SIAM J. SCI. COMPUT.
Vol. 43, No. 4, pp. A2949–A2980

ADAPTIVE HAMILTONIAN VARIATIONAL INTEGRATORS AND APPLICATIONS TO SYMPLECTIC ACCELERATED OPTIMIZATION*

VALENTIN DURUISSEAUX†, JEREMY SCHMITT†, AND MELVIN LEOK†

Abstract. It is well known that symplectic integrators lose their near energy preservation properties when variable time-steps are used. The most common approach to combining adaptive time-steps and symplectic integrators involves the Poincaré transformation of the original Hamiltonian. In this article, we provide a framework for the construction of variational integrators using the Poincaré transformation. Since the transformed Hamiltonian is typically degenerate, the use of Hamiltonian variational integrators based on Type II or Type III generating functions is required instead of the more traditional Lagrangian variational integrators based on Type I generating functions. Error analysis is provided, and numerical tests based on the Taylor variational integrator approach in [J. M. Schmitt, T. Shingel, and M. Leok, *BIT*, 58 (2018), pp. 457–488] to time-adaptive variational integration of Kepler's 2-body problem are presented. Finally, we use our adaptive framework together with the variational approach to accelerated optimization presented in [A. Wibisono, A. Wilson, and M. Jordan, *Proc. Natl. Acad. Sci. USA*, 113 (2016), pp. E7351–E7358] to design efficient variational and nonvariational explicit integrators for symplectic accelerated optimization.

Key words. accelerated optimization, symplectic integrators, time adaptivity, variational integrators

AMS subject classifications. 37N40, 65K10, 65P10, 70H15

DOI. 10.1137/20M1383835

1. Introduction.

Symplectic integrators form a class of geometric numerical integrators of interest since, when applied to Hamiltonian systems, they yield discrete approximations of the flow that preserve the symplectic 2-form (see [19]). The preservation of symplecticity results in the preservation of many qualitative aspects of the underlying dynamical system. In particular, when applied to conservative Hamiltonian systems, symplectic integrators show excellent long-time near-energy preservation. However, when symplectic integrators were first used in combination with variable time-steps, the near-energy preservation was lost and the integrators performed poorly (see [8, 17]). Backward error analysis provided justification both for the excellent long-time near-energy preservation of symplectic integrators and for the poor performance experienced when using variable time-steps (see Chapter IX of [19]). Backward error analysis shows that symplectic integrators can be associated with a modified Hamiltonian in the form of a power series in terms of the time-step. The use of a variable time-step results in a different modified Hamiltonian at every iteration where the time-step is changed, which is the source of the poor energy conservation. There has been a great effort to circumvent this problem, and there have been many successes. However, there has yet to be a unified general framework for constructing adaptive symplectic integrators. In this paper, we contribute to this effort by demon-

*Submitted to the journal's Methods and Algorithms for Scientific Computing section December 2, 2020; accepted for publication (in revised form) June 4, 2021; published electronically August 19, 2021.

https://doi.org/10.1137/20M1383835

Funding: This work was partially supported by the NSF under grants DMS-1411792, DMS-1345013, and DMS-1813635, by the AFOSR under grant FA9550-18-1-0288, and by the DoD under grant 13106725 (Newton Award for Transformative Ideas during the COVID-19 Pandemic).

†Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 USA (vduruiss@ucsd.edu, j2schmit@gmail.com, mleok@math.ucsd.edu).

strating how Hamiltonian variational integrators [33] can be used to systematically construct symplectic integrators that allow for the use of variable time-steps.

The use of variable time-steps is motivated by the observation that the global error estimates for a numerical method depend in part on the maximum local truncation error, and this in turn is related to both the time-step and the magnitude of the $(r+1)$-derivatives of the solution for an r-order numerical method. For a fixed number of time-steps, the maximum local truncation error is minimized if the local truncation error is equidistributed over the time intervals. In turn, this can be achieved if, for example, the time-step is chosen to be an appropriate function of the reciprocal of the relevant derivative of the solution. This derivative can be estimated a posteriori by comparing methods with different orders of accuracy, or methods with the same order of accuracy but different error constants. Alternatively, in the Kepler 2-body problem, for example, Kepler's second law states that the line joining the planet and the Sun sweeps out equal areas during equal intervals of time, so the angular velocity of the planet is proportional to the reciprocal of the radius squared, which gives an a priori bound. In essence, variable time-steps are chosen to control the error incurred at each time-step, which in turn affects the global accuracy of the numerical trajectory.

The goal of this paper is to develop an analogue of the methods derived using the framework of [18, 47], but directly in terms of generating functions of symplectic maps. These prior results are based on symplectic (partitioned) Runge–Kutta methods, which are related to Type I generating functions [54], but we desire an explicit characterization of the flow maps of time-adaptive Hamiltonian systems so that we can employ the Hamiltonian variational integrator framework instead.

Variational integrators provide a systematic method for constructing symplectic integrators of arbitrarily high order based on the discretization of Hamilton's principle [34, 20] or, equivalently, by the approximation of generating functions, but there has not been a systematic attempt to incorporate time-adaptivity into the setting of variational integrators. This is due to the fact that the Poincaré transformed Hamiltonian that is used is in general degenerate, so there is no corresponding Lagrangian analogue, which prevents the use of traditional variational integrators that are based on a Lagrangian formulation of mechanics and involve the construction of a discrete Lagrangian that approximates a Type I generating function given by Jacobi's solution of the Hamilton–Jacobi equation. Instead, we propose the use of Hamiltonian variational integrators [33], which are based on Type II and Type III generating functions that have no difficulty with this degeneracy.

After a brief introduction to variational integrators in section 2.1, we will review the construction of Type II and Type III Hamiltonian Taylor variational integrators from [50] and present a new theorem concerning their order of accuracy in section 2.2. We will then present a framework for variable time-step variational integrators in section 3.1, derive corresponding error analysis results in section 3.2, and test our approach with Hamiltonian Taylor variational integrators on Kepler's 2-body problem in section 3.3. Finally, in section 4, we will design efficient variational and nonvariational explicit integrators for symplectic accelerated optimization, using our adaptive approach applied to the variational framework for accelerated optimization introduced in [58].

2. Hamiltonian variational integrators.

2.1. Variational integration.

Variational integrators are derived by discretizing Hamilton's principle, instead of discretizing Hamilton's equations directly. As a result, variational integrators are symplectic, preserve many invariants and momentum maps, and have excellent long-time near-energy preservation (see [34]).

Type I. Traditionally, variational integrators have been designed based on the Type I generating function known as the discrete Lagrangian, $L_d : Q \times Q \to \mathbb{R}$. The exact discrete Lagrangian of the true flow of Hamilton's equations can be represented in both a variational form and in a boundary-value form. The latter is given by

$$L_d^E(q_0, q_1; h) = \int_0^h L(q(t), \dot{q}(t))dt, \tag{2.1}$$

where $q(0) = q_0$, $q(h) = q_1$, and q satisfies the Euler–Lagrange equations over the time interval $[0, h]$. A variational integrator is defined by constructing an approximation $L_d : Q \times Q \to \mathbb{R}$ to L_d^E and then applying the discrete Euler–Lagrange equations,

$$p_k = -D_1 L_d(q_k, q_{k+1}), \qquad p_{k+1} = D_2 L_d(q_k, q_{k+1}), \tag{2.2}$$

which implicitly define the integrator $\tilde{F}_{L_d} : (q_k, p_k) \mapsto (q_{k+1}, p_{k+1})$, where D_i denotes a partial derivative with respect to the ith argument. The error analysis is greatly simplified via Theorem 2.3.1 of [34], which states that if a discrete Lagrangian, $L_d : Q \times Q \to \mathbb{R}$, approximates the exact discrete Lagrangian $L_d^E : Q \times Q \to \mathbb{R}$ to order r, i.e.,

$$L_d(q_0, q_1; h) = L_d^E(q_0, q_1; h) + \mathcal{O}(h^{r+1}), \tag{2.3}$$

then the discrete Hamiltonian map $\tilde{F}_{L_d} : (q_k, p_k) \mapsto (q_{k+1}, p_{k+1})$, viewed as a one-step method, has order of accuracy r. Many other properties of the integrator, such as momentum conservation properties of the method, can be determined by analyzing the associated discrete Lagrangian, as opposed to analyzing the integrator directly.

More recently, variational integrators have been extended to the framework of Type II and Type III generating functions, commonly referred to as discrete Hamiltonians (see [29, 33, 49]). Hamiltonian variational integrators are derived by discretizing Hamilton's phase space principle.

Type II. The boundary-value formulation of the exact Type II generating function of the time-h flow of Hamilton's equations is given by the exact discrete right Hamiltonian,

$$H_d^{+,E}(q_0, p_1; h) = p_1^\top q_1 - \int_0^h \left[p(t)^\top \dot{q}(t) - H(q(t), p(t))\right] dt, \tag{2.4}$$

where (q, p) satisfies Hamilton's equations with boundary conditions $q(0) = q_0$ and $p(h) = p_1$. A Type II Hamiltonian variational integrator is constructed by using an approximate discrete Hamiltonian H_d^+, and applying the discrete right Hamilton's equations,

$$p_0 = D_1 H_d^+(q_0, p_1), \qquad q_1 = D_2 H_d^+(q_0, p_1), \tag{2.5}$$

which implicitly defines the integrator, $\tilde{F}_{H_d^+} : (q_0, p_0) \mapsto (q_1, p_1)$.

Theorem 2.3.1 of [34], which simplified the error analysis for Lagrangian variational integrators, has an analogue for Hamiltonian variational integrators. Theorem 2.2 in [49] states that if a discrete right Hamiltonian H_d^+ approximates the exact discrete right Hamiltonian $H_d^{+,E}$ to order r, i.e.,

$$H_d^+(q_0, p_1; h) = H_d^{+,E}(q_0, p_1; h) + \mathcal{O}(h^{r+1}), \tag{2.6}$$

then the discrete right Hamilton's map $\tilde{F}_{H_d^+} : (q_k, p_k) \mapsto (q_{k+1}, p_{k+1})$, viewed as a one-step method, is order r accurate.

Type III. The boundary-value formulation of the exact Type III generating function of the time-h flow of Hamilton's equations is given by the exact discrete left Hamiltonian,

$$H_d^{-,E}(q_1, p_0; h) = -p_0^\top q_0 - \int_0^h \left[p(t)^\top \dot{q}(t) - H(q(t), p(t))\right] dt, \tag{2.7}$$

where (q, p) satisfies Hamilton's equations with boundary conditions $q(h) = q_1$ and $p(0) = p_0$. A Type III Hamiltonian variational integrator is constructed by using an approximate discrete left Hamiltonian H_d^-, and applying the discrete left Hamilton's equations,

$$p_1 = -D_1 H_d^-(q_1, p_0), \qquad q_0 = -D_2 H_d^-(q_1, p_0), \tag{2.8}$$

which implicitly defines the integrator, $\tilde{F}_{H_d^-} : (q_0, p_0) \mapsto (q_1, p_1)$. As mentioned in [49], the proof of Theorem 2.2 in [49] can be easily adjusted to prove an equivalent theorem for the discrete left Hamiltonian case, which states that if a discrete left Hamiltonian H_d^- approximates the exact discrete left Hamiltonian $H_d^{-,E}$ to order r, i.e.,

$$H_d^-(q_1, p_0; h) = H_d^{-,E}(q_1, p_0; h) + \mathcal{O}(h^{r+1}), \tag{2.9}$$

then the discrete left Hamilton's map $\tilde{F}_{H_d^-} : (q_k, p_k) \mapsto (q_{k+1}, p_{k+1})$, viewed as a one-step method, is order r accurate.

Examples of Hamiltonian variational integrators include Galerkin variational integrators [33], prolongation-collocation variational integrators [32], and Taylor variational integrators [50]. In many cases, the Type I and Type II/III approaches will produce equivalent integrators. This equivalence has been established in [50] for Taylor variational integrators provided the Lagrangian is hyperregular, and in [33] for generalized Galerkin variational integrators constructed using the same choices of basis functions and numerical quadrature formula provided the Hamiltonian is hyperregular. However, Hamiltonian and Lagrangian variational integrators are not always equivalent. In particular, it was shown in [49] that even when the Hamiltonian and Lagrangian integrators are analytically equivalent, they might still have different numerical properties because of numerical conditioning issues. Even more to the point, Lagrangian variational integrators cannot always be constructed when the underlying Hamiltonian is degenerate, and in that situation, Hamiltonian variational integrators are the more natural choice. Depending on the form of the Hamiltonian and the method used to design the corresponding approximate discrete Hamiltonian, one of the Type II or Type III approaches might be more convenient than the other, in the sense that it might allow for an explicit algorithm or might allow for higher-order methods given some constraints on the type of methods permitted. In section 3, we will examine a transformation commonly used to construct variable time-step symplectic integrators, which results in a degenerate Hamiltonian in most cases of interest, such as the optimization application considered in section 4. We will apply Hamiltonian variational integrators to the resulting transformed Hamiltonian system. For the optimization application presented in section 4, we will prefer Type II Hamiltonian Taylor variational integrators to their Type III analogues, and this choice will be justified carefully based on the order and explicitness of the resulting methods.

2.2. Hamiltonian Taylor variational integrators (HTVIs). We now present HTVIs [50], together with a new theorem concerning their order of accuracy, which is analogous to Theorem 3.1 in [50] for their Lagrangian counterpart. A discrete approximate Hamiltonian is constructed by approximating the flow map and the trajectory associated with the boundary values using a Taylor method and by approximating the integral by a quadrature rule. The HTVI is then generated by the discrete Hamilton's equations associated with that discrete Hamiltonian. More explicitly, we first construct the r-order and $(r+1)$-order Taylor methods $\Psi_h^{(r)}$ and $\Psi_h^{(r+1)}$ approximating the exact time-h flow map $\Phi_h : T^*Q \to T^*Q$ corresponding to Hamilton's equation $\dot{z} = \varphi(z)$, where $z = (q, p)$:

$$\Psi_h^{(r)}(z_0) = z_0 + \sum_{k=1}^{r} \frac{h^k}{k!} \varphi^{(k-1)}(z_0). \tag{2.10}$$

Let $\pi_{T^*Q} : (q, p) \mapsto p$ and $\pi_Q : (q, p) \mapsto q$. Given a quadrature rule of order s with weights and nodes (b_i, c_i) for $i = 1, \dots, m$, the Type II and Type III integrators are then constructed as follows.

Type II:

(i) Approximate $p(0) = p_0$ by the solution $\tilde{p}_0$ of the problem $p_1 = \pi_{T^*Q} \circ \Psi_h^{(r)}(q_0, \tilde{p}_0)$.

(ii) Generate approximations $(q_{c_i}, p_{c_i}) \approx (q(c_i h), p(c_i h))$ via $(q_{c_i}, p_{c_i}) = \Psi_{c_i h}^{(r)}(q_0, \tilde{p}_0)$.

(iii) Approximate q_1 via $\tilde{q}_1 = \pi_Q \circ \Psi_h^{(r+1)}(q_0, \tilde{p}_0)$.

(iv) Use the continuous Legendre transform to obtain $\dot{q}_{c_i} = \frac{\partial H}{\partial p_{c_i}}$.

(v) Apply the quadrature rule to obtain the associated discrete right Hamiltonian

$$H_d^+(q_0, p_1; h) = p_1^\top \tilde{q}_1 - h \sum_{i=1}^{m} b_i \left[p_{c_i}^\top \dot{q}_{c_i} - H(q_{c_i}, p_{c_i}) \right].$$

(vi) The variational integrator is then defined by the implicit discrete right Hamilton's equations

$$q_1 = D_2 H_d^+(q_0, p_1), \qquad p_0 = D_1 H_d^+(q_0, p_1).$$

Type III:

(i) Approximate $q(0) = q_0$ by the solution $\tilde{q}_0$ of the problem $q_1 = \pi_Q \circ \Psi_h^{(r+1)}(\tilde{q}_0, p_0)$.

(ii) Generate approximations $(q_{c_i}, p_{c_i}) \approx (q(c_i h), p(c_i h))$ via $(q_{c_i}, p_{c_i}) = \Psi_{c_i h}^{(r)}(\tilde{q}_0, p_0)$.

(iii) Use the continuous Legendre transform to obtain $\dot{q}_{c_i} = \frac{\partial H}{\partial p_{c_i}}$.

(iv) Apply the quadrature rule to obtain the associated discrete left Hamiltonian

$$H_d^-(q_1, p_0; h) = -p_0^\top \tilde{q}_0 - h \sum_{i=1}^{m} b_i \left[p_{c_i}^\top \dot{q}_{c_i} - H(q_{c_i}, p_{c_i}) \right].$$

(v) The variational integrator is then defined by the implicit discrete left Hamilton's equations

$$p_1 = -D_1 H_d^-(q_1, p_0), \qquad q_0 = -D_2 H_d^-(q_1, p_0).$$

Taylor variational integrators were inspired by a resurgence of interest in high-order Taylor methods for celestial mechanics that has been fueled by the continued

progress in automatic differentiation software (see [25, 4, 1, 45, 3, 37, 39]). Implicit modified Taylor methods have been proposed to deal with stiff ODEs [28], while Taylor variational integrators provide a class of Taylor-based integrators to deal with conservative Hamiltonian systems, and can be viewed as a predictor-corrector method that applies a symplectic correction to the Taylor method. For high-order Taylor methods, the key to an efficient implementation relies upon efficient automatic differentiation software to compute higher-order gradients.

We now present a theorem specifying the order of accuracy of the resulting HTVIs.

THEOREM 2.1. *If the Hamiltonian H and $\frac{\partial H}{\partial p}$ are Lipschitz continuous in both variables, then the discrete Hamiltonian $H_d^{\pm}$ obtained using the above construction approximates $H_d^{\pm,E}$ with at least order of accuracy* $\min(r+1, s)$. *By Theorem* 2.2 *in* [49] *(or its analogue for the left Hamiltonian case), the associated discrete Hamiltonian map has the same order of accuracy.*

Proof. See Appendix A. ∎

3. Adaptive integrators and variational error analysis.

3.1. The Poincaré transformation and discrete Hamiltonians. Given an autonomous Hamiltonian $H(q,p)$, and a desired transformation of time $t \mapsto \tau$ described by the monitor function $g(q,p)$ via

$$\frac{dt}{d\tau} = g(q,p), \tag{3.1}$$

a new Hamiltonian system is constructed using the Poincaré transformation,

$$\bar{H}(\bar{q},\bar{p}) = g(q,p)\left(H(q,p) + p^t\right), \tag{3.2}$$

where $\bar{q} = \left[\begin{smallmatrix} q \\ q^t \end{smallmatrix}\right]$ and $\bar{p} = \left[\begin{smallmatrix} p \\ p^t \end{smallmatrix}\right]$. We will make the common choice $q^t = t$ and $p^t = -H(q(0),p(0))$, so that $\bar{H}(\bar{q},\bar{p}) = 0$ along all integral curves through $(\bar{q}(0),\bar{p}(0))$. The time t shall be referred to as the physical time, while τ will be referred to as the fictive time.

In general, along an integral curve through $(\bar{q}(0),\bar{p}(0))$,

$$\frac{\partial^2 \bar{H}}{\partial \bar{p}^2} = \begin{bmatrix} \frac{\partial H}{\partial p}\nabla_p g(q,p)^\top + g(q,p)\frac{\partial^2 H}{\partial p^2} + \nabla_p g(q,p)\frac{\partial H}{\partial p}^\top & \nabla_p g(q,p) \\ \nabla_p g(q,p)^\top & 0 \end{bmatrix}, \tag{3.3}$$

which can be singular for many initial Hamiltonians H and choices of monitor function g.

Most of the prior literature on variable time-step symplectic integrators cited in this paper focuses exclusively on monitor functions that are only a function of position, in which case $\frac{\partial^2 \bar{H}}{\partial \bar{p}^2}$ is singular, and the associated Legendre transformation, $\mathbb{F}\bar{H} : T^*Q \to TQ$ is noninvertible, which is to say that the resulting transformed Hamiltonian is degenerate and there is no corresponding Lagrangian formulation. Therefore, the Type II and Type III Hamiltonian variational integrator frameworks are the most general and natural way to derive variable time-step variational integrators.

The exact Type II generating function for the transformed Hamiltonian is given by

$$\bar{H}_d^{+,E}(\bar{q}_0,\bar{p}_1;h) = \bar{p}_1^\top \bar{q}_1 - \int_0^h \left(\bar{p}(\tau)^\top \dot{\bar{q}}(\tau) - \bar{H}(\bar{q}(\tau),\bar{p}(\tau))\right) d\tau, \tag{3.4}$$

where $(\bar{q}(\tau), \bar{p}(\tau))$ satisfy the Hamilton's equations corresponding to the Poincaré transformed Hamiltonian $\bar{H}$ with boundary conditions $\bar{q}(0) = \bar{q}_0$ and $\bar{p}(h) = \bar{p}_1$. This exact discrete right Hamiltonian implicitly defines a symplectic map with respect to the symplectic form $\bar{\omega}(\bar{p}_k, \bar{q}_k)$ on $T^*\bar{Q}$ via the discrete Legendre transforms given by

$$\bar{p}_0 = \frac{\partial \bar{H}_d^{+,E}}{\partial \bar{q}_0}, \qquad \bar{q}_1 = \frac{\partial \bar{H}_d^{+,E}}{\partial \bar{p}_1}. \tag{3.5}$$

Similarly, the exact Type III generating function for the transformed Hamiltonian is given by

$$\bar{H}_d^{-,E}(\bar{q}_0, \bar{p}_1; h) = -\bar{p}_0^\top \bar{q}_0 - \int_0^h \left(\bar{p}(\tau)^\top \dot{\bar{q}}(\tau) - \bar{H}(\bar{q}(\tau), \bar{p}(\tau))\right) d\tau, \tag{3.6}$$

where $(\bar{q}(\tau), \bar{p}(\tau))$ satisfy the Hamilton's equations corresponding to the Poincaré transformed Hamiltonian $\bar{H}$ with boundary conditions $\bar{q}(h) = \bar{q}_1$ and $\bar{p}(0) = \bar{p}_0$. This exact discrete left Hamiltonian implicitly defines a symplectic map with respect to the symplectic form $\bar{\omega}(\bar{p}_k, \bar{q}_k)$ on $T^*\bar{Q}$ via the discrete Legendre transforms given by

$$\bar{p}_1 = -\frac{\partial \bar{H}_d^{-,E}}{\partial \bar{q}_1}, \qquad \bar{q}_0 = -\frac{\partial \bar{H}_d^{-,E}}{\partial \bar{p}_0}. \tag{3.7}$$

Our approach is to construct Hamiltonian variational integrators using a discrete Hamiltonian $\bar{H}_d^{\pm}$ that approximates the corresponding exact discrete Hamiltonian $\bar{H}_d^{\pm,E}$ to order r. The resulting integrator will be symplectic with constant time-step in fictive time τ and more importantly with the desired variable time-step in physical time t via $\frac{dt}{d\tau} = g(q, t, p)$. It is important to note that this method will be symplectic in two different ways. It will be symplectic both with respect to the symplectic form $d\bar{p} \wedge d\bar{q}$ and with respect to the symplectic form $dp \wedge dq$. Since $\dot{p}^t = 0$, p^t is constant and $dp_k^t \wedge dq_k^t = 0$, the symplectic form in generalized coordinates is given by

$$\bar{\omega}(\bar{p}_k, \bar{q}_k) = d\bar{p}_k \wedge d\bar{q}_k = \sum_{i=1}^{n+1} d\bar{p}_{k,i} \wedge d\bar{q}_{k,i} = \sum_{i=1}^{n} dp_{k,i} \wedge dq_{k,i} = \omega(p_k, q_k). \tag{3.8}$$

A symplectic variable time-step method was proposed independently in [18] and [47], which applied a symplectic integrator to the Poincaré transformed Hamiltonian. In [18], it is noted that one of the first applications of the Poincaré transformation was by Levi-Civita, who applied it to the three-body problem. A more in-depth discussion of such time transformations can be found in [52]. Further work using this type of transformation has been published, such as [5, 6], which focused on developing symplectic, explicit, splitting methods with variable time-steps.

The novelty of our approach consists in discretizing the Type II or Type III generating function for the flow of Hamilton's equations, where the Hamiltonian is given by the Poincaré transformation. Therefore, we are constructing variational integrators, and in particular Hamiltonian variational integrators (see [29, 33]). The use of Type II or Type III integrators is justified by the degeneracy of the Hamiltonian, which implies that there is no corresponding Type I Lagrangian formulation. This approach works seamlessly with existing methods and theorems of Hamiltonian variational integrators, but now the system under consideration is the transformed Hamiltonian system resulting from the Poincaré transformation. The methods of [18, 47] include the possibility of applying a given variational integrator to the transformed differential equations. Our approach gives a framework for constructing variational integrators at the level of the generating function by using the Poincaré transformed discrete right Hamiltonian.

Remark 3.1. Other approaches to variable time-step variational integrators can be found in [26, 35, 36]. In particular, [26] is inspired by the result of [16], which states that constant time-step symplectic integrators of autonomous Hamiltonian systems cannot exactly conserve the energy unless it agrees with the exact flow map up to a time reparametrization. Therefore, they sought a variable time-step energy-conserving symplectic integrator in an expanded nonautonomous system. However, symplecticity is with respect to the space-time symplectic form $dp \wedge dq + dH \wedge dt$. The time-step is determined by enforcing discrete energy conservation, which arises as a consequence of the fact that energy is the Noether quantity associated with time translational symmetry. An extended Hamiltonian is used, similar in spirit to the Poincaré transformation. An approach that builds off this idea and space-time symplecticity was presented in [36], and a less constrained choice of time-step was allowed. In [35], adaptive variational integrators are constructed using a transformation of the Lagrangian, which is motivated by the Poincaré transformation, but it is not equivalent. The lack of equivalence is not surprising since the Poincaré transformed Hamiltonian is degenerate for their choice of monitor functions. As a consequence, the phase space path is not preserved.

Note that our framework can be extended to the case where the original Hamiltonian H and the chosen monitor function g depend explicitly on time t (inspired by [18]). Given a time-dependent Hamiltonian $H(q,t,p)$, consider a desired transformation of time $t \mapsto \tau$, given by $\frac{dt}{d\tau} = g(q,t,p)$. Then, we can define $\bar{q} = \left[\begin{smallmatrix} q \\ q^t \end{smallmatrix}\right]$, where $q^t = t$ and $\bar{p} = \left[\begin{smallmatrix} p \\ p^t \end{smallmatrix}\right]$, where p^t is the conjugate momentum for $q^t = t$ with $p^t(0) = -H(q(0),0,p(0))$. Consider the new Hamiltonian system given by the Poincaré transformation

$$\bar{H}(\bar{q},\bar{p}) = g(q,q^t,p)\left(H(q,q^t,p) + p^t\right). \tag{3.9}$$

The corresponding equations of motion in the extended phase space are then given by

$$\dot{\bar{q}} = \frac{\partial \bar{H}}{\partial \bar{p}}, \qquad \dot{\bar{p}} = -\frac{\partial \bar{H}}{\partial \bar{q}}. \tag{3.10}$$

Suppose $(\bar{Q}(\tau), \bar{P}(\tau))$ are solutions to these extended equations of motion, and let $(q(t), p(t))$ solve Hamilton's equations for the original Hamiltonian H. Then,

$$\bar{H}(\bar{Q}(\tau), \bar{P}(\tau)) = \bar{H}(\bar{Q}(0), \bar{P}(0)) = 0. \tag{3.11}$$

Therefore, the components $(Q(\tau), P(\tau))$ in the original phase space of the solutions $(\bar{Q}(\tau), \bar{P}(\tau))$ satisfy

$$H(Q(\tau),\tau,P(\tau)) = -p^t(\tau), \qquad H(Q(0),0,P(0)) = -p^t(0) = H(q(0),0,p(0)). \tag{3.12}$$

Then, $(Q(\tau), P(\tau))$ and $(q(t), p(t))$ both satisfy Hamilton's equations for the original Hamiltonian H with the same initial values, so they must be the same. As before,

$$\frac{\partial^2 \bar{H}}{\partial \bar{p}^2} = \begin{bmatrix} \frac{\partial H}{\partial p}\nabla_p g(\bar{q},p)^\top + g(\bar{q},p)\frac{\partial^2 H}{\partial p^2} + \nabla_p g(\bar{q},p)\frac{\partial H}{\partial p}^\top & \nabla_p g(\bar{q},p) \\ \nabla_p g(\bar{q},p)^\top & 0 \end{bmatrix} \tag{3.13}$$

will be singular in many cases, so Hamiltonian variational integrators are the most general and natural way to derive variable time-step variational integrators.

3.2. Variational error analysis. The standard error analysis for Hamiltonian variational integrators assumes a nondegenerate Hamiltonian, i.e., $\det\left(\frac{\partial^2 \bar{H}}{\partial \bar{p}^2}\right) \neq 0$ (see [49]), which might not be the case for the Poincaré transformed Hamiltonian. The nondegeneracy of the Hamiltonian ensures that we can apply the usual implicit function theorem to the discrete Hamilton's equations, and the proof of the standard error analysis theorem relies upon Lemma 2.3 of [49].

LEMMA 3.2. *Let $f_1, g_1, e_1, f_2, g_2, e_2 \in C^r$ (r-times continuously differentiable) be such that*

$$f_1(x,h) = g_1(x,h) + h^{r+1}e_1(x,h), \qquad f_2(x,h) = g_2(x,h) + h^{r+1}e_2(x,h). \tag{3.14}$$

Then, there exist functions e_{12} and $\bar{e}_1$ bounded on compact sets such that

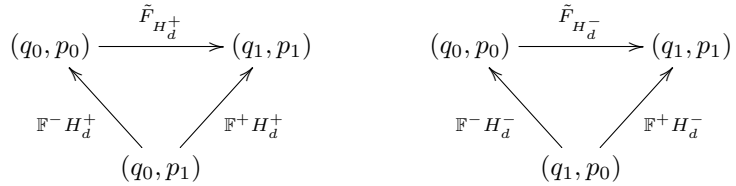
$$f_2(f_1(x,h),h) = g_2(g_1(x,h),h) + h^{r+1}e_{12}(g_1(x,h),h), \tag{3.15}$$

$$f_1^{-1}(y) = g_1^{-1}(y) + h^{r+1}\bar{e}_1(y). \tag{3.16}$$

Given a discrete Hamiltonian $H_d^{\pm}$, we introduce the discrete fiber derivatives (or discrete Legendre transforms), $\mathbb{F}^{\pm}H_d^{\pm}$:

$$\begin{aligned}
&\mathbb{F}^{+}H_d^{+}(q_0,p_1) : (q_0,p_1) \mapsto (D_2H_d^{+}(q_0,p_1),p_1),\\
&\mathbb{F}^{+}H_d^{-}(q_1,p_0) : (q_1,p_0) \mapsto (q_1,-D_1H_d^{-}(q_1,p_0)),\\
&\mathbb{F}^{-}H_d^{+}(q_0,p_1) : (q_0,p_1) \mapsto (q_0,D_1H_d^{+}(q_0,p_1)),\\
&\mathbb{F}^{-}H_d^{-}(q_1,p_0) : (q_1,p_0) \mapsto (-D_2H_d^{-}(q_1,p_0),p_0).
\end{aligned}$$

We observe that the following diagrams commute:



As such, the discrete left and right Hamiltonian maps can be expressed in terms of the discrete fiber derivatives,

$$\tilde{F}_{H_d^{\pm}}(q_0,p_0) = \mathbb{F}^{+}H_d^{\pm} \circ (\mathbb{F}^{-}H_d^{\pm})^{-1}(q_0,p_0) = (q_1,p_1), \tag{3.17}$$

and this observation together with Lemma 3.2 ensures that the order of accuracy of the integrator is at least of the order to which the discrete Hamiltonian $H_d^{\pm}$ approximates the exact discrete Hamiltonian $H_d^{\pm,E}$.

However, the Poincaré transformed Hamiltonian might be degenerate, so we cannot apply the usual implicit function theorem, and we need to establish the invertibility of the discrete Legendre transform $\mathbb{F}^{-}H_d^{\pm}$ in a different way.

The strongest general result we have been able to establish involves the case where the original Hamiltonian is autonomous, i.e., $H = H(q,p)$, and nondegenerate, and the monitor function is autonomous as well. These assumptions hold for an interesting and useful class of problems, and we will show that the exact discrete left and right Hamiltonians can be reduced to a particular form and that the extended variables p_1^t

and q_1^t can be solved for explicitly. As a result, the implicit function theorem is not needed with respect to these variables.

Hamilton's equations of the Poincaré transformed Hamiltonian

$$\bar{H}(\bar{q},\bar{p}) = g(q,p)\left(H(q,p)+p^t\right)$$

are given by

$$\dot{\bar{q}} = \begin{bmatrix} \nabla_p g(q,p)(H(q,p)+p^t) + \frac{\partial H}{\partial p} g(q,p) \\ g(q,p) \end{bmatrix}, \quad \dot{\bar{p}} = -\begin{bmatrix} \nabla_q g(q,p)(H(q,p)+p^t) + \frac{\partial H}{\partial q} g(q,p) \\ 0 \end{bmatrix}.$$

Using these equations, the corresponding exact discrete Hamiltonians are of the form

(3.18)
$$\bar{H}_d^{+,E}(\bar{q}_0,\bar{p}_1;h) = p_1^\top q_1 + p_1^t q_1^t - \int_0^h \left(p(\tau)^\top \dot{q}(\tau) - g(q(\tau),p(\tau))H(q(\tau),p(\tau))\right) d\tau,$$

(3.19)
$$\bar{H}_d^{-,E}(\bar{q}_1,\bar{p}_0;h) = -p_0^\top q_0 - p_0^t q_0^t - \int_0^h \left(p(\tau)^\top \dot{q}(\tau) - g(q(\tau),p(\tau))H(q(\tau),p(\tau))\right) d\tau.$$

As a result, only one part of these exact discrete left and right Hamiltonians requires approximations of the extended variable q^t and p^t. Furthermore, $\dot{p}^t = 0$ implies that $p_1^t = p_0^t$.

Now, let $\bar{H}_d^{\pm}$ be approximations to the exact discrete left and right Hamiltonians of the form

$$\begin{aligned} \bar{H}_d^{+}(\bar{q}_0,\bar{p}_1;h) &= p_1^\top \hat{q}_1(q_0,p_1;h) + p_1^t \hat{q}_1^t(q_0^t,q_0,p_1;h) - I_1(q_0,p_1;h), \\ \bar{H}_d^{-}(\bar{q}_1,\bar{p}_0;h) &= -p_0^\top \hat{q}_0(q_1,p_0;h) - p_0^t \hat{q}_0^t(q_1^t,q_1,p_0;h) - I_2(q_1,p_0;h), \end{aligned} \tag{3.20}$$

where $\hat{\cdot}$ denotes an approximation and where $I_1(q_0,p_1;h)$ and $I_2(q_1,p_0;h)$ both approximate

$$\int_0^h \left(p(\tau)^\top \dot{q}(\tau) - g(q(\tau),p(\tau))H(q(\tau),p(\tau))\right) d\tau. \tag{3.21}$$

Then, the discrete right Legendre transforms give the following relations for p_1^t and q_1^t:

(3.22)
$$\begin{bmatrix} p_0 \\ p_0^t \end{bmatrix} = \begin{bmatrix} \frac{\partial \hat{q}_1}{\partial q_0}^\top p_1 + p_1^t \frac{\partial \hat{q}_1^t}{\partial q_0} - \frac{\partial I_1}{\partial q_0} \\ \frac{\partial \hat{q}_1^t}{\partial q_0^t} p_1^t \end{bmatrix}, \qquad \begin{bmatrix} q_1 \\ q_1^t \end{bmatrix} = \begin{bmatrix} \hat{q}_1 + \frac{\partial \hat{q}_1}{\partial p_1}^\top p_1 + \frac{\partial \hat{q}_1}{\partial p_1}^\top p_1^t - \frac{\partial I_1}{\partial p_1} \\ \hat{q}_1^t \end{bmatrix}.$$

Now, since the analytic solution satisfies $p_1^t = p_0^t$, there is no need to approximate p_1^t. Therefore, $\frac{\partial \hat{q}_1^t}{\partial q_0^t} = 1$. The resulting two systems can both be solved by first setting $p_1^t = p_0^t$, then implicitly solving for p_1 in terms of (q_0^t, q_0, p_1^t, p_1), explicitly solving for q_1, and finally explicitly solving for q_1^t. Since p_1 is not determined by q_1^t, the implicit function theorem is simply needed for finding p_1. Therefore, we need $\det\left(\frac{\partial^2 \bar{H}}{\partial p^2}\right) \neq 0$, and from (3.3), this is the same as $\det\left(\frac{\partial H}{\partial p}\nabla_p g(q,p)^\top + g(q,p)\frac{\partial^2 H}{\partial p^2} + \nabla_p g(q,p)\frac{\partial H}{\partial p}^\top\right) \neq 0$. Note that this holds for nondegenerate Hamiltonians H and p-independent monitor functions.

Similarly, the discrete left Legendre transforms give the following relations for p_1^t and q_1^t:

(3.23)
$$\begin{bmatrix} p_1 \\ p_1^t \end{bmatrix} = \begin{bmatrix} \frac{\partial \hat{q}_0}{\partial q_1}^\top p_0 + p_0^t \frac{\partial \hat{q}_0^t}{\partial q_1} + \frac{\partial I_2}{\partial q_1} \\ \frac{\partial \hat{q}_0^t}{\partial q_1^t} p_0^t \end{bmatrix}, \qquad \begin{bmatrix} q_0 \\ q_0^t \end{bmatrix} = \begin{bmatrix} \hat{q}_0 + \frac{\partial \hat{q}_0}{\partial p_0}^\top p_0 + \frac{\partial \hat{q}_0}{\partial p_0}^\top p_0^t + \frac{\partial I_2}{\partial p_0} \\ \hat{q}_0^t \end{bmatrix},$$

which can be solved provided $\det\left(\frac{\partial H}{\partial p}\nabla_p g(q,p)^\top + g(q,p)\frac{\partial^2 H}{\partial p^2} + \nabla_p g(q,p)\frac{\partial H}{\partial p}^\top\right) \neq 0$.

The results that we have established are summarized in the following theorem.

THEOREM 3.3. *Consider a nondegenerate Hamiltonian H, and a monitor function $g \in C^1([0,h])$, such that* $\det\left(\frac{\partial H}{\partial p}\nabla_p g(q,p)^\top + g(q,p)\frac{\partial^2 H}{\partial p^2} + \nabla_p g(q,p)\frac{\partial H}{\partial p}^\top\right) \neq 0$. *Then, if the discrete Hamiltonian $\bar{H}_d^{\pm}$ approximates the exact discrete Hamiltonian $\bar{H}_d^{\pm,E}$ to order r, i.e.,*

$$\bar{H}_d^{\pm}(\bar{q}_0, \bar{p}_1; h) = \bar{H}_d^{\pm,E}(\bar{q}_0, \bar{p}_1; h) + \mathcal{O}(h^{r+1}), \tag{3.24}$$

then the discrete Hamiltonian map $\tilde{F}_{\bar{H}_d^{\pm}} : (\bar{q}_k, \bar{p}_k) \mapsto (\bar{q}_{k+1}, \bar{p}_{k+1})$, viewed as one-step method, is order r accurate.

Remark 3.4. It should be noted that the assumptions that the original Hamiltonian is nondegenerate and autonomous fail to hold in the application we consider of time-adaptive variational integrators to the discretization of the Bregman Hamiltonian associated with accelerated optimization, as it is time-dependent. This is unavoidable, as it models a system with dissipation, which cannot be described with an autonomous Hamiltonian, as the Hamiltonian would otherwise be an integral of motion, as it is the Noether quantity associated with time translational symmetry. In the cases when the original Hamiltonian is degenerate or nonautonomous, we need to analyze the solvability of the discrete Hamiltonian equations on a case-by-case basis, but as we demonstrate, this can be done in the case of the Bregman Hamiltonian with the given choices of monitor function $g(t)$ and discrete Hamiltonians that we consider.

3.3. Numerical tests on Kepler's planar 2-body problem. We will now demonstrate the approach using HTVIs, presented in section 2.2, on Kepler's planar 2-body problem. For a lucid exposition, we will at first assume that $g(q,p) = g(q)$ and $H(q,p) = \frac{1}{2}p^\top M^{-1} p + V(q)$. Consider the discrete right Hamiltonian given by approximating $\bar{q}_1$ with a first-order Taylor method about $\bar{q}_0$, approximating $\bar{p}_0$ with a zeroth-order Taylor expansion about $\bar{p}_0$, and using the rectangular quadrature rule about the initial point:

$$\bar{H}_d^{+} = p_1^\top \left(q_0 + \frac{1}{2}hg(q_0)M^{-1}p_1\right) + p_1^t(q_0^t + hg(q_0)) + hg(q_0)V(q_0). \tag{3.25}$$

The corresponding variational integrator is given by

$$\begin{aligned} \bar{p}_1 &= \begin{bmatrix} p_0 - hg(q_0)\nabla V(q_0) - h\nabla g(q_0)\left(\frac{1}{2}p_1^\top M^{-1}p_1 + V(q_0) + p_0^t\right) \\ p_0^t \end{bmatrix}, \\ \bar{q}_1 &= \begin{bmatrix} q_0 + hg(q_0)M^{-1}p_1 \\ q_0^t + hg(q_0) \end{bmatrix}. \end{aligned} \tag{3.26}$$

This integrator is merely symplectic Euler-B applied to the transformed Hamiltonian system

$$\bar{q}_1 = \bar{q}_0 + h\frac{\partial \bar{H}(\bar{q}_0, \bar{p}_1)}{\partial \bar{p}}, \qquad \bar{p}_1 = \bar{p}_0 - h\frac{\partial \bar{H}(\bar{q}_0, \bar{p}_1)}{\partial \bar{q}}. \tag{3.27}$$

In fact, this is precisely the adaptive symplectic integrator first proposed in [18] and presented in [31, page 254]. Most existing symplectic integrators can be interpreted as variational integrators, but there are also new methods that are most naturally derived as variational integrators. We will also consider a fourth-order Hamiltonian Taylor variational integrator (HTVI4), which is distinct from any existing symplectic method.

One of the most important aspects of implementing a variable time-step symplectic integrator of this form is a well-chosen monitor function, $g(q)$. We need g to be positive-definite, so that we never stall or march backward in time. Noting that the above integrator is first-order, a natural choice is to use the second-order truncation error given by $-\frac{(q_1^t-q_0^t)^2}{2}M^{-1}\nabla V(q_0)$. Let tol be some desired level of accuracy. Then, using $q_1^t - q_0^t = hg(q_0)$, one choice for g would be

$$g(q_0) = \frac{tol}{\|\frac{(q_1^t-q_0^t)^2}{2}g(q_0)M^{-1}\nabla V(q_0)\|} = \frac{tol}{\|\frac{h^2 g(q_0)^3}{2}M^{-1}\nabla V(q_0)\|}. \tag{3.28}$$

Thus,

$$g(q_0) = \left(\frac{tol}{\|\frac{h^2}{2}M^{-1}\nabla V(q_0)\|}\right)^{\frac{1}{4}}. \tag{3.29}$$

Experimentally, the fourth root did not affect results very much, but required messier computations, which is the reason why we have chosen the simpler yet very similar monitor function

$$g(q_0) = \frac{tol}{\|\frac{h^2}{2}M^{-1}\nabla V(q_0)\|}, \tag{3.30}$$

which achieves an error which is comparable to the chosen value of tol.

Alternative choices for the monitor function $g(q)$, proposed in [18], include the p-independent arclength parametrization

$$g(q) = (2(H_0 - V(q)) + \nabla V(q)^\top M^{-1}\nabla V(q))^{-\frac{1}{2}}, \tag{3.31}$$

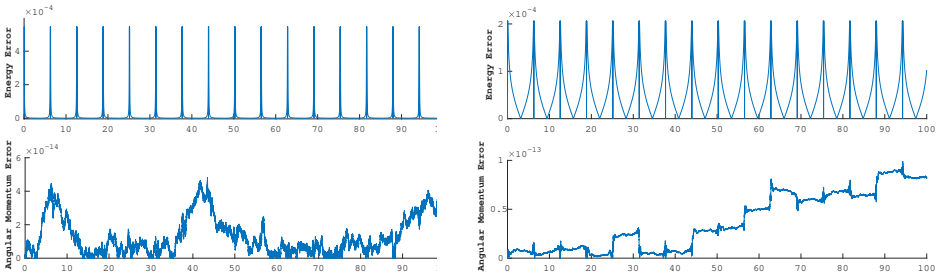
and a choice particular to Kepler's 2-body problem,

$$g(q) = q^\top q, \tag{3.32}$$

which is motivated by Kepler's second law, which states that a line segment joining the two bodies sweeps out equal areas during equal intervals of time.

We have tested the algorithm given by (3.26) on Kepler's planar 2-body problem, with an eccentricity of 0.9, using the three choices of monitor function g given by (3.30), (3.31), and (3.32). Of these three choices, (3.32) is specific to Kepler's 2-body problem, while (3.30) and (3.31) are more general choices. Unlike (3.31), which is independent of the order of the method, (3.30) is based on the truncation error, and thus the corresponding cost of computing this function will increase as the order of the method increases. Simulations using Kepler's 2-body problem with an eccentricity of 0.9 over a time interval of $[0, 1000]$ were run using the three different choices of g and the usual symplectic Euler-B. Results indicate that symplectic Euler-B takes the most steps and computational time to achieve a level of accuracy around 10^{-5}. To achieve this level of accuracy, the choice of the truncation error monitor function,

(3.30), resulted in the least number of steps and the second lowest computational time. The lowest computational time belonged to (3.32), but it used significantly more steps than (3.30). The lower computational cost can be attributed to the cheaper evaluation cost of the monitor function and its derivative. Finally, the monitor function (3.31) required the most steps and computational time of the adaptive algorithms, but it is still a good choice in general given its broad applicability. Figures 1 and 2 present the energy and angular momentum errors for the fixed time-step method versus the adaptive time-step method and the time-steps for the different monitor functions, respectively.



(a) Symplectic Euler-B (b) Adaptive method with monitor function (3.30)

FIG. 1. *Symplectic Euler-B and the adaptive algorithm with monitor function given by* (3.30) *were applied to Kepler's planar* 2*-body problem over a time interval of* $[0, 100]$ *with an eccentricity of* 0.9.

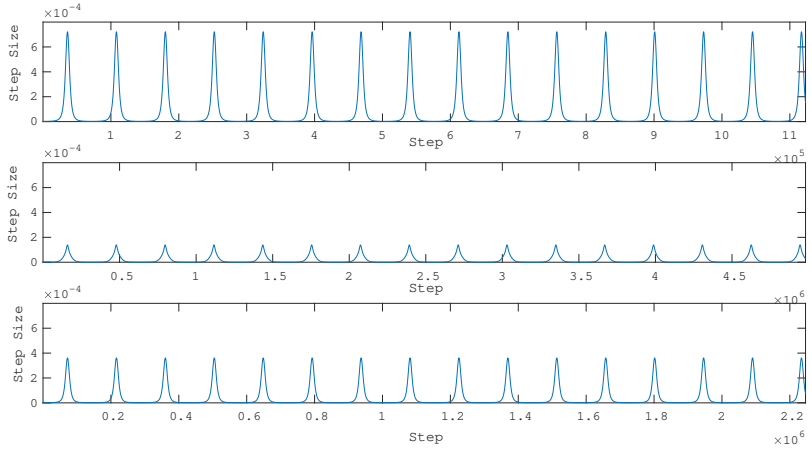


FIG. 2. *Time-steps taken for the various choices of monitor functions. The top, middle, and bottom plots correspond to the monitor functions* (3.30), (3.31), *and* (3.32), *respectively. All of the monitor functions appear to increase and decrease the time-step at the same points along the trajectory, but clearly* (3.30) *allowed for the larger steps to be taken.*

Next, we consider a Type II HTVI4 constructed using the strategy from section 2.2 and the automatic differentiation package from [38, 44]. We will drop the assumption of p-independent monitor functions and consider $g(q, p)$. The following monitor

functions were considered:

$$g(q) = \left(q^\top q\right)^\gamma \text{ for } \gamma = 0.5 \text{ and } 1 \qquad \text{(Gamma)}, \tag{3.33}$$

$$g(q) = \left(2(H_0 - V(q)) + \nabla V(q)^\top M^{-1} \nabla V(q)\right)^{-\frac{1}{2}} \qquad \text{(Arclength)}, \tag{3.34}$$

$$g(q,p) = \|p^t - L(q, M^{-1}p)\|_2^{-1} \qquad \text{(Energy)}. \tag{3.35}$$

The monitor function (3.35) was originally intended to be $\|p^t + H(q,p)\|_2^{-1}$, but experimental results suggested that (3.35) is the better choice. We will discuss the shortcomings of using the inverse energy error in the next paragraph. Note that $\|L(q, M^{-1}p)\|_2^{-1}$ also performs decently, but the addition of $p^t = -H(q_0, p_0)$ showed noticeable improvement. It was noted in [18] that the inverse Lagrangian has been considered as a possible choice for g in the Poincaré transformation but not in the framework of symplectic integration. While the choice of (3.33) was generally the most efficient, (3.35) was very close in terms of efficiency and offers a more general monitor function. This also implies that efficiency is not limited to only q- or p-independent monitor functions. However, various attempts to construct separable transformed Hamiltonians (see [5, 6]) required the use of q- or p-independent monitor functions, so this is where such monitor functions are most useful.

In the case of monitor functions involving the gradient, higher-order derivatives will be required for higher-order Taylor variational integrators, but there are efficiencies to be had when leveraging the higher-order derivatives already being calculated for the underlying Taylor method and Hessian-vector multiplication that can be done efficiently without needing to explicitly construct the full Hessian [10]. The calculation of higher-order derivatives do come with a higher cost, and in the case of Kepler's 2-body problem there is a clear computational advantage in using the gradient-free gamma monitor function (3.33), as shown in Tables 1 and 2. However, the gamma monitor function (3.33) is more specific to Kepler's 2-body, while the energy and arclength monitor functions are applicable to a wider range of problems. Monitor functions that are both general and efficient would be highly desirable.

Figure 3 displays the time-steps taken for the different choices of monitor functions for this HTVI4.

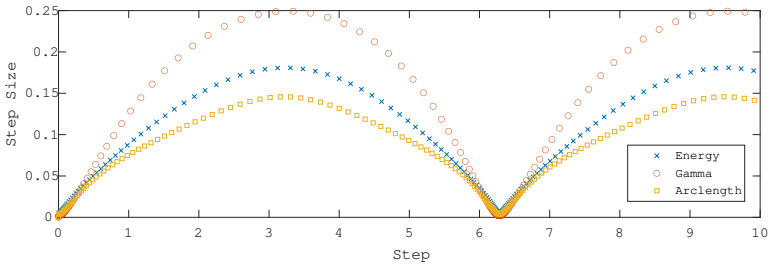


FIG. 3. *Time-steps taken for the various choices of monitor functions. The energy* (3.35) *and gamma* (3.33) *monitor functions performed better, in terms of fewest steps, lowest computational cost, and lowest global error, than the arclength monitor function* (3.34). *Note that* (3.35) *did not take the largest or the smallest steps.*

The truncation error monitor function, (3.30), performed quite well for first-order methods, and this motivated the choice of using Taylor variational integrators, since derivatives would be readily available. However, its success cannot as easily be applied to higher-order methods. This is due to the fact that for higher-order truncation

errors, one obtains an implicit differential-algebraic definition of the monitor function. This deviates from the first-order case, where the monitor function can be solved for explicitly. Another seemingly natural choice for the monitor function is the inverse of the energy error. However, Taylor variational integrators are constructed using Taylor expansions about the initial point, and consequently the monitor function is mostly evaluated at or near the initial point. If the initial point is at a particularly tricky part of the dynamics and requires a small first step, then the energy error at the first step will not reflect this since initially the energy error is zero. In contrast, the inverse Lagrangian will be small at an initial point that requires a small first step. The inverse energy error may work well for methods that primarily evaluate the energy error at the end point rather than the initial point. It is also often advantageous to bound the time-step below or above. As noted in [31, page 248], this can be done by setting $a = \frac{\Delta t_{\min}}{\Delta \tau}$ and $b = \frac{\Delta t_{\max}}{\Delta \tau}$ and defining the new monitor function as $\hat{g} = b\frac{g+a}{g+b}$.

Tables 1 and 2 display a comparison of bounds, computational time, steps, and error. We note that for methods such as the Taylor variational integrator, bounding $g(q,p)$ bounds the time-step, but not directly. Also, compared to nonadaptive variational integrators, such as the nonadaptive Taylor variational integrator and the Störmer–Verlet (SV) method, the adaptive methods showed a significant gain in efficiency for Kepler's 2-body planar problem with high eccentricity, while low eccentricity models do not need nor do they benefit from adaptivity. A Hamiltonian dynamical system with regions of high curvature in the vector field and its norm will in general benefit from an adaptive scheme such as the one outlined here.

TABLE 1
Kepler's planar 2-body problem; eccentricity $= 0.9$.

Method	Monitor $g(q,p)$	h	min step	max step	min g	max g	Energy error	Global error	Steps	Time
HTVI4	Gamma	0.1	0.0020	0.2493	0.01	8	1.43E-05	7.09E-06	181	26.9
HTVI4	Energy	0.1	0.0051	0.1809	1E-04	2	1.93E-06	4.76E-06	146	28.3
HTVI4	Arclength	0.1	0.0040	0.1458	3E-03	0.3	1.10E-04	3.69E-05	185	70.2
HTVI4	-	0.0025	0.0025	0.0025	-	-	2.50E-06	2.89E-05	4000	120
SV	-	5E-05	5E-05	5E-05	-	-	3.12E-06	4.68E-05	2E05	1.9

TABLE 2
Kepler's planar 2-body problem; eccentricity $= 0.99$.

Method	Monitor $g(q,p)$	h	min step	max step	min g	max g	Energy error	Global error	Steps	Time
HTVI4	Gamma	0.1	6E-05	0.2648	5E-04	8	4.88E-05	5.60E-06	372	49.3
HTVI4	Energy	0.03	1.5E-04	0.1462	1E-06	5	9.13E-06	4.63E-06	383	58.4
HTVI4	Arclength	0.1	5E-05	0.1379	8E-04	10	1.31E-05	1.49E-05	691	146.0
HTVI4	-	5E-04	5E-04	5E-04	-	-	1.38E-01	7.83E-01	2E04	525.2
SV	-	5E-07	5E-07	5E-07	-	-	3.34E-06	2.68E-05	2E07	189.2

4. Application to symplectic accelerated optimization.

4.1. Accelerated optimization. Efficient optimization has become one of the major concerns in data analysis. Many machine learning algorithms are designed around the minimization of a loss function or the maximization of a likelihood function. Due to the ever-growing scale of the data sets and size of the problems, there has been a lot of focus on first-order optimization algorithms because of their low cost per iteration. The first gradient descent algorithm was proposed in [9] by Cauchy to deal with the very large systems of equations he was facing when trying to simulate orbits of celestial bodies, and many gradient-based optimization methods have

been proposed since Cauchy's work in 1847. In 1983, Nesterov's accelerated gradient method was introduced in [41],

$$x_k = y_{k-1} - h\nabla f(y_{k-1}), \qquad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}), \tag{4.1}$$

which converges in $\mathcal{O}(1/k^2)$ to the minimum of the convex objective function f, improving on the $\mathcal{O}(1/k)$ convergence rate exhibited by the standard gradient descent methods. This $\mathcal{O}(1/k^2)$ convergence rate was shown in [42] to be optimal among first-order methods using only information about ∇f at consecutive iterates. This phenomenon in which an algorithm displays this improved rate of convergence is referred to as acceleration, and other accelerated algorithms have been derived since Nesterov's algorithm, such as accelerated mirror descent [40] and accelerated cubic-regularized Newton's method [43]. More recently, it was shown in [53] that Nesterov's accelerated gradient method limits to the second order ODE,

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \tag{4.2}$$

as $h \to 0$. The authors also proved that the objective function $f(x(t))$ converges to its optimal value at a rate of $\mathcal{O}(1/t^2)$ along the trajectories of this ODE. It was then shown in [58] that in continuous time, the convergence rate of $f(x(t))$ can be accelerated to an arbitrary convergence rate $\mathcal{O}(1/t^p)$, by considering flow maps generated by time-dependent Lagrangian and Hamiltonian systems. We will present this result in more detail in the next section together with the variational framework introduced in [58] for accelerated optimization, which will be at the heart of our approach.

4.2. Variational framework for accelerated optimization. In this section, we will review the variational framework introduced in [58] for accelerated optimization which will be the basis for the methods we will design. In a general space $\mathcal{X}$, given a convex, continuously differentiable function $h : \mathcal{X} \to \mathbb{R}$ such that $\|\nabla h(x)\| \to \infty$ as $\|x\| \to \infty$, its corresponding Bregman divergence is given by

$$D_h(x, y) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle. \tag{4.3}$$

We then define the Bregman Lagrangian and Hamiltonian

$$\mathcal{L}_{\alpha,\beta,\gamma}(x, v, t) = e^{\alpha(t)+\gamma(t)} \left[D_h(x + e^{-\alpha(t)}v, x) - e^{\beta(t)} f(x) \right], \tag{4.4}$$

$$\mathcal{H}_{\alpha,\beta,\gamma}(x, r, t) = e^{\alpha(t)+\gamma(t)} \left[D_{h^*}(\nabla h(x) + e^{-\gamma(t)}r, \nabla h(x)) + e^{\beta(t)} f(x) \right], \tag{4.5}$$

which are scalar valued functions of position $x \in \mathcal{X}$, velocity $v \in \mathbb{R}^d$, momentum $r \in \mathbb{R}^d$, and time t, which are parametrized by smooth functions of time, α, β, γ, and where $h^* = \sup_{v \in \mathcal{X}} [\langle r, v \rangle - h(v)]$ is the Legendre transform (or convex dual function) of h. These parameters α, β, γ are said to satisfy the ideal scaling conditions if

$$\dot{\beta}(t) \leq e^{\alpha(t)} \qquad \text{and} \qquad \dot{\gamma}(t) = e^{\alpha(t)}. \tag{4.6}$$

If the ideal scaling conditions are satisfied, then by Theorem 1.1 in [58],

$$f(x(t)) - f(x^*) \leq \mathcal{O}(e^{-\beta(t)}). \tag{4.7}$$

Another very important property of this family of Bregman Lagrangians is its closure under time-dilation, proven in Theorem 1.2 of [58].

THEOREM 4.1. *If $x(t)$ satisfies the Euler–Lagrange equations corresponding to the Bregman Lagrangian $\mathcal{L}_{\alpha,\beta,\gamma}$, then the reparametrized curve $y(t) = x(\tau(t))$ satisfies the Euler–Lagrange equations corresponding to the Bregman Lagrangian $\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}$, where*

$$\tilde{\alpha}(t) = \alpha(\tau(t)) + \log \dot{\tau}(t), \qquad \tilde{\beta}(t) = \beta(\tau(t)), \qquad \tilde{\gamma}(t) = \gamma(\tau(t)), \tag{4.8}$$

and

$$\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}(x, v, t) = \dot{\tau}(t)\mathcal{L}_{\alpha,\beta,\gamma}\left(x, \frac{1}{\dot{\tau}(t)}v, \tau(t)\right). \tag{4.9}$$

Furthermore, α, β, γ satisfy the ideal scaling equation (4.6) *if and only if $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ do.*

A subfamily of Bregman Lagrangians of interest, indexed by a parameter $p > 0$, is given by the choice of functions

$$\alpha(t) = \log p - \log t, \qquad \beta(t) = p \log t + \log C, \qquad \gamma(t) = p \log t, \tag{4.10}$$

where $C > 0$ is a constant. The Bregman Lagrangian and Hamiltonian become

$$\mathcal{L}(x, v, t) = pt^{p-1}\left[D_h\left(x + \frac{t}{p}v, x\right) - Ct^p f(x)\right], \tag{4.11}$$

$$\mathcal{H}(x, r, t) = pt^{p-1}\left[D_{h^*}(\nabla h(x) + t^p r, \nabla h(x)) + Ct^p f(x)\right]. \tag{4.12}$$

These parameter functions are of interest since they satisfy the ideal scaling equation (4.6), and the resulting evolution $x(t)$ of the corresponding dynamical system was shown in [58] to satisfy the aforementioned $\mathcal{O}(1/t^p)$ convergence rate,

$$f(x(t)) - f(x^*) \leq \mathcal{O}(1/t^p), \tag{4.13}$$

where x^* is the desired minimizer of the objective function f.

4.3. Adaptive variational integrators for symplectic accelerated optimization. For simplicity of exposition, we will consider the case where $h(x) = \frac{1}{2}\langle x, x\rangle$. Our new approaches will make use of the adaptive framework developed in section 3 via carefully chosen Poincaré transformations. Recalling the discussion of section 3.1, there might not be a Lagrangian formulation for the future Poincaré transformed systems, so we will need to work from the Hamiltonian point of view to design variational integrators. When $h(x) = \frac{1}{2}\langle x, x\rangle$, the Bregman Hamiltonian with parameters α, β, γ given by (4.10) for a specific value of $p > 0$ becomes

$$\mathcal{H}(q, r, t) = \frac{p}{2t^{p+1}}\langle r, r\rangle + Cpt^{2p-1}f(q). \tag{4.14}$$

As mentioned in the previous section, the solution to the corresponding Hamilton's equations was shown in [58] to satisfy the convergence rate

$$f(q(t)) - f(q^*) \leq \mathcal{O}(1/t^p), \tag{4.15}$$

where q^* is the desired minimizer of the objective function f. Together with the time-dilation result from Theorem 4.1, this implies that this entire subfamily of Bregman trajectories indexed by the parameter p can be obtained by speeding up or slowing

down along the Bregman curve in spacetime corresponding to any specific value of p. We will present two new approaches based on the adaptive framework developed in section 3 to integrate the Bregman Hamiltonian dynamics, thereby solving the optimization problem, and then compare their performance.

Direct approach. Our first approach will use our adaptive framework with monitor function $g(q,r) = 1$ to design a variational integrator for the Bregman Hamiltonian given in (4.14) for a given value of $p > 0$. This choice of monitor function will convert the time-dependent Bregman Hamiltonian into an autonomous Hamiltonian in extended phase space. More precisely, given a value of $p > 0$, the time transformation $t \mapsto \tau$ given by $\frac{dt}{d\tau} = g(q,t,r) = 1$ generates the Poincaré transformed Hamiltonian,

$$\bar{H}(\bar{q},\bar{r}) = \frac{p}{2(q^t)^{p+1}}\langle r, r\rangle + Cp(q^t)^{2p-1}f(q) + r^t, \tag{4.16}$$

in the phase space with extended coordinates $(\bar{q},\bar{r})$. This strategy is equivalent to the usual trick to remove time-dependency by considering time t as an additional position variable and adding a corresponding conjugate momentum variable, which is the energy (see [2] for an example with Hamiltonian given by (4.5)). This shows that our adaptive framework is very general and can also be used for purposes other than solely enforcing a desired variable time-stepping.

Adaptive approach. Our second approach will exploit the time-dilation property of the Bregman dynamics together with our adaptive framework with a carefully tuned monitor function. More precisely, we will use adaptivity to transform the Bregman Hamiltonian corresponding to a specific value of $p > 0$ into an autonomous version of the Bregman Hamiltonian corresponding to a smaller value $\mathring{p} < p$ in extended phase space. This will allow us to integrate the higher-order Bregman dynamics corresponding to the value p while benefiting from the computational efficiency of integrating the lower-order Bregman dynamics corresponding to the value $\mathring{p} < p$. Explicitly, solving (4.8) for $\tau(t)$ to transform the Bregman dynamics corresponding to the values of α, β, γ as in (4.10) for a given value of p into the Bregman dynamics corresponding to the values of $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ as in (4.10) for a given value $\mathring{p} < p$ yields $\tau(t) = t^{\mathring{p}/p}$. The corresponding monitor function is given by

$$\frac{dt}{d\tau} = g(q,t,r) = \frac{p}{\mathring{p}}t^{1-\frac{\mathring{p}}{p}} \tag{4.17}$$

and generates the Poincaré transformed Hamiltonian

$$\bar{H}(\bar{q},\bar{r}) = \frac{1}{\mathring{p}}\left[\frac{p^2}{2(q^t)^{p+\frac{\mathring{p}}{p}}}\langle r, r\rangle + Cp^2(q^t)^{2p-\frac{\mathring{p}}{p}}f(q) + pr^t(q^t)^{1-\frac{\mathring{p}}{p}}\right]. \tag{4.18}$$

4.4. Presentation of numerical methods. In this section, we will test the performance of our new adaptive framework by implementing variational and nonvariational integrators in the case where $\mathcal{X} = \mathbb{R}^d$ and $\langle x, x\rangle = x^\top x$, and we will discuss the results obtained. Keeping in mind the machine learning application where data sets are very large, we will restrict ourselves to explicit first-order optimization algorithms.

Looking at the forms of Hamilton's equations in both the direct and the adaptive approaches, we note that the objective function f and its gradient ∇f only appear in the expression for $\dot{\bar{r}}$ and are functions of q only. Looking back at the construction of HTVIs from section 2.2, denoting the order of the Taylor methods by ρ instead of r to avoid confusion with the extended momentum variable $\bar{r}$, we note that both the Type II and the Type III approaches require ρ-order Taylor approximations of r and $(\rho+1)$-order Taylor approximations of q. This means that the highest value of ρ that we can choose to obtain a gradient-based algorithm is $\rho = 1$. Now, the starting point of the Type III approach is a $(\rho+1)$-order Taylor approximation $\tilde{q}_0$ of q_0 around q_0. As a consequence, the subsequent steps in the Type III method with $\rho = 0$ and $\rho = 1$ will contain evaluations of the objective function f and its gradient ∇f at this approximation $\tilde{q}_0$. Aside from the inconvenience of the function evaluations not being at the iterates q_0 and q_1 themselves, if f is a nonlinear function, this will also generate nonlinearity in the equations for the updates. As a result, we will not be able to design an explicit algorithm, or at least not a general explicit algorithm that would work for all the functions f considered. On the other hand, the starting point of the Type II approach is a ρ-order Taylor approximation $\tilde{p}_0$ of p_0 around p_0. A similar issue as for the Type III case arises when $\rho = 1$ due to the approximations $(q_{c_i}, p_{c_i}) = \Psi^{(\rho)}_{c_i h}(\tilde{q}_0, p_0)$. Therefore, we cannot design a general explicit algorithm for the Type II case with $\rho = 1$. The remaining possibility is to construct a Type II HTVI using $\rho = 0$. This will produce explicit gradient-based algorithms, where all the evaluations of the objective function f and its gradient ∇f are performed at the iterates q_0 and q_1. Note that when $\rho = 0$, we have $(q_{c_i}, p_{c_i}) = \Psi^{(0)}_{c_i h}(\tilde{q}_0, p_0) = (\tilde{q}_0, p_0)$ for all i, so for given values of p and $\mathring{p}$, every quadrature rule generates the same integrator. Following the method of section 2.2, with time-step h, we will derive explicit gradient-based HTVIs.

Type II Hamiltonian Taylor variational integrators (HTVIs) with $\rho = 0$. As mentioned earlier, since $\rho = 0$, the choice of quadrature rule does not matter, so we can take the rectangular quadrature rule about the initial point ($c_1 = 0$ and $b_1 = 1$). We approximate $\bar{r}(0) = \tilde{\bar{r}}_0$ via $\bar{r}_1 = \pi_{T^*\bar{Q}} \circ \Psi^{(0)}_h(q_0, \tilde{\bar{r}}_0) = \tilde{\bar{r}}_0$ and generate approximations $(\bar{q}_{c_1}, \bar{r}_{c_1}) = \Psi^{(0)}_{c_1 h}(\bar{q}_0, \tilde{\bar{r}}_0) = (\bar{q}_0, \tilde{\bar{r}}_0)$.

Direct approach	**Adaptive approach**
$\tilde{\bar{q}}_1 = \pi_{\bar{Q}} \circ \Psi^{(1)}_h(\bar{q}_0, \tilde{\bar{r}}_0) = \begin{bmatrix} q_0 + h\frac{pr_0}{(q_0^t)^{p+1}} \\ q_0^t + h \end{bmatrix}$,	$\tilde{\bar{q}}_1 = \pi_{\bar{Q}} \circ \Psi^{(1)}_h(\bar{q}_0, \tilde{\bar{r}}_0) = \begin{bmatrix} q_0 + h\frac{p^2 r_0}{\mathring{p}}(q_0^t)^{-p-\frac{\mathring{p}}{p}} \\ q_0^t + h\frac{p}{\mathring{p}}(q_0^t)^{1-\frac{\mathring{p}}{p}} \end{bmatrix}$,
$H_d^+(\bar{q}_0, \bar{r}_1; h) = r_1^\top q_0 + r_1^t q_0^t + h\frac{p}{2(q_0^t)^{p+1}} r_1^\top r_1 + hCp(q_0^t)^{2p-1} f(q_0) + hr_1^t$.	$H_d^+(\bar{q}_0, \bar{r}_1; h) = r_1^\top q_0 + r_1^t q_0^t + h\frac{p^2}{2\mathring{p}(q_0^t)^{p+\frac{\mathring{p}}{p}}} r_1^\top r_1 + hC\frac{p^2}{\mathring{p}}(q_0^t)^{2p-\frac{\mathring{p}}{p}} f(q_0) + h\frac{p}{\mathring{p}}(q_0^t)^{1-\frac{\mathring{p}}{p}} r_1^t$.
The discrete right Hamilton's equations (2.5) yield the explicit variational integrator	The discrete right Hamilton's equations (2.5) yield the explicit variational integrator
$r_1 = r_0 - hCp(q_0^t)^{2p-1}\nabla f(q_0)$, $r_1^t = r_0^t + h\frac{p(p+1)}{2(q_0^t)^{p+2}} r_1^\top r_1 - hCp(2p-1)(q_0^t)^{2p-2} f(q_0)$, $q_1 = q_0 + h\frac{p}{(q_0^t)^{p+1}} r_1$, $q_1^t = q_0^t + h$.	$r_1 = r_0 - \frac{p^2}{\mathring{p}} hC(q_0^t)^{2p-\frac{\mathring{p}}{p}} \nabla f(q_0)$, $r_{1/2}^t = \frac{p^3 + p\mathring{p}}{2\mathring{p}(q_0^t)^{p+\frac{\mathring{p}}{p}+1}} hr_1^\top r_1 + \frac{p\mathring{p} - 2p^3}{\mathring{p}(q_0^t)^{\frac{\mathring{p}}{p}+1-2p}} hCf(q_0)$, $r_1^t = \left[1 - h(q_0^t)^{-\frac{\mathring{p}}{p}}\left(1 - \frac{p}{\mathring{p}}\right)\right]^{-1}\left(r_0^t + r_{1/2}^t\right)$, $q_1 = q_0 + \frac{p^2}{\mathring{p}} h(q_0^t)^{-p-\frac{\mathring{p}}{p}} r_1$, $q_1^t = q_0^t + \frac{p}{\mathring{p}} h(q_0^t)^{1-\frac{\mathring{p}}{p}}$.

Other types of first-order variational integrators can be constructed for Poincaré transformed Hamiltonians, such as prolongation-collocation variational integrators [32], Galerkin variational integrators [33], and higher-order HTVIs. We will not consider these integrators here since they require that one solves systems of nonlinear equations and cannot be implemented explicitly in general. Having said that, in practice, implicit methods for the numerical solution of ODEs that can be solved using fixed-point iterations (as opposed to Newton iterations) can be quite competitive, as there is a good initial guess which may allow them to converge in a small number of iterations, and the iterations are inexpensive, as they do not require the assembly of a Jacobian. The convergence of the fixed-point iteration depends on the conditioning of the system of equations and may impose a stringent time-step restriction. This can be overcome by the use of exponential integrators [21], in particular, symplectic and energy-preserving exponential integrators [51].

We have also implemented nonvariational methods based on these direct and adaptive approaches (the classical 4th-order explicit Runge–Kutta method and the explicit adaptive ODE solvers (`ode23`, `ode45`) of MATLAB), and more traditional optimization methods have been tested as well, such as Nesterov's accelerated gradient (4.1) and adaptive optimization algorithms.

Nonvariational symplectic integrators based on the direct and adaptive approaches.

(i) **Direct and adaptive approaches with splitting of the Hamiltonian**. The direct approach with splitting of the Hamiltonian is the approach presented in [2]. The three terms of the Poincaré transformed Hamiltonian (4.16) are considered separately. They generate dynamics in the extended phase space via six vector fields, and a symmetric leapfrog composition of the corresponding component dynamics is constructed to obtain a symplectic integrator (referred to in the numerical results section as "direct splitting"). A new symplectic integrator can also be obtained by adapting the approach presented in [2] to the adaptive Poincaré transformed Hamiltonian (4.18) to obtain a symplectic integrator (referred to in the numerical results section as "adaptive splitting").

Direct approach:

$$
\begin{aligned}
t &= t + \frac{h}{2},\\
r^t &= r^t + \frac{h}{2}\frac{p(p+1)}{2t^{p+2}} r^\top r - \frac{h}{2} Cp(2p-1)t^{2p-2} f(q),\\
r &= r - \frac{h}{2} Cpt^{2p-1} \nabla f(q),\\
q &= q + h\frac{p}{t^{p+1}} r,\\
r &= r - \frac{h}{2} Cpt^{2p-1} \nabla f(q),\\
r^t &= r^t + \frac{h}{2}\frac{p(p+1)}{2t^{p+2}} r^\top r - \frac{h}{2} Cp(2p-1)t^{2p-2} f(q),\\
t &= t + \frac{h}{2}.
\end{aligned}
$$

Adaptive approach:

$$
\begin{aligned}
t &= \left(t^{\mathring{p}/p} + \frac{h}{2}\right)^{p/\mathring{p}},\\
\theta &= \frac{\mathring{p}}{\mathring{p}-p}\left[\left(\frac{p^3}{2\mathring{p}} + \frac{p}{2}\right) t^{-p-1} r^\top r + \left(p - \frac{2p^3}{\mathring{p}}\right) t^{2p-1} C f(q)\right],\\
r^t &= (r^t + \theta)\exp\left(\left(1 - \frac{p}{\mathring{p}}\right)\frac{h}{2} t^{-\mathring{p}/p}\right) - \theta,\\
r &= r - h\frac{Cp^2}{2\mathring{p}} t^{2p-\mathring{p}/p}\nabla f(q),\\
q &= q + h\frac{p^2}{\mathring{p}t^{p+\mathring{p}/p}} r,\\
r &= r - h\frac{Cp^2}{2\mathring{p}} t^{2p-\mathring{p}/p}\nabla f(q),\\
\theta &= \frac{\mathring{p}}{\mathring{p}-p}\left[\left(\frac{p^3}{2\mathring{p}} + \frac{p}{2}\right) t^{-p-1} r^\top r + \left(p - \frac{2p^3}{\mathring{p}}\right) t^{2p-1} C f(q)\right],\\
r^t &= (r^t + \theta)\exp\left(\left(1 - \frac{p}{\mathring{p}}\right)\frac{h}{2} t^{-\mathring{p}/p}\right) - \theta,\\
t &= \left(t^{\mathring{p}/p} + \frac{h}{2}\right)^{p/\mathring{p}}.
\end{aligned}
$$

(ii) **Phase-space cloning and splitting**. A very natural approach to integrate these nonseparable Hamiltonian dynamics consists in defining a new Hamiltonian via two copies of the Poincaré transformed Hamiltonian in an extended phase space of dimension twice as large [46]:

$$\tilde{H}(\bar{q}, \tilde{\bar{q}}, \bar{r}, \tilde{\bar{r}}) = \bar{H}_1(\bar{q}, \tilde{\bar{r}}) + \bar{H}_2(\tilde{\bar{q}}, \bar{r}), \tag{4.19}$$

where $\bar{H}_1 = \bar{H}_2 = \bar{H}$. Hamilton's equations are then given by

$$\dot{\bar{q}} = \nabla_{\bar{r}}\bar{H}_2, \qquad \dot{\tilde{\bar{q}}} = \nabla_{\tilde{\bar{r}}}\bar{H}_1, \qquad \dot{\bar{r}} = -\nabla_{\bar{q}}\bar{H}_1, \qquad \dot{\tilde{\bar{r}}} = \nabla_{\tilde{\bar{q}}}\bar{H}_2. \tag{4.20}$$

We can then integrate this new Hamiltonian system explicitly using a Strang splitting or a Yoshida 4 or Yoshida 6 splitting, for instance (referred to as "CloningStrang," "CloningY4," and "CloningY6" in the numerical results section). This approach will usually require a larger number of evaluations of the objective function f and of its gradient at each step.

4.5. Numerical results. The numerical methods presented in the previous section have been conducted to minimize the quartic function

$$f(x) = \left[(x-1)^\top \Sigma (x-1)\right]^2, \tag{4.21}$$

where $x \in \mathbb{R}^{50}$ and $\Sigma_{ij} = 0.9^{|i-j|}$. This convex function achieves its minimum value 0 at $x^* = 1$.

Unless specified otherwise, the termination criterion used was

$$|f(x_k) - f(x_{k-1})| < \delta \quad \text{and} \quad \|\nabla f(x_k)\| < \delta, \qquad \text{where} \quad \delta = 10^{-10}. \tag{4.22}$$

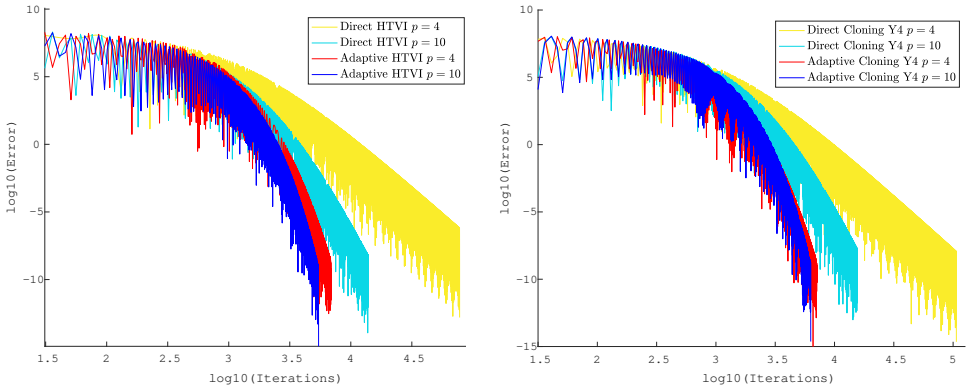


FIG. 4. *Comparison of the rates of convergence between the direct and adaptive approaches for the HTVI method and the cloning method with a Yoshida* 4 *splitting. We can clearly see that the adaptive approach outperforms the direct approach.*

TABLE 3

Comparison of the direct and adaptive approaches for the HTVI method and the cloning method with a Yoshida 4 *splitting. The adaptive approach clearly outperforms the direct approach in terms of number of iterations required.*

Approach	p	$\mathring{p}$	h	Iterations	Approach	p	$\mathring{p}$	h	Iterations
Direct HTVI	4	-	8.00E-04	77 878	Direct CloneY4	4	-	9.00E-04	106 530
Adap. HTVI	4	0.5	1.21E-04	6 867	Adap. CloneY4	4	0.5	1.15E-04	7 101
Direct HTVI	10	-	4.00E-04	13 872	Direct CloneY4	10	-	3.30E-04	15 498
Adap. HTVI	10	0.5	1.95E-05	5 564	Adap. CloneY4	10	0.5	1.62E-05	6 300

4.5.1. Adaptive versus direct approach. Numerical experiments conducted with all the symplectic algorithms presented in section 4.4 showed that a carefully tuned adaptive approach enjoys a significantly better rate of convergence and a much smaller number of steps required to achieve convergence than the direct approach, as can be seen in Figure 4 and Table 3 for the HTVI and CloningY4 methods. Although the adaptive approach requires a smaller fictive time-step h than the direct approach, the physical time-steps resulting from $t = \tau^{p/\mathring{p}}$ in the adaptive approach grow rapidly to values larger than the physical time-step of the direct approach.

4.5.2. Comparison of methods within direct and adaptive approaches. Numerical experiments were conducted to compare the various algorithms presented in section 4.4, and the results are presented in Figure 5 and Table 4. Although the number of iterations for all methods were of the same order of magnitude, the HTVI method and the splitting method based on the idea of [2] performed much better than the methods based on the phase-space cloning idea of [46]. This is mostly due to the fact that these phase-space cloning methods require several evaluations of the objective function f and of its gradient ∇f at each iteration (3 for Strang splitting, 7 for Yoshida's 4th order splitting, and 19 for Yoshida's 6th order splitting), while the HTVI and splitting methods only required one such evaluation at each iteration. As a result, these phase-space cloning methods also required much more computational time to achieve convergence. It might be possible to improve the performance of these phase-space cloning methods by adding an extra term in the final Hamiltonian which binds the two copies of the Poincaré transformed Hamiltonian, as was done in [55].

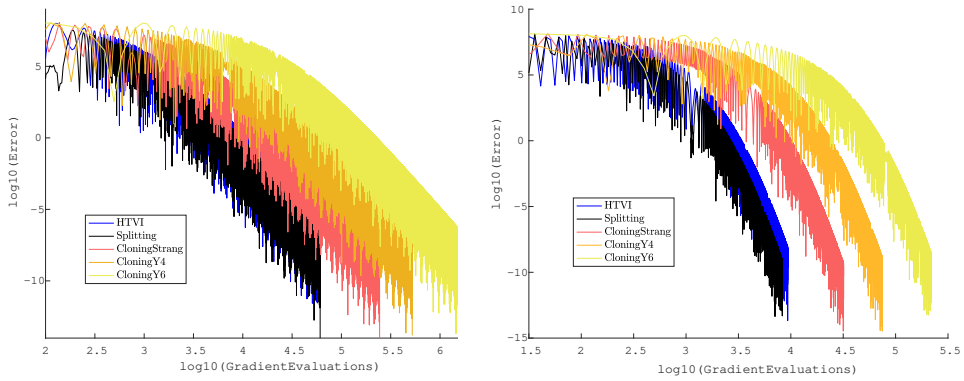


FIG. 5. *Comparison of the convergence, in terms of gradient evaluations needed, of the different symplectic integrators within the direct (on the left) and adaptive (on the right) approaches.*

TABLE 4
Number of iterations needed until convergence of the different symplectic integrators within the direct (on the left) and adaptive (on the right) approaches.

Method	p	h	Iterations	Method	p	$\mathring{p}$	h	Iterations
Direct HTVI	4	8.7E-04	57 504	Adaptive HTVI	4	1	2.4E-04	9 361
Direct splitting	4	9.5E-04	60 881	Adaptive splitting	4	1	2.9E-04	8 313
Direct CloneStrang	4	9.7E-04	81 367	Adaptive CloneStrang	4	1	2.4E-04	10 638
Direct CloneY4	4	8.9E-04	74 747	Adaptive CloneY4	4	1	2.9E-04	10 721
Direct CloneY6	4	7.9E-04	87 075	Adaptive CloneY6	4	1	2.0E-04	11 549

However, this additional term is likely to require a larger number of compositions when splitting the final Hamiltonian, which would require more gradient evaluations of the objective function at each step. Thus, even though the trick presented in [55] could reduce the number of iterations required to achieve convergence, it seems very unlikely that the resulting algorithm would be competitive against the HTVI and splitting methods, in terms of computational time and total number of gradient evaluations needed.

4.5.3. Dependence on p and $\mathring{p}$ in the adaptive approach. We conducted numerical experiments with the HTVI method to study the evolution of the performance of the adaptive approach as the parameters p and $\mathring{p}$ are varied. We can see from the results presented in Figure 6 and Table 5 that the adaptive HTVI method becomes more and more efficient as p is increased and $\mathring{p}$ is decreased. The improvement in efficiency is very important as we increase p from $p = 2$ to $p = 4$, while it is minor but still noticeable as we increase p from $p = 4$ to $p = 8$. A possible explanation for this behavior is that the integrator might not be of high enough order to distinguish between the $p = 6$ and $p = 8$ Bregman dynamics. Note that the fictive time-step h must be reduced as p increases or $\mathring{p}$ decreases, but the time relation $t = \tau^{p/\mathring{p}}$ ensures that the resulting physical time-steps do not become significantly smaller.

4.5.4. Comparison to nonsymplectic integrators. We will now present the results of numerical experiments investigating the role that symplecticity plays when integrating the Bregman dynamics in the direct and adaptive approaches.

We first implemented fixed time-step integrators such as the 4th-order explicit Runge–Kutta method, but these failed to converge both in the direct and adaptive

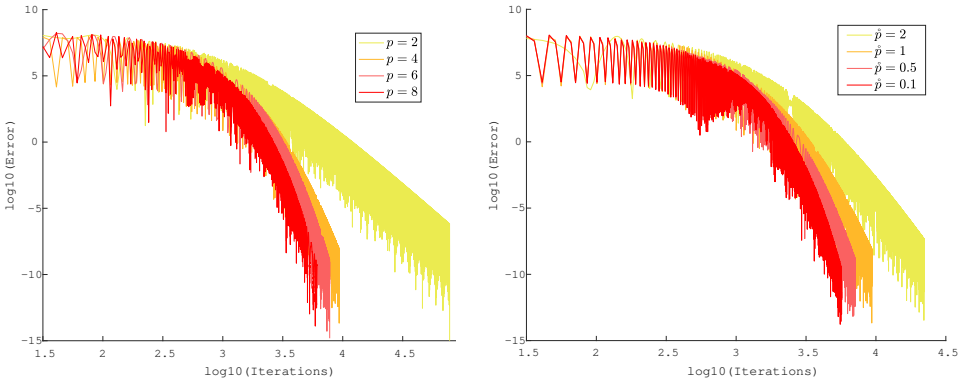


FIG. 6. *Evolution of the rates of convergence of the HTVI method as p is increased (on the left) and as $\mathring{p}$ is decreased (on the right).*

TABLE 5

Evolution of the fictive time-step h and number of iterations until convergence for the HTVI method as p increases (left) and as $\mathring{p}$ decreases (right).

Method	p	$\mathring{p}$	h	Iterations	Method	p	$\mathring{p}$	h	Iterations
Adaptive HTVI	2	1	8.0E-04	77 855	Adaptive HTVI	4	2	3.8E-04	22 128
Adaptive HTVI	4	1	2.4E-04	9 361	Adaptive HTVI	4	1	2.4E-04	9 361
Adaptive HTVI	6	1	9.4E-05	7 785	Adaptive HTVI	4	0.5	1.2E-04	7 099
Adaptive HTVI	8	1	6.1E-05	6 133	Adaptive HTVI	4	0.1	2.4E-05	5 689

approaches. The reason why convergence cannot be achieved may have to do with the nonautonomous aspect of the differential equation. More precisely, explicit Runge–Kutta methods are conditionally stable, where stability intervals for the time-steps depend on the expansivity of the differential equation. Since the differential equations considered here are not autonomous, the stability intervals are time-dependent, and thus any fixed choice of time-step may eventually violate the stability condition. It might be possible to achieve low accuracy convergence using these methods, but the fact that they cannot achieve higher accuracy and are likely to lose stability eventually makes them undesirable.

We then considered variable time-step explicit Runge–Kutta methods. To this end, we tested the differential equation solvers `ode45` and `ode23` of MATLAB, which are explicit variable time-step Runge–Kutta pairs, and the corresponding numerical results are presented in Figure 7. The HTVI method required a significantly smaller number of iterations than the MATLAB solvers. Furthermore, an inherent part of the time-step control in embedded Runge–Kutta methods is that, at each iteration, the underlying Runge–Kutta method may be executed several times to determine the appropriate time-step that satisfies the prescribed tolerances. For this reason, the MATLAB solvers require more evaluations of f and ∇f at each iteration, and since they also required more iterations than the HTVI method, these MATLAB solvers are much less competitive.

It should also be noted that the MATLAB solvers did not exhibit any improvements when used with the adaptive approach instead of the direct approach, while the HTVI method improved significantly. This is not surprising since the MATLAB solvers `ode23` and `ode45` both use a variable time-step strategy, regardless of the

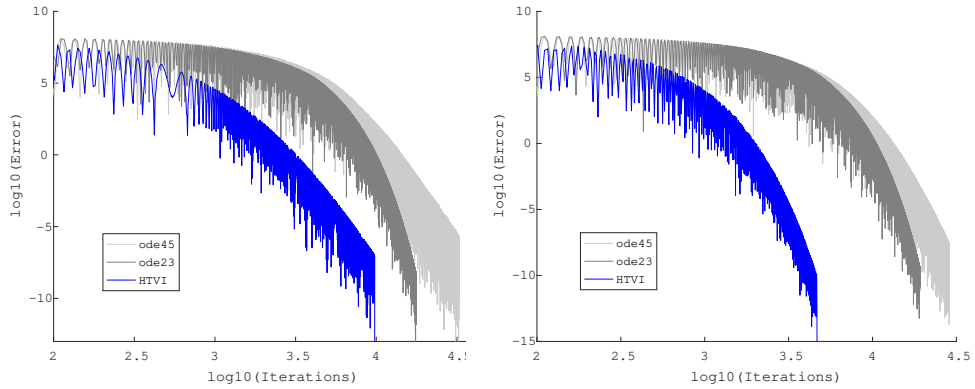


FIG. 7. *Comparison of the HTVI method with the* `ode23` *and* `ode45` *MATLAB functions in the direct (left) and adaptive (right) approaches with* $p = 10$ *and* $\mathring{p} = 0.5$. *The HTVI method outperforms the MATLAB solvers.*

approach chosen.

Note that our adaptive approach and the embedded Runge–Kutta methods use adaptivity in two fundamentally different ways. Our approach uses a priori adaptivity based on known global properties of the family of differential equations considered (i.e., the time-dilation symmetry of the family of Bregman dynamics). In contrast, embedded Runge–Kutta methods use adaptivity based on a posteriori local error estimates. This could explain why the embedded Runge–Kutta methods do not perform as well as our adaptive approach: a posteriori estimators might focus mostly on the fast local oscillations of the Bregman dynamics and not on the slower global decay, and these fast oscillations might be forcing the embedded Runge–Kutta methods to adaptively take smaller time-steps than necessary.

We can also see from Figure 7 that for both the symplectic and the nonsymplectic adaptive methods, a significant number of iterations are needed before the error effectively starts decaying. The fact that this slow initial behavior persists with those two approaches, which use time-adaptivity in the two fundamentally different ways described in the previous paragraph, suggests that this behavior might be intrinsic to the continuous trajectory being discretized and that time-adaptivity might not be able to help accelerate this initial phase.

4.5.5. Comparison to popular optimization methods. Finally, we have compared the performance of our adaptive HTVI method to Nesterov's accelerated gradient (NAG) (4.1) with the same initial time-step $h = 2 \times 10^{-6}$, and to popular adaptive optimization algorithms such as trust region steepest descent (TRUST), ADAM [27], AdaGrad [12], and RMSprop [57].

Figure 8 and Table 6 present the numerical results obtained when applying these algorithms to the quartic objective function (4.21). Although the adaptive HTVI method is not the most efficient method, we can see that it significantly outperformed certain popular optimization algorithms on this particular convex problem. This suggests that the adaptive HTVI method might be a competitive first-order explicit algorithm, and that it might be worth considering it as one of several possible options to use in practice, as the relative performance often depends on the specific choice of objective function.

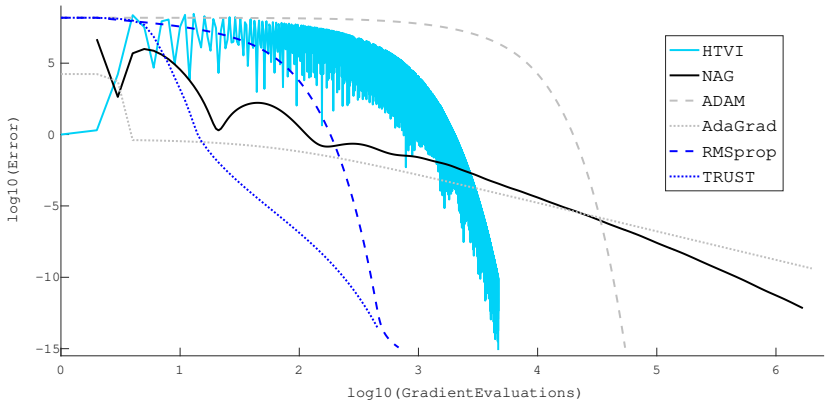


FIG. 8. *Comparison of HTVI, NAG, and other adaptive optimization algorithms to achieve convergence on the quartic objective function* (4.21)*, with different values of* δ *as termination criterion* (4.22)*. Note that HTVI and NAG were implemented with the same initial time-step.*

TABLE 6

Comparison of the number of iterations needed for HTVI, for NAG, and for other adaptive optimization algorithms to achieve convergence on the quartic objective function (4.21)*, with different values of* δ *as termination criterion* (4.22)*. Note that HTVI and NAG were implemented with the same initial time-step. For all these algorithms, the number of gradient evaluations equals the number of iterations.*

$\delta =$	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
HTVI	2 182	2 486	2 750	3 233	3 434	3 593	4 014	4 097	4 566
NAG	4 143	10 949	27 660	65 724	154 258	341 928	745 292	1.7E6	>1E10
ADAM	29 665	32 733	35 802	38 871	41 939	45 008	48 076	51 145	54 215
AdaGrad	520 482	2.4E06	1.1E07	5.2E07	2.4E08	—	—	—	—
RMSprop	276	305	334	363	393	422	452	498	682
TRUST	32	48	71	106	154	215	288	366	455

Remark 4.2. In [2], the authors noted that Nesterov's accelerated gradient algorithm transitions into an exponential rate of convergence once it is sufficiently close to the minimum of certain objective functions and suggested that this behavior requires strong convexity of the objective function in the neighborhood of the minimum. Similarly to the strategy presented in [2], a gradient flow can be incorporated into the updates of the direct and adaptive algorithms presented so that for certain objective functions, the same exponential rate of convergence can be achieved close to the minimum.

Remark 4.3. In very high dimensional nonconvex optimization problems of practical interest, it has been noted empirically that a main source of difficulty is not the presence of local minima but rather the ubiquity of saddle points surrounded by high error plateaux [11, 23]. These saddle points can significantly slow down gradient-based algorithms and give the illusion of the existence of a local minima. It was demonstrated in [24] via a variant of Nesterov's accelerated gradient algorithm that momentum techniques can escape saddle points faster than standard gradient methods and can thereby accelerate convergence in the nonconvex setting as well. This suggests that the variational framework for accelerated optimization and our adaptive approach to obtain symplectic optimization algorithms may also be promising with regards to nonconvex optimization.

5. Conclusions. Due to the degeneracy of the Hamiltonian, adaptive variational integrators based on the Poincaré transformation should be constructed using discrete Hamiltonians, which are Type II or III generating functions. This has potential implications for the numerical properties of such integrators and might explain why there has only been a limited amount of work on the construction of adaptive variational integrators based on the traditional Lagrangian perspective. The efficiency of the resulting integrator is largely based upon a proper choice of the monitor function g, and more research is needed to find a general choice of g that maintains a decent level of efficiency.

We have also noted that the gain in efficiency provided by adaptivity depends on the properties of the Hamiltonian dynamical system and tends to be more significant in regions of high curvature of the Hamiltonian in the vector field. We focused primarily on Taylor variational integrators, but Galerkin variational integrators are likely to be very promising as well since the cost of evaluating the monitor function and its derivatives should be low. In addition, the Galerkin approximation scheme may help inform a better choice of monitor function, due to the extensive literature on efficient a posteriori error estimation. A posteriori error estimation, in general, would be a nice addition to give some guarantees on global accuracy.

Finally, we used our adaptive framework together with the variational approach to accelerated optimization presented in [58] to design efficient Hamiltonian variational and nonvariational explicit integrators for symplectic accelerated optimization. We noted that a careful use of adaptivity and symplecticity can result in significantly faster algorithms, and it could also play an important role for nonconvex optimization problems. It would be desirable to understand at a theoretical level the role that adaptivity and symplecticity plays in the accurate and stable discretization of flows that correspond to accelerated optimization algorithms, which could better inform the choice of monitor functions, and the convex function used to define the Bregman divergence that arises in the construction of the Bregman Hamiltonian. This direction seems particularly promising for constructing novel optimization algorithms with superior computational efficiency and performance. Another possible future research direction is to consider how these variational and adaptive frameworks extend to more general spaces such as Lie groups [56, 30] and Riemannian manifolds [15, 13, 14]. It could also be interesting to consider the implications of this work for stochastic gradient descent methods [48], by considering it in the context of a Bregman Lagrangian or Hamiltonian, but with a stochastic perturbation of the potential. This naturally leads to considering stochastic generalizations of the adaptive Hamiltonian variational integrators considered in this paper, by extending the existing work on stochastic variational integrators [7, 22].

Appendix A. Proof of Theorem 2.1. The proof of Theorem 2.1 is similar to the one presented in the Appendix of [50] for Lagrangian Taylor variational integrators. We first start with the right HTVI case. Let $q(t)$ and $p(t)$ denote the solutions of Hamilton's boundary-value problem

$$\dot{q}(t) = g(q(t), p(t), t), \qquad \dot{p}(t) = f(q(t), p(t), t), \qquad q(0) = q_0, \quad p(h) = p_1.$$

Let $q_1 = q(h)$ and $p_0 = p(0)$.

LEMMA A.1. *Given an r-order Taylor method $\Psi_h^{(r)}$ approximating the exact time-h flow map corresponding to Hamilton's equations, let $\tilde{p}_0$ solve the problem $p_1 = \pi_{T^*Q} \circ \Psi_h^{(r)}(q_0, \tilde{p}_0)$. Then,*

$$\tilde{p}_0 = p_0 + \mathcal{O}(h^{r+1}).$$

Proof. Solving the equation $p_1 = \pi_{T^*Q} \circ \Psi_h^{(r)}(q_0, \tilde{p}_0)$ for $\tilde{p}_0$ yields

$$\tilde{p}_0 = p_1 - \sum_{k=1}^{r} \frac{h^k}{k!} f^{(k-1)}(q_0, \tilde{p}_0, 0).$$

The exact solution $p(t)$ belongs to $C^{r+1}([0,h])$, so Taylor's theorem gives

$$p_0 = p_1 - \sum_{k=1}^{r} \frac{h^k}{k!} f^{(k-1)}(q_0, p_0, 0) + R_r(h).$$

Now, since $p(t)$ belongs to $C^{r+1}([0,h])$, $f^{(k-1)}$ is Lipschitz continuous in its arguments for $k = 1, \dots, r-1$. Let M be the largest of the corresponding $(r-1)$ Lipschitz constants with respect to the second argument over the compact interval $[0, C]$. Then, using the triangle inequality,

$$\begin{aligned} \|\tilde{p}_0 - p_0\| &= \left\| R_r(h) - \sum_{k=1}^{r} \frac{h^k}{k!} \left[f^{(k-1)}(q_0, \tilde{p}_0, 0) - f^{(k-1)}(q_0, p_0, 0) \right] \right\| \\ &\leq M \sum_{k=1}^{r} \frac{h^k}{k!} \|\tilde{p}_0 - p_0\| + \|R_r(h)\|. \end{aligned}$$

Thus, $\left(1 - M \sum_{k=1}^{r} \frac{h^k}{k!}\right) \|\tilde{p}_0 - p_0\| \leq \|R_r(h)\| = \mathcal{O}(h^{r+1})$, and by continuity, $\exists \tilde{C} \in (0, C)$ such that $\forall h \in (0, \tilde{C})$, the term $\left(1 - M \sum_{k=1}^{r} \frac{h^k}{k!}\right)$ is bounded away from zero, which concludes the proof. ∎

We now show that starting the r-order Taylor method with initial conditions $(q_0, \tilde{p}_0)$ rather than (q_0, p_0) will not affect the order of accuracy of the method.

LEMMA A.2. *The r-order Taylor method $\Psi_h^{(r)}$ with initial conditions $(q_0, \tilde{p}_0)$ and where $\tilde{p}_0$ solves $p_1 = \pi_{T^*Q} \circ \Psi_h^{(r)}(q_0, \tilde{p}_0)$ is accurate to at least $\mathcal{O}(h^{r+1})$ for the Hamiltonian boundary-value problem.*

Proof. Let $(\tilde{q}(t), \tilde{p}(t))$ denote the exact solution to Hamiltonian's equations with initial values $(q_0, \tilde{p}_0)$, and let $(q_d(t), p_d(t))$ denote the values generated by the r-order Taylor method with initial conditions $(q_0, \tilde{p}_0)$. The Hamiltonian initial-value problem is well-posed, so denoting the Lipschitz constant with respect to the second argument by M, we get

$$\begin{aligned} \|(q(t), p(t)) - (q_d(t), p_d(t))\| &\leq \|(q(t), p(t)) - (\tilde{q}(t), \tilde{p}(t))\| + \|(\tilde{q}(t), \tilde{p}(t)) - (q_d(t), p_d(t))\| \\ &\leq M \|p_0 - \tilde{p}_0\| + \mathcal{O}(h^{r+1}) \leq \mathcal{O}(h^{r+1}), \end{aligned}$$

where we have used the triangle inequality and the fact that the local truncation error of an r-order Taylor method is $\mathcal{O}(h^{r+1})$ to bound $\|(\tilde{q}(t), \tilde{p}(t)) - (q_d(t), p_d(t))\|$. ∎

We are now ready to prove Theorem 2.1 for right HTVIs.

THEOREM A.3. *Consider a Hamiltonian H such that H and $\frac{\partial H}{\partial p}$ are Lipschitz continuous in both variables. Given an r-order accurate Taylor method $\Psi_h^{(r)}$ and an s-order accurate quadrature formula with weights and nodes (b_i, c_i), define the associated Taylor discrete right Hamiltonian*

$$H_d^+(q_0, p_1; h) = p_1^\top \tilde{q}_1 - h \sum_{i=1}^{m} b_i \left[p_{c_i}^\top \dot{q}_{c_i} - H(q_{c_i}, p_{c_i}) \right],$$

where $\tilde{p}_0$ *solves* $p_1 = \pi_{T^*Q} \circ \Psi_h^{(r)}(q_0, \tilde{p}_0)$*, where* $\tilde{q}_1 = \pi_Q \circ \Psi_h^{(r+1)}(q_0, \tilde{p}_0)$ *and* $(q_{c_i}, p_{c_i}) = \Psi_{c_i h}^{(r)}(q_0, \tilde{p}_0)$*, and where we use the continuous Legendre transform to obtain* $\dot{q}_{c_i}$*.*

Then, H_d^+ *approximates* $H_d^{+,E}$ *with order of accuracy at least* $\min(r+1, s)$*. By Theorem* 2.2 *in* [49]*, the associated discrete right Hamiltonian map has the same order of accuracy.*

Proof. From Lemma A.2 we have that $q(c_i h) = q_{c_i} + \mathcal{O}(h^{r+1})$ and $p(c_i h) = p_{c_i} + \mathcal{O}(h^{r+1})$, and since $\frac{\partial H}{\partial p}$ is Lipschitz in both variables, $\dot{q}(c_i h) - \dot{q}_{c_i} = \frac{\partial H}{\partial p}(q(c_i h), p(c_i h)) - \frac{\partial H}{\partial p}(q_{c_i}, p_{c_i}) = \mathcal{O}(h^{r+1})$. Since the quadrature formula is of order s accurate, (2.4) for $H_d^{+,E}(q_0, p_1; h)$ gives

$$H_d^{+,E}(q_0, p_1; h)$$
$$= p_1^\top q_1 - h \sum_{i=1}^m b_i \left[p(c_i h)^\top \dot{q}(c_i h) - H\left(q_{c_i} + \mathcal{O}(h^{r+1}), p_{c_i} + \mathcal{O}(h^{r+1})\right)\right] + \mathcal{O}(h^{s+1}).$$

Now, since $\tilde{q}_1 = \pi_Q \circ \Psi_h^{(r+1)}(q_0, \tilde{p}_0)$, it follows from Lemma A.2 that $\tilde{q}_1 = q_1 + \mathcal{O}(h^{r+2})$. Therefore, combining this with the fact that H is Lipschitz continuous in both variables yields

$$\begin{aligned} H_d^{+,E}(q_0, p_1; h) &= p_1^\top \tilde{q}_1 - h \sum_{i=1}^m b_i \left[p_{c_i}^\top \dot{q}_{c_i} - H\left(q_{c_i}, p_{c_i}\right)\right] + \mathcal{O}(h^{r+2}) + \mathcal{O}(h^{s+1}) \\ &= H_d^+(q_0, p_1; h) + \mathcal{O}(h^{\min(r+1,s)+1}). \end{aligned}$$

Therefore, H_d^+ approximates $H_d^{+,E}$ with order of accuracy at least $\min(r+1, s)$. ∎

Theorem 2.1 can be proven in a similar way for left HTVIs. Now, $q(t)$ and $p(t)$ denote the solutions of the Hamilton's boundary-value problem

$$\dot{q}(t) = g(q(t), p(t), t), \qquad \dot{p}(t) = f(q(t), p(t), t), \qquad q(h) = q_1, \quad p(0) = p_0,$$

and we let $q_0 = q(0)$ and $p_1 = p(h)$. Lemma A.1 is replaced by the following.

LEMMA A.4. *Given an* $(r+1)$*-order Taylor method* $\Psi_h^{(r+1)}$ *approximating the exact time-h flow map corresponding to Hamilton's equations, let* $\tilde{q}_0$ *solve the problem* $q_1 = \pi_Q \circ \Psi_h^{(r+1)}(\tilde{q}_0, p_0)$*. Then,*

$$\tilde{q}_0 = q_0 + \mathcal{O}(h^{r+2}).$$

Proof. We proceed as in the proof of Lemma A.1. We first solve $q_1 = \pi_Q \circ \Psi_h^{(r+1)}(\tilde{q}_0, p_0)$ for $\tilde{q}_0$ and then Taylor expand the exact solution $q(t)$ which belongs to $C^{r+2}([0, h])$. Now, $q(t)$ is Lipschitz continuous in its arguments for $k = 1, \ldots, r$, so we can let M be the largest of the corresponding r Lipschitz constants with respect to the first argument over the compact interval $[0, C]$. Then, as before, the triangle inequality can be used to get that $\left(1 - M \sum_{k=1}^{r+1} \frac{h^k}{k!}\right) \|\tilde{q}_0 - q_0\| = \mathcal{O}(h^{r+2})$, and by continuity, the term inside the parentheses is bounded away from zero. ∎

In analogy to Lemma A.2, we now show that starting the r-order Taylor method with initial conditions $(\tilde{q}_0, p_0)$ rather than (q_0, p_0) will not affect the order of accuracy of the method.

LEMMA A.5. *The* r*-order Taylor method* $\Psi_h^{(r)}$ *with initial conditions* $(\tilde{q}_0, p_0)$ *and where* $\tilde{q}_0$ *solves* $q_1 = \pi_Q \circ \Psi_h^{(r)}(\tilde{q}_0, p_0)$ *is accurate to at least* $\mathcal{O}(h^{r+1})$ *for the Hamiltonian boundary-value problem.*

Proof. Let $(\tilde{q}(t), \tilde{p}(t))$ denote the exact solution to Hamiltonian's equations with initial values $(\tilde{q}_0, p_0)$, and let $(q_d(t), p_d(t))$ denote the values generated by the r-order Taylor method with initial conditions $(\tilde{q}_0, p_0)$. The Hamiltonian initial-value problem is well-posed, so denoting the Lipschitz constant with respect to the first argument by M, we get

$$\begin{aligned}\|(q(t), p(t)) - (q_d(t), p_d(t))\| &\leq \|(q(t), p(t)) - (\tilde{q}(t), \tilde{p}(t))\| + \|(\tilde{q}(t), \tilde{p}(t)) - (q_d(t), p_d(t))\| \\ &\leq M\|q_0 - \tilde{q}_0\| + \mathcal{O}(h^{r+1}) \leq \mathcal{O}(h^{r+1}),\end{aligned}$$

where we have used the triangle inequality and the fact that the local truncation error of an r-order Taylor method is $\mathcal{O}(h^{r+1})$ to bound $\|(\tilde{q}(t), \tilde{p}(t)) - (q_d(t), p_d(t))\|$. ∎

We are now ready to prove Theorem 2.1 for left HTVIs.

THEOREM A.6. *Consider a Hamiltonian H such that H and $\frac{\partial H}{\partial p}$ are Lipschitz continuous in both variables. Given an r-order accurate Taylor method $\Psi_h^{(r)}$ and an s-order accurate quadrature formula with weights and nodes (b_i, c_i), define the associated Taylor discrete left Hamiltonian*

$$H_d^-(q_1, p_0; h) = -p_0^\top \tilde{q}_0 - h\sum_{i=1}^m b_i \left[p_{c_i}^\top \dot{q}_{c_i} - H(q_{c_i}, p_{c_i})\right],$$

where $\tilde{q}_0$ solves the problem $q_1 = \pi_Q \circ \Psi_h^{(r+1)}(\tilde{q}_0, p_0)$, where $(q_{c_i}, p_{c_i}) = \Psi_{c_i h}^{(r)}(q_0, \tilde{p}_0)$, and where we use the continuous Legendre transform to obtain $\dot{q}_{c_i}$.

Then, H_d^- approximates $H_d^{-,E}$ with order of accuracy at least $\min(r+1, s)$. By a result analogous to Theorem 2.2 in [49], the associated discrete left Hamiltonian map has the same order of accuracy.

Proof. From Lemma A.5 we have that $q(c_i h) = q_{c_i} + \mathcal{O}(h^{r+1})$, and $p(c_i h) = p_{c_i} + \mathcal{O}(h^{r+1})$, and since $\frac{\partial H}{\partial p}$ is Lipschitz in both variables, $\dot{q}(c_i h) - \dot{q}_{c_i} = \frac{\partial H}{\partial p}(q(c_i h), p(c_i h)) - \frac{\partial H}{\partial p}(q_{c_i}, p_{c_i}) = \mathcal{O}(h^{r+1})$. Since the quadrature formula is of order s accurate, (2.7) for $H_d^{-,E}(q_1, p_0; h)$ gives

$$\begin{aligned}&H_d^{-,E}(q_1, p_0; h) \\ &= -p_0^\top q_0 - h\sum_{i=1}^m b_i \left[p(c_i h)^\top \dot{q}(c_i h) - H\left(q_{c_i} + \mathcal{O}(h^{r+1}), p_{c_i} + \mathcal{O}(h^{r+1})\right)\right] + \mathcal{O}(h^{s+1}).\end{aligned}$$

Now, since $q_1 = \pi_Q \circ \Psi_h^{(r+1)}(\tilde{q}_0, p_0)$, it follows from Lemma A.4 that $\tilde{q}_0 = q_0 + \mathcal{O}(h^{r+2})$. Therefore, combining this with the fact that H is Lipschitz continuous in both variables yields

$$H_d^{-,E}(q_1, p_0; h) = H_d^-(q_1, p_0; h) + \mathcal{O}(h^{\min(r+1, s)+1}).$$ ∎

REFERENCES

[1] R. BARRIO, *Performance of the Taylor series method for ODEs/DAEs*, Appl. Math. Comput., 163 (2005), pp. 525–545.

[2] M. BETANCOURT, M. I. JORDAN, AND A. WILSON, *On Symplectic Optimization*, preprint, https://arxiv.org/abs/1802.03653, 2018.

[3] J. BETTENCOURT, M. JOHNSON, AND D. DUVENAUD, *Taylor-Mode Automatic Differentiation for Higher-Order Derivatives in JAX*, preprint, https://openreview.net/forum?id=SkxEF3FNPH, 2019.

[4] F. Biscani and D. Izzo, *Revisiting high-order Taylor methods for astrodynamics and celestial mechanics*, Mon. Not. R. Astron. Soc., 504 (2021), pp. 2614–2628, https://doi.org/10.1093/mnras/stab1032.
[5] S. Blanes and C. J. Budd, *Explicit adaptive symplectic (easy) integrators: A scaling invariant generalisation of the Levi-Civita and KS regularisations*, Celest. Mech. Dyn. Astron., 89 (2004), pp. 383–405.
[6] S. Blanes and A. Iserles, *Explicit adaptive symplectic integrators for solving Hamiltonian systems*, Celest. Mech. Dyn. Astron., 114 (2012), pp. 297–317.
[7] N. Bou-Rabee and H. Owhadi, *Stochastic variational integrators*, IMA J. Numer. Anal., 29 (2009), pp. 421–443.
[8] M. P. Calvo and J. M. Sanz-Serna, *The development of variable-step symplectic integrators, with application to the two-body problem*, SIAM J. Sci. Comput., 14 (1993), pp. 936–952, https://doi.org/10.1137/0914057.
[9] A. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, Acad. Sci. Paris, 25 (1847), pp. 536–538.
[10] B. Christianson, *Automatic Hessians by reverse accumulation*, IMA J. Numer. Anal., 12 (1992), pp. 135–150, https://doi.org/10.1093/imanum/12.2.135.
[11] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, *Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-convex Optimization*, Adv. Neural Inform. Process. Syst. 27, MIT Press, Cambridge, MA, 2014.
[12] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.
[13] V. Duruisseaux and M. Leok, *Accelerated Optimization on Riemannian Manifolds via Discrete Constrained Variational Integrators*, preprint, https://arxiv.org/abs/2104.07176, 2021.
[14] V. Duruisseaux and M. Leok, *Accelerated Optimization on Riemannian Manifolds via Projected Variational Integrators*, manuscript, 2021.
[15] V. Duruisseaux and M. Leok, *A variational formulation of accelerated optimization on Riemannian manifolds*, preprint, https://arxiv.org/abs/2101.06552, 2021.
[16] Z. Ge and J. E. Marsden, *Lie–Poisson Hamilton–Jacobi theory and Lie-Poisson integrators*, Phys. Lett. A, 133 (1988), pp. 134–139.
[17] B. Gladman, M. Duncan, and J. Candy, *Symplectic integrators for long-time integrations in celestial mechanics*, Celest. Mech. Dyn. Astron., 52 (1991), pp. 221–240.
[18] E. Hairer, *Variable time step integration with symplectic methods*, Appl. Numer. Math., 25 (1997), pp. 219–227.
[19] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, 2nd ed., Springer Ser. Comput. Math. 31, Springer-Verlag, Berlin, 2006.
[20] J. Hall and M. Leok, *Spectral variational integrators*, Numer. Math., 130 (2015), pp. 681–740.
[21] M. Hochbruck and A. Ostermann, *Exponential integrators*, Acta Numer., 19 (2010), pp. 209–286.
[22] D. D. Holm and T. M. Tyranowski, *Stochastic discrete Hamiltonian variational integrators*, BIT, 58 (2018), pp. 1009–1048.
[23] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, *On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points*, J. ACM, 68 (2021), 11, https://doi.org/10.1145/3418526, 11.
[24] C. Jin, P. Netrapalli, and M. I. Jordan, *Accelerated gradient descent escapes saddle points faster than gradient descent*, in Proceedings of the 31st Conference on Learning Theory, Proc. Mach. Learn. Res.. 75, 2018, pp. 1042–1085.
[25] A. Jorba and M. Zou, *A Software Package for the Numerical Integration of ODEs by Means of High-Order Taylor Methods*, Experiment. Math., 14 (2005), pp. 99–117, https://doi.org/10.1080/10586458.2005.10128904.
[26] C. Kane, J. Marsden, and M. Ortiz, *Symplectic-energy-momentum preserving variational integrators*, J. Math. Phys., 40 (1999), pp. 3353–3371.
[27] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in Proceedings of the International Conference on Learning Representations, 2014.
[28] G. Kirlinger and G. F. Corliss, *On implicit Taylor series methods for stiff ODEs*, in Computer Arithmetic and Enclosure Methods, North-Holland, Amsterdam, 1992, pp. 271–379.
[29] S. Lall and M. West, *Discrete variational Hamiltonian mechanics*, J. Phys. A, 39 (2006), pp. 5509–5519.
[30] T. Lee, M. Tao, and M. Leok, *Variational Symplectic Accelerated Optimization on Lie Groups*, preprint, https://arxiv.org/abs/2103.14166, 2021.
[31] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monogr. Appl.

Comput. Math. 14, Cambridge University Press, Cambridge, UK, 2004.
[32] M. Leok and T. Shingel, *Prolongation-collocation variational integrators*, IMA J. Numer. Anal., 32 (2012), pp. 1194–1216.
[33] M. Leok and J. Zhang, *Discrete Hamiltonian variational integrators*, IMA J. Numer. Anal., 31 (2011), pp. 1497–1532.
[34] J. E. Marsden and M. West, *Discrete mechanics and variational integrators*, Acta Numer., 10 (2001), pp. 357–514.
[35] K. Modin and C. Führer, *Time-step adaptivity in variational integrators with application to contact problems*, ZAMM Z. Angew. Math. Mech., 86 (2006), pp. 785–794.
[36] S. Nair, *Time adaptive variational integrators: A space-time geodesic approach*, Phys. D, 241 (2012), pp. 315–325.
[37] R. D. Neidinger, *Directions for computing truncated multivariate Taylor series*, Math. Comp., 74 (2005), pp. 321–340.
[38] R. D. Neidinger, *Introduction to automatic differentiation and MATLAB object-oriented programming*, SIAM Rev., 52 (2010), pp. 545–563, https://doi.org/10.1137/080743627.
[39] R. D. Neidinger, *Efficient recurrence relations for univariate and multivariate Taylor series coefficients*, in Proceedings of the 9th AIMS International Conference (Orlando, FL, USA) 2013, pp. 587–596.
[40] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.
[41] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $\mathcal{O}(1/k^2)$, Soviet Math. Dokl., 27 (1983), pp. 372–376.
[42] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic Publishers, Boston, MA, 2004.
[43] Y. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Math. Program., 112 (2008), pp. 159–181.
[44] M. A. Patterson, M. Weinstein, and A. V. Rao, *An efficient overloaded method for computing derivatives of mathematical functions in MATLAB*, ACM Trans. Math. Software, 39 (2013), 17, https://doi.org/10.1145/2450153.2450155.
[45] B. A. Pearlmutter, *Lazy multivariate higher-order forward-mode AD*, in Proceedings of the Symposium on Principles of Programming Languages, 2007.
[46] P. Pihajoki, *Explicit methods in extended phase space for inseparable Hamiltonian problems*, Celest. Mech. Dyn. Astron., 121 (2015), pp. 211–231.
[47] S. Reich, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal., 36 (1999), pp. 1549–1570, https://doi.org/10.1137/S0036142997329797.
[48] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statistics, 22 (1951), pp. 400–407.
[49] J. M. Schmitt and M. Leok, *Properties of Hamiltonian variational integrators*, IMA J. Numer. Anal., 38 (2018), pp. 377–398.
[50] J. M. Schmitt, T. Shingel, and M. Leok, *Lagrangian and Hamiltonian Taylor variational integrators*, BIT, 58 (2018), pp. 457–488.
[51] X. Shen and M. Leok, *Geometric exponential integrators*, J. Comput. Phys., 382 (2019), pp. 27–42.
[52] J. Struckmeier, *Hamiltonian dynamics on the symplectic extended phase space for autonomous and non-autonomous systems*, J. Phys. A, 38 (2005), pp. 1257–1278.
[53] W. Su, S. Boyd, and E. Candes, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res., 17 (2016), 153.
[54] Y. B. Suris, *Hamiltonian methods of Runge-Kutta type and their variational interpretation*, Mat. Model., 2 (1990), pp. 78–87.
[55] M. Tao, *Explicit symplectic approximation of nonseparable Hamiltonians: Algorithm and long time performance*, Phys. Rev. E (3), 94 (2016), 043303.
[56] M. Tao and T. Ohsawa, *Variational Optimization on Lie Groups, with Examples of Leading (Generalized) Eigenvalue Problems*, in Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proc. Mach. Learn. Res. 108, 2020, pp. 4269–4280.
[57] T. Tieleman and G. Hinton, *Coursera: Neural networks for machine learning - Lecture* 6.5*: RMSprop*, University of Toronto, Toronto, ON, Canada, 2012.
[58] A. Wibisono, A. Wilson, and M. Jordan, *A variational perspective on accelerated methods in optimization*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. E7351–E7358.