SIAM J. NUMER. ANAL.
Vol. 0, No. 0, pp. 000–000

THE CONNECTIONS BETWEEN LYAPUNOV FUNCTIONS FOR SOME OPTIMIZATION ALGORITHMS AND DIFFERENTIAL EQUATIONS.*

J. M. SANZ SERNA† AND K. C. ZYGALAKIS‡

Abstract. In this manuscript we study the properties of a family of a second-order differential equations with damping, its discretizations, and their connections with accelerated optimization algorithms for m-strongly convex and L-smooth functions. In particular, using the linear matrix inequality (LMI) framework developed by Fazlyab et. al. (2018), we derive analytically a (discrete) Lyapunov function for a two-parameter family of Nesterov optimization methods, which allows for the complete characterization of their convergence rate. In the appropriate limit, this family of methods may be seen as a discretization of a family of second-order ODEs for which we construct (continuous) Lyapunov functions by means of the LMI framework. The continuous Lyapunov functions may alternatively be obtained by studying the limiting behavior of their discrete counterparts. Finally, we show that the majority of typical discretizations of the of the family of ODEs, such as the heavy ball method, do not possess Lyapunov functions with properties similar to those of the Lyapunov function constructed here for the Nesterov method.

Key words. Nesterov's method, Lyapunov function, linear matrix inequalities, convex optimization

AMS subject classifications. 65L06, 65L20, 90C25, 93C15

DOI. 10.1137/20M1364138

1. Introduction.

This paper studies Lyapunov functions for differential equations with damping, their discretizations, and optimization algorithms.

The simplest algorithm for solving

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

is the gradient descent (GD) method

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

which is of course the result of applying Euler's rule, with step-size $\alpha_k > 0$, to the gradient system

$$\frac{dx}{dt} = -\nabla f(x), \qquad x(0) = x_0.$$

The value of f decreases along solutions $x(t)$ of this system, and, correspondingly, it may be hoped that, for GD, $f(x_{k+1}) \le f(x_k)$ for sufficiently small α_k. In fact, that is the case for $\alpha_k < 2/L$ if f is L-smooth; i.e., if $\nabla f(x)$ is L-Lipschitz continuous.

*Received by the editors September 2, 2020; accepted for publication (in revised form) March 15, 2021; published electronically DATE.
https://doi.org/10.1137/20M1364138
Funding: The work of the first author was supported by the projects MTM2016-77660-P(AEI/FEDER, UE) MINECO and PID2019-104927GB- C21 (AEI/FEDER, UE) (Spain). The work of the second author was supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1.
†Departamento de Matemáticas, Universidad Carlos III de Madrid, Leganés (Madrid), E-28911, Spain (jmsanzserna@gmail.com).
‡School of Mathematics, University of Edinburgh, Edinburgh, EH9 3FD, UK (k.zygalakis@ed.ac.uk).

In this paper we are mainly interested in problems where f belongs the set $\mathcal{F}_{m,L}$ of m-strongly convex and L-smooth functions, a class that plays an important role in optimization [19]. For f in this class and the constant step-size $\alpha = 2/(m+L)$, GD has a bound [19, Theorem 2.1.15]

$$f(x_k) - f(x^\star) \leq \frac{L}{2}\left(\frac{1-1/\kappa}{1+1/\kappa}\right)^{2k} \|x_0 - x^\star\|^2, \tag{1.1}$$

where $x^\star$ is the (unique) minimizer of f and $\kappa = L/m \geq 1$ is the condition number of f.

The $1-\mathcal{O}(1/\kappa)$ rate of decay in f in the preceding bound is unsatisfactory because in many applications of interest one has $\kappa \gg 1$. It is possible to improve on GD by resorting to *accelerated* algorithms with rates $1-\mathcal{O}(1/\sqrt{\kappa})$; for instance, for the method

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k), \tag{1.2a}$$

$$y_k = x_k + \frac{1-\sqrt{1/\kappa}}{1+\sqrt{1/\kappa}}(x_k - x_{k-1}), \tag{1.2b}$$

introduced by Nesterov, it may be shown [19, Theorem 2.2.3] that, if $y_0 = x_0$,

$$f(x_k) - f(x^\star) \leq \left(1-\sqrt{1/\kappa}\right)^k \left(f(x_0) - f(x^\star) + \frac{m}{2}\|x_0 - x^\star\|^2\right). \tag{1.3}$$

The factor $1-\sqrt{1/\kappa}$ here is close to the optimal possible factor $(1-\sqrt{1/\kappa})^2/(1+\sqrt{1/\kappa})^2$ one can achieve for minimization algorithms when $f \in \mathcal{F}_{m,L}$ [19, Theorem 2.1.13]. The algorithm (1.2) is also related to ODEs, because it may be seen as a discretization of of the Polyak damped oscillator equation [21]

$$\ddot{x} + 2\sqrt{m}\dot{x} + \nabla f(x) = 0, \tag{1.4}$$

whose solutions $x(t)$ approach $x^\star$ as $t \to \infty$ if f is m-strongly convex [32, Proposition 3].

In recent years, there has been a revived interest, beginning with [30], in the connections between differential equations and optimization algorithms (see also [27]). In particular, there has been several papers (see, e.g., [31, 13]) that proposed accelerated algorithms, both in Euclidean and non-Euclidean geometry, based on discretizations of second-order dissipative ODEs. The structure of these ODEs and the fact that they can been viewed as describing Hamiltonian systems with dissipation led to a number of research works that tried to construct or explain optimization algorithms using concepts such as shadowing [20], symplecticity [2, 4, 17, 18, 29], discrete gradients [7], and backward error analysis [9].

A common feature of the analysis presented in many of the papers mentioned above was the construction of a discrete Lyapunov function that was used in order to deduce the convergence rate of the underlying algorithm. In [32] a general analysis of optimization methods based on the derivation of Lyapunov functions that mimic ODE Lyapunov functions was carried out; that paper presents a Lyapunov function for (1.4). A Lyapunov function for (1.2) may be seen in [14], where it was also used to study stochastic versions of the algorithm. The paper [28], among other contributions, constructs a Lyapunov function for a one-parameter family of optimization algorithms that includes (1.2) as a particular case. Outside the field of optimization, Lyapunov functions are important in establishing ergodicity of random dynamical systems [24], as well as ergodicity of Markov Chain Monte Carlo algorithms; see, for example, [16, 3]. The construction of Lyapunov functions for optimization algorithms from the perspective of control theory was the subject of study in [8]. The authors extend the work in [15] and derive linear matrix inequalities (LMIs) that guarantee the existence of suitable Lyapunov functions that may be used to establish the convergence rate of the algorithm under study. In addition, [8] develops an LMI framework to construct Lyapunov functions for systems of ODEs. Typically, the LMIs that appear in this context have been solved numerically in the literature.

In this work,

1. For $f \in \mathcal{F}_{m,L}$, we use the LMI framework from [8] to derive *analytically* Lyapunov functions for a two-parameter family of Nesterov optimization methods (see (3.1) below); this family includes the one-parameter family of algorithms in [28]. In this way we find, as a function of the two parameters in (3.1), a convergence rate for the methods in the family. It turns out that the best convergence rate is achieved when the parameters are chosen as in (1.2). The relation between the Lyapunov function constructed in the present work and its counterpart in [28] is discussed in Remark 3.5.
2. By taking an appropriate limit of the parameters as in, e.g., [26, 2, 28, 4, 17, 18, 29, 9] the optimization algorithms in the family may be seen as discretizations of second-order ODEs of the form

$$\ddot{x} + \bar{b}\sqrt{m}\dot{x} + \nabla f(x) = 0, \tag{1.5}$$

   where $\bar{b} > 0$ is a friction parameter. We obtain analytically Lyapunov functions for (1.5) and determine, as a function of $\bar{b}$, a convergence rate of f to $f(x^\star)$ along solutions $x(t)$. We prove that the value $\bar{b} = 2$ in the Polyak ODE (1.4) yields the *optimal convergence rate if f is m-strongly convex*. Additionally we show that if one is to take explicitly into account the value of L into this calculation, the optimal value of $\bar{b}$ becomes strictly larger than 2 and yields slightly better convergence rates.
3. We show that, in the limit where the optimization algorithms approximate the ODEs, the discrete Lyapunov functions converge to the ODE Lyapunov function. Using this correspondence we show, by means of the heavy ball method [21] and other examples that typically optimization algorithms that are discretizations of (1.5) do not possess discrete Lyapunov functions that mimic the Lyapunov function of the differential equation in item 2 above and lead to acceleration. This emphasizes the well-known fact that, when designing optimization methods, it is not sufficient to ensure that the algorithm may be seen as a consistent discretization of a well-behaved ODE. Unfortunately, discretizations do not necessarily inherit the good long-time properties of the differential equation, as seen, for example, in the case of discretization of gradient flows [25], and Hamiltonian problems [23].

The rest of the paper is organized as follows. In section 2 we briefly review the approach in [8] that provides a basis for our constructions. In section 3 we find analytically Lyapunov functions/rates of convergence for a two-parameter family of optimization methods that contains (1.2) as a particular case. Section 4 analyzes the ODE (1.5), and section 5 studies the connection between the discrete and continuous Lyapunov functions. The heavy ball method and other methods that do not possess suitable Lyapunov functions are discussed in section 6. Finally, we present in the appendix the calculations that allows us to deduce that while the choice $\bar{b} = 2$ in (1.5) is optimal if f is only assumed to be m-strongly convex, slightly better rates of convergence may be achieved for $f \in \mathcal{F}_{m,L}$ by taking $\bar{b} > 2$.

2. Preliminaries.

We will now briefly describe the framework introduced in [8] for the construction of Lyapunov functions of optimization methods and differential equations. The presentation here is adapted from the material in [8] to suit our specific needs.

Remark 2.1. The following material is limited to results needed to study strongly convex optimization. However the LMI approach in [8] also works in convex optimization.

2.1. Optimization methods.

Optimization algorithms can often be represented as linear dynamical systems interacting with one or more static nonlinearities (see [15]). In this paper we will consider first-order algorithms that have the following state-space representation:

$$\xi_{k+1} = A\xi_k + Bu_k, \tag{2.1a}$$
$$u_k = \nabla f(y_k), \tag{2.1b}$$
$$y_k = C\xi_k, \tag{2.1c}$$
$$x_k = E\xi_k, \tag{2.1d}$$

where $\xi_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^d$ is the input ($d \leq n$), $y_k \in \mathbb{R}^d$ is the feedback output that is mapped to u_k by the nonlinear map ∇f. From the perspective of the optimization, x_k is the approximation to the mimimizer $x^\star$.

As example, consider algorithms of the well-known form ([15, 8])

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(y_k), \tag{2.2a}$$
$$y_k = x_k + \gamma(x_k - x_{k-1}), \tag{2.2b}$$

where $\alpha > 0, \beta, \gamma$ are scalar parameters that specify the algorithm within the family. For $\beta = \gamma = 0$ we recover GD. For $\beta = \gamma$, we have Nesterov's method; (1.2) corresponds to a particular choice of α and β. The heavy ball method has $\gamma = 0$, $\beta \neq 0$. By defining the state vector $\xi_k = [x_{k-1}^\mathsf{T}, x_k^\mathsf{T}]^T \in \mathbb{R}^{2d}$ we can represent (2.2) in the form (2.1) with the matrices A, B, C, E given by

$$A = \begin{bmatrix} 0 & I_d \\ -\beta I_d & (\beta+1)I_d \end{bmatrix}, \; B = \begin{bmatrix} 0 \\ -\alpha I_d \end{bmatrix}, \; C = \begin{bmatrix} -\gamma I_d & (\gamma+1)I_d \end{bmatrix}, \; E = \begin{bmatrix} 0 & I_d \end{bmatrix}.$$

Fixed points of (2.1) satisfy

$$\xi^\star = A\xi^\star + Bu^\star, \quad y^\star = C\xi^\star, \quad u^\star = \nabla f(y^\star), \quad x^\star = E\xi^\star;$$

in the optimization context $u^\star = 0$, and $y^\star = x^\star$ is the minimizer sought.

To study the convergence rate of optimization algorithms, [8] considers functions of the form

$$V_k(\xi) = \rho^{-2k}\left(a_0(f(x) - f(x^\star)) + (\xi - \xi^\star)^\mathsf{T} P(\xi - \xi^\star)\right), \tag{2.3}$$

where $a_0 > 0$ and P is positive semidefinite (denoted by $P \succeq 0$). If along the trajectories of (2.1)

$$V_{k+1}(\xi_{k+1}) \leq V_k(\xi_k), \tag{2.4}$$

we can conclude that $\rho^{-2k}a_0(f(x_k) - f(x^\star)) \leq V_k(\xi_k) \leq V_0(\xi_0)$ or

$$f(x_k) - f(x^\star) \leq \rho^{2k}\frac{V_0(\xi_0)}{a_0}.$$

If $\rho < 1$, we have found a convergence rate for $f(x_k)$ towards the optimal value $f(x^\star)$. The following theorem defines an LMI that, when $f \in \mathcal{F}_{m,L}$, guarantees that the property (2.4) holds, and therefore (2.3) provides a Lyapunov function for the system.

THEOREM 2.2 (see [8, Theorem 3.2]). *Suppose that, for* (2.1)*, there exist* $a_0 > 0, P \succeq 0$*,* $\ell > 0$*, and* $\rho \in [0, 1)$ *such that*

$$T = M^{(0)} + a_0\rho^2 M^{(1)} + a_0(1-\rho^2)M^{(2)} + \ell M^{(3)} \preceq 0, \tag{2.5}$$

where

$$M^{(0)} = \begin{bmatrix} A^\mathsf{T}PA - \rho^2 P & A^\mathsf{T}PB \\ B^\mathsf{T}PA & B^\mathsf{T}PB \end{bmatrix},$$

and

$$M^{(1)} = N^{(1)} + N^{(2)}, \quad M^{(2)} = N^{(1)} + N^{(3)}, \quad M^{(3)} = N^{(4)},$$

with

$$N^{(1)} = \begin{bmatrix} EA - C & EB \\ 0 & I_d \end{bmatrix}^\mathsf{T} \begin{bmatrix} \frac{L}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix} \begin{bmatrix} EA - C & EB \\ 0 & I_d \end{bmatrix},$$

$$N^{(2)} = \begin{bmatrix} C - E & 0 \\ 0 & I_d \end{bmatrix}^\mathsf{T} \begin{bmatrix} -\frac{m}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix} \begin{bmatrix} C - E & 0 \\ 0 & I_d \end{bmatrix},$$

$$N^{(3)} = \begin{bmatrix} C^\mathsf{T} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{m}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix},$$

$$N^{(4)} = \begin{bmatrix} C^\mathsf{T} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{mL}{m+L}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & -\frac{1}{m+L}I_d \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix}.$$

Then, for $f \in \mathcal{F}_{m,L}$*, the sequence* $\{x_k\}$ *satisfies*

$$f(x_k) - f(x^\star) \leq \frac{a_0(f(x_0) - f(x^\star)) + (\xi_0 - \xi^\star)^\mathsf{T}P(\xi_0 - \xi^\star)}{a_0}\rho^{2k}.$$

2.2. Continuous-time systems. We also consider continuous-time dynamical systems in state space form (throughout the paper we often use a bar over symbols related to ODEs)

$$\dot{\xi}(t) = \bar{A}\xi(t) + \bar{B}u(t), \quad y(t) = \bar{C}\xi(t), \quad u(t) = \nabla f(y(t)) \quad \text{for all } t \geq 0, \tag{2.6}$$

where $\xi(t) \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^d (d \leq n)$ the output, and $u(t) = \nabla f(y(t))$ the continuous feedback input. Fixed points of (2.6) satisfy

$$0 = \bar{A}\xi^\star, \quad y^\star = \bar{C}\xi^\star, \quad u^\star = \nabla f(y^\star);$$

in our context $u^\star = 0$ and $y^\star = x^\star$. We can replicate the convergence analysis of the discrete case using now functions of the form

$$\bar{V}(\xi(t)) = e^{\lambda t}\left(f(y(t)) - f(y^\star) + (\xi(t) - \xi^\star)^\mathsf{T}\bar{P}(\xi(t) - \xi^\star)\right), \tag{2.7}$$

where $\lambda > 0$. If $\bar{P} \succeq 0$ and, along solutions, $(d/dt)\bar{V}(\xi(t)) \leq 0$, then we have $\bar{V}(\xi(t)) \leq \bar{V}(\xi(0))$ which in turns implies

$$f(y(t)) - f(y^\star) \leq e^{-\lambda t}\bar{V}(\xi(0)).$$

The following theorem similarly to the discrete time case, formulates an LMI that guarantees the existence of such a Lyapunov function.

THEOREM 2.3. *Suppose that, for* (2.6), *there exist* $\lambda > 0$, $\bar{P} \succeq 0$, *and* $\sigma \geq 0$ *that satisfy*

$$\bar{T} = \bar{M}^{(0)} + \bar{M}^{(1)} + \lambda \bar{M}^{(2)} + \sigma \bar{M}^{(3)} \preceq 0, \tag{2.8}$$

where

$$\bar{M}^{(0)} = \begin{bmatrix} \bar{P}\bar{A} + \bar{A}^\mathsf{T}\bar{P} + \lambda\bar{P} & \bar{P}\bar{B} \\ \bar{B}^\mathsf{T}\bar{P} & 0 \end{bmatrix},$$

$$\bar{M}^{(1)} = \frac{1}{2}\begin{bmatrix} 0 & (\bar{C}\bar{A})^\mathsf{T} \\ \bar{C}\bar{A} & \bar{C}\bar{B} + \bar{B}^\mathsf{T}\bar{C}^\mathsf{T} \end{bmatrix},$$

$$\bar{M}^{(2)} = \begin{bmatrix} \bar{C}^\mathsf{T} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{m}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix} \begin{bmatrix} \bar{C} & 0 \\ 0 & I_d \end{bmatrix},$$

$$\bar{M}^{(3)} = \begin{bmatrix} \bar{C}^\mathsf{T} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{mL}{m+L}I_d & \frac{1}{2}I_d, \\ \frac{1}{2}I_d & -\frac{1}{m+L}I_d \end{bmatrix} \begin{bmatrix} \bar{C} & 0 \\ 0 & I_d \end{bmatrix}.$$

Then the following inequality holds for $f \in \mathcal{F}_{m,L}$, $t \geq 0$,

$$f(y(t)) - f(y^\star) \leq e^{-\lambda t}\left(f(y(0)) - f(y^\star) + (\xi(0) - \xi^\star)^\mathsf{T}\bar{P}(\xi(0) - \xi^\star)\right).$$

3. A Lyapunov function for Nesterov's optimization algorithm.

We study the optimization method (cf. (2.2))

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(y_k), \tag{3.1a}$$

$$y_k = x_k + \beta(x_k - x_{k-1}), \tag{3.1b}$$

$k = 0, 1, \ldots$, with parameters $\alpha > 0$ and β. As noted before, the choice $\beta = 0$ gives GD, and $\beta \neq 0$ corresponds to Nesterov's accelerated algorithm.

3.1. The construction.

After introducing

$$\delta = \sqrt{m\alpha},$$

and the divided difference, $k = 0, 1, \ldots$,

$$d_k = \frac{1}{\delta}(x_k - x_{k-1}), \tag{3.2}$$

the recursion (3.1) may be rewritten ($k = 0, 1, \ldots$)

$$d_{k+1} = \beta d_k - \frac{\alpha}{\delta}\nabla f(y_k), \tag{3.3a}$$

$$x_{k+1} = x_k + \delta\beta d_k - \alpha\nabla f(y_k), \tag{3.3b}$$

$$y_k = x_k + \delta\beta d_k. \tag{3.3c}$$

Remark 3.1. For future reference, it is useful to observe that, from a dimensional analysis point of view, m, L, and $1/\alpha$ have the dimensions of the quotient $f/\|x\|^2$. Therefore δ is a *nondimensional* version of $\sqrt{\alpha}$. The parameter β is nondimensional. The divided difference (3.2) shares the dimensions of x.

Equation (3.3) can now be written in the form (2.1) with $\xi_k = [d_k^\mathsf{T}, x_k^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{2d}$ and

$$A = \begin{bmatrix} \beta I_d & 0 \\ \delta\beta I_d & I_d \end{bmatrix}, \quad B = \begin{bmatrix} -(\alpha/\delta) I_d \\ -\alpha I_d \end{bmatrix}, \quad C = \begin{bmatrix} \delta\beta I_d & I_d \end{bmatrix}, \quad E = \begin{bmatrix} 0 & I_d \end{bmatrix}. \tag{3.4}$$

In the preceding section, as in [8], the state ξ_k was taken to be $[x_{k-1}^\mathsf{T}, x_k^\mathsf{T}]^\mathsf{T}$ rather than $[d_k^\mathsf{T}, x_k^\mathsf{T}]^\mathsf{T}$. While both choices are of course mathematically equivalent, the new ξ_k is more convenient for our purposes. In addition, when looking numerically for Lyapunov functions by solving LMIs, it leads to problems that are better conditioned for large condition numbers κ.

Remark 3.2. For $\beta = 0$ (gradient descent), the first equation in (3.3) is a reformulation of the second: It would be more natural to use the simpler state $\xi_k = x_k$.

According to Theorem 2.2, in order to find a Lyapunov function of the form (2.3), it is sufficient to find a matrix $P \succeq 0$ and numbers $a_0 > 0$, $0 < \rho < 1$, $\ell \geq 0$, such that the matrix T in (2.5) is negative semidefinite. At the outset, we choose $\ell = 0$ in order to simplify the subsequent analysis. As we will discuss in the Appendix, this simplification does not have a significant impact on the value of the convergence rate ρ that results from the analysis. With $\ell = 0$, (2.5) is homogeneous in P and a_0, and we may divide accross by a_0. In other words, without loss of generality, we may take $a_0 = 1$. Then T is a function of P and ρ (and the method parameters β and δ).

The matrix A in (3.4) is a Kronecker product of a 2×2 matrix and I_d,

$$A = \begin{bmatrix} \beta & 0 \\ \delta\beta & 1 \end{bmatrix} \otimes I_d;$$

the factor I_d originates from the dimensionality of the decision variable x and the 2×2 factor is independent of d and arises from the optimization algorithm. The matrices B, C, and E have a similar Kronecker product structure. It is then natural to consider symmetric matrices P of the form

$$P = \widehat{P} \otimes I_d, \qquad \widehat{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}, \tag{3.5}$$

and then T will also have a Kronecker product structure

$$T = \widehat{T} \otimes I_d, \qquad \widehat{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{12} & t_{22} & t_{23} \\ t_{13} & t_{23} & t_{33} \end{bmatrix}, \tag{3.6}$$

where the t_{ij} are explicitly given by the following complicated expressions obtained from (3.4) and the recipes for $M^{(0)}$, $M^{(1)}$, and $M^{(2)}$ in Theorem 2.2:

$$t_{11} = \beta^2 p_{11} + 2\delta\beta^2 p_{12} + \delta^2\beta^2 p_{22} - \rho^2 p_{11} - \delta^2\beta^2 m/2, \tag{3.7a}$$

$$t_{12} = \beta p_{12} + \delta\beta p_{22} - \rho^2 p_{12} - \delta\beta m/2 + \rho^2 \delta\beta m/2, \tag{3.7b}$$

$$t_{13} = -\delta^{-1}\alpha\beta p_{11} - 2\alpha\beta p_{12} - \delta\alpha\beta p_{22} + \delta\beta/2, \tag{3.7c}$$

$$t_{22} = p_{22} - \rho^2 p_{22} - m/2 + \rho^2 m/2, \tag{3.7d}$$

$$t_{23} = -\delta^{-1}\alpha p_{12} - \alpha p_{22} + 1/2 - \rho^2/2, \tag{3.7e}$$

$$t_{33} = \delta^{-2}\alpha^2 p_{11} + 2\delta^{-1}\alpha^2 p_{12} + \alpha^2 p_{22} + \alpha^2 L/2 - \alpha. \tag{3.7f}$$

Our task is to find $\rho \in [0,1)$, p_{11}, p_{12}, and p_{22} that lead to $\widehat{T} \preceq 0$ and $\widehat{P} \succeq 0$ (which imply $T \preceq 0$ and $P \succeq 0$). The algebra becomes simpler if we represent β and ρ^2 as

$$\beta = 1 - b\delta, \qquad \rho^2 = 1 - r\delta. \tag{3.8}$$

Note that we are interested in $r \in (0, 1/\delta]$ so as to get $\rho^2 \in [0,1)$. We proceed in steps as follows.

First step. Impose the condition $t_{23} = 0$. This leads to

$$p_{12} = \frac{m}{2} r - \delta p_{22}. \tag{3.9}$$

Second step. Impose the condition $t_{13} = 0$. This results in

$$p_{11} = \frac{m}{2} - 2\delta p_{12} - \delta^2 p_{22},$$

which in tandem with (3.9) yields

$$p_{11} = \frac{m}{2} - mr\delta + \delta^2 p_{22}. \tag{3.10}$$

Third step. Impose the condition $\det(\widehat{P}) = p_{11}p_{22} - p_{12}^2 = 0$. Using (3.9) and (3.10), we have a linear equation for p_{22} with solution

$$p_{22} = \frac{m}{2} r^2.$$

We now take this value to (3.9) and (3.10) and get

$$\widehat{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} = \frac{m}{2} \begin{bmatrix} (1-r\delta)^2 & r(1-r\delta) \\ r(1-r\delta) & r^2 \end{bmatrix}, \tag{3.11}$$

a matrix that is positive semidefinite (but not positive definite).

Fourth step. Impose $t_{33} \leq 0$. After using (3.11) in the expression for t_{33} in (3.7), this condition is seen to be equivalent to $\alpha^2 L - \alpha \leq 0$ or

$$\alpha \leq \frac{1}{L}$$

(for $\alpha = 1/L$, t_{33} actually vanishes). In what follows we assume that this bound on α holds; note that then $\delta = \sqrt{m\alpha} \leq \sqrt{m/L} < 1$.

Fifth step. We impose $t_{22} \leq 0$. This may be written as $(p_{22} - m/2)r\delta \leq 0$, which leads to $p_{22} \leq m/2$. From (3.11)

$$r \leq 1,$$

which sets a lower limit $\rho^2 \leq 1 - \delta$ for the rate of convergence. For $r^2 < 1$, $t_{22} < 0$.

Sixth step. Impose $t_{11}t_{22} - t_{12}^2 = 0$. From (3.11) and (3.7), some algebra yields

$$t_{11}t_{22} - t_{12}^2 = -\frac{m^3}{4} r(1 - r\delta)\,\Xi$$

with

$$\Xi = \Xi_\delta(r,b) = (r+\delta)(1-\delta^2)b^2 - 2(1+r^2)(1-\delta^2)b + (r^3 - 3r^2\delta + 3r - \delta). \tag{3.12}$$

Since $\delta < 1$ and, after step five, $r \in (0,1]$, we must have $\Xi = 0$. For fixed $\delta \in (0,1)$, the condition $\Xi_\delta = 0$ establishes a relation between the values of r and b or, in other words, the rate of convergence ρ^2 and the parameter β in (3.1). In order to study this relation, we now make a digression and describe, for fixed $\delta \in (0,1)$, the algebraic curve of equation $\Xi_\delta(r,b) = 0$ in the real plane (r,b); in this description we allow arbitrary real values of r and b (even though in our problem $r \in (0,1]$).

The formula for the roots of a quadratic equation yields

$$b_\pm = \frac{(1+r^2)(1-\delta^2) \pm (1-r\delta)\sqrt{(1-r^2)(1-\delta^2)}}{(r+\delta)(1-\delta^2)}. \tag{3.13}$$

For $r^2 \neq 1$ and $r \neq -\delta$ there are two distinct real roots b_+ and b_-. For $r = \pm 1$ there is a double root $b = 2/(r+\delta)$. As $r \downarrow -\delta$, we have $b_+ \uparrow +\infty$ and $b_- \downarrow -2\delta/(1-\delta^2)$. By using (3.13) it is not difficult to prove that $\Xi_\delta(r,b) = 0$ defines r as a single-valued function of the variable $b \in \mathbb{R}$. (We could find an explicit expression for r in terms of b by means of the formula for the roots of a cubic equation, but this is not necessary for our purposes.) Figure 3.1 provides a plot of the curve $\Xi_\delta(r,b) = 0$ when $\delta = 1/2$.

We now return to the construction of T. Recall that for our purposes, we need $r > 0$ (so as to have $\rho < 1$); this requirement holds for $b \in (b_{\min}, b_{\max})$, where

$$b_{\min} = \frac{1-\delta^2 - \sqrt{1-\delta^2}}{\delta(1-\delta^2)} < 0, \qquad b_{\max} = \frac{1-\delta^2 + \sqrt{1-\delta^2}}{\delta(1-\delta^2)} > 0$$

are the intersections of the curve $\Xi_\delta = 0$ with the vertical axis. As $\delta \downarrow 0$,

$$b_{\min} \uparrow 0, \qquad b_{\max} \uparrow +\infty. \tag{3.14}$$

The limits on b just found are equivalent to

$$-\sqrt{1-\delta^2} < \beta < +\sqrt{1-\delta^2}. \tag{3.15}$$

For the maximum value $r = 1$ found in step five above, (3.13) gives the double root $b = 2/(1+\delta)$ or $\beta = (1-\delta)/(1+\delta)$. Values $r \in (0,1)$ correspond to two different choices of $b \in (b_{\min}, b_{\max})$.

We are now ready to present the following result.

THEOREM 3.3. *Consider the minimization algorithm* (3.1) *(or* (3.3)*) with parameters subject to*

$$\alpha \leq 1/L, \qquad -\sqrt{1-m\alpha} \leq \beta \leq \sqrt{1-m\alpha}.$$

Set $\delta = \sqrt{m\alpha}$, and let $r > 0$ be the value determined by $\Xi_\delta(r,b) = 0$ (see (3.12)*); set $\rho^2 = 1 - r\delta < 1$, and define the positive semidefinite matrix P by* (3.5) *and* (3.11)*. Then the matrix T in* (3.6)*–*(3.7) *is negative semidefinite.*

As a result, for any x_{-1}, x_0, the sequence

$$\rho^{-2k}\Big(f(x_k) - f(x_\star) + [d_k^T, x_k^T - x_\star^T]\, P\, [d_k^T, x_k^T - x_\star^T]^T\Big) \tag{3.16}$$

decreases monotonically, which, in particular, implies

$$f(x_k) - f(x_\star) \leq C\rho^{2k}$$

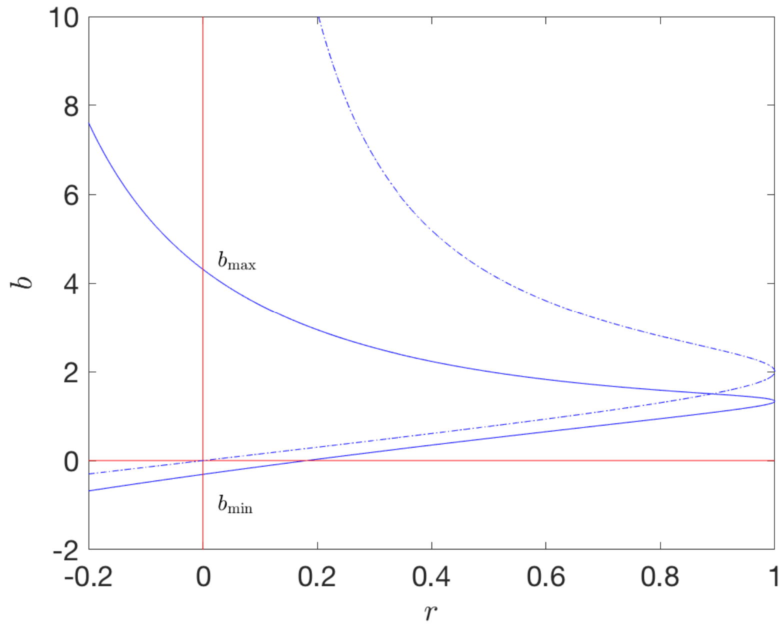

FIG. 3.1. *The solid curve corresponds to the equation $\Xi_\delta(r,b)=0$ when $\delta=1/2$. It has a vertical asymptote at $r=-\delta$ (not shown). To each real b there corresponds a single value of r. For $b\in(b_{\min},b_{\max})$, we have $0<r\leq 1$ that corresponds to $1>\rho^2\geq 1-\delta$. The best rate $\rho^2=1-\delta$ is achieved for $b=2\delta/(1+\delta)$; i.e., $\beta=(1-\delta)/(1+\delta)$. The discontinuous curve corresponds to the equation $\Xi_\delta(r,b)=0$ in the limit $\delta\to 0$; again to each real b there corresponds a single value of r. This curve is symmetric with respect to the origin (changing b into $-b$ changes r into $-r$) and has a vertical asymptote at $r=0$. Positive values of b correspond to positive values of r. The maximum value $r=1$ is achieved when $b=2$.*

with

$$C=f(x_0)-f(x^\star)+\frac{m}{2}\left\|\frac{1-r\delta}{\delta}(x_0-x_{-1})+r(x_0-x^\star)\right\|^2.$$

Proof. Using Theorem 2.2, we only have to prove that $\widehat{T}\preceq 0$. The second, first, and fourth steps of our construction, respectively, ensure that $t_{13}=t_{23}=0$ and $t_{33}\leq 0$, and therefore we are left with the task of checking that the 2×2 matrix $\widehat{T}^{12}$ obtained by suppressing the last row and last column of $\widehat{T}$ is $\preceq 0$. If $r<1$, we know from step five that $t_{22}<0$ and from step six that the determinant of $\widehat{T}^{12}$ vanishes, and therefore $\widehat{T}^{12}\preceq 0$. For $r=1$, $t_{22}=0$, but again $\widehat{T}^{12}\preceq 0$, because in this case $t_{11}=-(m/2)\delta(1-\delta)^3/(1+\delta)<0$. ∎

For fixed $\alpha\leq 1/L$, as noted above, ρ^2 is minimized by the choice

$$\beta=(1-\sqrt{m\alpha})/(1+\sqrt{m\alpha});$$

then

$$\rho^2=1-\sqrt{m\alpha}.$$

When α is allowed to vary in the interval $(0,1/L]$, increasing α results in an improvement of ρ^2, so that the best rate $\rho^2=1-\sqrt{m/L}=1-\sqrt{1/\kappa}$ is obtained by setting $\alpha=1/L$, and then (3.1) coincides with (1.2). The parameter values $\alpha=1/L$, $\beta=(1-\sqrt{1/\kappa})/(1+\sqrt{1/\kappa})$ in (1.2) are of course the "standard" choice for Nesterov's algorithm (see, e.g., [15, Proposition 12]). For this choice of parameters and

$x_{-1} = x_0$, the bound in Theorem 3.3 exactly coincides (including the value of C) with that in (1.3), which is derived in [19, Theorem 2.2.3] without using Lyapunov functions. Numerical experiments in [15] show that for $\kappa^{-1} = m/L$ small the rate of convergence $\rho^2 = 1 - \sqrt{1/\kappa}$ is essentially the best that the algorithm achieves.

The theorem may also be applied to the GD algorithm with $\beta = 0$ and $b = 1/\delta$, even though (see Remark 3.2) in this case the preceding treatment is unnatural. One finds $r = \delta$, so that the decay per step in $f(x_k) - f(x_\star)$ provided by Theorem 3.3 is $\rho^2 = 1 - \delta^2 = 1 - m\alpha$ for $\alpha \leq 1/L$. When $\alpha = 2/(m+L)$, the decay per step guaranteed by Theorem 3.3 is $\rho^2 = 1 - 1/\kappa/1 + 1/\kappa$; this is worse than the bound in (1.1) valid for the same value of α.

Remark 3.4. The decay rate ρ^2 provided by the theorem is a nondimensional quantity that only depends on the nondimensional variables b and δ. The bound $\alpha \leq 1/L$ may be rewritten in the nondimensional form as $\delta^2 \leq m/L = 1/\kappa$. These facts guarantee that the theorem is equivariant with respect to changes in scale of f and x. The Lyapunov function in (3.16) has the dimensions of f because, according to (3.11), P has the dimensions of m, i.e., those of $f/\|x\|^2$.

Remark 3.5. For the particular choice of α and β leading to (1.2), the Lyapunov function in the theorem above was derived in [14] by means of an alternative technique (see Remark 5.2). In [28] a Lyapunov function that contains the gradient $\nabla f(x)$ is constructed analytically for the situation where the learning rate α in (3.1) is a free parameter and the momentum parameter is fixed as $\beta = (1 - \sqrt{m\alpha})/(1 + \sqrt{m\alpha})$ (i.e., at the value that according to the analysis above optimizes ρ^2). The analysis in [28] requires (see Lemma 3.4 in that reference) $\alpha \leq 1/(4L)$, while here $\alpha \leq 1/L$. In addition for $\alpha = 1/(4L)$, [28, Theorem 3] proves a rate $1/(1 + (1/12)\sqrt{m/L})$ which, while establishing acceleration, compares unfavourably with the value $1-(1/2)\sqrt{m/L}$ provided by Theorem 3.3.

3.2. Optimality. The path leading to Theorem 3.3 has a degree of arbitrariness, and it may be asked whether, by following an alternative construction, it is possible to determine the parameters ρ, p_{11}, p_{12}, p_{22}, and in such a way that $\widehat{T} \preceq 0$, $\widehat{P} \succeq 0$ and the value of ρ is larger than the value provided in Theorem 3.3. We conclude this section by presenting a result in this direction. We fix the parameters in the algorithm at the standard choices, i.e., $\alpha = 1/L$, $\beta = (1-\delta)/(1+\delta)$, $\delta = \sqrt{m/L}$, and denote by $\rho^\star = \sqrt{1-\delta}$, $p_{11}^\star = (m/2)(1-\delta)^2$, $p_{12}^\star = (m/2)(1-\delta)$, $p_{22}^\star = m/2$ the values yielded by Theorem 3.3. In the space of the decision variables ρ, p_{11}, p_{22}, p_{33} we pose the convex optimization problem of minimizing ρ subject to the constraints $\widehat{T} \preceq 0$, $\widehat{P} \succeq 0$. We then have the following result that shows that the rate provided in Theorem 3.3 cannot be improved with an alternative choice of $\widehat{P}$.

THEOREM 3.6. *With the notation just described, the unique solution of the minimization problem is* $(\rho^\star, p_{11}^\star, p_{12}^\star, p_{22}^\star)$.

Proof. We use the notation $\sigma = \rho^2$, $\sigma^\star = (\rho^\star)^2$ and write $\sigma = \sigma^\star + \widetilde{\sigma}$, $p_{11} = p_{11}^\star + \widetilde{p}_{11}$, $p_{12} = p_{12}^\star + \widetilde{p}_{12}$, $p_{22} = p_{22}^\star + \widetilde{p}_{22}$. Since the minimization problem is convex, it is sufficient to show that $\rho^\star$, $p_{11}^\star$, $p_{12}^\star$, $p_{22}^\star$ provide a local minimum; i.e., that if the increments $\widetilde{\sigma} \leq 0$, $\widetilde{p}_{11}$, $\widetilde{p}_{12}$, $\widetilde{p}_{22}$ are of sufficiently small magnitude and $(\sigma, p_{11}, p_{12}, p_{22})$ is feasible, then $\sigma = \sigma^\star$, $p_{11} = p_{11}^\star$, $p_{12} = p_{12}^\star$, $p_{22} = p_{22}^\star$.

We study three requirements that feasibility imposes on $\widetilde{\sigma}$, $\widetilde{p}_{11}$, $\widetilde{p}_{12}$, $\widetilde{p}_{22}$.

(1) First, the constraint $\widehat{P} \succeq 0$ implies that $p_{11}p_{22} - p_{12}^2 \geq 0$ or

$$p_{22}^\star \widetilde{p}_{11} - 2p_{12}^\star \widetilde{p}_{12} + p_{11}^\star \widetilde{p}_{22} + \widetilde{p}_{11}\widetilde{p}_{22} - (\widetilde{p}_{12})^2 \geq 0.$$

Because we are carrying a local study, we replace the constraint by its linearization

$$p^\star_{22}\widetilde{p}_{11} - 2p^\star_{12}\widetilde{p}_{12} + p^\star_{11}\widetilde{p}_{22} \geq 0,$$

or, after using the known values of the symbols with a star,

$$\widetilde{p}_{11} - 2(1-\delta)\widetilde{p}_{12} + (1-\delta)^2\widetilde{p}_{22} \geq 0. \tag{3.17}$$

(2) Then, the constraint $\widehat{T} \preceq 0$ implies $t_{22}t_{33} - t_{23}^2 \geq 0$ or, using (3.7),

$$-\Big(\frac{1}{2}\widetilde{\sigma} + \frac{\delta}{m}\widetilde{p}_{12} + \frac{\delta^2}{m}\widetilde{p}_{22}\Big)^2 + \frac{\delta^3}{m^2}\widetilde{p}_{22}\big(\widetilde{p}_{11} + 2\delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\big)$$
$$- \frac{\delta^2}{m^2}\widetilde{\sigma}\widetilde{p}_{22}\big(\widetilde{p}_{11} + 2\delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\big) \geq 0.$$

This time the leading terms in the right-hand side are quadratic in the increments, and we discard the cubic terms to get

$$-\Big(\frac{m}{2}\widetilde{\sigma} + \delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\Big)^2 + \delta^3\widetilde{p}_{22}\big(\widetilde{p}_{11} + 2\delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\big) \geq 0. \tag{3.18}$$

By completing the square in the quadratic form, this may be equivalently rewritten as

$$\Big(\frac{m}{2}\widetilde{\sigma} + \delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\Big)^2 + \delta\Big(\frac{1}{2}\widetilde{p}_{11} + \delta\widetilde{p}_{12}\Big)^2 \leq \delta\Big(\frac{1}{2}\widetilde{p}_{11} + \delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\Big)^2. \tag{3.19}$$

(3) Finally $\widehat{T} \preceq 0$ requires $t_{22} \leq 0$ or $\widetilde{p}_{22}(\delta - \widetilde{\sigma}) \leq 0$; discarding the quadratic term, we get

$$\widetilde{p}_{22} \leq 0. \tag{3.20}$$

The proof concludes by applying the lemma below. ∎

LEMMA 3.7. *If the increments $\widetilde{\sigma} \leq 0$, $\widetilde{p}_{11}$, $\widetilde{p}_{12}$, $\widetilde{p}_{22}$ satisfy the constraints* (3.17)–(3.20)*, then $\widetilde{\sigma} = 0$, $\widetilde{p}_{11} = 0$, $\widetilde{p}_{12} = 0$, $\widetilde{p}_{22} = 0$.*

Proof. The relation (3.19) obviously implies

$$\Big(\frac{1}{2}\widetilde{p}_{11} + \delta\widetilde{p}_{12}\Big)^2 \leq \Big(\frac{1}{2}\widetilde{p}_{11} + \delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\Big)^2,$$

and therefore, in view of (3.20),

$$\frac{1}{2}\widetilde{p}_{11} + \delta\widetilde{p}_{12} \leq 0. \tag{3.21}$$

We combine this inequality with (3.17) to get

$$0 \leq -2\widetilde{p}_{12} + (1-\delta)^2\widetilde{p}_{22}$$

so that

$$\widetilde{p}_{12} \leq 0. \tag{3.22}$$

Since the three quantities being added in the first bracket in (3.19) are now known to be ≤ 0, it is enough to consider hereafter the worst case $\widetilde{\sigma} = 0$:

$$\Big(\delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\Big)^2 \leq \delta\Big(\frac{1}{2}\widetilde{p}_{11} + \delta\widetilde{p}_{12} + \delta^2\widetilde{p}_{22}\Big)^2.$$

Since $\delta\widetilde{p}_{12}+\delta^2\widetilde{p}_{22}\le 0$, we must have

$$\widetilde{p}_{11}\le 0. \tag{3.23}$$

From (3.17)

$$\widetilde{p}_{11}+2\delta\widetilde{p}_{12}+\delta^2\widetilde{p}_{22}\ge 2\widetilde{p}_{12}+(-1+2\delta)\widetilde{p}_{22},$$

which implies (see (3.20), (3.22), (3.23))

$$\widetilde{p}_{22}(\widetilde{p}_{11}+2\delta\widetilde{p}_{12}+\delta^2\widetilde{p}_{22})\le 2\widetilde{p}_{12}\widetilde{p}_{22}+(-1+2\delta)\widetilde{p}_{22}^2.$$

By combining this inequality and (3.18) (with $\widetilde{\sigma}=0$), we obtain a relation

$$\delta^2\widetilde{p}_{12}^2+\delta^3(1-\delta)\widetilde{p}_{22}^2\le 0$$

that shows that $\widetilde{p}_{12}=0$. Then comparing (3.17), (3.20), and (3.23), we conclude that $\widetilde{p}_{11}=\widetilde{p}_{22}=0$, which in turn concludes the proof. ∎

4. The differential equation.

Let us now set $h=\sqrt{\alpha}$ (so that $\delta=\sqrt{m}h$) and assume that in (3.1) the parameter $\beta=\beta_h$ changes smoothly with h in such a way that, for some constant $\bar{b}\in\mathbb{R}$, $\beta_h=1-\bar{b}\sqrt{m}h+o(h)$ as $h\downarrow 0$. Then, (3.1) may be written as

$$\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1})+\frac{1-\beta_h}{\sqrt{m}h}\sqrt{m}\frac{1}{h}(x_k-x_{k-1})+\nabla f(y_k)=0,$$

which, if x_k is seen as an approximation to $x(kh)$, provides a consistent discretization of the differential equation (1.5). An example is provided by the choice $\beta=(1-\delta)/(1+\delta)=(1-\sqrt{m}h)/(1+\sqrt{m}h)$, where $\bar{b}=2$ and (1.5) is the equation (1.4) used by Polyak.

Remark 4.1. In general, this two-step discretization is not a linear multistep formula. Note the following:

- ∇f is evaluated at y_k, a linear combination of x_k and x_{k-1}. In this regard, (3.1) is similar to the *one-leg* methods introduced by Dahlquist in his study of the long-time properties of multistep methods applied to nonlinear differential equations (see, e.g., [6, 5, 12])
- The unconventional factor $(1-\beta_h)/(\sqrt{m}h)$ that converges to $\bar{b}$ as $h\downarrow 0$. From the point of view of discretization methods for ODEs having $\bar{b}$ instead of this factor, or equivalently having $\beta=1-\bar{b}\sqrt{m}h$, would be more natural. But note that, when $\beta=(1-\sqrt{m}h)/(1+\sqrt{m}h)$, the algorithm (3.1) becomes GD for $h=1/\sqrt{L}$ and $\kappa=1$; the choice $\beta=1-\bar{b}\sqrt{m}h$ does not share this favorable property.

4.1. The construction.

We now define

$$v=\frac{1}{\sqrt{m}}\dot{x}$$

and rewrite (1.5) as a first-order system

$$\dot{v} = -\bar{b}\sqrt{m}v - \frac{1}{\sqrt{m}}\nabla f(x), \tag{4.1a}$$

$$\dot{x} = \sqrt{m}v. \tag{4.1b}$$

Remark 4.2. In a dimensional analysis as in Remarks 3.1 and 3.4, h has the same units as t. It is then a dimensional time-step, to be comparable with the nondimensional δ. The units of v are those of x. Of course, the divided difference (3.2) is a discrete version of $v = \dot{x}/\sqrt{m}$.

If we set $\xi = [v^{\mathsf{T}}, x^{\mathsf{T}}]^{\mathsf{T}}$, then (4.1) is of the form (2.6) with

$$\bar{A} = \begin{bmatrix} -\bar{b}\sqrt{m}I_d & 0_d \\ \sqrt{m}I_d & 0_d \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} -(1/\sqrt{m})I_d \\ 0_d \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} 0_d & I_d \end{bmatrix}.$$

Now according to Theorem 2.3, in order to find a Lyapunov function of the form (2.7) it is sufficient to find a matrix $\bar{P} \succeq 0$ and parameters $\lambda > 0$, $\sigma \geq 0$ such that the matrix $\bar{T}$ in (2.8) is negative semidefinite. Similarly to the discrete case, we will simplify the subsequent analysis by considering the case $\sigma = 0$. (The case $\sigma > 0$ is studied in the Appendix.) The Lipschitz constant L only enters T in Theorem 2.3 through $\bar{M}^{(3)}$; under the assumption $\sigma = 0$, $\bar{T}$ is independent of L. This has an important implication: The analysis in this section applies to f strongly m-convex but *not necessarily L-smooth.*

We look for $\bar{P}$ of the form

$$\bar{P} = \widehat{\bar{P}} \otimes I_d, \qquad \widehat{\bar{P}} = \begin{bmatrix} \bar{p}_{11} & \bar{p}_{12} \\ \bar{p}_{12} & \bar{p}_{22} \end{bmatrix}, \tag{4.2}$$

and then $\bar{T}$ is found to be

$$\bar{T} = \widehat{\bar{T}} \otimes I_d, \qquad \widehat{\bar{T}} = \begin{bmatrix} \bar{t}_{11} & \bar{t}_{12} & \bar{t}_{13} \\ \bar{t}_{12} & \bar{t}_{22} & \bar{t}_{23} \\ \bar{t}_{13} & \bar{t}_{23} & \bar{t}_{33} \end{bmatrix}, \tag{4.3}$$

where the $\bar{t}_{ij}$ have the following expressions:

$$\begin{aligned}
\bar{t}_{11} &= -2\bar{b}\bar{p}_{11} + 2\sqrt{m}\bar{p}_{12} + \lambda\bar{p}_{11}, \\
\bar{t}_{12} &= -\bar{b}\sqrt{m}\bar{p}_{12} + \sqrt{m}\bar{p}_{22} + \lambda\bar{p}_{12}, \\
\bar{t}_{13} &= -(1/\sqrt{m})\bar{p}_{11} + \sqrt{m}/2, \\
\bar{t}_{22} &= \lambda\bar{p}_{22} - (m/2)\lambda, \\
\bar{t}_{23} &= -(1/\sqrt{m})\bar{p}_{12} + \lambda/2, \\
\bar{t}_{33} &= 0.
\end{aligned}$$

We now determine λ and $\widehat{\bar{P}}$. The algebra is simplified if we set $\lambda = \sqrt{m}\,\bar{r}$.

First step. Since $\bar{t}_{33} = 0$, the requirement $\widehat{\bar{T}} \preceq 0$ implies $\bar{t}_{13} = 0$ and $\bar{t}_{23} = 0$ and accordingly

$$\bar{p}_{11} = m/2, \qquad \bar{p}_{12} = (m/2)\bar{r}. \tag{4.4}$$

Second step. We choose $\bar{p}_{22}$ to ensure $\det(\widehat{\bar{P}}) = \bar{p}_{11}\bar{p}_{22} - \bar{p}_{12}^2 = 0$. This yields

$$\bar{p}_{22} = (m/2)\bar{r}^2$$

and leads to

$$(4.5) \qquad \hat{\bar{P}} = \frac{m}{2}\begin{bmatrix} 1 & \bar{r} \\ \bar{r} & \bar{r}^2 \end{bmatrix},$$

a matrix that is positive semidefinite (but not positive definite).

Third step. Since, $\hat{\bar{T}} \preceq 0$ implies $\bar{t}_{22} \leq 0$, we may write $0 \geq \bar{p}_{22} - m/2 = (m/2)(\bar{r}^2 - 1)$, and therefore we have

$$\bar{r} \leq 1;$$

this imposes a bound $\lambda \leq \sqrt{m}$ on the convergence rate.

Fourth step. We impose the condition $\bar{t}_{11}\bar{t}_{22} - \bar{t}_{12}^2 = 0$. This results in an equation $\bar{\Xi} = 0$,

$$(4.6) \qquad \bar{\Xi}(\bar{r}, \bar{b}) = \bar{r}b^2 - 2(\bar{r}^2 + 1)b + \bar{r}^3 + 3\bar{r},$$

that relates $\bar{r}$ (or equivalently the rate λ) and the parameter $\bar{b}$ in the differential equation (1.5).

We observe that the polynomial $\bar{\Xi}$ is the limit as $\delta \downarrow 0$ of the polynomial Ξ_δ in (3.12) (except of course for the symbols used to denote the variables: r and b for Ξ_δ and $\bar{r}$ and $\bar{b}$ for $\bar{\Xi}$). As a consequence, the discontinuous line in Figure 3.1, presented there as a limit of curves $\Xi_\delta = 0$, also describes the curve $\bar{\Xi} = 0$ (again after renaming the variables).

The curve of equation $\bar{\Xi}(\bar{r}, \bar{b}) = 0$ in the $(\bar{r}, \bar{b})$ plane is invariant with respect to the symmetry $(\bar{r}, \bar{b}) \mapsto (-\bar{r}, -\bar{b})$ (this is a consequence of the fact that changing $\bar{b}$ into $-\bar{b}$ in the differential equation is equivalent to reversing the sign of independent variable t).[1] The formula for the roots of a quadratic equation gives

$$\bar{b}_\pm = \frac{1 + \bar{r}^2 \pm \sqrt{1 - \bar{r}^2}}{\bar{r}}.$$

From here one may prove that to each real $\bar{b}$ there corresponds a unique $\bar{r}$ such that $\bar{\Xi}(\bar{r}, \bar{b}) = 0$. The maximum value $\bar{r} = 1$ ($\lambda = \sqrt{m}$) is achieved only for $\bar{b} = 2$ (i.e., for Polyak's (1.4)), and values $\bar{r} \in (0, 1)$ correspond to two different real values of $\bar{b}$.

We now have the following result that is proved as in the discrete case.

THEOREM 4.3. *Consider the differential equation* (1.5) *(or the equivalent system* (4.1)*) with parameter $\bar{b} > 0$, and assume that f is m-strongly convex. Let $\lambda = \sqrt{m}\bar{r}$, where $\bar{r} > 0$ is the value determined by the relation $\bar{\Xi}(\bar{r}, \bar{b}) = 0$ (see* (4.6)*), and define the positive semidefinite matrix $\bar{P}$ by* (4.2) *and* (4.5)*. Then the matrix $\bar{T}$ in* (4.3) *is negative semidefinite.*

As a result, if $x(t)$ is a solution of (1.5)*, the function*

$$(4.7) \qquad \exp(\lambda t)\Big(f(x(t)) - f(x_\star) + [v(t)^\mathsf{T}, x(t)^\mathsf{T} - x_\star^\mathsf{T}]\, \bar{P}\, [v(t)^\mathsf{T}, x(t)^\mathsf{T} - x_\star^\mathsf{T}]^\mathsf{T}\Big)$$

decreases monotonically as t increases, which implies

$$f(x(t)) - f(x_\star) \leq \bar{C}\exp(-\lambda t)$$

[1] The curves $\Xi_\delta(r, b) = 0$, $\delta > 0$ do not possess any symmetry because in the discrete algorithm (3.1), x_{k+1} and x_{k-1} do not play a symmetric role (or in the terminology of differential equation integrators we are not dealing with time-symmetric algorithms).

with

$$\bar{C} = f(x(0)) - f(x^\star) + \frac{m}{2}\left\|\frac{1}{\sqrt{m}}\dot{x}(0) + \bar{r}(x(0) - x^\star)\right\|^2.$$

Remark 4.4. For $\bar{b} = 0$, the construction leading to the theorem yields $r = 0$, i.e., $\lambda = 0$, and,

$$(\xi(t) - \xi_\star)^\mathsf{T} \bar{P}(\xi(t) - \xi_\star) = \frac{m}{2}\|v\|^2.$$

In addition, $\bar{T} = 0$, and therefore the factor in round brackets in (4.7) is an invariant of motion. In this case the system (4.1) is Hamiltonian, and the invariant we have found equals $\sqrt{m}$ times the corresponding Hamiltonian function.

Remark 4.5. The value $\bar{b} = 2$, in addition to maximizing the decay rate in $f(x(t))$ in Theorem 4.3 for arbitrary m-strongly convex f, has another optimality property in the simple one-dimensional case with $f(x) = mx^2/2$, when (1.5) or (4.1) describe a damped harmonic oscillator. An elementary computation (see, e.g., [33]) shows that $\bar{b} = 2$ is the value of the friction coefficient that ensures the *fastest dissipation of the energy* $(\dot{x})^2/2 + mx^2/2$.

It will be proved in the Appendix that if f, in addition to being strongly convex has Lipschitz continuous gradient, then better decay rates in $f(x(t))$ may be obtained by choosing $\bar{b}$ to be larger than 2. Therefore $(\dot{x})^2/2 + mx^2/2$ is not the best Lyapunov function to study the rate of decay of $f(x)$ in the damped harmonic oscillator. This is in agreement with Theorem 4.6 below.

Reference [22] gives a Lyapunov function for (1.5) or (4.1) that includes a cross-term $v^T\nabla f(x)$ and does not require the strong convexity of f. However, the presence of the gradient in the Lyapunov function makes it necessary that f be demanded to be twice-differentiable (the Hessian of f appears when differentiating the Lyapunov function with respect to t).

4.2. Optimality. Steps 2 and 4 in the construction above imply a degree of arbitrariness and it is of interest to ask whether there are alternative choices of λ and $\widehat{\bar{P}} \succeq 0$ that, while ensuring $\widehat{\bar{T}} \preceq 0$, furnish better decay rates. We conclude this section by proving that this is not the case.

In the theorem below we use the notation $\bar{r}^\star$ and $\widehat{\bar{P}}^\star$ for the values obtained, for given $\bar{b} > 0$, in the construction leading to Theorem 4.3. (These are functions $\bar{r}^\star = \bar{r}^\star(b)$ and $\widehat{\bar{P}}^\star = \widehat{\bar{P}}^\star(b)$, but the dependence on $\bar{b}$ will be dropped from the notation.) In particular, $\bar{p}^\star_{22} = m\bar{r}^{\star 2}/2$ and $\bar{\Xi}(\bar{r}^\star, \bar{b}) = 0$. The symbols λ and $\widehat{\bar{P}}$ are used in the theorem to refer to an arbitrary real number and an arbitrary 2×2 symmetric matrix. Finally, we set $\lambda^\star = \sqrt{m}\,\bar{r}^\star$ and $\lambda = \sqrt{m}\,\bar{r}$.

THEOREM 4.6. *With the notation as described, for each fixed* $\bar{b} > 0$*,* $\lambda^\star = \max\lambda$*, subject to the constraints* $\widehat{\bar{T}}(\lambda, \widehat{\bar{P}}) \preceq 0$*,* $\widehat{\bar{P}} \succeq 0$.

Proof. Since we are solving a convex optimization problem, it is sufficient to show that $(\lambda^\star, \widehat{\bar{P}}^\star)$ provides a *local* maximum.

We observed in step 1 above that $\widehat{\bar{T}} \preceq 0$ determines the values of $\bar{p}_{11}$, $\bar{p}_{12}$ as in (4.4). This leaves us with λ (or equivalently $\bar{r}$) and $\bar{p}_{22}$ as decision variables. For simplicity we hereafter omit the subindices in $\bar{p}_{22}$.

The constraint $\widehat{\bar{P}} \succeq 0$ implies $\det(\widehat{\bar{P}}) \geq 0$ or (after using the values of $\bar{p}_{11}$, $\bar{p}_{12}$) $\bar{p} \geq (m/2)\bar{r}^2$. The constraint $\widehat{\bar{T}} \preceq 0$ implies $\bar{t}_{11}\bar{t}_{22} - \bar{t}_{12}^2 \geq 0$. We use (4.4) to write $\bar{t}_{11}\bar{t}_{22} - \bar{t}_{12}^2 \geq 0$ as a function $\Delta(\bar{r}, \bar{p})$; tedious algebra leads to the expression:

$$\Delta(\bar r,\bar p) = -\frac{m^3}{2}\bar r^4 + \frac{\bar b m^3}{2}\bar r^3 + \left(\frac{m^2\bar p}{2} - \frac{3m^3+\bar b^2 m^3}{4}\right)\bar r^2 + \frac{bm^3}{2}\bar r - m\bar p^2.$$

We will be done if we prove that the pair $(\bar r^\star,\bar p^\star)$ is a local maximum for the problem

$$\max \bar r \quad \text{subject to} \quad \bar p - m\bar r^2/2 \geq 0,\ \ \Delta(\bar r,\bar p)\geq 0.$$

At the point $(\bar r^\star,\bar p^\star)$ both constraints are active (in fact they were chosen to be so at steps 2 and 4). If we define the Lagrangian

$$\mathcal{L}(\bar r,\bar p) = \bar r + \zeta_1\,(\bar p - m\bar r^2/2) + \zeta_2\,\Delta(\bar r,\bar p),$$

where ζ_1, ζ_2 are the multipliers, the proof concludes by showing that the gradient of $\mathcal{L}$ at $(\bar r^\star,\bar p^\star)$ may be annihilated for a suitable choice of *positive* multipliers.

We impose the requirements

$$0 = \frac{\partial}{\partial \bar r}\mathcal{L}\Big|^\star = 1 - \zeta_1 m\bar r^\star + \zeta_2\,\frac{\partial}{\partial\bar r}\Delta\Big|^\star$$

($|^\star$ means evaluation at at $(\bar r^\star,\bar p^\star)$) and

$$0 = \frac{\partial}{\partial \bar p}\mathcal{L}\Big|^\star = \zeta_1 + \zeta_2\left(\frac{m^2}{2}\bar r^{\star 2} - 2m\bar p^\star\right) = \zeta_1 - \zeta_2\frac{m^2}{2}\bar r^{\star 2},$$

(which implies that ζ_1 and ζ_2 have the same sign) and eliminate ζ_1 to get

$$1 + \zeta_2\left(\frac{m^3}{2}\bar r^{\star 3} + \frac{\partial}{\partial \bar r}\Delta\Big|^\star\right) = 0.$$

In this way we are left with the task of proving that

$$\frac{m^3}{2}\bar r^{\star 3} + \frac{\partial}{\partial\bar r}\Delta\Big|^\star < 0,$$

or, after using the expression for Δ and some simplification,

$$-2\bar r^{\star 3} + 3\bar b\bar r^{\star 2} - (3+\bar b^2)\bar r^\star + \bar b < 0.$$

Let us denote by $\Lambda = \Lambda(\bar r^\star,\bar b)$ the left-hand side of this inequality. When $\bar b = 2$ and $\bar r^\star = 1$, we have $\Lambda = -1$. On the other hand, we know that

$$\bar\Xi = \bar b^2\bar r - 2(\bar r^{\star 2}+1)\bar b + \bar r^{\star 3} + 3\bar r^\star = 0,$$

and this relation makes it impossible for Λ to change sign as $\bar b > 0$ and the corresponding $\bar r^\star(b)\in(0,1]$ vary. In fact, if Λ were to vanish, we would have

$$\Lambda + \bar\Xi = \big(\bar r^{\star 2} - 1\big)\bar b - \bar r^{\star 3} = 0,$$

something that cannot happen because $\bar r^\star < 1$ for $\bar b \neq 2$. $\square$

5. Connecting the differential equations with optimization algorithms. The second-order differential equation (1.5) provides a limit for the algorithm (3.1) when β changes smoothly with $h = \sqrt{\alpha}$ in such a way that $\beta_h = 1 - \bar{b}\sqrt{m}h + o(h)$ as $h \downarrow 0$. In this section we study this limit when $\bar{b} > 0$. As in (3.8) write $\beta_h = 1 - b_h\delta = 1 - b_h\sqrt{m}h$. Clearly, $b_h \to \bar{b}$ and, in addition, for h sufficiently small $b_h \in (b^h_{\min}, b^h_{\max})$ (see (3.14)). The application of Theorem 3.3 then gives a rate $\rho_h^2 = 1 - r_h\delta = 1 - r_h\sqrt{m}h$. As noted before, the polynomial $\bar{\Xi}$ in (4.6) is the limit of Ξ_δ in (3.12) as h (or δ) approaches zero, and, accordingly, $r_h \to \bar{r}$, where $\bar{r}$ solves $\bar{\Xi}(\bar{r}, \bar{b}) = 0$. Then Theorem 3.3 guarantees that, over one step $k \mapsto k+1$ of the algorithm, $f(x_k) - f(x^\star)$ decays by a factor $\rho_h^2 = 1 - \sqrt{m}\bar{r}h + o(h)$. Over k steps the decay factor will be $(1 - \sqrt{m}\bar{r}h + o(h))^k$, a quantity that in the limit $kh \to t$ converges to $\exp(-\sqrt{m}\bar{r}t) = \exp(-\lambda t)$. This is exactly the decay guaranteed by Theorem 4.3 for $f(x(t)) - f(x^\star)$ over an interval of length t.

In addition, the matrices P_h in the discrete Lyapunov function converge to the matrix $\widehat{P}$ in the differential equation, because from the expression for the entries in (3.11) and (4.5)

$$p^h_{11} \to \bar{p}_{11}, \qquad p^h_{12} \to \bar{p}_{12}, \qquad p^h_{22} \to \bar{p}_{22}.$$

The above discussion and standard results on the convergence of discretizations of ODEs imply the following result.

THEOREM 5.1. *Fix the parameter $\bar{b} > 0$ and the initial conditions $x(0)$, $\dot{x}(0)$ for the differential equation* (1.5). *For small $h > 0$, consider the optimization algorithm* (3.1) *with parameters $\alpha = h^2$ and $\beta = \beta_h = 1 - \bar{b}\sqrt{m}h + o(h)$. Assume that the initial points x_{-1}, x_0 are such that, as $h \downarrow 0$, $x_0 \to x(0)$ and $(1/h)(x_0 - x_{-1}) \to \dot{x}(0)$. Then, in the limit $kh \to t$,*

1. *$x_k \to x(t)$ and $(1/h)(x_{k+1} - x_k) \to \dot{x}(t)$.*
2. *The discrete Lyapunov function in* (3.16) *converges to the Lyapunov function in* (4.7).

Remark 5.2. As a consequence of this theorem, the Lyapunov function of the differential equation could have been derived alternatively by first finding the Lyapunov function for the discrete optimization algorithm and then taking limits. In our research we first investigated the discrete case and then studied the differential equations; in hindsight we saw it would have been easier to first deal with the differential equation and then carry out the analysis of the algorithm by mimicking the treatment of the continuous case. References [28, 29, 14] find Lyapunov functions for different optimization algorithms by first constructing Lyapunov functions for suitable so-called high-resolution differential equations. In our context, this would mean perturbing (4.1) with suitable h-dependent terms so as to obtain an (h-dependent) differential equation for which the algorithm has a high order of consistency. The idea behind those high-resolution equations is very old in the numerical analysis of ordinary and partial differential equations, where they are known as *modified equations*; see, e.g., [11] or [23, Chapter 10] and, for the stochastic case, [34].

6. Heavy ball and other methods. The paper [30] has given rise to a number of contributions that aim to understand the behavior of optimization methods by seeing them as discretizations of differential equations. However it is well known that the long-time properties of a differential equation are not automatically inherited by their discretizations, regardless of the value of the step-size chosen. A very simple example is provided by the application of Euler's rule to the harmonic oscillator: For all step-sizes the discrete trajectories grow while the continuous solutions stay bounded. A more relevant example in an optimization context may be seen in [25]. On the other hand properties of the discretizations may often be extrapolated to the continuous limit; a general discussion of these points in different settings may be seen in [1].

In the setting of the preceding section, it is not true that discretizing a dissipative differential equation with a known a Lyapunov function will always yield an optimization algorithm with a "suitable" Lyapunov function. We now illustrate this fact by means of the heavy ball algorithm obtained by choosing $\gamma = 0$ and $\beta \neq 0$ in (2.2).

We proceed as in section 3: rewrite the algorithm in terms of d_k and x_k and then cast it in the general format (2.1). We will presently prove that a discrete Lyapunov *with properties similar to the Lyapunov function for Nesterov's method in Theorem* 3.3 *does not exist*. We argue by contradiction. With the notation as in section 3, we consider

- $p_{ij} = m\,\phi_{ij}(\beta,\delta)$, $(i,j) = (1,1),(1,2),(2,2)$ such that $\widehat{P} \succeq 0$,
- $r = \psi(\beta,\delta) > 0$,
- $c > 0$

and suppose that the corresponding $T(\lambda, P)$ is $\preceq 0$ for each $\delta < c/\sqrt{\kappa}$. As in Remark 3.4 to ensure equivariance with respect to changes of scale, the number c and functions ϕ_{ij} and ψ are assumed to be independent of the constants m and L associated with f and the values of the parameters α and β in the heavy ball algorithm.

For future reference, the element t_{11} is found to have the expression

$$t_{11} = (\beta^2 - \rho^2)p_{11} + 2\delta\beta^2 p_{12} + \delta^2\beta^2 p_{22} + \delta^2(L-m)\beta^2/2.$$

This has to be ≤ 0 for $\delta < c/\sqrt{\kappa}$.

Next, as in the preceding section, we assume that β changes smoothly with h in such a way that, for some $\bar{b} > 0$, $\beta = \beta_h = 1 - \bar{b}\delta + o(h) = 1 - \bar{b}\sqrt{m}h + o(h)$. Clearly the algorithm is then a consistent discretization of the differential equation (1.5), and we assume that r_h, p_{ij}^h converge to their differential equation counterparts $\bar{r}$ and $\bar{p}_{ij}$.[2]

In this situation

$$0 \geq \delta^{-1} t_{11}^h = \frac{\beta_h^2 - \rho_h^2}{\delta} p_{11}^h + 2\beta_h^2 p_{12}^h + \delta\beta_h^2 p_{22}^h + \frac{c}{2}\sqrt{\frac{m}{L}}(L-m)\beta_h^2,$$

and, taking limits,

$$0 \geq -2\frac{\bar{b}-\lambda}{\sqrt{m}}\bar{p}_{11} + 2\bar{p}_{12} + \frac{c}{2}\sqrt{\frac{m}{L}}(L-m). \tag{6.1}$$

This cannot happen because L may be arbitrarily large.

Remark 6.1. The heavy ball algorithm is a "more natural" discretization of (1.5) than Nesterov's, in that, as conventional linear multistep methods, it does not evaluate ∇f at a linear combination of x_k, x_{k-1} (cf. Remark 4.1).

Remark 6.2. The contradiction in (6.1) arises because we insisted on T being $\preceq 0$ for "large" nondimensional stepsizes $\delta = \sqrt{m}h < c/\sqrt{\kappa}$. For optimization algorithms that, in the limit $h \downarrow 0$, approximate a differential equation with decay $\exp(-\lambda h) = \exp(-\bar{r}\delta)$ in a time-interval of length h, such large stepsizes seem to be necessary to achieve accelerated rates $1 - \mathcal{O}(\sqrt{\kappa})$ rather than rates $1 - \mathcal{O}(\kappa)$.

[2]This hypothesis is not necessarily in the argument that follows. It is enough to suppose that r_h, p_{ij}^h have finite limits.

The reference [28] constructs a Lyapunov function for the heavy ball method, but it only operates for $\delta = \mathcal{O}(1/\kappa)$ and, while useful in showing convergence, does not provide acceleration. For an additional convergence proof of the heavy ball algorithm see [10]; again this reference does not prove acceleration.

The three-parameter family of methods (2.2) contains algorithms, like Nesterov's, that "inherit" the ODE Lyapunov function for stepsizes $\delta < c/\sqrt{\kappa}$ and algorithms, like the heavy ball, that do not. In fact the situation for the heavy ball is arguably the rule rather than the exception. For (2.2),

$$t_{11} = (\beta^2 - \rho^2)p_{11} + 2\delta\beta^2 p_{12} + \delta^2\beta^2 p_{22} + \delta^2(L-m)(\beta-\gamma)^2/2 - m\gamma^2\delta^2/2,$$

where we observe the unwelcome presence of the factor $L - m$ that created the difficulties in the analysis of the heavy ball algorithm. If we look at a situation where β changes with h as above and in addition γ is also allowed to change with h and approaches a limit, a Lyapunov function that has the form envisaged and works for $\delta < c/\sqrt{\kappa}$ may only exist if $\beta_h - \gamma_h$ vanishes (at least in the limit $h \downarrow 0$) to offset the factor, i.e., if the algorithm is not far away from Nesterov's.

Appendix. In Theorem 4.6 we proved that, for each $\bar{b} > 0$, the rate of decay λ provided by Theorem 4.3 is the best one may obtain by using Theorem 2.3 *if one chooses* $\sigma = 0$. In this appendix we investigate whether λ may be improved by a suitable choice of $\sigma > 0$. Since for $\sigma \neq 0$, the matrix $\bar{M}^{(3)}$ that contains the constant L contributes to T, the following results require that f, in addition to being m-strongly convex (as in Theorem 4.3) is L-smooth; i.e., they hold for $f \in \mathcal{F}_{m,L}$.

When $\sigma \neq 0$ the expressions for the t_{ij} in section 4 have to be replaced by

$$\begin{aligned}
\bar{t}_{11} &= -2\bar{b}\bar{p}_{11} + 2\sqrt{m}\bar{p}_{12} + \lambda\bar{p}_{11},\\
\bar{t}_{12} &= -\bar{b}\sqrt{m}\bar{p}_{12} + \sqrt{m}\bar{p}_{22} + \lambda\bar{p}_{12},\\
\bar{t}_{13} &= -(1/\sqrt{m})\bar{p}_{11} + \sqrt{m}/2,\\
\bar{t}_{22} &= \lambda\bar{p}_{22} - (m/2)\lambda - \sigma m L/(m+L),\\
\bar{t}_{23} &= -(1/\sqrt{m})\bar{p}_{12} + \lambda/2 + \sigma/2,\\
\bar{t}_{33} &= -\sigma/(m+L).
\end{aligned}$$

As in section 4, we set $\lambda = \sqrt{m}\,\bar{r}$ and, in addition, $\sigma = m\bar{s}$ (the variable $\bar{s}$ is, as $\bar{r}$, nondimensional). We shall show that it is possible, for given m and L, to find values of the six parameters $\bar{p}_{11}$, $\bar{p}_{12}$, $\bar{p}_{22}$, $\bar{b}$, $\bar{s}$, $\bar{r}$, in such a way that the constraints $\hat{\bar{T}} \preceq 0$, $\hat{\bar{P}} \succeq 0$, $\bar{s} \geq 0$ are satisfied and, at the same time, $\bar{r} > 1$, so that by using the matrix $\bar{M}^{(3)}$ it is possible to improve on the best value $\bar{r} = 1$ (associated with $\bar{b} = 2$ and leading to $\lambda = \sqrt{m}$) that may be achieved in Theorem 4.3.

For given m and L, we determine the values of the six parameters as follows.

First step. We impose $\bar{t}_{22} = 0$, a requirement that leads to the relation

$$\frac{\bar{p}_{22}}{m} = \frac{1}{2} + \frac{\bar{s}}{\bar{r}}\frac{\kappa}{\kappa+1}.$$

Second step. We impose $\bar{t}_{23} = 0$ and get

$$\frac{\bar{p}_{12}}{m} = \frac{\bar{r}+\bar{s}}{2}.$$

Third step. We require $\det(\widehat{\bar{P}}) = 0$. Therefore

$$\frac{\bar{p}_{11}}{m} = \frac{(\bar{p}_{12}/m)^2}{\bar{p}_{22}/m}.$$

Note that for $\bar{r}, \bar{s} \geq 0$ we have $\bar{p}_{22} > 0$, and thus the third step guarantees that $\widehat{\bar{P}} \succeq 0$.

Fourth step. We next demand that $\bar{t}_{12} = 0$ and obtain

$$\bar{b} = \bar{r} + \frac{\bar{p}_{22}/m}{\bar{p}_{12}/m}.$$

The four preceding displayed formulas allow us to express the parameters $\bar{p}_{12}$, $\bar{p}_{22}$, and $\bar{b}$ as known functions of $\bar{s}$ and $\bar{r}$.

Fifth step. At this stage, we have ensuble that $\bar{t}_{12}$, $\bar{t}_{22}$, $\bar{t}_{23}$ vanish. As a result, the condition $\widehat{\bar{T}} \preceq 0$ is equivalent to $\widehat{\bar{T}}^{13} \preceq 0$ where $\widehat{\bar{T}}^{13}$ is the 2×2 matrix obtained by suppressing from $\widehat{\bar{T}}$ its second row and column. Furthermore $\bar{t}_{33} < 0$ for $\bar{s} > 0$ and then we shall have $\widehat{\bar{T}}^{13} \preceq 0$ if we impose that $\det(\widehat{\bar{T}}^{13}) = 0$, or

$$\bar{t}_{11}\bar{t}_{33} - \bar{t}_{13}^2 = 0.$$

By using the displayed formulas above, the last equation becomes a relation $F(\bar{r}, \bar{s}) = 0$, between $\bar{r}$ and $\bar{s}$, with

$$F = \frac{\bar{r}^2 \bar{s} (\bar{r} + \bar{s})^2}{2(\kappa+1)\bar{r} + 4\kappa\bar{s}} - \frac{1}{4}\left(\frac{(\kappa+1)\bar{r}(\bar{r}+\bar{s})^2}{(\kappa+1)\bar{r} + 2\kappa\bar{s}} - 1\right)^2.$$

We next show that the rational curve $F(\bar{r}, \bar{s}) = 0$ in the $(\bar{r}, \bar{s})$ real plane has points with $\bar{s} > 0$ and $\bar{r} > 1$.

It is easily checked that the point $\bar{r} = 1$, $\bar{s} = 0$ lies on the curve $F = 0$ and has $\bar{b} = 0$. This could have been anticipated because, if $\bar{s} = 0$ and $\bar{b} = 2$, the construction in this appendix just reproduces the construction in section 4, which yields $\bar{r} = 1$.

By removing the denominator in the rational function F so as to have a polynomial equation for the curve and looking at the Newton diagram at $\bar{r} = 1$, $\bar{s} = 0$, one sees that in the neighborhood of this point the curve consists of a single branch that may be parameterized by $\bar{r}$. A Taylor expansion reveals that

$$\bar{s} = 2(\kappa+1)(\bar{r}-1)^2 + \mathcal{O}((\bar{r}-1)^3).$$

In this way, choosing a sufficiently small value of the parameter $\bar{s} > 0$, there are two possible values of the rate $\bar{r}$

$$\bar{r} \approx 1 \pm \sqrt{\frac{\bar{s}}{2(\kappa+1)}},$$

one of which is > 1. In conclusion we have proved analytically that the introduction of σ and $\bar{M}^{(3)}$ in T makes it possible to *achieve rates* $\bar{r} > 1$ (or $\lambda > \sqrt{m}$).

We next determined the value of $\bar{s}$ that leads to the largest possible $\bar{r}$ on the curve $F = 0$. In view of the involved expression of F, we proceeded numerically and found this largest value by continuation along the curve, starting from $\bar{r} = 1$, $\bar{s} = 0$. The results, for different values of κ, are given in Table 6.1. For the small condition number $\kappa = 10$, the table shows that it is possible to achieve a decay $\approx \exp(-1.086\sqrt{m}t)$ by fixing the dissipation coefficient at the value $\bar{b} \approx 2.35$ rather than at $\bar{b} = 2$ as in

TABLE 6.1

Value of the dissipation parameter $\bar{b}$ in the differential equation that leads to the best rate of decay $\bar{r}$ for different choices of the condition number κ. The table also gives the values of the parameters to construct the matrices $\widehat{\overline{T}} \preceq 0$, $\widehat{\overline{P}} \succeq 0$.

κ	$\bar{b}-2$	$\bar{r}-1$	$\bar{s}$	$\frac{\bar{p}_{11}}{m}-\frac{1}{2}$	$\frac{\bar{p}_{12}}{m}-\frac{1}{2}$	$\frac{\bar{p}_{22}}{m}-\frac{1}{2}$
10^1	3.5(-1)	8.6(-2)	4.1(-1)	1.6(-1)	2.5(-1)	3.4(-1)
10^2	2.2(-1)	1.8(-2)	1.3(-1)	2.7(-2)	7.6(-2)	1.3(-1)
10^3	1.0(-1)	3.9(-3)	5.5(-2)	5.2(-3)	2.9(-2)	5.5(-2)
10^4	4.7(-2)	8.2(-4)	2.4(-2)	1.1(-3)	1.3(-2)	2.4(-2)
10^5	2.1(-2)	1.8(-4)	1.1(-2)	2.3(-4)	5.5(-3)	1.1(-2)
10^6	9.9(-3)	3.8(-5)	5.0(-3)	5.0(-5)	2.5(-3)	5.0(-3)
10^7	4.6(-3)	8.1(-6)	2.3(-3)	1.1(-5)	1.2(-3)	2.3(-3)
10^8	2.2(-3)	1.7(-6)	1.1(-3)	2.3(-6)	5.4(-4)	1.1(-3)
10^9	9.9(-4)	3.8(-7)	5.0(-4)	5.0(-7)	2.5(-4)	5.0(-4)

Polyak's (1.4)—this is a marginal improvement on the best decay $\exp(-\sqrt{m}t)$ that one may insure without using $\bar{M}^{(3)}$. In addition the improvement quickly decreases as the condition number grows: for $\kappa = 10^3$ the decay is $\exp(-1.0039\sqrt{m}t)$. In fact, we observe in the table that, as $\kappa \uparrow \infty$, $\bar{r} \approx 1 + 0.38\kappa^{-2/3}$. Of course as κ increases, $\bar{r}$ and $\bar{b}$ approach the values 1 and 2 that correspond to the situation studied in section 4, where f is not assumed to possess Lipschitz gradients. A similar convergence obtains for the matrix $\widehat{\overline{P}} \succeq 0$. Also note that $\bar{s} \approx 0.50\kappa^{-1/3}$: As the condition number increases the parameter $\sigma = \sqrt{m}\bar{s}$ that multiplies $\bar{M}^{(3)}$ decreases, as it may have been expected.

The results in the appendix and the connection between discrete and continuous Lyapunov functions strongly suggest that there would have been no substantial gain in the rate ρ^2 found in section 3 if we had allowed $\ell \neq 0$ there.

Acknowledgment. We are thankful to an anonymous referee for helping us to improve the discussion of our results.

REFERENCES

[1] U. M. ASCHER, *Discrete processes and their continuous limits*, J. Dyn. Games, 7 (2020), p. 123.

[2] M. BETANCOURT, M. I. JORDAN, AND A. C. WILSON, *On Symplectic Optimization*, preprint, arXiv:1802.03653, 2018.

[3] N. BOU-RABEE AND J. M. SANZ-SERNA, *Randomized hamiltonian monte carlo*, Ann. Appl. Probab., 27 (2017), pp. 2159–2194, https://doi.org/10.1214/16-AAP1255.

[4] A. BRAVETTI, M. L. DAZA-TORRES, H. FLORES-ARGUEDAS, AND M. BETANCOURT, *Optimization Algorithms Inspired by the Geometry of Dissipative Systems*, preprint, arXiv:1912.02928, 2019, https://arxiv.org/abs/arXiv:1912.02928.

[5] J. C. BUTCHER, *Numerical Methods for Ordinary Differential Equations*, 3rd ed., John Wiley & Sons, Ltd., Chichester, 2016, https://doi.org/10.1002/9781119121534.

[6] G. G. DAHLQUIST, *Error analysis for a class of methods for stiff non-linear initial value problems*, in Numerical Analysis, G. A. Watson, ed., Springer Berlin Heidelberg, 1976, pp. 60–72.

[7] M. J. EHRHARDT, E. S. RIIS, T. RINGHOLM, AND C.-B. SCHÖNLIEB, *A Geometric Integration Approach to Smooth Optimisation: Foundations of the Discrete Gradient Method*, preprint, arXiv:1805.06444, 2018, https://arxiv.org/abs/arXiv:1805.06444.

[8] M. FAZLYAB, A. RIBEIRO, M. MORARI, AND V. M. PRECIADO, *Analysis of optimization algorithms via integral quadratic constraints: nonstrongly convex problems*, SIAM J. Optim., 28 (2018), pp. 2654–2689, https://doi.org/10.1137/17M1136845.

[9] G. FRANÇA, M. I. JORDAN, AND R. VIDAL, *On Dissipative Symplectic Integration with Applications to Gradient-Based Optimization*, preprint, arXiv:2004.06840, 2020, https://arxiv.org/abs/arXiv:2004.06840.

[10] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, *Global convergence of the heavy-ball method for convex optimization*, in Proceedings of the 2015 European Control Conference (ECC), 2015, pp. 310–315, https://doi.org/10.1109/ECC.2015.7330562.

[11] D. F. Griffiths and J. M. Sanz-Serna, *On the scope of the method of modified equations*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 994–1008, https://doi.org/10.1137/0907067.

[12] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin and Heidelberg, 1996.

[13] W. Krichene, A. Bayen, and P. L. Bartlett, *Accelerated mirror descent in continuous and discrete time*, in Adv. Neural Inform. Process. Syst. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2015, pp. 2845–2853.

[14] M. Laborde and A. Oberman, *A lyapunov analysis for accelerated gradient methods: From deterministic to stochastic case*, Proceedings of the International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, 2020, pp. 602–612, http://proceedings.mlr.press/v108/laborde20a.html.

[15] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM J. Optim., 26 (2016), pp. 57–95, https://doi.org/10.1137/15M1009597.

[16] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[17] M. Muehlebach and M. Jordan, *A dynamical systems perspective on Nesterov acceleration*, in Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, 2019, pp. 4656–4662, http://proceedings.mlr.press/v97/muehlebach19a.html.

[18] M. Muehlebach and M. I. Jordan, *Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives*, preprint, arXiv:2002.12493, 2020, https://arxiv.org/abs/arXiv:2002.12493.

[19] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, 1st ed., Springer, New York, 2014.

[20] A. Orvieto and A. Lucchi, *Shadowing properties of optimization algorithms*, in Adv. Neural Inform. Process. Syst. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, eds., Curran Associates, Red Hook, NY, 2019, pp. 12692–12703.

[21] B. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys., (1964), pp. 1–17.

[22] B. Polyak and P. Shcherbakov, *Lyapunov functions: An optimization theory perspective*, IFAC, 50 (2017), pp. 7456–7461, https://doi.org/https://doi.org/10.1016/j.ifacol.2017.08.1513.

[23] J. Sanz-Serna and M. Calvo, *Numerical Hamiltonian Problems*, Dover Books on Mathematics, Dover Publications, 2018.

[24] J. Sanz-Serna and A. Stuart, *Ergodic properties of dissipative differential equations subject to random impulses*, J. Diff. Eq., 155 (1999), pp. 262–284.

[25] J. M. Sanz Serna and K. C. Zygalakis, *Contractivity of Runge–Kutta methods for convex gradient systems*, SIAM J. Numer. Anal., 58 (2020), pp. 2079–2092, https://doi.org/10.1137/19M1299256.

[26] D. Scieur, A. d'Aspremont, and F. Bach, *Regularized nonlinear acceleration*, in Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 712–720.

[27] D. Scieur, V. Roulet, F. R. Bach, and A. d'Aspremont, *Integration methods and optimization algorithms*, in Proceedings of the Annual Conference on Neural Information Processing Systems 2017, pp. 1109–1118.

[28] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, *Understanding the Acceleration Phenomenon via High-Resolution Differential Equations*, preprint, arXiv:1810.08907, 2018, https://arxiv.org/abs/arXiv:1810.08907.

[29] B. Shi, S. S. Du, W. Su, and M. I. Jordan, *Acceleration via symplectic discretization of high-resolution differential equations*, in Adv. Neural Inform. Process. Syst., H. Wallach, H. Larochelle, A. Beygelzimer, F. eds., vol. 32, Curran Associates, Red Hook, NY, 2019, pp. 5744–5752.

[30] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res., 17 (2016), pp. 1–43.

[31] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Nat. Acad. Sci. USA, 113 (2016), pp. E7351–E7358, https://doi.org/10.1073/pnas.1614734113.

[32] A. C. Wilson, B. Recht, and M. I. Jordan, *A Lyapunov analysis of momentum methods in optimization*, preprint, arXiv:1611.02635, 2016, arXiv:1611.02635.
[33] L. Yang, R. Arora, V. braverman, and T. Zhao, *The physical systems behind optimization algorithms*, in Adv. Neural Inform. Process. Syst., S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., vol. 31, Curran Associates, Red Hook, NY, 2018, pp. 4372–4381.
[34] K. C. Zygalakis, *On the existence and the applications of modified equations for stochastic differential equations*, SIAM J. Sci. Comput., 33 (2011), pp. 102–130, https://doi.org/10.1137/090762336.