Journal of Machine Learning Research 24 (2023) 1-33 Submitted 11/21; Revised 11/22; Published 1/23

Discrete Variational Calculus for Accelerated Optimization

Cédric M. Campos CEDRIC.MCAMPOS@URJC.ES
Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica
Universidad Rey Juan Carlos
Calle Tulipán s/n, 28933 Móstoles, Spain

Alejandro Mahillo ALMAHILL@UNIZAR.ES
Departamento de Matemáticas
Instituto Universitario de Matemáticas y Aplicaciones
Universidad de Zaragoza
Calle de Pedro Cerbuna 12, 50009 Zaragoza, Spain

David Martín de Diego DAVID.MARTIN@ICMAT.ES
Instituto de Ciencias Matemáticas (CSIC-UAM-UC3M-UCM)
Calle Nicolás Cabrera 13-15, 28049 Madrid, Spain

Editor: Sayan Mukluk

Abstract

Many of the new developments in machine learning are connected with gradient-based optimization methods. Recently, these methods have been studied using a variational perspective (Betancourt et al., 2018). This has opened up the possibility of introducing variational and symplectic methods using geometric integration. In particular, in this paper, we introduce variational integrators (Marsden and West, 2001) which allow us to derive different methods for optimization. Using both Hamilton's and Lagrange-d'Alembert's principle, we derive two families of optimization methods in one-to-one correspondence that generalize Polyak's heavy ball (Polyak, 1964) and Nesterov's accelerated gradient (Nesterov, 1983), the second of which mimics the behavior of the latter reducing the oscillations of classical momentum methods. However, since the systems considered are explicitly time-dependent, the preservation of symplecticity of autonomous systems occurs here solely on the fibers. Several experiments exemplify the result.

Keywords: Polyak's heavy ball, Nesterov's accelerated gradient, momentum methods, variational integrators, Bregman Lagrangians

1. Introduction

Many of the literature on machine learning and data analysis is connected with gradient-based optimization methods (see Polak, 1997; Nesterov, 2018; and references therein). The computations often involve large data and parameter sets and then, not only the computational efficiency is a crucial point, but the optimization theory also plays a fundamental role. A typical optimization problem is:

$$\operatorname{argmin} f(x)\,, \quad x \in Q\,, \tag{1.1}$$

where we assume that Q is a convex set in $\mathbb{R}^n$ and f is a continuously differentiable convex function with Lipschitzian gradient. In this case, one of the most extended algorithms to achieve (1.1) is Nesterov's accelerated gradient (Nesterov, 1983; Su et al., 2016) which may take the following form:

$$y_{k+1} = x_k - \eta \nabla f(x_k)\,,$$
$$x_{k+1} = y_{k+1} + \frac{k}{k+3}(y_{k+1} - y_k)\,,$$

starting from an initial condition x_0 (see more details in Sections 2 and 7). An important observation was made by Su et al. (2016) showing that the continuous limit of Nesterov's method is a time-dependent second order differential equation. Moreover, Wibisono et al. (2016) show that this system of differential equations has a variational origin (see also Wibisono, 2016). In particular, they take as point of departure this variational approach that captures acceleration in continuous time considering a particular type of time-dependent Lagrangian functions, called Bregman Lagrangians (see Section 3).

In a recent paper, Betancourt et al. (2018) introduce symplectic (and presymplectic) integrators for the differential equations associated with accelerated optimizations methods (see Sanz-Serna and Calvo, 1994; Hairer et al., 2010; Blanes and Casas, 2016, for an introduction to symplectic integration). They use the Hamiltonian formalism since it is possible to extend the phase space to turn the system into a time-independent Hamiltonian system and apply there standard symplectic techniques (see Marthinsen and Owren, 2016; Celledoni et al., 2020). For recent improvements of this approach using adaptive Hamiltonian variational integrators, see Duruisseaux et al. (2021).

In our paper we set an alternative route: The idea is to use variational integrators adapted to an explicit time-dependent framework and external forces (see Marsden and West, 2001, and references therein) to derive a whole family of optimizations methods. The theory of discrete variational mechanics has reached maturity in recent years by combining results of differential geometry, classical mechanics, and numerical integration. Roughly speaking, the continuous Lagrangian $L\colon TQ \to \mathbb{R}$ is substituted by a discrete Lagrangian $L_d\colon Q \times Q \to \mathbb{R}$. Observe that, by replacing the standard velocity phase space TQ with $Q \times Q$, we are discretizing a velocity vector by two (in principle) close points. With the unique information of the discrete Lagrangian we can define the discrete action sum and, applying standard variational techniques, we derive a system of second order difference equations known as discrete Euler-Lagrange equations. The numerical order of the methods is obtained using variational error analysis (see Marsden and West, 2001; Patrick and Cuell, 2009). Moreover, it is possible to derive a discrete version of Noether's theorem relating the symmetries of the discrete Lagrangian with conserved quantities. The derived methods are automatically symplectic and, perhaps more importantly, easily adapted to other situations as, for instance, Lie group integrators, time-dependent Lagrangians, forced systems, optimal control theory, holonomic and nonholonomic mechanics, field theories, etc.

The Lagrangian functions described in Section 3, Bregman Lagrangians, are those explicitly time-dependent that typically arise on accelerated optimization. The geometry for time-dependent systems is different from symplectic geometry, in particular, the phase space is odd dimensional. In this case, an appropriate geometric framework is given by cosymplectic geometry (see Libermann, 1959; Cappelletti-Montano et al., 2013; and ref-

erences therein). In Section 4, we introduce the cosymplectic structure associated with a time-dependent Hamiltonian system (induced by a time-dependent Lagrangian) and also an interesting symplectic preservation property associated with the restriction of the Hamiltonian flow to the fibers of the projection onto the time variable (Theorem 1). Having in mind this geometrical framework, we introduce in Section 5 discrete variational mechanics for time-dependent Lagrangians with fixed time step (compare with Marsden and West, 2001, for variable time step). Moreover, we recover the symplectic character on fibers of the continuous Hamiltonian flow. We show the feasibility of constructing variational integrators using similar techniques to the developed for the autonomous case that, in some interesting cases, are in addition explicit and, consequently, reduce the computational cost. An example of such methods is the second-order difference equation

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

a type of momentum-descent method widely studied in the literature and whose origin goes back to Polyak (1964). Momentum methods allow to accelerate gradient descent by taking into account the "speed" achieved by the method at the last update. However, because of that speed, momentum methods can overpass the minimum. Nesterov's method tries to anticipate future information reducing the typical oscillations of classical momentum methods towards the minimum. In Section 6 we adapt our construction of variational integrators to add external forces using discrete Lagrange-d'Alembert's principle (see Marsden and West, 2001). Upon this machinery, we derive in Section 7 two families of momentum methods in mutual bijective correspondence, one of which corresponds to Nesterov's method (see Theorem 6). Finally, for Section 8, many methods and numerical simulations have been implemented in Julia v1.8.2. We optimize several test functions with our methodology and other methods that appeared recently in the literature. One of the test functions is reused afterwards for a machine learning example.

2. From Gradient Descent to Nesterov's Accelerated Gradient

In this section we give a historical perspective of Nesterov's accelerated gradient from gradient descent with a threefold objective: First, properly introduce the methods of interest and their properties, second, give an overall view of the elements to take under consideration, and, third, set some of the notation.

Although the first method that comes to mind to solve the optimization problem (1.1) is Newton-Raphson, the first "dynamical" one is due to Cauchy (1847). His method, known as Gradient Descent (GD), is the one-step method

$$x_{k+1} = x_k - \eta \nabla f(x_k) \,, \tag{2.1}$$

where η is the step size parameter or, as it is referred in the machine learning community, the learning rate. It is readily seen that this method is a simple discretization of the first order ODE

$$\dot{x} = -\nabla f(x) \,, \tag{2.2}$$

from which it takes its dynamical nature. What is perhaps not so readily seen is that, given an initial condition x_0, the trajectories obtained from both equations, x_k and $x(t)$, converge

to the argument of minima x^*. In particular, x_k converges linearly to x^* while the function values $f(x_k)$ do so to the global minimum $f(x^*)$ at a rate of $\mathcal{O}(1/k)$ (Polyak, 1964, 1987).

An initial improvement over GD was given by Polyak (1964): He introduced a novel two-step method, Polyak's Heavy Ball (PHB), also known as Classical Momentum (CM) after Sutskever et al. (2013). As it was originally presented, PHB/CM takes the form of the two-step method

$$x_{k+1} = x_k - \eta P(x_k) + \mu(x_k - x_{k-1})\,, \tag{2.3}$$

where P is a functional operator for which a root is sought and μ, η are "small" positive constants that condition the convergence of the method. In comparison with (2.1), (2.3) adds a new term, $x_k - x_{k-1}$, the momentum of the discrete motion that incorporates past information in an amount controlled by μ, the so called momentum coefficient. When P is conservative, that is, when $P = \nabla f$, Polyak showed that, although the method's trajectory still converges linearly as GD's, it does so faster than GD's, that is, with a smaller geometric ratio (Polyak, 1964, 1987).

The continuous analogue of (2.3) is the second order ODE

$$\ddot{x} + \nu(t)\dot{x} + \eta(t)P(x) = 0\,, \tag{2.4}$$

that turns out to be the equation of motion of a Lagrangian system when $P = \nabla f$ (Lemma 4). Then $x(t)$ traces the motion of a point mass in a well given by f. We therefore drop P and stick hereon with ∇f.

A further an crucial step towards improving GD (and PHB/CM) was given in 1983 by Nesterov, a former student of Polyak. He presented a new method, coined after him as Nesterov's Accelerated Gradient (NAG), similar to PHB/CM but with a slight change of unexpected consequences. A naive derivation from (2.3) is almost immediate: Introduce a new variable y_k in (2.3) so it can be easily rewritten as the equivalent method

$$y_{k+1} = x_k - \eta_k \nabla f(x_k)\,, \tag{2.5a}$$
$$x_{k+1} = y_{k+1} + \mu_k(x_k - x_{k-1})\,, \tag{2.5b}$$

where discrete-time dependence has been added to the coefficients μ, η for convenience. Replace the x's of the momentum term (right hand side of the second equation, Eq. 2.5b) by y's to get the new and non-equivalent method

$$\bar{y}_{k+1} = \bar{x}_k - \eta_k \nabla f(\bar{x}_k)\,, \tag{2.6a}$$
$$\bar{x}_{k+1} = \bar{y}_{k+1} + \mu_k(\bar{y}_{k+1} - \bar{y}_k)\,, \tag{2.6b}$$

where the bars are added to distinguish more easily both methods and underline that the sequences of points that they define are in fact different. This latter method (2.6) is NAG as it is usually presented. Nesterov showed in turn that his method accelerates the convergence rate of the function values down to $\mathcal{O}(1/k^2)$ (see Nesterov, 1983, 2018).

The original values of η_k, μ_k given by Nesterov are rather intricate, a simpler and commonly used version is

$$\bar{y}_{k+1} = \bar{x}_k - \eta \nabla f(\bar{x}_k)\,, \tag{2.7a}$$
$$\bar{x}_{k+1} = \bar{y}_{k+1} + \frac{k}{k+3}(\bar{y}_{k+1} - \bar{y}_k)\,, \tag{2.7b}$$

with $\eta > 0$ constant. As it is shown in Su et al. (2016), a continuous analogue of (2.7) is

$$\ddot{x} + \frac{3}{t}\dot{x} + \nabla f(x) = 0\,, \tag{2.8}$$

which is but a particular case of PHB/CM's continuous analogue (2.4). Besides that, Su et al. (2016) also show that the function values converge to the minimum at an inverse quadratic rate, that is, $f(x(t)) = f(x^*) + \mathcal{O}(1/t^2)$.

More generally (Remark 13), (2.6) is a natural discretization of a perturbed ODE of the form

$$\ddot{x} + \nu(t)\dot{x} + \eta(t)\nabla f(x) = \varepsilon F(x, \dot{x}, t)\,, \tag{2.9}$$

which also is the equation of motion of a Lagrangian system (Lemma 4). In fact, it is this variational origin that Wibisono et al. (2016) take as point of departure. Once a particular type of time-dependent Lagrangian functions is considered, a subfamily of the so called Bregman Lagrangians, the variational approach captures acceleration in continuous-time into the derived discrete schemes achieving, in this case, a function value convergence rate of $\mathcal{O}(t^{1-n})$ with $n \geq 3$ (see also Wibisono, 2016).

3. Bregman Lagrangians

A Bregman Lagrangian is roughly speaking a time-dependent mechanical Lagrangian whose kinetic part is close to be a metric. They are built upon Bregman divergences (Brègman, 1967), a particular case of divergence functions. Bregman Lagrangians allow to define variational problems whose solutions minimize an objective function at an exponential rate (Betancourt et al., 2018).

A **divergence function** over a manifold Q is a twice differentiable function $\mathcal{B}\colon Q \times Q \to \mathbb{R}_+$ such that for all $x, y \in Q$ we have:

- $\mathcal{B}(x, y) \geq 0$ and $\mathcal{B}(x, x) = 0$;
- $\partial_x \mathcal{B}(x, x) = \partial_y \mathcal{B}(x, x)$; and
- $\partial^2_{xy} \mathcal{B}(x, x)$ is negative-definite.

Divergence functions appear as pseudo-distances that are non-negative but are not, in general, symmetric. A typical divergence function over $Q = \mathbb{R}^n$ associated with a differentiable strictly convex function $\Phi\colon \mathbb{R}^n \to \mathbb{R}$ is the **Bregman divergence**:

$$\mathcal{B}_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle d\Phi(y), x - y\rangle\,.$$

Observe that it is the remainder of the first order Taylor expansion of Φ around y evaluated at x, a sort of Hessian metric.

Given a Bregman divergence over $\mathbb{R}^n$, let us consider the time-dependent kinetic energy

$$K(x, \dot{x}, t) = \mathcal{B}_\Phi(x + e^{-\alpha(t)}\dot{x}, x)\,,$$

and the time-dependent potential energy

$$U(x, t) = e^{\beta(t)} f(x)\,,$$

from which we define the **Bregman Lagrangian** $L\colon T\mathbb{R}^n \times T\mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} L(x,\dot{x},t) &= e^{\alpha(t)+\gamma(t)}\big(K(x,\dot{x},t) - U(x,t)\big) \\ &= e^{\alpha(t)+\gamma(t)}\left(\Phi(x+e^{-\alpha(t)}\dot{x}) - \Phi(x) - e^{-\alpha(t)}\langle d\Phi(x),\dot{x}\rangle - e^{\beta(t)}f(x)\right), \end{aligned}$$

where the time-dependent functions $\alpha(t), \beta(t), \gamma(t)$ are chosen to produce different algorithms. These functions verify what Wibisono et al. (2016) refer to as **ideal scaling conditions**, namely,

$$\dot{\gamma}(t) = e^{\alpha(t)} \qquad \text{and} \qquad \dot{\beta}(t) \leq e^{\alpha(t)}\,. \tag{3.1}$$

The first condition greatly simplifies several expressions that can be derived from the Bregman Lagrangian. For instance, when $\dot{\gamma}(t) = e^{\alpha(t)}$ is met, the associated Euler-Lagrange equations reduce to

$$\nabla^2\Phi\left(x+e^{-\alpha(t)}\dot{x}\right)\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(x+e^{-\alpha(t)}\dot{x}\right)\right] + e^{\alpha(t)+\beta(t)}\nabla f(x) = 0\,.$$

The second condition ensures convergence of the underlying trajectories to the minimum at a rate not slower than $\mathcal{O}(e^{-\beta(t)})$.

In the particular case where $\Phi(x) = \frac{1}{2}\|x\|^2$, for which $\mathcal{B}_\Phi(x,y) = \frac{1}{2}\|x-y\|^2$, the Bregman Lagrangian takes the simple form

$$L(x,\dot{x},t) = a(t)\tfrac{1}{2}\|\dot{x}\|^2 - b(t)f(x)\,, \tag{3.2}$$

with $a(t) = e^{\gamma(t)-\alpha(t)}$ and $b(t) = e^{\alpha(t)+\beta(t)+\gamma(t)}$.

4. Geometry of the Time-Dependent Lagrangian and Hamiltonian Formalisms

Since Bregman Lagrangians are time-dependent, in this section, we introduce some needed geometric ingredients about non-autonomous mechanics and highlight some of their main invariance properties (see Abraham and Marsden, 1978; Libermann and Marle, 1987; de León and R. Rodrigues, 1987).

Let Q be a manifold and TQ its tangent bundle. Coordinates (x^i) on Q induce coordinates $(x^i, \dot{x}^i)$ on TQ. Therefore we have natural coordinates $(x^i, \dot{x}^i, t)$ on $TQ \times \mathbb{R}$ which is the velocity phase space for time-dependent systems.

Given two instants (time values) $a, b \in \mathbb{R}$, with $a < b$, and corresponding positions $x_a, x_b \in Q$, consider the set of curves:

$$\mathcal{C}^2_{a,b} = \mathcal{C}^2([a,b], x_a, x_b) = \{\sigma\colon [a,b] \to Q \mid \sigma \in \mathcal{C}^2 \text{ with } \sigma(a) = x_a,\ \sigma(b) = x_b\}\,.$$

Given a time-dependent Lagrangian function $L\colon TQ \times \mathbb{R} \to \mathbb{R}$, define the action functional $\mathcal{J}_L\colon \mathcal{C}^2_{a,b} \to \mathbb{R}$

$$\mathcal{J}_L(\sigma) = \int_a^b L(\sigma'(t),t)\,\mathrm{d}t\,, \tag{4.1}$$

where $\sigma'\colon [a,b] \to TQ$.

Using variational calculus, the critical points of $\mathcal{J}_L$ are locally characterized by the solutions of the **Euler-Lagrange equations**:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}^i}\right) - \frac{\partial L}{\partial x^i} = 0\,, \quad 1 \leq i \leq n = \dim Q\,. \tag{4.2}$$

For time-dependent Lagrangians it is possible to check that the energy $E_L \colon TQ \times \mathbb{R} \to \mathbb{R}$,

$$E_L = \Delta L - L = \dot{x}^i \frac{\partial L}{\partial \dot{x}^i} - L\,,$$

where Δ is the Liouville vector field on TQ (Libermann and Marle, 1987), is not, in general, preserved since

$$\frac{\mathrm{d}E_L}{\mathrm{d}t} = \frac{\partial L}{\partial t}\,.$$

We now pass to the Hamiltonian formalism using the **Legendre transformation**

$$\mathcal{F}L \colon TQ \times \mathbb{R} \longrightarrow T^*Q \times \mathbb{R}\,,$$

where T^*Q is the cotangent bundle of Q whose natural coordinates are (x^i, p_i). The Legendre transformation is locally given by

$$\mathcal{F}L(x^i, \dot{x}^i, t) = \left(x^i, \frac{\partial L}{\partial \dot{x}^i}, t\right)\,.$$

We assume that the Legendre transformation is a diffeomorphism (that is, the Lagrangian is hyperregular) and define the Hamiltonian function $H \colon T^*Q \times \mathbb{R} \to \mathbb{R}$ by

$$H = E_L \circ (\mathcal{F}L)^{-1}\,,$$

which induces the cosymplectic structure $(\Omega_H, \eta_{\mathbb{R}})$ on $T^*Q \times \mathbb{R}$ with

$$\mathrm{d}t := \mathrm{pr}_2^*\,\mathrm{d}t\,, \qquad \Omega_H = -\mathrm{d}(\mathrm{pr}_1^*\,\theta_Q - H\mathrm{d}t) = \Omega_Q + \mathrm{d}H \wedge \mathrm{d}t\,,$$

where pr_i, $i = 1, 2$, are the projections to each Cartesian factor and θ_Q denotes the Liouville 1-form on T^*Q (Abraham and Marsden, 1978), given in induced coordinates by $\theta_Q = p_i\,\mathrm{d}x^i$. We also denote by $\Omega_Q = -\mathrm{d}\,\mathrm{pr}_1^*\,\theta_Q$ the pullback of the canonical symplectic 2-form $\omega_Q = -\mathrm{d}\theta_Q$ on T^*Q. In coordinates, $\Omega_Q = \mathrm{d}x^i \wedge \mathrm{d}p_i$. (Observe that now Ω_Q is presymplectic since $\ker \Omega_Q = \mathrm{span}\{\partial/\partial t\}$.) Therefore in induced coordinates (x^i, p_i, t):

$$\Omega_H = \mathrm{d}x^i \wedge \mathrm{d}p_i + \mathrm{d}H \wedge \mathrm{d}t\,, \qquad \eta_{\mathbb{R}} = \mathrm{d}t\,.$$

We define the **evolution vector field** $E_H \in \mathfrak{X}(T^*Q \times \mathbb{R})$ by

$$i_{E_H}\Omega_H = 0\,, \qquad i_{E_H}\mathrm{d}t = 1\,. \tag{4.3}$$

In local coordinates the evolution vector field is:

$$E_H = \frac{\partial}{\partial t} + \frac{\partial H}{\partial p_i}\frac{\partial}{\partial x^i} - \frac{\partial H}{\partial x^i}\frac{\partial}{\partial p_i}\,.$$

The integral curves of E_H are given by:

$$\dot{t} = 1\,, \qquad \dot{x}^i = \frac{\partial H}{\partial p_i}\,, \qquad \dot{p}_i = -\frac{\partial H}{\partial x^i}\,. \tag{4.4}$$

The integral curves of E_H are precisely the curves of the form $t \mapsto \mathcal{F}L(\sigma'(t), t)$ where $\sigma\colon I \to Q$ is a solution of the Euler-Lagrange equations for $L\colon TQ \times \mathbb{R} \to \mathbb{R}$.

From Equation (4.3) we deduce that the flow of E_H verifies the following preservation properties

$$\mathcal{L}_{E_H}\Omega_H = \mathcal{L}_{E_H}(\Omega_Q + \mathrm{d}H \wedge \mathrm{d}t) = 0\,, \qquad \mathcal{L}_{E_H}\mathrm{d}t = 0\,. \tag{4.5}$$

Denote by $\Psi_s\colon \mathcal{U} \subset T^*Q \times \mathbb{R} \longrightarrow T^*Q \times \mathbb{R}$ the flow of the evolution vector field E_H, where $\mathcal{U}$ is an open subset of $T^*Q \times \mathbb{R}$. Observe that

$$\Psi_s(\alpha_q, t) = (\Psi_{t,s}(\alpha_q), t + s)\,,\ \alpha_q \in T^*_qQ\,,$$

where $\Psi_{t,s}(\alpha_q) = \mathrm{pr}_1(\Psi_s(\alpha_q, t))$. Therefore from the flow of E_H we induce a map

$$\Psi_{t,s}\colon \mathcal{U}_t \subseteq T^*Q \to T^*Q\,,$$

where $\mathcal{U}_t = \{\alpha_q \in T^*Q \mid (\alpha_q, t) \in \mathcal{U}\}$. Observe that if we know $\Psi_{t,s}$ for all t, we can recover the flow Ψ_s of E_H.

From Equations (4.5) we deduce that

$$\Psi_s^*(\Omega_Q + \mathrm{d}H \wedge \mathrm{d}t) = \Omega_Q + \mathrm{d}H \wedge \mathrm{d}t\,, \qquad \Psi_s^*(\mathrm{d}t) = \mathrm{d}t\,. \tag{4.6}$$

The following theorem relates the preservation properties (4.6) with the symplecticity of the map family $\{\Psi_{t,s}\colon T^*Q \to T^*Q\}$.

Theorem 1 *We have that $\Psi_{t,s}\colon \mathcal{U}_t \subseteq T^*Q \to T^*Q$ is a symplectomorphism, that is, $\Psi_{t,s}^*\omega_Q = \omega_Q$.*

Proof First, observe that any vector $Y_{(\alpha_q,t)} \in T_{(\alpha_q,t)}(T^*Q \times \mathbb{R})$ admits a unique decomposition:

$$Y_{(\alpha_q,t)} = Y_{\alpha_q}(t) + Y_t(\alpha_q)\,,$$

where $Y_{\alpha_q}(t) \in T_{\alpha_q}T^*Q$ and $Y_t(\alpha_q) \in T_t\mathbb{R}$. Moreover, we have that $\langle \mathrm{d}t, Y_{\alpha_q}(t)\rangle = 0$.

Therefore, if we restrict ourselves to vectors tangent to the pr_2-fibers $T_{(\alpha_q,t)}\,\mathrm{pr}_2^{-1}(t) = V_{(\alpha_q,t)}\,\mathrm{pr}_2$ then we have the decomposition

$$Y_{(\alpha_q,t)} = Y_{\alpha_q}(t) + 0_t = Y_{\alpha_q}(t) \in V_{(\alpha_q,t)}\,\mathrm{pr}_2 \equiv T_{\alpha_q}T^*Q\,.$$

From the second preservation property given in (4.5) we deduce that

$$0 = \langle (\mathrm{d}t)_{(\alpha_q,t)}, Y_{\alpha_q}(t)\rangle = \langle (\Psi_s^*\mathrm{d}t)_{(\alpha_q,t)}, Y_{\alpha_q}(t)\rangle = \langle (\mathrm{d}t)_{\Psi_s(\alpha_q,t)}, T\Psi_s(Y_{\alpha_q}(t))\rangle\,.$$

Therefore $T\Psi_s(Y_{\alpha_q}(t)) \in V_{(\Psi_{s,t}(\alpha_q),t+s)}pr_2 \equiv T_{\Psi_{t,s}(\alpha_q)}T^*Q$ and

$$T\Psi_s(Y_{\alpha_q}(t)) = T\Psi_{t,s}(Y_{\alpha_q}(t)) + 0_{t+s} \equiv T\Psi_{t,s}(Y_{\alpha_q}(t))\,.$$

Using the first identity in (4.5) we deduce that

$$\begin{aligned}(\omega_Q)_{\alpha_q}(Y_{\alpha_q}(t),\tilde{Y}_{\alpha_q}(t)) &= (\Omega_Q + \mathrm{d}H\wedge\mathrm{d}t)_{(\alpha_q,t)}(Y_{\alpha_q}(t),\tilde{Y}_{\alpha_q}(t))\\ &= (\Omega_Q + \mathrm{d}H\wedge\mathrm{d}t)_{(\Psi_{s,t}(\alpha_q),t+s)}(T\Psi_s(Y_{\alpha_q}(t)),T\Psi_s(\tilde{Y}_{\alpha_q}(t)))\\ &= (\omega_Q)_{\Psi_{s,t}(\alpha_q)}(T\Psi_{t,s}(Y_{\alpha_q}(t)),T\Psi_{t,s}(\tilde{Y}_{\alpha_q}(t)))\end{aligned}$$

where $Y_{\alpha_q}(t),\tilde{Y}_{\alpha_q}(t)\in T_{\alpha_q}T^*Q\equiv T_{(\alpha_q,t)}\,\mathrm{pr}_2^{-1}(t)$. We conclude that $\Psi_{t,s}^*\omega_Q=\omega_Q$. ■

5. Discrete Variational Methods for Time-Dependent Lagrangian Systems

Consider the set of discrete paths (or sequences) on Q for a fixed number of steps $N\in\mathbb{N}$, that is, the set

$$\mathcal{C}_d(Q)=\{x_d\colon\{0,1,\dots,N\}\to Q\}=Q\times\overset{(N+1)}{\cdots}\times Q\,.$$

Then an appropriate discrete interpretation of the velocities are pairs in $Q\times Q$, a discrete version of TQ.

A **discrete time-dependent Lagrangian** is a family of maps

$$L_d^k\colon Q\times Q\to\mathbb{R}\,,\qquad k\in\mathbb{N}\,,$$

for which we define the **discrete action map** on the space of sequences as

$$S_d(x_d)=\sum_{k=0}^{N-1}L_d^k(x_k,x_{k+1})\,,\quad x_d\in\mathcal{C}_d(Q)\,.$$

If we consider variations of x_d with fixed end points x_0 and x_N and extremize S_d over $x_1,\dots,x_{N-1}$, we obtain the **discrete Euler-Lagrange equations** (DEL for short)

$$\partial_{x_k}S_d(x_d)=D_1L_d^k(x_k,x_{k+1})+D_2L_d^{k-1}(x_{k-1},x_k)=0\,,\quad\text{for all } k=1,\dots,N-1,$$

where D_1 and D_2 denote the partial derivatives with respect to the first and second components, respectively.

If, for all k, L_d^k is regular, that is, the matrix

$$D_{12}L_d^k=\left(\frac{\partial^2L_d^k}{\partial x_k\partial x_{k+1}}\right)$$

is non-singular, then we locally obtain a well defined family of discrete Lagrangian maps:

$$\begin{array}{rccc}F_{k,k+1}\colon & Q\times Q & \longrightarrow & Q\times Q\\ & (x_k,x_{k+1}) & \longmapsto & (x_{k+1},x_{k+2}(x_k,x_{k+1},k))\,.\end{array}$$

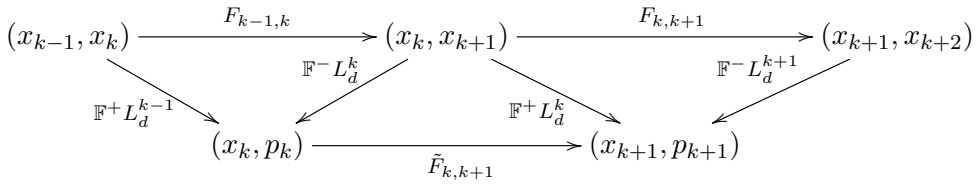
where the value of x_{k+2} is determined in terms of x_k, x_{k+1} and k. In this setting, we can define two discrete Legendre transformations associated with L_d^k, $\mathbb{F}^\pm L_d^k\colon Q\times Q\to T^*Q$, by the expressions

$$\begin{aligned}\mathbb{F}^+L_d^k\colon(x_k,x_{k+1})&\longmapsto(x_{k+1},D_2L_d^k(x_k,x_{k+1}))\,,\\ \mathbb{F}^-L_d^k\colon(x_k,x_{k+1})&\longmapsto(x_k,-D_1L_d^k(x_k,x_{k+1}))\,.\end{aligned}$$

We can also define the evolution of the discrete system on the Hamiltonian side, $\tilde{F}_{k,k+1}\colon T^*Q \to T^*Q$, by any of the formulas:

$$\tilde{F}_{k,k+1} = \mathbb{F}^+L_d^k \circ (\mathbb{F}^-L_d^k)^{-1} = \mathbb{F}^+L_d^k \circ F_{k-1,k} \circ (\mathbb{F}^+L_d^{k-1})^{-1} = \mathbb{F}^-L_d^{k+1} \circ F_{k,k+1} \circ (\mathbb{F}^-L_d^k)^{-1}\,,$$

thanks to the commutativity of the following diagram.



Proposition 2 *The discrete Hamiltonian map $\tilde{F}_{k,k+1}\colon (T^*Q, \omega_Q) \longrightarrow (T^*Q, \omega_Q)$ is a* ***symplectic transformation****, that is,*

$$(\tilde{F}_{k,k+1})^*\omega_Q = \omega_Q\,.$$

Proof Using similar arguments to the autonomous case (Marsden and West, 2001), we deduce that

$$(\mathbb{F}^+L_d^k)^*\omega_Q = (\mathbb{F}^-L_d^k)^*\omega_Q\,.$$

From the definition of $\tilde{F}_{k,k+1} : T^*Q \to T^*Q$ we deduce that

$$(\tilde{F}_{k,k+1})^*\omega_Q = (\mathbb{F}^+L_d^k \circ (\mathbb{F}^-L_d^k)^{-1})^*\omega_Q = ((\mathbb{F}^-L_d^k)^*)^{-1}((\mathbb{F}^+L_d^k)^*\omega_Q) = \omega_Q\,.$$

■

Given the map $\tilde{F}_{k,k+1}(q_k, p_k) = (q_{k+1}, p_{k+1})$, we immediately have the map

$$(x_k, p_k, kh) \mapsto (x_{k+1}, p_{k+1}, (k+1)h)$$

on $T^*Q \times \mathbb{R}$ where we now give explicit information of the evolution of discrete time.

Let see the relation of these discrete maps $F_{k,k+1}\colon Q \times Q \to Q \times Q$ and $\tilde{F}_{k,k+1}\colon T^*Q \to T^*Q$ with the Euler-Lagrange equations and Hamilton equations of a time-dependent Lagrangian system. Given a regular Lagrangian function $L\colon TQ \times \mathbb{R} \to \mathbb{R}$ and a sufficiently small time step $h > 0$, we are going to define an h-and-k-dependent family of discrete Lagrangian functions $L_{d,h}^k\colon Q \times Q \to \mathbb{R}$ as an infinitesimal approximation to the continuous action $\mathcal{J}_L$ defined in expression (4.1). As intermediate step, we first consider the **exact time-dependent discrete Lagrangian** associated with a regular Lagrangian L which is given by the expression

$$L_{d,h}^{k,E}(x_0, x_1) = \frac{1}{h}\int_{kh}^{(k+1)h} L(x_{0,1}(t), \dot{x}_{0,1}(t), t)\,\mathrm{d}t\,,$$

where $x_{0,1}(t)$ is the unique solution of the Euler-Lagrange equations for L with $x_{0,1}(kh) = x_0$ and $x_{0,1}((k+1)h) = x_1$, (see Hartman, 2002; Marrero et al., 2016). Then for a sufficiently

small h, the solutions of the DEL for $L_{d,h}^{k,E}$ lie on the solutions of the Euler-Lagrange equations for L (Theorem 1.6.4, Marsden and West, 2001).

In practice, $L_{d,h}^{k,E}(x_0, x_1)$ will not be available, therefore we take an approximation,

$$L_{d,h}^{k}(x_0, x_1) \approx L_{d,h}^{k,E}(x_0, x_1)\,,$$

using some quadrature rule. Then, as we have seen, the scheme derived from the DEL will be geometric integrators for Equations (4.4) preserving the symplectic form in the sense of Theorem 1 (see Patrick and Cuell, 2009).

Remark 3 *As we have commented in the Introduction, one of the main advantages of the proposed approach is the possibility to use other options to derive different numerical methods for optimization by only discretizing a unique function, the action functional. Of course, there are many different ways to do it (Marsden and West, 2001). For instance, we can combine several discrete Lagrangians together to obtain a new discrete Lagrangian with higher order (composition methods) or similarly obtaining splitting methods (Campos and Sanz-Serna, 2017). Also, we can easily derive symplectic partitioned Runge-Kutta methods or symplectic Garlekin methods using polynomial approximations of the trajectories and a numerical quadrature to approximate the action integral (Campos, 2014). Moreover, it is possible to adapt the variational integrators to a non-euclidean setting using appropriate retraction maps.*

6. Discretization of Lagrangian Systems with Forces

Our intention here is to continue looking for numerical approximations to the time-dependent Euler-Lagrange equations but considering additionally an external force that decreases jointly with the time-step parameter h. With it, we will obtain a whole family of algorithms whose behavior resembles that of the Nesterov method. Fortunately, discrete mechanics is also adapted to the case of external forces (see Marsden and West, 2001). To this end, in addition to a time-dependent Lagrangian function $L\colon TQ \times \mathbb{R} \to \mathbb{R}$, we have an external force given by a fiber preserving map $F\colon TQ \times \mathbb{R} \to T^*Q$ given locally by

$$F(x, \dot{x}, t) = (x^i, F_i(x, \dot{x}, t))\,.$$

Given the force f, we derive the equations of motion of the forced system modifying Hamilton's principle to the **Lagrange-d'Alembert principle**, which seeks curves $\sigma \in \mathcal{C}_{a,b}^2$ satisfying

$$\delta \int_a^b L(\sigma'(t), t)\,\mathrm{d}t + \int_a^b F(\sigma'(t), t)\delta\sigma(t)\,\mathrm{d}t = 0\,, \tag{6.1}$$

for all $\delta\sigma \in T_\sigma \mathcal{C}_{a,b}^2$. Using integration by parts, we derive the forced Euler-Lagrange equations, which have the following coordinate expression:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{x}^i}\right) - \frac{\partial L}{\partial x^i} = F_i\,.$$

To discretize these equations, we consider, as before, a family of Lagrangian functions $L_d^k\colon Q \times Q \to \mathbb{R}$ and two discrete forces $(F_d^k)^{\pm}\colon Q \times Q \to T^*Q$, which are fiber preserving

in the sense that $\pi_Q \circ (F_d^k)^{\pm} = \mathrm{pr}_{\pm}$, where $\mathrm{pr}_{\pm}(x_-, x_+) = x_{\pm}$. Combing both forces, we obtain $F_d^k \colon Q \times Q \to T^*(Q \times Q)$ by

$$\langle F_d^k(x_k, x_{k+1}), (\delta x_k, \delta x_{k+1}) \rangle = (F_d^k)^-(x_k, x_{k+1})\delta x_k + (F_d^k)^+(x_k, x_{k+1})\delta x_{k+1}\,.$$

As in (6.1), we have a discrete version of the Lagrange-d'Alembert principle for the discrete forced system given by L_d^k and F_d^k:

$$\delta \sum_{k=0}^{N-1} L_d^k(x_k, x_{k+1}) + \sum_{k=0}^{N-1} \langle F_d^k(x_k, x_{k+1}), (\delta x_k, \delta x_{k+1}) \rangle = 0\,,$$

for all variations $\{\delta x_k\}_{k=0}^N$ vanishing at the endpoints, that is, $\delta x_0 = \delta x_N = 0$. This is equivalent to the forced discrete Euler-Lagrange equations:

$$D_1 L_d^k(x_k, x_{k+1}) + D_2 L_d^{k-1}(x_{k-1}, x_k) + (F_d^k)^-(x_k, x_{k+1}) + (F_d^{k-1})^+(x_{k-1}, x_k) = 0\,, \quad (6.2)$$

for all $k = 1, \ldots, N-1$.

7. The Variational Derivation of PHB/CM and NAG

As seen in Section 2, NAG can be derived naively from PHB/CM. Besides, under suitable conditions on μ, η and the starting points, both methods converge to a minimum of f, the latter, NAG, doing so faster (Polyak, 1964; Nesterov, 1983). Questions arise: What makes NAG faster than PHB/CM? Can this be exploited to obtain even faster methods? Can it be generalized? Questions that boil down to how NAG is fundamentally derived from PHB/CM.

To begin with, note that the NAG equations (2.6) can be rewritten only in terms of the x's, as in (2.3), or only in terms of the y's, yielding the equations

$$\Delta \bar{x}_k = \mu_k \quad \Delta \bar{x}_{k-1} - \eta_k \nabla f(\bar{x}_k) - \mu_k \big(\eta_k \nabla f(\bar{x}_k) - \eta_{k-1} \nabla f(\bar{x}_{k-1})\big)\,, \quad (7.1a)$$

$$\Delta \bar{y}_k = \mu_{k-1} \Delta \bar{y}_{k-1} - \eta_k \nabla f(\bar{y}_k + \mu_{k-1} \Delta \bar{y}_{k-1})\,. \quad (7.1b)$$

The first, Eq. (7.1a), when compared with (2.3) shows an extra term,

$$\mu_k \big(\eta_k \nabla f(\bar{x}_k) - \eta_{k-1} \nabla f(\bar{x}_{k-1})\big)\,,$$

that in fact points to the very origin of the method: an additional forcing term. The second, Eq. (7.1b), compared again with (2.3) shows that the y-trajectory is obtained almost as if it was computed by PHB/CM but evaluating ∇f at a "future" point, $\bar{y}_k + \mu_{k-1} \Delta \bar{y}_{k-1}$, which "informs better" the method on how to advance towards the minimum.

Besides the convergence towards the minimum, it can be shown that both methods are a time discretization of the second order differential equation (Su et al., 2016)

$$\ddot{x} + \nu(t)\dot{x} + \eta(t) \nabla f(x) = 0\,, \quad (7.2)$$

a well known fact in the literature, equation that furthermore is variational in general (see Lemma 4). A fact that is not so well known is that NAG discretizes better the equation

when including a force term proportional to the underlying time step and, moreover, it can be derived, as well as PHB/CM, from a variational approach, in the geometric integration sense (see Sections 5 and 6).

We first give a rather simple and direct result whose purpose is to establish properly the continuous setting over which the discretizations will be built and from which the methods can be derived.

Lemma 4 *Given a vector field $P\colon \mathbb{R}^n \to \mathbb{R}^n$, consider the second order differential equation*

$$\ddot{x} + \nu(t)\dot{x} + \eta(t)P(x) = \varepsilon \frac{\mathrm{d}}{\mathrm{d}t}\Big[\eta(t)P(x)\Big]\,, \tag{7.3}$$

where $\nu, \eta\colon \mathbb{R}_+ \to \mathbb{R}$ are continuous time-dependent real valued functions and where $\varepsilon \in \mathbb{R}$ is a constant. If P is conservative, that is, if $P = \nabla f$, then (7.3) *corresponds to the equation of motion of the (forced) time-dependent Lagrangian system:*

$$L(x, \dot{x}, t) = a(t)\tfrac{1}{2}\|\dot{x}\|^2 - b(t)f(x)\,, \tag{7.4a}$$

$$F(x, \dot{x}, t) = \varepsilon a(t)\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{b(t)}{a(t)}P(x)\right]\,, \tag{7.4b}$$

in which f is the field's potential and where

$$a(t) = \exp(\textstyle\int_0^t \nu(s)\mathrm{d}s)\,, \quad \textit{and} \quad b(t) = a(t)\eta(t)\,, \tag{7.5}$$

for $t \geq 0$.

Proof Assume P is conservative and let f denote its potential. Then the Euler-Lagrange equation for (7.4) is

$$a(t)\ddot{x} + a'(t)\dot{x} + b(t)P(x) = \varepsilon a(t)\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{b(t)}{a(t)}P(x)\right]\,. \tag{7.6}$$

Dividing by $a(t)$, and taking into account that, from (7.5) we have that

$$\nu(t) = \frac{a'(t)}{a(t)} \quad \text{and} \quad \eta(t) = \frac{b(t)}{a(t)}\,, \tag{7.7}$$

we obtain (7.3). ∎

Remark 5 *P being conservative is not a necessary condition for* (7.3) *to be derived from the Lagrange-d'Alembert principle. In order to be variational, Equation* (7.6) *requires a Lagrangian of the form*

$$L(x, \dot{x}, t) = a(t)\tfrac{1}{2}\dot{x}^2 + \langle c(x,t), \dot{x}\rangle + d(x,t)$$

for some unknown functions c, d, which implies

$$\frac{\partial d}{\partial x} = \frac{\partial c}{\partial t} + b(t)P(x)\,.$$

A vector field P satisfying this last relation need not be conservative.

Next, the result that links the previous continuous equation (7.3) with PHB/CM and NAG, showing, in particular, that NAG is a forced version of PHB/CM, between which the transition is immediate.

Theorem 6 *Given a real valued function $f\colon \mathbb{R}^n \to \mathbb{R}$ and a vector field $P\colon \mathbb{R}^n \to \mathbb{R}^n$, consider the time-dependent discrete Lagrangian and forces*

$$L_d^k(z_0, z_1) = a_k \tfrac{1}{2} \|z_1 - z_0\|^2 - b_k^- f(z_0) - b_{k+1}^+ f(z_1)\,, \tag{7.8a}$$

$$(F_d^k)^-(z_0, z_1) = -\tfrac{a_{k-1}}{a_k}(b_k^- + b_k^+)P(z_0)\,, \ \textit{and} \tag{7.8b}$$

$$(F_d^k)^+(z_0, z_1) = \qquad (b_k^- + b_k^+)P(z_0)\,. \tag{7.8c}$$

where $\{a_k\}_{k\geq 0}$, $\{b_k^-\}_{k\geq 0}$, $\{b_k^+\}_{k\geq 0}$, are arbitrary sequences of real numbers. If f is regular enough, so that $P = \nabla f$, and a_k is never null, then the free and forced equations of motion for L_d^k and $(L_d^k, (F_d^k)^-, (F_d^k)^+)$ are, respectively, equivalent to the following recursive schemes

$$y_{k+1} = x_k - \eta_k P(x_k)\,, \qquad \bar{y}_{k+1} = \bar{x}_k - \eta_k P(\bar{x}_k)\,, \tag{7.9a}$$

$$x_{k+1} = y_{k+1} + \mu_k(x_k - x_{k-1})\,, \qquad \bar{x}_{k+1} = \bar{y}_{k+1} + \mu_k(\bar{y}_{k+1} - \bar{y}_k)\,, \tag{7.9b}$$

where

$$\mu_{k+1} = \frac{a_k}{a_{k+1}} \quad \textit{and} \quad \eta_k = \frac{b_k^- + b_k^+}{a_k}\,, \tag{7.10}$$

for $k \geq 0$.

Conversely, given a vector field $P\colon \mathbb{R}^n \to \mathbb{R}^n$ and two arbitrary sequences of real numbers $\{\mu_{k+1}\}_{k\geq 0}$ and $\{\eta_k\}_{k\geq 0}$, consider the sequences of pairs of points given in equations (7.9a)-(7.9b)*. If P is conservative and μ_{k+1} is never null, then both schemes are variational. Moreover, they are equivalent to the equations of motions for the free and forced time-dependent discrete Lagrangian systems given in* (7.8a)-(7.8c)*, for which f is the field's potential and*

$$a_0 = 1,\ a_{k+1} = a_k/\mu_{k+1},\ \forall k \geq 0\,, \qquad b_k^{\pm} = \tfrac{1}{2} a_k \eta_k,\ \forall k \geq 0\,. \tag{7.11}$$

Proof Partial differentiation of the Lagrangian gives

$$\begin{aligned} D_1 L_d^k(z_0, z_1) &= -a_k \Delta z_0 - b_k^- \nabla f(z_0)\,, \\ D_2 L_d^k(z_0, z_1) &= \quad a_k \Delta z_0 - b_{k+1}^+ \nabla f(z_1)\,, \end{aligned}$$

from where it is readily seen that the DEL equations with forces (6.2) are

$$\begin{aligned} -a_{k+1}\Delta z_1 + a_k \Delta z_0 - (b_{k+1}^- + b_{k+1}^+)\nabla f(z_1) = \\ = \frac{a_k}{a_{k+1}}(b_{k+1}^- + b_{k+1}^+)\nabla f(z_1) - (b_k^- + b_k^+)\nabla f(z_0)\,, \end{aligned}$$

where the RHS is null for the non-forced DEL equations. Dividing by $-a_{k+1}$ and using the relations (7.10), we get

$$\Delta z_1 - \mu_{k+1}\Delta z_0 + \eta_{k+1}\nabla f(z_1) = -\mu_{k+1}\left(\eta_{k+1}\nabla f(z_1) - \eta_k \nabla f(z_0)\right)\,, \tag{7.12}$$

where, again, the right hand side is null for the non-forced case. Replacing z_i with x_{k+i} in the non-forced case, and with $\bar{x}_{k+i}$ in the forced one, taking into account that $P = \nabla f$, and using the equations in (7.9a), we obtain those in (7.9b).

The converse is immediate. ■

Now remarks are in order that will summarize some points that have been mentioned earlier and underline others that haven not been yet.

Remark 7 (One-to-one correspondence) *Note that both equations in* (7.9a) *are formally the same, whereas in* (7.9b) *there is a difference in the last term: While PHB/CM uses x's, NAG considers y's. This is a slight change that nonetheless defines different schemes and in which the forcing term is hidden. This result not only shows that NAG is a forced version of PHB/CM, but that the schemes are in a natural bijective correspondence.*

Remark 8 (Strategies) *A specific method is defined by a strategy: a pair of coefficients (μ, η). This work and especially Theorem 6 focus on variational strategies, those pairs $\mu, \eta\colon \mathbb{N} \to \mathbb{R}$ that can be derived variationally from a time-dependent Lagrangian of the form* (3.2)*. Being the original strategy of Polyak's method constant, it falls within this class of variational momentum-descent methods, whereas the original method by Nesterov is non-variational and belongs to a broader class of generalized momentum-descent methods where the coefficients might depend on the objective function itself (confer with Polyak, 1964; Nesterov, 1983; see also Eq. 8.12).*

Remark 9 (Initial conditions) *Usually the initial condition, (x_0, v_0) or (x_0, p_0) in phase space, or (x_0, x_1) in configuration space, is crucial for the proper simulation of the dynamical system. Here, however, the dynamics are a tool and generally an initial condition so that $x_1 = x_0$ or $\bar{y}_0 = \bar{x}_0$, where $x_0 = \bar{x}_0$ is close to the minimum, will suffice, which corresponds to sticking the ball to the bowl's wall and leave it roll.*

Remark 10 (Natural trajectory) *From the schemes' definitions, the sequences $\{x_k\}_{k=0}^{\infty}$ and $\{\bar{x}_k\}_{k=0}^{\infty}$ are the natural (on track) dynamical trajectories towards the minimum of f, while $\{y_k\}_{k=1}^{\infty}$ and $\{\bar{y}_k\}_{k=1}^{\infty}$ are off road marks that limit these trajectories like slalom flags. The latter are, however, asymptotically close to the former and, hence, to the minimum.*

Remark 11 (Discrete flow) *If one wants to compare the trajectories obtained from both methods, perhaps Equations* (7.9) *are better suited. If one is solely interested in a simple implementation to compute the minimum, then Equations* (2.3) *and* (7.1b) *are a good alternative since they can easily be rewritten to give discrete flow updates in the form "momentum first, then position" as in Sutskever et al. (2013):*

$$\Delta x_k = \mu_{k-1}\Delta x_{k-1} - \eta_k \nabla f(x_k)\,, \quad \Delta \bar{y}_k = \mu_{k-1}\Delta \bar{y}_{k-1} - \eta_k \nabla f(\bar{y}_k + \mu_{k-1}\Delta \bar{y}_{k-1})\,, \tag{7.13a}$$

$$x_{k+1} = x_k + \Delta x_k\,, \qquad \bar{y}_{k+1} = \bar{y}_k + \Delta \bar{y}_k\,, \tag{7.13b}$$

where both methods should be initialized with $x_0 = \bar{y}_0$ and $\Delta x_{-1} = \Delta \bar{y}_{-1} = 0$. Both approaches, Equations (7.9) *and Equations* (7.13)*, have been considered for the simulations of Section 8.*

Remark 12 (Force approximation) *The action induced by the discrete forces in* (7.8) *is a second order approximation to the action induced by the continuous force* (7.4b)*. Indeed, given continuous coefficients $a(t), b(t)$, define $a_k = a(kh)/h^2$ and $b_k^{\pm}$ so that $b_k^- + b_k^+ = b(kh)$, and similarly for a continuous path $x(t)$ and a variation $\delta x(t)$ of it. Then*

$$
\begin{aligned}
h \sum_{k=0}^{N-1} & F_d^k(x_k, x_{k+1}) \cdot (\delta x_k, \delta x_{k+1}) = \\
&= -h \sum_{k=1}^{N-1} \frac{a_{k-1}}{a_k} (b_k^- + b_k^+) \nabla f(x_k) \cdot \delta x_k + h \sum_{k=1}^{N-1} (b_{k-1}^- + b_{k-1}^+) \nabla f(x_{k-1}) \cdot \delta x_k \\
&= \sum_{k=1}^{N-1} \int_{t_k - \frac{h}{2}}^{t_k + \frac{h}{2}} \left(-\frac{a(t-h)}{a(t)} b(t) \nabla f(x(t)) + b(t-h) \nabla f(x(t-h)) \right) \cdot \delta x(t) \, \mathrm{d}t + \mathcal{O}(h^2) \\
&= \int_{\frac{h}{2}}^{T - \frac{h}{2}} -h \, a(t-h) \frac{\mathrm{d}}{\mathrm{d}t} \left[\frac{b(t)}{a(t)} \nabla f(x(t)) \right] \cdot \delta x(t) \, \mathrm{d}t + \mathcal{O}(h^2) \\
&= \int_0^T -h \, a(t) \frac{\mathrm{d}}{\mathrm{d}t} \left[\frac{b(t)}{a(t)} \nabla f(x(t)) \right] \cdot \delta x(t) \, \mathrm{d}t + \mathcal{O}(h^2)
\end{aligned}
$$

where we have considered a mid point quadrature rule to establish the second equality, the rest follows.

Remark 13 (Force evanescence) *As remarked, NAG can be viewed as an approximation to a forced continuous Lagrangian system, where the force is proportional to the time step that is used* a posteriori *for the discretization. Although its contribution is non-null along the whole trajectory, it however vanishes not only locally, when h decreases, but also asymptotically in time, when t or k increase.*

8. Simulations

Numerical experiments are performed considering different elements, namely:

- The time-dependent coefficients that appear in the Lagrangian, $a(t)$ and $b(t)$, for the simple case (3.2);
- The discretization scheme used to approximate the Lagrangian action; and obviously,
- The objective function to be minimized, $f(x)$.

8.1 Lagrangian Coefficients

We consider three different Lagrangians or, more precisely, three different pairs of time-dependent coefficients $(a(t), b(t))$ given, as in (3.2), by time-dependent exponents triples $(\alpha(t), \beta(t), \gamma(t))$ satisfying the ideal scaling conditions (3.1), which ensure that $\operatorname{argmin} f$ is an attractor of the underlying dynamical system.

8.1.1 Potential Dilation

The first Lagrangian under consideration is a potential dilation of a mechanical Lagrangian, namely,

$$L(x, \dot{x}, t) = t^n \left(\frac{1}{2} \|\dot{x}\|^2 - f(x) \right) , \tag{8.1}$$

whose Euler-Lagrange equation is

$$\ddot{x} + \frac{n}{t} \dot{x} + \nabla f(x) = 0 \, , \tag{8.2}$$

which is the equation considered in Su et al. (2016).

A naive choice of exponents from which to obtain the time-dependent coefficients $a(t) = b(t) = t^n$ is

$$\alpha(t) = 0, \ \beta(t) = 0, \ \gamma(t) = n \log t. \tag{8.3}$$

However, they do not satisfy the ideal scaling conditions. A triple that does meet this requirement is

$$\alpha(t) = \log \mathsf{p} - \log t, \ \beta(t) = 2(\log t - \log \mathsf{p}), \ \gamma(t) = \mathsf{p} \log t + \log \mathsf{p}, \tag{8.4}$$

where $\mathsf{p} = n - 1$, but only when $n \geq 3$, thus, showing the "magic" $n = 3$ in (8.2) of NAG (Su et al., 2016). It is, in fact, the only choice satisfying the scaling conditions and gives an optimal rate of convergence not slower than $\mathcal{O}(1/t^2)$.

8.1.2 Modified Potential Dilation

The Lagrangian

$$L(x, \dot{x}, t) = t^n \left(\frac{1}{2} \|\dot{x}\|^2 - D t^{n-3} f(x) \right) , \tag{8.5}$$

whose Euler-Lagrange equation is

$$\ddot{x} + \frac{n}{t} \dot{x} + D t^{n-3} \nabla f(x) = 0 \, ,$$

is the Lagrangian considered by Wibisono et al. (2016) for the metric case and corresponds to the exponents

$$\alpha(t) = \log \mathsf{p} - \log t, \ \beta(t) = \mathsf{p} \log t + \log C, \ \gamma(t) = \mathsf{p} \log t + \log \mathsf{p}, \tag{8.6}$$

where $C = D/\mathsf{p}^2$ and $\mathsf{p} = n - 1$. They satisfy the ideal scaling conditions for $n > 1$, giving a rate of convergence not slower than $\mathcal{O}(1/t^{n-1})$ (confer with Wibisono et al., 2016, in particular for specific details on the constant C).

8.1.3 Exponential Dilation

Finally, the exponentially dilated Lagrangian

$$L(x, \dot{x}, t) = e^{\lambda t} \left(\frac{1}{2} \|\dot{x}\|^2 - f(x) \right) , \tag{8.7}$$

whose equation of motion is precisely the one of a mechanical system with linear damping,

$$\ddot{x} + \lambda\dot{x} + \nabla f(x) = 0\,, \tag{8.8}$$

corresponds to the choice exponents

$$\alpha(t) = \log\lambda,\ \beta(t) = -2\log\lambda,\ \gamma(t) = \lambda t + \log\lambda. \tag{8.9}$$

We note here three points. First, it is the unique choice for $a(t) = b(t) = e^{\lambda t}$ that satisfies the ideal scaling conditions. Second, in principle this choice gives a theoretical convergence rate of $\mathcal{O}(1)$, fortunately it can be reduced down to $o(1)$ (Attouch et al., 2021, Th. 2.5). Third and more notably, although the Lagrangian (8.7) is explicitly time-dependent, the Euler-Lagrange equation (8.8) is autonomous and it introduces a linear time-independent damping term in the equation.

8.2 Discretizations

Using the trapezoidal rule to approximate the action of the above Lagrangians, we retrieve common NAG coefficients that appear in the literature. With some abuse of notation, we write in general

$$a(k) \coloneqq \frac{a(t_k) + a(t_{k+1})}{2h^2}\,, \qquad \text{and} \qquad b^{\pm}(k) \coloneqq \frac{b(t_k)}{2}\,,$$

where $t_k = kh$.

8.2.1 Bounded Coefficients from the Potential Dilation

For the continuous time-dependent coefficients $a(t) = b(t) = t^n$, with $n = \mathsf{p} + 1$, the trapezoidal rule yields the discrete time-dependent coefficients

$$a(k) = \frac{t_k^n + t_{k+1}^n}{2h^2}, \quad b^{\pm}(k) = \frac{t_k^n}{2},$$

from which we obtain the coefficients

$$\mu(k) = \frac{k^n + (k-1)^n}{k^n + (k+1)^n}, \quad \eta(k) = \frac{2k^n}{k^n + (k+1)^n}h^2. \tag{8.10}$$

Both coefficients are strictly increasing and bounded above by 1 and h^2, respectively. To avoid integer overflow and slightly reduce computational cost in final implementations, these expressions can be simplified to

$$\mu(k) = \frac{2k-n}{2k+n} + o(1), \quad \eta(k) = \left(\frac{2k}{2k+n} + o(1)\right)h^2.$$

For the particular case $n = 3$, that is $\mathsf{p} = 2$, we get $\mu(k + {}^3\!/_2) = \frac{k}{k+3} + \mathcal{O}({}^1\!/_{k^3})$ as in (2.7).

8.2.2 Unbounded Coefficients from the Modified Potential Dilation

In this case, the continuous time-dependent coefficients are $a(t) = t^n$ and $b(t) = Dt^{2n-3}$, as in Wibisono et al. (2016), which yield

$$\mu(k) = \frac{k^n + (k-1)^n}{k^n + (k+1)^n}, \quad \eta(k) = D\frac{2k^n}{k^n + (k+1)^n} t_k^{n-3} h^2. \tag{8.11}$$

As earlier,

$$\mu(k) = \frac{2k-n}{2k+n} + o(1), \quad \eta(k) = D\left(\frac{2k}{2k+n} + o(1)\right) t_k^{n-3} h^2.$$

This time, η is unbounded when $n \geq 4$, and bounded above by Dh^2 when $n = 3$.

Similarly, Betancourt et al. (2018) suggest a palindromic split Hamiltonian method with 7 stages that can be recovered from the proposed perspective considering the discrete Lagrangian

$$L_d^k(x_k, x_{k+1}) = t_{k+1/2}^n \left(\frac{1}{2}\left\|\frac{x_{k+1} - x_k}{h}\right\|^2 - Dt_{k+1/2}^{n-3} \frac{f(x_k) + f(x_{k+1})}{2}\right).$$

Note that it is not obtained by a trapezoidal approximation of the Lagrangian (8.5), but it is still a discretization of it for which

$$\mu(k) = \left(\frac{2k-1}{2k+1}\right)^n, \quad \eta(k) = D\frac{(2k+1)^{2n-3} + (2k-1)^{2n-3}}{2(2k+1)^{2n-3}} t_{k+1/2}^{n-3} h^2.$$

As before,

$$\eta(k) = D\left(\frac{2k}{2k+2n-3} + o(1)\right) t_{k+1/2}^{n-3} h^2\,.$$

8.2.3 Constant Coefficients from the Exponential Dilation

Taking $a(t) = b(t) = e^{\lambda t}$ yields

$$\mu(k) = \frac{1 + e^{-\lambda h}}{1 + e^{+\lambda h}}, \quad \eta(k) = \frac{2}{1 + e^{\lambda h}} h^2. \tag{8.12}$$

For $\lambda = 1$ and $h = 0.1024$, $\mu \approx 0.9$ and $\eta \approx 0.01$, values that often appear in the literature, as in Sutskever et al. (2013). In general, any pair of constant coefficients $\mu, \eta > 0$ can be obtained from values $\lambda \in \mathbb{R}, h > 0$.

8.2.4 The Actual Method by Wibisono, Wilson, and Jordan

Strictly speaking, the method by Wibisono et al. (2016) is based on NAG, but it differs from it. Although they consider the previous Lagrangian (8.5) and experiment with it in Betancourt et al. (2018); Jordan (2018), what is proposed in Wibisono et al. (2016) is the

3-phase scheme

$$x_{k+1} = \frac{\mathsf{p}}{k+\mathsf{p}} z_k + \frac{k}{k+\mathsf{p}} y_k \,, \tag{8.13a}$$

$$y_k = \operatorname*{argmin}_y \left\{ f_{\mathsf{p}-1}(y; x_k) + \frac{N}{\mathsf{p} h^{\mathsf{p}}} \|y - x_k\|^{\mathsf{p}} \right\} , \tag{8.13b}$$

$$z_k = z_{k-1} - D \frac{k}{\mathsf{p}} t_k^{\mathsf{p}-2} h^2 \nabla f(y_k) \,, \tag{8.13c}$$

with $z_0 = y_0 = x_0$ and where $\mathsf{p} = n - 1$, $f_{\mathsf{p}-1}(y; x_k)$ is the $(\mathsf{p}-1)$-th Taylor expansion of f about x_k and N is a constant related to D and p that ensures convergence. Note that for $n = 3$ ($\mathsf{p} = 2$), the optimization problem (8.13b) is explicit and reduces to

$$y_k = x_k - \frac{1}{N} h^2 \nabla f(x_k) \,, \tag{8.14}$$

but it is implicit in general, which increases the cost of the method, aside of having to compute the Hessian and higher derivatives of f, either explicitly or by autodifferentiation.

In Section 8.4, we will refer to method (8.13) by WWJ, after the authors, and consider it as a (modified) NAG method.

8.3 Objective Functions

Several objective functions are considered: a highly dimensional quadratic function with tridiagonal matrix representation, a generalized Rosenbrock function, yet another test function for momentum-descent methods, and one that combines a generalized logistic function with a mean loss and that is often used in Neural Networks.

8.3.1 Highly Dimensional Quadratic Function

In Betancourt et al. (2018), they consider the quadratic map on $\mathbb{R}^n$

$$f(x) = \tfrac{1}{2} \left\langle x, \Sigma^{-1} x \right\rangle \,, \tag{8.15}$$

where Σ is the matrix whose elements are $\Sigma_{ij} = \rho^{|i-j|}$, with $\rho = 0.9$ and $n = 50$, and whose inverse Σ^{-1} is the tridiagonal matrix

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & & & \\ -\rho & 1+\rho^2 & -\rho & & \\ & \ddots & \ddots & \ddots & \\ & & -\rho & 1+\rho^2 & -\rho \\ & & & -\rho & 1 \end{pmatrix} .$$

8.3.2 Generalized Rosenbrock Function

The Rosenbrock function (Rosenbrock, 1960/61), whose expression is

$$f(x, y) = (a - x)^2 + b(y - x^2)^2 \,,$$

with $a, b > 0$, represents a banana-shaped flat-valley surrounded by steep walls with a unique critical point and global minimum at a, a^2, whose search by numerical means is difficult, hence its use to test and benchmark optimizers. We consider here its generalization to higher dimensions, $n > 2$, namely

$$f(x) = \sum_{i=1}^{n-1} \left[(1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2 \right] . \tag{8.16}$$

As in the two-dimensional case, the function has a global minimum at $(1, 1, \ldots, 1)$ but, unlike it, also has a local minimum close to $(-1, 1, \ldots, 1)$ (the higher is the dimension, the closer it gets).

8.3.3 Yet Another Test Function (YATF)

Another example that might hinder the search of a minimum is the following

$$f(x, y) = \sin(2x^2 - y^2 + 3) \cdot \cos(x + 1 - \exp(2y)) , \tag{8.17}$$

which has a local minimum close to $(0.32, 1.60)$.

8.3.4 Multinomial Logistic Regression

In artificial neural networks (ANN), the activation function of a node (or neuron) defines the output of that node given an input (or set of inputs). Supervised learning is a learning paradigm of the training process of the ANN. Different choices are available for the activation function and the training process, which is a subject that might be in some cases controversial within the ANN community, but that is not the object of this work. We consider a shallow neural network with a single layer for classification in n classes or targets given m features. Upon reaching the neurons, the inputs (features) are weighted and possibly biased, which in fact is the model to be determined through the learning process,

$$w \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^n \colon\ x \in \mathbb{R}^m \mapsto w \cdot x + b \in \mathbb{R}^n .$$

The chosen activation function is a classifier, a generalization of the logistic function, the multinomial logistic function (a.k.a. softmax),

$$\sigma(z) = \left(\frac{e^{z_j}}{\sum_k e^{z_k}} \right) .$$

As loss function, we choose the cross-entropy or log-loss,

$$\mathrm{H}(\hat{y}, y) = - \langle y, \log \hat{y} \rangle ,$$

where $\hat{y} \in \mathbb{R}^n$ is the computed output, $y \in \{0, 1\}^n$ with $\sum_j y_j = 1$ is the expected output (class), and log is applied componentwise. We take as objective function the average loss over a training dataset $\mathcal{D}$ of length $|\mathcal{D}|$, namely,

$$f(w, b) = -\frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \langle y, (\log \circ \sigma)(w \cdot x + b) \rangle . \tag{8.18}$$

It is important to note here that f is the sum of convex functions and therefore itself convex, however f need not have a global minimum but an asymptotic infimum: 0 as, for instance, $\exp(t)$ opposed to $\cosh(t)$. This depends on the data fed for the training process.

8.4 Experiments

Many experiments have been performed, we optimized each test function using the aforementioned methods (and others) set with different parameters and initial guesses, from which we present here a small but suggestive sample. In summary, we minimize the test functions (8.16) and (8.17), the quadratic function (8.15), and the loss (8.18) with the PHB/CM and NAG methods (7.9) given by the constant, bounded and unbounded coefficients (8.12), (8.10), and (8.11), respectively, and WWJ's method (8.13), the latter three for $n = 3, 4$, for a total of seven methods. Each objective is minimized using its own initial guess, time-step, and number of epochs, which are fixed for the seven simulations.

We set Rosenbrock's test function with 30 dimensions and seek for its global minimum at $(1, \dots, 1)$ from $(0, \dots, 0)$ at a pace of $h = 0.01$ for 20000 epochs. In the case of the YATF, there is a local minimum near $(0.32, 1.60)$ which we seek from $(-0.25, 0.35)$ with time-step $h = 0.01$ for 3800 epochs. A random point on a sphere of radius 50 is the initial guess for the 50-dimensional quadratic function, whose global minimum, sought for 10000 epochs with $h = 0.1$, is clearly at the origin. For the convergence tests, the log-loss function is fed with 10 arbitrary samples, with 4 features and 3 targets each, which defines an optimization problem of dimension 10 with no global minimum, hence we will seek for the infimum for 12000 epochs at a pace of $h \approx 0.945$ from a random weight distribution with null biases. For an actual ANN test, the log-loss function is fed with the widely used Iris dataset (Dua and Graff, 2017; see below for more details).

The methods have been implemented in Julia v1.8.2 (Bezanson et al., 2017), using solely as nonnative libraries NLsolve.jl (Mogensen et al., 2020) to solve the side problem (8.13c), Plots.jl (Breloff, 2021) and PGFPlotsX.jl for plotting, and Pluto.jl for an interactive notebook. Methods, functions, and simulations are available online (Campos, 2022a,b). All plots except Figures 1 and 7 represent the objective function residual against the epoch in log-log scale.

In Figure 1, we can see how the trajectories computed by PHB/CM and NAG for the YATF pass by its minimum about $(0.32, 1.60)$. They start at the bottom left and go upward until they "realize" they have overreached the minimum and, about $(0.7, 1.9)$, they back up. Although not shown in the figure, this motion is repeated successively, but each time they back up, they do so sooner and closer to the minimum, like a heavy ball in a bowl. The trajectories shown in the figure correspond to the iterations 765 to 1450 of the methods.

Similarly we observe in Figure 2 that NAG oscillates less than PHB/CM. Each downward peak corresponds to the trajectory passing by the minimum of the YATF. In fact, the trajectories shown in Fig. 1 correspond to the first peak in blue of Fig. 2. These oscillations where the trajectories pass by the minimum of the objective function back and forth, and the fact that NAG oscillates less than PHB/CM slightly outperforming it are common aspects of all the simulations performed, reason why in the remaining figures we focus solely on NAG methods, where several aspects are worth noting.

Figures 3-6 compare the seven methods enumerated above, one figure for each objective function presented in Section 8.3, and following the same order. Each figure is made up of two plots: the top one is composed of methods with $n = 3$, the bottom one of methods with $n = 4$, whereas both include the NAG method with constant coefficients for proper reference.

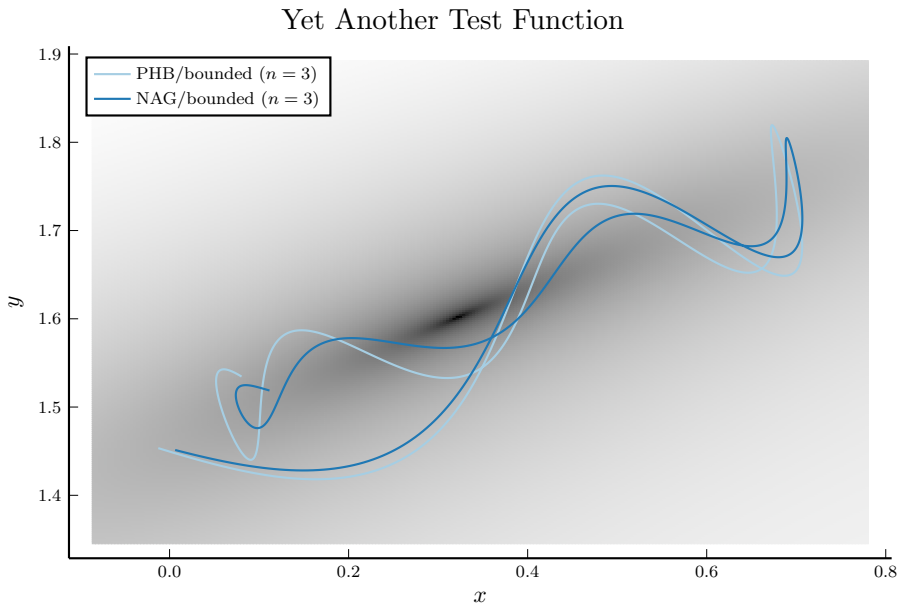


Figure 1: Trajectory slices nearby the local minimum of the YATF using PHB/CM (pale) and NAG (strong) with the bounded coefficients from the Lagrangian's polynomial dilation with $n = 3$. A nonlinear grayscale gradient indicates the minimum's location in black.

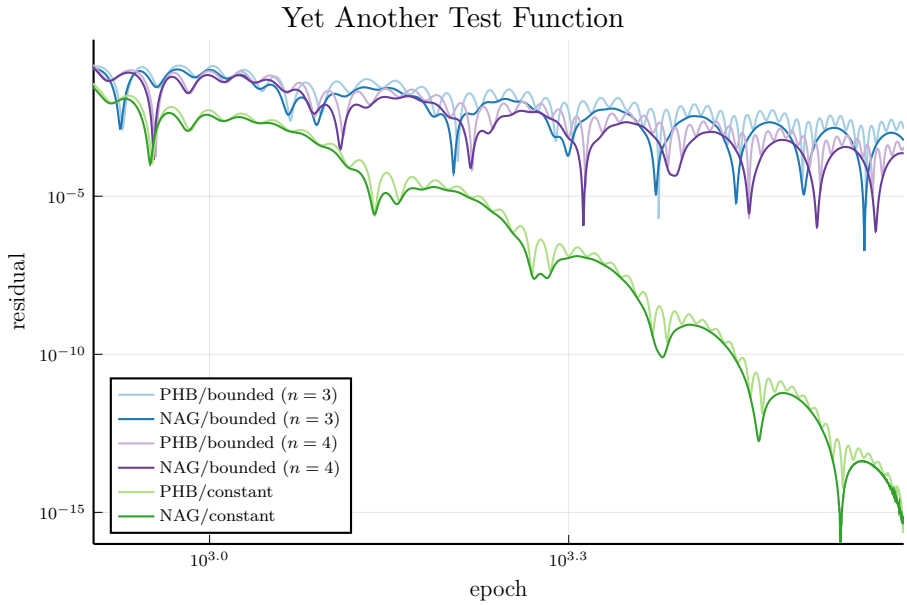


Figure 2: YATF residual values along PHB/CM (pale) and NAG (strong) trajectories for coefficients from the polynomially dilated Lagrangian with $n = 3$ (blue) and $n = 4$ (violet), and from the exponentially dilated Lagrangian (green).

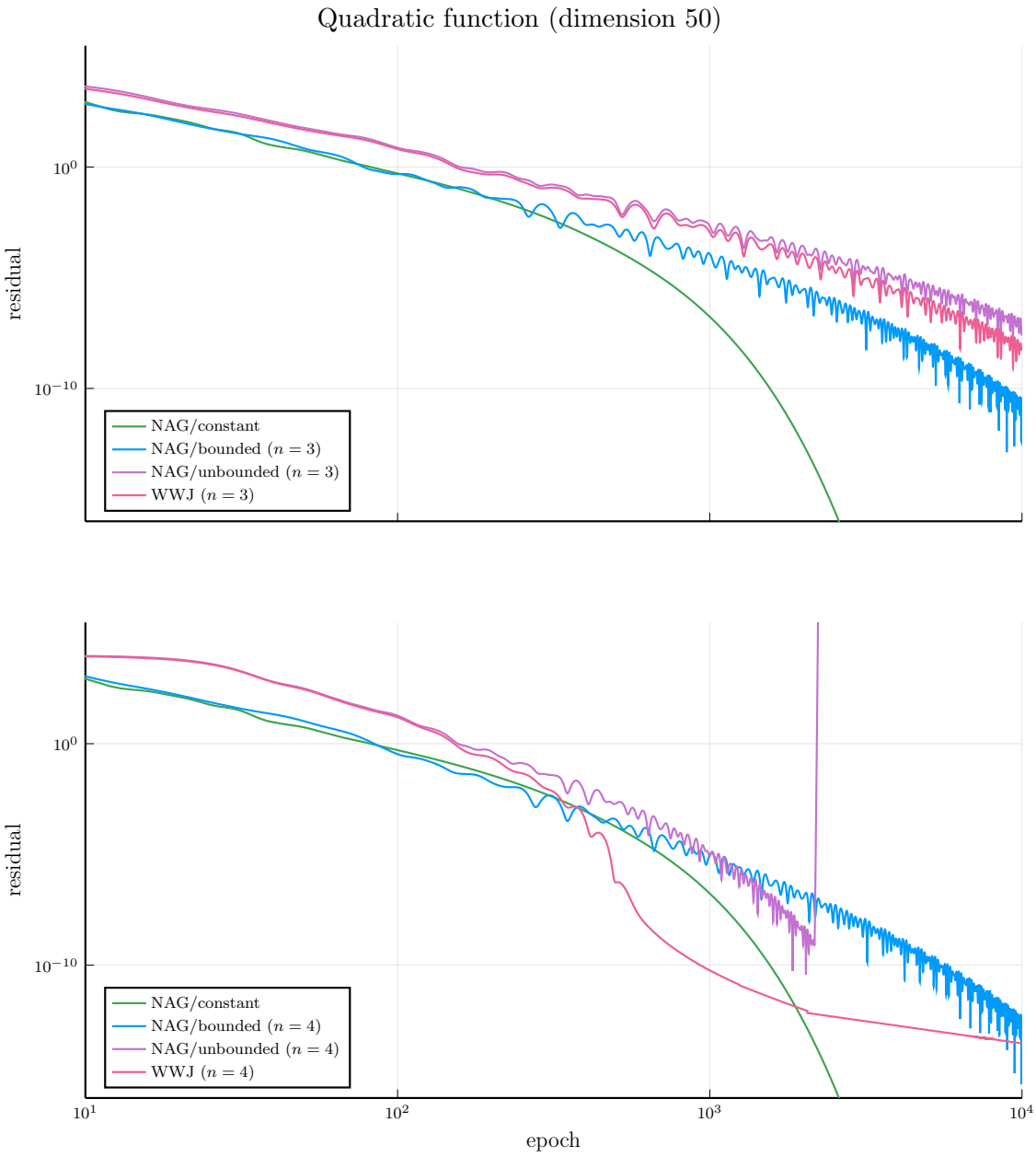


Figure 3: Quadratic test function values along trajectories computed with NAG for constant (green), bounded (blue), and unbounded (violet) coefficients, and WWJ (red); the latter three set with $n = 3$ (top) and $n = 4$ (bottom).

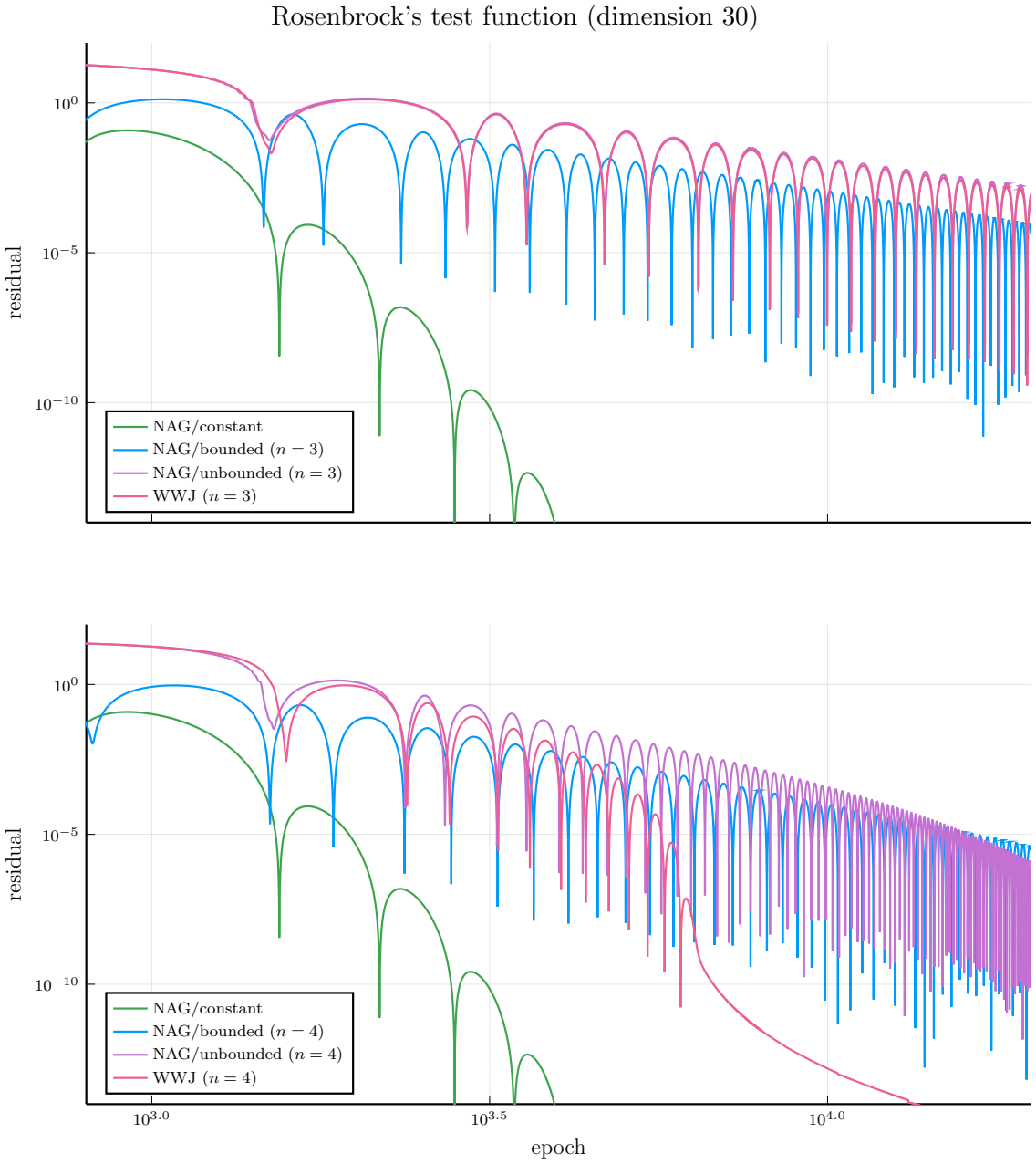


Figure 4: Rosenbrock's test function values along trajectories computed with NAG for constant (green), bounded (blue), and unbounded (violet) coefficients, and WWJ (red); the latter three set with $n = 3$ (top) and $n = 4$ (bottom).

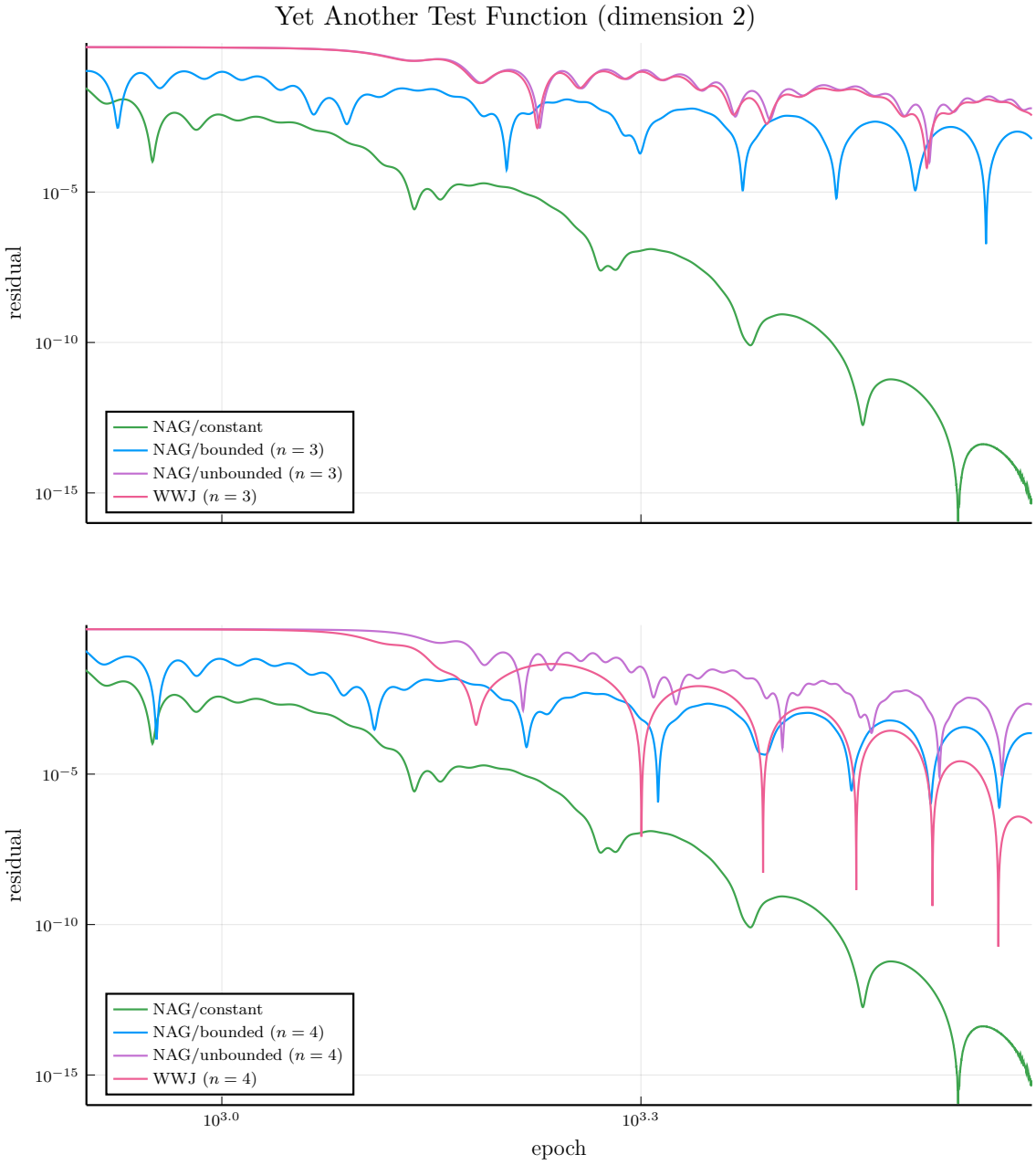


Figure 5: YATF residual values along trajectories computed with NAG for constant (green), bounded (blue), and unbounded (violet) coefficients, and WWJ (red); the latter three set with $n = 3$ (top) and $n = 4$ (bottom).

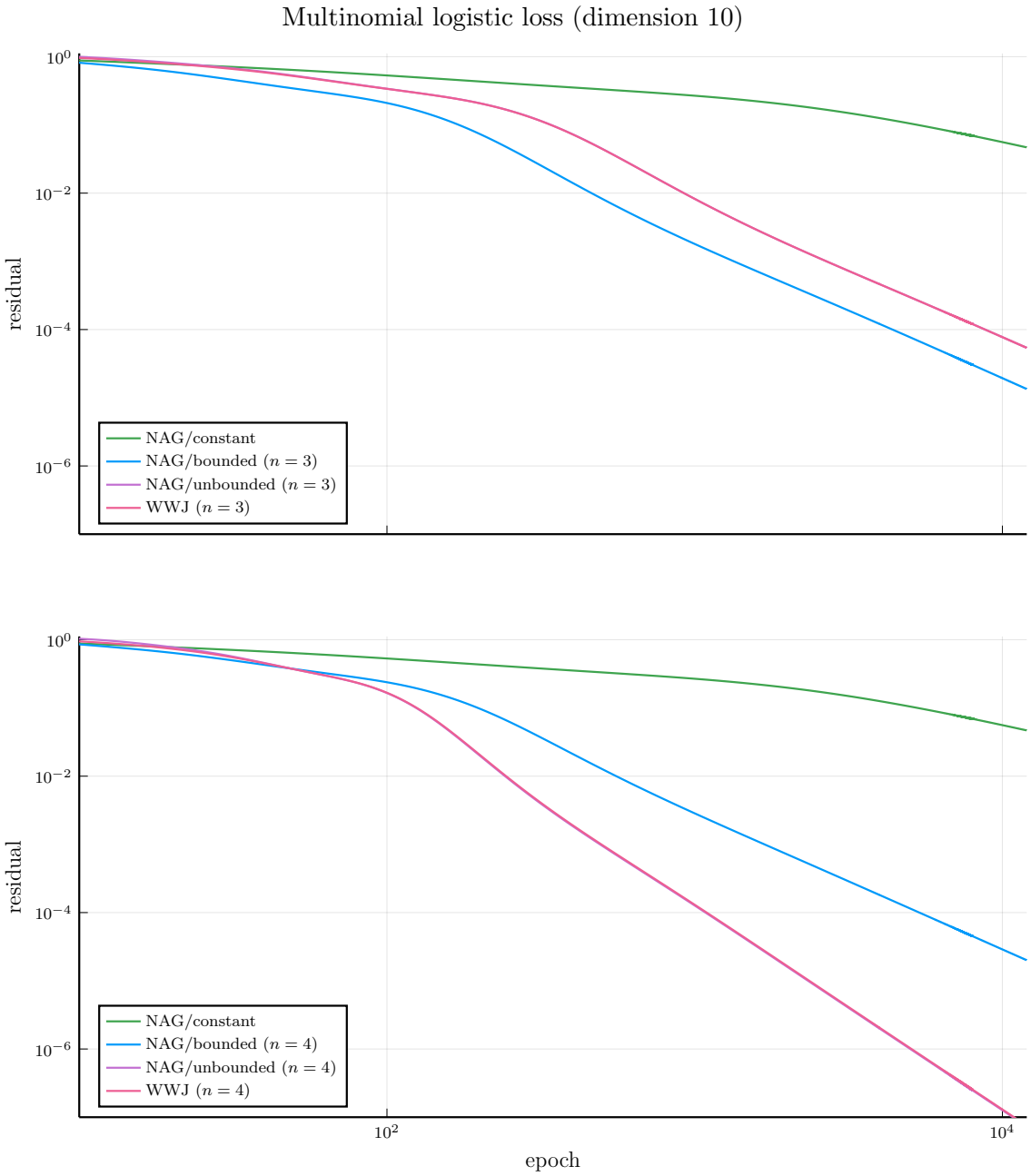


Figure 6: Logistic loss values along trajectories computed with NAG for constant (green), bounded (blue), and unbounded (violet) coefficients, and WWJ (red); the latter three set with $n = 3$ (top) and $n = 4$ (bottom).

Observe first that, the three methods with $n = 3$ behave similarly for the four objective functions. This is not surprising since the Lagrangians considered reduce to a potentially dilated mechanical Lagrangian in which the objective function, in the cases of unbounded NAG and WWJ, is coupled with a further constant coefficient, $D = 1/4$ when $n = 3$, which slightly delays convergence when compared to bounded NAG. This coefficient needs to be small in order to ensure convergence according to Wibisono et al. (2016).

When $n = 4$, these three methods clearly show their differences. Most importantly, unbounded NAG (8.11) blows up in Figure 3 (highly dimensional quadratic function). This is due to the increasing and unbounded learning rate coefficient η that is "ignored" when the gradient is almost null but brakes the method when it steps at a point where the gradient is not negligible. Therefore we infer that unbounded NAG is not suited for long runs when $n \geq 4$ and should be restarted to avoid blow-up. Surprisingly WWJ is not affected by this problem even though the same coefficient appears in its definition, Eq. (8.13). The method is numerically stable thanks to the fast convergence of the trajectory towards the minimum.

In fact, WWJ really shows up when $n = 4$, where it clearly outperforms its NAG counterpart, unbounded NAG, as well as bounded NAG, but may "stall" when the simulation has advanced, Fig. 3. Besides, there is a trick into its performance, when $n \geq 4$ the method must solve a side optimization problem, (8.13c), which is solved here using the NLsolve.jl library, something that it somewhat redundant and suffers from the curse of dimensionality: it is around 10 times slower than the other methods for the low dimensional case (YATF) and more than 100 times slower in the high dimensional ones (Rosenbrock and quadratic function). Nonetheless, this disadvantage might decrease or even disappear when the Bregman divergence is not the usual Euclidean norm (something worth exploring), since the associated methods are unlikely to be explicit.

With regards to constant NAG, the results are somehow paradoxical. Recall that the exponentially dilated Lagrangian is the only case in which a good convergence rate is not ensured, Eq. (8.9), however it is the method that in general performs the best, Figures 3-5. Furthermore, although within the Machine Learning community it is perhaps the most popular method among the analyzed, it is the one that has performed the worst in the purely ML scenario, Figure 6.

Fig. 6 is where the theoretical convergence rates are better seen. In this figure, the values obtained by unbounded NAG and WWJ coalesce for both, $n = 3$ and $n = 4$.

In addition to the previous optimization problems, we apply the analyzed methods to a simple classification problem using Firsher's Iris dataset (Dua and Graff, 2017). The dataset consists of 150 entries (or samples) with 7 fields: 4 features and 3 targets. Therefore we consider a shallow neural network with a single layer of 3 neurons with 4 inputs. The input data (the features) are weighted and biases upon entry into the network, the computed output is compared with the expected output (targets) using the multinomial logistic function. This process is summarized by the objective function (8.18).

For simplicity, we only consider four methods: constant, bounded, and unbounded NAG, and WWJ, the latter three with $n = 3$. We train the network (or optimize the objective function) at an increasing number of epochs (from 25 up to 250) for 1000 runs. At each run and for each number of epochs, the samples are randomly split in two: 100 samples for training and 50 samples for testing. At the same time, the initial weights of the network are drawn from a normal distribution with null mean and standard deviation $\sigma = 10$. This

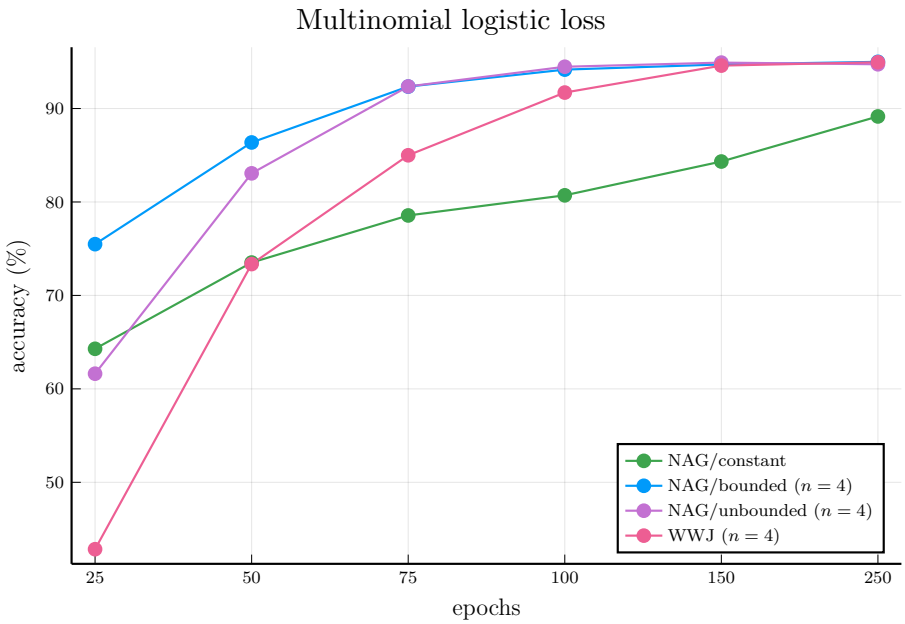


Figure 7: Average accuracies for Iris dataset achieved by NAG with constant (green), bounded (blue), and unbounded (violet) coefficients, and WWJ (red); the latter three set with $n = 4$.

initial guess is kept for the four methods. Then the network is trained and the accuracy of network with the computed weights is measured (percentage of correct matches for the testing data).

Figure 7 shows the average accuracy along the 1000 runs for each method *versus* the number of epochs. This figure is clearly consistent with what is obtained in Fig. 6.

9. Conclusions and Future work

In this paper, we have studied the relation between accelerated optimization and discrete variational calculus, proving a symplecticity property for the continuous differential equation in Theorem 1 which is also preserved by the corresponding discrete variational methods. We have derived Classical Momentum (CM) or Polyak's Heavy Ball (PHB), and Nesterov's Accelerated Gradient methods (NAG) from the discrete Hamiltonian and Lagrange-d'Alembert principles in Theorem 6 adding forces in the picture and proven a one-to-one correspondence. Several simulations were performed showing the applicability of our techniques to optimization. Among all the methods, NAG with the constant coefficients from the exponentially dilated Lagrangian, aside from being the simplest choice, it also seems to be the best one for general purpose applications according to the simulations.

In a future paper, we will study whether the proposed optimization algorithm generated by using Lagrange-d'Alembert principle achieves the accelerated rate for minimizing both strongly convex functions and convex functions (Wibisono et al., 2016; Shi et al., 2019). The

main idea is to discretize, using discrete variational calculus, the continuous Euler-Lagrange equations (with or without forces) while maintaining their convergence rates (see Vaquero et al., 2021, for recent advances in this topic). Moreover, the extension to problems of accelerated optimization in manifolds will be given using discrete variational calculus and well-know optimization techniques with retraction maps (Absil et al., 2008; see also Lee et al., 2021).

Acknowledgments

D. Martín de Diego acknowledges financial support from the Spanish Ministry of Science and Innovation, under grants PID2019-106715GB-C21, from the Spanish National Research Council (CSIC), through the "Ayuda extraordinaria a Centros de Excelencia Severo Ochoa" R&D (CEX2019-000904-S). A. Mahillo would like to thank CSIC for its financial support through a JAE Intro scholarship.

References

Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics*. AMS Chelsea Publishing, Redwood City, CA, 2 edition, 1978.

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. ISBN 978-0-691-13298-3. doi: 10.1515/9781400830244. URL `https://doi.org/10.1515/9781400830244`. With a foreword by Paul Van Dooren.

Hedy Attouch, Zaki Chbani, and Hassan Riahi. Fast convex optimization via time scaling of damped inertial gradient dynamics. *Pure Appl. Funct. Anal.*, 6(6):1081–1117, 2021. ISSN 2189-3756.

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

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL `https://doi.org/10.1137/141000671`.

Sergio Blanes and Fernando Casas. *A concise introduction to geometric numerical integration*. Monographs and Research Notes in Mathematics. CRC Press, Boca Raton, FL, 2016. ISBN 978-1-4822-6342-8.

L. M. Brègman. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat i Mat. Fiz.*, 7:620–631, 1967. ISSN 0044-4669.

Tom Breloff. Plots.jl, May 2021. URL `https://doi.org/10.5281/zenodo.4776893`.

Cédric M. Campos. High order variational integrators: a polynomial approach. In *Advances in differential equations and applications*, volume 4 of *SEMA SIMAI Springer Ser.*, pages

249–258. Springer, Cham, 2014. doi: 10.1007/978-3-319-06953-1_24. URL https://doi.org/10.1007/978-3-319-06953-1_24.

Cédric M. Campos and J. M. Sanz-Serna. Palindromic 3-stage splitting integrators, a roadmap. *J. Comput. Phys.*, 346:340–355, 2017. ISSN 0021-9991. doi: 10.1016/j.jcp.2017.06.006. URL https://doi.org/10.1016/j.jcp.2017.06.006.

Cédric M. Campos. Personal website, 2022a. URL https://cmcampos.xyz.

Cédric M. Campos. Research repository, 2022b. URL https://github.com/cmcampos-xyz.

Beniamino Cappelletti-Montano, Antonio De Nicola, and Ivan Yudin. A survey on cosymplectic geometry. *Rev. Math. Phys.*, 25(10):1343002, 55, 2013. ISSN 0129-055X. doi: 10.1142/S0129055X13430022. URL https://doi.org/10.1142/S0129055X13430022.

Agustin-Louis Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *C. R. Acad. Sci.*, 25:536—-538, 1847. URL https://gallica.bnf.fr/ark:/12148/bpt6k2982c/f540.item.

Elena Celledoni, Matthias J. Ehrhardt, Etmann Christian, Robert I McLachlan, Brynjulf Owren, Carola-Bibiane Schönlieb, and Sherry Ferdia. Structure preserving deep learning. *arXiv*, 2006.03364, 2020. URL https://arxiv.org/abs/2006.03364.

Manuel de León and Paulo R. Rodrigues. *Methods of Differential Geometry in Analytical Mechanics*, volume 158. Elsevier, Amsterdam, 1987. ISBN 0-08-087269-7.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Valentin Duruisseaux, Jeremy Schmitt, and Melvin Leok. Adaptive Hamiltonian variational integrators and applications to symplectic accelerated optimization. *SIAM J. Sci. Comput.*, 43(4):A2949–A2980, 2021. ISSN 1064-8275. doi: 10.1137/20M1383835. URL https://doi.org/10.1137/20M1383835.

E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Heidelberg, 2010. ISBN 978-3-642-05157-9. Structure-preserving algorithms for ordinary differential equations, Reprint of the second (2006) edition.

P. Hartman. *Ordinary differential equations*, volume 38 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. ISBN 0-89871-510-5. doi: 10.1137/1.9780898719222. URL http://dx.doi.org/10.1137/1.9780898719222. Corrected reprint of the second (1982) edition [Birkhäuser, Boston, MA; MR0658490 (83e:34002)], With a foreword by Peter Bates.

Michael I. Jordan. Dynamical symplectic and stochastic perspectives on gradient-based optimization. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. I. Plenary lectures*, pages 523–549. World Sci. Publ., Hackensack, NJ, 2018.

Taeyoung Lee, Molei Tao, and Melvin Leok. Variational symplectic accelerated optimization on lie groups. *arXiv*, 2103.14166, 2021. URL https://arxiv.org/abs/2103.14166.

Paulette Libermann. Sur les automorphismes infinitésimaux des structures symplectiques et des structures de contact. In *Colloque Géom. Diff. Globale (Bruxelles, 1958)*, pages 37–59. Centre Belge Rech. Math., Louvain, 1959.

Paulette Libermann and Charles-Michel Marle. *Symplectic geometry and analytical mechanics*, volume 35 of *Mathematics and its Applications*. D. Reidel Publishing Co., Dordrecht, 1987. ISBN 90-277-2438-5. doi: 10.1007/978-94-009-3807-6. URL https://doi.org/10.1007/978-94-009-3807-6. Translated from the French by Bertram Eugene Schwarzbach.

J. C. Marrero, D. Martín de Diego, and E. Martínez. On the exact discrete lagrangian function for variational integrators: theory and applications, 2016. URL https://arxiv.org/abs/1608.01586.

J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numer.*, 10:357–514, 2001. ISSN 0962-4929. doi: 10.1017/S096249290100006X. URL http://dx.doi.org/10.1017/S096249290100006X.

Håkon Marthinsen and Brynjulf Owren. Geometric integration of non-autonomous linear Hamiltonian problems. *Adv. Comput. Math.*, 42(2):313–332, 2016. ISSN 1019-7168. doi: 10.1007/s10444-015-9425-0. URL https://doi.org/10.1007/s10444-015-9425-0.

Patrick Kofod Mogensen, Kristoffer Carlsson, Sébastien Villemot, Spencer Lyon, Matthieu Gomez, Christopher Rackauckas, Tim Holy, David Widmann, Tony Kelman, Daniel Karrasch, Antoine Levitt, Asbjørn Nilsen Riseth, Carlo Lucibello, Changhyun Kwon, David Barton, Julia TagBot, Mateusz Baran, Miles Lubin, Sarthak Choudhury, Simon Byrne, Simon Christ, Takafumi Arakaki, Troels Arnfred Bojesen, benneti, and Miguel Raz Guzmán Macedo. Julianlsolvers/nlsolve.jl: v4.5.1, December 2020. URL https://doi.org/10.5281/zenodo.4404703.

Yu. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983. ISSN 0002-3264.

Yurii Nesterov. *Lectures on convex optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2018. ISBN 978-3-319-91577-7; 978-3-319-91578-4. doi: 10.1007/978-3-319-91578-4. URL https://doi.org/10.1007/978-3-319-91578-4. Second edition of [MR2142598].

G. W. Patrick and C. Cuell. Error analysis of variational integrators of unconstrained Lagrangian systems. *Numer. Math.*, 113(2):243–264, 2009. ISSN 0029-599X. doi: 10.1007/s00211-009-0245-3. URL http://dx.doi.org/10.1007/s00211-009-0245-3.

Elijah Polak. *Optimization*, volume 124 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1997. ISBN 0-387-94971-2. doi: 10.1007/978-1-4612-0663-7. URL https://doi.org/10.1007/978-1-4612-0663-7. Algorithms and consistent approximations.

Boris T. Polyak. Some methods of speeding up the convergence of iterative methods. *Ž. Vyčisl. Mat i Mat. Fiz.*, 4:791–803, 1964. ISSN 0044-4669.

Boris T. Polyak. *Introduction to optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987. ISBN 0-911575-14-6. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *Comput. J.*, 3:175–184, 1960/61. ISSN 0010-4620. doi: 10.1093/comjnl/3.3.175. URL https://doi.org/10.1093/comjnl/3.3.175.

J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian problems*, volume 7 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1994. ISBN 0-412-54290-0.

B Shi, Du S. S., Jordan M.I., and Su J.U. Acceleration via symplectic discretization of high-resolution differential equations. *arXiv*, 2019. URL https://arxiv.org/pdf/1902.03694.pdf.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016. URL http://jmlr.org/papers/v17/15-084.html.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Advances in Neural Information Processing Systems*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/sutskever13.html.

M. Vaquero, Mestres P., and J: Cortés. Resource-aware discretization of accelerated optimization flows. *Preprint*, 2021. URL http://carmenere.ucsd.edu/jorge/publications/data/2020_VaMeCo-tac.pdf.

Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proc. Natl. Acad. Sci. USA*, 113(47):E7351–E7358, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1614734113. URL https://doi.org/10.1073/pnas.1614734113.

Andre Yohannes Wibisono. *Variational and Dynamical Perspectives On Learning and Optimization*. ProQuest LLC, Ann Arbor, MI, 2016. ISBN 978-1369-05764-5. Thesis (Ph.D.)–University of California, Berkeley.