BIT Numerical Mathematics (2023) 63:29
https://doi.org/10.1007/s10543-023-00966-y

Discretization of inherent ODEs and the geometric integration of DAEs with symmetries

Peter Kunkel[1] · Volker Mehrmann[2]

Received: 17 May 2022 / Accepted: 20 March 2023 / Published online: 29 April 2023

Abstract
Discretization methods for differential-algebraic equations (DAEs) are considered that are based on the integration of an associated inherent ordinary differential equation (ODE). This allows to make use of any discretization scheme suitable for the numerical integration of ODEs. For DAEs with symmetries it is shown that the inherent ODE can be constructed in such a way that it inherits the symmetry properties of the given DAE and geometric properties of its flow. This in particular allows the use of geometric integration schemes with a numerical flow that has analogous geometric properties.

Keywords Differential-algebraic equation · Inherent ordinary differential equation · Geometric integration · Symplectic flow · Orthogonal flow

Mathematics Subject Classification 37J06 · 65L80 · 65L05 · 65P10

1 Introduction

We consider the numerical solution of general nonlinear systems of differential-algebraic equations (DAEs)

$$F(t, x, \dot{x}) = 0, \quad F \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}, \mathbb{R}^n) \text{ sufficiently smooth,} \tag{1.1}$$

Communicated by Antonella Zanna Munthe-Kaas.

Partially supported by the Research In Pairs program of *Mathematisches Forschungsinstitut Oberwolfach*, whose hospitality is gratefully acknowledged.

✉ Volker Mehrmann
mehrmann@math.tu-berlin.de

Peter Kunkel
kunkel@math.uni-leipzig.de

[1] Mathematisches Institut, Universität Leipzig, Augustusplatz 10, D-04109 Leipzig, Germany

[2] Institut für Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, Germany

where $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ are open domains and $\mathbb{I} \subseteq \mathbb{R}$ is a compact non-trivial interval, together with a given initial condition

$$x(t_0) = x_0, \quad t_0 \in \mathbb{I},\ x_0 \in \mathbb{D}_x. \tag{1.2}$$

For this task, numerous discretization schemes that work directly on (1.1) or on some index-reduced reformulation have been given in the literature, see e.g. [7, 10, 13]. In this paper, we consider discretization schemes that work on a so-called *inherent ordinary differential equation* (ODE) of the given DAE. The advantage of such an approach is that we can make use of any discretization scheme suitable for the numerical integration of ODEs. In particular, if we are able to choose the inherent ODE in such a way that it inherits symmetry properties of the given DAE and thus special properties of its flow, we may be in the situation to use *geometric integration*, i.e., to use special discretization schemes whose numerical flow possesses similar geometric properties, see [8].

In the context of geometric integration, we concentrate in this paper on linear problems. The basic principle of geometric integration in this special case is as follows. Given a linear initial value problem

$$\dot{x} = A(t)x + f(t), \quad x(t_0) = x_0, \tag{1.3}$$

its solution can be written by means of the variation of constant formula as

$$x(t; t_0; x_0) = \Phi(t)x_0 + \int_{t_0}^{t} \Phi(t)\Phi(s)^{-1} f(s)\, ds,$$

where $\Phi \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ is the solution of the matricial initial value problem

$$\dot{\Phi} = A(t)\Phi, \quad \Phi(t_0) = I_n. \tag{1.4}$$

In particular, we have that $\frac{d}{dx_0} x(t; t_0, x_0) = \Phi(t)$. If A now lies pointwise in a Lie algebra then the flow Φ lies pointwise in the corresponding Lie group. The idea of geometric integration is then to construct numerical integration methods that inherit this geometric property. If we write the numerical solution after one step as $x_h(t_0 + h; t_0, x_0)$, we therefore require $\frac{d}{dx_0} x_h(t_0 + h; t_0, x_0)$ also to lie in this Lie group.

In the case of a *quadratic Lie group*

$$\mathbb{G} = \{G \in \mathrm{GL}(n) | G^T X G = X\}, \tag{1.5}$$

with some given $X \in \mathbb{R}^{n,n}$ and $\mathrm{GL}(n)$ denoting the general linear group of invertible matrices in $\mathbb{R}^{n,n}$, and its associated *Lie algebra*

$$\mathbb{A} = \{A \in \mathbb{R}^{n,n} | A^T X + XA = 0\}, \tag{1.6}$$

the above property that Φ lies pointwise in $\mathbb{G}$ if A lies pointwise in $\mathbb{A}$ follows from $\Phi(t_0)^T X\Phi(t_0) = X$ and

$$\tfrac{d}{dt}(\Phi^T X\Phi) = \dot{\Phi}^T X\Phi + \Phi^T X\dot{\Phi} = \Phi^T A^T X\Phi + \Phi^T XA\Phi = \Phi^T(A^T X + XA)\Phi = 0.$$

In particular, $\Phi^T X\Phi$ is a quadratic invariant of (1.4). It can then be shown that all Runge–Kutta methods that conserve quadratic invariants constitute geometric integration methods in the case of quadratic Lie groups and their Lie algebras, see [8]. A prominent class of Runge–Kutta methods that conserve quadratic invariants are given by *Gauß collocation*, see again [8].

Writing linear time-varying DAEs in the form

$$E(t)\dot{x} = A(t)x + f(t), \quad E, A \in C(\mathbb{I}, \mathbb{R}^{n,n}),\ f \in C(\mathbb{I}, \mathbb{R}^n) \text{ sufficiently smooth,} \tag{1.7}$$

we are interested in this paper in the following symmetry properties.

Definition 1.1 The DAE (1.7) and its associated pair (E, A) of matrix functions are called *self-adjoint* if

$$E^T = -E, \quad A^T = A + \dot{E} \tag{1.8}$$

as equality of functions.

Definition 1.2 The DAE (1.7) and its associated pair (E, A) of matrix functions are called *skew-adjoint* if

$$E^T = E, \quad A^T = -A - \dot{E} \tag{1.9}$$

as equality of functions.

For these two cases, it has been shown in [15] that the inherent ODE can be chosen in such a way that its flow possesses certain geometric properties posing the question of possible geometric integration. In particular, we are here concerned with the quadratic Lie group $\mathrm{Sp}(2p)$ of symplectic matrices related to

$$X = J, \quad J = \begin{bmatrix} 0 & I_p \\ -I_p & 0 \end{bmatrix} \tag{1.10}$$

and the associated Lie algebra of *Hamiltonian matrices* in the case of self-adjoint DAEs and with the quadratic Lie group $\mathrm{O}(p, q)$ of *generalized orthogonal matrices* related to

$$X = S, \quad S = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} \tag{1.11}$$

in the case of skew-adjoint DAEs.

2 Preliminaries

In the following, we give a concise overview of the relevant theory on DAEs that we make use of, see e.g. [13]. The basis are the so-called *derivative array equations*

$$F_\ell(t,x,\dot{x},\dots,x^{(\ell+1)})=0, \tag{2.1}$$

see [3], where F_ℓ has the form

$$F_\ell(t,x,\dot{x},\dots,x^{(\ell+1)})=\begin{bmatrix} F(t,x,\dot{x}) \\ \frac{d}{dt}F(t,x,\dot{x}) \\ \vdots \\ \left(\frac{d}{dt}\right)^\ell F(t,x,\dot{x}) \end{bmatrix}$$

with Jacobians (denoting the derivative of F with respect to the variable x by F_x and accordingly)

$$\begin{aligned} M_\ell(t,x,\dot{x},\dots,x^{(\ell+1)}) &= F_{\ell;\dot{x},\dots,x^{(\ell+1)}}(t,x,\dot{x},\dots,x^{(\ell+1)}), \\ N_\ell(t,x,\dot{x},\dots,x^{(\ell+1)}) &= -[\, F_{\ell;x}(t,x,\dot{x},\dots,x^{(\ell+1)}) \;\; 0 \;\; \dots \;\; 0\,]. \end{aligned} \tag{2.2}$$

The following hypothesis then states sufficient conditions for the given DAE to describe a *regular problem*.

Hypothesis 2.1 *There exist integers μ, a, and d such that the set*

$$\mathbb{L}_\mu=\{(t,x,\dot{x},\dots,x^{(\mu+1)})\in\mathbb{R}^{(\mu+2)n+1} | F_\mu(t,x,\dot{x},\dots,x^{(\mu+1)})=0\} \tag{2.3}$$

associated with F is nonempty and such that for every $(t_0,x_0,\dot{x}_0,\dots,x_0^{(\mu+1)})\in\mathbb{L}_\mu$, there exists a (sufficiently small) neighborhood in which the following properties hold:

1. *We have* rank $M_\mu(t,x,\dot{x},\dots,x^{(\mu+1)})=(\mu+1)n-a$ *on $\mathbb{L}_\mu$ such that there exists a smooth matrix function Z_2 of size $(\mu+1)n\times a$ and pointwise maximal rank, satisfying $Z_2^T M_\mu=0$ on $\mathbb{L}_\mu$.*
2. *We have* rank $\hat{A}_2(t,x,\dot{x},\dots,x^{(\mu+1)})=a$*, where $\hat{A}_2=Z_2^T N_\mu[I_n\, 0\,\cdots\, 0]^T$ such that there exists a smooth matrix function T_2 of size $n\times d$, $d=n-a$, and pointwise maximal rank, satisfying $\hat{A}_2T_2=0$.*
3. *We have* rank $F_{\dot{x}}(t,x,\dot{x})T_2(t,x,\dot{x},\dots,x^{(\mu+1)})=d$ *such that there exists a smooth matrix function Z_1 of size $n\times d$ and pointwise maximal rank, satisfying* rank $\hat{E}_1T_2=d$*, where $\hat{E}_1=Z_1^TF_{\dot{x}}$.*

Note that the local existence of functions Z_2, T_2, Z_1 can be guaranteed by the application of the implicit function theorem, see [13, Theorem 4.3]. Moreover, we may assume that they possess (pointwise) orthonormal columns. Note also that due to the full rank requirement we may choose Z_1 to be constant.

Following the presentation in [11], we use the shorthand notation $y = (\dot{x}, \ldots, x^{(\mu+1)})$ and similarly $y_0 = (\dot{x}_0, \ldots, x_0^{(\mu+1)})$. The system of nonlinear equations

$$H(t, x, y, \alpha) = \begin{bmatrix} F_\mu(t, x, y) - Z_{2,0}\alpha \\ T_{1,0}^T(y - y_0) \end{bmatrix}, \tag{2.4}$$

with the columns of $T_{1,0}$ forming an orthonormal basis of kernel $F_{\mu;y}(t_0, x_0, y_0)$ and $Z_{2,0} = Z_2(t_0, x_0, y_0)$ according to Hypothesis 2.1, is then locally solvable for y, α in terms of (t, x) due to the implicit function theorem. In particular, $\alpha = \hat{F}_2(t, x)$ with some function $\hat{F}_2$. One can show that $\hat{F}_2(t, x) = 0$ describes the whole set of algebraic constraints implied by the original DAE. Setting furthermore $\hat{F}_1(t, x, \dot{x}) = Z_1^T F(t, x, \dot{x})$ yields a so-called *reduced DAE*

$$\begin{aligned} &\hat{F}_1(t, x, \dot{x}) = 0, && \text{(d differential equations)} \\ &\hat{F}_2(t, x) = 0, && \text{(a algebraic equations)} \end{aligned} \tag{2.5}$$

in the sense that it satisfies Hypothesis 2.1 with $\mu = 0$.

Moreover, one can show that $\hat{F}_{2;x}$ possesses full row rank implying that we can split x possibly after a renumeration of the components according to $x = (x_1, x_2)$ such that $\hat{F}_{2;x_2}$ is nonsingular. The implicit function theorem then yields $x_2 = \mathscr{R}(t, x_1)$ with some function $\mathscr{R}$. Differentiating this relation to eliminate x_2 and $\dot{x}_2$ in the first equation of (2.5), we can apply the implicit function theorem once more (requiring the solvability of the DAE) yielding $\dot{x}_1 = \mathscr{L}(t, x_1)$, a so-called *inherent ODE*, with some function $\mathscr{L}$. Putting both parts together, we end up with a second kind of reduced DAE

$$\begin{aligned} &\dot{x}_1 = \mathscr{L}(t, x_1), && \text{(d differential equations)} \\ &x_2 = \mathscr{R}(t, x_1). && \text{(a algebraic equations)} \end{aligned} \tag{2.6}$$

Note that, once we have fixed the splitting of the variables, the constructed functions $\mathscr{L}$ and $\mathscr{R}$ are unique. In particular, the set $\mathbb{L}_{\mu+1}$ can be locally parameterized according to

$$F_{\mu+1}(t, x_1, \mathscr{R}(t, x_1), \mathscr{L}(t, x_1), \mathscr{R}_t(t, x_1) + \mathscr{R}_{x_1}(t, x_1)\mathscr{L}(t, x_1), \mathscr{W}(t, x_1, p)) \equiv 0 \tag{2.7}$$

with a suitable parameter $p \in \mathbb{R}^a$ and a related function $\mathscr{W}$.

Under some technical assumptions, see [13], the original DAE and the reduced DAEs (2.5) and (2.6) possess the same solutions. As a consequence, we may discretize the reduced DAEs instead of the original DAE utilizing the better properties of the latter ones. But this requires the possibility to evaluate the implicitly defined functions. In the case of $\hat{F}_2$ in (2.5) the standard approach, see [13], is to go back to the definition of $\hat{F}_2$ in such a way that we replace $\hat{F}_2(t, x) = 0$ by $F_\mu(t, x, y) = 0$.

In the special case of linear time-varying DAEs (1.7), the Jacobians M_μ, N_μ used in Hypothesis 2.1 only depend on t such that the functions Z_2, T_2, Z_1 can be chosen to depend also only on t. The corresponding reduced DAE (2.5) then takes the form

$$\begin{aligned} \hat{E}_1(t)\dot{x} &= \hat{A}_1(t)x + \hat{f}_1(t), \text{ (d differential equations)} \\ 0 &= \hat{A}_2(t)x + \hat{f}_2(t), \text{ (a algebraic equations)} \end{aligned} \tag{2.8}$$

where

$$\begin{aligned} \hat{E}_1 = Z_1^T E,\ \hat{A}_1 &= Z_1^T A, & \hat{f}_1 &= Z_1^T f, \\ \hat{A}_2 &= Z_2^T N_\mu [I_n\ 0\ \cdots\ 0]^T, & \hat{f}_2 &= Z_2^T g_\mu \end{aligned} \tag{2.9}$$

with

$$M_\mu = \begin{bmatrix} E & & & \\ \dot{E} - A & E & & \\ \ddot{E} - 2\dot{A} & 2\dot{E} - A & E & \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix},\ N_\mu = \begin{bmatrix} A & 0 & \cdots & 0 \\ \dot{A} & 0 & \cdots & 0 \\ \ddot{A} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \end{bmatrix},\ g_\mu = \begin{bmatrix} f \\ \dot{f} \\ \ddot{f} \\ \vdots \end{bmatrix}. \tag{2.10}$$

The splitting of the variables as $x = (x_1, x_2)$ that leads to the second form of a reduced DAE corresponds to a splitting of $\hat{A}_2 = [\ A_{21}\ \ A_{22}\ ]$ with the requirement that A_{22} is pointwise nonsingular. It is then obvious that we can solve the second equation of (2.8) for x_2 in terms of x_1, differentiate, and eliminate x_2 and $\dot{x}_2$ in the first equation of (2.8) to obtain a linear version of (2.6).

In order to utilize global canonical forms as they were presented in [15], we observe that the construction of (2.8) transforms covariantly with global equivalence transformations as follows. Let $(\tilde{E}, \tilde{A})$ be globally equivalent to (E, A), i.e., let sufficiently smooth, pointwise nonsingular matrix functions $P \in C(\mathbb{I}, \mathbb{R}^{n,n})$ and $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ be given such that

$$\tilde{E} = PEQ, \quad \tilde{A} = PAQ - PE\dot{Q}, \tag{2.11}$$

describing scalings of the DAE (1.7) and the unknown x, respectively. The corresponding Jacobians are then related by

$$\tilde{M}_\mu = \Pi_\mu M_\mu \Theta_\mu, \quad \tilde{N}_\mu = \Pi_\mu N_\mu \Theta_\mu - \Pi_\mu M_\mu \Psi_\mu \tag{2.12}$$

with

$$\Pi_\mu = \begin{bmatrix} P & & & \\ \dot{P} & P & & \\ \ddot{P} & 2\dot{P} & P & \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix},\ \Theta_\mu = \begin{bmatrix} Q & & & \\ 2\dot{Q} & Q & & \\ 3\ddot{Q} & 3\dot{Q} & Q & \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix},\ \Psi_\mu = \begin{bmatrix} \dot{Q} & 0 & \cdots & 0 \\ \ddot{Q} & 0 & \cdots & 0 \\ \dddot{Q} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \end{bmatrix}. \tag{2.13}$$

With given choices Z_2, T_2, Z_1 for (E, A) along Hypothesis 2.1 we may choose $\tilde{Z}_2, \tilde{T}_2, \tilde{Z}_1$ for $(\tilde{E}, \tilde{A})$ as

$$\tilde{Z}_2^T = Z_2^T \Pi_\mu^{-1}, \quad \tilde{T}_2 = Q^{-1} T_2, \quad \tilde{Z}_1^T = Z_1^T P^{-1}. \tag{2.14}$$

Having summarized the theory for general nonlinear and linear time-varying DAEs, the next section deals with the construction of suitable inherent ODEs for a given DAE.

3 Construction and evaluation of an inherent ODE

To get more flexibility into the choice of an inherent ODE, we introduce a (linear but in general time-dependent) transformation of the unknown x before we perform the splitting, i.e., we consider

$$x = Q(t) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{3.1}$$

where $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ is sufficiently smooth and pointwise nonsingular. According to [13, Lemma 4.6] the so transformed DAE (1.1) satisfies Hypothesis 2.1 as well with the same characteristic values μ, a, d. As before, the only requirement for Q is that we can solve the algebraic constraints for x_2 in terms of x_1. Writing

$$Q = [\, T_2 \;\; T_2' \,], \tag{3.2}$$

the algebraic constraints read

$$\hat{F}_2(t, T_2 x_1 + T_2' x_2) = 0.$$

Hence, in order to be able to solve for x_2 we need $\hat{F}_{2;x} T_2'$ to be pointwise nonsingular. If this is the case, then the chosen Q fixes a reduced DAE of the form (2.6) satisfying

$$\begin{gathered} F_{\mu+1}\left(t, Q(t)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, Q(t)\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} + \dot{Q}(t)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \mathscr{W}(t, x_1, p)\right) \equiv 0, \\ \dot{x}_1 = \mathscr{L}(t, x_1),\; x_2 = \mathscr{R}(t, x_1),\; \dot{x}_2 = \mathscr{R}_t(t, x_1) + \mathscr{R}_{x_1}(t, x_1)\mathscr{L}(t, x_1) \end{gathered} \tag{3.3}$$

with a suitable parameter $p \in \mathbb{R}^a$ and a related function $\mathscr{W}$.

For a numerical realization, we are confronted with two problems. First, we must be able to evaluate the implicitly defined functions $\mathscr{L}$ and $\mathscr{R}$. Second, for a nontrivial choice of Q we must have access to $\dot{Q}$.

In the next subsections, we discuss how to overcome these problems.

3.1 Numerical evaluation of the inherent ODE

The first problem can be dealt with by solving the system of (nonlinear) equations

$$F_{\mu+1}(t,x,\dot{x},w)=0,\quad [\,I_d\;\;0\,]Q(t)^{-1}x=x_1 \tag{3.4}$$

for given (t,x_1). Because of the first part in (3.4), at a solution, the resulting $(t,x,\dot{x},w)$ must satisfy

$$x=Q(t)\begin{bmatrix} x_1\\ \mathscr{R}(t,x_1)\end{bmatrix},\;\dot{x}=Q(t)\begin{bmatrix} \mathscr{L}(t,x_1)\\ \mathscr{R}_t(t,x_1)+\mathscr{R}_{x_1}(t,x_1)\mathscr{L}(t,x_1)\end{bmatrix}$$
$$+\dot{Q}(t)\begin{bmatrix} x_1\\ \mathscr{R}(t,x_1)\end{bmatrix}.$$

Because of the second part in (3.4), we regain the prescribed x_1. Furthermore, we observe that

$$\mathscr{R}(t,x_1)=[\,0\;\;I_a\,]Q(t)^{-1}x,\;\mathscr{L}(t,x_1)=[\,I_d\;\;0\,]Q(t)^{-1}(\dot{x}-\dot{Q}(t)Q(t)^{-1}x)$$

yielding the required evaluations of $\mathscr{L}$ and $\mathscr{R}$.

Since (3.4) constitutes an underdetermined system of equations, the method of choice to solve (3.4) numerically is the *Gauß-Newton method*. In order to show that the Gauß-Newton method will convergence quadratically for sufficiently good starting values, we need to show that the Jacobian at a solution possesses full row rank, see e.g. [5].

Theorem 3.1 *Let (1.1) satisfy Hypothesis 2.1 both with μ,a,d and with $\mu+1,a,d$. Then, the Jacobian of (3.4) possesses full row rank at every solution provided that $\hat{F}_{2;x}T_2'$ is pointwise nonsingular.*

Proof Due to (2.4) for $\mu+1$ replacing μ we have

$$F_{\mu+1;x}-Z_{2,0}\hat{F}_{2;x}=0,$$

omitting for convenience the arguments here and later. Hence,

$$\hat{F}_{2;x}=(Z_2^TZ_{2,0})^{-1}Z_2^TF_{\mu+1;x}$$

in a sufficiently small neighborhood. Completing Z_2 to a pointwise nonsingular matrix function $[\,Z_2'\;\;Z_2\,]$, elementary row operations of the Jacobian of the first part in (3.4) yield

$$\begin{bmatrix} F_{\mu+1;x} & F_{\mu+1;\dot{x},\ldots,x^{(\mu+2)}}\end{bmatrix}\to\begin{bmatrix} Z_2'^TF_{\mu+1;x} & Z_2'^TF_{\mu+1;\dot{x},\ldots,x^{(\mu+2)}}\\ Z_2^TF_{\mu+1;x} & 0\end{bmatrix}.$$

According to Hypothesis 2.1 the entry $Z_2'^TF_{\mu+1;\dot{x},\ldots,x^{(\mu+2)}}$ possesses full row rank such that we are left with the entry $Z_2^TF_{\mu+1;x}$ together with the Jacobian $[\,I_d\;\;0\,]Q^{-1}$

of the second equation in (3.4). Multiplying the first part with $(Z_2^T Z_{2,0})^{-1}$ from the left and both parts with Q from the right yields the matrix function

$$\begin{bmatrix} \hat{F}_{2;x}T_2 & \hat{F}_{2;x}T_2' \\ I_d & 0 \end{bmatrix}$$

which is pointwise nonsingular provided that $\hat{F}_{2;x}T_2'$ is pointwise nonsingular. □

3.2 Numerical construction of the transformation

It remains the question how we can deal with $\dot{Q}$ in extracting the evaluation of $\mathscr{L}(t, x_1)$. In particular, we are interested in applications where a trivial choice as constant Q or beforehand given Q with implemented functions to evaluate both $Q(t)$ and $\dot{Q}(t)$ is not possible but where Q has to be chosen numerically during the integration of the DAE. The main problem in this context is that we must choose Q in a smooth way, at least on the current interval $[t_0, t_0 + h]$ of the numerical integration with $h > 0$ sufficiently small, and that we must be able to evaluate $\dot{Q}$.

The approach we will follow here is automatic differentiation, see [6]. This means that we work not only with the value of a variable but with a pair of numbers that represent the value and the derivative of a variable. Operations on such pairs are then defined by means of the known differentiation rules. If we use the notation $\langle x, \dot{x}\rangle$ for such a pair, the typical operations used in linear algebra then read

$$\begin{array}{ll} \text{(a)} & \langle x, \dot{x}\rangle + \langle y, \dot{y}\rangle = \langle x + y, \dot{x} + \dot{y}\rangle, \\ \text{(b)} & \langle x, \dot{x}\rangle - \langle y, \dot{y}\rangle = \langle x - y, \dot{x} - \dot{y}\rangle, \\ \text{(c)} & \langle x, \dot{x}\rangle \cdot \langle y, \dot{y}\rangle = \langle x \cdot y, \dot{x} \cdot y + x \cdot \dot{y}\rangle, \\ \text{(d)} & \langle x, \dot{x}\rangle / \langle y, \dot{y}\rangle = \langle x/y, (\dot{x} - x \cdot \dot{y}/y)/y\rangle, \\ \text{(e)} & \sqrt{\langle x, \dot{x}\rangle} = \langle \sqrt{x}, \tfrac{1}{2}\dot{x}/\sqrt{x}\rangle. \end{array} \tag{3.5}$$

These operations can be obviously extended in a componentwise way to vector and matrix operations.

Note that in a programming language like C++ this approach can be implemented by defining a corresponding new class and overloading the above operations to work with this class. In this way it is possible to perform tasks of linear algebra like Cholesky decomposition $A = L \cdot L^T$ in a smooth way yielding $\langle L, \dot{L}\rangle$ for given $\langle A, \dot{A}\rangle$. This is valid for all numerical algorithms that do not include if-clauses. If there are if-clauses, as for example in the QR decomposition $A \cdot \Pi = Q \cdot R$, then we can at least locally get a smooth version. To do this for the QR decomposition, we may proceed as follows. For a reference point, typically t_0, we perform a standard QR decomposition $A(t_0) \cdot \Pi_0 = Q_0 \cdot R_0$. We then freeze all if-clauses and use automatic differentiation in the evaluation of the QR decomposition $A \cdot \Pi_0 = Q \cdot R$. In this way, we get $\langle Q, \dot{Q}\rangle$ and $\langle R, \dot{R}\rangle$ for given $\langle A, \dot{A}\rangle$.

In particular, we can use this approach to perform the construction of reduced DAEs for linear time-varying systems as described in Sect. 2 with the aim to get not only values for the involved transformations but also values for their derivatives.

To start the construction of the reduced system (2.8), we need $\dot{M}_\mu, \dot{N}_\mu, \dot{g}_\mu$ besides M_μ, N_μ, g_μ. Writing M, N, g for the formally infinite extensions of M_μ, N_μ, g_μ and defining

$$S = \begin{bmatrix} 0 & & & \\ I_n & 0 & & \\ & I_n & 0 & \\ & & \ddots & \ddots \end{bmatrix}, \quad V = \begin{bmatrix} I_n \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

we have the relations

$$\dot{M} = S^T M - M S^T + N, \quad \dot{N} = S^T N, \quad \dot{g} = S^T g,$$

see [4]. Hence, from the evaluations $M_{\mu+1}, N_{\mu+1}, g_{\mu+1}$ we can actually retrieve the desired $\langle M_\mu, \dot{M}_\mu \rangle$, $\langle N_\mu, \dot{N}_\mu \rangle$,and $\langle g_\mu, \dot{g}_\mu \rangle$. A first locally smooth QR decomposition then yields $\langle Z_2, \dot{Z}_2 \rangle$ and thus $\langle \hat{A}_2, \frac{d}{dt}\hat{A}_2 \rangle$. A second locally smooth QR decomposition then gives $\langle T_2, \dot{T}_2 \rangle$ and with a third locally smooth QR decomposition for $\langle E, \dot{E} \rangle \cdot \langle T_2, \dot{T}_2 \rangle$ we finally get $\langle Z_1, \dot{Z}_1 \rangle$. In the latter case we can also use a standard QR decomposition once at t_0 and use the so obtained $Z_{1,0}$ to set $\langle Z_1, \dot{Z}_1 \rangle = \langle Z_{1,0}, 0 \rangle$ if it seems more suited. The remaining quantities of the reduced DAE are then given by automatic differentiation along the lines of (2.9).

With a given choice $\langle Q, \dot{Q} \rangle$ for fixing an inherent ODE, transforming the reduced DAE (2.5) by means of (3.1) yields

$$\begin{aligned} \hat{E}_{11}(t)\dot{x}_1 + \hat{E}_{12}(t)\dot{x}_2 &= \hat{A}_{11}(t)x_1 + \hat{A}_{12}(t)x_2 + \hat{f}_1(t), \\ 0 &= \hat{A}_{21}(t)x_1 + \hat{A}_{22}(t)x_2 + \hat{f}_2(t), \end{aligned}$$

where

$$\begin{aligned} &\hat{E}_{11} = \hat{E}_1 T_2, && \hat{E}_{12} = \hat{E}_1 T_2', \\ &\hat{A}_{11} = \hat{A}_1 T_2 - \hat{E}_1 \dot{T}_2, && \hat{A}_{12} = \hat{A}_1 T_2' - \hat{E}_1 \dot{T}_2', \\ &\hat{A}_{21} = \hat{A}_2 T_2, && \hat{A}_{22} = \hat{A}_2 T_2', \end{aligned}$$

and we are in the same situation as in the special case described in Sect. 2. In particular, we can solve for x_2, differentiate, eliminate, and solve for $\dot{x}_1$ to get the fixed inherent ODE.

A special choice of Q can be obtained by a locally smooth QR decomposition of $\langle \hat{E}_1^T, \frac{d}{dt}\hat{E}_1^T \rangle$ leading to $\hat{E}_{12} = 0$. Hypothesis 2.1 then guarantees that $\hat{A}_{22}$ is pointwise nonsingular. If we set $Q_0 = Q(t_0)$ and $\dot{Q}_0 = \dot{Q}(t_0)$, we may also replace Q by the constant version $Q(t) = Q_0$ or by the linearized version $Q(t) = Q_0 + (t - t_0)\dot{Q}_0$. The latter corresponds to the construction of so-called *spin-stabilized integrators* introduced in [14]. In the case that $\mu = 0$, the constructions can be simplified by using E instead of $\hat{E}_1$ since no construction of a reduced system is required.

4 Symmetries and geometric integration

In this section we treat linear time-varying DAEs that are self-adjoint or skew-adjoint. The aim is to utilize the symmetry in the construction of a suitable inherent ODE such that it inherits certain properties of the original DAE. Note that self-adjointness and skew-adjointness are invariant under so-called congruence, i.e., under global equivalence (2.11) with $P = Q^T$, see e.g. [15]. As there, we will write $(\tilde{E}, \tilde{A}) \equiv (E, A)$ to indicate that the pairs are congruent. Note also that regularity of a pair (E, A) of sufficiently smooth matrix function $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ is necessary and sufficient for the asscociated DAE (1.7) to satisfy Hypothesis 2.1, see e.g. [13].

4.1 Self-adjoint DAEs

Assuming (1.8) for (1.7), we will make use of the following global canonical form taken from [15] in a slightly rephrased version.

Theorem 4.1 *Let (E, A) with $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be sufficiently smooth and let the associated DAE (1.7) satisfy Hypothesis 2.1. If (E, A) is self-adjoint, then we have that*

$$(E, A) \equiv \left(\begin{bmatrix} 0 & I_p & 0 \\ -I_p & 0 & 0 \\ 0 & 0 & E_{33} \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix} \right), \tag{4.1}$$

where

$$E_{33}(t)\dot{x}_3 = A_{33}(t)x_3 + f_3(t), \tag{4.2}$$

is uniquely solvable for every sufficiently smooth f_3 without specifying initial conditions. Furthermore,

$$E_{33}^T = -E_{33}, \quad A_{22}^T = A_{22}, \quad A_{32}^T = A_{23}, \quad A_{33}^T = A_{33} + \dot{E}_{33}. \tag{4.3}$$

In order to construct a suitable reduced DAE (2.8), we follow the lines of Hypothesis 2.1 for the global canonical form, indicated by tildes, and start with

$$\tilde{M}_\mu = \left[\begin{array}{ccc|ccc|ccc|c} 0 & I_p & 0 & & & & & & & \\ -I_p & 0 & 0 & & & & & & & \\ 0 & 0 & E_{33} & & & & & & & \\ \hline 0 & 0 & 0 & 0 & I_p & 0 & & & & \\ 0 & -A_{22} & -A_{23} & -I_p & 0 & 0 & & & & \\ 0 & -A_{32} & \dot{E}_{33} - A_{33} & 0 & 0 & E_{33} & & & & \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_p & 0 & \\ 0 & -2\dot{A}_{22} & -2\dot{A}_{23} & 0 & -A_{22} & -A_{23} & -I_p & 0 & 0 & \\ 0 & -2\dot{A}_{32} & \ddot{E}_{33} - 2\dot{A}_{33} & 0 & -A_{32} & 2\dot{E}_{33} - A_{33} & 0 & 0 & E_{33} & \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array} \right].$$

Due to the identities, the only possible rank-deficiency is related to the part belonging to the pair (E_{33}, A_{33}). The properties of (4.2) then imply that $d = 2p$ and $a = n-2p$ in Hypothesis 2.1. Furthermore, the left null space of $\tilde{M}_\mu$ is described by

$$\tilde{Z}_2^T = \left[* \; 0 \; \tilde{Z}_{2,0}^T \middle| * \; 0 \; \tilde{Z}_{2,1}^T \middle| * \; 0 \; \tilde{Z}_{2,2}^T \middle| \cdots \right].$$

Observing that

$$\tilde{N}_\mu [I_n \; 0 \; \cdots \; 0]^T = \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \\ \hline 0 & 0 & 0 \\ 0 & \dot{A}_{22} & \dot{A}_{23} \\ 0 & \dot{A}_{32} & \dot{A}_{33} \\ \hline 0 & 0 & 0 \\ 0 & \ddot{A}_{22} & \ddot{A}_{23} \\ 0 & \ddot{A}_{32} & \ddot{A}_{33} \\ \hline \vdots & \vdots & \vdots \end{array}\right],$$

we get

$$\hat{A}_2 = \left[0 \; \hat{A}_{32} \; I_a \right]$$

for the second part of Hypothesis 2.1, where the identity comes from a special choice of $\tilde{Z}_2^T$. Choosing

$$\tilde{T}_2 = \begin{bmatrix} I_p & 0 \\ 0 & I_p \\ 0 & -\hat{A}_{32} \end{bmatrix}$$

and $\tilde{Z}_1 = \tilde{T}_2$ yields

$$\tilde{Z}_1^T \tilde{E} \tilde{T}_2 = \begin{bmatrix} I_p & 0 & 0 \\ 0 & I_p & -\hat{A}_{32}^T \end{bmatrix} \begin{bmatrix} 0 & I_p & 0 \\ -I_p & 0 & 0 \\ 0 & 0 & E_{33} \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & I_p \\ 0 & -\hat{A}_{32} \end{bmatrix} = \begin{bmatrix} 0 & I_p \\ -I_p & \hat{A}_{32}^T E_{33} \hat{A}_{32} \end{bmatrix},$$

which is indeed pointwise nonsingular, thus satisfying the third part of Hypothesis 2.1. In particular, the special choice $\tilde{Z}_1 = \tilde{T}_2$ is possible. According to (2.14) with $P = Q^T$ we can also choose $Z_1 = T_2$ for the original pair such that the reduced DAE inherits some symmetry properties of the original DAE. Note also that we may assume that T_2 possesses pointwise orthonormal columns.

By construction, the matrix function $T_2^T E T_2$ is not only pointwise skew-symmetric but also pointwise nonsingular. We can then proceed similar to [16]. Setting

$$T_2^T E T_2 = \begin{bmatrix} \bar{E} & c \\ -c^T & 0 \end{bmatrix},$$

there exists a smooth pointwise orthogonal transformation U with $U^T c = \alpha e_1, \alpha \neq 0$, where e_1 denotes the first canonical basis vector of appropriate size, see e.g. [13, Theorem 3.9]. It follows that

$$\begin{bmatrix} U & \\ & 1 \end{bmatrix}^T \begin{bmatrix} \bar{E} & c \\ -c^T & 0 \end{bmatrix} \begin{bmatrix} U & \\ & 1 \end{bmatrix} = \begin{bmatrix} U^T \bar{E} U & \alpha e_1 \\ -\alpha e^T & 0 \end{bmatrix} = \begin{bmatrix} * & * & \alpha \\ * & \bar{\bar{E}} & 0 \\ -\alpha & 0 & 0 \end{bmatrix},$$

where $\bar{\bar{E}}$ is again skew-symmetric and pointwise nonsingular. Thus, inductively after p steps, we arrive at

$$W_1^T T_2^T E T_2 W_1 = \begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ -\tilde{E}_{12}^T & 0 \end{bmatrix},$$

where W_1 collects all the applied transformations. By construction, $\tilde{E}_{11}$ is skew-symmetric and $\tilde{E}_{12}$ is anti-triangular and pointwise nonsingular. Finally, setting

$$W_2 = \begin{bmatrix} I_p & 0 \\ -\frac{1}{2}\tilde{E}_{12}^{-1}\tilde{E}_{11} & \tilde{E}_{12}^{-1} \end{bmatrix}$$

yields

$$W_2^T W_1^T T_2^T E T_2 W_1 W_2 = \begin{bmatrix} 0 & I_p \\ -I_p & 0 \end{bmatrix} = J.$$

For convenience, we write again T_2 instead of the transformed $T_2 W_1 W_2$. Completing T_2 to a pointwise nonsingular Q according to (3.2), we get

$$Q^T E Q = \begin{bmatrix} J & \hat{E}_{12} \\ * & * \end{bmatrix}, \quad Q^T A Q - Q^T E \dot{Q} = \begin{bmatrix} C & \hat{A}_{12} \\ * & * \end{bmatrix}.$$

Since self-adjointness is invariant under congruence and J is constant, the matrix function C is pointwise symmetric. With (3.1) the reduced DAE transforms to

$$\begin{aligned} J\dot{x}_1 + \hat{E}_{12}(t)\dot{x}_2 &= C(t)x_1 + \hat{A}_{12}(t)x_2 + T_2(t)^T f(t), \\ 0 &= \hat{A}_{22}(t)x_2 + \hat{f}_2(t), \end{aligned}$$

where $\hat{A}_{22} = \hat{A}_2 T_2'$ is pointwise nonsingular. Solving the second equation for x_2, differentiating, and eliminating x_2 and $\dot{x}_2$ from the first equation yields the inherent ODE

$$\dot{x}_1 = J^{-1}C(t)x_1 + \tilde{f}_1(t) \tag{4.4}$$

with some transformed inhomogeneity $\tilde{f}_1$.

Theorem 4.2 *Let (E, A) with $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be sufficiently smooth and let the associated DAE (1.7) satisfy Hypothesis 2.1. If (E, A) is self-adjoint, then Q in (3.1) can be chosen from a restricted class of transformations in such a way that the ODE (1.4) belonging to the so constructed inherent ODE possesses a symplectic flow.*

Proof The above construction shows that it is possible to fix an inherent ODE such that the associated ODE (1.4) has a symplectic flow. It is special in the sense that it works with pointwise orthogonal transformations with the exception of W_2 which transforms within one half of the variables and adapts the other half to obtain the matrix J and thus a set of variables for which the inherent ODE is Hamiltonian. □

In the special case $\mu = 0$ a slightly simplified construction is possible. Here, Hypothesis 2.1 says that E has constant rank allowing to choose Q in the form (3.2) such that

$$Q^T E Q = \begin{bmatrix} \hat{E}_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

with $\hat{E}_{11} = T_2^T E T_2$ pointwise nonsingular. Then, the same modifications of T_2 as before are possible leading to a modified T_2 with $\hat{E}_{11} = J$. With the corresponding modified Q, observing $ET_2' = 0$, we get that

$$Q^T E Q = \begin{bmatrix} J & 0 \\ 0 & 0 \end{bmatrix}, \quad Q^T A Q - Q^T E \dot{Q} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}.$$

Since congruence conserves self-adjointness, see e.g. [16], we have $\hat{A}_{11}^T = \hat{A}_{11}$, $\hat{A}_{12}^T = \hat{A}_{21}$, and $\hat{A}_{22}^T = \hat{A}_{22}$. Moreover, Hypothesis 2.1 with $\mu = 0$ requires that $\hat{A}_{22}$ is pointwise nonsingular. The corresponding reduced DAE, which is here just the original DAE, transforms to

$$\begin{aligned} J\dot{x}_1 &= \hat{A}_{11}(t)x_1 + \hat{A}_{12}(t)x_2 + T_2(t)^T f(t), \\ 0 &= \hat{A}_{12}(t)^T x_1 + \hat{A}_{22}(t)x_2 + T_2'(t)^T f(t). \end{aligned}$$

Solving the second equation for x_2 and eliminating it from the first equation, we again obtain an inherent ODE of the form (4.4), where

$$C = \hat{A}_{11} - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{A}_{12}^T$$

is pointwise symmetric.

Theoretically, all constructions can be performed globally. For a numerical realization one typically uses locally smooth variants as described in Sect. 3, which in this case is straightforward on the basis of locally smooth QR decompositions.

4.2 Skew-adjoint DAEs

Assuming (1.9) for (1.7), we will make use of the following global canonical form taken from [15] in a slightly rephrased version.

Theorem 4.3 *Let (E, A) with $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be sufficiently smooth and let the associated DAE (1.7) satisfy Hypothesis 2.1. If (E, A) is skew-adjoint, then we have that*

$$(E, A) \equiv \left(\begin{bmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & E_{33} \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & A_{33} \end{bmatrix} \right), \tag{4.5}$$

where

$$E_{33}(t)\dot{x}_3 = A_{33}(t)x_3 + f_3(t) \tag{4.6}$$

is uniquely solvable for every sufficiently smooth f_3 without specifying initial conditions. Furthermore,

$$E_{33}^T = E_{33}, \quad A_{33}^T = -A_{33} - \dot{E}_{33} \tag{4.7}$$

In order to construct a suitable reduced DAE (2.8), we proceed as in the self-adjoint case using the same notation. For the canonical form, we have

$$\tilde{M}_\mu = \left[\begin{array}{ccc|ccc|ccc|c} I_p & 0 & 0 & & & & & & & \\ 0 & -I_q & 0 & & & & & & & \\ 0 & 0 & E_{33} & & & & & & & \\ \hline 0 & 0 & 0 & I_p & 0 & & & & & \\ 0 & 0 & 0 & 0 & -I_q & 0 & & & & \\ 0 & 0 & \dot{E}_{33} - A_{33} & 0 & 0 & E_{33} & & & & \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & I_p & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_q & 0 & 0 \\ 0 & 0 & \ddot{E}_{33} - 2\dot{A}_{33} & 0 & 0 & 2\dot{E}_{33} - A_{33} & 0 & 0 & E_{33} & \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array} \right].$$

Due to the identities, the only possible rank-deficiency is related to the part belonging to the pair (E_{33}, A_{33}). The properties of (4.6) then imply that $d = p + q$ and $a = n - (p + q)$ in Hypothesis 2.1. Furthermore, the left null space of $\tilde{M}_\mu$ is described by

$$\tilde{Z}_2^T = \left[0 \; 0 \; \tilde{Z}_{2,0}^T \middle| 0 \; 0 \; \tilde{Z}_{2,1}^T \middle| 0 \; 0 \; \tilde{Z}_{2,2}^T \middle| \cdots \right].$$

Observing that

$$\tilde{N}_\mu [I_n \; 0 \; \cdots \; 0]^T = \left[\begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & A_{33} \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \dot{A}_{33} \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \ddot{A}_{33} \\ \hline \vdots & \vdots & \vdots \end{array}\right],$$

we get that

$$\hat{A}_2 = \begin{bmatrix} 0 & 0 & I_a \end{bmatrix}$$

for the second part of Hypothesis 2.1, where the identity comes from a special choice of $\tilde{Z}_2^T$. Choosing

$$\tilde{T}_2 = \begin{bmatrix} I_p & 0 \\ 0 & I_q \\ 0 & 0 \end{bmatrix}$$

and $\tilde{Z}_1 = \tilde{T}_2$ yields

$$\tilde{Z}_1^T \tilde{E} \tilde{T}_2 = \begin{bmatrix} I_p & 0 & 0 \\ 0 & I_p & 0 \end{bmatrix} \begin{bmatrix} I_p & 0 & 0 \\ 0 & -I_q & 0 \\ 0 & 0 & E_{33} \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & I_q \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix},$$

which is indeed pointwise nonsingular, thus satisfying the third part of Hypothesis 2.1. In particular, the special choice $\tilde{Z}_1 = \tilde{T}_2$ is possible. According to (2.14) with $P = Q^T$ we can also choose $Z_1 = T_2$ for the original pair such that the reduced DAE inherits some symmetry properties of the original DAE. Note also that we may assume that T_2 possesses pointwise orthonormal columns.

By construction, the matrix function $T_2^T E T_2$ is not only pointwise symmetric but also pointwise nonsingular. We can then apply the results of [12], which guarantee the existence of a smooth matrix function W with

$$W^T T_2^T E T_2 W = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} = S.$$

For convenience, we write again T_2 instead of the transformed T_2W. Completing T_2 to a pointwise nonsingular Q according to (3.2), we get

$$Q^TEQ = \begin{bmatrix} S & \hat{E}_{12} \\ * & * \end{bmatrix}, \quad Q^TAQ - Q^TE\dot{Q} = \begin{bmatrix} J & \hat{A}_{12} \\ * & * \end{bmatrix}.$$

Since skew-adjointness is invariant under congruence, see [1, 15], and S is constant, the matrix function J is pointwise skew-symmetric. With (3.1) the reduced DAE transforms to

$$\begin{aligned} S\dot{x}_1 + \hat{E}_{12}(t)\dot{x}_2 &= J(t)x_1 + \hat{A}_{12}(t)x_2 + T_2(t)^T f(t), \\ 0 &= \hat{A}_{22}(t)x_2 + \hat{f}_2(t), \end{aligned}$$

where $\hat{A}_{22} = \hat{A}_2T_2'$ is pointwise nonsingular. Solving the second equation for x_2, differentiating, and eliminating x_2 and $\dot{x}_2$ from the first equation yields the inherent ODE

$$\dot{x}_1 = S^{-1}J(t)x_1 + \tilde{f}_1(t) \tag{4.8}$$

with a transformed inhomogeneity $\tilde{f}_1$.

Theorem 4.4 *Let (E, A) with $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be sufficiently smooth and let the associated DAE* (1.7) *satisfy Hypothesis* (2.1)*. If (E, A) is skew-adjoint, then Q in* (3.1) *can be chosen from a restricted class of transformations in such a way that the ODE* (1.4) *belonging to the so constructed inherent ODE possesses a generalized orthogonal flow.*

Proof The above construction shows that it is possible to fix an inherent ODE such that the associated ODE (1.4) has a generalized orthogonal flow. It is special in the sense that it works with pointwise orthogonal transformations with the exception of W. □

In the special case that $\mu = 0$, a slightly simplified construction is possible. Here, Hypothesis 2.1 implies that E has constant rank allowing to choose Q in the form (3.2) such that

$$Q^TEQ = \begin{bmatrix} \hat{E}_{11} & 0 \\ 0 & 0 \end{bmatrix}$$

with $\hat{E}_{11} = T_2^TET_2$ pointwise nonsingular. Then, the same modifications of T_2 as before are possible leading to a modified T_2 with $\hat{E}_{11} = S$. With the corresponding modified Q, observing $ET_2' = 0$, we get

$$Q^TEQ = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}, \quad Q^TAQ - Q^TE\dot{Q} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix}.$$

Since congruence transformations conserve skew-adjointness, we have $\hat{A}_{11}^T = -\hat{A}_{11}$, $\hat{A}_{12}^T = -\hat{A}_{21}$, and $\hat{A}_{22}^T = -\hat{A}_{22}$. Moreover, Hypothesis 2.1 with $\mu = 0$ requires that $\hat{A}_{22}$ is pointwise nonsingular. The corresponding reduced DAE, which is here just the original DAE, transforms to

$$\begin{aligned} S\dot{x}_1 &= \hat{A}_{11}(t)x_1 + \hat{A}_{12}(t)x_2 + T_2(t)^T f(t), \\ 0 &= \hat{A}_{12}(t)^T x_1 + \hat{A}_{22}(t)x_2 + T_2'(t)^T f(t). \end{aligned}$$

Solving the second equation for x_2 and eliminating it from the first equation, we again obtain an inherent ODE of the form (4.4), where

$$J = \hat{A}_{11} - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{A}_{12}^T$$

is pointwise skew-symmetric.

Theoretically, all constructions can be performed globally. For a numerical realization one typically uses locally smooth variants as described in Sect. 3. The only exception is the construction of a suitable W, where we are still in need of a locally smooth variant to be used within an integration. One possibility is given in the following, cp. [12].

We start with a reference factorization

$$W_0^T \hat{E}_{11}(t_0) W_0 = S$$

which may be obtained by solving the symmetric eigenvalue problem and then scaling the eigenvalues by congruence to ±1 or by a Cholesky-like factorization for indefinite matrices as given by [2]. We then consider the matrix function

$$W_0^T \hat{E}_{11} W_0 = \begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ \tilde{E}_{21} & \tilde{E}_{22} \end{bmatrix},$$

where $\tilde{E}_{11}^T = \tilde{E}_{11}$, $\tilde{E}_{12}^T = \tilde{E}_{21}$, and $\tilde{E}_{22}^T = \tilde{E}_{22}$. In a sufficiently small neighborhood, the entry $\tilde{E}_{11}$ is close to I_p, the entry $\tilde{E}_{22}$ is close to $-I_q$, and the entry $\tilde{E}_{12}$ is small in norm. In particular, the entry $\tilde{E}_{11}$ is symmetric positive definite allowing for a Cholesky factorization

$$\tilde{E}_{11} = L_{11} L_{11}^T,$$

which is a smooth process. We then get

$$\begin{bmatrix} L_{11}^{-1} & 0 \\ -\tilde{E}_{12}^T \tilde{E}_{11}^{-1} & I_q \end{bmatrix} \begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} \\ \tilde{E}_{12}^T & \tilde{E}_{22} \end{bmatrix} \begin{bmatrix} L_{11}^{-T} & -\tilde{E}_{11}^{-1}\tilde{E}_{12} \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & \tilde{E}_{22} - \tilde{E}_{12}^T \tilde{E}_{11}^{-1} \tilde{E}_{12} \end{bmatrix}.$$

In a sufficiently small neighborhood, the Schur complement $\tilde{E}_{22} - \tilde{E}_{12}^T \tilde{E}_{11}^{-1} \tilde{E}_{12}$ is symmetric negative definite allowing for a Cholesky factorization

$$-(\tilde{E}_{22} - \tilde{E}_{12}^T \tilde{E}_{11}^{-1} \tilde{E}_{12}) = L_{22} L_{22}^T,$$

such that

$$\begin{bmatrix} I_p & 0 \\ 0 & L_{22}^{-1} \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & \tilde{E}_{22} - \tilde{E}_{12}^T \tilde{E}_{11}^{-1} \tilde{E}_{12} \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & L_{22}^{-T} \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} = S.$$

Gathering all transformations gives the locally smooth

$$W = W_0 \begin{bmatrix} L_{11}^{-T} & -\tilde{E}_{11}^{-1} \tilde{E}_{12} \\ 0 & I_q \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & L_{22}^{-T} \end{bmatrix}$$

and all steps can be executed numerically in a smooth way using automatic differentiation.

5 Numerical experiments

The presented numerical method has been implemented using automatic differentiation in order to be able to evaluate all needed derivatives and Jacobians. For the determination of $\langle Q, \dot{Q} \rangle$ on the current interval $[t_0, t_0 + h]$ one can choose between the following possibilities.

INHERENT	$Q(t) = Q_0$
SPIN_STABILIZED	$Q(t) = Q_0 + (t - t_0)\dot{Q}_0$
ROTATED	$Q = [\, T_2 \;\; T_2' \,], \;\; \hat{E}_1 T_2' = 0$
SELF_ADJOINT	Q as described in Subsection 4.1
SKEW_ADJOINT	Q as described in Subsection 4.2
PRESCRIBED	Q by user-provided routine

In all cases except for the last one, one can choose between the general approach, which includes transformation to a reduced DAE, and the simplified approach assuming that no such transformation is necessary. Schemes based on the direct discretization of (2.5) are labelled as DIRECT. As numerical integration methods we use the following discretization methods, see e.g. [10, 13].

GAUSS-LOBATTO	collocation methods for DAEs based on Gauß nodes for the differential part and Lobatto nodes for the algebraic part, see [17]
RADAU	collocation methods for DAEs based on Radau nodes the simplest of which is the implicit Euler method
DORMAND-PRINCE	Runge-Kutta-Fehlberg methods for ODEs, see [9]
GAUSS	collocation methods for ODEs based on Gauß nodes

Experiment 5.1 The linear DAE

$$\begin{bmatrix} \delta - 1 & \delta t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -\eta(\delta - 1) & -\eta\delta t \\ \delta - 1 & \delta t - 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix},$$

cp. [18], with real parameters η and $\delta \neq 1$ is constructed in such a way that direct discretization by the implicit Euler method corresponds to the discretization of an inherent ODE by the explicit Euler method. Setting $\delta = -10^5$, $\eta = 0$ yields a stiff inherent ODE and we expect stability problems when working directly with the implicit Euler method. For our numerical experiments we have chosen f_1, f_2 and the initial condition so that the solution is given by $x_1(t) = x_2(t) = \exp(-t)$. Integration interval was $[0, 1]$ and tolerance was 10^{-5}. The following table gives the cpu times and the number of integration steps for the various versions of the implicit Euler method.

version	cpu time	steps
DIRECT	10.31	97840
INHERENT	0.73	10
SPIN_STABILIZED	0.60	10
ROTATED	0.67	10

The stabilizing effect of discretizing an inherent ODE is obvious. The three different versions in the choice of the inherent ODE do not differ significantly.

Experiment 5.2 A mathematical model of a pendulum is given by the DAE

$$\begin{aligned} \dot{x}_3 &= x_1, \\ \dot{x}_4 &= x_2, \\ -\dot{x}_1 &= 2x_3x_5, \\ -\dot{x}_2 &= 1 + 2x_4x_5, \\ 0 &= x_3^2 + x_4^2 - 1, \end{aligned}$$

which is known to satisfy Hypothesis 2.1 with $\mu = 2, a = 3$, and $d = 2$. The equations and unknowns are ordered in such a way that

$$F_{\dot{x}}(t, x, \dot{x}) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad F_x(t, x, \dot{x}) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2x_5 & 0 & 2x_3 \\ 0 & 0 & 0 & 2x_5 & 2x_4 \\ 0 & 0 & 2x_3 & 2x_4 & 0 \end{bmatrix}.$$

Hence, $(F_{\dot{x}}, F_x)$ is self-adjoint for all arguments. The constructions of Sect. 4, however, are only valid for linear DAEs and therefore not applicable. The only valid use of an inherent ODE as presented here is by the versions INHERENT and PRESCRIBED, since in the nonlinear case the Jacobians do not only depend on t. The following table shows the performance of various discretization schemes when integrating over the interval [0, 10] with stepsize control starting with $x(0) = (0, 0, 1, 0, 0)^T$ and using a tolerance of 10^{-5}.

method	version	stages	order	cpu time	steps
GAUSS-LOBATTO	DIRECT	2-3	4	0.93	55
RADAU	DIRECT	4	7	0.99	28
DORMAND-PRINCE	INHERENT	7	4	1.33	47
DORMAND-PRINCE	INHERENT	13	7	1.29	28
GAUSS	INHERENT	2	4	11.76	55
RADAU	INHERENT	4	7	12.92	34

In particular, we observe that we are able to solve the given problem by explicit schemes for the chosen inherent ODE with nearly the same efficiency as the standard direct methods. As in the standard ODE case, implicit methods applied to the inherent ODE are for this problem outperformed by explicit methods since the inherent ODE is non-stiff.

In the following experiments we measure the geometric error in the flow Φ with respect to a quadratic Lie group (1.5), i. e. the deviation of the flow Φ from being pointwise in the corresponding Lie group, by $\|\Phi(t)^T X \Phi(t) - X\|$, where we use the matrix norm defined by $\|\Delta\| = \max_{i,j=1,\dots,n} |\Delta_{ij}|$ for a matrix $\Delta = [\Delta_{ij}] \in \mathbb{R}^{n,n}$.

Experiment 5.3 The self-adjoint DAE $E(t)\dot{x} = A(t)x$ given by

$$E = Q^T \hat{E} Q, \quad A = Q^T \hat{A} Q - Q^T \hat{E} \dot{Q},$$

where

$$\hat{E} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & s & 0 \\ s & 1 & s \\ 0 & s & 1 \end{bmatrix},$$

with $s(t) = \frac{1}{2} \sin \omega t$, $\omega = 1$, possesses a symplectic flow with respect to the first two components of the transformed unknown $\hat{x} = Qx$.

The following table shows the performance and the maximal geometric error in the flow for various discretization schemes when integrating over the interval $[0, 200\pi]$ using $1,000$ equidistant steps. We used the simplified approach due to $\mu = 0$.

method	version	stages	order	cpu time	error
GAUSS-LOBATTO	DIRECT	2-3	4	1.44	1.380e-02
DORMAND-PRINCE	INHERENT	7	4	5.36	2.468e-01
GAUSS	ROTATED	2	4	23.68	7.281e-04
GAUSS	SELF_ADJOINT	2	4	24.88	1.927e-11

In particular, we see that the last method given in the table shows a geometric error only governed by roundoff effects, thus constituting a geometric integrator for this class of problems.

Experiment 5.4 The skew-adjoint DAE $E(t)\dot{x} = A(t)x$ given by

$$E = Q^T \hat{E} Q, \quad A = Q^T \hat{A} Q - Q^T \hat{E} \dot{Q},$$

where

$$\hat{E} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & s & 0 & 0 \\ s & 1 & s & 0 \\ 0 & s & 1 & s \\ 0 & 0 & s & 1 \end{bmatrix},$$

with $s(t) = \frac{1}{2} \sin \omega t, \ \omega = 1$, possesses an orthogonal flow with respect to the first two components of the transformed unknown $\hat{x} = Qx$.

The following table shows the performance and the maximal geometric error in the flow for various discretization schemes when integrating over the interval $[0, 200\pi]$ using $1,000$ equidistant steps. We used the simplified approach due to $\mu = 0$.

method	version	stages	order	cpu time	error
GAUSS-LOBATTO	DIRECT	$2-3$	4	2.05	1.226e-01
DORMAND-PRINCE	INHERENT	7	4	12.52	1.965e-02
GAUSS	ROTATED	2	4	74.85	9.363e+00
GAUSS	SKEW_ADJOINT	2	4	80.64	3.814e-12

In particular, we see that the last method given in the table shows a geometric error only governed by roundoff effects, thus constituting a geometric integrator for this class of problems.

Experiment 5.5 The skew-adjoint DAE $E(t)\dot{x} = A(t)x$ given by

$$E = Q^T \hat{E} Q, \quad A = Q^T \hat{A} Q - Q^T \hat{E} \dot{Q},$$

where

$$\hat{E} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{bmatrix}, \quad Q = \begin{bmatrix} 1 & s & 0 & 0 & 0 \\ s & 1 & s & 0 & 0 \\ 0 & s & 1 & s & 0 \\ 0 & 0 & s & 1 & s \\ 0 & 0 & 0 & s & 1 \end{bmatrix},$$

with $s(t) = \frac{1}{2} \sin \omega t, \ \omega = 1$, possesses a generalized orthogonal flow in O(2, 1) with respect to the first three components of the transformed unknown $\hat{x} = Qx$.

The following table shows the performance and the maximal geometric error in the flow for various discretization schemes when integrating over the interval $[0, 200\pi]$

using $1,000$ equidistant steps. We used the simplified approach due to $\mu = 0$.

method	version	stages	order	cpu time	error
GAUSS-LOBATTO	DIRECT	2-3	4	3.33	4.548e-01
DORMAND-PRINCE	INHERENT	7	4	32.40	8.957e-01
GAUSS	ROTATED	2	4	226.55	6.912e-01
GAUSS	SKEW_ADJOINT	2	4	288.55	1.096e-11

In particular, we see that again the last method given in the table shows a geometric error only governed by roundoff effects, thus constituting a geometric integrator for this class of problems.

6 Conclusions

We have presented discretization methods for DAEs that are based on the integration of an inherent ODE which is extracted from the derivative array equations associated with the given DAE utilizing automatic differentiation. We have shown that for this inherent ODE we can use classical discretization schemes for the numerical integration of ODEs that cannot be used for DAEs directly. For self-adjoint and skew-adjoint linear time-varying DAEs we have shown that the inherent ODE can be constructed in such a way that it inherits these symmetry properties of the given DAE and thus also the geometric properties of its flow. We then have exploited this property to construct geometric integration schemes with a numerical flow that preserves these geometric properties.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding was provided by Deutsche Forschungsgemeinschaft (CRC 910 Project A2), Mathematisches Forschungsinstitut Oberwolfach (Research In Pairs)

Declarations

Conflicts of interest The authors have no conflicts of interests to declare that are relevant to the content of this article.

References

1. Beattie, C., Mehrmann, V., Xu, H., Zwart, H.: Port-Hamiltonian descriptor systems. Math. Control Signals Syst. **30**, 1–27 (2018)

2. Bunch, J.R., Kaufman, L.: On smooth decompositions of matrices. Math. Comput. **31**, 163–179 (1977)
3. Campbell, S.L.: A general form for solvable linear time varying singular systems of differential equations. SIAM J. Math. Anal. **18**, 1101–1115 (1987)
4. Campbell, S.L., Kunkel, P.: Completions of nonlinear DAE flows based on index reduction techniques and their stabilization. J. Comput. Appl. Math. **233**, 1021–1034 (2009)
5. Deuflhard, P.: Newton Methods for Nonlinear Problems Affine Invariance and Adaptive Algorithms. Springer, Berlin, Germany (2004)
6. Griewank, A.: On Automatic Differentiation. In: Mathematical Programming: Recent Development and Applications, pp. 83–108. Kluwer Academic Publishers, Stuttgart, Germany (1989)
7. Hairer, E., Lubich, C., Roche, M.: Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations. BIT **28**, 678–700 (1988)
8. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations. Springer-Verlag, Berlin, Germany (2002)
9. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems, 1st edn. Springer, Berlin (1987)
10. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems, 2nd edn. Springer, Berlin, Germany (1996)
11. Kunkel, P.: Differential-Algebraic Equations: Theory and Simulation. In: P. Benner, M. Bollhöfer, D. Kressner, C. Mehl, T. Stykel (eds.) Numerical Algebra, Matrix Theory, Differential-Algebraic Equations and Control Theory. Springer (2015)
12. Kunkel, P.: A smooth version of Sylvester's law of inertia and its numerical realization. Electr. Trans. Num. Anal. **36**, 542–560 (2020)
13. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations Analysis and Numerical Solution. EMS Publishing House, Switzerland (2006)
14. Kunkel, P., Mehrmann, V.: Stability properties of differential-algebraic equations and spin-stabilized discretizations. Electron. Trans. Numer. Anal. **26**, 385–420 (2007)
15. Kunkel, P., Mehrmann, V.: Local and global canonical forms for differential-algebraic equations with symmetries. Vietnam J. Mathemat. **51**, 177–198 (2023)
16. Kunkel, P., Mehrmann, V., Scholz, L.: Self-adjoint differential-algebraic equations. Math. Control Signals Syst. **26**, 47–76 (2014). https://doi.org/10.1007/s00498-013-0109-3
17. Kunkel, P., Mehrmann, V., Stöver, R.: Symmetric collocation for unstructured nonlinear differential-algebraic equations of arbitrary index. Numer. Math. **98**, 277–304 (2004)
18. März, R., Rodriguez-Santiesteban, A.R.: Analyzing the stability behaviour of solutions and their approximations in case of index-2 differential-algebraic systems. Math. Comp. **71**, 605–632 (2001)