AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 4(6): 1841–1853.
DOI:10.3934/math.2019.6.1841
Received: 31 January 2020
Accepted: 07 February 2020
Published: 19 February 2020

Research article

On theoretical upper limits for valid timesteps of implicit ODE methods

Kevin R. Green[1], **George W. Patrick**[2] **and Raymond J. Spiteri**[1,*]

[1] Department of Computer Science, University of Saskatchewan, 110 Science Place, Saskatoon, SK, S7N 5C9, Canada

[2] Department of Mathematics and Statistics, University of Saskatchewan, 106 Wiggins Road, Saskatoon, SK, S7N 5E6, Canada

* **Correspondence:** Email: spiteri@cs.usask.ca; Tel: +1-306-966-2909; Fax: +1-306-966-4884.

Abstract: Implicit methods for the numerical solution of initial-value problems may admit multiple solutions at any given time step. Accordingly, their nonlinear solvers may converge to any of these solutions. Below a critical timestep, exactly one of the solutions (the consistent solution) occurs on a solution branch (the principal branch) that can be continuously and monotonically continued back to zero timestep.

Standard step-size control can promote convergence to consistent solutions by adjusting the timestep to maintain an error estimate below a given tolerance. However, simulations for symplectic systems or large physical systems are often run with constant timesteps and are thus more susceptible to convergence to inconsistent solutions. Because simulations cannot be reliably continued from inconsistent solutions, the critical timestep is a theoretical upper bound for valid timesteps.

Keywords: implicit method; bifurcation; initial-value problem; double pendulum
Mathematics Subject Classification: 34A09

1. Introduction

Many mathematical models take the form of a system of ordinary differential equations (ODEs) for a vector of unknowns $\boldsymbol{q}(t) \in \mathbb{R}^m$ subject to boundary data:

$$\begin{aligned} \dot{\boldsymbol{q}}(t) &= \boldsymbol{f}(\boldsymbol{q}(t)), \qquad t \in (t_0, t_f), \\ 0 &= g_i(\boldsymbol{q}(\tau_i)), \quad \tau_i \in \{t_0, t_f\},\ i = 1, 2, \ldots, m. \end{aligned}$$

Standard transformations reduce systems that are higher order, non-autonomous, or subject to interior-point data to this first-order autonomous form with boundary data at the cost of increased system

dimension [9]. When $\tau_i \equiv \tau_0$, $i = 1, 2, \ldots, m$, we have an initial-value problem (IVP); otherwise, we have a two-point boundary-value problem (BVP).

In practice, numerical methods for the solution for such ODEs involve successive approximations at successive timesteps and are either *implicit* or *explicit*. Implicit methods typically involve the iterative solution, at each time step, of systems of nonlinear algebraic equations, generally a theoretically infinite process with a potentially non-unique or non-existent result. Explicit methods, in contrast, can be implemented directly, generally a theoretically finite process with a unique result. The iterative solution process of an implicit method can incur a significant run-time cost, but the use of such methods may result in greater overall efficiency or fidelity. For example, the increase in the timestep afforded by an implicit method when solving stiff ODEs typically offsets the increased cost per step. Also, when integrating a Hamiltonian system, an implicit method may be used to arrange that the simulation itself preserves energy or is symplectic [3, 23].

The existence and uniqueness theory for IVPs is much more decisive than for BVPs. IVPs have unique solutions under mild assumptions that are typically satisfied in practice, whereas BVPs may have from zero to uncountably many solutions. However, when an implicit method is involved in approximating the solution of an IVP, the possibility emerges of divergence or convergence to one of multiple solutions. Convergence to spurious solutions is well recognized ([5, 11] and references therein), particularly in the numerical solution of BVPs ([9, 16–18] and references therein). Less attention is typically given to the context of solving IVPs ([14, 15, 19] and references therein), where there is theoretically a unique solution and where the presence of a "good" initial guess is taken for granted. In this article, we consider the context of an implicit IVP method that has multiple solutions at a given timestep and how to choose from among them, as opposed to a qualitatively incorrect numerical solution of an IVP or BVP.

Standard methods for error estimation and control via timestep selection tend to adjust timesteps such that, in practice, any ambiguity arising from multiple solutions is avoided, but they are not usually specifically designed to do so. But there are two specific scenarios in which constant timesteps are often used in practice: simulations of symplectic systems using a symplectic method [3] and simulations of large physical systems. Efforts toward adaptive symplectic methods have been made, but they tend to be specialized and require a significant amount of user judgment [20–22, 33]. Software packages for the simulation of large physical systems, especially on distributed architectures [29–32], often use constant timesteps because of the large relative expense of estimating the error and changing the timestep. However, at constant timestep, a simulation may unexpectedly encounter a time-localized region of complex dynamics, and convergence can easily be construed as nominal when in fact it is not. A large number of independent simulations, e.g., explorations of a parameter space, may not be feasible or efficient at small constant timestep. It may not be easy to automate the detection of anomalous behaviour in the absence of convergence failure, and a manual inspection may not be feasible.

Numerical methods for solution of ODEs can also be classified according to how many past steps are stored and used in computing the next step. In general, the next step can be a function of k past steps, leading to *multi-step methods*. If only the current state is stored, then the method is *one-step*. Although a multistep method may be regarded as a one-step method on a Cartesian product of state spaces [28], the theory of multistep methods is complicated by the possibility of uncontrolled growth of the error in the past states [25]. This is not the focus of this article; we consider only one-step methods.

To formalize, the time-advanced state $\boldsymbol{q}^{n+1} \approx \boldsymbol{q}(t_{n+1})$ after one step of a one-step implicit method is obtained from the given current state $\boldsymbol{q}^n \approx \boldsymbol{q}(t_n)$ by solving a generally nonlinear equation of the form

$$\boldsymbol{F}(\boldsymbol{y}; \boldsymbol{x}, h) = \boldsymbol{0}, \qquad \boldsymbol{y} = \boldsymbol{q}^{n+1}, \qquad \boldsymbol{x} = \boldsymbol{q}^n, \tag{1.1}$$

where h is the given timestep. For example, the backward Euler method has $\boldsymbol{F}(\boldsymbol{y}; \boldsymbol{x}, h) = \boldsymbol{y} - \boldsymbol{x} - h\,\boldsymbol{f}(\boldsymbol{y})$.

We assume, for all $\boldsymbol{x}$, that

$$\boldsymbol{F}(\boldsymbol{x}; \boldsymbol{x}, 0) = \boldsymbol{0}, \qquad \boldsymbol{F}_{\boldsymbol{y}}(\boldsymbol{x}; \boldsymbol{x}, 0) = \boldsymbol{1}, \qquad \boldsymbol{F}_h(\boldsymbol{x}; \boldsymbol{x}, 0) = -\boldsymbol{f}(\boldsymbol{x}),$$

where $\boldsymbol{1}$ is the identity matrix and subscripts of $\boldsymbol{F}$ denote partial derivatives. Then, by the implicit function theorem, for any fixed $\boldsymbol{x}$, there is a unique smooth solution $\boldsymbol{y}(h)$ defined for sufficiently small h, and $\boldsymbol{y} = \boldsymbol{x} + h\boldsymbol{f}(\boldsymbol{x}) + O(h^2)$, as is required for consistency.

Two solutions $\boldsymbol{y}_1(h)$ and $\boldsymbol{y}_2(h)$, defined on open intervals containing $h = 0$ and satisfying the condition that $\boldsymbol{F}_{\boldsymbol{y}}(\boldsymbol{y}_i(h); \boldsymbol{x}, h)$ is nonsingular, are equal on the intersection of their domains (the set on which $\boldsymbol{y}_1(h) = \boldsymbol{y}_2(h)$ is nonempty, closed, and open by the implicit function theorem, and the intersection of intervals is connected). Therefore, there is a maximal such solution, which we call the *principal solution branch*, and there is generally a critical timestep h_c after which the principal branch ceases to exist. The condition that the solution set of $F(\boldsymbol{y}; \boldsymbol{x}, h) = \boldsymbol{0}$ is smooth in the space $\{\boldsymbol{y}, h\}$ is weaker than the condition that it defines $\boldsymbol{y}$ as a function of h. Solutions may be continuously connected after a fold bifurcation, for example, where the solution manifold turns backwards from the direction of increasing timestep [26, 27]. Continuing through such a bifurcation leads to multiple solutions at smaller timesteps than the critical one at which the bifurcation occurs. These solutions co-exist with the solutions from the principal branch, and they may persist even as the timestep approaches zero.

A simple example demonstrating the existence of a critical timestep h_c at which a fold bifurcation occurs is the application of the backward Euler method to the scalar IVP

$$\dot{q} = q^2, \quad q(0) = q_0 > 0.$$

With the backward Euler method, the update equation (1.1) can be written as

$$hy^2 - y + x = 0,$$

having solutions

$$y = \frac{1 \pm \sqrt{1 - 4hx}}{2h}.$$

Evidently, there are two solutions for $h < h_c = (4x)^{-1}$, a single solution at $h = h_c$, and no solutions for $h > h_c$; see Figure 1. We note the existence of two solutions for all $0 < h < h_c$. By the Newton–Kantarovich Theorem [24], convergence to either solution is possible for an appropriately chosen initial guess.

The principal solution branch contains the initial condition at zero timestep, which is exactly correct, and it contains all solutions near zero timestep that continuously emanate from the initial condition. If a more complicated bifurcation occurs, say a pitchfork, as opposed to a fold (e.g., bifurcations of the solutions of $y^3 - (h - h_c)y = 0$ and $y^2 - (h - h_c) = 0$, respectively), then there are multiple solutions

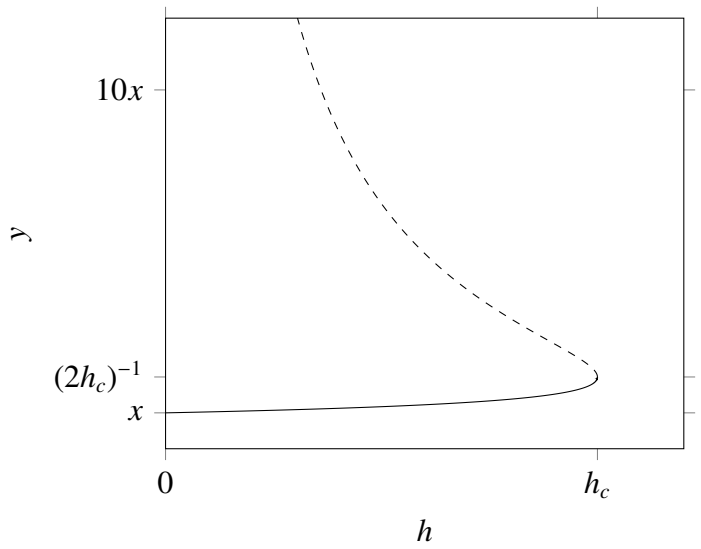


Figure 1. The backward Euler method for the equation $\dot{q} = q^2$ has a fold point at $(h_c, (2h_c)^{-1})$, where $h_c = (4x)^{-1}$. There are two solutions for $h < h_c$, a unique solution at $h = h_c$, and no solutions for $h > h_c$. Solutions on the lower branch, which we refer to as the *principal solution branch*, are preferred because solutions they are naturally considered to be consistent with the initial condition.

beyond the critical timestep, none of which can be continuously and monotonically continued back to the initial condition without passage through the bifurcation itself. There may be two successive fold bifurcations, resulting in an "S"-shaped solution manifold, with timesteps larger than h_c, for which there is a unique solution. But in that situation, the two additional solutions at timesteps $h \lesssim h_c$ would be rejected in favour of that on the principal branch, and for $h \gtrsim h_c$, the unique solutions are near an inconsistent one and so would also be rejected.

Therefore, if there are multiple solutions at a given timestep smaller than the critical one, then the principal branch should be chosen, and the timestep cannot be validly increased to be larger than the critical one, irrespective of whether or not there are unique solutions there. Simulations should generally not be carried on from solutions that are not on the principal branch, regardless of whether the timestep used is larger or smaller the critical timestep, because they tend to result in non-negligible perturbations to the solution. So as to have a concise term, we make the following definitions.

Definition 1.1. Given an initial condition $\boldsymbol{x}$, a solution $\boldsymbol{y}$ with timestep h is called *consistent* if $(\boldsymbol{y}, h)$ is on the principal solution branch.

Definition 1.2. The smallest timestep h such that there is a bifurcation of the principal branch is called the critical timestep h_c. Timesteps $h > h_c$ are called *invalid*. By definition, any solutions obtained with invalid timesteps cannot be on the principal branch.

Another basic example relevant to Lagrangian systems that shows the existence of a fold bifurcation for a critical timestep h_c involves the Lagrangian

$$L(x, \dot{x}) = \frac{1}{2}\dot{x}^2 - \frac{1}{3}x^3,$$

which models the dynamics of a nonlinear spring that has a stiffness proportional to its linear displacement from equilibrium. Applying the first-order Variational Taylor method (described in Section 3 below) yields an update equation of the form

$$\begin{bmatrix} h(\dot{x}+\dot{y}) - 2(x-y) \\ h\left(x^2+y^2\right) + 2(\dot{x}-\dot{y}) \end{bmatrix} = \mathbf{0},$$

to be solved for y and $\dot{y}$. The solution to this update equation is

$$\begin{aligned} y &= h^{-2}\left(-2 \pm \sqrt{4 - h^2\left(h^2x^2 + 4h\dot{x} - 4x\right)}\right), \\ \dot{y} &= -\dot{x} + 2h^{-1}(x-y), \end{aligned} \tag{1.2}$$

which again shows the potential for multiple solutions. Taking, for example, the initial state $x = 1$, $\dot{x} = 0$, it is straightforward to show that (1.2) has two real solutions for $h < h_c = \sqrt{2+2\sqrt{2}}$, a single solution at $h = h_c$, and no real solutions for $h > h_c$.

2. Numerical continuation of the implicit solution

We consider one-parameter families of solutions to (1.1) in the hyperplane defined by the vector $(\boldsymbol{y}, h)$ for fixed $\boldsymbol{x}$. For practical computation, we parameterize these families by arclength, which monotonically increases throughout the computation.

Suppose we have two points on the solution curve $(\boldsymbol{y}_0, h_0)$ and $(\boldsymbol{y}_1, h_1)$ with known tangent vector $\boldsymbol{\mathcal{T}}_0$ at the first point. Our goal is to find a next point on the solution curve in the same arclength direction following $(\boldsymbol{y}_1, h_1)$. This *pseudo-arclength continuation* is accomplished in a predictor-corrector fashion. The prediction is performed via a linear approximation to the curve at the point $(\boldsymbol{y}_1, h_1)$ and the correction by computing the minimum-norm solution to the system with rank-1 deficiency.

The tangent vector $\boldsymbol{\mathcal{T}}_1$ at $(\boldsymbol{y}_1, h_1)$ satisfies

$$\begin{bmatrix} \boldsymbol{F}_{\boldsymbol{y}} & \boldsymbol{F}_h \end{bmatrix} \boldsymbol{\mathcal{T}}_1 = \mathbf{0}$$

and determines $\boldsymbol{\mathcal{T}}_1$ up to a scalar factor. To preserve the direction of the orientation of the branch, we require

$$\boldsymbol{\mathcal{T}}_0^T \boldsymbol{\mathcal{T}}_1 = 1.$$

Using a natural decomposition $\boldsymbol{\mathcal{T}}_i = (\boldsymbol{\mathcal{T}}_i^{(\boldsymbol{y})}, \mathcal{T}_i^{(h)})$, $i = 0, 1$, we write the above as a single system

$$\begin{bmatrix} \boldsymbol{F}_{\boldsymbol{y}} & \boldsymbol{F}_h \\ \boldsymbol{\mathcal{T}}_0^{(\boldsymbol{y})T} & \mathcal{T}_0^{(h)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mathcal{T}}_1^{(\boldsymbol{y})} \\ \mathcal{T}_1^{(h)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix},$$

the solution to which uniquely determines $\boldsymbol{\mathcal{T}}_1$.

Using a step of length Δs, we create a predictor

$$\hat{\boldsymbol{y}}_2 = \boldsymbol{y}_1 + \frac{\Delta s}{\|\boldsymbol{\mathcal{T}}_1\|} \boldsymbol{\mathcal{T}}_1^{(\boldsymbol{y})}, \quad \hat{h}_2 = h_1 + \frac{\Delta s}{\|\boldsymbol{\mathcal{T}}_1\|} \mathcal{T}_1^{(h)},$$

to approximate the next point on the curve. The name *pseudo-arclength* comes from the fact that Δs measures arclength along the tangent line.

From the predictor, a Newton-like method is applied to obtain the next solution point on the curve $(\boldsymbol{y}_2, h_2)$. Our specific choice of algorithm is the Gauss–Newton method. This particular approach for path-following was first used in [10]. It is identified as being the Gauss–Newton method in [8]. A practical description is given in [2] and summarized as follows.

We seek a solution to

$$\boldsymbol{F}(\hat{\boldsymbol{y}} + \Delta\boldsymbol{y}; \boldsymbol{x}, \hat{h} + \Delta h) = \boldsymbol{0}$$

such that $\|(\Delta\boldsymbol{y}, \Delta h)\|$ is minimal. Because $\boldsymbol{F}$ is nonlinear, we set up the Newton iteration as $(\boldsymbol{y}^0, h^0) = (\hat{\boldsymbol{y}}, \hat{h})$, and $\boldsymbol{y}^{k+1} = \boldsymbol{y}^k + \Delta\boldsymbol{y}$, $h^{k+1} = h^k + \Delta h$, where $(\Delta\boldsymbol{y}, \Delta h)$ is the minimum-norm solution to

$$\boldsymbol{F}_y(\boldsymbol{y}^k; \boldsymbol{x}, h^k)\Delta\boldsymbol{y} + \boldsymbol{F}_h(\boldsymbol{y}^k; \boldsymbol{x}, h^k)\Delta h = -\boldsymbol{F}(\boldsymbol{y}^k; \boldsymbol{x}, h^k).$$

The minimum-norm solution is obtained by solving the matrix system

$$\begin{bmatrix} \boldsymbol{F}_y & \boldsymbol{F}_h \\ \boldsymbol{\mathcal{T}}_1^{(y)T} & \mathcal{T}_1^{(h)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mathcal{T}}^{(y)} & \Delta_1\boldsymbol{y} \\ \mathcal{T}^{(h)} & \Delta_1 h \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{F} \\ 1 & 0 \end{bmatrix}$$

and constructing

$$\Delta\boldsymbol{y} = \Delta_1\boldsymbol{y} + \eta\boldsymbol{\mathcal{T}}^{(y)}, \quad \Delta h = \Delta_1 h + \eta\mathcal{T}^{(h)},$$

with

$$\eta = -\frac{(\Delta_1\boldsymbol{y})^T\boldsymbol{\mathcal{T}}^{(y)} + (\Delta_1 h)\mathcal{T}^{(h)}}{\|\boldsymbol{\mathcal{T}}\|^2}.$$

Quadratic convergence of the above algorithm is proven in [8], and in our experiments, we iterate solutions to a tolerance $\|\boldsymbol{F}\| < F_{\text{tol}} = 10^{-9}$, where $\|\cdot\| = \|\cdot\|_2$.

An initial step length Δs_0 must be chosen to initialize the continuation. For subsequent steps along the continuation curve, we choose the next step length according to

$$\Delta s_k = \sqrt{\frac{2\epsilon}{\|\boldsymbol{w}_k\|}}, \quad \boldsymbol{w}_k = \frac{1}{\Delta s_{k-1}}\left(\boldsymbol{\mathcal{T}}_k - \boldsymbol{\mathcal{T}}_{k-1}\right), \quad k = 1, 2, \ldots,$$

where ϵ is a user-defined tolerance for the absolute error in $(\hat{\boldsymbol{y}}, \hat{h})$; we set $\epsilon = 100\, F_{\text{tol}}$.

3. The double pendulum

We applied pseudo-arclength continuation as described in Section 2 to the double pendulum system of two bobs connected by a frictionless pin joint, the first moving on a circle with fixed center, the second moving on a circle centered at the first, and both moving in the presence of gravity; see Figure 2. The system is Lagrangian (and so symplectic), with Lagrangian function

$$L^{\text{DP}}_{\text{ang}} = \dot{\alpha}^2 + \frac{1}{2}\dot{\beta}^2 + \dot{\alpha}\dot{\beta}\cos(\alpha - \beta) + g\left(2\cos\alpha + \cos\beta\right),$$

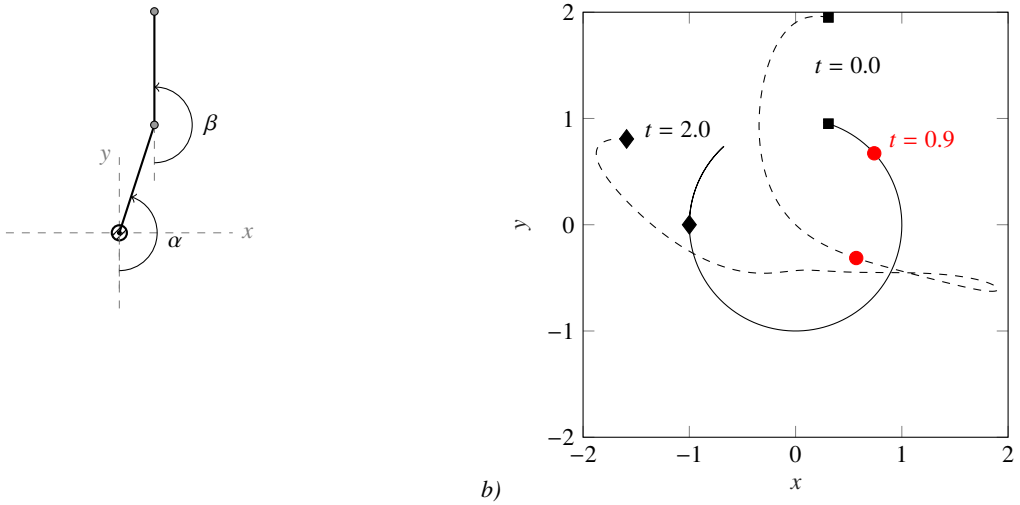
in terms of the angular coordinates (α, β) shown in Figure 2a, and $g = 9.81$. The system evolution is in accord with Euler–Lagrange differential equations of the form

$$\dot{\boldsymbol{q}}(t) = \boldsymbol{f}(\boldsymbol{q}), \quad \boldsymbol{q} = \left(\alpha, \beta, \dot{\alpha}, \dot{\beta}\right)^T.$$

For initial conditions, we use the near-vertical arrangement as in Dharmaraja et al. [1],

$$\alpha(0) = \frac{9\pi}{10}, \qquad \dot{\alpha}(0) = 0.7, \qquad \beta(0) = \pi, \qquad \dot{\beta}(0) = 0.4.$$

We compute a reference solution trajectory, $\boldsymbol{q}_{\text{ref}}$; details are given at the end of this section. In the resulting motion, the system quickly goes through a fast loop, as seen in the lower right of Figure 2b. Events such as this, where the dynamics change quickly relative to the initial or overall dynamics, seem to readily induce the bifurcations in (1.1) in which we are interested. Accordingly, we chose a point on the trajectory that is just before the loop for our $\boldsymbol{x}$ value, specifically the point corresponding to time $t = 0.9$.



a) *b)*

Figure 2. *a)* Double pendulum in its initial configuration. *b)* Reference trajectory of the double pendulum bobs used in this study. Starting from the initial configuration (black squares), the outer pendulum swings down performing a fast loop before arriving at the final configuration (black diamonds). We look at the behaviour of integrators starting from the configuration corresponding to time $t = 0.9$ (red circles), just before entering the fast loop.

The fold points we seek are a property of the implicit method used for the integration as defined by $\boldsymbol{F}(\boldsymbol{y}; \boldsymbol{x}, h)$. We now investigate the properties of these fold points for various implicit integration methods.

We first consider a Variational Taylor (VT) method [6, 7], which uses the first-order Taylor expansion of trajectories $\boldsymbol{x} \to \boldsymbol{x} + t\dot{\boldsymbol{x}}$ to obtain the *discrete Lagrangian*,

$$L_h(\boldsymbol{x}) = \int_{-h/2}^{h/2} L_{\text{ang}}^{DP}(\boldsymbol{x} + t\dot{\boldsymbol{x}})\, \mathrm{d}t.$$

The resulting one-step method is obtained by finding the critical points of the *discrete action* $L_h(\boldsymbol{y}) + L_h(\boldsymbol{x})$ subject to the constraints

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{h}{2}\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \boldsymbol{a}_1, \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \frac{h}{2}\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \boldsymbol{a}_2, \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{h}{2}\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \frac{h}{2}\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix},$$

where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are constants, the values of which are not required for the method implementation. The first two of these constraints are discrete analogues of the fixed-endpoint constraint for the continuous variational principle. The last, which connects the discrete trajectories, is a discrete analogue of the continuous constraint that the derivative of configuration is velocity. The resulting constrained optimization problem is equivalent to *discrete Euler–Lagrange equations*, and, in the case at hand, the associated Lagrange multipliers may be explicitly eliminated to obtain equations of the form (1.1). Such methods extend to any order, but they become much more complicated, and it suffices for our purpose here to consider only the first-order method [12, 13].

In addition to this, we investigate the fold points of a variety of implicit Runge–Kutta methods of the form

$$\boldsymbol{q}^{n+1} = \boldsymbol{q}^n + h\sum_{i=1}^{s} b_i \boldsymbol{k}_i, \qquad \boldsymbol{k}_i = \boldsymbol{f}\left(t_n + c_i h, \boldsymbol{q}_n + h\sum_{j=1}^{s} a_{ij}\boldsymbol{k}_j\right).$$

Specifically, using the usual Butcher tableau notation [4], we use the methods below (in row order): backward Euler (BE); trapezoidal rule (TR); the trapezoidal rule backward differentiation formula 2 split-step method, with optimized split-step size $\gamma = 2 - \sqrt{2}$ as in [1] (TRB); third- and fifth-order Radau IIA methods (R3, R5); and fourth- and sixth-order Gauss–Legendre methods (GL4, GL6). In the experiments described below, the nonlinear systems associated with any implicit method were solved using a classical Newton iteration.

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \gamma & \frac{\gamma}{2} & \frac{\gamma}{2} & 0 \\ 1 & \frac{1}{2(2-\gamma)} & \frac{1}{2(2-\gamma)} & \frac{1-\gamma}{2-\gamma} \\ \hline & \frac{1}{2(2-\gamma)} & \frac{1}{2(2-\gamma)} & \frac{1-\gamma}{2-\gamma} \end{array}$$

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array} \qquad \begin{array}{c|ccc} \frac{2}{5}-\frac{\sqrt{6}}{10} & \frac{11}{45}-\frac{7\sqrt{6}}{360} & \frac{37}{225}-\frac{169\sqrt{6}}{1800} & -\frac{2}{225}+\frac{\sqrt{6}}{75} \\ \frac{2}{5}+\frac{\sqrt{6}}{10} & \frac{37}{225}+\frac{169\sqrt{6}}{1800} & \frac{11}{45}+\frac{7\sqrt{6}}{360} & -\frac{2}{225}-\frac{\sqrt{6}}{75} \\ 1 & \frac{4}{9}-\frac{\sqrt{6}}{36} & \frac{4}{9}+\frac{\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{4}{9}-\frac{\sqrt{6}}{36} & \frac{4}{9}+\frac{\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

$$\begin{array}{c|cc} \frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\ \frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad \begin{array}{c|ccc} \frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\ \frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

A reference solution for this problem was computed using timestep $h_{\text{ref}} = 2 \times 10^{-5}$ with the VT method. This reference solution was compared to that obtained from a VT integration with timestep size $h_{\text{ref}}/2$. The state of the system at $t = 2.0$ is given in Table 1, where we see agreement of 8 significant digits between these solutions (Table 1).

Table 1. State of the reference solution at final time $t = 2.0$ compared to that obtained by halving the timestep. Matching digits are in bold font.

	h_{ref}	$h_{\text{ref}}/2$	**Digits**
α	**−1.570737**451319846	**−1.570737**439361913	8
β	**3.773018**950076258	**3.773018**944217077	8
$\dot{\alpha}$	**4.118116**660671203	**4.118116**638219623	8
$\dot{\beta}$	**−6.273626**026547350	**−6.273626**000367146	8

To demonstrate the potential effects of convergence to non-principal-branch solutions from an implicit one-step method, we show inconsistent solutions obtained with the VT method and timestep size $h = 0.1225$ in Figure 3. The nonlinear solver used in computing the solution at the next timestep converges in all steps displayed, yet the inconsistent solutions appear markedly different from the reference solution of Figure 2. There is even a difference in trajectories with inconsistent solutions when all that is changed is the initialization of the nonlinear solver, in this case, from the solution at the previous timestep to a linear extrapolation based on the solution at the previous two timesteps; this implies the existence of at least two non-principal solution branches at some point along the time integration.

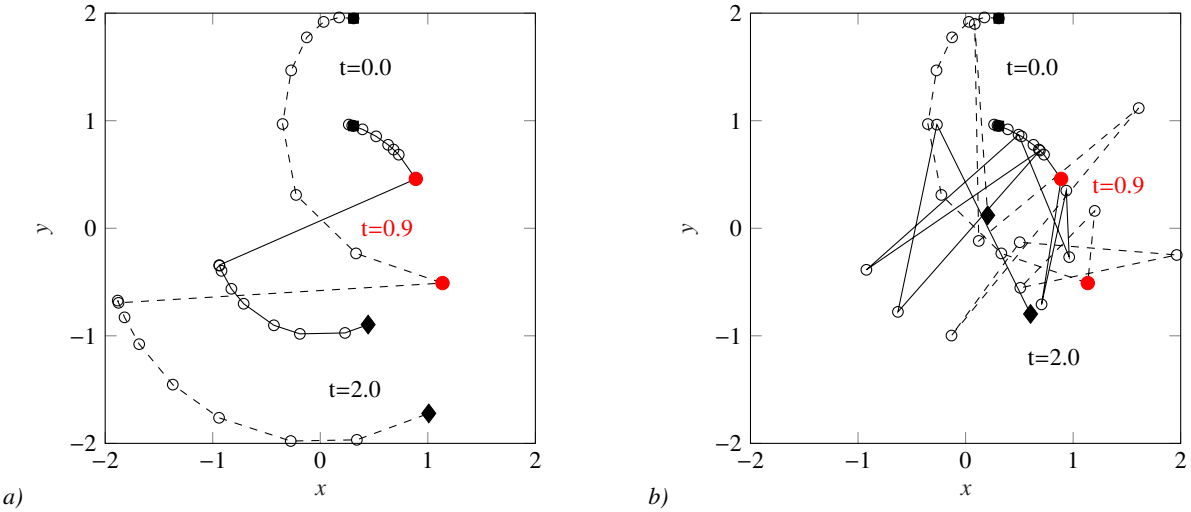


Figure 3. Simulations of the double pendulum using a timestep of $h = 0.1225$ with the VT method. *a)* The nonlinear solver of a VT timestep (initialized with the solution at the previous timestep) converges to a solution, but the trajectory differs substantially from the true trajectory as shown in Figure 2. *b)* Changing the initialization of the nonlinear solver to a linear extrapolation based on the solution at the previous two timesteps, the simulation still converges, but the resulting trajectory diverges even further.

All of the methods tested exhibit fold points before $h = 0.33$, as seen in Figure 4. Four of the methods (BE, TR, TRB, and R3) have two fold points, resulting in a continuous connection between the $h = 0$ solution and the $h = 0.35$ solution. Three of the methods (R5, GL4, and GL6) have a single fold point, beyond which we were unable to find other solutions. The solution obtained with the VT method exhibits two separate solution branches. The main branch originating from the $h = 0$ solution

folds back, with norm approaching infinity as h returns to zero. A second branch extends beyond the fold point, allowing for solutions for larger h that are not continuously and monotonically connected to the $h = 0$ solution.

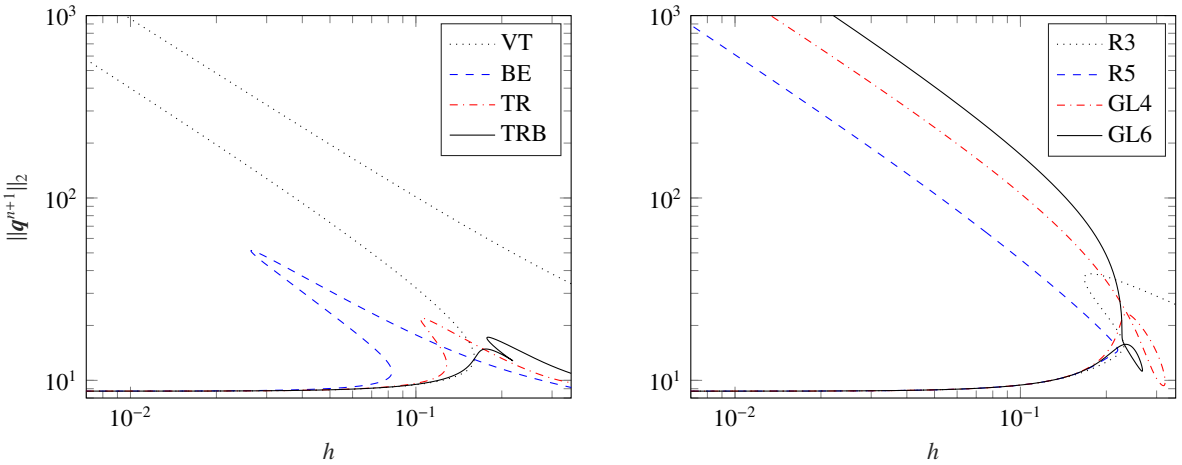


Figure 4. Norm of the computed state for various integrators. Point on the trajectory used for initialization is at $t = 0.9$ from the reference solution pictured in Figure 2b.

The relative error in the trajectories, computed as

$$e_{\text{rel}} = \frac{\|q^{n+1} - q^{n+1}_{\text{ref}}\|_2}{\|q^{n+1}_{\text{ref}}\|_2},$$

is visualized for all solution curves in Figure 5. We notice that the computed solutions near the fold points for each of the methods have relatively large (typically $O(1)$) errors in this instance and that the methods are out of their regions of asymptotic convergence. Passing through the fold points in arclength, the error generally increases as the timestep decreases.

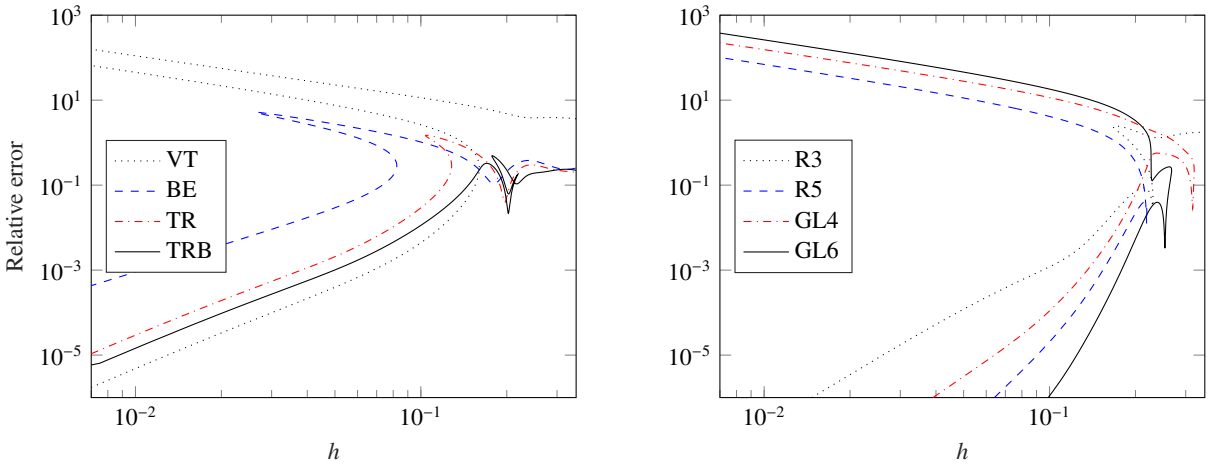


Figure 5. Relative errors in the state q^{n+1}. These errors are large (typically $O(1)$) at the fold points.

4. Discussion

Bifurcations, as a function of the timestep, are to be expected in the solutions of the update equations defined by implicit numerical methods for nonlinear IVPs. As demonstrated, such bifurcations can occur in purely mathematical examples, such as the backward Euler method applied to the differential equation $\dot{q} = q^2$, as well as in more physically interesting models, such as the double pendulum.

For simulations where the accuracy of single trajectories is important and variable step sizes are used, standard step-size control may reduce the timestep to help prevent convergence to inconsistent solutions. Such convergence, however, may persist even at small timesteps, depending on how the iterative solver for a given method is initialized. Accordingly, convergence of an implicit method solver alone is unreliable for assessing whether or not a solution is consistent or whether a simulation should be carried on. In fact, it may be possible to identify a critical timestep size h_c, corresponding to the first point at which a bifurcation of the principal solution branch occurs, such that timesteps $h > h_c$ can be considered to be invalid and solutions obtained are by definition inconsistent.

In other instances, such as the long-time integration of symplectic systems, where timesteps are often constant or more generally where consideration of backward error is a driving motivation, specific trajectories may not be as important as the general behaviour of the system. Even though the $O(1)$ relative errors we observe in trajectories at fold points in the double pendulum example imply that the solutions may be inconsistent for timesteps smaller than the critical timestep, it is helpful to be aware of the presence of fundamental method- and state-dependent limitations on the size of the valid timesteps. Simulations with timestep sizes that are larger than the critical timestep cannot be expected to generate reasonable and robust results in practice.

We used pseudo-arclength continuation to follow the solution branches of the numerical solution to the double pendulum problem using several common implicit numerical methods. Pseudo-arclength continuation can be used in this fashion as a post hoc solution validation technique. That is, given a simulation, we can use the pseudo-arclength continuation procedure described to verify that the solutions at each timestep lie on the principal solution branch. In principle, this can be done in parallel with the computations of the time-advanced state. It would be interesting to develop such a tool to monitor timesteps of implicit methods and to flag if an inconsistent solution is generated.

Acknowledgements

This work was funded by the National Science and Engineering Research Council of Canada under grant number 228090-2013. The authors thank the anonymous referees for their insightful comments.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. S. Dharmaraja, Y. Wang, G. Strang, *Optimal stability for trapezoidal-backward difference split-steps*, IMA J. Numer. Anal., **30** (2010), 141–148.

2. W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamical Equilibria*, Society for Industrial and Applied Mathematics, 2000.
3. E. Hairer, C. Lubich, G. Wanner, *Geometric numerical integration: Structure-preserving algorithms for ordinary differential equations*, Springer-Verlag, 2006.
4. E. Hairer, S. P. Nørsett, G. Wanner, *Solving ordinary differential equations I: Nonstiff problems.*, Springer-Verlag, 1993.
5. A. Stuart, A.R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 1998.
6. J. E. Marsden and M. West, *Discrete Mechanics and Variational Integrators*, Acta Numer., **10** (2001), 357–514.
7. G. W. Patrick, C. Cuell, *Error analysis of variational integrators of unconstrained Lagrangian systems*, Numer. Math., **113** (2009), 243–264.
8. P. Deuflhard, B. Fiedler, P. Kunkel, *Efficient Numerical Pathfollowing Beyond Critical Points*, SIAM J. Numer. Anal., **24** (1987), 912–927.
9. U. M. Ascher, R. M. M. Mattheij, R. D. Russell, *Numerical solution of boundary value problems for ordinary differential equations*, Industrial and Applied Mathematics (SIAM), 1995.
10. C. B. Haselgrove, *The Solution of Non-Linear Equations and of Differential Equations with Two-Point Boundary Conditions*, The Computer Journal, **4** (1961), 255–259.
11. P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer Publishing Company, 2011.
12. G. W. Patrick, C. Cuell, R. J. Spiteri, et al. *On converting any one-step method to a variational integrator of the same order*, ASME 2009 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, **4** (2009), 341–349.
13. G. W. Patrick, *Variational discretizations: discrete tangent bundles, local error analysis, and arbitrary order variational integrators*, AIP Conference Proceedings, **1168** (2009), 1013–1016.
14. A. R. Humphries, *Spurious solutions of numerical methods for initial value problems*, IMA J. Numer. Anal., **13** (1993), 263–290.
15. A. Iserles, A. T. Peplow, A. M. Stuart, *A unified approach to spurious solutions introduced by time discretisation. I: Basic theory*, SIAM J. Numer. Anal., **28** (1991), 1723–1751.
16. R. Schreiber, H. B. Keller, *Spurious solutions in driven cavity calculations*, J. Comput. Phys., **49** (1983), 165–172.
17. T. Murdoch, C. J. Budd, *Convergent and spurious solutions of nonlinear elliptic equations*, IMA J. Numer. Anal., **12** (1992), 365–386.
18. A. B. Stephens, G. R. Shubin, *Multiple Solutions and Bifurcation of Finite Difference Approximations to Some Steady Problems of Fluid Dynamics*, SIAM Journal on Scientific and Statistical Computing, **2** (1981), 404–415.
19. D. F. Griffiths, P. K. Sweby, H. C. Yee, *On spurious asymptotic numerical solutions of explicit Runge-Kutta methods*, IMA J. Numer. Anal., **12** (1992), 319–338.

20. E. Hairer, *Variable time step integration with symplectic methods*, Appl. Numer. Math., **25** (1997), 219–227.

21. S. Blanes, C. J. Budd, *Adaptive Geometric Integrators for Hamiltonian Problems with Approximate Scale Invariance*, SIAM Journal on Scientific Computing, **26** (2005), 1089–1113.

22. E. Hairer, G. Söderlind, *Explicit, time reversible, adaptive step size control*, SIAM J. Sci. Comput., **26** (2005), 1838–1851.

23. B. Leimkuhler, S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, 2004.

24. M. Schatzman, *Numerical analysis: A mathematical introduction*, Clarendon Press, Oxford, 2002.

25. A. Iserles, *A first course in the numerical analysis of differential equations*, Cambridge University Press, Cambridge, 2009.

26. J. Guchenheimer and P. Holmes, *Nonlinear oscillations, dynamical systems, and bifurcation of vector fields*, Springer-Verlag, 1983.

27. Y. A. Kuznetsov, *Elements of applied bifurcation theory*, Springer-Verlag, 2004.

28. U. Kirchgraber, *Multistep methods are essentially one-step methods*, Numerische Mathematik, **48** (1986), 85–90.

29. J. C. Phillips, R. Braun, W. Wang, et al. *Scalable molecular dynamics with NAMD*, J. Comput. Chem., **26** (2005), 1781–1802.

30. C. D. Cantwell, D. Moxey, A. Comerford, et al. *Nektar plus plus : An open-source spectral/hp element framework*, Comput. Phys. Commun., **192** (2015), 205–219.

31. J. Pitt-Francis, M. O. Bernabeu, J. Cooper, et al. *Chaste: using agile programming techniques to develop computational biology software*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, **366** (1878), 3111–3136.

32. J. Juno, A. Hakim, J. TenBarge, et al. *Discontinuous Galerkin algorithms for fully kinetic plasmas*, J. Comput. Phys., **353** (2018), 110–147.

33. S. Reich, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal., **36** (1999), 1549–1570.

AIMS Press