Numerical event location techniques in discontinuous differential algebraic equations

L. Lopez [a], S. Maset [b,*]

[a] Dipartimento di Matematica, Università di Bari, Italy
[b] Dipartimento di Matematica e Geoscienze, Università di Trieste, Italy

ARTICLE INFO

Article history:
Received 28 February 2022
Accepted 14 March 2022
Available online 29 March 2022

Keywords:
Discontinuous differential algebraic equation
Event location
Continuous extension

ABSTRACT

In this paper, we consider numerical methods for the location of events of differential algebraic equations of index one. These events correspond to cross a discontinuity surface, beyond which another differential algebraic equation holds. Convergence theorems of the numerical event time and event point to the true event time and event point are given. It is proved that, for integrations by semi-implicit methods or Rosenbrock methods, the order of convergence of the numerical event location is the order of convergence of the continuous extension and, for integration by implicit Runge-Kutta methods, the order of convergence is the order of convergence of the implicit RK method.

1. Introduction

Several problems in applications may be described by the Differential Algebraic Equation (DAE)

$$\begin{cases} y'(t) = f(y(t), z(t)), \ t \in [0,T], \\ g(y(t), z(t)) = 0, \ t \in [0,T], \\ (y(0), z(0)) = (y_0, z_0), \end{cases} \tag{1}$$

where $(y(t), z(t)) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $(y_0, z_0) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $(f, g) : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. The spaces $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ are equipped with norms, both denoted by $\|\cdot\|$.

We assume that the initial values (y_0, z_0) are *consistent*, i.e. $g(y_0, z_0) = 0$, f and g are sufficiently smooth functions and the DAE (1) has a unique solution on $[0, T]$.

Moreover, we assume that the DAE (1) has *index* 1 (see for instance [14]). This means that there exist neighborhoods U_1 of $\{y(t) : t \in [0,T]\}$ and U_2 of $\{z(t) : \ t \in [0,T]\}$ such that, for any $(y, z) \in U_1 \times U_2$, the jacobian matrix $g_z(y, z)$ of g at (y, z) is invertible and the inverse is uniformly bounded on $U_1 \times U_2$. Then, by the Implicit Function Theorem, there exists a smooth function $G : U_1 \to U_2$ such that

$$g(y, z) = 0 \ \Leftrightarrow \ z = G(y), \ (y, z) \in U_1 \times U_2. \tag{2}$$

In this paper we are interested in a DAE (1) whose solution (y, z) meets, during its evolution in the state space $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, a surface Σ at a certain time $t^* \in [0, T)$, i.e. we have $(y(t), z(t)) \notin \Sigma$ for $t \in [0, t^*)$ and $(y(t^*), z(t^*)) \in \Sigma$.

* Corresponding author.
E-mail addresses: luciano.lopez@uniba.it (L. Lopez), maset@units.it (S. Maset).

This surface Σ is a considered as a discontinuity surface, since we suppose that beyond Σ another DAE with a different (f, g) holds.

Several real systems may be modeled by DAEs of this type, see for instance the applications in Chemical Engineering [1,4,26,27] and electrical systems [20,21]. In recent years a growing interest has been observed concerning the theoretical aspects (see for example [4,7,11]) together with the numerical questions that arise in such DAEs (see for example [15,16,18, 19,23]). In literature such DAEs, are called in several ways: *non-smooth DAEs*, *hybrid DAEs*, *discontinuous DAEs*, *DAEs of Filippov type*. We will call these DAEs as DDAEs (Discontinuous DAEs).

What happens to the solution for $t > t^*$, that is once the solution has met the discontinuity surface Σ, is not the aim of this paper. In general, (y, z) could cross Σ, or could slide on Σ, or could not exist, or could not be unique (see for instance [8] where a deep study of all these questions is addressed).

In this paper, we focus on the important computational task of the *numerical event location* for DDAEs, which is the computational detection of the *event time* t^* and the *event point* $(y(t^*), z(t^*)) \in \Sigma$. The accurate computation of the event point on Σ is a crucial question in the construction of numerical schemes for this kind of problems, in particular when the solution $(y(t), z(t))$ slides on the discontinuity surface for $t > t^*$. For the similar task in case of discontinuous Ordinary Differential Equations (ODEs), techniques based on one-step methods (see for instance [9,10,17]) or multistep methods (see for instance [3]) are used. In case of DDAEs we have the additional difficulty that the event point must be consistent with the algebraic constraint.

In this paper, we consider standard numerical integrations of the DAE (1) given by semi-implicit methods, Rosenbrock methods and implicit Runge-Kutta (RK) methods. For the integration, we consider a mesh

$$0 = t_0 < t_1 < \cdots < t_N = T$$

over the interval $[0, T]$ of stepsizes

$$\tau_{n+1} := t_{n+1} - t_n,\ n = 0, 1, \ldots, N-1.$$

The maximun stepsize is denoted by $\tau_{\max}$. The methods provide approximations (y_n, z_n) of $(y(t_n), z(t_n))$, $n = 0, 1, \ldots, N$.

We show how the numerical event location can be accomplished during the integration by these three families of methods, and convergence theorems of the numerical event time and event point to the true event time and event point are given. Some test examples of numerical event location are presented. The appendix gives a background on the numerical integration of DAEs and continuous extensions.

2. Event location for DAEs

We begin with some definitions better defining the event location problem.

Let $h : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ be a sufficiently smooth function partitioning the state space $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ in the three subsets:

$$S^- = \{(y, z) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} : h(y, z) < 0\}$$
$$\Sigma = \{(y, z) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} : h(y, z) = 0\}$$
$$S^+ = \{(y, z) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} : h(y, z) > 0\}.$$

Observe that the regions S^- and S^+ are separated by the surface Σ. We can suppose that the DAE (1) holds in $S^- \cup \Sigma$, while in $S^+ \cup \Sigma$ a similar DAE holds but with a different (f, g). For the aim of this paper we can consider just the DAE in $S^- \cup \Sigma$, ignoring what happens in $S^+ \cup \Sigma$. However, we assume that (f, g) in $S^- \cup \Sigma$ can be smoothly extended to a neighborhood of Σ in S^+.

Suppose to have an initial value (y_0, z_0) such that $(y_0, z_0) \in S^-$. We want to integrate the DAE (1) up to the first time $t^* > 0$ such that

$$h\left(y\left(t^*\right), z\left(t^*\right)\right) = 0. \tag{3}$$

The time t^* is called the *event time* and $(y(t^*), z(t^*))$ is called the *event point* or *event state*. The *event location* is the determination of the event time t^* and the event point $(y(t^*), z(t^*))$.

In the following we assume that the event time t^* exists and it is a simple root of the event equation (3), i.e.

$$\frac{d}{dt} h(y(t), z(t))\Big|_{t=t^*} > 0$$

holds. Observe that this is the generic case.

2.1. The numerical event equation

We rewrite the equation (3), determining the event time and the event point, as

$$h(\varphi(t^*)) = 0, \tag{4}$$

where $\varphi(t)$ is the solution $(y(t), z(t))$ of (1). Observe that t^* is the event time and $\varphi(t^*)$ is the event point.

In the numerical event location, we do not solve the equation (4) but its approximation

$$h\left(\varphi_\tau\left(t_\tau^*\right)\right) = 0, \tag{5}$$

where $\varphi_\tau(t)$ is a numerical solution of (1) defined at all times $t \in [0, T]$. Observe that t_τ^* is the numerical event time and $\varphi_\tau\left(t_\tau^*\right)$ is the numerical event point.

The function φ_τ is obtained by the numerical method with which we are integrating the DAE (1): φ_τ can be a continuous extension of the numerical solution or the numerical solution whose value at $t \in (t_n, t_{n+1})$, $n = 0, 1, \ldots, N-1$, is obtained by taking a stepsize $t - t_n$ from (t_n, y_n).

Next theorem says how close are the numerical event time and event point to the true event time and event point. In the theorem and in the following, we use the norm $\|(y, z)\| = \max\{\|y\|, \|z\|\}$ on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.

Theorem 1. *Let $t^* \in (0, T)$ be an event time which is a simple root of (4). Consider a mesh*

$$0 = t_0 < t_1 < \cdots < t_N = T$$

over $[0, T]$ and assume that φ_τ is continuous on $[0, T]$ and, for any mesh interval $[t_n, t_{n+1}]$, $n = 0, 1, \ldots, N$, the restriction of φ_τ to $[t_n, t_{n+1}]$ is continuously differentiable. If

$$\begin{aligned} &\max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\| \to 0 \\ &\max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\varphi_\tau'(t) - \varphi'(t)\right\| \to 0, \end{aligned} \tag{6}$$

as $\tau_{\max} \to 0$, then there exist $\varepsilon > 0$ and $\overline{\tau} > 0$ such that, for any mesh with $\tau_{\max} \le \overline{\tau}$, the equation (5) has a unique solution t_τ^* in $[t^* - \varepsilon, t^* + \varepsilon]$ and we have

$$\begin{aligned} &\left|t_\tau^* - t^*\right| = O\left(\max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\|\right) \\ &\left\|\varphi_\tau(t_\tau^*) - \varphi(t^*)\right\| = O\left(\max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\|\right) \end{aligned} \tag{7}$$

as $\tau_{\max} \to 0$.

Proof. Assume (6). Let $R, R_1 > 0$ be such that

$$\max_{t\in[0,T]} \|\varphi(t)\| \le R \quad \text{and} \quad \max_{t\in[0,T]} \left\|\varphi'(t)\right\| \le R_1.$$

Since (6) holds, there exists $\overline{\tau}_0 > 0$ such that, for $\tau_{\max} \le \overline{\tau}_0$, we have

$$\max_{t\in[0,T]} \|\varphi_\tau(t)\| \le 2R \quad \text{and} \quad \max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\varphi_\tau'(t)\right\| \le 2R_1. \tag{8}$$

Suppose $\tau_{\max} \le \overline{\tau}_0$.

Let $F = h \circ \varphi$ and $F_\tau = h \circ \varphi_\tau$. Both F and F_τ are functions defined in $[0, T]$ with values in $\mathbb{R}$. We have

$$\max_{t\in[0,T]} |F_\tau(t) - F(t)| \le \max_{\substack{x\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2} \\ \|x\|\le 2R}} \left\|h'(x)\right\| \cdot \max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\|. \tag{9}$$

Moreover, for any mesh interval $[t_n, t_{n+1}]$, $n = 0, 1, \ldots, N-1$, the restriction of F_τ to $[t_n, t_{n+1}]$ is continuously differentiable and we have

$$\begin{aligned} &\max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left|F_\tau'(t) - F'(t)\right| \\ &\le \max_{\substack{x\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2} \\ \|x\|\le 2R}} \left\|h'(x)\right\| \cdot \max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\varphi_\tau'(t) - \varphi'(t)\right\| \\ &\quad + \max_{\substack{x\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2} \\ \|x\|\le 2R}} \left\|h''(x)\right\| \cdot \max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\| \cdot R_1. \end{aligned}$$

So, we have

$$\max_{t\in[0,T]} |F_\tau(t) - F(t)| \to 0$$
$$\max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left|F'_\tau(t) - F'(t)\right| \to 0 \tag{10}$$

as $\tau_{\max} \to 0$ by (6).

The equations (4) and (5) read $F(t^*) = 0$ and $F_\tau(t^*_\tau) = 0$, respectively. Since $t^* \in (0,T)$ is a simple root of (4), we have $F'(t^*) > 0$. Then, there exists $\varepsilon > 0$ such that

$$F'(t) \geq \frac{F'(t^*)}{2},\ t \in \left[t^* - \varepsilon, t^* + \varepsilon\right],$$
$$F\left(t^* - \varepsilon\right) < 0,\ F\left(t^* + \varepsilon\right) > 0.$$

By reminding (10), there exists $\overline{\tau} > 0$ such that, for any $\tau_{\max} \leq \overline{\tau}$, we have

$$F'_\tau(t) \geq \frac{F'(t^*)}{4},\ t \in \left[t^* - \varepsilon, t^* + \varepsilon\right], \tag{11}$$
$$F_\tau\left(t^* - \varepsilon\right) \leq \frac{F(t^* - \varepsilon)}{2},\ F_\tau\left(t^* + \varepsilon\right) \geq \frac{F(t^* + \varepsilon)}{2}.$$

(When t is a mesh point t_n, $n = 1, \ldots, N-1$, the first inequality is true for both values $F'_\tau(t)$ coming from the restrictions to the intervals $[t_{n-1}, t_n]$ and $[t_n, t_{n+1}]$.) This shows that, for $\tau_{\max} \leq \overline{\tau}$, the function F_τ is strictly increasing and changes sign in $[t^* - \varepsilon, t^* + \varepsilon]$ and so the equation (5) has a unique solution in $[t^* - \varepsilon, t^* + \varepsilon]$.

Now, we show how the conclusions (7) follow. Suppose $\tau_{\max} \leq \min\{\overline{\tau}_0, \overline{\tau}\}$. We have (by (11))

$$\left|F_\tau\left(t^*\right)\right| = \left|F_\tau\left(t^*\right) - F_\tau\left(t^*_\tau\right)\right| \geq \frac{F'(t^*)}{4}\left|t^*_\tau - t^*\right|$$

and then

$$\left|t^*_\tau - t^*\right| \leq \frac{4\,|F_\tau(t^*)|}{F'(t^*)}.$$

Since (recall (9))

$$\left|F_\tau\left(t^*\right)\right| \leq \max_{\substack{x\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\\ \|x\|\leq 2R}} \left\|h'(x)\right\| \cdot \max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\|$$

(really

$$\left|F_\tau\left(t^*\right)\right| \leq \max_{\substack{x\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\\ \|x\|\leq 2R}} \left\|h'(x)\right\| \cdot \left\|\varphi_\tau(t^*) - \varphi(t^*)\right\|$$

holds), we obtain

$$\left|t^*_\tau - t^*\right| = O\left(\max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\|\right),\ \tau_{\max} \to 0.$$

Moreover, we have (recall (8))

$$\begin{aligned}\left\|\varphi_\tau(t^*_\tau) - \varphi(t^*)\right\| &\leq \left\|\varphi_\tau(t^*_\tau) - \varphi_\tau(t^*)\right\| + \left\|\varphi_\tau(t^*) - \varphi(t^*)\right\| \\ &\leq 2R_1\left|t^*_\tau - t^*\right| + \left\|\varphi_\tau(t^*) - \varphi(t^*)\right\|\end{aligned}$$

and then

$$\left\|\varphi_\tau(t^*_\tau) - \varphi(t^*)\right\| = O\left(\max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\|\right),\ \tau_{\max} \to 0. \quad \square$$

Note that convergence results for the numerical event location of ODEs (not DAEs) are given in [25]. However, that paper lacks of the proof of existence and uniqueness for the root of the numerical event equation and, as a consequence, conditions like (6) do not appear. Most of our proof is for showing existence and uniqueness for the solution of the numerical event equation (5). The assumption on the derivative in (6) guarantees the uniqueness of the solution.

In the next subsections we deal with the numerical event location when the DAE (1) is integrated by semi-implicit methods, Rosenbrock methods and implicit Runge-Kutta (RK) methods, which are introduced in Appendix along with their continuous extensions.

2.2. Event location for semi-implicit methods

Suppose that the DAE (1) is integrated by a semi-implicit method presented in Appendix A.1 up to an index $n = n^*$ such that $h(y_{n^*}, z_{n^*})$ and $h(y_{n^*+1}, z_{n^*+1})$ have different signs. The numerical event time t_τ^* and event point (y_τ^*, z_τ^*) are obtained by solving the nonlinear system in the unknowns t_τ^* and z_τ^*:

$$\begin{cases} g(y_\tau^*, z_\tau^*) = 0 \\ h(y_\tau^*, z_\tau^*) = 0, \end{cases} \tag{12}$$

where

$$y_\tau^* = \eta(t_\tau^*) = y_{n^*} + \tau_{n^*+1} \sum_{i=1}^{s} b_i \left(\frac{t_\tau^* - t_{n^*}}{\tau_{n^*+1}} \right) f(y_{n^*i}, z_{n^*i}), \tag{13}$$

with η the continuous extension (54). Observe that the stage values (y_{n^*i}, z_{n^*i}), $i = 1, \ldots, s$, in (13) are known since they are obtained during the step from t_{n^*} to t_{n^*+1}.

The system (12) of unknowns t_τ^* and z_τ^* has dimension $1 + d_2$. This should be compared with a step of the semi-implicit method, where s systems of dimension d_2 need to be solved.

Observe that the numerical event point (y_τ^*, z_τ^*) is consistent and it is on the surface Σ, since we require both these conditions in the equations (12).

Next theorem shows that the order of convergence of the numerical event time and event point is equal to the order of convergence of the continuous extension, whenever the event time t^* is a simple root of the event equation (3).

Theorem 2. *Suppose that the DAE (1) is integrated by a semi-implicit method over a mesh*

$$0 = t_0 < t_1 < \cdots < t_N = T$$

on the interval $[0, T]$. *Let* $t^* \in (0, T)$ *be an event time which is a simple root of (3). Suppose that the numerical event time and event point are obtained by solving the system (12), once an index* $n = n^*$ *such that* $h(y_{n^*}, z_{n^*})$ *and* $h(y_{n^*+1}, z_{n^*+1})$ *have different signs is detected.*

Suppose that the semi-implicit method has order p *and the continuous extension (54) has order* q, *where* q *is a positive integer. So, the continuous extension has convergence order* $\min\{p, q+1\}$.

Then, there exists a neighborhood $\mathcal{U}$ *of* $(t^*, y(t^*), z(t^*))$ *such that, for any mesh with* $\tau_{\max}$ *sufficiently small, there exists a unique numerical event time and event point* $(t_\tau^*, y_\tau^*, z_\tau^*)$ *in* $\mathcal{U}$ *and we have*

$$\begin{aligned} \left| t_\tau^* - t^* \right| &= O\left(\tau_{\max}^{\min\{p,q+1\}} \right) \\ \| y_\tau^* - y(t^*) \| &= O\left(\tau_{\max}^{\min\{p,q+1\}} \right) \\ \| z_\tau^* - z(t^*) \| &= O\left(\tau_{\max}^{\min\{p,q+1\}} \right) \end{aligned}$$

as $\tau_{\max} \to 0$.

Proof. The continuous extension η is given by

$$\begin{aligned} &\eta(t) = y_n + \tau_{n+1} \sum_{i=1}^{s} b_i \left(\frac{t - t_n}{\tau_{n+1}} \right) f(y_{ni}, z_{ni}) \\ &n = 0, 1, \ldots, N-1 \text{ and } t \in [t_n, t_{n+1}]. \end{aligned}$$

Since the continuous extension has order $q \geq 1$, we have

$$\sum_{i=1}^{s} b_i(\theta) = \theta, \ \theta \in [0, 1],$$

and then

$$\sum_{i=1}^{s} b_i'(\theta) = 1, \ \theta \in [0, 1]. \tag{14}$$

Let U_1 and U_2 be the neighborhoods of $\{y(t) : t \in [0, T]\}$ and $\{z(t) : t \in [0, T]\}$, respectively, defined in Section 1 and let $G : U_1 \to U_2$ be the smooth function also defined in Section 1.

If $y_{ni} \in U_1$ and $y_{n+1} \in U_1$, we consider $z_{ni} = G(y_{ni})$ and $z_{n+1} = G(y_{n+1})$ as the roots of the equations (41) and (43), respectively, of the semi-implicit method (other roots not in U_2 are not considered). Moreover, if $y_\tau^* = \eta(t_\tau^*) \in U_1$, we rewrite the first equation in (12) as

$$z_\tau^* = G(y_\tau^*).$$

For a sufficiently small $\tau_{\max}$, we have

$$y_{ni} \in U_1,\ z_{ni} = G(y_{ni}),\ y_{n+1} \in U_1 \text{ and } z_{n+1} = G(y_{n+1})$$
$$n = 0, 1, \ldots, N-1 \text{ and } i = 1, \ldots, s.$$

Moreover, since the continuous extension has convergence order $\min\{p, q+1\} \geq q \geq 1$, we have

$$\max_{t\in[0,T]} \|\eta(t) - y(t)\| \to 0,\ \tau_{\max} \to 0, \tag{15}$$

and then

$$\eta(t) \in U_1,\ t \in [0,T],$$

in particular $y_\tau^* = \eta(t_\tau^*) \in U_1$, for a sufficiently small $\tau_{\max}$. We define

$$F(y) = f(y, G(y)),\ y \in U_1.$$

The functions φ and φ_τ appearing in (4) and (5) are

$$\varphi(t) = (y(t), G(y(t))),\ t \in [0,T],$$

and

$$\varphi_\tau(t) = (\eta(t), G(\eta(t))),\ t \in [0,T].$$

The function φ_τ is continuous in $[0,T]$ and, for any mesh interval $[t_n, t_{n+1}]$, $n = 0, 1, \ldots, N-1$, the restriction of φ_τ to $[t_n, t_{n+1}]$ is continuously differentiable: we have

$$\varphi_\tau'(t) = \left(\eta'(t), G'(\eta(t))\eta'(t)\right),\ t \in [t_n, t_{n+1}],$$

where

$$\eta'(t) = \sum_{i=1}^{s} b_i'\left(\frac{t-t_n}{\tau_{n+1}}\right) F(y_{ni}),\ t \in [t_n, t_{n+1}].$$

Observe that

$$\max_{t\in[t_n,t_{n+1}]} \left\|\eta'(t)\right\| \leq \max_{\theta\in[0,1]} \sum_{i=1}^{s} \left|b_i'(\theta)\right| \cdot \sup_{y\in U_1} \|F(y)\|.$$

We have

$$\max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\| \leq \max\{1, \mathrm{Lip}(G)\} \max_{t\in[0,T]} \|\eta(t) - y(t)\|, \tag{16}$$

where $\mathrm{Lip}(G)$ is the Lipschitz constant of the function G. Moreover,

$$\begin{aligned}
&\max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\varphi_\tau'(t) - \varphi'(t)\right\| \\
&\leq \max\Bigg\{ \max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\eta'(t) - y'(t)\right\|, \\
&\qquad \mathrm{Lip}\left(G'\right) \max_{t\in[0,T]} \|\eta(t) - y(t)\| \max_{\theta\in[0,1]} \sum_{i=1}^{s} \left|b_i'(\theta)\right| \cdot \sup_{y\in U_1} \|F(y)\| \\
&\qquad + \sup_{t\in[0,T]} \left\|G'(y(t))\right\| \cdot \max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\eta'(t) - y'(t)\right\| \Bigg\},
\end{aligned} \tag{17}$$

where $\mathrm{Lip}\left(G'\right)$ is the Lipschitz constant of the function G'.

Beside (15), we also have

$$\max_{n=0,1,\dots,N-1}\ \max_{t\in[t_n,t_{n+1}]}\left\|\eta'(t)-y'(t)\right\|\to 0,\ \tau_{\max}\to 0.$$

In fact, for $n=0,1,\dots,N-1$ and $t\in[t_n,t_{n+1}]$, we have (recall (14))

$$\eta'(t)-y'(t)=\sum_{i=1}^{s}b_i'\left(\frac{t-t_n}{\tau_{n+1}}\right)\left(F(y_{ni})-y'(t)\right).$$

Then

$$\max_{n=0,1,\dots,N-1}\ \max_{t\in[t_n,t_{n+1}]}\left|\eta'(t)-y'(t)\right|$$

$$\le\max_{\theta\in[0,1]}\sum_{i=1}^{s}\left|b_i'(\theta)\right|\max_{n=0,1,\dots,N-1}\ \max_{t\in[t_n,t_{n+1}]}\ \max_{i=1,\dots,s}\left\|F(y_{ni})-y'(t)\right\|$$

with

$$\max_{n=0,1,\dots,N-1}\ \max_{t\in[t_n,t_{n+1}]}\ \max_{i=1,\dots,s}\left\|F(y_{ni})-y'(t)\right\|\to 0,\ \tau_{\max}\to 0. \tag{18}$$

Regarding (18), observe that

$$F(y_{ni})=F(y_n)+R_{ni}$$

and

$$y'(t)=F(y(t))=F(y_n)+S_n(t)$$

with

$$\max_{n=0,1,\dots,N-1}\ \max_{i=1,\dots,s}\|R_{ni}\|\to 0,\ \tau_{\max}\to 0,$$

and

$$\max_{n=0,1,\dots,N-1}\ \max_{t\in[t_n,t_{n+1}]}\|S_n(t)\|\to 0,\ \tau_{\max}\to 0.$$

Now, by (16) and (17) we have

$$\max_{t\in[0,T]}\|\varphi_\tau(t)-\varphi(t)\|\to 0$$

$$\max_{n=0,1,\dots,N-1}\ \max_{t\in[t_n,t_{n+1}]}\left\|\varphi_\tau'(t)-\varphi'(t)\right\|\to 0,$$

as $\tau_{\max}\to 0$. The thesis follows by Theorem 1. □

When the DAE (1) is integrated by a semi-implicit method, the numerical event location has the drawback that the order of convergence of the numerical event time and event point is $\min\{p,q+1\}$, which is in general less than the order of convergence p of the semi-implicit method.

2.3. Event location for Rosenbrock methods

Now, suppose that the DAE (1) is integrated by a Rosenbrock method of Appendix A.2.2 up to an index $n=n^*$ such that $h(y_{n^*},z_{n^*})$ and $h(y_{n^*+1},z_{n^*+1})$ have different signs. The numerical event time t_τ^* and event point (y_τ^*,z_τ^*) are obtained by solving the equation

$$h(y_\tau^*,z_\tau^*)=0, \tag{19}$$

where

$$y_\tau^*=\eta(t_\tau^*)=y_{n^*}+\sum_{i=1}^{s}b_i\left(\frac{t_\tau^*-t_{n^*}}{\tau_{n^*+1}}\right)l_{ni}$$

$$z_\tau^*=\mu(t_\tau^*)=z_{n^*}+\sum_{i=1}^{s}b_i\left(\frac{t_\tau^*-t_{n^*}}{\tau_{n^*+1}}\right)k_{ni}$$

with (η, μ) the continuous extension (56).

The equation (19) is a scalar equation in the unknown t_τ^*. So, the computational cost of the numerical event location is a small fraction of the computational cost of a step of the Rosenbrock method, where s linear systems of dimension $d_1 + d_2$ need to be solved.

Observe that, unlike the case of semi-implicit methods, the numerical event point (y_τ^*, z_τ^*) is not consistent, in general.

The order of convergence of the numerical event time and event point is equal to the order of convergence of the continuous extension, as in case of semi-implicit methods. However, we have to assume

$$\frac{\tau_{\max}^p}{\tau_{\min}} \to 0, \ \tau_{\max} \to 0, \tag{20}$$

where

$$\tau_{\min} := \min_{n=0,\ldots,N-1} \tau_{n+1}$$

and p is the differential algebraic order of the Rosenbrock method. The reason for introducing this assumption is that, in general, the numerical solutions (y_n, z_n), $n = 0, \ldots, N$, are not consistent. The assumption (20) is an assumption on the type of mesh we are using for the integration of the DAE. For a constant stepsize mesh, it holds if and only if $p > 1$. For a variable stepsize mesh with $\tau_{\min} = O\left(\tau_{\max}^\alpha\right)$, $\tau_{\max} \to 0$, for some $\alpha \geq 1$, it holds if and only if $\alpha < p$.

Theorem 3. *Suppose that the DAE (1) is integrated by a Rosenbrock method over a mesh*

$$0 = t_0 < t_1 < \cdots < t_N = T$$

on the interval $[0, T]$. *Let* $t^* \in (0, T)$ *be an event time which is a simple root of (4). Suppose that the numerical event time and event point are obtained by solving the equation (19), once an index* $n = n^*$ *such that* $h\left(y_{n^*}, z_{n^*}\right)$ *and* $h\left(y_{n^*+1}, z_{n^*+1}\right)$ *have different signs is detected.*

Suppose that the Rosenbrock method has differential algebraic order p *and stability function* R *such that* $|R(\infty)| < 1$ *and the continuous extension (56) has differential algebraic order* q, *where* q *is a positive integer. So, the continuous extension has convergence order* $\min\{p, q+1\}$.

Suppose that the assumption (20) holds.

Then, there exists a neighborhood $\mathcal{U}$ *of* $(t^*, y(t^*), z(t^*))$ *such that, for any mesh with* $\tau_{\max}$ *sufficiently small, there exists a unique numerical event time and event point* $(t_\tau^*, y_\tau^*, z_\tau^*)$ *in* $\mathcal{U}$ *and we have*

$$\left|t_\tau^* - t^*\right| = O\left(\tau_{\max}^{\min\{p,q+1\}}\right)$$
$$\left\|y_\tau^* - y(t^*)\right\| = O\left(\tau_{\max}^{\min\{p,q+1\}}\right)$$
$$\left\|z_\tau^* - z(t^*)\right\| = O\left(\tau_{\max}^{\min\{p,q+1\}}\right)$$

as $\tau_{\max} \to 0$.

Proof. The continuous extension (η, μ) is given by

$$(\eta(t), \mu(t)) = (y_n, z_n) + \sum_{i=1}^{s} b_i\left(\frac{t - t_n}{\tau_{n+1}}\right)(l_{ni}, k_{ni}),$$
$$n = 0, 1, \ldots, N-1 \ \text{ and } t \in [t_n, t_{n+1}].$$

Since the continuous extension has differential algebraic order $q \geq 1$, we have

$$\sum_{i=1}^{s} b_i(\theta) = \theta, \ \theta \in [0, 1],$$

and then

$$\sum_{i=1}^{s} b_i'(\theta) = 1, \ \theta \in [0, 1]. \tag{21}$$

The functions φ and φ_τ appearing in (4) and (5) are

$$\varphi(t) = (y(t), z(t)), \ t \in [0, T],$$

and

$$\varphi_\tau(t) = (\eta(t), \mu(t)), \ t \in [0, T].$$

Since the continuous extension has convergence order $\min\{p, q+1\} \geq q \geq 1$, we have

$$\max_{t\in[0,T]} \|\varphi_\tau(t) - \varphi(t)\| \to 0, \ \tau \to 0. \tag{22}$$

Moreover, we also have

$$\max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\varphi'_\tau(t) - \varphi'(t)\right\| \to 0, \ \tau \to 0. \tag{23}$$

In fact, for $n = 0, 1, \ldots, N-1$ and $t \in [t_n, t_{n+1}]$, we have (recall (21))

$$\varphi'_\tau(t) - \varphi'(t) = \sum_{i=1}^{s} b'_i\left(\frac{t-t_n}{\tau_{n+1}}\right)\left(\frac{(l_{ni}, k_{ni})}{\tau_{n+1}} - \left(y'(t), z'(t)\right)\right)$$

and then

$$\begin{aligned}
&\max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \left\|\eta'(t) - y'(t)\right\| \\
&\leq \max_{\theta\in[0,1]} \sum_{i=1}^{s} \left|b'_i(\theta)\right| \max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \max_{i=1,\ldots,s} \left\|\frac{(l_{ni}, k_{ni})}{\tau_{n+1}} - \left(y'(t), z'(t)\right)\right\|
\end{aligned}$$

with

$$\max_{t\in[t_n,t_{n+1}]} \max_{i=1,\ldots,s} \left\|\frac{(l_{ni}, k_{ni})}{\tau_{n+1}} - \left(y'(t), z'(t)\right)\right\| \to 0, \ \tau \to 0. \tag{24}$$

Regarding (24), observe that

$$\frac{(l_{ni}, k_{ni})}{\tau_{n+1}} = \left(f(y_n, z_n), -\left(g_z^n\right)^{-1} g_y^n f(y_n, z_n)\right) + R_{ni}$$

and

$$\begin{aligned}
\left(y'(t), z'(t)\right) &= \left(f(y(t), z(t)), -g_z^{-1}\left(y(t), z(t)\right) g_y\left(y(t), z(t)\right) f(y(t), z(t))\right) \\
&= \left(f(y_n, z_n), -\left(g_z^n\right)^{-1} g_y^n f(y_n, z_n)\right) + S_n(t)
\end{aligned}$$

with

$$\max_{n=0,1,\ldots,N-1} \max_{i=1,\ldots,s} \|R_{ni}\| \to 0, \ \tau_{\max} \to 0, \tag{25}$$

and

$$\max_{n=0,1,\ldots,N-1} \max_{t\in[t_n,t_{n+1}]} \|S_n(t)\| \to 0, \ \tau_{\max} \to 0.$$

Observe that (25) holds if

$$\max_{n=0,1,\ldots,N-1} \frac{\|g(y_n, z_n)\|}{\tau_{n+1}} \to 0, \ \tau_{\max} \to 0, \tag{26}$$

and (26) holds if (20) holds.

Since (22) and (23) hold, the thesis follows by Theorem 1. □

When the DAE (1) is integrated by a Rosenbrock method, we have the same drawback encountered for a semi-implicit method, namely the order of convergence of the numerical event time and event point is $\min\{p, q+1\}$, which is in general less that the order of convergence p of the Rosenbrock method.

Finally, suppose that the DAE (1) is integrated by an implicit RK method of Appendix A.2.1 up to an index $n = n^*$ such that $h(y_{n^*}, z_{n^*})$ and $h(y_{n^*+1}, z_{n^*+1})$ have different signs. The numerical event time t^*_τ and event point (y^*_τ, z^*_τ) are not obtained by using a continuous extension of the numerical solution (y_n, z_n), but simply by seeing (y^*_τ, z^*_τ) as new values (y_{n^*+1}, z_{n^*+1}) corresponding to a new unknown step $\tau^* = t^*_{n^*+1} - t_{n^*}$. In other words, we have (remind (48)):

$$
\begin{aligned}
t^*_\tau &= t_{n^*} + \tau^* \\
y^*_\tau &= y_{n^*} + \tau^* \sum_{i=1}^{s} b_i f(y^*_i, z^*_i) \\
z^*_\tau &= \left(1 - \sum_{i,j=1}^{s} b_i \omega_{ij}\right) z_{n^*} + \sum_{i,j=1}^{s} b_i \omega_{ij} z^*_j,
\end{aligned}
$$

where τ^* and (y^*_i, z^*_i), $i = 1, \ldots, s$, satisfy

$$
\begin{aligned}
& y^*_i = y_{n^*} + \tau^* \sum_{j=1}^{s} a_{ij} f(y^*_j, z^*_j), \ i = 1, \ldots, s, \\
& g(y^*_i, z^*_i) = 0, \ i = 1, \ldots, s, \\
& h(y^*_\tau, z^*_\tau) = 0.
\end{aligned} \tag{27}
$$

The non-linear system (27) in the unknowns τ^* and (y^*_i, z^*_i), $i = 1, \ldots, s$, has dimension $1 + s(d_1 + d_2)$. So, the computational cost of the numerical event location is essentially the same as the computational cost of a step in the implicit RK method, where a non-linear system of dimension $s(d_1 + d_2)$ has to be solved.

Observe that this technique for the numerical event location was used in [12] in the context of detecting breaking points for state-dependent delay differential equations integrated by implicit schemes.

Unlike the case of semi-implicit methods and similarly to the case of Rosenbrock methods, the numerical event point (y^*_τ, z^*_τ) is not consistent, in general. However, it is guaranteed to be consistent if the implicit RK method is stiffly accurate.

When the numerical event location problem is accomplished in this fully implicit way, we are able to overcome the drawback of the lower order of convergence of the numerical event time and event point with respect to the order of the method. We obtain, an order of convergence equal to the order of convergence of the method.

Theorem 4. *Suppose that the DAE (1) is integrated by an implicit RK method with matrix A non-singular over a mesh*

$$0 = t_0 < t_1 < \cdots < t_N = T$$

on the interval $[0, T]$. Let $t^* \in (0, T)$ be an event time which is a simple root of (4). Suppose that the numerical event time and event point are obtained by solving the system (27), once an index $n = n^*$ such that $h(y_{n^*}, z_{n^*})$ and $h(y_{n^*+1}, z_{n^*+1})$ have different signs is detected.

Suppose that the implicit RK method has differential algebraic order p and stability function R such that $|R(\infty)| < 1$, where p is a positive integer.

Then, there exists a neighborhood $\mathcal{U}$ of $(t^*, y(t^*), z(t^*))$ such that, for any mesh with $\tau_{\max}$ sufficiently small, there exists a unique numerical event time and event point $(t^*_\tau, y^*_\tau, z^*_\tau)$ in $\mathcal{U}$ and we have

$$
\begin{aligned}
& \left|t^*_\tau - t^*\right| = O\left(\tau^p_{\max}\right) \\
& \left\|y^*_\tau - y(t^*)\right\| = O\left(\tau^p_{\max}\right) \\
& \left\|z^*_\tau - z(t^*)\right\| = O\left(\tau^p_{\max}\right)
\end{aligned}
$$

as $\tau_{\max} \to 0$.

Proof. Since the implicit RK method has differential algebraic order $p \geq 1$, we have

$$\sum_{i=1}^{s} b_i = 1. \tag{28}$$

The functions φ and φ_τ appearing in (4) and (5) are

$$\varphi(t) = (y(t), z(t)), \ t \in [0, T],$$

and

$$\varphi_\tau(t) = (y_\tau(t), z_\tau(t)), \ t \in [0, T],$$

where, for $n = 0, 1, \ldots, N-1$ and $t \in [t_n, t_{n+1}]$, we have

$$\begin{aligned} y_\tau(t) &= y_n + (t - t_n) \sum_{i=1}^{s} b_i f(y_{ni}(t), z_{ni}(t)) \\ z_\tau(t) &= \left(1 - \sum_{i,j=1}^{s} b_i \omega_{ij}\right) z_n + \sum_{i,j=1}^{s} b_i \omega_{ij} z_{nj}(t), \end{aligned} \tag{29}$$

with $(y_i(t), z_i(t))$, $i = 1, \ldots, s$, given by

$$\begin{aligned} &y_{ni}(t) = y_n + (t - t_n) \sum_{j=1}^{s} a_{ij} f(y_{nj}(t), z_{nj}(t)), \ i = 1, \ldots, s, \\ &g(y_{ni}(t), z_{ni}(t)) = 0, \ i = 1, \ldots, s. \end{aligned} \tag{30}$$

Let U_1 and U_2 be the neighborhoods of $\{y(t) : t \in [0, T]\}$ and $\{z(t) : t \in [0, T]\}$, respectively, defined in Section 1 and let $G : U_1 \to U_2$ be the smooth function also in Section 1.

For a sufficiently small stepsize τ, we have

$$y_{ni}(t) \in U_1 \text{ and } z_{ni}(t) = G(y_{ni}(t)), \ n = 0, 1, \ldots, N-1, \ t \in [t_n, t_{n+1}] \text{ and } i = 1, \ldots, s.$$

Now, we can rewrite (29) as

$$\begin{aligned} y_\tau(t) &= y_n + (t - t_n) \sum_{i=1}^{s} b_i f(y_{ni}(t), G(y_{ni}(t))) \\ z_\tau(t) &= \left(1 - \sum_{i,j=1}^{s} b_i \omega_{ij}\right) z_n + \sum_{i,j=1}^{s} b_i \omega_{ij} G(y_{nj}(t)) \end{aligned}$$

and (30) as

$$y_{ni}(t) = y_n + (t - t_n) \sum_{j=1}^{s} a_{ij} f(y_{nj}(t), G(y_{nj}(t))), \ i = 1, \ldots, s.$$

Since the implicit RK method has convergence order $p \geq 1$, we have

$$\max_{t \in [0,T]} \|\varphi_\tau(t) - \varphi(t)\| \to 0, \ \tau \to 0. \tag{31}$$

Moreover, we also have

$$\max_{n=0,1,\ldots,N-1} \max_{t \in [t_n, t_{n+1}]} \left\|\varphi'_\tau(t) - \varphi'(t)\right\| \to 0, \ \tau \to 0. \tag{32}$$

In fact, for $n = 0, 1, \ldots, N$ and $t \in [t_n, t_{n+1}]$, we have (recall (28))

$$\begin{aligned} y'_\tau(t) - y'(t) &= \sum_{i=1}^{s} b_i \left(f(y_{ni}(t), G(y_{ni}(t))) - y'(t)\right) \\ &+ (t - t_n) \sum_{i=1}^{s} b_i \frac{d}{dt} f(y_{ni}(t), G(y_{ni}(t))) \end{aligned}$$

and then

$$\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\left\|y'_\tau(t)-y'(t)\right\|$$
$$\leq\sum_{i=1}^{s}|b_i|\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\max_{i=1,\ldots,s}\left\|f(y_{ni}(t),G(y_{ni}(t)))-y'(t)\right\|$$
$$+\tau_{\max}\sum_{i=1}^{s}|b_i|\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\max_{i=1,\ldots,s}\left\|\frac{d}{dt}f(y_{ni}(t),G(y_{ni}(t)))\right\|$$

with

$$\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\max_{i=1,\ldots,s}\left\|f(y_{ni}(t),G(y_{ni}(t)))-y'(t)\right\|\to 0,\ \tau_{\max}\to 0,$$

and

$$\limsup_{\tau_{\max}\to 0}\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\max_{i=1,\ldots,s}\left\|\frac{d}{dt}f(y_{ni}(t),G(y_{ni}(t)))\right\|<\infty.$$

In addition, for $n=0,1,\ldots,N$ and $t\in[t_n,t_{n+1}]$, we have (recall (28))

$$z'_\tau(t)-z'(t)=\sum_{i,j=1}^{s}b_i\omega_{ij}G'(y_{nj}(t))y'_{nj}(t)-z'(t)$$
$$=\sum_{i=1}^{s}b_i\left(\sum_{j=1}^{s}\omega_{ij}G'(y_{nj})y'_{nj}(t)-z'(t)\right),$$

and then

$$\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\left\|z'_\tau(t)-z'(t)\right\|$$
$$\leq\sum_{i=1}^{s}|b_i|\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\max_{i=1,\ldots,s}\left\|\sum_{j=1}^{s}\omega_{ij}G'(y_{nj})y'_{nj}(t)-z'(t)\right\|$$

with

$$\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\max_{i=1,\ldots,s}\left\|\sum_{j=1}^{s}\omega_{ij}G'(y_{nj})y'_{nj}(t)-z'(t)\right\|,\ \tau_{\max}\to 0. \tag{33}$$

Regarding (33) observe that, for $i=1,\ldots,s$, we have

$$\sum_{j=1}^{s}\omega_{ij}G'(y_{nj})y'_{nj}(t)$$
$$=\sum_{j=1}^{s}\omega_{ij}G'(y_{nj})$$
$$\cdot\left(\sum_{k=1}^{s}a_{jk}f(y_{nk}(t),G(y_{nk}(t)))+(t-t_n)\sum_{k=1}^{s}a_{jk}\frac{d}{dt}f(y_{nk}(t),G(y_{nk}(t)))\right)$$
$$=\sum_{j=1}^{s}\sum_{k=1}^{s}\omega_{ij}a_{jk}G'(y_n)f(y_n,G(y_n))+R_{ni}(t)$$
$$=G'(y_n)f(y_n,G(y_n))+R_{ni}(t)$$

(in the last equality recall that ω_{ij} are the elements of A^{-1}), where

$$\max_{n=0,1,\ldots,N-1}\max_{t\in[t_n,t_{n+1}]}\max_{i=1,\ldots,s}\|R_{ni}(t)\|=0,\ \tau_{\max}\to 0,$$

and

$$z'(t) = G'(y(t))y'(t) = G'(y_n)f(y_n, G(y_n)) + S_n(t),$$

where

$$\max_{n=0,1,\dots,N-1} \max_{t\in[t_n,t_{n+1}]} \|S_n(t)\| = 0,\ \tau_{\max} \to 0.$$

Since (31) and (32) hold, the thesis follows by Theorem 1. □

3. Numerical tests

Theorems 2, 3 and 4 say what is the convergence order of the numerical event location for semi-implicit methods, Rosenbrock methods and implicit RK methods, respectively. Some numerical tests on simple DAEs are now presented with the aim to experimentally confirm these convergence orders.

3.1. Numerical methods

The tests involve the following methods.
As semi-implicit methods, we consider:

- the explicit two-stage improved (or modified) Euler method of order $p = 2$ equipped by the linear continuous extension

  $$b_1(\theta) = \frac{1}{2}\theta,\ b_2(\theta) = \frac{1}{2}\theta$$

  of order $q = 1$;
- the explicit four-stage classical RK method of order $p = 4$ equipped by the continuous extension

  $$b_1(\theta) = -\frac{1}{2}\theta^2 + \frac{2}{3}\theta,\ b_2(\theta) = b_3(\theta) = \frac{1}{3}\theta,\ b_4(\theta) = \frac{1}{2}\theta^2 - \frac{1}{3}\theta$$

  of order $q = 2$ or the continuous extension

  $$b_1(\theta) = \frac{2}{3}\theta^3 - \frac{3}{2}\theta^2 + \theta,\ b_2(\theta) = b_3(\theta) = -\frac{2}{3}\theta^3 + \theta^2,\ b_4(\theta) = \frac{2}{3}\theta^3 - \frac{1}{2}\theta^2$$

  of order $q = 3$ (see [2]).

By Theorem 2, the numerical event location has convergence order $\min\{p, q+1\}$.

As a Rosenbrock method, we use:

- the two-stage method given by

  $$b_1 = 0,\ b_2 = 1,\ a_{21} = \frac{1}{12},\ \gamma_{11} = \frac{1}{4},\ \gamma_{21} = \frac{1}{12},\ \gamma_{22} = \frac{1}{3}$$

  of differential algebraic order $p = 2$ and satisfying $R(\infty) = 0$, equipped by the linear continuous extension

  $$b_1(\theta) = 0,\ b_2(\theta) = \theta$$

  of differential algebraic order $q = 1$.

By Theorem 3, the numerical event location has convergence order $\min\{p, q+1\}$.

Finally, as implicit RK methods, we see:

- the stiffly accurate two-stage Lobatto III C method of differential algebraic order $p = 2$.
- the stiffly accurate three-stage Radau II A method of differential algebraic order $p = 5$.

By Theorem 4, the numerical event location has convergence order p.

We present three test, where three different simple DAEs are considered. These DAEs are all integrated with constant stepsize

$$\tau_k = r_0 q^{-k} \tag{34}$$

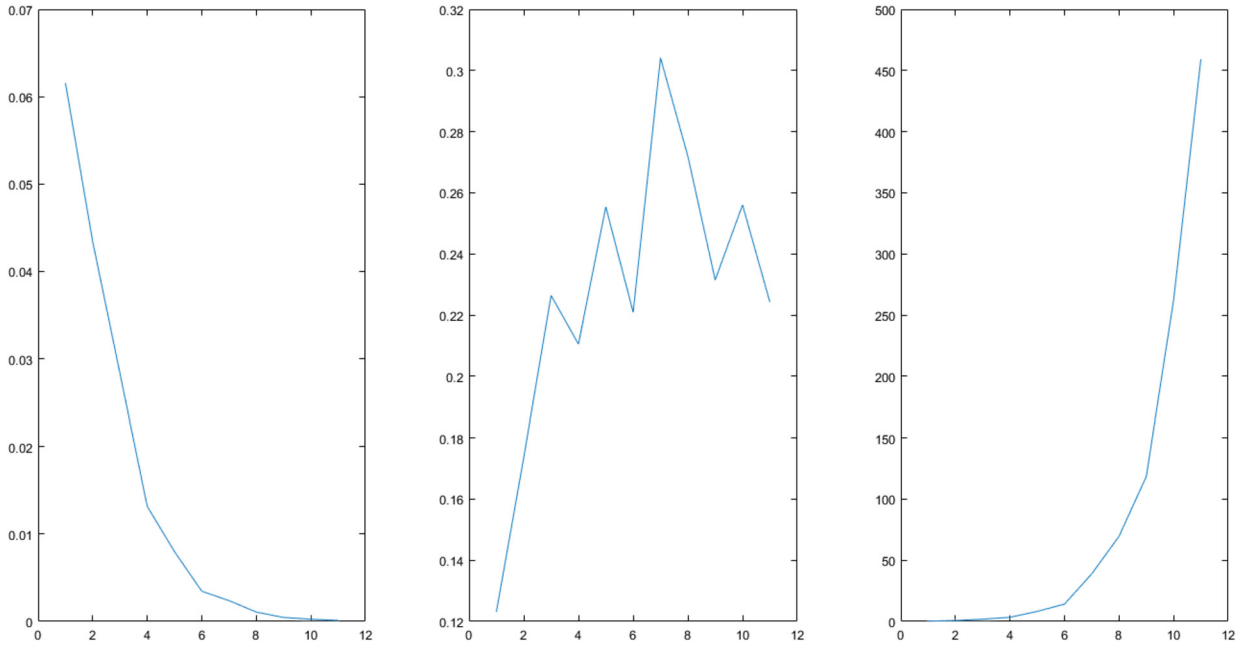

Fig. 1. Improved Euler method.

where $r_0 > 0$ and $q > 1$ are fixed and the integer k varies from 0 to $k_{\max}$. We check the convergence order r by looking to the ratios

$$\frac{M_k}{\tau_k^{r-1}}, \quad \frac{M_k}{\tau_k^{r}} \quad \text{and} \quad \frac{M_k}{\tau_k^{r+1}}, \tag{35}$$

where

$$M_k = \max\left\{|t^*_{\tau_k} - t^*|, \|y^*_{\tau_k} - y^*\|_2, \|z^*_{\tau_k} - z^*\|_2\right\},$$

for the stepsizes (34) with $k = 0, \ldots, k_{\max}$. If the convergence order of the numerical event location is (exactly) r, then, as τ_k goes to zero, $\frac{M_k}{\tau_k^{r-1}}$ is expected to go to zero, $\frac{M_k}{\tau_k^{r+1}}$ is expected to go to infinite, whereas $\frac{M_k}{\tau_k^{r}}$ is expected to stay away from zero and infinite.

3.2. First test

We consider the scalar DAE

$$\begin{cases} y'(t) = z(t), \ t \geq 1, \\ y(t)^2 - z(t)^2 - 1 = 0, \ t \geq 1, \\ (y(1), z(1)) = (\cosh 1, \sinh 1), \end{cases} \tag{36}$$

whose solution is

$$(y(t), z(t)) = (\cosh t, \sinh t), \ t \geq 1.$$

The event equation is

$$h(y, z) = 2yz - 100,$$

with event time $t^* = \frac{1}{2}\operatorname{arcsinh}(100) = 2.6492$ and event point $(\cosh t^*, \sinh t^*)$.

For this DAE, we check the convergence order of the numerical event location for the improved Euler method (as semi-implicit method), the Rosenbrock method and the Lobatto III C method (as implicit RK method). For all three methods the convergence order is $r = 2$.

In Figs. 1, 2 and 3, we see, from the left to the right, the ratios (35) for the stepsizes

$$\tau_k = 0.5 \cdot 2^{-k}, \ k = 0, \ldots, k_{\max} = 10.$$

The abscissas in these figures are the values of k. The order $r = 2$ is confirmed.

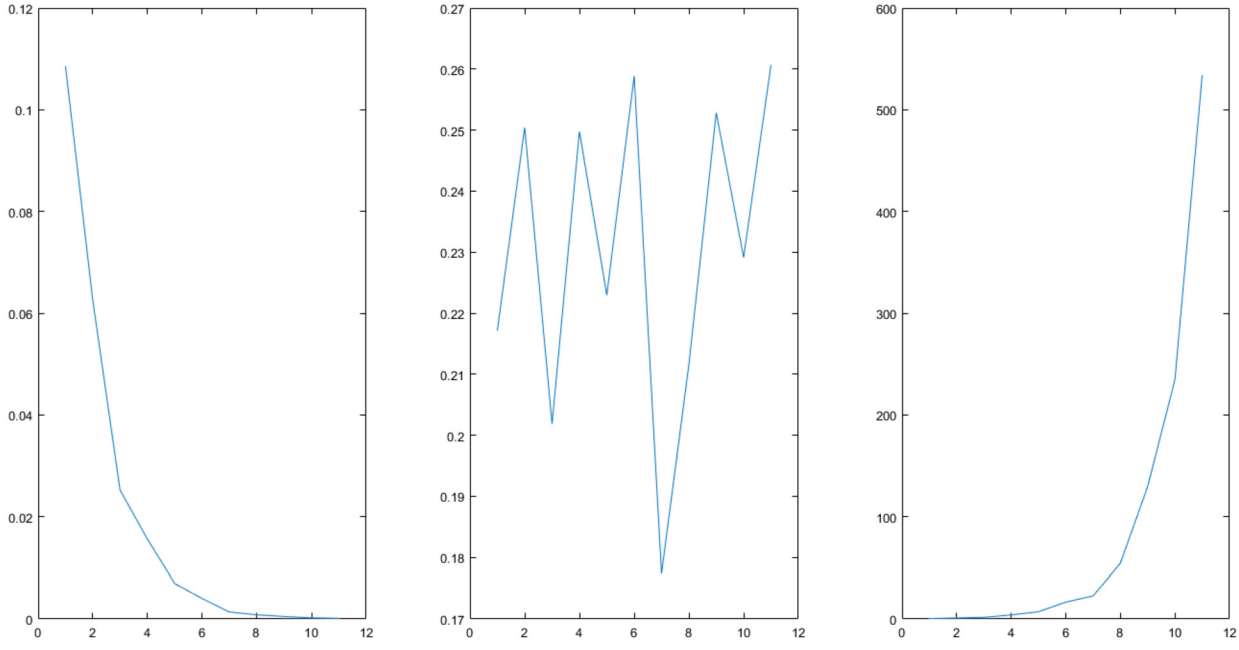

Fig. 2. Rosenbrock method.

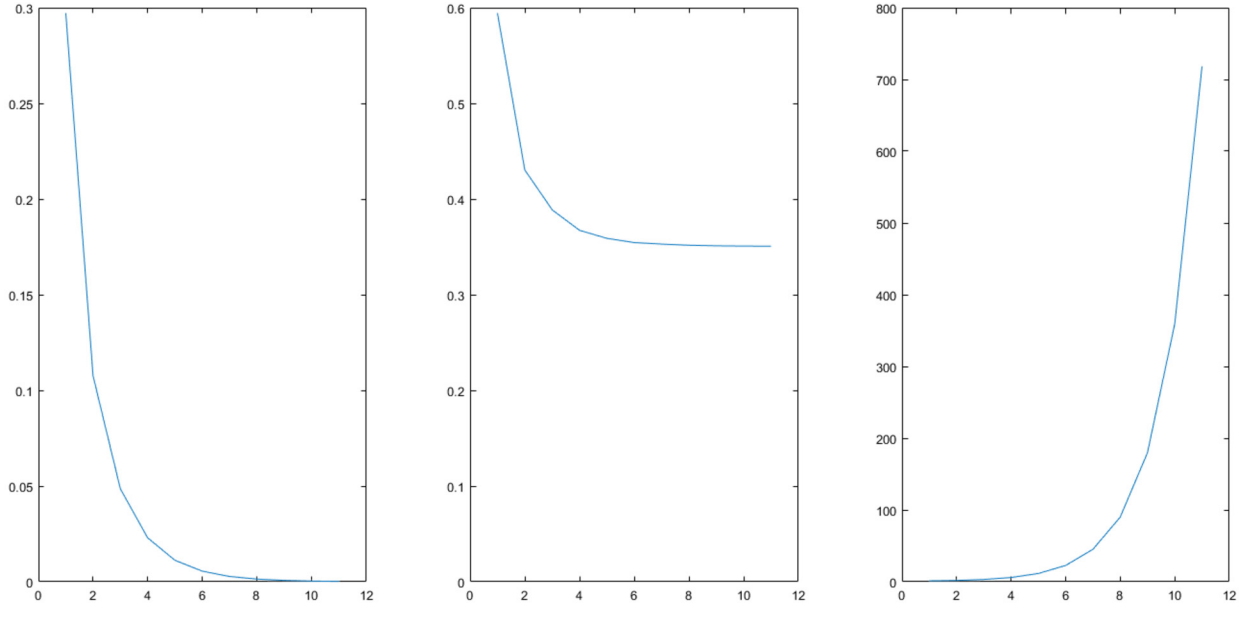

Fig. 3. Two-stage Lobatto III C method.

Just for completing the information of the previous figures, we give in Table 1 the errors M_k for $k = 0$ and $k_{\max} = 10$.

3.3. Second test

We consider the linear DAE

$$\begin{cases} y'(t) = Ay(t) + Bz(t), \ t \geq 0 \\ Cy(t) + Dz(t) = 0, \ t \geq 0, \\ (y(0), z(0)) = (y_0, z_0), \end{cases} \tag{37}$$

Table 1
Improved Euler method (top), Rosenbrock method (middle) and two-stage Lobatto III C method (bottom).

k	τ_k	M_k
0	0.5	3.08e-02
$k_{\max} = 10$	4.88e-04	5.35e-08

k	τ_k	M_k
0	0.5	5.43e-02
$k_{\max} = 10$	4.88e-04	6.22e-08

k	τ_k	M_k
0	0.5	1.49e-01
$k_{\max} = 10$	4.88e-04	8.36e-08

where $A, B, C, D \in \mathbb{R}^{n\times n}$, with D invertible, and (y_0, z_0) are consistent initial values. The solution is

$$z(t) = -D^{-1}Cy(t)$$

and

$$y(t) = \mathrm{e}^{t(A-BD^{-1}C)}y_0.$$

In particular, we consider $n = 10$, the matrices

$$A = D = \begin{bmatrix} -2 & 1 & & & & \\ 1 & \ddots & \ddots & & & \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & 1 \\ & & & & 1 & -2 \end{bmatrix}, \; B = -C = \begin{bmatrix} -1 & & & & & \\ 1 & \ddots & & & & \\ & \ddots & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 1 & -1 \end{bmatrix},$$

and the consistent initial values

$$y_0 = (1, 0, \ldots, 0), z_0 = -D^{-1}Cy_0.$$

The event equation is

$$h(y(t), z(t)) = a^T y(t) + b^T z(t) + c = 0,$$

where

$$a = \frac{1}{5}(1, \ldots, 10) \in \mathbb{R}^n, \; b = -\frac{1}{2}a$$

and the number c is chosen in order to have an event time at $t^* = \frac{\sqrt{2}}{2}$. Then, the event point is

$$y^* = \mathrm{e}^{t^*(A-BD^{-1}C)}y_0, \; z^* = -D^{-1}Cy^*.$$

First, as semi-implicit method, we check the convergence order of the numerical event location for the classical RK method with both continuous extensions. The order of convergence is $r = 3$ for the second order continuous extension and $r = 4$ for the third order continuous extension.

In Figs. 4 and 5, we see, from the left to the right, the ratios (35) for the stepsizes

$$\tau_k = 0.25 \cdot 1.1^{-k}, \; k = 0, \ldots, k_{\max}, \tag{38}$$

for the second order continuous extension ($k_{\max} = 70$) and third order continuous extension ($k_{\max} = 60$), respectively. The convergence orders $r = 3$ and $r = 4$ of the numerical event location are confirmed.

Now, as implicit method, we check the convergence order of the numerical event location for the three-stage Radau II A method. The order of convergence is $r = 5$. The ratios (35) for the stepsizes (38) with $k_{\max} = 40$ are given in Fig. 6. The order of convergence $r = 5$ is confirmed.

Table 2 gives the errors M_k for $k = 0$ and $k = k_{\max}$.

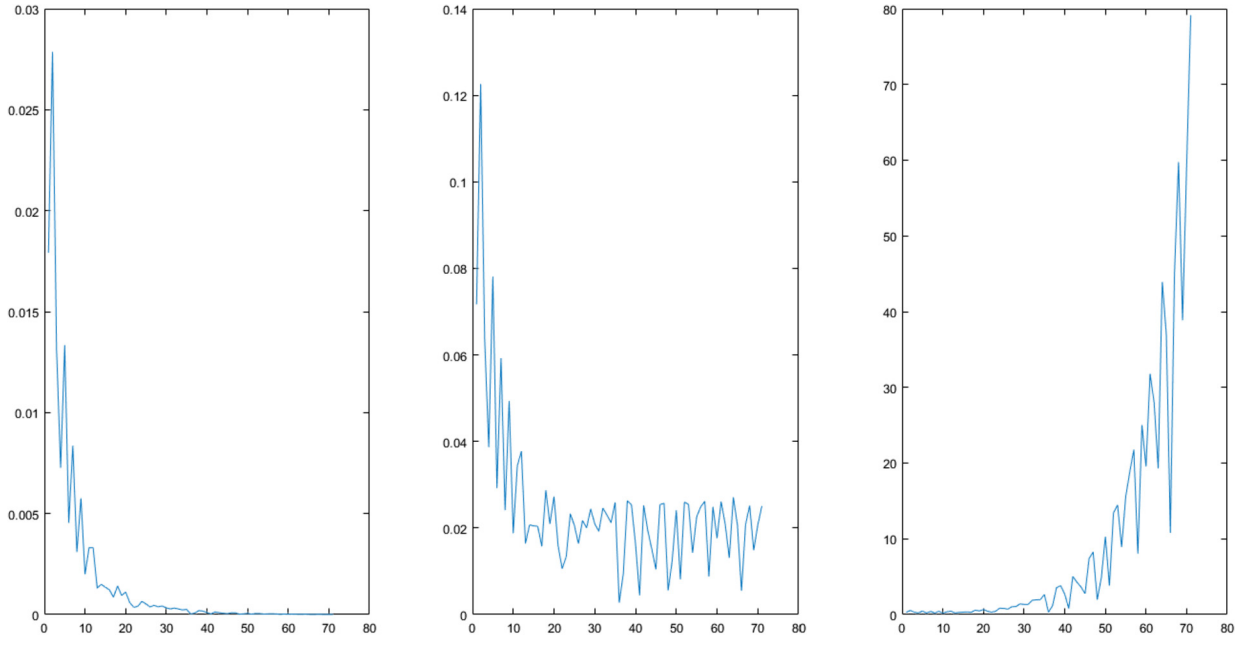

Fig. 4. Classical RK method with second order continuous extension ($r = 3$).

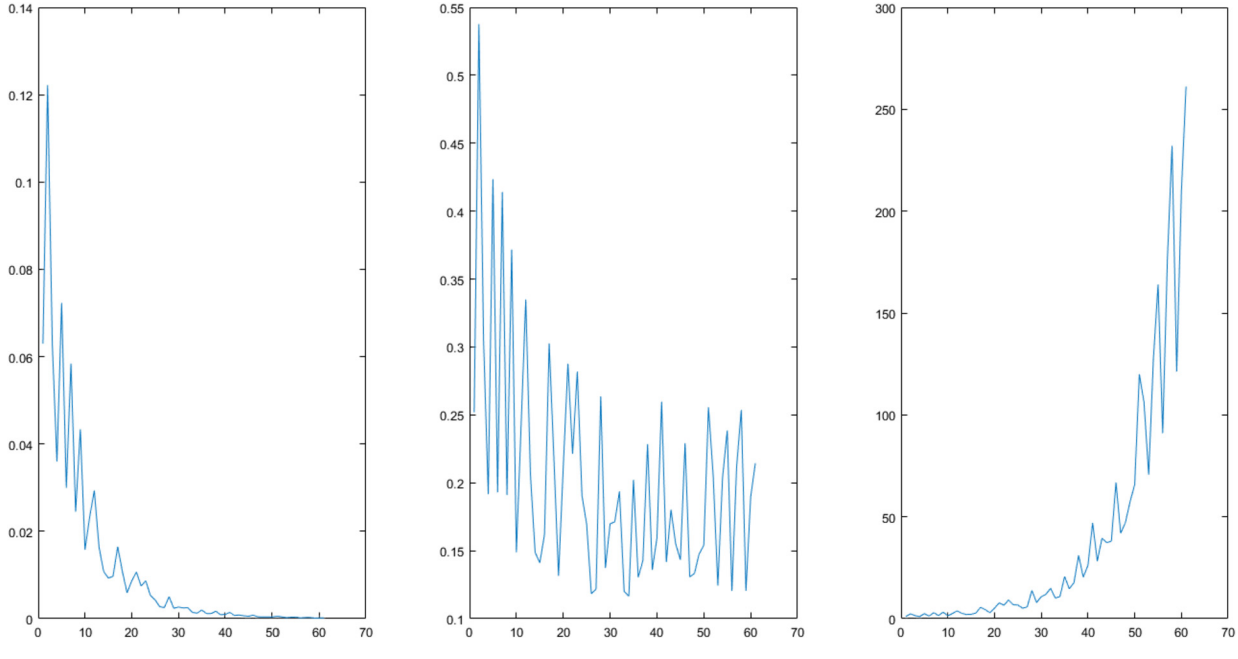

Fig. 5. Classical RK method with third order continuous extension ($r = 4$).

3.4. Third test

The following DAE describes the gas-phase in a model of a soft-drink production (see [8]) and it reads

$$\begin{cases} y_1'(t) = F_1 - z(t) - k_c \frac{y_1(t) y_2(t)}{V} \\ y_2'(t) = F_2 - k_c \frac{y_1(t) y_2(t)}{V} \\ y_3'(t) = k_c \frac{y_1(t) y_2(t)}{V} \\ z(t) = k_g X \left(P(y(t)) - P_{\text{out}} \right), \end{cases} \tag{39}$$

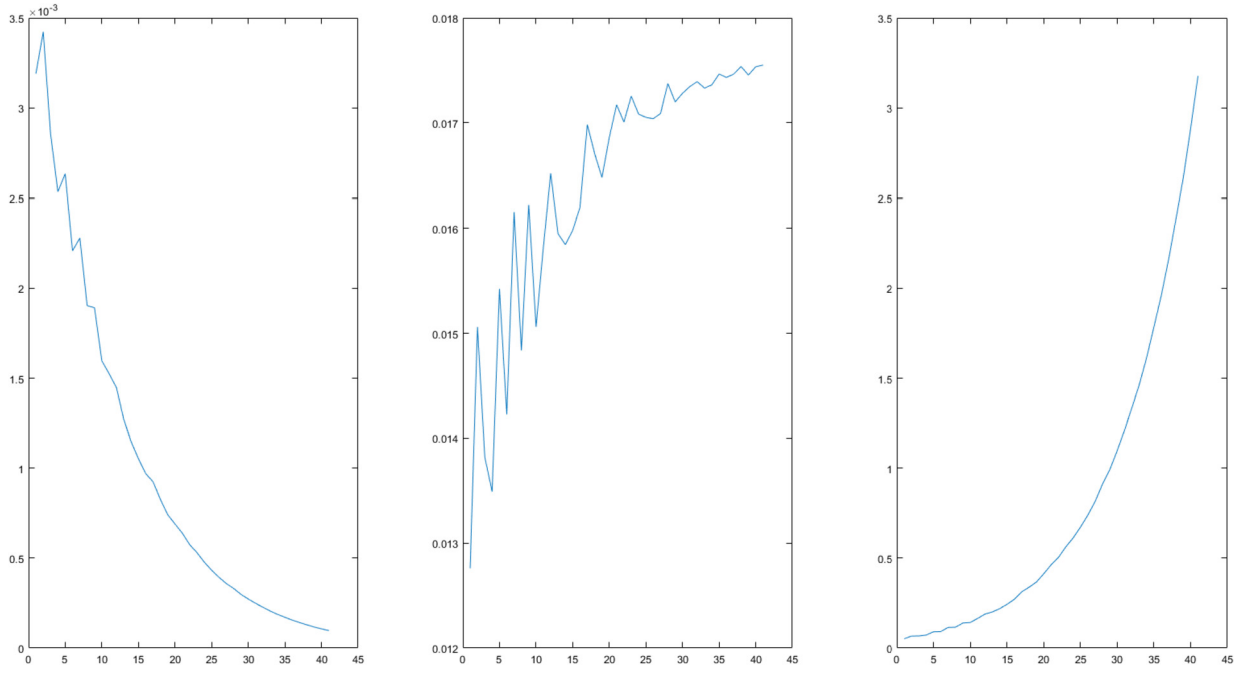

Fig. 6. Three-stage Radua II A method ($r = 5$).

Table 2
Classical RK method with second order continuous extension (top), classical RK method with third order continuous extension (middle) and three-stage Radau II A method (bottom).

k	τ_k	M_k
0	2.50e-01	1.12e-03
$k_{\max} = 70$	3.17e-04	7.95e-13

k	τ_k	M_k
0	2.50e-01	9.84e-04
$k_{\max} = 60$	8.21e-04	9.75e-14

k	τ_k	M_k
0	2.50e-01	1.25e-05
$k_{\max} = 40$	5.52e-03	9.03e-14

where

$$P(y(t)) = \frac{y_1(t)RT}{V - \frac{y_2(t)}{\rho_l} - \frac{y_3(t)}{\rho_a}}.$$

Here y_1, y_2 and y_3 are molar concentrations of CO_2, H_2O and H_2CO_3, respectively, in a tank and z is the flow rate of CO_2 leaving the tank by a valve. The values of constants and parameters in suitable units are

$$F_1 = 0.5,\ F_2 = 7.5,\ k_c = \frac{0.433}{4000},\ V = 10,$$
$$k_g = 3,\ X = 1,\ P_{\text{out}} = 1$$
$$R = 0.0820574587,\ T = 293,\ \rho_a = 16,\ \rho_l = 50.$$

The consistent initial values are given by

$$(y_1(0), y_2(0), y_3(0)) = (0.72, 95, 0).$$

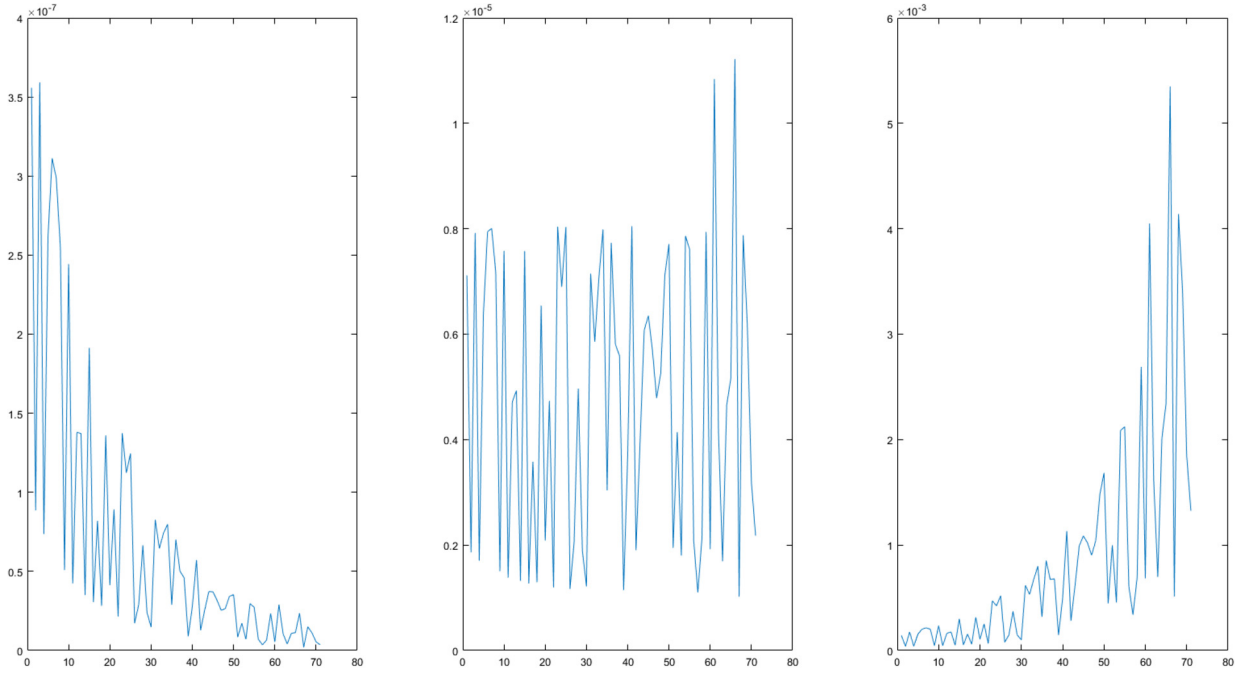

Fig. 7. Improved Euler method.

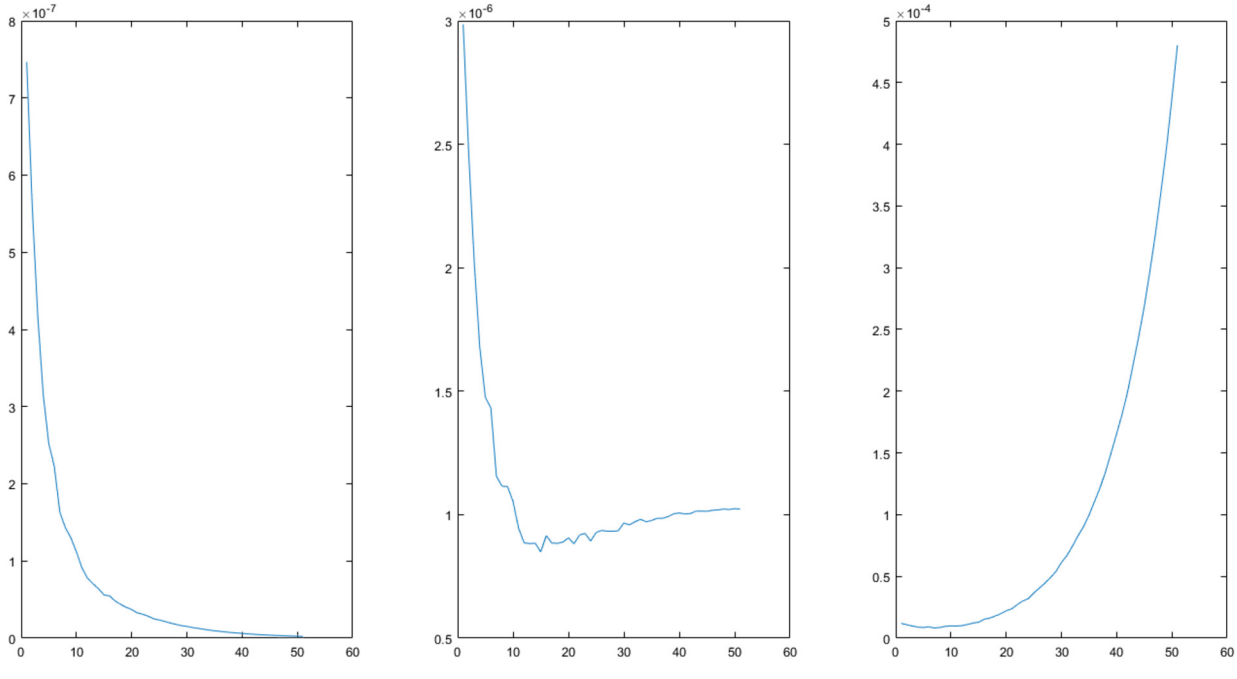

Fig. 8. Lobatto III C method.

The event equation involves only $y(t)$ and it is given by

$$h(y(t), z(t)) = \frac{y_2(t)}{\rho_l} + \frac{y_3(t)}{\rho_a} - V_d = 0$$

with $V_d = 2.25$. After the event, we have a transition to the liquid-phase, where another DAE holds.

We check the convergence order of numerical event location for the improved Euler method, as semi-implicit method, and the two-stage Lobatto III C method, as implicit method. The order of convergence is $r = 2$ for both methods. In Figs. 7 and 8, we see the ratios (35) for the stepsizes

Table 3
Improved Euler method (top) and two-stage Lobatto III C method (bottom).

k	τ_k	M_k
0	5.00e-02	1.78e-08
$k_{\max} = 70$	1.64e-03	5.88e-12

k	τ_k	M_k
0	5.00e-02	1.87e-07
$k_{\max} = 50$	4.36e-03	4.64e-12

$$\tau_k = 0.05 \cdot 1.05^{-k},\ k = 0, 1, 2, \ldots, k_{\max},$$

for the improved Euler method ($k_{\max} = 70$) and the Lobatto III C method ($k_{\max} = 50$), respectively. We take the numerical event time and event point corresponding to $k = k_{\max} + 30$ as the exact event time and event point.

The convergence order $r = 2$ for the numerical event location is confirmed.

In Table 3 we see the errors M_k for $k = 0$ and $k = k_{\max}$. Although the convergence order is only 2, we have very small errors M_k, $k = 0, \ldots, k_{\max}$. This is explained by a very small constant C in the error estimate $C\tau_k^2$, as we can see in the middle parts of Figs. 7 and 8: the order of magnitude of the constant C is 10^{-6}.

4. Conclusion

In this paper, we have studied numerical methods for the event location of Discontinuous DAEs (DDAEs), namely DAEs for which the trajectory meets a discontinuity surface Σ. We have considered some standard methods for the numerical solution of DAEs and convergence theorems of the numerical event time and event point to the true event time and event point are given. Of course the event location is a step of a more sophisticated procedure: that is, once the numerical trajectory meets the discontinuity surface Σ one should ask if the trajectory has to cross or slide on Σ: if the trajectory crosses Σ, then we need to change the vector field; while if the discontinuity surface has to slide on Σ, then a unique vector field has to be chosen and the numerical solution has to be computed until eventually it leaves Σ. The theory of DDAEs, concerning for instance the existence of solutions, the conditions of crossing/sliding for the trajectories and the definition of a unique sliding vector, is enough different with respect to the case of discontinuous ODEs, where Filippov theory helps. In the paper [8], there is a deep study of these aspects which should be considered in order to develop effective algorithms for solving such a type of problems.

In future the authors wish to study transformation techniques, which consists in time re-parametrizations that seem to be particularly effective for problems of this type (see [10,17]) and other problems (see [5,6]).

Acknowledgement

The authors acknowledge that this research was supported by funds from the Italian MUR (Ministero dell'Università e della Ricerca) within the PRIN 2017 Project "Discontinuous dynamical systems: theory, numerics and applications" and by the INdAM Research group GNCS (Gruppo Nazionale di Calcolo Scientifico).

Appendix A. Background on numerical methods for DAEs

In this appendix, we recall some classical numerical methods for solving DAEs, which will be used to derive numerical event location techniques for DDAEs. The results in this section may be found in standard books on DAEs, see for example [13,14], and they are given here for better understanding the numerical event location techniques.

A.1. Semi-implicit methods

Starting from an s-stage explicit RK method (A, b, c), we can construct the following method for the DAE (1).

Given the approximation y_n of $y(t_n)$, $n = 0, 1, \ldots, N-1$, we obtain the approximation (y_{n+1}, z_{n+1}) of $(y(t_{n+1}), z(t_{n+1}))$ by

$$y_{n+1} = y_n + \tau_{n+1} \sum_{i=1}^{s} b_i f(y_{ni}, z_{ni}) \tag{40}$$

$$g(y_{n+1}, z_{n+1}) = 0, \tag{41}$$

where the stage values (y_{ni}, z_{ni}), $i = 1, \ldots, s$, are recursively obtained by

$$y_{ni} = y_n + \tau_{n+1} \sum_{j=1}^{i-1} a_{ij} f(y_{nj}, z_{nj}) \tag{42}$$

$$g(y_{ni}, z_{ni}) = 0. \tag{43}$$

The equations (40)-(41)-(42)-(43) define a *semi-implicit method* for the DAE (1). Observe that:

- for any stage index $i = 1, \ldots, s$, given the already computed stage values (y_{nj}, z_{nj}), $j = 1, \ldots, i-1$, first we compute explicitly y_{ni} by the formula (42) and then we compute implicitly z_{ni} by solving the equation (43);
- finally, given all the stage values (y_{ni}, z_{ni}), $i = 1, \ldots, s$, first we compute explicitly y_{n+1} by the formula (40) and then we compute implicitly z_{n+1} by solving the equation (41).

Thus, at any step, the method requires to solve $s+1$ non-linear systems of dimension d_2.

Observe that all the stage values (y_{ni}, z_{ni}), $i = 1, \ldots, s$, as well as the approximation (y_{n+1}, z_{n+1}) are consistent, i.e. $g(y_{ni}, z_{ni}) = 0$, $i = 1, \ldots, s$, and $g(y_{n+1}, z_{n+1}) = 0$, by construction. So, the approximations (y_n, z_n), $n = 0, 1, \ldots, N$, obtained by a semi-implicit method are all consistent.

The order of convergence of this semi-implicit method is the order p of the explicit RK method: we have

$$\max_{n=0,1,\ldots,N} \|y_n - y(t_n)\| = O(\tau_{\max}^p)$$

and

$$\max_{n=0,1,\ldots,N} \|z_n - z(t_n)\| = O(\tau_{\max}^p)$$

as $\tau_{\max} \to 0$. The proof of this fact is a trivial consequence of (2).

A.2. Singularly perturbed problems

We consider the solution of the DAE (1) as the limit, as $\varepsilon \to 0^+$, of the solution (y, z) of the Singularly Perturbed Problem (SPP)

$$\begin{cases} y'(t) = f(y(t), z(t)), \ t \in [0, T], \\ \varepsilon z'(t) = g(y(t), z(t)), \ t \in [0, T], \\ (y(0), z(0)) = (y_0, z_0). \end{cases} \tag{44}$$

We assume that the SPP (44) has a unique solution on $[0, T]$ for any sufficiently small $\varepsilon > 0$.

Numerical methods for the DAE (1) can be derived by applying to the SPP (44) a numerical method for Ordinary Differential Equations (ODEs) and then by setting ε to zero. As numerical method for ODEs, we consider implicit RK methods and Rosenbrock methods.

A.2.1. Implicit RK methods

Consider an implicit RK method (A, b, c) with A non-singular as applied to the SPP (44). We obtain, for $n = 0, 1, \ldots, N-1$,

$$\begin{aligned} y_{n+1} &= y_n + \tau_{n+1} \sum_{i=1}^{s} b_i f(y_{ni}, z_{ni}) \\ \varepsilon z_{n+1} &= \varepsilon z_n + \tau_{n+1} \sum_{i=1}^{s} b_i g(y_{ni}, z_{ni}), \end{aligned} \tag{45}$$

where the stage values (y_{ni}, z_{ni}), $i = 1, \ldots, s$, are obtained by solving the system of equations

$$\begin{aligned} y_{ni} &= y_n + \tau_{n+1} \sum_{j=1}^{s} a_{ij} f(y_{nj}, z_{nj}), \ i = 1, \ldots, s, \\ \varepsilon z_{ni} &= \varepsilon z_n + \tau_{n+1} \sum_{j=1}^{s} a_{ij} g(y_{nj}, z_{nj}), \ i = 1, \ldots, s. \end{aligned} \tag{46}$$

Since A is invertible, we can rewrite (46) as

$$\tau_{n+1} g(y_{ni}, z_{ni}) = \varepsilon \sum_{j=1}^{s} \omega_{ij}(z_{nj} - z_n),\ i = 1, \ldots, s, \tag{47}$$

where ω_{ij}, $i, j = 1, \ldots, s$, are the elements of A^{-1}, and then (45) as

$$z_{n+1} = \left(1 - \sum_{i,j=1}^{s} b_i \omega_{ij}\right) z_n + \sum_{i,j=1}^{s} b_i \omega_{ij} z_{nj}.$$

By setting $\varepsilon = 0$, the equation (47) becomes

$$g(y_{ni}, z_{ni}) = 0,\ i = 1, \ldots, s.$$

Thus, for the DAE (1), we have the *implicit RK method*

$$\begin{aligned} y_{n+1} &= y_n + \tau_{n+1} \sum_{i=1}^{s} b_i f(y_{ni}, z_{ni}) \\ z_{n+1} &= \left(1 - \sum_{i,j=1}^{s} b_i \omega_{ij}\right) z_n + \sum_{i,j=1}^{s} b_i \omega_{ij} z_{nj}, \end{aligned} \tag{48}$$

where the stage values (y_{ni}, z_{ni}), $i = 1, \ldots, s$, are obtained by solving the non-linear system of equations

$$\begin{aligned} &y_{ni} = y_n + \tau_{n+1} \sum_{j=1}^{s} a_{ij} f(y_{nj}, z_{nj}),\ i = 1, \ldots, s, \\ &g(y_{ni}, z_{ni}) = 0,\ i = 1, \ldots, s. \end{aligned} \tag{49}$$

At each step, the method needs to solve a non-linear system of dimension $s(d_1 + d_2)$. When compared to semi-implicit methods for DAEs, implicit RK methods for DAEs require more computational effort.

Observe that the stage values (y_{ni}, z_{ni}), $i = 1, \ldots, s$, are consistent by construction but, in general, the approximation (y_{n+1}, z_{n+1}) given in (48) could be non-consistent.

The consistency of (y_{n+1}, z_{n+1}) is automatically obtained in case of a *stiffly accurate* implicit RK method (for example, a Radau IIA method or a Lobatto IIIC method): for such a method, we have

$$a_{si} = b_i,\ i = 1, \ldots, s,$$

and then $(y_{n+1}, z_{n+1}) = (y_{ns}, z_{ns})$.

Regarding the order of convergence of implicit RK methods for DAEs, we introduce the following notion of differential algebraic order.

Definition 5. Let (A, b, c) be a RK method with A non-singular. We say that the RK method has *differential algebraic order* p if

$$\max_{t \in [0,T]} \|y_{n+1} - y(t + \tau)\| = O(\tau^{p+1})$$

and

$$\max_{t \in [0,T]} \|z_{n+1} - z(t + \tau)\| = O(\tau^{p})$$

as $\tau \to 0$, where y_{n+1} and z_{n+1} are obtained by (48)-(49) with $(t_n, y_n, z_n) = (t, y(t), z(t))$.

The method (48) is convergent of order p if the implicit RK method has differential algebraic order p and satisfies an additional condition.

Theorem 6. *Let (A, b, c) be a RK method with A non-singular. If the RK method has differential algebraic order p and $|R(\infty)| < 1$, where R is the stability function of the RK method, then the method (48)-(49) has convergence order p, i.e. we have*

$$\max_{n=1,\ldots,N} \|y_n - y(t_n)\| = O(\tau_{\max}^{p})$$

and

$$\max_{n=1,\dots,N} \|z_n - z(t_n)\| = O(\tau_{\max}^p)$$

as $\tau_{\max} \to 0$.

A stiffly accurate implicit RK method of order p has differential algebraic order p and its stability function R satisfies $R(\infty) = 0$. So, the method has convergence order p when it is applied to DAEs.

A.2.2. Rosenbrock methods

Stiffly accurate RK methods for DAEs work fine, but they are implicit methods: at any step the fully nonlinear system (49) needs to be solved. To reduce the computational effort, we can use *Rosenbrock methods* instead of implicit RK methods for solving the SPP (44), since the Rosenbrock methods constitute a good compromise between the cost and the stability requirement. The application of such methods to SPPs has been extensively studied, for instance, in [14,22,24].

Unlike RK methods, Rosenbrock methods make use of the jacobian matrix of the right-hand side of an ODE. A Rosenbrock method as applied to the SSP (44) reads, for $n = 0, 1, \dots, N-1$,

$$\begin{aligned} y_{n+1} &= y_n + \sum_{i=1}^{s} b_i l_{ni} \\ z_{n+1} &= z_n + \sum_{i=1}^{s} b_i k_{ni}, \end{aligned} \tag{50}$$

where (l_{ni}, k_{ni}), $i = 1, \dots, s$, are recursively obtained by

$$\begin{aligned} l_{ni} &= \tau_{n+1} f(y_{ni}, z_{ni}) + \tau_{n+1} \sum_{j=1}^{i} \gamma_{ij} \left(f_y^n l_{nj} + f_z^n k_{nj} \right), \ i = 1, \dots, s, \\ \varepsilon k_{ni} &= \tau_{n+1} g(y_{ni}, z_{ni}) + \tau_{n+1} \sum_{j=1}^{i} \gamma_{ij} \left(g_y^n l_{nj} + g_z^n k_{nj} \right), \ i = 1, \dots, s, \end{aligned} \tag{51}$$

with (y_{ni}, z_{ni}), $i = 1, \dots, s$, recursively and explicitly given by

$$\begin{aligned} y_{ni} &= y_n + \sum_{j=1}^{i-1} a_{ij} l_{nj}, \ i = 1, \dots, s, \\ z_{ni} &= z_n + \sum_{j=1}^{i-1} a_{ij} k_{nj}, \ i = 1, \dots, s. \end{aligned} \tag{52}$$

Thus, the Rosenbrock method is defined by an explicit RK method (A, b, c) and additional parameters γ_{ij}, $i = 1, \dots, s$ and $j = 1, \dots, i$. The method uses the jacobian matrices f_y^n, f_z^n, g_y^n and g_z^n of f and g evaluated at (y_n, z_n).

By setting $\varepsilon = 0$, the second equation in (51) becomes

$$0 = \tau_{n+1} g(y_{ni}, z_{ni}) + \tau_{n+1} \sum_{j=1}^{i} \gamma_{ij} \left(g_y^n l_{nj} + g_z^n k_{nj} \right), \ i = 1, \dots, s,$$

and the first and second equations of (51) read

$$\begin{aligned} &\begin{bmatrix} I - \tau_{n+1}\gamma_{ii} f_y^n & -\tau_{n+1}\gamma_{ii} f_z^n \\ -\tau_{n+1}\gamma_{ii} g_y^n & -\tau_{n+1}\gamma_{ii} g_z^n \end{bmatrix} \begin{bmatrix} l_{ni} \\ k_{ni} \end{bmatrix} \\ &= \begin{bmatrix} \tau_{n+1} f(y_{ni}, z_{ni}) + \tau_{n+1} \sum_{j=1}^{i-1} \gamma_{ij} \left(f_y^n l_{nj} + f_z^n k_{nj} \right) \\ \tau_{n+1} g(y_{ni}, z_{ni}) + \tau_{n+1} \sum_{j=1}^{i-1} \gamma_{ij} \left(g_y^n l_{nj} + g_z^n k_{nj} \right) \end{bmatrix} \\ &i = 1, \dots, s. \end{aligned} \tag{53}$$

For the DAE (1), the *Rosenbrock method* is given by the equations (50)-(52)-(53):

- for each stage index $i = 1, \ldots, s$, given the previously computed (l_{nj}, k_{nj}), $j = 1, \ldots, i-1$, first we compute explicitly (y_{ni}, z_{ni}) by using the formula (52) and then (l_{ni}, k_{ni}) is obtained by solving the 2×2 block linear system (53);
- finally, given all values (l_{ni}, k_{ni}), $i = 1, \ldots, s$, we compute (y_{n+1}, z_{n+1}) by the formula (50).

At each step of this method, s linear systems of dimension $d_1 + d_2$ need to be solved. This is a big computational saving with respect to implicit RK methods, where a fully non-linear system of dimension $s(d_1 + d_2)$ has to be solved.

Observe that, in general, the stage values (y_{ni}, z_{ni}), $i = 1, \ldots, s$, and the approximation (y_{n+1}, z_{n+1}) are not consistent.

As in case of implicit RK method for DAEs, we can introduce for Rosenbrock methods for DAEs the notion of differential algebraic order and prove that if a Rosenbrock method has differential algebraic order p and its stability function R satisfies $|R(\infty)| < 1$, then it has convergence order p (see [13]).

Appendix B. Continuous extensions

This other appendix contains a brief review of continuous extensions for the methods introduced in the previous appendix.

Often, it is important to have an approximated solution of the DAE (1) defined not only at mesh times t_n, $n = 0, 1, \ldots, N$, but at all times $t \in [0, T]$. Approximations of this type are given by *continuous extensions* of numerical solutions defined at mesh times.

We consider continuous extensions of numerical solutions given by semi-implicit methods or Rosenbrock methods, since such numerical solutions require continuous extensions for the numerical event location. On the other hand, numerical solutions given by implicit RK methods do not require continuous extensions for this task.

When the DAE (1) is integrated by a semi-implicit method (as described in Subsection A.1), the numerical event location requires a continuous extension of the numerical solution defined in (40), as we will see in Subsection 2.2. This continuous extension is a function η defined at all times $t \in [0, T]$ and given by

$$\begin{aligned} &\eta(t_n + \theta\tau_{n+1}) = y_n + \tau_{n+1} \sum_{i=1}^{s} b_i(\theta) f(y_{ni}, z_{ni}), \\ &n = 0, 1, \ldots, N-1 \text{ and } \theta \in [0, 1] \end{aligned} \tag{54}$$

where $b_i(\cdot)$, $i = 1, \ldots, s$, are polynomial functions such that:

$$b_i(0) = 0 \text{ and } b_i(1) = b_i. \tag{55}$$

The continuous extension is said to have order q if

$$\max_{\substack{t \in [0,T] \\ \theta \in [0,1]}} \|\eta(t + \theta\tau) - y(t + \theta\tau)\| = O(\tau^{q+1})$$

as $\tau \to 0$, where η is given by (54) with $(t_n, y_n) = (t, y(t))$.

One can prove (see [2]) that if the continuous extension has order q and the semi-implicit method has order p, then the continuous extension has convergence order $\min\{p, q+1\}$, i.e. we have

$$\max_{t \in [0,T]} \|\eta(t) - y(t)\| = O(\tau_{\max}^{\min\{p, q+1\}})$$

as $\tau_{\max} \to 0$.

It is possible to prove that there exists a continuous extension of order $[(p+1)/2]$ (see [2]).

When the DAE (1) is integrated by a Rosenbrock method (as described in Subsection A.2.2), the numerical event location requires a continuous extension (η, μ) of the numerical solution in (50), as we will see in Subsection 2.3. It is defined as follows:

$$\begin{aligned} &\eta(t_n + \theta\tau) = y_n + \sum_{i=1}^{s} b_i(\theta) l_{ni} \\ &\mu(t_n + \theta\tau) = z_n + \sum_{i=1}^{s} b_i(\theta) k_{ni} \\ &n = 0, 1, 2, \ldots \text{ and } \theta \in [0, 1], \end{aligned} \tag{56}$$

where $b_i(\cdot)$, $i = 1, \ldots, s$, are polynomial functions satisfying (55).

The continuous extension is said to have differential algebraic order q if

$$\max_{\substack{t\in[0,T]\\ \theta\in[0,1]}} \|\eta(t+\theta\tau) - y(t+\theta\tau)\| = O(\tau^{q+1})$$

and

$$\max_{\substack{t\in[0,T]\\ \theta\in[0,1]}} \|\mu(t+\theta\tau) - z(t+\theta\tau)\| = O(\tau^{q}),$$

as $\tau \to 0$, where (η, μ) is given by (56) with $(t_n, y_n, z_n) = (t, y(t), z(t))$.

Similarly to continuous extensions of semi-implicit methods, one can prove (see [22]) that if the continuous extension has differential algebraic order q, the Rosenbrock method has differential algebraic order p and $|R(\infty)| < 1$, where R is the stability function of the Rosenbrock method, then the continuous extension has convergence order $\min\{p, q+1\}$, i.e. we have

$$\max_{t\in[0,T]} \|\eta(t) - y(t)\| = O(\tau_{\max}^{\min\{p,q+1\}})$$

and

$$\max_{t\in[0,T]} \|\mu(t) - z(t)\| = O(\tau_{\max}^{\min\{p,q+1\}})$$

as $\tau_{\max} \to 0$.

It is possible to prove that there exists a continuous extension of differential algebraic order $[(p+1)/2]$ (see [22]).

References

[1] J. Agrawal, K.M. Moudgalya, A.K. Pani, Sliding motion of discontinuous dynamical systems described by semi-implicit index one differential algebraic equations, Chem. Eng. Sci. 61 (2006) 4722–4731.
[2] A. Bellen, M. Zennaro, Numerical Methods for Delay Differential Equations, Oxford Science Publications, 2002.
[3] M. Berardi, L. Lopez, On the continuous extension of Adams-Bashforth methods and the event location in discontinuous ODEs, Appl. Math. Lett. 25 (2015) 995–999.
[4] M. Biak, T. Hanus, D. Janovska, Some applications of Filippov's dynamical systems, J. Comput. Appl. Math. 254 (2013) 132–143.
[5] H. Brunner, S. Maset, Time transformation for delay differential equations, Discrete Contin. Dyn. Syst. 25 (2009) 751–775.
[6] H. Brunner, S. Maset, Time transformation for state-dependent delay differential equations, Commun. Pure Appl. Anal. 9 (2010) 23–45.
[7] N. Del Buono, C. Elia, L. Lopez, On the equivalence between the sigmoidal approach and Utkin's approach for piecewise-linear models of gene regulatory networks, SIAM J. Appl. Dyn. Syst. 13 (2014) 1270–1292.
[8] L. Dieci, C. Elia, L. Lopez, On Filippov solutions of discontinuous DAEs of index 1, Commun. Nonlinear Sci. Numer. Simul. 95 (2021) 105656.
[9] L. Dieci, L. Lopez, A survey of numerical methods for IVPs of ODEs with discontinuous right-hand side, J. Comput. Appl. Math. 236 (2012) 3967–3991.
[10] L. Dieci, L. Lopez, One-sided direct event location techniques in the numerical solution of discontinuous differential systems, BIT Numer. Math. 55 (2015) 987–1003.
[11] D.S. Galan, P.I. Barton, Dynamic optimization of hybrid systems, Comput. Chem. Eng. 22 (Suppl.) (1998) S183–S190.
[12] N. Guglielmi, E. Hairer, Computing breaking points in implicit delay differential equations, Adv. Comput. Math. 29 (2008) 229–247.
[13] E. Hairer, C. Lubich, M. Roche, The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods, Lecture Notes in Mathematics, Springer, 1989.
[14] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II. Stiff Differential-Algebraic Problems, second revised edition, Springer-Verlag Berlin Heidelberg, 1996.
[15] P. Kunkel, V. Mehrmann, Numerical solution of hybrid systems of differential-algebraic equations, Comput. Methods Appl. Mech. Eng. 197 (2008) 693–705.
[16] P. Kunkel, V. Mehrmann, Regular solutions of DAE hybrid systems and regularization techniques, BIT Numer. Math. 58 (2018) 1049–1077.
[17] L. Lopez, S. Maset, Time-transformations for the event location in discontinuous ODEs, Math. Comput. 87 (2018) 2321–2341.
[18] C. Majer, W. Marquardt, E.D. Gilles, Reinitialization of DAEs after discontinuities, Comput. Chem. Eng. 6 (Suppl.) (1995) 8507–8512.
[19] G. Mao, L.R. Petzold, Efficient integration over discontinuities for differential-algebraic systems, Comput. Math. Appl. 43 (2002) 65–79.
[20] M. Murad, M. Liu, F. Milano, Modeling and simulation of variable limits on conditional anti-windup PI controllers for VSC-based devices, IEEE Trans. Circuits Syst. I, Regul. Pap. 68 (7) (2021) 3079–3088.
[21] M. Najafi, R. Nikoukhah, Modeling and simulation of differential equations in Scicos, in: Modelica, The Modelica Association, 2006, pp. 177–185.
[22] A. Ostermann, Continuous extensions of Rosenbrock-type methods, Computing 44 (1990) 59–68.
[23] T. Park, P.I. Barton, State event location in differential-algebraic models, ACM Trans. Model. Comput. Simul. 6 (1996) 137–165.
[24] M. Roche, Rosenbrock methods for differential algebraic equations, Numer. Math. 52 (1987) 45–63.
[25] L. Shampine, R. Thompson, Event location for ordinary differential equations, Comput. Math. Appl. 39 (2000) 45–54.
[26] P. Stechlinski, M. Patrascu, P.I. Barton, Nonsmooth differential-algebraic equations in chemical engineering, Comput. Chem. Eng. 114 (2017) 52–68.
[27] P. Stechlinski, P.I. Barton, Nonsmooth Hessenberg differential-algebraic equations, J. Math. Anal. Appl. 495 (2021) 124721.