm{EE}} = (1-q\sqrt{s})n_{\mathrm{SIE}}$ and $m_{\mathrm{EE}} = m_{\mathrm{SIE}} + \sqrt{s}n_{\mathrm{SIE}}$. Combining these conditions with the theorem requirements on n_{SIE}, we get:

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

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

Stability of EE and SIE. For the integration of Hamiltonian systems, semi-implicit Euler is provably more stable than explicit Euler (Hairer et al., 2006). For example, a linearized pendulum integrated with explicit Euler diverges in phase space, while the semi-implicit Euler method is stable and conserves the structure of the ODE system (energy, volume). In Fig. 5, we show that for a dissipative (hence not Hamiltonian) system such as GM-ODE the situation can be very different: in complete agreement with our equivalence result in Lemma 2, there exists parameter configurations for which EE is stable but SIE is not, and vice versa.

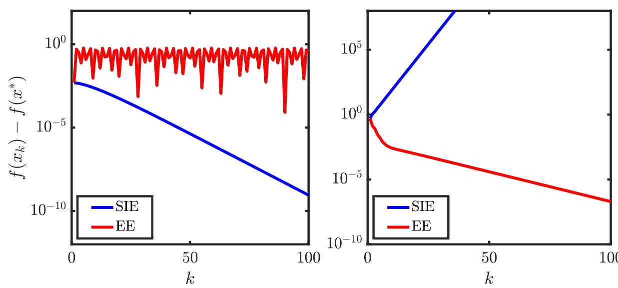

Figure 5: EE vs. SIE. To show that SIE and EE are neither superior nor inferior to each other, in each subplot, we use the same parameters m, n, q for both SIE and EE discretization. We observe very different behaviours. This suggests the stability and convergence is determined by the joint choice of parameters and numerical integrator together. The objective function here is a 2-dimensional quadratic with $\mu = 0.01, L = 1$ and the step-size is $s = 1$. 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 (Wilson et al., 2016; Shi et al., 2019; Muehlebach and Jordan, 2019, 2020) claims that semi-implicit integration *is somehow more natural* for the approximation of partitioned dissipative systems such as GM-ODE. Indeed, recent works (França et al., 2020a; Muehlebach and Jordan, 2020) showed that the geometric properties of semi-implicit methods combined with backward error analysis (Hairer et al., 2006) can be used to successfully prove the preservation of continuous-time rates of convergence up to a controlled error. Instead, our results in Thm. 6 show that explicit Euler discretization — of a proper ODE — also leads to an accelerated method (see also Tb. 1). To conclude our study, 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: local error analysis for the general case. Consider the following discretization errors:

$$\Delta_k^{(\mathrm{EE})} := \|X(k\sqrt{s}) - x_k^{(\mathrm{EE})}\|, \tag{5}$$

$$\Delta_k^{(\mathrm{SIE})} := \|X(k\sqrt{s}) - x_{k+1}^{(\mathrm{SIE})}\|. \tag{6}$$

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 $m = O(\sqrt{s})$, 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[16]. 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 (Orvieto and Lucchi, 2019; Muehlebach and Jordan, 2020; França et al., 2020a) studied the behavior of the discretization error of NAG-ODE as $k \to \infty$, showing interesting *shadowing*[17] 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. The next lemma proves that the integration error of semi-implicit and explicit Euler discretization of GM-ODE decays exponentially fast if one properly chooses the parameters.

[16]It is well known (Hairer et al., 2006) that these methods actually have the same order, since they are $O(s)$ in the velocity.

[17]That is, the discretization bound does *not* explode exponentially due to error accumulation (Chow and Van Vleck, 1994) if the objective is convex, due to the contraction provided by the landscape.

Lemma 8. *Let f be μ-strongly-convex and L-smooth. For EE discretization of GM-ODE obeying Eq. 4, 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. 3 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 rates.

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 following old notion: explicit Euler is inferior to semi-implicit (a.k.a symplectic) Euler because of its unstable nature. Instead, we show that the stability of these integrators is tied to the underlying accelerated ODE. At a deeper level, our methodology provides new challenging insights on the link between accelerated optimization, and numerical integration.

Acknowledgements. We are grateful for the enlightening discussions with Christian Lubich and Aurelien Lucchi on the connection between integration accuracy and optimization speed.

References

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

Cristian Daniel Alecsa. The long time behavior and the rate of convergence of symplectic convex algorithms obtained via splitting discretizations of inertial damping systems. *arXiv preprint arXiv:2001.10831*, 2020.

Cristian Daniel Alecsa, Szilárd Csaba László, and Titus Pinta. An extension of the second order dynamical system that models Nesterov's convex gradient method. *arXiv preprint arXiv:1908.02574*, 2019.

Foivos Alimisis, Antonio Orvieto, Gary Bécigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1297–1307. PMLR, 2020.

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

Felipe Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000.

Mahmoud Assran and Michael Rabbat. On the convergence of Nesterov's accelerated gradient method in stochastic settings. *arXiv preprint arXiv:2002.12414*, 2020.

Hedy Attouch and Felipe Alvarez. The heavy ball with friction dynamical system for convex constrained minimization problems. In *Optimization*, pages 25–35. Springer, 2000.

Hedy Attouch, Xavier Goudou, and Patrick Redont. The heavy-ball with friction method, i. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(01):1–34, 2000.

Pascal Bégout, Jérôme Bolte, and Mohamed Ali Jendoubi. On damped second-order gradient systems. *Journal of Differential Equations*, 259(7):3115–3143, 2015.

Giancarlo Benettin and Antonio Giorgilli. On the Hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms. *Journal of Statistical Physics*, 74(5-6):1117–1143, 1994.

Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

Alessandro Bravetti, Hans Cruz, and Diego Tapias. Contact Hamiltonian mechanics. *Annals of Physics*, 376:17–39, 2017.

Alessandro Bravetti, Maria L Daza-Torres, Hugo Flores-Arguedas, and Michael Betancourt. Optimization algorithms inspired by the geometry of dissipative systems. *arXiv preprint arXiv:1912.02928*, 2019.

Alexandre Cabot, Hans Engler, and Sébastien Gadat. On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the American Mathematical Society*, 361(11):5983–6017, 2009.

Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

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.

Manuel de León and Manuel Lainz Valcázar. Contact Hamiltonian systems. *Journal of Mathematical Physics*, 60(10):102902, 2019.

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

Jelena Diakonikolas and Michael I Jordan. Generalized momentum-based methods: A Hamiltonian perspective. *arXiv preprint arXiv:1906.00436*, 2019.

Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. *arXiv preprint arXiv:1706.04680*, 2017.

Mahyar Fazlyab, Alejandro Ribeiro, Manfred Morari, and Victor M Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018.

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

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

Guilherme França, Jeremias Sulam, Daniel Robinson, and René Vidal. Conformal symplectic and relativistic optimization. *Advances in Neural Information Processing Systems*, 33, 2020b.

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

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.

Ernst Hairer. Backward analysis of numerical integrators and symplectic methods. *Annals of Numerical Mathematics*, 1:107–132, 1994.

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

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.

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

Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in neural information processing systems*, 28:2845–2853, 2015.

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.

Christian Lubich. *From quantum to classical molecular dynamics: reduced models and numerical analysis*. European Mathematical Society, 2008.

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

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

Robert I McLachlan and G Reinout W Quispel. Splitting methods. *Acta Numerica*, 11:341, 2002.

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

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

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

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Antonio Orvieto and Aurelien Lucchi. Shadowing properties of optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 12692–12703, 2019.

Antonio Orvieto, Jonas Kohler, and Aurelien Lucchi. The role of memory in stochastic optimization. In *Uncertainty in Artificial Intelligence*, pages 356–366. PMLR, 2020.

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

JM Sanz-Serna and Konstantinos C Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *arXiv preprint arXiv:2009.00673*, 2020.

JM Sanz Serna and Konstantinos C Zygalakis. Contractivity of Runge–Kutta methods for convex gradient systems. *SIAM Journal on Numerical Analysis*, 58 (4):2079–2092, 2020.

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.

Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, pages 5744–5752, 2019.

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.

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

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

Pan Xu, Tianhao Wang, and Quanquan Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *International Conference on Machine Learning*, pages 5492–5501, 2018.

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.

Jingzhao Zhang, Suvrit Sra, and Ali Jadbabaie. Acceleration in first order quasi-strongly convex optimization by ode discretization. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1501–1506. IEEE, 2019.