A Lie Algebra Approach to Lie Group Time Integration of Constrained Systems

Martin Arnold, Alberto Cardona and Olivier Brüls

Abstract Lie group integrators preserve by construction the Lie group structure of a nonlinear configuration space. In multibody dynamics, they support a representation of (large) rotations in a Lie group setting that is free of singularities. The resulting equations of motion are differential equations on a manifold with tangent spaces being parametrized by the corresponding Lie algebra. In the present paper, we discuss the time discretization of these equations of motion by a generalized-α Lie group integrator for constrained systems and show how to exploit in this context the linear structure of the Lie algebra. This linear structure allows a very natural definition of the generalized-α Lie group integrator, an efficient practical implementation and a very detailed error analysis. Furthermore, the Lie algebra approach may be combined with analytical transformations that help to avoid an undesired order reduction phenomenon in generalized-α time integration. After a tutorial-like step-by-step introduction to the generalized-α Lie group integrator, we investigate its convergence behaviour and develop a novel initialization scheme to achieve second-order accuracy in the application to constrained systems. The theoretical results are illustrated by a comprehensive set of numerical tests for two Lie group formulations of a rotating heavy top.

1 Introduction

Structure-preserving integrators overcome limitations of classical time integration methods from the fields of ordinary differential equations (ODEs) and differential-algebraic equations (DAEs). They are known for their favourable nonlinear stability

M. Arnold (✉)
Martin Luther University Halle-Wittenberg, Halle (Saale), Germany
e-mail: martin.arnold@mathematik.uni-halle.de

A. Cardona
Universidad Nacional Litoral - CONICET, Santa Fe, Argentina

O. Brüls
University of Liège, Liège, Belgium

P. Betsch (ed.), *Structure-preserving Integrators in Nonlinear Structural Dynamics and Flexible Multibody Dynamics*, CISM International Centre for Mechanical Sciences 565, DOI 10.1007/978-3-319-31879-0_3

properties for the long-term integration of conservative systems, see, e.g., (Hairer et al. 2006).

The focus of the present paper is slightly different since we consider a class of time integration methods that is tailored to flexible multibody system models with dissipative terms resulting, e.g., from friction forces or control structures. The methods are applied to constrained systems with a nonlinear configuration space with Lie group structure. They preserve this structural property of the equations of motion in the sense that the numerical solution remains by construction in this nonlinear configuration space.

The Lie group setting allows a representation of (large) rotations that is globally free of singularities. Local parametrizations could be used to transform the system in each time step in a linear configuration space such that classical time integration methods could be used. As an alternative to such local parametrizations, Simo and Vu-Quoc (1988) proposed a Newmark-type method that is directly based on the equations of motion in a nonlinear configuration space with Lie group structure.

Starting with the work of Crouch and Grossman (1993) and Munthe-Kaas (1995, 1998), the time discretization of ordinary differential equations on Lie groups has found much interest in the numerical analysis community. This work was summarized in the comprehensive survey paper by Iserles et al. (2000). In that time, the application of Lie group time integration methods to multibody system models was studied, e.g., by Bottasso and Borri (1998) and Celledoni and Owren (2003).

In 2010, the combination of Lie group time integration with the time discretization by generalized-α methods was proposed, see (Brüls and Cardona 2010). Generalized-α methods are Newmark type methods that go back to the work of Chung and Hulbert (1993). They may be considered as a generalization of Hilber–Hughes–Taylor (HHT) methods, see (Hilber et al. 1977), and have found new interest in industrial multibody system simulation since they avoid the very strong damping of high-frequency solution components that is characteristic of other integrators in this field, see, e.g., (Negrut et al. 2005; Lunk and Simeon 2006; Jay and Negrut 2007, 2008; Arnold and Brüls 2007).

Cardona and Géradin (1994) investigated systematically the stability and convergence of HHT methods for constrained systems. This analysis may be extended to generalized-α methods, see Géradin and Cardona (2001, Sect. 10.5), and shows a risk of order reduction and large transient errors in the Lagrange multipliers and constrained forces. Numerical test results for the generalized-α Lie group integrator illustrate that this undesired numerical effect is strongly related to the specific Lie group formulation of the equations of motion, see (Brüls et al. 2011).

Therefore, the error analysis for the Lie group integrator has to consider the global errors in long-term integration as well as the transient behaviour of the numerical solution. In a series of papers, we developed a strategy for defining, implementing and analysing the Lie group integrator that is based on the observation that the increments of the configuration variables in each time step are parametrized by elements of the Lie algebra, i.e., by elements of a *linear* space, see (Arnold et al. 2011b, 2014, 2015) and (Brüls et al. 2011, 2012). In the present paper, we follow this *Lie algebra approach* and consider local and global discretization errors of the Lie group integrator as elements of the corresponding Lie algebra.

We introduce the Lie group setting in a tutorial like style and show how to discretize the equations of motion by a generalized-α Lie group integrator. There is a specific focus on practical aspects like corrector iteration and initialization of the integrator. In a comprehensive numerical test series, we consider different Lie group formulations of a heavy top benchmark problem. For the convergence analysis, we follow to a large extent the presentation in the recently published paper (Arnold et al. 2015).

The remaining part of the paper is organized as follows: Basic aspects of Lie group theory in the context of multibody dynamics and the equations of motion of constrained systems are introduced in Sect. 2. Furthermore, we discuss two different Lie group formulations of a rotating heavy top that will be used as benchmark problem throughout the paper.

In Sect. 3, we consider the generalized-α Lie group DAE integrator and study its asymptotic behaviour for time step sizes $h \to 0$. Classical results of Hilber and Hughes (1978) on "overshooting" of Newmark type methods in the application to linear problems with high-frequency solutions are shown to result in an order reduction phenomenon for the constrained case, see (Cardona and Géradin 1994). In Sect. 3.3, the large first-order error terms are illustrated by numerical tests for the heavy top benchmark problem. They may be reduced drastically by index reduction and a modification of the generalized-α Lie group integrator that is based on the so-called stabilized index-2 formulation of the equations of motion, see Sect. 3.4. Implementation aspects and some technical details are discussed in Sects. 3.5 and 3.6.

For the convergence analysis, we discuss in Sect. 4.1 a one-step error recursion of generalized-α methods for constrained systems. The coupled error propagation in differential and algebraic solution components may be studied extending the convergence analysis of ODE one-step methods to the Lie group DAE case, see Sect. 4.2. The convergence theorem for the generalized-α Lie group DAE integrators is given in Sect. 4.3. It provides the basis for an optimal initialization using perturbed starting values that guarantee second-order convergence in all solution components such that order reduction may be avoided.

2 Constrained Systems in a Configuration Space with Lie Group Structure

The main interest of this paper is in time integration methods for constrained mechanical systems that have a configuration space with Lie group structure. In the present section, we introduce this Lie group setting by studying the configuration space of a rigid body (Sect. 2.1). Lie groups are differentiable manifolds that are in a very natural way parametrized locally by elements of the corresponding Lie algebra (Sect. 2.2).

Lie groups may be used to represent large rotations in $\mathbb{R}^3$ without singularities. They are part of the mathematical framework for a generic finite element approach to flexible multibody dynamics that has been applied successfully for more than two decades (Géradin and Cardona 1989, 2001). In Sect. 2.3, we consider constrained systems and discuss the general structure of the equations of motion. As a typical example, two different Lie group formulations of a heavy top benchmark problem are introduced in Sect. 2.4. Finally, some technical details of the Lie group setting are discussed in Sect. 2.5.

2.1 The Configuration Space of a Rigid Body in $\mathbb{R}^3$

The position of a rigid body in an inertial frame is represented by a vector $\mathbf{x} \in \mathbb{R}^3$, i.e., by an element of a linear space. There are three additional degrees of freedom that describe the orientation of this rigid body but these degrees of freedom may not be represented globally by elements of a three-dimensional linear space. In engineering, small deviations from a nominal state are often characterized by three angles of rotation like Euler angles or Bryant angles (Géradin and Cardona 2001, Sect. 4.8) that suffer, however, from singularities in the case of large rotations.

Alternative representations that are free of singularities are provided, e.g., by Euler parameters that are also known as quaternions (Betsch and Siebert 2009 and Géradin and Cardona 2001, Sect. 4.5) or by the rotation matrix

$$\mathbf{R} \in \mathrm{SO}(3) := \{\, \mathbf{R} \in \mathbb{R}^{3\times 3} \,:\, \mathbf{R}^\top \mathbf{R} = \mathbf{I}_3\,,\ \det \mathbf{R} = +1 \,\}\,.$$

The set SO(3) is a three-dimensional differentiable manifold in $\mathbb{R}^{3\times 3}$ and may be combined in two alternative ways with the linear space $\mathbb{R}^3$ to describe the configuration of the rigid body by an element $q := (\mathbf{R}, \mathbf{x})$ of a six-dimensional group G (Brüls et al. 2011; Müller and Terze 2014a): In the direct product $G = \mathrm{SO}(3) \times \mathbb{R}^3$, the group operation $\circ$ is defined by

$$(\mathbf{R}_a, \mathbf{x}_a) \circ (\mathbf{R}_b, \mathbf{x}_b) = (\mathbf{R}_a \mathbf{R}_b, \mathbf{x}_a + \mathbf{x}_b)$$

and results in kinematic relations

$$\dot{\mathbf{R}} = \mathbf{R}\widetilde{\mathbf{\Omega}}\,, \qquad \dot{\mathbf{x}} = \mathbf{u} \tag{1}$$

with $\mathbf{u} \in \mathbb{R}^3$ denoting the translation velocity in the inertial frame and a skew symmetric matrix

$$\widetilde{\mathbf{\Omega}} := \begin{pmatrix} 0 & -\Omega_3 & \Omega_2 \\ \Omega_3 & 0 & -\Omega_1 \\ -\Omega_2 & \Omega_1 & 0 \end{pmatrix} \in \mathbb{R}^{3\times 3} \tag{2}$$

that represents the angular velocity $\mathbf{\Omega} = (\,\Omega_1,\ \Omega_2,\ \Omega_3\,)^\top \in \mathbb{R}^3$. The semi-direct product $G = \mathrm{SO}(3) \ltimes \mathbb{R}^3$ is known as the *special Euclidean group* SE(3) with the group operation

$$(\mathbf{R}_a, \mathbf{x}_a) \circ (\mathbf{R}_b, \mathbf{x}_b) = (\mathbf{R}_a \mathbf{R}_b, \mathbf{R}_a \mathbf{x}_b + \mathbf{x}_a)\,,$$

kinematic relations

$$\dot{\mathbf{R}} = \mathbf{R}\widetilde{\boldsymbol{\Omega}}\,, \quad \dot{\mathbf{x}} = \mathbf{R}\mathbf{U} \tag{3}$$

and $\mathbf{U} \in \mathbb{R}^3$ denoting the translation velocity in the body-attached frame.

For group elements $q = (\mathbf{R}, \mathbf{x})$, the group operations in $\mathrm{SO}(3) \times \mathbb{R}^3$ and in $\mathrm{SE}(3)$ are equivalent to the matrix multiplication of non-singular block-structured matrices in $\mathbb{R}^{7\times 7}$ and in $\mathbb{R}^{4\times 4}$, respectively, that are defined by

$$\mathrm{SO}(3) \times \mathbb{R}^3 \,:\, \begin{pmatrix} \mathbf{R} & \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 3} & \mathbf{I}_3 & \mathbf{x} \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} & 1 \end{pmatrix}, \quad \mathrm{SE}(3) \,:\, \begin{pmatrix} \mathbf{R} & \mathbf{x} \\ \mathbf{0}_{1\times 3} & 1 \end{pmatrix}. \tag{4}$$

Therefore, the groups $\mathrm{SO}(3) \times \mathbb{R}^3$ and $\mathrm{SE}(3)$ as well as the group $\mathrm{SO}(3)$ of all rotation matrices $\mathbf{R}$ are isomorphic to a subset of a general linear group $\mathrm{GL}(r) = \{\, \mathbf{A} \in \mathbb{R}^{r\times r} \,:\, \det \mathbf{A} \neq 0 \,\}$ of suitable degree $r > 0$. The structure of the block matrices in (4) and the orthogonality condition $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_3$ imply that the groups $\mathrm{SO}(3) \times \mathbb{R}^3$, $\mathrm{SE}(3)$ and $\mathrm{SO}(3)$ are isomorphic to differentiable manifolds in $\mathrm{GL}(7)$, $\mathrm{GL}(4)$ and $\mathrm{GL}(3)$, respectively.

2.2 *Differential Equations on Manifolds: Matrix Lie Groups*

A group G with group operation $\circ$ and neutral element $e \in G$ is called a *Lie group* if G is a differentiable manifold and the group operation $\circ \,:\, G \times G \to G$ as well as the map $q \mapsto q^{-1}$ are differentiable ($q \circ q^{-1} = e$). Lie groups that are subgroups of $\mathrm{GL}(r)$ for some $r > 0$ are called *matrix Lie groups* if the group operation $\circ$ is given by the matrix multiplication. For a compact introduction to analytical and numerical aspects of such matrix Lie groups, the interested reader is referred to (Hairer et al. 2006, Sect. IV.6).

It is a trivial observation that a continuously differentiable function $q(t)$ with $q(t_0) \in G$ will remain in a Lie group G if and only if its time derivative $\dot{q}(t)$ is in the tangent space $T_q G$ at the point $q = q(t)$: $\dot{q}(t) \in T_{q(t)} G$, $(t \geq t_0)$. The tangent space at the neutral element e defines the *Lie algebra* $\mathfrak{g} := T_e G$. As a linear space, it is isomorphic to a finite dimensional linear space $\mathbb{R}^k$ with an invertible linear mapping $\widetilde{(\bullet)} \,:\, \mathbb{R}^k \to \mathfrak{g}$, $\mathbf{v} \mapsto \widetilde{\mathbf{v}}$.

The group structure of G makes it possible to represent the elements of $T_q G$ at *any* element $q \in G$ by the elements $\widetilde{\mathbf{v}}$ of the Lie algebra: The left translation

$$L_q \,:\, G \to G\,, \quad y \mapsto L_q(y) := q \circ y$$

defines a bijection in G. Its derivative $DL_q(y)$ at $y = e$ represents the corresponding bijection between the tangent spaces $\mathfrak{g} := T_e G$ and $T_q G$, i.e.,

$$T_q G = \{\, DL_q(e) \cdot \widetilde{\mathbf{v}} \,:\, \widetilde{\mathbf{v}} \in \mathfrak{g} \,\} = \{\, DL_q(e) \cdot \widetilde{\mathbf{v}} \,:\, \mathbf{v} \in \mathbb{R}^k \,\}\,. \tag{5}$$

With these notations, kinematic relations like (1) and (3) may be summarized in compact form:

$$\dot{q}(t) = DL_{q(t)}(e) \cdot \widetilde{\mathbf{v}}(t) \tag{6}$$

with a velocity vector $\mathbf{v}(t) \in \mathbb{R}^k$. In (6), the left translation L_q as well as the tilde operator $\widetilde{(\bullet)}$ depend on the specific Lie group setting.

For constant velocity $\mathbf{v}$, the kinematic relation (6) yields locally

$$q(t) = q(t_0) \circ \exp\big((t - t_0)\widetilde{\mathbf{v}}\big) \in G \tag{7}$$

with the exponential map $\exp : \mathfrak{g} \to G$. For matrix Lie groups, this exponential map is given by

$$\exp(\widetilde{\mathbf{v}}) = \sum_{i=0}^{\infty} \frac{1}{i!}\, \widetilde{\mathbf{v}}^i\,. \tag{8}$$

It is a local diffeomorphism, i.e., for any $q_a \in G$ there are neighbourhoods $U_{q_a} \subset G$ and $V_{\widetilde{\mathbf{0}}} \subset \mathfrak{g}$ such that any $q \in U_{q_a}$ may be expressed by

$$q = q_a \circ \exp(\widetilde{\boldsymbol{\Delta}}_q) \tag{9}$$

with a uniquely defined element $\widetilde{\boldsymbol{\Delta}}_q \in V_{\widetilde{\mathbf{0}}}$.

Example 2.1 (a) Using the block matrix representation (4), the groups $\mathrm{SO}(3) \times \mathbb{R}^3$, SE(3) and SO(3) are seen to be matrix Lie groups. The Lie algebra corresponding to Lie group $G = \mathrm{SO}(3)$ is given by the set

$$\mathfrak{so}(3) := \{\, \mathbf{A} \in \mathbb{R}^{3\times 3} \,:\, \mathbf{A} + \mathbf{A}^\top = \mathbf{0} \,\}$$

of all skew symmetric matrices in $\mathbb{R}^{3\times 3}$. As a linear space, this Lie algebra is isomorphic to $\mathbb{R}^3$ with the tilde operator being defined in (2). In SO(3), the exponential map (8) may be evaluated very efficiently by Rodrigues' formula

$$\exp_{\mathrm{SO}(3)}(\widetilde{\boldsymbol{\Omega}}) = \mathbf{I}_3 + \frac{\sin \Phi}{\Phi}\, \widetilde{\boldsymbol{\Omega}} + \frac{1 - \cos \Phi}{\Phi^2}\, \widetilde{\boldsymbol{\Omega}}^2 \tag{10}$$

with $\Phi := \|\boldsymbol{\Omega}\|_2$ since powers $\widetilde{\boldsymbol{\Omega}}^i$ with $i \geq 3$ may be expressed in terms of $\mathbf{I}_3$, $\widetilde{\boldsymbol{\Omega}}$ and $\widetilde{\boldsymbol{\Omega}}^2$ because each matrix $\widetilde{\boldsymbol{\Omega}} \in \mathbb{R}^{3\times 3}$ is a zero of its characteristic polynomial $\chi_\mu(\widetilde{\boldsymbol{\Omega}}) = \det(\mu\mathbf{I}_3 - \widetilde{\boldsymbol{\Omega}}) = \mu^3 + \|\boldsymbol{\Omega}\|_2^2\, \mu = \mu^3 + \Phi^2\mu$, i.e., $\widetilde{\boldsymbol{\Omega}}^3 = -\Phi^2\, \widetilde{\boldsymbol{\Omega}}$ (Cayley-Hamilton theorem).

According to (1), (3) and (6), the Lie algebras of SO(3) × $\mathbb{R}^3$ and SE(3) are parametrized by vectors $\mathbf{v} = (\boldsymbol{\Omega}^\top, \mathbf{u}^\top)^\top$ and $\mathbf{v} = (\boldsymbol{\Omega}^\top, \mathbf{U}^\top)^\top$, respectively. In block matrix form, they are represented by (Brüls et al. 2011)

$$\mathfrak{so}(3) \times \mathbb{R}^3 \, : \, \widetilde{\mathbf{v}} = \begin{pmatrix} \widetilde{\boldsymbol{\Omega}} & \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 3} & \mathbf{u} \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} & 0 \end{pmatrix}, \quad \mathfrak{se}(3) \, : \, \widetilde{\mathbf{v}} = \begin{pmatrix} \widetilde{\boldsymbol{\Omega}} & \mathbf{U} \\ \mathbf{0}_{1\times 3} & 0 \end{pmatrix}$$

with exponential maps

$$\exp_{\mathrm{SO}(3)\times\mathbb{R}^3}(\widetilde{\mathbf{v}}) = \begin{pmatrix} \exp_{\mathrm{SO}(3)}(\widetilde{\boldsymbol{\Omega}}) & \mathbf{0}_{3\times 3} & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{3\times 3} & \mathbf{I}_3 & \mathbf{u} \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} & 1 \end{pmatrix}, \tag{11a}$$

$$\exp_{\mathrm{SE}(3)}(\widetilde{\mathbf{v}}) = \begin{pmatrix} \exp_{\mathrm{SO}(3)}(\widetilde{\boldsymbol{\Omega}}) & \mathbf{T}^\top_{\mathrm{SO}(3)}(\boldsymbol{\Omega})\,\mathbf{U} \\ \mathbf{0}_{1\times 3} & 1 \end{pmatrix} \tag{11b}$$

and the so-called tangent operator $\mathbf{T}_{\mathrm{SO}(3)} \, : \, \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$, see (33), that will be discussed in more detail in Remark 2.8(b) below.

(b) The linear space $\mathbb{R}^k$ with vector addition $+$ as group operation $\circ$ is a trivial example of a matrix Lie group since $\mathbf{x} \in \mathbb{R}^k$ may be identified with the non-singular 2 × 2 block matrix

$$\begin{pmatrix} \mathbf{I}_k & \mathbf{x} \\ \mathbf{0}_{1\times k} & 1 \end{pmatrix} \in \mathrm{GL}(k+1)\,. \tag{12}$$

Substituting vector $\mathbf{x}$ by $\mathbf{u} \in \mathbb{R}^k$ and the main diagonal blocks by $\mathbf{0}_{k\times k}$ and by 0, respectively, we get the block matrix representation of the corresponding Lie algebra that is parametrized by $\mathbf{u}$:

$$\widetilde{\mathbf{u}} = \begin{pmatrix} \mathbf{0}_{k\times k} & \mathbf{u} \\ \mathbf{0}_{1\times k} & 0 \end{pmatrix}, \quad \exp_{\mathbb{R}^k}(\widetilde{\mathbf{u}}) = \begin{pmatrix} \mathbf{I}_k & \mathbf{u} \\ \mathbf{0}_{1\times k} & 1 \end{pmatrix}. \tag{13}$$

Alternatively, the exponential map may be expressed directly in terms of $\mathbf{u} \in \mathbb{R}^k$ using $\exp_{\mathbb{R}^k} = \mathrm{id}_{\mathbb{R}^k}$, i.e., $\mathbf{x} \circ \exp_{\mathbb{R}^k}(\widetilde{\mathbf{u}}) = \mathbf{x} + \mathbf{u}$.

(c) The block matrix representation of $(\mathbf{R}, \mathbf{x})_{\mathrm{SO}(3)\times\mathbb{R}^3}$ in (4) is block-diagonal with diagonal blocks for $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{x} \in \mathbb{R}^3$, see (12). The same block-diagonal structure is observed for the elements of the corresponding Lie algebra $\mathfrak{so}(3) \times \mathbb{R}^3$, for the tilde operator and for $\exp_{\mathrm{SO}(3)\times\mathbb{R}^3}$, see (11a) and (13). It is typical for direct products of Lie groups and may be used as well for Lie groups $G^N = G \times G \times \cdots \times G$ that are direct products of $N \geq 2$ factors G with $G = \mathrm{SO}(3) \times \mathbb{R}^3$ or $G = \mathrm{SE}(3)$. In particular, we have

$$\exp_{G^N}\big((\widetilde{\mathbf{v}}_1, \widetilde{\mathbf{v}}_2, \, \ldots \, , \widetilde{\mathbf{v}}_N)\big) = \mathrm{blockdiag}_{1\leq i\leq N} \exp_G(\widetilde{\mathbf{v}}_i)\,.$$

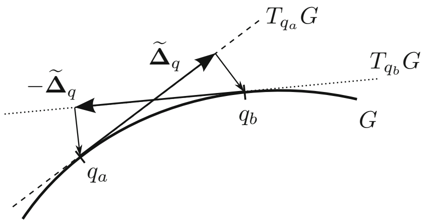

$T_{q_a}G$
$\widetilde{\mathbf{\Delta}}_q$
$T_{q_b}G$
$-\widetilde{\mathbf{\Delta}}_q$
q_b
G
q_a

Fig. 1 Interpolation in Lie groups: $q_b = q_a \circ \exp(\widetilde{\mathbf{\Delta}}_q)$

Hence, the exponential map $\exp_{G^N} : \mathfrak{g}^N \to G^N$ in the direct product G^N may be evaluated as efficiently as the one in its factors G, see (10) and (11). In flexible multibody dynamics, the configuration spaces $(\mathrm{SO}(3) \times \mathbb{R}^3)^N$ and $(\mathrm{SE}(3))^N$ are of special interest since they allow to represent the configuration of an articulated system of rigid and flexible bodies in the nonlinear finite element method by $N \geq 1$ pairs of absolute nodal translation and rotation variables, see (Brüls et al. 2012; Géradin and Cardona 2001).

Remark 2.2 The parametrization (9) offers a generic way to interpolate between q_a and any point q_b in a sufficiently small neighbourhood $U_{q_a} \subset G$, see Fig. 1: If $q_b = q_a \circ \exp(\widetilde{\mathbf{\Delta}}_q)$ with a vector $\mathbf{\Delta}_q \in \mathbb{R}^k$ of sufficiently small norm $\| \mathbf{\Delta}_q \|$, then $\exp(\vartheta \widetilde{\mathbf{\Delta}}_q)$ is well defined for any $\vartheta \in [0, 1]$ and $q_a, q_b \in G$ are connected by the path

$$\{ \, q(\vartheta; q_a, \mathbf{\Delta}_q) = q_a \circ \exp(\vartheta \widetilde{\mathbf{\Delta}}_q) \; : \; \vartheta \in [0, 1] \, \} \subset G \, .$$

Because of $q_a = q_b \circ \exp(-\widetilde{\mathbf{\Delta}}_q)$ the parametrization of this path by $\vartheta \in [0, 1]$ is symmetric in the sense that $q(\vartheta; q_a, \mathbf{\Delta}_q) = q(1 - \vartheta; q_b, -\mathbf{\Delta}_q)$. This expression is the Lie group equivalent to the identity $\mathbf{q}_a + \vartheta \mathbf{\Delta}_q = \mathbf{q}_b - (1 - \vartheta) \mathbf{\Delta}_q$ that is trivially satisfied for a path that interpolates two points $\mathbf{q}_a, \mathbf{q}_b \in \mathbb{R}^k$.

In the Lie group setting, the *nonlinear* structure of the configuration space G makes it possible to represent large rotations *globally* without singularities. Under reasonable smoothness assumptions, there are smooth functions $q : [t_0, t_{\mathrm{end}}] \to G$ solving the equations of motion on a time interval $[t_0, t_{\mathrm{end}}]$ of finite length, see Sect. 2.3 below. *Locally*, for a fixed time $t = t^* \in [t_0, t_{\mathrm{end}}]$, the configuration space in a sufficiently small neighbourhood of $q(t^*)$ may nevertheless be parametrized by elements of the *linear* space $\mathfrak{g}$ that is independent of t^* and $q(t^*)$, see (9).

The local parametrization of G by elements $\widetilde{\mathbf{v}} \in \mathfrak{g}$ provides the basis for an efficient implementation of Lie group time integration methods and for the analysis of discretization errors, see Sects. 3 and 4 below. Using the notation $\exp(\cdot)$ we will assume tacitly throughout the paper that the argument of the exponential map is in a small neighbourhood of $\widetilde{\mathbf{0}} \in \mathfrak{g}$ on which exp is a diffeomorphism.

The basic concepts of time discretization and error analysis in Lie group time integration are not limited to the specific parametrization by the exponential map, see, e.g., (Kobilarov et al. 2009) for an analysis of variational Lie group integrators that may be combined with the exponential map exp, with the Cayley transform

$\mathrm{cay}(\widetilde{\mathbf{v}}/2) = (\mathbf{I} - \widetilde{\mathbf{v}}/2)^{-1}(\mathbf{I} + \widetilde{\mathbf{v}}/2)$ or with other local parametrizations. In the present paper, we restrict ourselves, however, to the exponential map that reproduces the flow exactly if the velocity $\widetilde{\mathbf{v}} \in \mathfrak{g}$ is constant, see (7).

2.3 Configuration Space with Lie Group Structure: Equations of Motion

In a k-dimensional configuration space G with Lie group structure, the kinematic relations are given by (6) with position coordinates $q(t) \in G$ and the velocity vector $\mathbf{v}(t) \in \mathbb{R}^k$.

We consider constrained systems with $m \leq k$ linearly independent holonomic constraints $\boldsymbol{\Phi}(q) = \mathbf{0}$ that are coupled by constraint forces $-\mathbf{B}^\top(q)\boldsymbol{\lambda}$ to the equilibrium equations for forces and momenta. Here, $\boldsymbol{\lambda}(t) \in \mathbb{R}^m$ denotes a vector of Lagrange multipliers which is multiplied by the transposed of the constraint matrix $\mathbf{B}(q) \in \mathbb{R}^{m \times k}$ with rank $\mathbf{B}(q) = m$ that represents the constraint gradients in the sense that

$$D\boldsymbol{\Phi}(q) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big) = \mathbf{B}(q)\mathbf{w}\,,\ \ (\,\mathbf{w} \in \mathbb{R}^k\,)\,. \tag{14}$$

The notation $D\boldsymbol{\Phi}(q) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big)$ is used for the directional derivative of $\boldsymbol{\Phi} : G \to \mathbb{R}^m$ at $q \in G$ in the direction of $DL_q(e) \cdot \widetilde{\mathbf{w}} \in T_qG$.

Kinematic equations, equilibrium conditions and holonomic constraints are summarized in the equations of motion

$$\dot{q} = DL_q(e) \cdot \widetilde{\mathbf{v}}\,, \tag{15a}$$

$$\mathbf{M}(q)\dot{\mathbf{v}} = -\mathbf{g}(q, \mathbf{v}, t) - \mathbf{B}^\top(q)\boldsymbol{\lambda}\,, \tag{15b}$$

$$\boldsymbol{\Phi}(q) = \mathbf{0} \tag{15c}$$

that form a differential-algebraic equation (DAE) on Lie group G, see (Brüls and Cardona 2010). Matrix $\mathbf{M}(q)$ denotes the mass matrix that is supposed to be symmetric, positive definite. The force vector $-\mathbf{g}(q, \mathbf{v}, t)$ summarizes external, internal and complementary inertia forces. Throughout the present paper, we consider equations of motion (15) with functions $\mathbf{M}(q)$, $\mathbf{g}(q, \mathbf{v}, t)$ and $\boldsymbol{\Phi}(q)$ being smooth in the sense that they are as often continuously differentiable as required by the convergence analysis.

Remark 2.3 (a) For linear configuration spaces, the equations of motion (15) are well known from textbooks on DAE time integration, see, e.g., (Brenan et al. 1996, Sect. 6.2 and Hairer and Wanner 1996, Sect. VII.1). Model equations of constrained mechanical and mechatronic systems in industrial applications have often a more complex structure with additional first-order differential equations $\dot{\mathbf{c}} = \mathbf{h}_\mathbf{c}(q, \mathbf{v}, \mathbf{c}, t)$ or additional algebraic equations $\mathbf{0} = \mathbf{h}_\mathbf{s}(q, \mathbf{s})$ that are locally uniquely solvable w.r.t. $\mathbf{s} = \mathbf{s}(q)$ if the Jacobian $(\partial\mathbf{h}_\mathbf{s}/\partial\mathbf{s})(q, \mathbf{s})$ is non-singular. Other useful generalizations of (15) are rheonomic, i.e., explicitly time-dependent constraints $\mathbf{\Phi}(q,t)=\mathbf{0}$ and force vectors $\mathbf{g}=\mathbf{g}(q,\mathbf{v},\boldsymbol{\lambda},t)$ that contain friction forces depending nonlinearly on $\boldsymbol{\lambda}$, see (Arnold et al. 2011c and Brüls and Golinval 2006) for a more detailed discussion. All these additional model components may be considered straightforwardly in the convergence analysis of generalized-α Lie group integrators, see (Arnold et al. 2015).

(b) The full rank assumption on $\mathbf{B}(q)$ is essential for the analysis and numerical solution of (15) since otherwise the Lagrange multipliers $\boldsymbol{\lambda}(t)$ would not be uniquely defined, see (García de Jalón and Bayo 1994, Sect. 3.4) and the more recent material in (García de Jalón and Gutiérrez-López 2013). On the other hand, the assumptions on $\mathbf{M}(q)$ may be slightly relaxed considering symmetric, positive semi-definite mass matrices that are positive definite on $\ker \mathbf{B}(q)$, see (Géradin and Cardona 2001). The extension of the convergence analysis to this more complex class of model equations has recently been discussed in (Arnold et al. 2014).

The holonomic constraints (15c) imply hidden constraints at the level of velocity coordinates and at the level of acceleration coordinates. The first ones are obtained by differentiation of (15c) w.r.t. t:

$$\mathbf{0}=\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Phi}(q(t))=D\mathbf{\Phi}(q(t))\cdot\dot{q}(t)=D\mathbf{\Phi}(q)\cdot\big(DL_q(e)\cdot\widetilde{\mathbf{v}}\big)=\mathbf{B}(q)\mathbf{v}\,. \tag{16}$$

For the second time derivative of (15c), we have to consider partial derivatives of $\mathbf{\Theta}(q,\mathbf{z}):=\mathbf{B}(q)\mathbf{z}$ w.r.t. $q\in G$. Since $\mathbf{\Theta}\,:\,G\times\mathbb{R}^k\to\mathbb{R}^m$ is by construction linear in $\mathbf{z}$ we have

$$D_q\mathbf{\Theta}(q,\mathbf{z})\cdot\big(DL_q(e)\cdot\widetilde{\mathbf{w}}\big)=\mathbf{Z}(q)(\mathbf{z},\mathbf{w})\,,\quad(\,\mathbf{w}\in\mathbb{R}^k\,) \tag{17}$$

with a bilinear form $\mathbf{Z}(q)\,:\,\mathbb{R}^k\times\mathbb{R}^k\to\mathbb{R}^m$. Using these notations, the time derivative of (16) gets the form

$$\mathbf{0}=\frac{\mathrm{d}}{\mathrm{d}t}\big(\mathbf{B}(q(t))\mathbf{v}(t)\big)=\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{\Theta}\big(q(t),\mathbf{v}(t)\big)=\mathbf{B}(q)\dot{\mathbf{v}}+\mathbf{Z}(q)(\mathbf{v},\mathbf{v})\,. \tag{18}$$

It defines the hidden constraints at the level of acceleration coordinates.

The dynamical equations (15b) and the hidden constraints (18) are linear in $\dot{\mathbf{v}}(t)$ and $\boldsymbol{\lambda}(t)$ and may formally be used to eliminate $\boldsymbol{\lambda}(t)$ and to express $\dot{\mathbf{v}}(t)$ in terms of t, $q(t)$ and $\mathbf{v}(t)$, see (Hairer and Wanner 1996, Sect. VII.1):

$$\begin{pmatrix}\mathbf{M}(q) & \mathbf{B}^\top(q)\\ \mathbf{B}(q) & \mathbf{0}\end{pmatrix}\begin{pmatrix}\dot{\mathbf{v}}\\ \boldsymbol{\lambda}\end{pmatrix}=\begin{pmatrix}-\mathbf{g}(q,\mathbf{v},t)\\ -\mathbf{Z}(q)(\mathbf{v},\mathbf{v})\end{pmatrix}. \tag{19}$$

Initial value problems for the resulting analytically equivalent unconstrained system for functions $q\,:\,[t_0,t_{\mathrm{end}}]\to G$ and $\mathbf{v}\,:\,[t_0,t_{\mathrm{end}}]\to\mathbb{R}^k$ are uniquely solvable whenever its right-hand side satisfies a Lipschitz condition, see, e.g., (Walter 1998). This proves unique solvability of initial value problems for the constrained system

(15) if $q(t_0)$ and $\mathbf{v}(t_0)$ are *consistent* with the (hidden) constraints (15c) and (16), i.e., $\mathbf{\Phi}\big(q(t_0)\big) = \mathbf{B}\big(q(t_0)\big)\mathbf{v}(t_0) = \mathbf{0}$. The initial values $\dot{\mathbf{v}}(t_0)$ and $\boldsymbol{\lambda}(t_0)$ are given by (19) with $t = t_0$, $q = q(t_0)$ and $\mathbf{v} = \mathbf{v}(t_0)$.

The index analysis of Lie group DAE (15) follows step by step the classical index analysis for the equations of motion for constrained mechanical systems in linear configuration spaces, see (Hairer and Wanner 1996, Sect. VII.1). The algebraic variables $\boldsymbol{\lambda} = \boldsymbol{\lambda}(q, \mathbf{v}, t)$ are defined by the system of linear equations (19) that contains the second time derivative of (15c). A formal third differentiation step yields $\dot{\boldsymbol{\lambda}} = \dot{\boldsymbol{\lambda}}(q, \mathbf{v}, t)$ and illustrates that (15) is an index-3 Lie group DAE in $G \times \mathbb{R}^k \times \mathbb{R}^m$. Therefore, Eq. (15) is called the *index-3 formulation* of the equations of motion.

Remark 2.4 Block-structured systems of linear equations

$$\begin{pmatrix} \mathbf{M} & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_{\dot{\mathbf{v}}} \\ \mathbf{x}_{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_{\dot{\mathbf{v}}} \\ \mathbf{r}_{\boldsymbol{\lambda}} \end{pmatrix} \tag{20}$$

with a symmetric, positive definite matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$ and a rectangular matrix $\mathbf{B} \in \mathbb{R}^{m\times k}$ of full rank $m \le k$ are uniquely solvable since left multiplication of the upper block row by $\mathbf{B}\mathbf{M}^{-1}$ yields equations

$$\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top\mathbf{x}_{\boldsymbol{\lambda}} = \mathbf{B}\mathbf{M}^{-1}\mathbf{r}_{\dot{\mathbf{v}}} - \mathbf{B}\mathbf{x}_{\dot{\mathbf{v}}} = \mathbf{B}\mathbf{M}^{-1}\mathbf{r}_{\dot{\mathbf{v}}} - \mathbf{r}_{\boldsymbol{\lambda}}$$

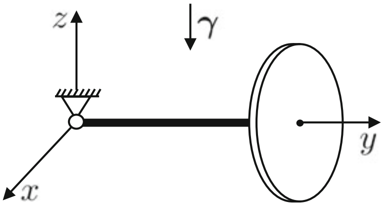
that may be solved w.r.t. $\mathbf{x}_{\boldsymbol{\lambda}} \in \mathbb{R}^m$ since $\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top$ is symmetric, positive definite. Inserting this vector $\mathbf{x}_{\boldsymbol{\lambda}}$ in the upper block row, we get $\mathbf{x}_{\dot{\mathbf{v}}} \in \mathbb{R}^k$ from $\mathbf{M}\mathbf{x}_{\dot{\mathbf{v}}} = \mathbf{r}_{\dot{\mathbf{v}}} - \mathbf{B}^\top\mathbf{x}_{\boldsymbol{\lambda}}$. The most time-consuming parts of this block Gaussian elimination are the Cholesky factorization of $\mathbf{M} \in \mathbb{R}^{k\times k}$ (to get $\mathbf{M}^{-1}\mathbf{B}^\top \in \mathbb{R}^{k\times m}$ and $\mathbf{M}^{-1}\mathbf{r}_{\dot{\mathbf{v}}} \in \mathbb{R}^k$) the evaluation of the matrix-matrix product $\mathbf{B}(\mathbf{M}^{-1}\mathbf{B}^\top) \in \mathbb{R}^{m\times m}$ and the Cholesky factorization of this matrix.

Alternatively, we could follow a nullspace approach that separates the nullspace of $\mathbf{B} \in \mathbb{R}^{m\times k}$ from a non-singular matrix $\bar{\mathbf{R}} \in \mathbb{R}^{m\times m}$: For any non-singular matrix $\mathbf{Q} \in \mathbb{R}^{k\times k}$ with $\mathbf{B}\mathbf{Q} = \big(\bar{\mathbf{R}}^\top,\ \mathbf{0}_{m\times(k-m)}\big)$, system (20) is equivalent to

$$\begin{pmatrix} \bar{\mathbf{M}}_{11} & \bar{\mathbf{M}}_{12} & \bar{\mathbf{R}} \\ \bar{\mathbf{M}}_{21} & \bar{\mathbf{M}}_{22} & \mathbf{0} \\ \bar{\mathbf{R}}^\top & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \bar{\mathbf{x}}_{\dot{\mathbf{v}},1} \\ \bar{\mathbf{x}}_{\dot{\mathbf{v}},2} \\ \mathbf{x}_{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{r}}_{\dot{\mathbf{v}},1} \\ \bar{\mathbf{r}}_{\dot{\mathbf{v}},2} \\ \mathbf{r}_{\boldsymbol{\lambda}} \end{pmatrix} \text{ with } \begin{pmatrix} \bar{\mathbf{M}}_{11} & \bar{\mathbf{M}}_{12} \\ \bar{\mathbf{M}}_{21} & \bar{\mathbf{M}}_{22} \end{pmatrix} = \mathbf{Q}^\top\mathbf{M}\mathbf{Q}\,,$$

$\bar{\mathbf{x}}_{\dot{\mathbf{v}}} = \mathbf{Q}^{-1}\mathbf{x}_{\dot{\mathbf{v}}}$ and $\bar{\mathbf{r}}_{\dot{\mathbf{v}}} = \mathbf{Q}^\top\mathbf{r}_{\dot{\mathbf{v}}}$. This block-structured system may be solved in three steps by block backward substitution to get $\bar{\mathbf{x}}_{\dot{\mathbf{v}},1}$, $\bar{\mathbf{x}}_{\dot{\mathbf{v}},2}$ and $\bar{\mathbf{x}}_{\boldsymbol{\lambda}}$ since matrices $\bar{\mathbf{R}}^\top$, $\bar{\mathbf{M}}_{22}$ and $\bar{\mathbf{R}}$ are non-singular. Betsch and Leyendecker (2006) discussed analytical nullspace representations of the constraint matrix $\mathbf{B}$ for typical types of constraints in engineering systems. If such analytical expressions are not available, then matrices $\mathbf{Q}$ and $\bar{\mathbf{R}}$ could be computed, e.g., by a QR-factorization of $\mathbf{B}^\top \in \mathbb{R}^{k\times m}$, see (Golub and van Loan 1996).

Fig. 2 Benchmark problem Heavy top (Brüls and Cardona 2010), see also (Géradin and Cardona 2001)



2.4 Benchmark Problem: Heavy Top

The Lie group formulation of the equations of motion is the backbone of a rather general finite element framework for flexible multibody dynamics (Géradin and Cardona 2001). In the present paper, we focus on basic aspects of Lie group time integration in multibody dynamics and restrict the numerical tests to the simulation of a single rigid body in a gravitation field. This *heavy top* has found much interest in mechanics and serves as a benchmark problem for Lie group methods (Géradin and Cardona 2001, Sect. 5.8). The simulation of more complex flexible structures by Lie group time integration methods is discussed, e.g., in (Brüls et al. 2012).

Figure 2 shows the configuration of the heavy top in $\mathbb{R}^3$ with $\mathbf{R}(t) \in \mathrm{SO}(3)$ characterizing its orientation and the position vector $\mathbf{x}(t) \in \mathbb{R}^3$ of the centre of mass in the inertial frame. In the body-attached frame, the centre of mass is given by $\mathbf{X} = (\,0,\ 1,\ 0\,)^\top$. Here and in the following, we omit all physical units. We consider a gravitation field with fixed acceleration vector $\boldsymbol{\gamma} = (0, 0, -9.81)^\top$. Mass and inertia tensor are given by $m = 15.0$ and $\mathbf{J} = \mathrm{diag}\,(0.234375, 0.46875, 0.234375)$ with $\mathbf{J}$ denoting the inertia tensor w.r.t. the centre of mass.

In the benchmark problem, the top rotates about a fixed point. Therefore, the configuration variables $(\mathbf{R}, \mathbf{x})$ are subject to holonomic constraints $\mathbf{x} = \mathbf{R}\mathbf{X}$. We consider an initial configuration being defined by $\mathbf{R}(0) = \mathbf{I}_3$ with an angular velocity $\boldsymbol{\Omega}(0) = (0, 150, -4.61538)^\top$. All other initial values are supposed to be consistent with $\mathbf{0} = \boldsymbol{\Phi}\big((\mathbf{R}, \mathbf{x})\big) := \mathbf{X} - \mathbf{R}^\top\mathbf{x}$ and with the corresponding hidden constraints (16) and (18) at the level of velocity and acceleration coordinates.

The equations of motion (15) of the rotating heavy top result from the principles of classical mechanics. In (Brüls et al. 2011), they were derived for configuration spaces $G = \mathrm{SO}(3) \times \mathbb{R}^3$ and $G = \mathrm{SE}(3)$ following an augmented Lagrangian method. In $\mathrm{SO}(3) \times \mathbb{R}^3$, we get hidden constraints

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{X} - \mathbf{R}^\top\mathbf{x}) = -\dot{\mathbf{R}}^\top\mathbf{x} - \mathbf{R}^\top\dot{\mathbf{x}} = -\widetilde{\boldsymbol{\Omega}}^\top\mathbf{R}^\top\mathbf{x} - \mathbf{R}^\top\mathbf{u} = -\widetilde{\mathbf{X}}\boldsymbol{\Omega} - \mathbf{R}^\top\mathbf{u}$$

and a constraint matrix $\mathbf{B} = (-\widetilde{\mathbf{X}} \quad -\mathbf{R}^\top)$. The equations of motion are given by

$$\mathbf{J}\dot{\boldsymbol{\Omega}} + \boldsymbol{\Omega} \times \mathbf{J}\boldsymbol{\Omega} + \mathbf{X} \times \boldsymbol{\lambda} = \mathbf{0}\,, \tag{21a}$$

$$m\dot{\mathbf{u}} - \mathbf{R}\boldsymbol{\lambda} = m\boldsymbol{\gamma}\,, \tag{21b}$$

$$\mathbf{X} - \mathbf{R}^\top\mathbf{x} = \mathbf{0} \tag{21c}$$

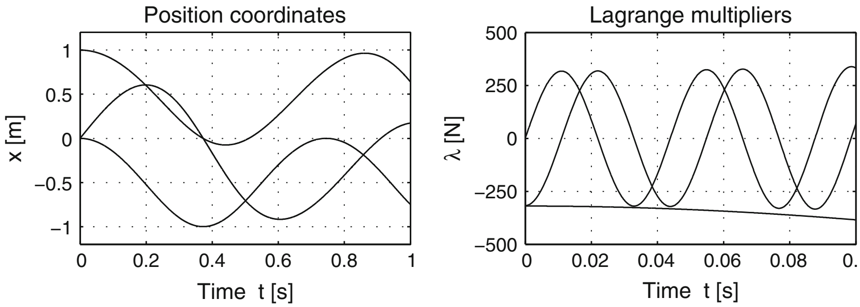


Fig. 3 Heavy top benchmark, $G = \mathrm{SO}(3) \times \mathbb{R}^3$: Reference solution

with kinematic relations (1). Figure 3 shows a reference solution that has been computed with the very small time step size $h = 2.5 \times 10^{-5}$. The position $\mathbf{x}(t) \in \mathbb{R}^3$ of the centre of mass varies slowly in the inertial frame. For the Lagrange multipliers $\boldsymbol{\lambda}(t) \in \mathbb{R}^3$, we observe much higher frequencies that reflect the fast rotation of the top being caused by the rather large initial velocity $\boldsymbol{\Omega}(0)$. Note, that the time scale in the right plot of Fig. 3 has been zoomed by a factor of 10.

In the configuration space SE(3), we have $\dot{\mathbf{x}} = \mathbf{RU}$ resulting in hidden constraints $\mathbf{0} = -\widetilde{\mathbf{X}}\boldsymbol{\Omega} - \mathbf{U}$ with a constraint matrix $\mathbf{B} = (-\widetilde{\mathbf{X}} \;\; -\mathbf{I}_3)$ that is constant and does not depend on $q \in G$. The equations of motion are given by

$$\mathbf{J}\dot{\boldsymbol{\Omega}} + \boldsymbol{\Omega} \times \mathbf{J}\boldsymbol{\Omega} + \mathbf{X} \times \boldsymbol{\lambda} = \mathbf{0}\,, \tag{22a}$$

$$m\dot{\mathbf{U}} + m\boldsymbol{\Omega} \times \mathbf{U} - \boldsymbol{\lambda} = \mathbf{R}^\top m\boldsymbol{\gamma}\,, \tag{22b}$$

$$\mathbf{X} - \mathbf{R}^\top \mathbf{x} = \mathbf{0} \tag{22c}$$

with kinematic relations (3). The position coordinates $q = (\mathbf{R}, \mathbf{x})$ coincide for both formulations (21) and (22) but there may be substantial differences between the velocity coordinates $\mathbf{u}(t)$ in the inertial frame and their counterparts $\mathbf{U}(t)$ in the body-attached frame. This is illustrated by the simulation results in Fig. 4 that have been obtained again with time step size $h = 2.5 \times 10^{-5}$. In $\mathrm{SO}(3) \times \mathbb{R}^3$, we observe low frequency changes of $\mathbf{u}(t)$ that correspond to the solution behaviour of $\mathbf{x}(t)$ in the left plot of Fig. 3. For the configuration space $G = \mathrm{SE}(3)$, we see in the right plot of Fig. 4 the dominating influence of the large initial velocity $\boldsymbol{\Omega}(0)$ on the qualitative solution behaviour of $\mathbf{U}(t)$.

Throughout the paper, we will use the two different formulations (21) and (22) of the heavy top benchmark problem for numerical tests to discuss various aspects of the convergence analysis for the generalized-α Lie group integrator.

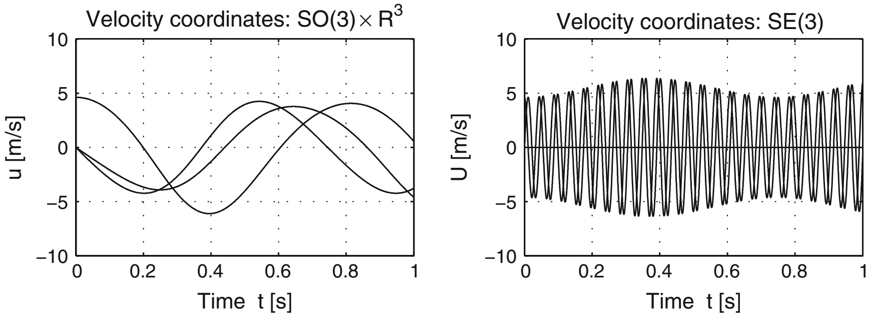


Fig. 4 Heavy top benchmark: Velocity coordinates in the inertial frame ($\mathbf{u}(t)$, *left plot*) and in the body-attached frame ($\mathbf{U}(t)$, *right plot*)

2.5 More on the Exponential Map

Equation (7) illustrates the crucial role of the exponential map for multibody system models that have a configuration space with Lie group structure. Since the numerical solution proceeds in time steps, we have to study the composition of exponential maps with different arguments in more detail. Furthermore, the proposed Lie group time integration methods are implicit and rely on a Newton–Raphson iteration that requires the efficient evaluation of Jacobians $(\partial\mathbf{h}/\partial\mathbf{v})\big(q \circ \exp(\widetilde{\mathbf{v}})\big)$ for vector-valued functions $\mathbf{h} : G \to \mathbb{R}^l$. In the present section, we follow the presentation in (Hairer et al. 2006, Sect. III.4) to discuss these rather technical aspects of Lie group time integration.

For matrix Lie groups, the exponential map exp is given by the matrix exponential. For $s \in \mathbb{R}$ and any matrices $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{r\times r}$, the series expansion (8) shows

$$\begin{aligned}\exp(s\mathbf{A})\exp(s\mathbf{C}) &= (\mathbf{I}_r + s\mathbf{A} + \frac{s^2}{2}\mathbf{A}^2)(\mathbf{I}_r + s\mathbf{C} + \frac{s^2}{2}\mathbf{C}^2) + \mathcal{O}(s^3)\\ &= \mathbf{I}_r + s(\mathbf{A}+\mathbf{C}) + \frac{s^2}{2}(\mathbf{A}^2 + 2\mathbf{A}\mathbf{C} + \mathbf{C}^2) + \mathcal{O}(s^3)\\ &= \mathbf{I}_r + s(\mathbf{A}+\mathbf{C}) + \frac{s^2}{2}(\mathbf{A}+\mathbf{C})^2 + \frac{1}{2}[s\mathbf{A}, s\mathbf{C}] + \mathcal{O}(s^3)\\ &= \exp\big(s\mathbf{A} + s\mathbf{C} + \frac{1}{2}[s\mathbf{A}, s\mathbf{C}]\big) + \mathcal{O}(s^3)\,,\ \ (s \to 0)\end{aligned}$$

with the *matrix commutator* $[\mathbf{A}, \mathbf{C}] := \mathbf{A}\mathbf{C} - \mathbf{C}\mathbf{A}$ that vanishes iff matrices $\mathbf{A}$ and $\mathbf{C}$ commute. For a slightly more detailed analysis of the product of matrix exponentials, we use the Baker–Campbell–Hausdorff formula, see (Hairer et al. 2006, Lemma III.4.3), to get the following estimate:

Lemma 2.5 *For* $s \to 0$*, the product of matrix exponentials* $\exp(s\mathbf{A})$ *and* $\exp(s\mathbf{C})$ *satisfies*

$$\exp(s\mathbf{A})\exp(s\mathbf{C}) = \exp\big(s\mathbf{A} + s\mathbf{C} + \frac{1}{2}[s\mathbf{A}, s\mathbf{C}] + \mathcal{O}(s)\|[s\mathbf{A}, s\mathbf{C}]\|\big)\,. \tag{23}$$

Proof The Baker–Campbell–Hausdorff formula defines the argument of the matrix exponential at the right-hand side of (23) by the solution of an initial value problem with zero initial values at $s = 0$. Solving this initial value problem by Picard iteration with starting guess $s\mathbf{A} + s\mathbf{C} + [s\mathbf{A}, s\mathbf{C}]/2$, we may show that all higher order terms result in a remainder term of size $\mathcal{O}(s)\|[s\mathbf{A}, s\mathbf{C}]\|$, see (23). □

For fixed argument $\mathbf{A}$, the matrix commutator defines a linear operator

$$\mathrm{ad}_{\mathbf{A}} \,:\, \mathbb{R}^{r\times r} \to \mathbb{R}^{r\times r}\,,\ \mathbf{C} \mapsto \mathrm{ad}_{A} := [\mathbf{A}, \mathbf{C}] \tag{24}$$

that is called the *adjoint operator*. By recursive application of $\mathrm{ad}_{\mathbf{A}}$ we may represent directional derivatives of the exponential map $\exp(\mathbf{A}) = \sum_i \mathbf{A}^i / i!$ in compact form: We denote $\mathrm{ad}^0_{\mathbf{A}}(\mathbf{C}) := \mathbf{C}$ and

$$\mathrm{ad}^{j+1}_{\mathbf{A}}(\mathbf{C}) := \mathrm{ad}_{\mathbf{A}}\big(\mathrm{ad}^{j}_{\mathbf{A}}(\mathbf{C})\big) = \mathbf{A}\,\mathrm{ad}^{j}_{\mathbf{A}}(\mathbf{C}) - \mathrm{ad}^{j}_{\mathbf{A}}(\mathbf{C})\,\mathbf{A}\,,\ (\,j \geq 1\,) \tag{25}$$

and consider powers $(\mathbf{A} + s\mathbf{C})^i$, $(\,i \geq 0\,)$, in the limit case $s \to 0$. For $i = 2$, we get

$$(\mathbf{A} + s\mathbf{C})^2 = \mathbf{A}^2 + s(\mathbf{AC} + \mathbf{CA}) + \mathcal{O}(s^2) = \mathbf{A}^2 + s\big(2\mathbf{AC} + \mathrm{ad}_{-\mathbf{A}}(\mathbf{C})\big) + \mathcal{O}(s^2)\,.$$

Here, the term $\mathrm{ad}_{-\mathbf{A}}(\mathbf{C})$ results from the non-commutativity of matrix multiplication and could be represented as well by the adjoint operator $\mathrm{ad}_{\mathbf{A}}$ itself since $2\mathbf{AC} + \mathrm{ad}_{-\mathbf{A}}(\mathbf{C}) = 2\mathbf{CA} + \mathrm{ad}_{\mathbf{A}}(\mathbf{C})$, see (Hairer et al. 2006). The use of $\mathrm{ad}_{-\mathbf{A}}$ corresponds, however, to the characterization of the tangent space $T_q G$ by left translations L_q, see (5) and the discussion in (Iserles et al. 2000). In multibody dynamics, this characterization implies that vector $\mathbf{v}$ in the kinematic relations (6) is a left-invariant velocity vector. These left-invariant vectors are favourable since the associated rotational inertia are defined in the body-attached frame and the body mass matrices remain constant during motion (Brüls et al. 2011).

Lemma 2.6 *For* $s \to 0$ *and matrices* $\mathbf{A}, \mathbf{C} \in \mathbb{R}^{r\times r}$*, the asymptotic behaviour of* $(\mathbf{A} + s\mathbf{C})^i$ *and* $\exp(\mathbf{A} + s\mathbf{C})$ *is characterized by*

$$(\mathbf{A} + s\mathbf{C})^i = \mathbf{A}^i + s\sum_{j=0}^{i-1}\binom{i}{j+1}\mathbf{A}^{i-j-1}\,\mathrm{ad}^{j}_{-\mathbf{A}}(\mathbf{C}) + \mathcal{O}(s^2)\,,\ (\,i \geq 1\,)\,, \tag{26}$$

and

$$\exp(\mathbf{A} + s\mathbf{C}) = \exp(\mathbf{A})\,\big(\mathbf{I}_r + s\ \mathrm{dexp}_{-\mathbf{A}}(\mathbf{C})\big) + \mathcal{O}(s^2) \tag{27}$$

with the matrix-valued function

$$\mathrm{dexp}_{-\mathbf{A}}(\mathbf{C}) := \sum_{j=0}^{\infty} \frac{1}{(j+1)!}\, \mathrm{ad}_{-\mathbf{A}}^{j}(\mathbf{C}) \tag{28}$$

that satisfies $\mathrm{dexp}_{-\mathbf{A}}(\mathbf{C}) = \mathbf{C}$ *whenever* $\mathbf{A}$ *and* $\mathbf{C}$ *commute.*

Proof To prove (26) by induction, we multiply this expression from the right by $(\mathbf{A} + s\mathbf{C})$ and observe that $\mathbf{A}^i\, s\mathbf{C} = s\mathbf{A}^{(i+1)-j-1}\, \mathrm{ad}_{-\mathbf{A}}^{j}(\mathbf{C})$ with $j = 0$. Taking into account the identity

$$\mathbf{A}^{i-j-1}\, \mathrm{ad}_{-\mathbf{A}}^{j}(\mathbf{C})\, \mathbf{A} = \mathbf{A}^{(i+1)-(j+1)-1}\, \mathrm{ad}_{-\mathbf{A}}^{j+1}(\mathbf{C}) + \mathbf{A}^{(i+1)-j-1}\, \mathrm{ad}_{-\mathbf{A}}^{j}(\mathbf{C})\,,$$

see (25), we get (26) with i being substituted by $i + 1$ since

$$\binom{i}{j} + \binom{i}{j+1} = \binom{i+1}{j+1}, \;\; (\, j = 0, 1, \ldots, i-1\,)\,.$$

For the proof of (27), we scale (26) by $1/i!$ and use the series expansion (8) to get

$$\begin{aligned}
\exp(\mathbf{A} + s\mathbf{C}) &= \sum_{i=0}^{\infty} \frac{1}{i!}\mathbf{A}^i + s \sum_{i=0}^{\infty} \frac{1}{i!} \sum_{j=0}^{i-1} \binom{i}{j+1} \mathbf{A}^{i-j-1}\; \mathrm{ad}_{-\mathbf{A}}^{j}(\mathbf{C}) + \mathcal{O}(s^2) \\
&= \sum_{i=0}^{\infty} \frac{1}{i!}\mathbf{A}^i + s \sum_{j=0}^{\infty} \frac{1}{(j+1)!} \underbrace{\sum_{i=j+1}^{\infty} \frac{1}{(i-j-1)!}\mathbf{A}^{i-j-1}}_{= \sum_{i=0}^{\infty} \frac{1}{i!}\mathbf{A}^i = \exp(\mathbf{A})}\; \mathrm{ad}_{-\mathbf{A}}^{j}(\mathbf{C}) + \mathcal{O}(s^2) \\
&= \exp(\mathbf{A}) \big(\mathbf{I}_r + s\, \mathrm{dexp}_{-\mathbf{A}}(\mathbf{C})\big) + \mathcal{O}(s^2)\,.
\end{aligned}$$

For commuting matrices $\mathbf{A}$ and $\mathbf{C}$, the iterated adjoint operators $\mathrm{ad}_{-\mathbf{A}}^{j}(\mathbf{C})$ vanish for all $j > 0$ resulting in $\mathrm{dexp}_{-\mathbf{A}}(\mathbf{C}) = \mathbf{C}$, see (28). □

Lemma 2.6 shows that the directional derivative of the matrix exponential is given by $(\partial/\partial\mathbf{A}) \exp(\mathbf{A})\mathbf{C} = \exp(\mathbf{A})\, \mathrm{dexp}_{-\mathbf{A}}(\mathbf{C})$. In the Lie group setting, we use this expression to study the Jacobian of vector-valued functions $\mathbf{h}\big(q \circ \exp(\widetilde{\mathbf{v}})\big)$ w.r.t. $\mathbf{v} \in \mathbb{R}^k$. For elements $\widetilde{\mathbf{v}}, \widetilde{\mathbf{w}} \in \mathfrak{g}$, the terms $\mathrm{ad}_{-\widetilde{\mathbf{v}}}(\widetilde{\mathbf{w}})$ and $\mathrm{dexp}_{-\widetilde{\mathbf{v}}}(\widetilde{\mathbf{w}})$ are linear in $\mathbf{w} \in \mathbb{R}^k$ and may be represented by matrix-vector products in $\mathbb{R}^k$ using the notation

$$\widehat{(\bullet)}\; : \; \mathbb{R}^k \to \mathbb{R}^{k \times k} \quad \text{with} \quad \widehat{\mathbf{v}}\mathbf{w} = \mathrm{ad}_{\widetilde{\mathbf{v}}}(\widetilde{\mathbf{w}}) = [\widetilde{\mathbf{v}}, \widetilde{\mathbf{w}}]\,, \;\; (\mathbf{v}, \mathbf{w} \in \mathbb{R}^k\,)\,. \tag{29}$$

With (29), the operators $\mathrm{ad}_{\widetilde{\mathbf{v}}}$, $\mathrm{ad}_{-\widetilde{\mathbf{v}}}$ and $\mathrm{ad}_{-\widetilde{\mathbf{v}}}^{j}$ correspond to $k \times k$-matrices $\widehat{\mathbf{v}}$, $-\widehat{\mathbf{v}}$ and $(-\widehat{\mathbf{v}})^j$, respectively, and the counterpart to $\widetilde{\mathbf{z}} = \mathrm{dexp}_{-\widetilde{\mathbf{v}}}(\widetilde{\mathbf{w}}) \in \mathfrak{g}$, see (28), is given by $\mathbf{z} = \mathbf{T}(\mathbf{v})\mathbf{w} \in \mathbb{R}^k$ with the *tangent operator*

$$\mathbf{T} : \mathbb{R}^k \to \mathbb{R}^{k\times k}, \quad \mathbf{T}(\mathbf{v}) = \sum_{i=0}^{\infty} \frac{(-1)^i}{(i+1)!} \widehat{\mathbf{v}}^i , \tag{30}$$

see (Iserles et al. 2000). Using the chain rule, we obtain

Corollary 2.7 *Consider a continuously differentiable function* $\mathbf{h}\colon G \to \mathbb{R}^l$ *and a matrix-valued function* $\mathbf{H}\colon G \to \mathbb{R}^{l\times k}$ *that represents the derivative of* $\mathbf{h}$ *in the sense that*

$$D\mathbf{h}(q) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big) = \mathbf{H}(q)\mathbf{w} , \ \ (\mathbf{w} \in \mathbb{R}^k) ,$$

see (14). The Jacobian of $\mathbf{h}\big(q \circ \exp(\widetilde{\mathbf{v}})\big)$ *w.r.t.* $\mathbf{v} \in \mathbb{R}^k$ *is given by*

$$\frac{\partial \mathbf{h}}{\partial \mathbf{v}}\big(q \circ \exp(\widetilde{\mathbf{v}})\big) = \mathbf{H}\big(q \circ \exp(\widetilde{\mathbf{v}})\big)\mathbf{T}(\mathbf{v}) . \tag{31}$$

Remark 2.8 (a) For commuting elements of the Lie algebra ($\widetilde{\mathbf{v}}, \widetilde{\mathbf{w}} \in \mathfrak{g}$ with $[\widetilde{\mathbf{v}}, \widetilde{\mathbf{w}}] = \widetilde{\mathbf{0}}$), the adjoint operator vanishes resulting in $\widehat{\mathbf{v}}\mathbf{w} = \mathbf{0}_k$ and $\mathbf{T}(\mathbf{v})\mathbf{w} = \mathbf{w}$. Therefore, the tangent operator satisfies $\mathbf{T}(\mathbf{v})\mathbf{v} = \mathbf{v}$, ($\mathbf{v} \in \mathbb{R}^k$), and Corollary 2.7 implies

$$\frac{\mathrm{d}\mathbf{h}}{\mathrm{d}\vartheta}\big(q \circ \exp(\vartheta\widetilde{\mathbf{v}})\big) = \mathbf{H}\big(q \circ \exp(\vartheta\widetilde{\mathbf{v}})\big)\mathbf{v} \tag{32}$$

with $\vartheta \in \mathbb{R}$ and any vector $\mathbf{v} \in \mathbb{R}^k$.

(b) The efficient evaluation of the tangent operator is essential for an efficient implementation of implicit Lie group integrators. In the Lie group $G = \mathrm{SO}(3)$, the hat operator maps $\mathbf{\Omega} \in \mathbb{R}^3$ to $\widehat{\mathbf{\Omega}} := \widetilde{\mathbf{\Omega}}$ with the skew symmetric matrix $\widetilde{\mathbf{\Omega}}$ being defined in (2). Similar to Rodrigues' formula (10), the tangent operator $\mathbf{T}_{\mathrm{SO}(3)}$ may be evaluated in closed form (Brüls et al. 2011):

$$\mathbf{T}_{\mathrm{SO}(3)}(\mathbf{\Omega}) = \mathbf{I}_3 + \frac{\cos\Phi - 1}{\Phi^2}\,\widetilde{\mathbf{\Omega}} + \frac{1 - \dfrac{\sin\Phi}{\Phi}}{\Phi^2}\,\widetilde{\mathbf{\Omega}}^2 . \tag{33}$$

For $G = \mathrm{SO}(3) \times \mathbb{R}^3$, the Lie algebra $\mathfrak{g} = \mathfrak{so}(3) \times \mathbb{R}^3$ is parametrized by vectors $\mathbf{v} = (\mathbf{\Omega}^\top, \mathbf{u}^\top)^\top \in \mathbb{R}^6$ and we get

$$\widehat{\mathbf{v}} = \mathrm{blockdiag}\,(\,\widetilde{\mathbf{\Omega}},\ \mathbf{0}_{3\times 3}\,) , \quad \mathbf{T}_{\mathrm{SO}(3)\times\mathbb{R}^3}(\mathbf{v}) = \mathrm{blockdiag}\,\big(\mathbf{T}_{\mathrm{SO}(3)}(\mathbf{\Omega}),\ \mathbf{I}_3\,\big) .$$

More complex expressions are obtained for the Lie group $G = \mathrm{SE}(3)$ and its Lie algebra $\mathfrak{se}(3)$ that is parametrized by vectors $\mathbf{v} = (\mathbf{\Omega}^\top, \mathbf{U}^\top)^\top \in \mathbb{R}^6$ with

$$\widehat{\mathbf{v}} = \begin{pmatrix} \widetilde{\mathbf{\Omega}} & \mathbf{0}_{3\times 3} \\ \widetilde{\mathbf{U}} & \widetilde{\mathbf{\Omega}} \end{pmatrix} . \tag{34}$$

Using the identities $\widetilde{\boldsymbol{\Omega}}^3 = -\Phi^2\widetilde{\boldsymbol{\Omega}}$, $\widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Omega}} = -(\boldsymbol{\Omega}^\top\mathbf{U})\mathbf{I}_3 + \boldsymbol{\Omega}\mathbf{U}^\top$, $\widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Omega}}^2 + \widetilde{\boldsymbol{\Omega}}^2\widetilde{\mathbf{U}} = -\Phi^2\widetilde{\mathbf{U}} - (\boldsymbol{\Omega}^\top\mathbf{U})\widetilde{\boldsymbol{\Omega}}$ and $\widetilde{\boldsymbol{\Omega}}\widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Omega}} = -(\boldsymbol{\Omega}^\top\mathbf{U})\widetilde{\boldsymbol{\Omega}}$ with $\Phi := \|\boldsymbol{\Omega}\|_2$, we prove by induction

$$\widehat{\mathbf{v}}^{2l+1} = \begin{pmatrix} (-\Phi^2)^l\,\widetilde{\boldsymbol{\Omega}} & \mathbf{0}_{3\times 3} \\ (-\Phi^2)^l\,\widetilde{\mathbf{U}} - 2l(-\Phi^2)^{l-1}(\boldsymbol{\Omega}^\top\mathbf{U})\widetilde{\boldsymbol{\Omega}} & (-\Phi^2)^l\,\widetilde{\boldsymbol{\Omega}} \end{pmatrix}$$

and

$$\widehat{\mathbf{v}}^{2l+2} = \begin{pmatrix} (-\Phi^2)^l\,\widetilde{\boldsymbol{\Omega}}^2 & \mathbf{0}_{3\times 3} \\ (-\Phi^2)^l\,(\widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Omega}} + \widetilde{\boldsymbol{\Omega}}\widetilde{\mathbf{U}}) - 2l(-\Phi^2)^{l-1}(\boldsymbol{\Omega}^\top\mathbf{U})\widetilde{\boldsymbol{\Omega}}^2 & (-\Phi^2)^l\,\widetilde{\boldsymbol{\Omega}}^2 \end{pmatrix}$$

for all $l \geq 0$ and get the tangent operator

$$\mathbf{T}_{\mathrm{SE}(3)}(\boldsymbol{\Omega}) = \begin{pmatrix} \mathbf{T}_{\mathrm{SO}(3)}(\boldsymbol{\Omega}) & \mathbf{0}_{3\times 3} \\ \mathbf{S}_{\mathrm{SE}(3)}(\boldsymbol{\Omega},\mathbf{U}) & \mathbf{T}_{\mathrm{SO}(3)}(\boldsymbol{\Omega}) \end{pmatrix} \tag{35}$$

with $\mathbf{S}_{\mathrm{SE}(3)}(\mathbf{0},\mathbf{U}) = -\widetilde{\mathbf{U}}/2$ and

$$\begin{aligned}\mathbf{S}_{\mathrm{SE}(3)}(\boldsymbol{\Omega},\mathbf{U}) = \frac{1}{\Phi^2}\Big(&-(1-\cos\Phi)\widetilde{\mathbf{U}} + \big(1 - \frac{\sin\Phi}{\Phi}\big)(\widetilde{\mathbf{U}}\widetilde{\boldsymbol{\Omega}} + \widetilde{\boldsymbol{\Omega}}\widetilde{\mathbf{U}}) + \\ &+ \big(2\,\frac{1-\cos\Phi}{\Phi^2} - \frac{\sin\Phi}{\Phi}\big)(\boldsymbol{\Omega}^\top\mathbf{U})\widetilde{\boldsymbol{\Omega}} + \\ &+ \frac{1}{\Phi^2}\big(1 - \cos\Phi - 3\,(1 - \frac{\sin\Phi}{\Phi})\big)(\boldsymbol{\Omega}^\top\mathbf{U})\widetilde{\boldsymbol{\Omega}}^2\Big)\end{aligned}$$

if $\boldsymbol{\Omega} \neq \mathbf{0}$, see (Brüls et al. 2011 and Sonneville et al. 2014, Appendix A).

(c) If $\mathbb{R}^k$ with the addition is considered as a Lie group, then we get $\widehat{\mathbf{v}} = \mathbf{0}_{k\times k}$ and $\mathbf{T}_{\mathbb{R}^k}(\mathbf{v}) = \mathbf{I}_k$ for any vector $\mathbf{v} \in \mathbb{R}^k$ since the group operation is commutative.

(d) Similar to the discussion in Example 2.1(c), we observe for direct products like $\mathrm{SO}(3) \times \mathbb{R}^3$ that the matrix $\widehat{\mathbf{v}}$ and the tangent operator $\mathbf{T}(\mathbf{v})$ are block-diagonal. In $(\mathrm{SO}(3) \times \mathbb{R}^3)^N$ and $(\mathrm{SE}(3))^N$, the tangent operators are given by

$$\mathbf{T}_{G^N}\big((\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N)\big) = \mathrm{blockdiag}_{1\leq i\leq N}\,\mathbf{T}_G(\mathbf{v}_i) \in \mathbb{R}^{6N\times 6N}$$

with $G = \mathrm{SO}(3) \times \mathbb{R}^3$ and $G = \mathrm{SE}(3)$, respectively.

3 Generalized-α Lie Group Time Integration

The time integration of the equations of motion (15) by Lie group methods is based on the observation that (15a) implies

$$q(t+h) = q(t) \circ \exp\big(h\widetilde{\mathbf{v}}(t) + \frac{h^2}{2}\dot{\widetilde{\mathbf{v}}}(t) + \mathcal{O}(h^3)\big)\,, \quad (\,h \to 0\,)\,. \tag{36}$$

In Sect. 3.1, a generalized-α Lie group method for the index-3 formulation (15) is introduced. In Sect. 3.2, we recall some well-known facts about order, stability and "overshooting" of generalized-α methods in linear spaces. For the heavy top benchmark problem, second-order convergence of the Lie group integrator and an order reduction phenomenon in the transient phase may be observed numerically (Sect. 3.3). In Sect. 3.4, we show that the error constant of the first-order error term may be reduced drastically by an analytical index reduction before time discretization. Implementation aspects and the discretization errors in hidden constraints are studied in Sects. 3.5 and 3.6.

3.1 The Lie Group Time Integration Method

As proposed by Brüls and Cardona (2010), we consider a generalized-α method for the index-3 formulation (15) of the equations of motion that updates the numerical solution $(q_n, \mathbf{v}_n, \mathbf{a}_n, \boldsymbol{\lambda}_n)$ in a time step $t_n \to t_n + h$ of step size h according to

$$q_{n+1} = q_n \circ \exp(h\widetilde{\boldsymbol{\Delta}\mathbf{q}}_n)\,, \tag{37a}$$

$$\boldsymbol{\Delta}\mathbf{q}_n = \mathbf{v}_n + (0.5-\beta)h\mathbf{a}_n + \beta h\mathbf{a}_{n+1}\,, \tag{37b}$$

$$\mathbf{v}_{n+1} = \mathbf{v}_n + (1-\gamma)h\mathbf{a}_n + \gamma h\mathbf{a}_{n+1}\,, \tag{37c}$$

$$(1-\alpha_m)\mathbf{a}_{n+1} + \alpha_m\mathbf{a}_n = (1-\alpha_f)\dot{\mathbf{v}}_{n+1} + \alpha_f\dot{\mathbf{v}}_n \tag{37d}$$

with vectors $\dot{\mathbf{v}}_{n+1}$, $\boldsymbol{\lambda}_{n+1}$ satisfying the equilibrium conditions

$$\mathbf{M}(q_{n+1})\dot{\mathbf{v}}_{n+1} = -\mathbf{g}(q_{n+1}, \mathbf{v}_{n+1}, t_{n+1}) - \mathbf{B}^\top(q_{n+1})\boldsymbol{\lambda}_{n+1}\,, \tag{37e}$$

$$\boldsymbol{\Phi}(q_{n+1}) = \mathbf{0}\,. \tag{37f}$$

The term *generalized-α method* refers to the coefficients α_m, α_f in the update formula (37d) for the acceleration like variables $\mathbf{a}_n$. These auxiliary variables $\mathbf{a}_n$ were introduced by Chung and Hulbert (1993) who studied the time integration of unconstrained linear systems in linear spaces and proposed a one-parametric set of algorithmic parameters α_m, α_f, β and γ that may be considered as a quasi-standard for this type of methods, see Sect. 3.2 below.

Method (37) is initialized with starting values $q_0 \in G$ and $\mathbf{v}_0 \in \mathbb{R}^k$ that approximate the (consistent) initial values $q(t_0)$, $\mathbf{v}(t_0)$ in (15). The starting values $\dot{\mathbf{v}}_0$, $\mathbf{a}_0$ at acceleration level are approximations of $\dot{\mathbf{v}}(t_0) \in \mathbb{R}^k$, see (19). The convergence analysis in Sect. 4 below will show that the starting values need to be selected carefully to guarantee second order convergence in all solution components and to avoid spurious oscillations in the numerical solution $\boldsymbol{\lambda}_n$.

In practical applications, variable step size implementations with error control are expected to be superior to methods with fixed time step size h. For constrained systems in linear configuration spaces, a step size control algorithm for generalized-α methods with $\alpha_m = 0$ (*HHT-methods*, see Hilber et al. 1977) was developed in

(Géradin and Cardona 2001, Chap. 11). For this problem class, Jay and Negrut (2007) proposed a linear update formula for the auxiliary variables $\mathbf{a}_n$ to compensate a first-order error term resulting from a step size change at $t = t_n$.

An alternative approach is based on the elimination of these variables $\mathbf{a}_n$ in the multi-step representation of generalized-α methods according to Erlicher et al. (2002). Here, the algorithmic parameters α_m, α_f, β and γ have to be updated in each time step considering the step size ratio h_{n+1}/h_n, see (Brüls and Arnold 2008).

There is no straightforward extension of the results of Erlicher et al. (2002) from linear configuration spaces to the Lie group setting of the present paper. Furthermore, the analysis of the error propagation in time integration is simplified substantially if the time step size h is *fixed* for all time steps. For both reasons, the convergence analysis for generalized-α Lie group integrators (37) with *variable* time step size h_n will be a topic of future research that is beyond the scope of the present paper.

3.2 The Generalized-α Method in Linear Spaces

For linear configuration spaces $G = \mathbb{R}^k$ and unconstrained systems (15) with constant mass matrix $\mathbf{M}$, the generalized-α Lie group method (37) coincides with the "classical" generalized-α method that goes back to the work of Chung and Hulbert (1993). Multiplying (37d) by the (constant) mass matrix $\mathbf{M}$ and eliminating vectors $\mathbf{\Delta q}_n$ and $\dot{\mathbf{v}}_{n+1}$, we get

$$\mathbf{q}_{n+1} = \mathbf{q}_n + h\mathbf{v}_n + (0.5-\beta)h^2\mathbf{a}_n + \beta h^2\mathbf{a}_{n+1}\,, \tag{38a}$$

$$\mathbf{v}_{n+1} = \mathbf{v}_n + (1-\gamma)h\mathbf{a}_n + \gamma h\mathbf{a}_{n+1}\,, \tag{38b}$$

$$\mathbf{0} = (1-\alpha_m)\mathbf{M}\mathbf{a}_{n+1} + \alpha_m\mathbf{M}\mathbf{a}_n + (1-\alpha_f)\mathbf{g}_{n+1} + \alpha_f\mathbf{g}_n \tag{38c}$$

with $\mathbf{g}_n := \mathbf{g}(\mathbf{q}_n, \mathbf{v}_n, t_n)$ and vectors $\mathbf{q}_n, \mathbf{q}_{n+1} \in \mathbb{R}^k$ that are typeset in boldface font to indicate the *linear* structure of the configuration space.

For a local error analysis, we suppose that $\mathbf{a}_n$ approximates $\dot{\mathbf{v}}(t_n + \Delta_\alpha h)$ with a fixed offset $\Delta_\alpha \in \mathbb{R}$, see (Jay and Negrut 2008, Sect. 2), and substitute in (38) the numerical solution vectors $\mathbf{q}_n$, $\mathbf{v}_n$, $\mathbf{a}_n$, $\mathbf{g}_n$ by $\mathbf{q}(t_n)$, $\mathbf{v}(t_n)$, $\dot{\mathbf{v}}(t_n + \Delta_\alpha h)$ and $-\mathbf{M}\dot{\mathbf{v}}(t_n)$, respectively. The resulting residuals define local truncation errors $\mathbf{l}_n^{\mathbf{q}}$, $\mathbf{l}_n^{\mathbf{v}}$ and $\mathbf{l}_n^{\mathbf{a}}$:

$$\begin{aligned}\mathbf{q}(t_{n+1}) = \mathbf{q}(t_n) + h\mathbf{v}(t_n) + (0.5-\beta)h^2\dot{\mathbf{v}}(t_n + \Delta_\alpha h) + \\ + \beta h^2\dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h) + \mathbf{l}_n^{\mathbf{q}}\,,\end{aligned} \tag{39a}$$

$$\mathbf{v}(t_{n+1}) = \mathbf{v}(t_n) + (1-\gamma)h\dot{\mathbf{v}}(t_n + \Delta_\alpha h) + \gamma h\dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h) + \mathbf{l}_n^{\mathbf{v}}\,, \tag{39b}$$

$$\begin{aligned}\mathbf{M}\mathbf{l}_n^{\mathbf{a}} = (1-\alpha_m)\mathbf{M}\dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h) + \alpha_m\mathbf{M}\dot{\mathbf{v}}(t_n + \Delta_\alpha h) - \\ - (1-\alpha_f)\mathbf{M}\dot{\mathbf{v}}(t_{n+1}) - \alpha_f\mathbf{M}\dot{\mathbf{v}}(t_n)\,.\end{aligned} \tag{39c}$$

For sufficiently smooth solutions $\mathbf{q}(t)$, the local truncation errors in (39) may be analysed by Taylor expansion of functions $\mathbf{q}(t)$, $\mathbf{v}(t)$ and $\dot{\mathbf{v}}(t)$ at $t = t_n$:

$$\mathbf{l}_n^{\mathbf{q}} = C_q h^3 \dddot{\mathbf{v}}(t_n) + \mathcal{O}(h^4) \ \text{ with } \ C_q := (1 - 6\beta - 3\Delta_\alpha)/6\,, \tag{40a}$$

$$\mathbf{l}_n^{\mathbf{v}} = (0.5 - \Delta_\alpha - \gamma) h^2 \ddot{\mathbf{v}}(t_n) + \mathcal{O}(h^3)\,, \tag{40b}$$

$$\mathbf{l}_n^{\mathbf{a}} = \big(\Delta_\alpha - (\alpha_m - \alpha_f)\big) h \ddot{\mathbf{v}}(t_n) + \mathcal{O}(h^2)\,. \tag{40c}$$

We get local truncation errors $\mathbf{l}_n^{\mathbf{q}} = \mathcal{O}(h^3)$, $\mathbf{l}_n^{\mathbf{v}} = \mathcal{O}(h^3)$ and $\mathbf{l}_n^{\mathbf{a}} = \mathcal{O}(h^2)$ if the algorithmic parameters satisfy the order condition

$$\gamma = 0.5 - \Delta_\alpha \ \text{ with } \ \Delta_\alpha := \alpha_m - \alpha_f\,. \tag{41}$$

Chung and Hulbert (1993) studied the scalar test equation $\ddot{q} + \omega^2 q = 0$ with periodic analytical solutions $q(t) = c_1 \sin \omega t + c_2 \cos \omega t$ and observed that (38) results in a frequency-dependent linear mapping $(q_n, v_n, a_n) \mapsto (q_{n+1}, v_{n+1}, a_{n+1})$. Scaling the update formulae (38a, 38b) by factors $1/h^2$ and $1/h$, respectively, we get

$$\underbrace{\begin{pmatrix} \frac{1}{(h\omega)^2} & 0 & -\beta \\ 0 & 1 & -\gamma \\ 1-\alpha_f & 0 & 1-\alpha_m \end{pmatrix}}_{=:\,\mathbf{T}_{h\omega}^{+}} \underbrace{\begin{pmatrix} \omega^2 q_{n+1} \\ \frac{1}{h} v_{n+1} \\ a_{n+1} \end{pmatrix}}_{=\,\mathbf{z}_{n+1}} = \underbrace{\begin{pmatrix} \frac{1}{(h\omega)^2} & 1 & 0.5-\beta \\ 0 & 1 & 1-\gamma \\ -\alpha_f & 0 & -\alpha_m \end{pmatrix}}_{=:\,\mathbf{T}_{h\omega}^{0}} \underbrace{\begin{pmatrix} \omega^2 q_n \\ \frac{1}{h} v_n \\ a_n \end{pmatrix}}_{=:\,\mathbf{z}_n}.$$

Recursive application yields $\mathbf{z}_n = \mathbf{T}_{h\omega}^n \mathbf{z}_0$ with $\mathbf{T}_{h\omega} := (\mathbf{T}_{h\omega}^{+})^{-1} \mathbf{T}_{h\omega}^{0}$. Therefore, the stability and (numerical) damping properties of the generalized-α method (38) applied to $\ddot{q} + \omega^2 q = 0$ may be characterized by an eigenvalue analysis of $\mathbf{T}_{h\omega} \in \mathbb{R}^{3\times 3}$. Chung and Hulbert (1993) propose to choose a user-defined parameter $\rho_\infty \in [0, 1]$ to characterize the numerical damping properties in the limit case $h\omega \to \infty$. They show that the algorithmic parameters $\alpha_m, \alpha_f, \beta$ and γ may be defined such that the order condition (41) is satisfied and the spectral radius $\varrho(\mathbf{T}_{h\omega})$ is monotonically decreasing for $h\omega \in (0, +\infty)$ with $\lim_{h\omega\to 0} \varrho(\mathbf{T}_{h\omega}) = 1$ and $\varrho(\mathbf{T}_\infty) = \rho_\infty$:

$$\alpha_m = \frac{2\rho_\infty - 1}{\rho_\infty + 1}\,,\ \alpha_f = \frac{\rho_\infty}{\rho_\infty + 1}\,,\ \gamma = \frac{1}{2} + \alpha_f - \alpha_m\,,\ \beta = \frac{1}{4}(\gamma + \frac{1}{2})^2\,. \tag{42}$$

For these parameters, all three eigenvalues of $\mathbf{T}_{h\omega} = \mathbf{T}_{h\omega}(\rho_\infty)$ coincide in the limit case $h\omega \to \infty$ and the Jordan canonical form of $\mathbf{T}_\infty(\rho_\infty) \in \mathbb{R}^{3\times 3}$ consists of a single 3×3 Jordan block for the eigenvalue $\mu := -\rho_\infty$, i.e., $\mathbf{T}_\infty(\rho_\infty) = \mathbf{X}(\mu)\mathbf{J}(\mu)\mathbf{X}^{-1}(\mu)$ with

$$\mathbf{J}(\mu) := \begin{pmatrix} \mu & 1 & 0 \\ 0 & \mu & 1 \\ 0 & 0 & \mu \end{pmatrix},\quad \mathbf{X}(\mu) := \begin{pmatrix} 1-\mu^2 & -(2+\mu) & 0 \\ 0 & \frac{1}{2}\,\frac{1+\mu}{1-\mu} & -\frac{1}{(1-\mu)^2} \\ 0 & 1 & 0 \end{pmatrix}.$$

With algorithmic parameters α_m, α_f, β and γ according to (42) and a damping parameter $\rho_\infty < 1$, the linear stability of the generalized-α method (38) is always guaranteed. For the test equation $\ddot{q} + \omega^2 q = 0$, the numerical solution $(q_n, v_n, a_n)^\top$

will finally be damped out for any starting values q_0, v_0, a_0 since $\mathbf{z}_n = \mathbf{T}^n_{h\omega}(\rho_\infty)\mathbf{z}_0$ and $\lim_{n\to\infty} \mathbf{T}^n_{h\omega}(\rho_\infty) = \mathbf{0}$ because $\varrho(\mathbf{T}_{h\omega}(\rho_\infty)) < 1$, ($h\omega \in (0,\infty)$).

In a transient phase, however, $\|\mathbf{z}_n\|$ may be much larger than $\|\mathbf{z}_0\|$ since $\|\mathbf{T}^n\|$ may be much larger than $(\varrho(\mathbf{T}))^n$ for matrices that are not diagonalisable (*non-normal* matrices). Typical values are $\max_n \|\mathbf{T}^n\|_2 = \|\mathbf{T}^3\|_2 = 7.4$ for $\mathbf{T} = \mathbf{T}_\infty(\rho_\infty)$ with $\rho_\infty = 0.6$ and $\max_n \|\mathbf{T}^n\|_2 = \|\mathbf{T}^{14}\|_2 = 34.3$ for $\mathbf{T} = \mathbf{T}_\infty(\rho_\infty)$ with $\rho_\infty = 0.9$. In structural dynamics, this phenomenon is called *overshooting* since $|q_n|$ may grow rapidly in a transient phase before the numerical dissipation results finally in $\lim_{n\to 0} q_n = 0$. Overshooting is a well-known problem of unconditionally stable Newmark-type methods with second-order accuracy (Hilber and Hughes 1978) and may be a motivation to prefer first-order accurate Newmark integrators in industrial multibody system simulation (Sanborn et al. 2014).

In the quantitative error analysis, we denote the global errors of the generalized-α method in linear spaces by $\mathbf{e}^{(\bullet)}_n$ with $(\bullet)(t_n) = (\bullet)_n + \mathbf{e}^{(\bullet)}_n$. For the auxiliary vectors $\mathbf{a}_n$ that do not have a corresponding component of the analytical solution, we take into account the offset parameter Δ_α from (41) and define the global error $\mathbf{e}^{\mathbf{a}}_n$ by $\dot{\mathbf{v}}(t_n + \Delta_\alpha h) = \mathbf{a}_n + \mathbf{e}^{\mathbf{a}}_n$. For the scalar test equation $\ddot{q} + \omega^2 q = 0$, these global errors as well as the local errors l^q_n, l^v_n, l^a_n are scalar quantities and $\mathbf{T}^+_{h\omega}\mathbf{z}_{n+1} = \mathbf{T}^0_{h\omega}\mathbf{z}_n$ implies

$$\mathbf{T}^+_{h\omega}\begin{pmatrix} \omega^2 e^q_{n+1} \\ \dfrac{1}{h}e^v_{n+1} \\ e^a_{n+1} \end{pmatrix} = \mathbf{T}^0_{h\omega}\begin{pmatrix} \omega^2 e^q_n \\ \dfrac{1}{h}e^v_n \\ e^a_n \end{pmatrix} + \begin{pmatrix} \dfrac{1}{h^2}l^q_n \\ \dfrac{1}{h}l^v_n \\ l^a_n \end{pmatrix}, \tag{43}$$

see (39). As before, the first and second row are scaled by $1/h^2$ and $1/h$, respectively. The resulting first-order error term $l^q_n/h^2 = C_q h\dddot{v}(t_n) + \mathcal{O}(h^2)$ may strongly affect the result accuracy.

This order reduction phenomenon is known from the convergence analysis for the application of Newmark-type methods to constrained mechanical systems in linear configuration spaces, see (Cardona and Géradin 1994). In the limit case $\omega \to \infty$, the transient solution behaviour is dominated by an oscillating first-order error term that is finally damped out by numerical dissipation. To study this qualitative solution behaviour in full detail, we introduce a new variable $\lambda := \omega^2 q$ and rewrite the test equation as a singular singularly perturbed problem with perturbation parameter $\varepsilon := 1/\omega$, see (Lubich 1993):

$$\ddot{q} + \omega^2 q = 0 \quad \Leftrightarrow \quad \left.\begin{aligned} \ddot{q} &= -\lambda \\ \frac{1}{\omega^2}\lambda &= q \end{aligned}\right\} \tag{44}$$

The corresponding reduced system ($\varepsilon = 0$, i.e., $\omega \to \infty$) is a constrained system (15) with $G = \mathbb{R}$ and $k = m = 1$:

$$\left.\begin{aligned} \ddot{q} &= -\lambda \\ 0 &= q \end{aligned}\right\} \tag{45}$$

With the notation $\lambda_n := \omega^2 q_n$, the generalized-$\alpha$ method (38) for the singularly perturbed system (44) converges for $\omega \to \infty$ to the generalized-α method (37) for the constrained system (45) and we get in (43) both for finite frequencies ω and in the limit case $\omega \to \infty$:

$$\mathbf{T}^{+}_{h\omega}\mathbf{e}^{\mathbf{r}}_{n+1} = \mathbf{T}^{0}_{h\omega}\mathbf{e}^{\mathbf{r}}_{n} + \mathbf{l}^{\mathbf{r}}_{n} \tag{46}$$

with

$$\mathbf{e}^{\mathbf{r}}_{n} := \begin{pmatrix} e^{\lambda}_{n} \\ r_n \\ e^{a}_{n} \end{pmatrix}, \quad \mathbf{l}^{\mathbf{r}}_{n} := \begin{pmatrix} 0 \\ \dfrac{1}{h} l^{v}_{n} + \dfrac{l^{q}_{n+1} - l^{q}_{n}}{h^2} \\ l^{a}_{n} \end{pmatrix} \tag{47}$$

and

$$r_n := \frac{1}{h}\big(e^{v}_{n} + \frac{1}{h} l^{q}_{n}\big) = \frac{1}{h}\big(e^{v}_{n} + C_q h^2 \ddot{v}(t_n)\big) + \mathcal{O}(h^2)\,. \tag{48}$$

The error recursion in terms of e^{λ}_{n}, r_n and e^{a}_{n} provides the basis for a detailed convergence analysis:

Theorem 3.1 *Consider the time discretization of the linear test equations (44) and (45) by a generalized-α method with parameters α_m, α_f, β and γ according to (42) for some numerical damping parameter $\rho_\infty \in [0, 1)$.*

(a) The discretization errors are bounded by

$$\|\mathbf{l}^{\mathbf{r}}_{n}\| = \mathcal{O}(h^2)\,, \quad \|\mathbf{e}^{\mathbf{r}}_{n+1} - \mathbf{T}_{h\omega}\mathbf{e}^{\mathbf{r}}_{n}\| = \mathcal{O}(h^2)\,, \tag{49}$$

$$\|\mathbf{e}^{\mathbf{r}}_{n} - \mathbf{T}^{n}_{h\omega}\mathbf{e}^{\mathbf{r}}_{0}\| = \mathcal{O}(h^2) \tag{50}$$

and

$$\|\mathbf{e}^{\mathbf{r}}_{n}\| \le \|\mathbf{T}^{n}_{h\omega}\|\,\|\mathbf{e}^{\mathbf{r}}_{0}\| + \mathcal{O}(h^2)\,. \tag{51}$$

(b) For starting values $\lambda_0 = \lambda(t_0) + \mathcal{O}(h^2)$, $a_0 = \dot{v}(t_0 + \Delta_\alpha h) + \mathcal{O}(h^2)$, we have $\|\mathbf{e}^{\mathbf{r}}_{0}\| = \mathcal{O}(h)$ if $v_0 = v(t_0) + \mathcal{O}(h^2)$. This error estimate may be improved by one power of h perturbing the starting value v_0 such that

$$v_0 = v(t_0) + C_q h^2 \ddot{v}(t_0) + \mathcal{O}(h^3)\,. \tag{52}$$

In that case, we get $\|\mathbf{e}^{\mathbf{r}}_{n}\| = \mathcal{O}(h^2)$, ($n \ge 0$).

Proof (a) Because of

$$l^{q}_{n+1} - l^{q}_{n} = C_q h^3\big(\ddot{v}(t_{n+1}) - \ddot{v}(t_n)\big) + \mathcal{O}(h^4) = \mathcal{O}(h^4)\,,$$

the local error term $\mathbf{l}_n^{\mathbf{r}}$ is of size $\mathcal{O}(h^2)$, see (40), and (49) is a direct consequence of the error recursion (46). The assumptions on parameters $\alpha_m, \alpha_f, \beta$ and γ imply $\varrho(\mathbf{T}_{h\omega}) < 1$ and the existence of a norm $\|\mathbf{T}\|_\rho$ with $\kappa := \|\mathbf{T}_{h\omega}\|_\rho < 1$, see, e.g., (Quarteroni et al. 2000, Sect. 1.11.1). Therefore,

$$\begin{aligned}\|\mathbf{e}_n^{\mathbf{r}} - \mathbf{T}_{h\omega}^n \mathbf{e}_0^{\mathbf{r}}\|_\rho &\le \|\mathbf{e}_n^{\mathbf{r}} - \mathbf{T}_{h\omega}\mathbf{e}_{n-1}^{\mathbf{r}}\|_\rho + \|\mathbf{T}_{h\omega}\mathbf{e}_{n-1}^{\mathbf{r}} - \mathbf{T}_{h\omega}^n \mathbf{e}_0^{\mathbf{r}}\|_\rho \\ &\le \|\mathbf{e}_n^{\mathbf{r}} - \mathbf{T}_{h\omega}\mathbf{e}_{n-1}^{\mathbf{r}}\|_\rho + \|\mathbf{T}_{h\omega}\|_\rho \, \|\mathbf{e}_{n-1}^{\mathbf{r}} - \mathbf{T}_{h\omega}^{n-1}\mathbf{e}_0^{\mathbf{r}}\|_\rho \\ &\le Ch^2 + \kappa \|\mathbf{e}_{n-1}^{\mathbf{r}} - \mathbf{T}_{h\omega}^{n-1}\mathbf{e}_0^{\mathbf{r}}\|_\rho\end{aligned}$$

with an appropriate constant $C > 0$, see (49). Recursive application of this error estimate results in

$$\|\mathbf{e}_n^{\mathbf{r}} - \mathbf{T}_{h\omega}^n \mathbf{e}_0^{\mathbf{r}}\|_\rho \le \sum_{i=0}^{n-1} \kappa^i \, Ch^2 + \kappa^n \|\mathbf{e}_0^{\mathbf{r}} - \mathbf{T}_{h\omega}^0 \mathbf{e}_0^{\mathbf{r}}\|_\rho < \frac{1}{1-\kappa} \, Ch^2$$

and (50) follows from the equivalence of all norms in the finite dimensional space $\mathbb{R}^3$. Error bound (51) is a straightforward consequence of the triangle inequality.

(b) We get $\|\mathbf{e}_0^{\mathbf{r}}\| = |r_0| + \mathcal{O}(h^2)$ and the estimates for $\|\mathbf{e}_0^{\mathbf{r}}\|$ and for $\|\mathbf{e}_n^{\mathbf{r}}\|$, $(\, n > 0\,)$, follow from the definition of r_n, see (48), and from part (a) of the theorem. □

The most natural choice of starting values $\lambda_0 := \lambda(t_0)$, $v_0 := v(t_0)$, $a_0 := \dot{v}(t_0 + \Delta_\alpha h)$ yields $\mathbf{e}_0^{\mathbf{r}} = (\, 0, \, r_0, \, 0\,)^\top$ with $r_0 = C_q h \ddot{v}(t_0) + \mathcal{O}(h^2)$, see (48). In error estimate (50), we obtain for $\ddot{v}(t_0) \neq 0$ a first-order error term being amplified by matrix-valued factors $\mathbf{T}_{h\omega}^n$ that are well known from the analysis of the "overshoot" phenomenon by Hilber and Hughes (1978). In the limit case $h\omega \to \infty$, this term may be studied in more detail using the Jordan canonical form of $\mathbf{T}_\infty$, see (Cardona and Géradin 1989, 1994). We get

$$\mathbf{T}_\infty^n \, \mathbf{e}_0^{\mathbf{r}} = \mathbf{X}(-\rho_\infty) \, \mathbf{J}^n(-\rho_\infty) \, \mathbf{X}^{-1}(-\rho_\infty) \, \mathbf{e}_0^{\mathbf{r}}$$

with the Jordan block $\mathbf{J}(-\rho_\infty) \in \mathbb{R}^{3\times 3}$. It may be verified by induction that the non-zero elements of $\mathbf{J}^n(-\rho_\infty)$ are given by $(-\rho_\infty)^n$, $n(-\rho_\infty)^{n-1}$ and $n(n-1)(-\rho_\infty)^{n-2}/2$. Straightforward computations show that the global error e_n^λ (that coincides up to a term of size $\mathcal{O}(h^2)$ with the first component of $\mathbf{T}_\infty^n \mathbf{e}_0^{\mathbf{r}}$) satisfies $e_n^\lambda = c_n h \ddot{v}(t_0) + \mathcal{O}(h^2)$ with

$$c_n := C_q(1+\rho_\infty)^2 \Big(\frac{n}{2}(n-1)(\rho_\infty^2 - 1)(-\rho_\infty)^{n-2} + n(2-\rho_\infty)(-\rho_\infty)^{n-1} \Big) \, . \tag{53}$$

After a transient phase, the first-order error term $c_n h \ddot{v}(t_0)$ is damped out since $\lim_{n\to\infty} c_n = 0$ for any $\rho_\infty \in [0, 1)$. In the transient phase, however, the error constants c_n may become very large with maximum absolute values of size $|c_3| = 6.8$ for $\rho_\infty = 0.6$, $|c_{15}| = 31.9$ for $\rho_\infty = 0.9$ and $|c_{161}| = 334.3$ for $\rho_\infty = 0.99$.

For the test equation (45) itself, this error analysis has not much practical relevance since $q(t) \equiv 0$ implies $\ddot{v}(t) \equiv 0$ and $\mathbf{e}_0^{\mathbf{r}} = \mathbf{0}$ for exact starting values $\lambda_0 = \lambda(t_0) = 0$,

$v_0 = v(t_0) = 0$, $a_0 = \dot{v}(t_0 + \Delta_\alpha h) = 0$. Substituting the trivial constraint $q = 0$ by a rheonomic constraint $q(t) = t^3/6$, we may construct, however, a slightly more complex test problem with non-vanishing first-order error term $r_0 = C_q h$ since $l_n^q = C_q h^3 \dddot{v}(t_n) = C_q h^3$ and the local truncation errors l_n^v, l_n^a vanish identically. For this test problem, the global error in λ really suffers from order reduction since $e_n^\lambda = c_n h$.

The convergence analysis for generalized-α methods shows that this order reduction phenomenon is typical for the initialization of method (37) with exact starting values $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}(t_0)$, $\mathbf{v}_0 = \mathbf{v}(t_0)$ and $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0 + \Delta_\alpha h)$, see (Arnold et al. 2015) and Sect. 4 below. For linear configuration spaces ($G = \mathbb{R}^k$), the global error in $\boldsymbol{\lambda}$ is bounded by

$$[\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top]\big(\mathbf{q}(t_n)\big)\,\mathbf{e}_n^{\boldsymbol{\lambda}} = c_n h\, \mathbf{B}\big(\mathbf{q}(t_0)\big)\ddot{\mathbf{v}}(t_0) + \mathcal{O}(h^2) \tag{54}$$

with the error constants c_n being defined in (53). The undesired first-order error term is nicely illustrated by numerical test results for the mathematical pendulum, see (Arnold et al. 2015, Sect. 2.3):

Example 3.2 Consider a mathematical pendulum of mass m and length l in Cartesian coordinates $\mathbf{q} = (x, y)^\top$ with constraint $(x^2 + y^2 - l^2)/2 = 0$, see (15c). In (15), we have $\mathbf{M} = m\mathbf{I}_2$, $\mathbf{g} = (\,0\,,\,g\,)^\top$ with $m = l = 1$, $g = 9.81$ (here and in the following, all physical units are omitted). We fix the total energy $E = m(\dot{x}_0^2 + \dot{y}_0^2)/2 + mgy_0$ to $E = m/2 - mgl$ and determine the consistent initial values x_0, y_0, $\dot{x}_0$, $\dot{y}_0$ and λ_0 by the initial deviation x_0 from the equilibrium position.

Method (37) is applied with algorithmic parameters according to (42) and damping parameter $\rho_\infty = 0.9$. The starting values are set to $\mathbf{q}_0 := (x_0, y_0)^\top$, $\mathbf{v}_0 := (\dot{x}_0, \dot{y}_0)^\top$ and $\dot{\mathbf{v}}_0 := (\ddot{x}_0, \ddot{y}_0)^\top$ with accelerations $\ddot{x}_0$, $\ddot{y}_0$ that are obtained from evaluating the equations of motion for the consistent initial values x_0, y_0, $\dot{x}_0$, $\dot{y}_0$, λ_0. The acceleration like variables $\mathbf{a}_n$ are initialized with $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0) + \Delta_\alpha h \ddot{\mathbf{v}}(t_0) + \mathcal{O}(h^2) = \dot{\mathbf{v}}(t_0 + \Delta_\alpha h) + \mathcal{O}(h^2)$ using the starting value $\dot{\mathbf{v}}_0 = \dot{\mathbf{v}}(t_0)$ and a difference approximation of $\ddot{\mathbf{v}}(t_0)$.

Figure 5 shows on a short time interval the global error in λ for initial values $x_0 = 0$ (marked by dots) and $x_0 = 0.2$ (marked by "+") for two different step sizes h. If we start in the equilibrium position, the error is very small but for $x_0 = 0.2$, the oscillating

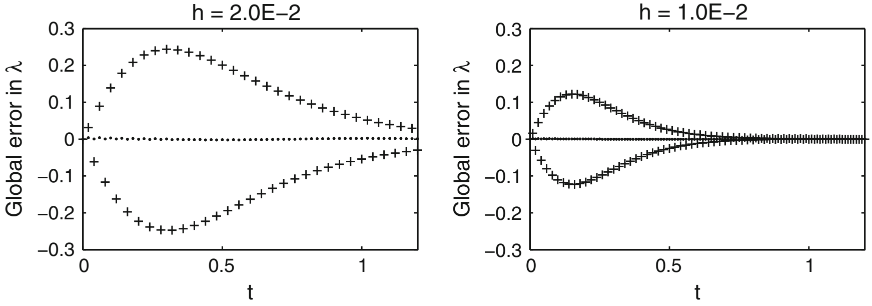


Fig. 5 Mathematical pendulum: Global error in λ for $x_0 = 0$ ("·") and $x_0 = 0.2$ ("+")

error in λ reaches a maximum amplitude of 2.48×10^{-1} for $h = 2.0 \times 10^{-2}$ and 1.23×10^{-1} for $h = 1.0 \times 10^{-2}$. After about 100 time steps these transient errors are damped out.

The numerical results in Fig. 5 show that in the transient phase the generalized-α method (37) may suffer from spurious oscillations of amplitude $\mathcal{O}(h)$. According to (54), this first-order error term is given by $c_n h\,\mathbf{B}\big(\mathbf{q}(t_0)\big)\ddot{\mathbf{v}}(t_0)$ with $\mathbf{B}\big(\mathbf{q}(t_0)\big)\ddot{\mathbf{v}}(t_0) = -3gx_0\dot{x}_0/y_0$. Therefore, the spurious oscillations and the order reduction disappear if we start at the equilibrium position $x_0 = 0$. Reducing the damping parameter ρ_∞ in (42), the oscillations are damped out more rapidly but may still be observed.

3.3 Numerical Tests for the Heavy Top Benchmark Problem

In the present section, we study the convergence behaviour of the generalized-α Lie group integrator (37) numerically. We use algorithmic parameters according to (42) with the numerical damping parameter $\rho_\infty = 0.9$ and apply (37) to the equations of motion (21), (22) of the heavy top benchmark problem in configuration spaces $G = \mathrm{SO}(3) \times \mathbb{R}^3$ and $G = \mathrm{SE}(3)$, respectively. Initial values $q(t_0)$, $\mathbf{v}(t_0)$ are given in Sect. 2.4. In the numerical tests, the integrator was initialized with starting values $q_0 := q(t_0)$, $\mathbf{v}_0 := \mathbf{v}(t_0)$, $\dot{\mathbf{v}}_0 := \dot{\mathbf{v}}(t_0)$ and $\mathbf{a}_0 := \dot{\mathbf{v}}(t_0)$ with $\dot{\mathbf{v}}(t_0)$ denoting the consistent acceleration vector being defined in (19).

In Fig. 6, the asymptotic behaviour of the global errors in q_n, $\mathbf{v}_n$ and $\boldsymbol{\lambda}_n$ for $h \to 0$ is visualized in terms of the maximum $\max_n \|\mathbf{e}_n^{(\bullet)}\| / \|(\bullet)_n\|$ of the norm of relative errors in the time interval $[t_0, t_{\mathrm{end}}] = [0, 1]$. Here, the numerical solutions for $h = 1.25 \times 10^{-4}$, $h = 2.5 \times 10^{-4}$, $h = 5.0 \times 10^{-4}, \ldots, h = 4.0 \times 10^{-3}$ are compared to a reference solution that has been obtained numerically with the very small time step size $h = 2.5 \times 10^{-5}$. In double logarithmic scale, the plots of global errors in q_n and $\mathbf{v}_n$ are straight lines of slope $+2$ (for both configuration spaces). These numerical test results indicate second-order convergence for components q and $\mathbf{v}$.

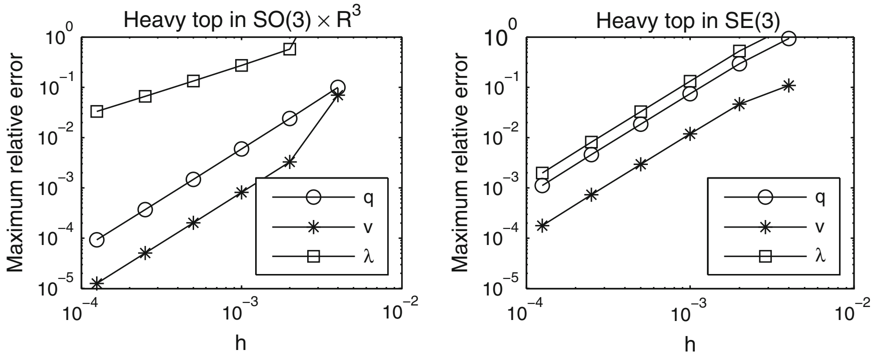


Fig. 6 Heavy top benchmark (index-3 formulation): Global error of integrator (37) versus h for $t \in [0, 1]$. *Left plot* $\mathrm{SO}(3) \times \mathbb{R}^3$, *right plot* $\mathrm{SE}(3)$

The error constants depend on model parameters, initial values and configuration space. With the test setup of Sect. 2.4, the velocity components $\mathbf{v}(t)$ vary much more rapidly for $G = \mathrm{SE}(3)$ than for $G = \mathrm{SO}(3) \times \mathbb{R}^3$, see Fig. 4. This might explain the substantially larger error constants for q_n and $\mathbf{v}_n$ in the right plot of Fig. 6. For other setups, much smaller error constants have been observed for the configuration space SE(3), see, e.g., the numerical test results of Brüls et al. (2011) for a slowly rotating top with an initial angular velocity $\boldsymbol{\Omega}(0)$ that has been reduced by a factor of 100.

Note, that Fig. 6 shows the norm of *relative* errors. The rather large nominal values of $\mathbf{v}(t)$ with $\|\boldsymbol{\Omega}(0)\| \approx 150.0$ result systematically in relative errors that have a substantially smaller norm than the ones in the position coordinates $q(t)$.

For the Lagrange multipliers $\boldsymbol{\lambda}(t)$, we observe order reduction since slope $+1$ of the curve for the global errors in $\boldsymbol{\lambda}_n$ in the left plot of Fig. 6 indicates first-order convergence. The test results for $G = \mathrm{SE}(3)$ in the right plot of Fig. 6 are qualitatively different from the ones in the left plot since they indicate second-order convergence for *all* solution components. A formal proof of this numerically observed convergence behaviour will be given in Theorem 4.18 and Example 4.19 below.

Guided by the test results for the mathematical pendulum in Example 3.2, we expect that the order reduction phenomenon might affect the numerical solution only in a transient phase and the first-order error terms in $\boldsymbol{\lambda}_n$ are finally damped out by numerical dissipation. This is nicely illustrated by Fig. 7 that shows the numerical solution $\lambda_{n,1}$ for $t \in [0, 0.1]$ and two different time step sizes. In the configuration space $G = \mathrm{SO}(3) \times \mathbb{R}^3$ (solid lines), spurious oscillations are observed that are damped out after about 50 time steps and have a maximum amplitude that depends linearly on h. Beyond this transient phase, the results coincide up to plot accuracy with the dashed lines showing simulation results for the configuration space $G = \mathrm{SE}(3)$ that do not suffer from order reduction.

Neglecting the transient behaviour, we observe for both Lie group formulations second-order convergence in all solution components, see Fig. 8 that shows the maximum of the norm of global errors in time interval [0.5, 1], i.e., beyond the transient phase.

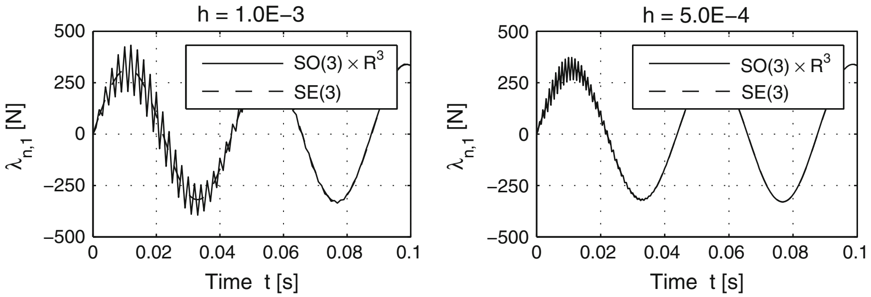


Fig. 7 Heavy top benchmark (index-3 formulation, $G = \mathrm{SO}(3) \times \mathbb{R}^3$ and $G = \mathrm{SE}(3)$): Numerical solution of Lagrange multiplier $\lambda_{n,1}$. *Left plot* $h = 1.0 \times 10^{-3}$, *right plot* $h = 5.0 \times 10^{-4}$

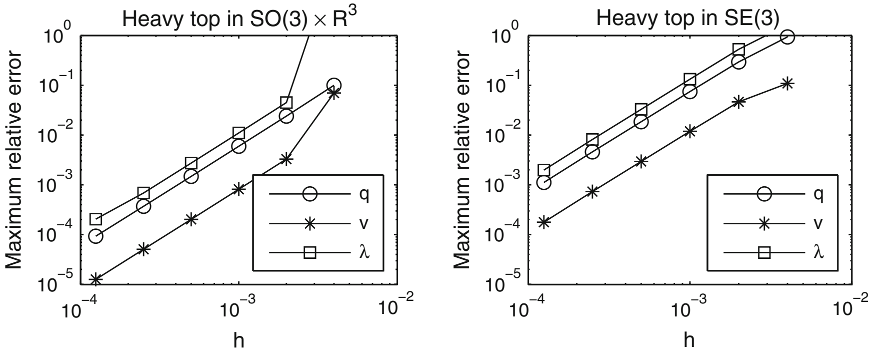


Fig. 8 Heavy top benchmark (index-3 formulation): Global error of integrator (37) versus h for $t \in [0.5, 1]$. *Left plot* $\mathrm{SO}(3) \times \mathbb{R}^3$, *right plot* $\mathrm{SE}(3)$

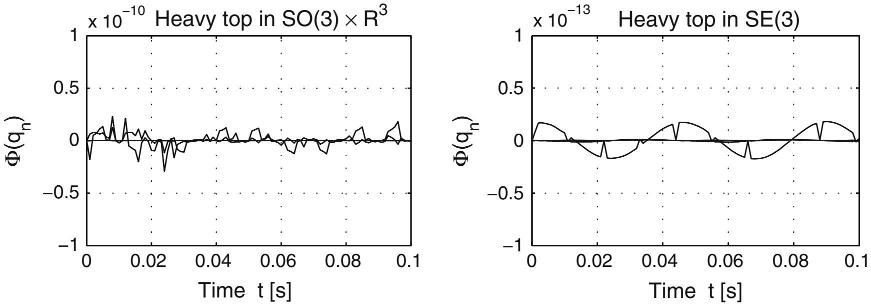


Fig. 9 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, index-3 formulation): Residuals in constraints (15c). *Left plot* $\mathrm{SO}(3) \times \mathbb{R}^3$, *right plot* $\mathrm{SE}(3)$

By construction, the Lie group integrator (37) defines a numerical solution q_n that satisfies the holonomic constraints $\mathbf{\Phi}(q) = \mathbf{0}$. In a practical implementation, the residuals remain in the size of the stopping bounds for the Newton method that is used to solve in each time step the system of nonlinear equations (37). For the numerical tests we applied a combined absolute and relative error criterion with tolerances $\mathrm{ATOL} = 10^{-10}$ for the absolute errors and $\mathrm{RTOL} = 10^{-8}$ for the relative errors and observe constraint residuals of size $\|\mathbf{\Phi}(q_n)\| \ll 10^{-10}$, see Fig. 9.

Situation is different for the residuals in the hidden constraints (16) that are in general of the size of global discretization errors since $\mathbf{B}\big(q(t)\big)\mathbf{v}(t) = \mathbf{0}$. The left plot of Fig. 10 shows these non-vanishing residuals $\mathbf{B}(q_n)\mathbf{v}_n$ for $h = 1.0 \times 10^{-3}$ and $G = \mathrm{SO}(3) \times \mathbb{R}^3$. They are of size $\|\mathbf{B}(q_n)\mathbf{v}_n\| \leq 0.025$ and suffer from the transient spurious oscillations being known from Fig. 7 above. For the configuration space $G = \mathrm{SE}(3)$, the constraint residuals are smaller by eight orders of magnitude with $\max_n \|\mathbf{B}(q_n)\mathbf{v}_n\| \approx 1.0 \times 10^{-10}$. This unexpected solution behaviour is visualized in the right plot of Fig. 10. It is closely related to the fact that the constraint Jacobian $\mathbf{B}(q)$ in (22) is constant along the analytical solution $q(t)$, see Sect. 3.6 below for a more detailed analysis.

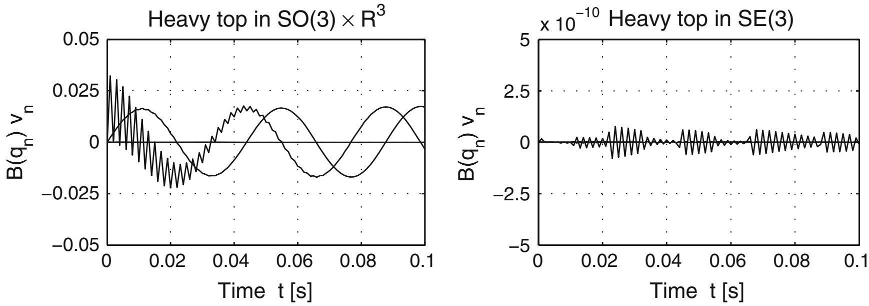


Fig. 10 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, index-3 formulation): Residuals in hidden constraints (16). *Left plot* $\mathrm{SO}(3) \times \mathbb{R}^3$, *right plot* $\mathrm{SE}(3)$

In all numerical tests of the present section, the numerical damping parameter was set to $\rho_\infty := 0.9$. The qualitative behaviour of the numerical solution in configuration spaces $\mathrm{SO}(3) \times \mathbb{R}^3$ and $\mathrm{SE}(3)$ is, however, not sensitive w.r.t. this algorithmic parameter, see, e.g., the results for $\rho_\infty = 0.8$ and the test setup of Fig. 7 in (Brüls et al. 2011) and the results for $\rho_\infty = 0.6$ and the test setup of Fig. 20 below in (Arnold et al. 2015).

3.4 Lie Group Time Integration and Index Reduction

The large amplitude of spurious oscillations in the numerical solution $\boldsymbol{\lambda}_n$, see Fig. 7, results from order reduction in Newmark-type methods that are directly applied to the index-3 formulation of the equations of motion for constrained mechanical systems, see (Cardona and Géradin 1994) and (Arnold et al. 2015). As an alternative to this direct time discretization of the index-3 Lie group DAE (15) we consider in the present section an analytical index reduction *before* time integration. We follow the approach of Gear et al. (1985) that is well known for equations of motion in linear spaces and was extended to the Lie group setting of the present paper in (Arnold et al. 2011a).

Gear et al. (1985) introduced an auxiliary vector $\boldsymbol{\eta}(t) \in \mathbb{R}^m$ in the kinematic equations to couple the hidden constraints at the level of velocity coordinates to the equations of motion. In the Lie algebra approach to Lie group time integration, these modified kinematic equations get the form $\dot{q}(t) = DL_{q(t)}(e) \cdot \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t)$ with $\widetilde{\boldsymbol{\Delta}\mathbf{q}} \in \mathfrak{g}$ being defined by $\boldsymbol{\Delta}\mathbf{q} = \mathbf{v} - \mathbf{B}^\top(q)\boldsymbol{\eta}$, see (6). The resulting *stabilized index-2 formulation* of the equations of motion is given by

$$\dot{q} = DL_q(e) \cdot \widetilde{\boldsymbol{\Delta}\mathbf{q}}\,, \tag{55a}$$

$$\boldsymbol{\Delta}\mathbf{q} = \mathbf{v} - \mathbf{B}^\top(q)\boldsymbol{\eta}\,, \tag{55b}$$

$$\mathbf{M}(q)\dot{\mathbf{v}} = -\mathbf{g}(q, \mathbf{v}, t) - \mathbf{B}^\top(q)\boldsymbol{\lambda}\,, \tag{55c}$$

$$\mathbf{\Phi}(q) = \mathbf{0}\,, \tag{55d}$$

$$\mathbf{B}(q)\mathbf{v} = \mathbf{0}\,. \tag{55e}$$

For the modified kinematic equations (55a), the time derivative of the holonomic constraints (55d) is given by $\mathbf{0} = \mathbf{B}(q)\mathbf{\Delta q}$, see (16). Therefore, Eqs. (55b) and (55e) yield $\mathbf{0} = [\mathbf{B}\mathbf{B}^\top](q)\boldsymbol{\eta}$ and $\boldsymbol{\eta}(t) \equiv \mathbf{0}$ since the full rank assumption on the constraint matrix $\mathbf{B} \in \mathbb{R}^{m\times k}$ implies that $\mathbf{B}\mathbf{B}^\top \in \mathbb{R}^{m\times m}$ is non-singular. Hence, $\mathbf{\Delta q}(t) = \mathbf{v}(t)$ and the stabilized index-2 formulation (55) is analytically equivalent to the original equations of motion (15).

The index analysis of Gear et al. (1985) is extended straightforwardly from linear spaces to the Lie group setting of the present paper and shows that the analytical transformation from (15) to (55) reduces the DAE index of the equations of motion from three to two.

The generalized-α method for the index-2 system (55) satisfies at $t = t_{n+1}$ the holonomic constraints (55d) as well as the hidden constraints (55e). An auxiliary vector $\boldsymbol{\eta}_n \in \mathbb{R}^m$ is added to the definition of the increment vector $\mathbf{\Delta q}_n$, see (55b):

$$q_{n+1} = q_n \circ \exp(h\widetilde{\mathbf{\Delta q}_n})\,, \tag{56a}$$

$$\begin{aligned}\mathbf{\Delta q}_n &= \mathbf{v}_n - \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n + \\ &\quad + (0.5-\beta)h\mathbf{a}_n + \beta h\mathbf{a}_{n+1}\,,\end{aligned} \tag{56b}$$

$$\mathbf{v}_{n+1} = \mathbf{v}_n + (1-\gamma)h\mathbf{a}_n + \gamma h\mathbf{a}_{n+1}\,, \tag{56c}$$

$$(1-\alpha_m)\mathbf{a}_{n+1} + \alpha_m\mathbf{a}_n = (1-\alpha_f)\dot{\mathbf{v}}_{n+1} + \alpha_f\dot{\mathbf{v}}_n\,, \tag{56d}$$

$$\mathbf{M}(q_{n+1})\dot{\mathbf{v}}_{n+1} = -\mathbf{g}(q_{n+1}, \mathbf{v}_{n+1}, t_{n+1}) - \mathbf{B}^\top(q_{n+1})\boldsymbol{\lambda}_{n+1}\,, \tag{56e}$$

$$\mathbf{\Phi}(q_{n+1}) = \mathbf{0}\,, \tag{56f}$$

$$\mathbf{B}(q_{n+1})\mathbf{v}_{n+1} = \mathbf{0}\,. \tag{56g}$$

Following the test scenario of Sect. 3.3, we study the asymptotic behaviour of integrator (56) for $h \to 0$ by numerical tests for the heavy top benchmark in configuration spaces $G = \mathrm{SO}(3) \times \mathbb{R}^3$ and $G = \mathrm{SE}(3)$, respectively. As before, we scale the norm of the (absolute) global errors by the norm of nominal values and consider the maximum of these relative errors in time interval $[t_0, t_{\mathrm{end}}] = [0, 1]$. Figure 11 shows these maximum values of the norm of global errors in q_n, $\mathbf{v}_n$ and $\boldsymbol{\lambda}_n$ versus time step size h. In double logarithmic scale, we get in the step size range $h \geq 2.5 \times 10^{-4}$ curves of slope $+2$ indicating second-order error terms in all solution components.

For the configuration space $\mathrm{SO}(3) \times \mathbb{R}^3$ (left plot) and very small time step sizes $h < 2.5 \times 10^{-4}$, the errors in $\boldsymbol{\lambda}_n$ are dominated by a first-order term. On the other hand, the error constants of the second-order error terms are slightly smaller than the ones in the corresponding plots for the index-3 integrator (37), see Figs. 6 and 8. The results for configuration space SE(3) in the right plot of Fig. 11 coincide up to plot accuracy with the ones in Figs. 6 and 8.

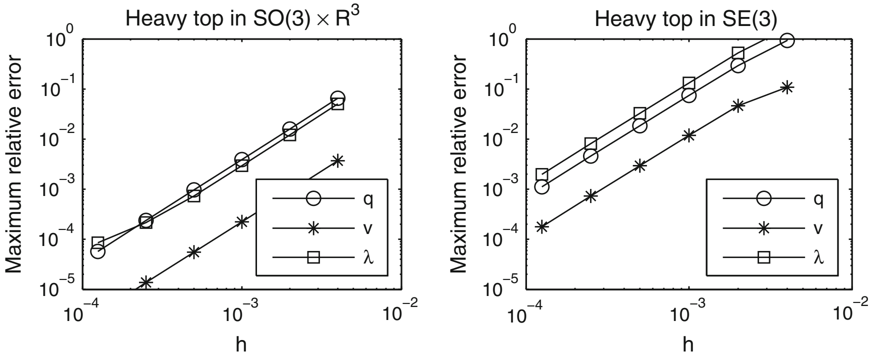


Fig. 11 Heavy top benchmark (stabilized index-2 formulation): Global error of integrator (56) versus h for $t \in [0, 1]$. *Left plot* $\mathrm{SO}(3) \times \mathbb{R}^3$, *right plot* $\mathrm{SE}(3)$

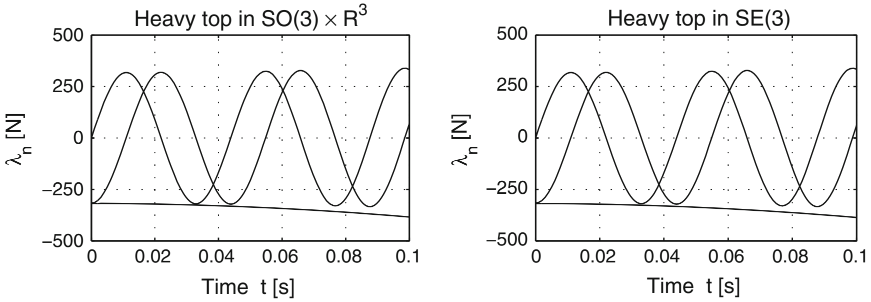


Fig. 12 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, stabilized index-2 formulation): Numerical solution $\boldsymbol{\lambda}_n$. *Left plot* $\mathrm{SO}(3) \times \mathbb{R}^3$, *right plot* $\mathrm{SE}(3)$

The comparison of time histories for $\boldsymbol{\lambda}_n$ in Figs. 7 and 12 shows that the spurious oscillations seem to disappear if hidden constraints are taken into account for time integration, see (56g). For a more detailed analysis, we consider in Fig. 13 the relative global error in $\lambda_{n,1}$ for $G = \mathrm{SO}(3) \times \mathbb{R}^3$ and two different time step sizes. There is an oscillating first-order error term of maximum amplitude $0.64\,h$ that is rapidly damped out. For time step sizes $h \geq 5.0 \times 10^{-4}$, it does not contribute significantly to the overall global error in $\boldsymbol{\lambda}_n$ on time interval $[0, 1]$ that is approximately of size $3.0 \times 10^3\, h^2$, see Fig. 11.

The test results in the right plot of Fig. 10 indicate that the index-3 integrator (37) yields for the heavy top benchmark in $G = \mathrm{SE}(3)$ a numerical solution q_n, $\mathbf{v}_n$ that satisfies the hidden constraints (56g) up to (very) small residuals. Therefore, the auxiliary variables $\boldsymbol{\eta}_n \in \mathbb{R}^m$ that represent the differences between integrators (37) and (56) vanish in that case identically, see also Sect. 3.6 below.

For the configuration space $G = \mathrm{SO}(3) \times \mathbb{R}^3$, we observed in the left plot of Fig. 10 non-vanishing constraint residuals $\mathbf{B}(q_n)\mathbf{v}_n$ for the index-3 integrator (37). In integrator (56), they are compensated by auxiliary variables $\boldsymbol{\eta}_n = \mathcal{O}(h^2)$ for the stabilized index-2 formulation of the equations of motion. Figure 14 shows $\boldsymbol{\eta}_n$ versus

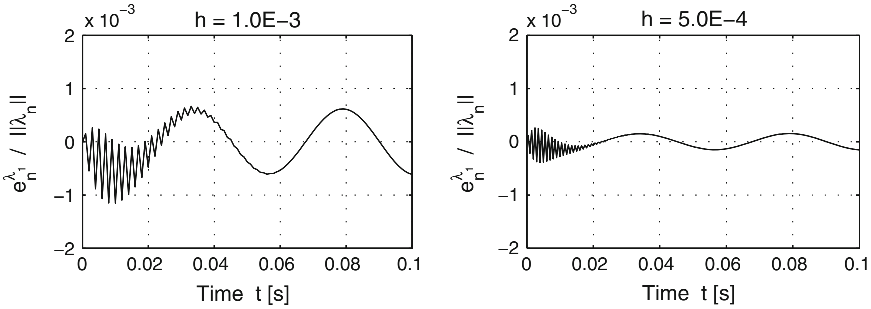


Fig. 13 Heavy top benchmark (stabilized index-2 formulation, $G = \text{SO}(3) \times \mathbb{R}^3$): Global error $e_n^{\lambda_1}/\|\boldsymbol{\lambda}_n\|$. *Left plot* $h = 1.0 \times 10^{-3}$, *right plot* $h = 5.0 \times 10^{-4}$

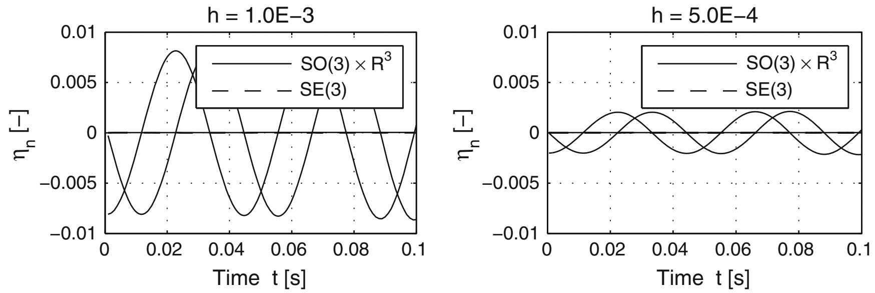


Fig. 14 Heavy top benchmark (stabilized index-2 formulation, $G = \text{SO}(3) \times \mathbb{R}^3$ and $G = \text{SE}(3)$): Numerical solution $\boldsymbol{\eta}_n$. *Left plot* $h = 1.0 \times 10^{-3}$, *right plot* $h = 5.0 \times 10^{-4}$

t_n for two different time step sizes. The maximum amplitudes of $\boldsymbol{\eta}_n$ differ by a factor of 4 if step sizes h and $h/2$ are considered, $h = 1.0 \times 10^{-3}$. Therefore, we expect second-order convergence for solution components $\boldsymbol{\eta}_n$.

Finally, we study the constraint residuals for a practical implementation of integrator (56). As before, the residuals in the holonomic constraints (15c) at the level of position coordinates are very small. For the hidden constraints (16) at the level of velocity coordinates, the residuals for integrator (56) are shown in Fig. 15. For the heavy top benchmark, they are of size 2.0×10^{-9} for $G = \text{SO}(3) \times \mathbb{R}^3$ and of size 2.0×10^{-15} for $G = \text{SE}(3)$.

In all these numerical tests for integrator (56), the extra effort for considering the hidden constraints (16) helps to reduce systematically shortcomings like spurious oscillations that were observed for the index-3 integrator (37) in Sect. 3.3.

3.5 Implementation Aspects

In each time step, the generalized-α method (37) defines the numerical solution $(q_{n+1}, \mathbf{v}_{n+1}, \dot{\mathbf{v}}_{n+1}, \mathbf{a}_{n+1}, \boldsymbol{\lambda}_{n+1})$ implicitly by a mixed system of linear and nonlinear

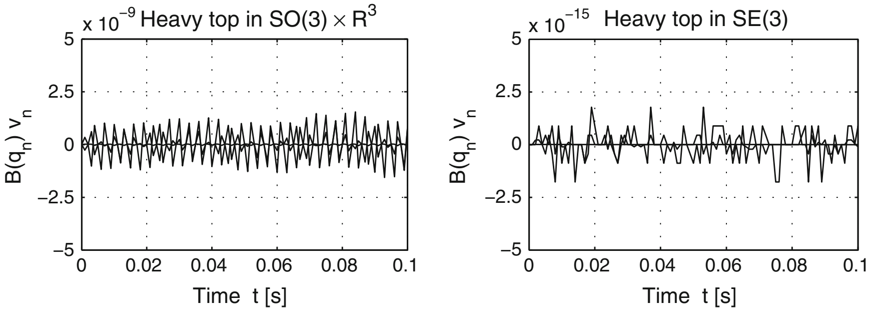


Fig. 15 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, stabilized index-2 formulation): Residuals in hidden constraints (16). *Left plot* $\mathrm{SO}(3) \times \mathbb{R}^3$, *right plot* $\mathrm{SE}(3)$

equations in $G \times \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^m$. Despite the nonlinear structure of the configuration space G, these equations may be solved numerically by a Newton–Raphson iteration in a *linear* space expressing $q_{n+1} \in G$ in terms of $\widetilde{\Delta \mathbf{q}}_n \in \mathfrak{g}$.

For the practical implementation of this Lie algebra approach, the Newton–Raphson method has to be combined with an appropriate scaling of equations and unknowns to guarantee that the condition number of the iteration matrix is bounded independently of h, see (Petzold and Lötstedt 1986) and the more recent discussion in (Bottasso et al. 2007). Denoting the scaled residual in the equilibrium conditions (15b) by

$$\mathbf{r}_h(q, \mathbf{v}, \dot{\mathbf{v}}, h\boldsymbol{\lambda}, t) := h\big(\mathbf{M}(q)\dot{\mathbf{v}} + \mathbf{g}(q, \mathbf{v}, t)\big) + \mathbf{B}^\top(q) \cdot h\boldsymbol{\lambda}\,,$$

we may rewrite the corrector equations (37) in the scaled and condensed form

$$\mathbf{0} = \boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}) := \begin{pmatrix} \mathbf{r}_h\big(q(\Delta\mathbf{q}_n), \mathbf{v}(\Delta\mathbf{q}_n), \dot{\mathbf{v}}(\Delta\mathbf{q}_n), h\boldsymbol{\lambda}_{n+1}, t_{n+1}\big) \\ \dfrac{1}{h}\,\boldsymbol{\Phi}\big(q(\Delta\mathbf{q}_n)\big) \end{pmatrix} \tag{57}$$

with $\boldsymbol{\xi}_{n+1} := \big((\Delta\mathbf{q}_n)^\top,\ h\boldsymbol{\lambda}_{n+1}^\top\big)^\top \in \mathbb{R}^{k+m}$ and

$$q_{n+1} = q(\Delta\mathbf{q}_n) := q_n \circ \exp(h\widetilde{\Delta\mathbf{q}}_n)\,, \tag{58a}$$

$$\mathbf{v}_{n+1} = \mathbf{v}(\Delta\mathbf{q}_n) := \frac{\gamma}{\beta}\Delta\mathbf{q}_n + (1 - \frac{\gamma}{\beta})\mathbf{v}_n + h(1 - \frac{\gamma}{2\beta})\mathbf{a}_n\,, \tag{58b}$$

$$\dot{\mathbf{v}}_{n+1} = \dot{\mathbf{v}}(\Delta\mathbf{q}_n) := \frac{1 - \alpha_m}{\beta(1 - \alpha_f)}\big(\frac{\Delta\mathbf{q}_n - \mathbf{v}_n}{h} - 0.5\mathbf{a}_n\big) + \frac{\mathbf{a}_n - \alpha_f\dot{\mathbf{v}}_n}{1 - \alpha_f}\,. \tag{58c}$$

The Newton–Raphson iteration

$$\boldsymbol{\xi}_{n+1}^{(k+1)} = \boldsymbol{\xi}_{n+1}^{(k)} + \Delta\boldsymbol{\xi}_{n+1}^{(k)} \quad \text{with} \quad \frac{\partial \boldsymbol{\Psi}_{n,h}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}_{n+1}^{(k)})\,\Delta\boldsymbol{\xi}_{n+1}^{(k)} = -\boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}^{(k)}) \tag{59}$$

may be started, e.g., with the initial guess $\boldsymbol{\xi}_{n+1}^{(0)} = \big(\mathbf{v}_n^\top + 0.5h\mathbf{a}_n^\top,\ h\boldsymbol{\lambda}_n^\top\big)^\top$, see also (Brüls et al. 2012, Table 1) for an alternative definition of $\boldsymbol{\xi}_{n+1}^{(0)}$ and for a more detailed description of the full algorithm. The iteration matrix $\partial\boldsymbol{\Psi}_{n,h}/\partial\boldsymbol{\xi}$ has a 2 × 2-block structure

$$\frac{\partial\boldsymbol{\Psi}_{n,h}}{\partial\boldsymbol{\xi}} = \begin{pmatrix} \dfrac{1-\alpha_m}{\beta(1-\alpha_f)}\mathbf{M} + h\,\dfrac{\gamma}{\beta}\mathbf{D} + h^2\,\mathbf{K}\,\mathbf{T} & \mathbf{B}^\top \\ \mathbf{B}\,\mathbf{T} & \mathbf{0} \end{pmatrix} \tag{60}$$

with mass matrix $\mathbf{M} = \mathbf{M}\big(q(\boldsymbol{\Delta}\mathbf{q}_n)\big) \in \mathbb{R}^{k\times k}$, damping matrix

$$\mathbf{D} = \frac{\partial\mathbf{g}}{\partial\mathbf{v}}\big(q(\boldsymbol{\Delta}\mathbf{q}_n), \mathbf{v}(\boldsymbol{\Delta}\mathbf{q}_n), t_{n+1}\big) \in \mathbb{R}^{k\times k}\,,$$

constraint matrix $\mathbf{B} = \mathbf{B}\big(q(\boldsymbol{\Delta}\mathbf{q}_n)\big) \in \mathbb{R}^{m\times k}$ and the tangent operator $\mathbf{T} = \mathbf{T}(h\boldsymbol{\Delta}\mathbf{q}_n) \in \mathbb{R}^{k\times k}$ that results from the derivative of the exponential map in (58a), see Corollary 2.7. The stiffness matrix $\mathbf{K} = \mathbf{K}(q, \mathbf{v}, \dot{\mathbf{v}}, \boldsymbol{\lambda}, t) \in \mathbb{R}^{k\times k}$ represents the partial derivatives of the equilibrium equations (15b) w.r.t. $q \in G$ in the sense that

$$D_q\big(\mathbf{M}(q)\dot{\mathbf{v}} + \mathbf{g}(q, \mathbf{v}, t) + \mathbf{B}^\top(q)\boldsymbol{\lambda}\big) \cdot \big(DL_q(e) \cdot \widetilde{\mathbf{w}}\big) = \mathbf{K}(q, \mathbf{v}, \dot{\mathbf{v}}, \boldsymbol{\lambda}, t)\,\mathbf{w}$$

for all $\mathbf{w} \in \mathbb{R}^k$. It is evaluated at $q = q(\boldsymbol{\Delta}\mathbf{q}_n)$, $\mathbf{v} = \mathbf{v}(\boldsymbol{\Delta}\mathbf{q}_n)$, $\dot{\mathbf{v}} = \dot{\mathbf{v}}(\boldsymbol{\Delta}\mathbf{q}_n)$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{n+1}$ and $t = t_{n+1}$.

The algorithmic parameters α_m, α_f and β in (37) satisfy $\alpha_m \neq 1, \alpha_f \neq 1$ and $\beta \neq 0$ since otherwise q_{n+1} would be independent of $\dot{\mathbf{v}}_{n+1}$ (and therefore also independent of the equilibrium equations (37e) at $t = t_{n+1}$). Hence, the iteration matrix $\partial\boldsymbol{\Psi}_{n,h}/\partial\boldsymbol{\xi}$ in (60) is non-singular for sufficiently small time step sizes h if the mass matrix $\mathbf{M}(q)$ is symmetric, positive definite and the constraint matrix $\mathbf{B}(q)$ has full rank (note, that $\mathbf{T}(h\boldsymbol{\Delta}\mathbf{q}_n) = \mathbf{I}_k + \mathcal{O}(h)$).

For sufficiently small time step sizes $h > 0$, the convergence of the Newton–Raphson iteration (59) may always be guaranteed under reasonable assumptions on $q_n, \mathbf{v}_n$:

Lemma 3.3 *If $\alpha_m \neq 1, \alpha_f \neq 1, \beta \neq 0$ and the numerical solution satisfies at $t = t_n$ the (hidden) constraints with residuals $\|\boldsymbol{\Phi}(q_n)\| \leq \gamma_0 h$ and $\|\mathbf{B}(q_n)\mathbf{v}_n\| \leq \gamma_0$ and a sufficiently small constant $\gamma_0 > 0$ then the generalized-α method (37) is well defined since the Newton–Raphson iteration (59) with initial guess $\boldsymbol{\xi}_{n+1}^{(0)} = (\mathbf{v}_n^\top, \mathbf{0}^\top)^\top + \mathcal{O}(h)$ converges for all sufficiently small time step sizes $h > 0$ to a locally uniquely defined solution of (57) with $\boldsymbol{\xi}_{n+1} = \boldsymbol{\xi}_{n+1}^{(0)} + \mathcal{O}(h) + \mathcal{O}(\gamma_0)$.*

Proof The assumptions on $\boldsymbol{\Phi}(q_n)$, $\mathbf{B}(q_n)\mathbf{v}_n$ and $\boldsymbol{\xi}_{n+1}^{(0)}$ are sufficient to prove $\boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}^{(0)}) = \mathcal{O}(h) + \mathcal{O}(\gamma_0)$ since $\mathbf{r}_h = \mathcal{O}(h)$ by definition and $q(\boldsymbol{\Delta}\mathbf{q}_n^{(0)}) = q(\mathbf{v}_n) + \mathcal{O}(h) = q_n \circ \exp(h\widetilde{\mathbf{v}}_n) + \mathcal{O}(h)$ resulting in

$$\begin{aligned}\frac{1}{h}\,\|\boldsymbol{\Phi}\big(q(\boldsymbol{\Delta}\mathbf{q}_n^{(0)})\big)\| &= \frac{1}{h}\,\|\boldsymbol{\Phi}(q_n) + h\,\frac{\mathrm{d}}{\mathrm{d}h}\boldsymbol{\Phi}\big(q_n \circ \exp(h\widetilde{\mathbf{v}}_n)\big) + \mathcal{O}(h^2)\| \\ &\le \frac{1}{h}\,\|\boldsymbol{\Phi}(q_n)\| + \|\mathbf{B}\big(q_n \circ \exp(h\widetilde{\mathbf{v}}_n)\big)\mathbf{v}_n\| + \mathcal{O}(h) \\ &= \mathcal{O}(h) + \mathcal{O}(\gamma_0)\,,\end{aligned}$$

see (32). Therefore, the convergence of the Newton–Raphson iteration to a locally uniquely defined solution $\boldsymbol{\xi}_{n+1} = \boldsymbol{\xi}_{n+1}^{(0)} + \mathcal{O}(h) + \mathcal{O}(\gamma_0)$ of (57) is guaranteed whenever the constant $\gamma_0 > 0$ and the time step size $h > 0$ are sufficiently small (Kelley 1995). □

The corrector equations (56) of the Lie group integrator for the stabilized index-2 formulation (55) may be condensed as well replacing the left equations in (58b, 58c) by

$$\mathbf{v}_{n+1} = \mathbf{v}\big(\boldsymbol{\Delta}\mathbf{q}_n + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n\big)\,,\;\; \dot{\mathbf{v}}_{n+1} = \dot{\mathbf{v}}\big(\boldsymbol{\Delta}\mathbf{q}_n + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n\big)\,.$$

The resulting scaled system of nonlinear equations is given by

$$\mathbf{0} = \boldsymbol{\Psi}_{n,h}(\boldsymbol{\xi}_{n+1}) := \begin{pmatrix} \boldsymbol{\varrho}_h(\boldsymbol{\Delta}\mathbf{q}_n, h\boldsymbol{\lambda}_{n+1}, \boldsymbol{\eta}_n) \\ \dfrac{1}{h}\,\boldsymbol{\Phi}\big(q(\boldsymbol{\Delta}\mathbf{q}_n)\big) \\ \mathbf{B}\big(q(\boldsymbol{\Delta}\mathbf{q}_n)\big)\mathbf{v}\big(\boldsymbol{\Delta}\mathbf{q}_n + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n\big) \end{pmatrix} \tag{61}$$

with $\boldsymbol{\xi}_{n+1} := \big((\boldsymbol{\Delta}\mathbf{q}_n)^\top,\, h\boldsymbol{\lambda}_{n+1}^\top,\, \boldsymbol{\eta}_n^\top\big)^\top \in \mathbb{R}^{k+2m}$ and

$$\begin{aligned}&\boldsymbol{\varrho}_h(\boldsymbol{\Delta}\mathbf{q}_n, h\boldsymbol{\lambda}_{n+1}, \boldsymbol{\eta}_n) := \\ &\quad \mathbf{r}_h\big(q(\boldsymbol{\Delta}\mathbf{q}_n), \mathbf{v}\big(\boldsymbol{\Delta}\mathbf{q}_n + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n\big), \dot{\mathbf{v}}\big(\boldsymbol{\Delta}\mathbf{q}_n + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n\big), h\boldsymbol{\lambda}_{n+1}, t_{n+1}\big)\,.\end{aligned}$$

The scaling of equations and unknowns guarantees again that the condition number of the iteration matrix $\partial\boldsymbol{\Psi}_{n,h}/\partial\boldsymbol{\xi}$ is bounded for $h \to 0$. This iteration matrix has the 3×3-block structure

$$\frac{\partial\boldsymbol{\Psi}_{n,h}}{\partial\boldsymbol{\xi}} = \begin{pmatrix} \mathbf{M}^* + h^2\,\mathbf{K}\,\mathbf{T} & \mathbf{B}^\top & \mathbf{M}^*\,\mathbf{B}^\top(q_n) \\ \mathbf{B}\,\mathbf{T} & \mathbf{0} & \mathbf{0} \\ \dfrac{\gamma}{\beta}\mathbf{B} + h\,\mathbf{Z} & \mathbf{0} & \dfrac{\gamma}{\beta}\,\mathbf{B}\mathbf{B}^\top(q_n) \end{pmatrix} \tag{62}$$

with

$$\mathbf{M}^* := \frac{1-\alpha_m}{\beta(1-\alpha_f)}\mathbf{M} + h\,\frac{\gamma}{\beta}\mathbf{D}$$

and a matrix $\mathbf{Z} \in \mathbb{R}^{k\times k}$ that represents $\big(\partial/\partial(\boldsymbol{\Delta}\mathbf{q}_n)\big)\mathbf{B}\big(q(\boldsymbol{\Delta}\mathbf{q}_n)\big)\mathbf{v}$ in the sense that

$$\mathbf{Z}\mathbf{w} = \mathbf{Z}\big(q(\boldsymbol{\Delta}\mathbf{q}_n)\big)\big(\mathbf{v}\big(\boldsymbol{\Delta}\mathbf{q}_n + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n\big), \mathbf{T}(h\boldsymbol{\Delta}\mathbf{q}_n)\mathbf{w}\big)\,,\;\; (\,\mathbf{w} \in \mathbb{R}^k\,)\,,$$

see (17). Using the formal decomposition

$$\frac{\partial \boldsymbol{\Psi}_{n,h}}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \mathbf{I}_k & \mathbf{0} & \mathbf{M}^* \, \mathbf{B}^\top(q_n) \\ \mathbf{0} & \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \frac{\gamma}{\beta}\mathbf{I}_m & \frac{\gamma}{\beta}\mathbf{B}\mathbf{B}^\top(q_n) \end{pmatrix} \begin{pmatrix} \mathbf{M}^* + \mathcal{O}(h) & \mathbf{B}^\top & \mathbf{0} \\ \mathbf{B}\,\mathbf{T} & \mathbf{0} & \mathbf{0} \\ \mathcal{O}(h) & \mathbf{0} & \mathbf{I}_m \end{pmatrix},$$

see (62), we may verify that the iteration matrix is non-singular if $h > 0$ is sufficiently small. With the additional assumptions $\gamma \neq 0$ and $\boldsymbol{\eta}_n^{(0)} = \mathcal{O}(h)$, Lemma 3.3 applies also to the Lie group integrator (56) for the stabilized index-2 formulation. The method is well defined and the corresponding condensed system (61) may be solved by the Newton–Raphson method (59).

In the practical implementation of implicit ODE/DAE time integration methods, the Jacobian $(\partial \boldsymbol{\Psi}_{n,h}/\partial \boldsymbol{\xi})(\boldsymbol{\xi}_{n+1}^{(k)})$ in the Newton–Raphson step (59) is substituted by an approximation that is kept constant during integration as long as possible, see, e.g., (Brenan et al. 1996, Sect. 5.2.2). In (Brüls et al. 2011), the influence of different Lie group formulations on the number of Jacobian updates was studied by numerical tests for the Lie group integrator (37). A very small number of Jacobian evaluations were observed for equations of motion like (22) that are characterized by a constant mass matrix $\mathbf{M}$ and a constant constraint Jacobian $\mathbf{B}$, see also Lemma 3.5 below.

If the generalized-α integrators (37) and (56) are applied to non-stiff systems and the time step size h is sufficiently small, then we may neglect in (60) and (62) the terms $h\gamma\mathbf{D}/\beta$, $h^2\mathbf{KT}$ and $h\mathbf{Z}$. For the numerical tests in Sects. 3.3 and 3.4, this simplified Newton–Raphson method was combined with a damping strategy based on Armijo line search, see (Kelley 1995). Convergence problems in the corrector iteration were observed for just one simulation scenario (integrator (37) for the heavy top benchmark, $G = \mathrm{SO}(3) \times \mathbb{R}^3$, $h = 4.0 \times 10^{-3}$, see the left plots of Figs. 6 and 8). Here, we had to take into account a difference approximation of the term $h\gamma\mathbf{D}/\beta + h^2\mathbf{KT}$ in (60).

3.6 Constraint Residuals

Both generalized-α integrators (37) and (56) satisfy by construction the holonomic constraints (15c) at the level of position coordinates: $\boldsymbol{\Phi}(q_n) = \mathbf{0}$, $(\, n > 0\,)$. For the stabilized index-2 integrator (56), the hidden constraints (16) at velocity level are satisfied as well: $\mathbf{B}(q_n)\mathbf{v}_n = \mathbf{0}$, $(\, n > 0\,)$, see (56g). For the index-3 integrator (37), these residuals $\mathbf{B}(q_n)\mathbf{v}_n$ remain in general in the size of global discretization errors since $\mathbf{B}(q(t))\mathbf{v}(t) \equiv \mathbf{0}$. For some problem classes, the constraint residuals $\mathbf{B}(q_n)\mathbf{v}_n$ vanish, however, also for the index-3 integrator (37). Therefore, both integrators (37) and (56) define in that case one and the same numerical solution $(q_n, \mathbf{v}_n, \dot{\mathbf{v}}_n, \mathbf{a}_n, \boldsymbol{\lambda}_n)$ with auxiliary variables $\boldsymbol{\eta}_n = \mathbf{0}$, $(\, n \geq 0\,)$. In a practical implementation, the numerical solutions will coincide up to round-off errors and errors that are caused by stopping the Newton–Raphson iteration after a finite number of iteration steps.

In the present section, we show that the numerical solution of the index-3 integrator (37) will always satisfy the hidden constraints (16) at the level of velocity coordinates if the constraint Jacobian $\mathbf{B}$ is constant (Lemma 3.4). In Lemma 3.5, this result is extended to a special problem class in SE(3) with $\mathbf{B}(q) = \text{const}$ on the constraint manifold $\mathfrak{M} = \{\, q \in G \,:\, \mathbf{\Phi}(q) = \mathbf{0} \,\}$. This analysis gives the formal proof for the numerical test results in the right plot of Fig. 10 that were obtained for the heavy top benchmark in configuration space $G = \text{SE}(3)$.

Improved error estimates for certain configuration spaces are a topic of active current research on Lie group time integration methods, see also the recently published results of Müller and Terze (2014a, b).

Lemma 3.4 *Consider equations of motion (15) with constant constraint Jacobian* $\mathbf{B}$ *in the hidden constraints (16) at velocity level.*

(a) For this problem class, the curvature term $\mathbf{Z}(q)\big(\mathbf{v},\mathbf{v}\big)$ *in the hidden constraints (18) at acceleration level vanishes identically.*

(b) If $\mathbf{B} = \text{const}$ *and the starting values* q_0, $\mathbf{v}_0$, $\mathbf{a}_0$ *are consistent (* $\mathbf{0} = \mathbf{\Phi}(q_0) = \mathbf{B}\mathbf{v}_0 = \mathbf{B}\mathbf{a}_0$ *) then the numerical solution* $(q_n, \mathbf{v}_n, \dot{\mathbf{v}}_n, \mathbf{a}_n, \boldsymbol{\lambda}_n)$ *of the generalized-*α *method (37) satisfies for all* $n \geq 0$ *both the holonomic constraints (15c) at position level and the hidden constraints (16) at velocity level:* $\mathbf{\Phi}(q_n) = \mathbf{B}\mathbf{v}_n = \mathbf{0}$.

Proof (a) The time derivative of hidden constraints (16) with $\mathbf{B} = \text{const}$ is given by $\mathbf{0} = \mathbf{B}\dot{\mathbf{v}}(t)$. Comparing this expression with the hidden constraints (18), we get $\mathbf{Z}(q)\big(\mathbf{v},\mathbf{v}\big) = \mathbf{0}$.

(b) Because of $\mathbf{\Phi}(q_0) = \mathbf{0}$ and (37f), the numerical solution q_n satisfies the holonomic constraints (15c) for all $n \geq 0$. To prove $\mathbf{B}\mathbf{v}_n = \mathbf{B}\mathbf{a}_n = \mathbf{0}$ by induction, we observe that $\mathbf{\Phi}(q_{n+1}) = \mathbf{\Phi}(q_n) = \mathbf{0}$ and $q_{n+1} = q_n \circ \exp(h\widetilde{\Delta\mathbf{q}}_n)$, see (37a), imply $\mathbf{\Psi}(1) = \mathbf{\Psi}(0) = \mathbf{0}$ for the continuously differentiable function $\mathbf{\Psi} : [0,1] \to \mathbb{R}^m$, $\vartheta \mapsto \mathbf{\Phi}\big(q_n \circ \exp(\vartheta h \widetilde{\Delta\mathbf{q}}_n)\big)$. Therefore,

$$\begin{aligned} \mathbf{0} = \frac{\mathbf{\Psi}(1) - \mathbf{\Psi}(0)}{h} &= \frac{1}{h}\int_0^1 \frac{\mathrm{d}\mathbf{\Phi}}{\mathrm{d}\vartheta}\big(q_n \circ \exp(\vartheta h \widetilde{\Delta\mathbf{q}}_n)\big)\,\mathrm{d}\vartheta \\ &= \int_0^1 \mathbf{B}\big(q_n \circ \exp(\vartheta h \widetilde{\Delta\mathbf{q}}_n)\big)\,\Delta\mathbf{q}_n\,\mathrm{d}\vartheta\,, \end{aligned} \tag{63}$$

see (14) and (32). If $\mathbf{B} = \text{const}$, then the integrand in (63) is constant as well resulting in $\mathbf{B}\Delta\mathbf{q}_n = \mathbf{0}$. We get $\mathbf{B}\mathbf{a}_{n+1} = \mathbf{0}$ (if $\mathbf{B}\mathbf{v}_n = \mathbf{B}\mathbf{a}_n = \mathbf{0}$) from left multiplication of (37b) by matrix $\mathbf{B}$ and obtain finally $\mathbf{B}\mathbf{v}_{n+1} = \mathbf{0}$ multiplying also the velocity update (37c) from the left by the (constant) constraint Jacobian $\mathbf{B}$. □

Lemma 3.5 *Consider a rigid body with configuration space* SE(3) *and holonomic constraints (15c) of the form*

$$\mathbf{0} = \boldsymbol{\Phi}(q) = \boldsymbol{\Phi}\big((\mathbf{R}, \mathbf{x})_{\mathrm{SE}(3)}\big) = \mathbf{X} - \mathbf{R}^\top \mathbf{x} \tag{64}$$

with a constant vector $\mathbf{X} \in \mathbb{R}^3$.

(a) Along any solution $q(t)$ of the constrained equations of motion (15) matrix $\mathbf{B}\big(q(t)\big)$ is constant and the curvature term $\mathbf{Z}\big(q(t)\big)\big(\mathbf{v}(t), \mathbf{v}(t)\big)$ vanishes identically.

(b) If the generalized-α method (37) is applied with consistent starting values ($\mathbf{0} = \boldsymbol{\Phi}(q_0) = \mathbf{B}(q_0)\mathbf{v}_0 = \mathbf{B}(q_0)\mathbf{a}_0$) and with sufficiently small time step size $h > 0$ to equations of motion (15) in SE(3) *with holonomic constraints (64) then the numerical solution satisfies both the holonomic constraints at position level and the hidden constraints at velocity level: $\boldsymbol{\Phi}(q_n) = \mathbf{B}(q_n)\mathbf{v}_n = \mathbf{0}$, ($n \geq 0$).*

Proof (a) Straightforward differentiation of constraint (64) shows

$$\begin{aligned}\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Phi}(q(t)) &= -\dot{\mathbf{R}}^\top \mathbf{x} - \mathbf{R}^\top \dot{\mathbf{x}} = -(\mathbf{R}\widetilde{\boldsymbol{\Omega}})^\top \mathbf{x} - \mathbf{R}^\top \mathbf{R}\mathbf{U} \\ &= -\widetilde{\boldsymbol{\Omega}}^\top \mathbf{R}^\top \mathbf{x} - \mathbf{U} = \widetilde{\boldsymbol{\Omega}}\mathbf{R}^\top \mathbf{x} - \mathbf{U} = -\widetilde{\mathbf{R}^\top \mathbf{x}}\,\boldsymbol{\Omega} - \mathbf{U} = \mathbf{B}(q)\mathbf{v}\end{aligned}$$

with $q = (\mathbf{R}, \mathbf{x})_{\mathrm{SE}(3)} \in \mathrm{SE}(3)$ and $\mathbf{v} = (\,\boldsymbol{\Omega}^\top,\ \mathbf{U}^\top)^\top \in \mathbb{R}^6$. On the constraint manifold, we have $\mathbf{R}^\top \mathbf{x} = \mathbf{X}$, see (64), and the constraint Jacobian $\mathbf{B}(q)$ is constant: $\mathbf{B}\big((\mathbf{R}, \mathbf{x})_{\mathrm{SE}(3)}\big) = \mathbf{B}^{\mathbf{X}} := \big(-\widetilde{\mathbf{X}} \quad -\mathbf{I}_3\,\big)$. Therefore, the hidden constraints (16) and (18) are given by $\mathbf{B}^{\mathbf{X}}\mathbf{v}(t) = \mathbf{0}$ and $\mathbf{B}^{\mathbf{X}}\dot{\mathbf{v}}(t) = \mathbf{0}$ with $\mathbf{Z}\big(q(t)\big)\big(\mathbf{v}(t), \mathbf{v}(t)\big) \equiv \mathbf{0}$ along any solution $\big(q(t), \mathbf{v}(t)\big)$.

(b) This part of the proof is substantially more technical than the corresponding proof of Lemma 3.4(b) since $\mathbf{B}(q)$ is not constant beyond the constraint manifold $\mathfrak{M}$ and there is no straightforward way to prove that in (63) the argument $q_n \circ \exp(\vartheta h \widetilde{\boldsymbol{\Delta}\mathbf{q}}_n)$ of $\mathbf{B}$ will remain in $\mathfrak{M}$ for $\vartheta \in (0, 1)$.

In SE(3), the position update formula $q_{n+1} = q_n \circ \exp(h\widetilde{\boldsymbol{\Delta}\mathbf{q}}_n)$ gets the form

$$\mathbf{R}_{n+1} = \mathbf{R}_n \exp_{\mathrm{SO}(3)}(h\widetilde{\boldsymbol{\Delta}\mathbf{R}_n})\,, \quad \mathbf{x}_{n+1} = \mathbf{x}_n + h\mathbf{R}_n \mathbf{T}^\top_{\mathrm{SO}(3)}(h\boldsymbol{\Delta}\mathbf{R}_n)\boldsymbol{\Delta}\mathbf{x}_n$$

with $\boldsymbol{\Delta}\mathbf{q}_n = (\,\boldsymbol{\Delta}\mathbf{R}_n^\top,\ \boldsymbol{\Delta}\mathbf{x}_n^\top)^\top$, see Example 2.1(a). Because of $\boldsymbol{\Phi}(q_0) = \mathbf{0}$ and $\boldsymbol{\Phi}(q_{n+1}) = \mathbf{0}$, ($n \geq 0$), see (37f), we get $\mathbf{R}_n^\top \mathbf{x}_n - \mathbf{R}_{n+1}^\top \mathbf{x}_{n+1} = \mathbf{X} - \mathbf{X} = \mathbf{0}$, see (64), and

$$
\begin{aligned}
\mathbf{0} &= \exp_{\mathrm{SO}(3)}(h\widetilde{\mathbf{\Delta R}}_n)\,\frac{\mathbf{R}_n^\top \mathbf{x}_n - \mathbf{R}_{n+1}^\top \mathbf{x}_{n+1}}{h} \\
&= \frac{\exp_{\mathrm{SO}(3)}(h\widetilde{\mathbf{\Delta R}}_n)\mathbf{R}_n^\top \mathbf{x}_n - \mathbf{R}_n^\top\Big(\mathbf{x}_n + h\mathbf{R}_n\mathbf{T}_{\mathrm{SO}(3)}^\top(h\mathbf{\Delta R}_n)\mathbf{\Delta x}_n\Big)}{h} \\
&= \frac{\exp_{\mathrm{SO}(3)}(h\widetilde{\mathbf{\Delta R}}_n) - \mathbf{I}_3}{h}\,\mathbf{R}_n^\top \mathbf{x}_n - \mathbf{T}_{\mathrm{SO}(3)}^\top(h\mathbf{\Delta R}_n)\mathbf{\Delta x}_n
\end{aligned}
\tag{65}
$$

with

$$
\begin{aligned}
\exp_{\mathrm{SO}(3)}(h\widetilde{\mathbf{\Delta R}}_n) - \mathbf{I}_3 = \sum_{i=1}^{\infty}\frac{1}{i!}\big(h\widetilde{\mathbf{\Delta R}}_n\big)^i &= h\sum_{i=0}^{\infty}\frac{1}{(i+1)!}\big(h\widetilde{\mathbf{\Delta R}}_n\big)^i\,\widetilde{\mathbf{\Delta R}}_n \\
&= h\sum_{i=0}^{\infty}\frac{(-1)^i}{(i+1)!}\big(-h\widetilde{\mathbf{\Delta R}}_n\big)^i\,\widetilde{\mathbf{\Delta R}}_n\,.
\end{aligned}
$$

In SO(3), the $\widetilde{(\bullet)}$ operator maps $\mathbf{\Delta R}_n \in \mathbb{R}^3$ to the skew symmetric matrix $\widetilde{\mathbf{\Delta R}}_n$, see (2), and we have $\widehat{\mathbf{\Delta R}}_n = \widetilde{\mathbf{\Delta R}}_n$, see Remark 2.8(b). Therefore, $-\widetilde{\mathbf{\Delta R}}_n = (\widetilde{\mathbf{\Delta R}}_n)^\top = (\widehat{\mathbf{\Delta R}}_n)^\top$ and the series expansion (30) proves

$$
\exp_{\mathrm{SO}(3)}(h\widetilde{\mathbf{\Delta R}}_n) - \mathbf{I}_3 = h\left(\mathbf{T}_{\mathrm{SO}(3)}(h\mathbf{\Delta R}_n)\right)^\top \widetilde{\mathbf{\Delta R}}_n\,.
$$

Inserting this expression in (65), we get

$$
\mathbf{0} = \mathbf{T}_{\mathrm{SO}(3)}^\top(h\mathbf{\Delta R}_n)\big(\widetilde{\mathbf{\Delta R}}_n(\mathbf{R}_n^\top \mathbf{x}_n) - \mathbf{\Delta x}_n\big)
$$

and therefore also

$$
\mathbf{0} = \widetilde{\mathbf{\Delta R}}_n(\mathbf{R}_n^\top \mathbf{x}_n) - \mathbf{\Delta x}_n = -\widetilde{\mathbf{R}_n^\top \mathbf{x}_n}\,\mathbf{\Delta R}_n - \mathbf{\Delta x}_n = \mathbf{B}(q_n)\mathbf{\Delta q}_n
$$

since the tangent operator $\mathbf{T}_{\mathrm{SO}(3)}(h\mathbf{\Delta R}_n) = \mathbf{I}_3 + \mathcal{O}(h)$ is non-singular for sufficiently small time step sizes $h > 0$. Now, the proof may be completed following line by line the proof of Lemma 3.4(b) since $q_n \in \mathfrak{M}$ by construction and $\mathbf{B}(q)$ is constant on the constraint manifold, i.e., $\mathbf{B}(q_n) = \mathbf{B}^{\mathbf{X}} = \text{const.}$ □

4 Convergence Analysis

The convergence of generalized-α time integration methods for nonlinear unconstrained systems in linear configuration spaces was studied by Erlicher et al. (2002) using an equivalent multi-step representation. In the DAE Lie group case, this analysis has to be extended to constrained systems in nonlinear configuration spaces with Lie group structure, see (Brüls et al. 2012). In the present section, we follow the direct convergence analysis for the generalized-α method in one-step form (37) that was developed in (Arnold et al. 2015) to study the convergence in long-term integration as well as in the transient phase in full detail.

4.1 Local Truncation Errors, Global Errors and Error Recursion

For unconstrained systems in linear spaces, the local truncation errors were introduced in (39), see Sect. 3.2 above. Since there are no discretization errors in the holonomic constraints (15c), see (37f), these definitions may be used as well in the constrained case.

For configuration spaces with Lie group structure, the definition of the local truncation error $\mathbf{l}_n^{\mathbf{q}}$ in (39a) has to be adapted to the Lie group setting. In the Lie algebra approach to error analysis of Lie group time integration methods, we follow the proposal of Wensch (2001) to define local and global errors by elements of the corresponding Lie algebra, see also (Orel 2010):

Definition 4.1 For the solution components $q \in G$, the *local truncation error* $\widetilde{\mathbf{l}}_n^q \in \mathfrak{g}$ of the generalized-α Lie group method (37) is defined by

$$q(t_{n+1}) = q(t_n) \circ \exp(h \widetilde{\mathbf{\Delta q}}(t_n)) \circ \exp(\widetilde{\mathbf{l}}_n^q) \tag{66}$$

with $\mathbf{\Delta q}(t_n) := \mathbf{v}(t_n) + (0.5 - \beta) h \dot{\mathbf{v}}(t_n + \Delta_\alpha h) + \beta h \dot{\mathbf{v}}(t_{n+1} + \Delta_\alpha h)$.

To get an error estimate for $\widetilde{\mathbf{l}}_n^q$, we compare the asymptotic behaviour of $q(t_{n+1}) = q(t_n + h)$ and $q(t_n) \circ \exp(h \widetilde{\mathbf{\Delta q}}(t_n))$ for $h \to 0$. For any smooth function $\mathbf{v}(t)$, the flow of $\dot{q}(t) = DL_q(e) \cdot \widetilde{\mathbf{v}}(t)$ is locally represented by a smooth function $\widetilde{\boldsymbol{\nu}} : [-h_0, h_0] \times \mathbb{R} \times G \to \mathfrak{g}$:

$$q(t + h) = q(t) \circ \exp\bigl(h \widetilde{\boldsymbol{\nu}}(h; t, q(t))\bigr). \tag{67}$$

The asymptotic behaviour of $h\widetilde{\boldsymbol{\nu}}$ is characterized by the Magnus expansion

$$h\widetilde{\boldsymbol{\nu}}(h; t, q(t)) = h\widetilde{\mathbf{v}}(t) + \frac{h^2}{2}\dot{\widetilde{\mathbf{v}}}(t) + \frac{h^3}{6}\ddot{\widetilde{\mathbf{v}}}(t) + \frac{h^3}{12}[\widetilde{\mathbf{v}}(t), \dot{\widetilde{\mathbf{v}}}(t)] + \mathcal{O}(h^4)\,, \tag{68}$$

see (Hairer et al. 2006) and (Müller 2010). The matrix commutator $[\widetilde{\mathbf{v}}, \dot{\widetilde{\mathbf{v}}}]$ vanishes identically in linear spaces, see Sect. 2.5. In the Lie group setting, it introduces an additional local error term if the arguments $\widetilde{\mathbf{v}}(t)$ and $\dot{\widetilde{\mathbf{v}}}(t)$ do not commute, see Lemma 4.2 below.

Inserting (67) with $t = t_n$ into the (implicit) definition of $\widetilde{\mathbf{l}}_n^q$, see (66), we get $q(t_n) \circ \exp\bigl(h\widetilde{\boldsymbol{\nu}}(h; t_n, q(t_n))\bigr) = q(t_n) \circ \exp(h\widetilde{\mathbf{\Delta q}}(t_n)) \circ \exp(\widetilde{\mathbf{l}}_n^q)$. Therefore, the term $\exp(\widetilde{\mathbf{l}}_n^q)$ may be expressed as product of matrix exponentials:

$$\exp(\widetilde{\mathbf{l}}_n^q) = \exp(-h\widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)) \circ \exp\bigl(h\widetilde{\boldsymbol{\nu}}(h; t_n, q(t_n))\bigr)\,.$$

In (Arnold et al. 2015, Lemma 1), we used the Baker–Campbell–Hausdorff formula to show that $\widetilde{\mathbf{l}}_n^q$ and $h\bigl(\widetilde{\boldsymbol{\nu}}(h; t_n, q(t_n)) - \widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)\bigr)$ coincide up to higher order terms, see also Lemma 2.5. Comparing the Magnus expansion (68) with the Taylor expansion of $\widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)$, we get

Lemma 4.2 *With $\Delta_\alpha := \alpha_m - \alpha_f$ and $C_q := (1 - 6\beta - 3\Delta_\alpha)/6$, the local truncation error $\widetilde{\mathbf{l}}_n^q$ is given by*

$$\widetilde{\mathbf{l}}_n^q = C_q h^3 \widetilde{\ddot{\mathbf{v}}}(t_n) + h^3[\widetilde{\mathbf{v}}(t_n), \widetilde{\dot{\mathbf{v}}}(t_n)]/12 + \mathcal{O}(h^4)\,. \tag{69}$$

If the parameters γ, α_m, α_f satisfy the order condition (41) then the local truncation errors are bounded by

$$\|\mathbf{l}_n^q\| = \mathcal{O}(h^3)\,,\quad \|\mathbf{l}_{n+1}^q - \mathbf{l}_n^q\| = \mathcal{O}(h^4)\,,\quad \|\mathbf{l}_n^{\mathbf{v}}\| = \mathcal{O}(h^3)\,,\quad \|\mathbf{l}_n^{\mathbf{a}}\| = \mathcal{O}(h^2)\,. \tag{70}$$

The linear relations between $\mathbf{v}_n$, $\mathbf{a}_n$ and $\dot{\mathbf{v}}_n$ in (37) result in linear relations for the corresponding global errors. Here and in the following we will always assume that the algorithmic parameters γ, α_m and α_f satisfy the order condition (41) and the local truncation errors are bounded by (70).

Lemma 4.3 *Consider global errors $\mathbf{e}_n^{\mathbf{a}}$ with $\dot{\mathbf{v}}(t_n + \Delta_\alpha h) = \mathbf{a}_n + \mathbf{e}_n^{\mathbf{a}}$ and use $(\bullet)(t_n) = (\bullet)_n + \mathbf{e}_n^{(\bullet)}$ to define $\mathbf{e}_n^{(\bullet)}$ for all remaining solution components being elements of linear spaces. The order condition (41) implies*

$$\mathbf{e}_{n+1}^{\mathbf{v}} = \mathbf{e}_n^{\mathbf{v}} + (1-\gamma)h\mathbf{e}_n^{\mathbf{a}} + \gamma h\mathbf{e}_{n+1}^{\mathbf{a}} + \mathcal{O}(h^3)\,, \tag{71a}$$

$$(1-\alpha_m)\mathbf{e}_{n+1}^{\mathbf{a}} + \alpha_m\mathbf{e}_n^{\mathbf{a}} = (1-\alpha_f)\mathbf{e}_{n+1}^{\dot{\mathbf{v}}} + \alpha_f\mathbf{e}_n^{\dot{\mathbf{v}}} + \mathcal{O}(h^2)\,. \tag{71b}$$

For linear configuration spaces G, the global error in $\mathbf{q}$ is given by $\mathbf{q}(t_n) = \mathbf{q}_n + \mathbf{e}_n^{\mathbf{q}}$. In the nonlinear case, we take into account the Lie group structure of the configuration space G and consider global errors $\widetilde{\mathbf{e}}_n^q$ being elements of the corresponding Lie algebra $\mathfrak{g}$:

$$q(t_n) = q_n \circ \exp(\widetilde{\mathbf{e}}_n^q)\,. \tag{72}$$

This definition is compatible with the classical definition of $\mathbf{e}_n^{\mathbf{q}} \in \mathbb{R}^k$ if the configuration space G is linear.

The position update (37a) and the definition (66) of the local error $\widetilde{\mathbf{l}}_n^q$ yield a global error recursion for $\widetilde{\mathbf{e}}_n^q$ in terms of matrix exponentials:

$$\begin{aligned}\exp(\widetilde{\mathbf{e}}_{n+1}^q) &= (q_{n+1})^{-1} \circ q(t_{n+1})\\ &= \exp(-h\widetilde{\boldsymbol{\Delta}\mathbf{q}}_n) \circ \underbrace{(q_n)^{-1} \circ q(t_n)}_{=\exp(\widetilde{\mathbf{e}}_n^q)} \circ \exp(h\widetilde{\boldsymbol{\Delta}\mathbf{q}}(t_n)) \circ \exp(\widetilde{\mathbf{l}}_n^q)\,.\end{aligned}$$

This product of matrix exponentials may be studied by repeated application of the Baker–Campbell–Hausdorff formula using Lemma 2.5. Omitting all technical details, we get

Lemma 4.4 (Arnold et al. 2015, Lemma 2) *The global errors* $\mathbf{e}_n^q$ *satisfy*

$$\mathbf{e}_{n+1}^q = \mathbf{e}_n^q + h\,\boldsymbol{\Delta}_h \mathbf{e}_n^q \tag{73}$$

with

$$\begin{aligned}\boldsymbol{\Delta}_h \widetilde{\mathbf{e}}_n^q = \widetilde{\mathbf{e}}_n^{\mathbf{v}} + (0.5-\beta)h\widetilde{\mathbf{e}}_n^{\mathbf{a}} + \beta h \widetilde{\mathbf{e}}_{n+1}^{\mathbf{a}} + [\widetilde{\mathbf{e}}_n^q, \widetilde{\mathbf{v}}(t_n)] + \frac{1}{h}\widetilde{\mathbf{l}}_n^q + \\ + \mathcal{O}(h)(\varepsilon_n + h\|\mathbf{e}_{n+1}^{\mathbf{a}}\|)\end{aligned} \tag{74}$$

and the notation

$$\varepsilon_n := \|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| + h\|\mathbf{e}_n^{\mathbf{a}}\| + h\|\mathbf{e}_n^{\boldsymbol{\lambda}}\| \tag{75}$$

that is used to summarize higher order error terms in compact form. In particular, Eqs. (73) and (74) and the local error estimate (69) imply

$$\mathbf{e}_{n+1}^q = \mathbf{e}_n^q + \mathcal{O}(h)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^3)\,, \tag{76a}$$

$$\|\boldsymbol{\Delta}_h \mathbf{e}_n^q\| \le \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2)\,. \tag{76b}$$

Error estimates like the ones in Lemma 4.4 are valid if the numerical solution remains in a small neighbourhood of the analytical one. More precisely, we suppose that there are positive constants h_0 and C and a sufficiently small constant $\gamma_0 > 0$ such that

$$\|\mathbf{e}_r^q\| \le Ch\,, \quad \|\mathbf{e}_r^{\mathbf{v}}\| + \|\mathbf{e}_r^{\mathbf{a}}\| + \|\mathbf{e}_r^{\boldsymbol{\lambda}}\| \le \gamma_0 \tag{77}$$

is satisfied for all $h \in (0, h_0]$ and all r with $t_0 + rh \in [t_0, t_{\text{end}}]$. This technical assumption may be verified using the results of the convergence analysis in Sect. 4.3 below, see (Hairer and Wanner 1996, Theorem VII.3.5) and the slightly more detailed discussion in (Arnold et al. 2015, Sect. 3.1).

Linearizing the equilibrium conditions (37e), we may estimate $\mathbf{e}_n^{\dot{\mathbf{v}}}$ in terms of ε_n and $\mathbf{e}_n^{\boldsymbol{\lambda}}$:

Lemma 4.5 (Arnold et al. 2015, Lemma 3) *If the order condition (41) is satisfied then*

$$\mathbf{e}_n^{\dot{\mathbf{v}}} + \mathbf{e}_n^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} = \mathcal{O}(1)\varepsilon_n\,, \quad \|\mathbf{e}_n^{\dot{\mathbf{v}}}\| = \mathcal{O}(1)(\varepsilon_n + \|\mathbf{e}_n^{\boldsymbol{\lambda}}\|)\,, \tag{78a}$$

$$\mathbf{e}_{n+1}^{\dot{\mathbf{v}}} + \mathbf{e}_{n+1}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}} = \mathcal{O}(1)\varepsilon_n + \mathcal{O}(h)(\|\mathbf{e}_{n+1}^{\mathbf{a}}\| + \|\mathbf{e}_{n+1}^{\boldsymbol{\lambda}}\|) + \mathcal{O}(h^3)\,. \tag{78b}$$

Here we used the notation $\mathbf{e}_n^{(\mathbf{C}\,\bullet)} := \mathbf{C}(q(t_n), \mathbf{v}(t_n), \boldsymbol{\lambda}(t_n), t_n)\mathbf{e}_n^{(\bullet)}$ *for matrix-valued functions* $\mathbf{C} = \mathbf{C}(q, \mathbf{v}, \boldsymbol{\lambda}, t)$.

Inserting (78) into the error estimate (71b), we get a coupled error recursion

$$(1-\alpha_m)\mathbf{e}^{\mathbf{a}}_{n+1} + \alpha_m \mathbf{e}^{\mathbf{a}}_n + (1-\alpha_f)\mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_{n+1} + \alpha_f \mathbf{e}^{\mathbf{M}^{-1}\mathbf{B}^\top\boldsymbol{\lambda}}_n = \\ = \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2) \tag{79}$$

that has to be studied separately in tangential and normal direction of the constraint manifold $\mathfrak{M} := \{\, q \in G \,:\, \boldsymbol{\Phi}(q) = \mathbf{0}\,\}$ to get optimal error bounds, see (Hairer and Wanner 1996). The error component in tangential direction is obtained by multiplication with a matrix $\mathbf{P}(q)$ that projects into the tangential space $T_q\mathfrak{M} = \ker \mathbf{B}(q)$. Such a projector $\mathbf{P}(q)$ is given by

$$\mathbf{P}(q) := \mathbf{I} - [\mathbf{M}^{-1}\mathbf{B}^\top\mathbf{S}^{-1}\mathbf{B}](q) \ \text{ with } \ \mathbf{S}(q) := [\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top](q) \tag{80}$$

since $\mathbf{PP} = \mathbf{P}$ and $\mathbf{BP} = \mathbf{B} - \mathbf{BM}^{-1}\mathbf{B}^\top\mathbf{S}^{-1}\mathbf{B} = \mathbf{B} - \mathbf{SS}^{-1}\mathbf{B} = \mathbf{0}$. Taking into account that this projector satisfies $\mathbf{PM}^{-1}\mathbf{B}^\top \equiv \mathbf{0}$, we get an optimal error recursion in tangential direction by left multiplication of (79) with matrix $\mathbf{P}\big(q(t_{n+1})\big)$. The error propagation in normal direction to the constrained manifold may be characterized multiplying (79) by $\mathbf{B}\big(q(t_{n+1})\big)$:

Lemma 4.6 (Arnold et al. 2015, Lemma 5) *The errors* $\mathbf{e}^{\mathbf{a}}_n$, $\mathbf{e}^{\boldsymbol{\lambda}}_n$ *satisfy*

$$(1-\alpha_m)\mathbf{e}^{\mathbf{Pa}}_{n+1} + \alpha_m \mathbf{e}^{\mathbf{Pa}}_n = \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2)\,, \tag{81}$$

$$(1-\alpha_m)\mathbf{e}^{\mathbf{Ba}}_{n+1} + \alpha_m \mathbf{e}^{\mathbf{Ba}}_n + (1-\alpha_f)\mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_{n+1} + \alpha_f \mathbf{e}^{\mathbf{S}\boldsymbol{\lambda}}_n = \\ = \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2) \tag{82}$$

and $\|\mathbf{e}^{\mathbf{a}}_n\| \le \|\mathbf{e}^{\mathbf{Pa}}_n\| + \|\mathbf{M}^{-1}\mathbf{B}^\top\mathbf{S}^{-1}\|\|\mathbf{e}^{\mathbf{Ba}}_n\| \le \mathcal{O}(1)(\|\mathbf{e}^{\mathbf{Pa}}_n\| + \|\mathbf{e}^{\mathbf{Ba}}_n\|)$.

Estimate (81) defines a one-step recursion for the tangential error component $\mathbf{e}^{\mathbf{Pa}}_n$ in terms of ε_n, ε_{n+1} and local errors $\mathcal{O}(h^2)$.

The most crucial part of the convergence analysis are recursive estimates for the error component $\mathbf{e}^{\mathbf{Ba}}_n$ in normal direction to the constrained manifold. Similar to the discussion in Sect. 3.2, we may scale the error recursion (71a) by the factor $1/h$ to get

$$(1-\gamma)\mathbf{e}^{\mathbf{Ba}}_n + \gamma \mathbf{e}^{\mathbf{Ba}}_{n+1} = \frac{\mathbf{e}^{\mathbf{Bv}}_{n+1} - \mathbf{e}^{\mathbf{Bv}}_n}{h} + \mathcal{O}(1)\varepsilon_n + \mathcal{O}(h^2)\,. \tag{83}$$

The scaled error term $\mathbf{e}^{\mathbf{Bv}}_n/h$ in the right-hand side of (83) is studied considering error estimate (74) and its equivalent in $\mathbb{R}^k$. We get

$$\frac{1}{h}\Big(\mathbf{e}^{\mathbf{Bv}}_n + \frac{1}{h}\mathbf{B}\big(q(t_n)\big)\mathbf{l}^q_n\Big) = \mathbf{r}^{\mathbf{B}}_n - \mathbf{r}_h(t_n, \mathbf{e}^q_n) + \mathcal{O}(1)\varepsilon_n + \mathcal{O}(h)\|\mathbf{e}^{\mathbf{a}}_{n+1}\| \tag{84}$$

with the vector

$$\mathbf{r}_n^{\mathbf{B}} := \frac{1}{h}\Big(\mathbf{B}\big(q(t_n)\big)\boldsymbol{\Delta}_h\mathbf{e}_n^q + \mathbf{Z}(q(t_n))\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big)\Big) - \\ -\mathbf{B}\big(q(t_n)\big)\big((0.5-\beta)\mathbf{e}_n^{\mathbf{a}} - \beta\mathbf{e}_{n+1}^{\mathbf{a}}\big) \tag{85}$$

and a vector-valued function

$$\mathbf{r}_h(t_n, \mathbf{e}_n^q) := \frac{1}{h}\Big(\mathbf{Z}(q(t_n))\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big) + \widehat{\mathbf{e}}_n^q\mathbf{v}(t_n)\Big) \tag{86}$$

that is linear in $\mathbf{e}_n^q$. Here, the term $\widehat{\mathbf{e}}_n^q\mathbf{v}(t_n) \in \mathbb{R}^k$ represents the matrix commutator $[\widetilde{\mathbf{e}}_n^q, \widetilde{\mathbf{v}}(t_n)] \in \mathfrak{g}$, see (29). By purpose, the notation $\mathbf{r}_n^{\mathbf{B}}$ in (84) adopts the notation r_n that was introduced in (48) to denote a scaled linear combination of global errors in v and local errors in q for proving second-order convergence for the linear test equation, see Sect. 3.2.

The definitions of $\mathbf{r}_n^{\mathbf{B}}$ and $\mathbf{r}_h(t_n, \mathbf{e}_n^q)$ contain a term $\mathbf{Z}(q(t_n))\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big)/h$ with the bilinear form $\mathbf{Z}(q)$ that is known from the hidden constraints (18) at the level of acceleration coordinates. A time discrete approximation of these hidden constraints shows that the first term in the right-hand side of (85) is of size $\mathcal{O}(1)(\|\mathbf{e}_n^q\| + \|\boldsymbol{\Delta}_h\mathbf{e}_n^q\|)$, see (88):

Lemma 4.7 (Arnold et al. 2015, Lemma 4) *The global errors* $\mathbf{e}_n^q \in \mathbb{R}^k$ *satisfy*

$$\mathbf{B}\big(q(t_n)\big)\mathbf{e}_n^q = \mathcal{O}(h)\|\mathbf{e}_n^q\|\,, \tag{87}$$

$$\mathbf{B}\big(q(t_n)\big)\boldsymbol{\Delta}_h\mathbf{e}_n^q + \mathbf{Z}\big(q(t_n)\big)\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big) = \mathcal{O}(h)(\|\mathbf{e}_n^q\| + \|\boldsymbol{\Delta}_h\mathbf{e}_n^q\|)\,. \tag{88}$$

Proof Taking into account that $\boldsymbol{\Phi}(q(t_n)) = \boldsymbol{\Phi}(q_n) = \mathbf{0}$, we consider $\boldsymbol{\Phi}(q_{n,\vartheta})$ for $q_{n,\vartheta} := q(t_n) \circ \exp(-\vartheta\widetilde{\mathbf{e}}_n^q) \in G$, ($\vartheta \in [0,1]$), and get

$$\mathbf{0} = -\big(\boldsymbol{\Phi}(q_n) - \boldsymbol{\Phi}(q(t_n))\big) = -\big(\boldsymbol{\Phi}(q_{n,1}) - \boldsymbol{\Phi}(q_{n,0})\big) = \int_0^1 \mathbf{B}(q_{n,\vartheta})\mathbf{e}_n^q \, \mathrm{d}\vartheta \tag{89}$$

since $\mathbf{B}(q_{n,\vartheta})\mathbf{e}_n^q = -(\mathrm{d}/\mathrm{d}\vartheta)\boldsymbol{\Phi}(q_{n,\vartheta})$, see (14). Assertion (87) follows from (89) because $\mathbf{B}(q_{n,\vartheta}) = \mathbf{B}\big(q(t_n)\big) + \mathcal{O}(h)$, see (77).

The proof of (88) is technically much more complicated and starts with the observation that

$$\mathbf{0} = \int_0^1 \frac{\mathbf{B}(q_{n+1,\vartheta})\mathbf{e}_{n+1}^q - \mathbf{B}(q_{n,\vartheta})\mathbf{e}_n^q}{h}\, \mathrm{d}\vartheta\,,$$

see (89). The integrand may be split into terms $\mathbf{B}(q_{n+1,\vartheta})(\mathbf{e}_{n+1}^q - \mathbf{e}_n^q)/h$ and $\big(\mathbf{B}(q_{n+1,\vartheta})\mathbf{e}_n^q - \mathbf{B}(q_{n,\vartheta})\mathbf{e}_n^q\big)/h$ that yield in (88) the terms $\mathbf{B}\big(q(t_n)\big)\boldsymbol{\Delta}_h\mathbf{e}_n^q$ and $\mathbf{Z}\big(q(t_n)\big)$ $\big(\mathbf{e}_n^q, \mathbf{v}(t_n)\big)$, respectively. For the detailed proof, we refer to (Arnold et al. 2015). □

Lemma 4.8 (Arnold et al. 2015, Lemma 6) *If* $\alpha_m \neq 1$, $\alpha_f \neq 1$, $\beta \neq 0$ *and the order condition (41) is satisfied then*

$$\mathbf{r}_n^{\mathbf{B}} + (0.5-\beta)\mathbf{e}_n^{\mathbf{Ba}} + \beta\mathbf{e}_{n+1}^{\mathbf{Ba}} = \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2)\,, \tag{90}$$

$$(1-\gamma)\mathbf{e}_n^{\mathbf{Ba}} + \gamma\mathbf{e}_{n+1}^{\mathbf{Ba}} = \mathbf{r}_{n+1}^{\mathbf{B}} - \mathbf{r}_n^{\mathbf{B}} + \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2)\,. \tag{91}$$

Proof (a) Inserting error estimate (88) in (85), we get

$$\mathbf{r}_n^{\mathbf{B}} + (0.5-\beta)\mathbf{e}_n^{\mathbf{Ba}} + \beta\mathbf{e}_{n+1}^{\mathbf{Ba}} = \mathcal{O}(1)(\|\mathbf{e}_n^q\| + \|\Delta_h\mathbf{e}_n^q\| + h\|\mathbf{e}_{n+1}^{\mathbf{a}}\|)\,,$$

and (90) follows from (76b).

(b) With the assumptions on the algorithmic parameters α_m, α_f, β and γ, we may substitute in (84) the term $\mathcal{O}(h)\|\mathbf{e}_{n+1}^{\mathbf{a}}\|$ by its upper bound $\mathcal{O}(1)\varepsilon_n + \mathcal{O}(h^2)$, see (Arnold et al. 2015, Corollary 1a). In this modified form, estimate (84) implies

$$\frac{\mathbf{e}_{n+1}^{\mathbf{Bv}} - \mathbf{e}_n^{\mathbf{Bv}}}{h} = \mathbf{r}_{n+1}^{\mathbf{B}} - \mathbf{r}_n^{\mathbf{B}} + \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2) \tag{92}$$

since $\|\mathbf{l}_{n+1}^q - \mathbf{l}_n^q\| = \mathcal{O}(h^4)$, see Lemma 4.2, and $\mathbf{r}_h(t_n, \mathbf{e}_n^q) = (\ldots)/h$ varies smoothly in n in the sense that

$$\begin{aligned}
&\mathbf{r}_h(t_{n+1}, \mathbf{e}_{n+1}^q) - \mathbf{r}_h(t_n, \mathbf{e}_n^q) \\
&= \big(\mathbf{r}_h(t_{n+1}, \mathbf{e}_{n+1}^q) - \mathbf{r}_h(t_{n+1}, \mathbf{e}_n^q)\big) + \big(\mathbf{r}_h(t_{n+1}, \mathbf{e}_n^q) - \mathbf{r}_h(t_n, \mathbf{e}_n^q)\big) \\
&= h\mathbf{r}_h(t_{n+1}, \Delta_h\mathbf{e}_n^q) + h\dot{\mathbf{r}}_h(t_n + \vartheta h, \mathbf{e}_n^q) = \mathcal{O}(1)\|\Delta_h\mathbf{e}_n^q\| + \mathcal{O}(1)\|\mathbf{e}_n^q\|
\end{aligned}$$

with some $\vartheta \in (0,1)$, see also the more detailed discussion in (Arnold et al. 2015, Lemma 6). Inserting (92) into (83), we get estimate (91). □

Finally, a one-step error recursion for the generalized-α Lie group integrator (37) may be formulated in terms of $\mathbf{r}_n^{\mathbf{B}}$ and the vector-valued global errors $\mathbf{e}_n^q$, $\mathbf{e}_n^{\mathbf{v}}$, $\mathbf{e}_n^{\mathbf{Pa}}$, $\mathbf{e}_n^{\mathbf{Ba}}$, $\mathbf{e}_n^{\mathbf{S\lambda}}$ combining (71a), (76a), (81), (82), (90) and (91) to

$$\|\mathbf{E}_{n+1}^{\mathbf{y}} - \mathbf{T}_{\mathbf{y}}\mathbf{E}_n^{\mathbf{y}}\| \leq \mathcal{O}(h)(\varepsilon_n + \varepsilon_{n+1} + \|\mathbf{E}_n^{\mathbf{z}}\| + \|\mathbf{E}_{n+1}^{\mathbf{z}}\|) + \mathcal{O}(h^3)\,, \tag{93a}$$

$$\|\mathbf{E}_{n+1}^{\mathbf{z}} - \mathbf{T}_{\mathbf{z}}\mathbf{E}_n^{\mathbf{z}}\| \leq \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2) \tag{93b}$$

with

$$\mathbf{E}_n^{\mathbf{y}} := \begin{pmatrix} \mathbf{e}_n^q \\ \mathbf{e}_n^{\mathbf{v}} \end{pmatrix}, \quad \mathbf{E}_n^{\mathbf{z}} := \begin{pmatrix} \mathbf{e}_n^{\mathbf{Pa}} \\ \mathbf{E}_n^{\mathbf{r}} \end{pmatrix}, \quad \mathbf{E}_n^{\mathbf{r}} := \begin{pmatrix} \mathbf{e}_n^{\mathbf{S\lambda}} \\ \mathbf{r}_n^{\mathbf{B}} \\ \mathbf{e}_n^{\mathbf{Ba}} \end{pmatrix}, \tag{94}$$

$$\mathbf{T}_{\mathbf{y}} := \mathbf{I}_{2k}\,, \quad \mathbf{T}_{\mathbf{z}} := \text{blockdiag}\,(-\frac{\alpha_m}{1-\alpha_m})\mathbf{I}_k,\ (\mathbf{T}_+^{-1}\mathbf{T}_0 \otimes \mathbf{I}_m)\,) \tag{95}$$

and

$$\mathbf{T}_+ := \begin{pmatrix} 0 & 0 & -\beta \\ 0 & 1 & -\gamma \\ 1-\alpha_f & 0 & 1-\alpha_m \end{pmatrix}, \quad \mathbf{T}_0 := \begin{pmatrix} 0 & 1 & 0.5-\beta \\ 0 & 1 & 1-\gamma \\ -\alpha_f & 0 & -\alpha_m \end{pmatrix}.$$

The one-step error recursion (93) couples the convergence analysis for unconstrained systems (error components $\mathbf{e}_n^q$, $\mathbf{e}_n^{\mathbf{v}}$, $\mathbf{e}_n^{\mathbf{Pa}}$) to error bounds for the Lagrange multipliers and other algebraic variables (error components $\mathbf{e}_n^{\boldsymbol{\lambda}}$, $\mathbf{r}_n^{\mathbf{B}}$, $\mathbf{e}_n^{\mathbf{Ba}}$). The latter ones are closely related to the error analysis for the linear test equation $\ddot{q} + \omega^2 q = 0$ in the limit case $h\omega \to \infty$, see Eqs. (46)–(48) in Sect. 3.2.

The error bounds (93) are the key to the convergence analysis of the DAE Lie group integrator (37), see Sect. 4.2 and Theorem 4.18 below. In the following, we will call this integrator the *index-3 integrator* since it results from the direct time discretization of the original index-3 formulation (15) of the equations of motion. With a slightly different definition of vectors $\mathbf{E}_n^{\mathbf{r}}$ and matrix $\mathbf{T}_{\mathbf{z}}$, error bounds (93) may also be proved for the *stabilized index-2 integrator* (56) that is based on the stabilized index-2 formulation (55) of the equations of motion. For this integrator, the time discrete approximation of hidden constraints yields:

Lemma 4.9 (see Arnold et al. 2015, Theorem 2)

(a) The auxiliary variables $\boldsymbol{\eta}_n$ in (56b) are of size $\|\boldsymbol{\eta}_n\| = \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2)$. Therefore, error estimate (76a) applies as well to integrator (56).

(b) For integrator (56), the error bounds in (84) and (91) get the form

$$\frac{1}{h}\mathbf{e}_n^{\mathbf{Bv}} = -\bar{\mathbf{r}}_h(t_n, \mathbf{e}_n^q) + \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2)\,, \tag{96}$$

$$(1-\gamma)\mathbf{e}_n^{\mathbf{Ba}} + \gamma\mathbf{e}_{n+1}^{\mathbf{Ba}} = \mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2) \tag{97}$$

with

$$\bar{\mathbf{r}}_h(t_n, \mathbf{e}_n^q) := \frac{1}{h}\mathbf{Z}(q(t_n))\big(\mathbf{v}(t_n), \mathbf{e}_n^q\big)\,.$$

Proof We sketch the basic ideas of the proof and refer to the proof of (Arnold et al. 2015, Theorem 2) for a more detailed discussion.

(a) For the stabilized index-2 formulation, the scaled increment $\Delta_h\mathbf{e}_n^q$ in (73) and (88) has to be substituted by $\Delta_h\mathbf{e}_n^q + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n$, see (56b). In this modified form, estimate (88) yields

$$\mathbf{B}\big(q(t_n)\big)\mathbf{B}^\top(q_n)\boldsymbol{\eta}_n = \mathcal{O}(1)(\|\mathbf{e}_n^q\| + \|\Delta_h\mathbf{e}_n^q\|) \tag{98}$$

with a right-hand side that is of size $\mathcal{O}(1)(\varepsilon_n + \varepsilon_{n+1}) + \mathcal{O}(h^2)$, see (76b). The assertion may be proved solving (98) w.r.t. $\boldsymbol{\eta}_n$ since the full rank assumption on $\mathbf{B}(q)$ implies that $\mathbf{B}\big(q(t_n)\big)\mathbf{B}^\top(q_n) = [\mathbf{B}\mathbf{B}^\top](q_n) + \mathcal{O}(h)$ is non-singular. Using this upper bound for $\|\boldsymbol{\eta}_n\|$, we get error estimate (76a) from $\mathbf{e}_{n+1}^q = \mathbf{e}_n^q + h(\Delta_h\mathbf{e}_n^q + \mathbf{B}^\top(q_n)\boldsymbol{\eta}_n)$.

(b) For the stabilized index-2 formulation, analytical and numerical solution satisfy the hidden constraints (16) resulting in

$$\mathbf{0} = \frac{\mathbf{B}\big(q(t_n)\big)\mathbf{v}(t_n) - \mathbf{B}(q_n)\mathbf{v}_n}{h} = \frac{1}{h}\mathbf{B}(q_n)\mathbf{e}_n^{\mathbf{v}} + \frac{\mathbf{B}\big(q(t_n)\big) - \mathbf{B}(q_n)}{h}\mathbf{v}(t_n) \tag{99}$$

with $\mathbf{B}(q_n)\mathbf{e}_n^{\mathbf{v}} = \mathbf{e}_n^{\mathbf{Bv}} + \mathcal{O}(h)\varepsilon_n$. For the analysis of the second term in the right-hand side of (99), we use ideas of the proof of Lemma 4.7 and take into account that

$$\big(\mathbf{B}\big(q(t_n)\big) - \mathbf{B}(q_n)\big)\mathbf{v}(t_n) = -\big(\mathbf{B}(q_{n,1}) - \mathbf{B}(q_{n,0})\big)\mathbf{v}(t_n)$$

with $q_{n,\vartheta} := q(t_n) \circ \exp(-\vartheta\widetilde{\mathbf{e}}_n^q)$. Because of

$$-\frac{\mathrm{d}}{\mathrm{d}\vartheta}\big(\mathbf{B}(q_{n,\vartheta})\mathbf{v}(t_n)\big) = \mathbf{Z}(q_{n,\vartheta})\big(\mathbf{v}(t_n), \mathbf{e}_n^q\big) = h\bar{\mathbf{r}}_h(t_n, \mathbf{e}_n^q) + \mathcal{O}(h^2)\varepsilon_n\,,$$

we get $\bar{\mathbf{r}}_h(t_n, \mathbf{e}_n^q) = \big(\mathbf{B}\big(q(t_n)\big) - \mathbf{B}(q_n)\big)\mathbf{v}(t_n)/h + \mathcal{O}(h)\varepsilon_n$ and estimate (96) is seen to be a consequence of (99). With (96), the one-step recursion (97) for error vectors $\mathbf{e}_n^{\mathbf{Ba}}$ may be proved as in Lemma 4.8. □

Because of Lemma 4.9(b), there is no need to consider vectors $\mathbf{r}_n^{\mathbf{B}}$ in the global error analysis of the stabilized index-2 integrator (56). Summarizing error estimates (71a), (76a), (81), (82) and (97), we get the one-step error recursion (93) with

$$\mathbf{T}_{\mathbf{y}} := \mathbf{I}_{2k}\,,\quad \mathbf{T}_{\mathbf{z}} := \text{blockdiag}\,\Big(-\frac{\alpha_m}{1-\alpha_m}\mathbf{I}_k,\ (\bar{\mathbf{T}}_+^{-1}\bar{\mathbf{T}}_0 \otimes \mathbf{I}_m)\Big) \tag{100}$$

and

$$\mathbf{E}_n^{\mathbf{r}} := \begin{pmatrix} \mathbf{e}_n^{\mathbf{S}\boldsymbol{\lambda}} \\ \mathbf{e}_n^{\mathbf{Ba}} \end{pmatrix},\quad \bar{\mathbf{T}}_+ := \begin{pmatrix} 0 & -\gamma \\ 1-\alpha_f & 1-\alpha_m \end{pmatrix},\quad \bar{\mathbf{T}}_0 := \begin{pmatrix} 0 & 1-\gamma \\ -\alpha_f & -\alpha_m \end{pmatrix}.$$

4.2 *Coupled Error Propagation in Differential and Algebraic Solution Components*

The classical convergence analysis of ODE one-step methods provides the basis for investigating the coupled error propagation in differential and algebraic solution components of DAE Lie group integrators. We start this section with a perturbation analysis for ODE initial value problems (Theorem 4.10) and consider in Theorem 4.11 the corresponding convergence result for ODE one-step methods. The main new result of this section is the extension of this convergence analysis to the DAE case, see Theorem 4.16.

Theorem 4.10 (see Walter 1998) *Consider the initial value problem*

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)) \,, \;\; (\, t \in [t_0, t_{\text{end}}]\,) \,, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \tag{101}$$

with a continuous right-hand side $\mathbf{f}$ *that satisfies for all* $t \in [t_0, t_{\text{end}}]$ *a Lipschitz condition w.r.t.* $\mathbf{x}$ *with a Lipschitz constant* $L > 0$. *For functions* $\hat{\mathbf{x}} \in C^1[t_0, t_{\text{end}}]$ *with*

$$\dot{\hat{\mathbf{x}}}(t) = \mathbf{f}(t, \hat{\mathbf{x}}(t)) + \boldsymbol{\delta}(t) \,, \;\; (\, t \in [t_0, t_{\text{end}}]\,) \,, \tag{102}$$

the influence of perturbations $\boldsymbol{\delta}(t)$ *may be estimated by*

$$\|\hat{\mathbf{x}}(t) - \mathbf{x}(t)\| \le \mathrm{e}^{L(t-t_0)} \|\hat{\mathbf{x}}(t_0) - \mathbf{x}(t_0)\| + \frac{\mathrm{e}^{L(t-t_0)} - 1}{L} \max_{s \in [t_0, t_{\text{end}}]} \|\boldsymbol{\delta}(s)\| \,. \tag{103}$$

Proof For $t \in [t_0, t_{\text{end}}]$, we have

$$\begin{aligned} \hat{\mathbf{x}}(t) - \mathbf{x}(t) &= \hat{\mathbf{x}}(t_0) - \mathbf{x}(t_0) + \int_{t_0}^{t} \big(\dot{\hat{\mathbf{x}}}(s) - \dot{\mathbf{x}}(s)\big)\, \mathrm{d}s \\ &= \hat{\mathbf{x}}(t_0) - \mathbf{x}(t_0) + \int_{t_0}^{t} \big(\mathbf{f}(s, \hat{\mathbf{x}}(s)) - \mathbf{f}(s, \mathbf{x}(s))\big)\, \mathrm{d}s + \int_{t_0}^{t} \boldsymbol{\delta}(s)\, \mathrm{d}s \,. \end{aligned}$$

Therefore, the triangle inequality and the Lipschitz condition on $\mathbf{f}$ imply

$$\|\hat{\mathbf{x}}(t) - \mathbf{x}(t)\| \le \psi(t) \tag{104}$$

with the continuously differentiable function

$$\psi(t) := \|\hat{\mathbf{x}}(t_0) - \mathbf{x}(t_0)\| + L \int_{t_0}^{t} \|\hat{\mathbf{x}}(s) - \mathbf{x}(s)\|\, \mathrm{d}s + (t - t_0)\Delta$$

and $\Delta := \max_{s \in [t_0, t_{\text{end}}]} \|\boldsymbol{\delta}(s)\|$. Note, that $\max_s \|\boldsymbol{\delta}(s)\|$ is well defined since $\hat{\mathbf{x}} \in C^1[t_0, t_{\text{end}}]$ implies that $\boldsymbol{\delta}$ is continuous on the compact interval $[t_0, t_{\text{end}}]$.

Because of (104), the time derivative of ψ satisfies for all $t \in [t_0, t_{\text{end}}]$ the estimate

$$\dot{\psi}(t) = L\|\hat{\mathbf{x}}(t) - \mathbf{x}(t)\| + \Delta \le L\psi(t) + \Delta \,.$$

Hence, the derivative of $\sigma(\tau) := \mathrm{e}^{L(t-\tau)}\psi(\tau)$ is bounded by

$$\sigma'(\tau) = \mathrm{e}^{L(t-\tau)}\big(-L\psi(\tau) + \dot{\psi}(\tau)\big) \le \mathrm{e}^{L(t-\tau)}\,\Delta$$

and we get

$$\sigma(t) = \sigma(s) + \int_{s}^{t} \sigma'(\tau)\, \mathrm{d}\tau \le \sigma(s) + \int_{s}^{t} \mathrm{e}^{L(t-\tau)}\, \mathrm{d}\tau \cdot \Delta \,,$$

i.e.,

$$\psi(t) \le \mathrm{e}^{L(t-s)}\psi(s) + \int_s^t \mathrm{e}^{L(t-\tau)}\,\mathrm{d}\tau \cdot \Delta \tag{105}$$

for any $s \in [t_0, t_{\mathrm{end}}]$. Error bound (105) with $s = t_0$ proves (103) since

$$\int_{t_0}^t \mathrm{e}^{L(t-\tau)}\,\mathrm{d}\tau = \frac{\mathrm{e}^{L(t-t_0)} - 1}{L} \tag{106}$$

and $\psi(t_0) = \|\hat{\mathbf{x}}(t_0) - \mathbf{x}(t_0)\|$. □

For the numerical solution of ODE (101), we consider a one-step method that updates the numerical solution in time step $t_n \to t_{n+1} = t_n + h_n$ according to

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h_n \boldsymbol{\Phi}_n(t_n, \mathbf{x}_n; \mathbf{f}, h_n) \tag{107}$$

with a continuous increment function $\boldsymbol{\Phi}$ that satisfies a Lipschitz condition w.r.t. $\mathbf{x}_n$ with a Lipschitz constant $L_{\boldsymbol{\Phi}} > 0$, see, e.g., (Hairer et al. 1993). The time discretization error in one single time step defines the *local error*

$$\mathbf{le}_n := \mathbf{x}(t_{n+1}) - \big(\mathbf{x}(t_n) + h_n \boldsymbol{\Phi}(t_n, \mathbf{x}(t_n); \mathbf{f}, h_n)\big)\,.$$

In the global error analysis, the accumulation of these local errors during time integration is studied by a discrete counterpart to the perturbation analysis for the continuous problem (see Theorem 4.10).

Theorem 4.11 *The global errors* $\mathbf{e}_n := \mathbf{x}(t_n) - \mathbf{x}_n$ *satisfy the error recursion*

$$\|\mathbf{e}_{n+1} - \mathbf{e}_n\| \le L_{\boldsymbol{\Phi}} h_n \|\mathbf{e}_n\| + \|\mathbf{le}_n\| \tag{108}$$

that results in the global error estimate

$$\|\mathbf{e}_n\| \le \mathrm{e}^{L_{\boldsymbol{\Phi}}(t_n - t_0)}\|\mathbf{e}_0\| + \frac{\mathrm{e}^{L_{\boldsymbol{\Phi}}(t_n-t_0)} - 1}{L_{\boldsymbol{\Phi}}} \max_{0 \le l < n} \frac{1}{h_l}\|\mathbf{le}_l\|\,. \tag{109}$$

Proof (a) Using the definition of local and global errors, we get

$$\begin{aligned}
\mathbf{e}_{n+1} - \mathbf{e}_n &= \big(\mathbf{x}(t_{n+1}) - \mathbf{x}_{n+1}\big) - \big(\mathbf{x}(t_n) - \mathbf{x}_n\big) \\
&= \mathbf{x}(t_{n+1}) - \big(\mathbf{x}(t_n) + h_n \boldsymbol{\Phi}(t_n, \mathbf{x}(t_n); \mathbf{f}, h_n)\big) + \\
&\qquad\qquad + h_n \boldsymbol{\Phi}(t_n, \mathbf{x}(t_n); \mathbf{f}, h_n) - \big(\mathbf{x}_{n+1} - \mathbf{x}_n\big) \\
&= \mathbf{le}_n + h_n\big(\boldsymbol{\Phi}(t_n, \mathbf{x}(t_n); \mathbf{f}, h_n) - \boldsymbol{\Phi}(t_n, \mathbf{x}_n; \mathbf{f}, h_n)\big)\,.
\end{aligned}$$

Therefore, estimate (108) follows from the triangle inequality and from the Lipschitz condition on $\mathbf{\Phi}$:

$$\|\mathbf{e}_{n+1} - \mathbf{e}_n\| \le \|\mathbf{le}_n\| + h_n L_{\mathbf{\Phi}} \|\mathbf{x}(t_n) - \mathbf{x}_n\| = L_{\mathbf{\Phi}} h_n \|\mathbf{e}_n\| + \|\mathbf{le}_n\| \,.$$

(b) Estimate (108) with n being substituted by some $r \in \{0, 1, \dots, n\}$ implies

$$\|\mathbf{e}_{r+1}\| \le \|\mathbf{e}_r\| + L_{\mathbf{\Phi}} h_r \|\mathbf{e}_r\| + \|\mathbf{le}_r\| = (1 + L_{\mathbf{\Phi}} h_r)\|\mathbf{e}_r\| + \|\mathbf{le}_r\| \tag{110}$$

with $h_r = t_{r+1} - t_r$. For a recursive application of this error estimate, we substitute the coefficients of $\|\mathbf{e}_r\|$ and $\|\mathbf{le}_r\|$ in the right-hand side of (110) by upper bounds that are obtained from $1 + Lt \le \mathrm{e}^{Lt}$ and

$$1 = \frac{t_{r+1} - t_r}{h_r} = \frac{1}{h_r} \int_{t_r}^{t_{r+1}} \mathrm{d}\tau \le \frac{1}{h_r} \int_{t_r}^{t_{r+1}} \mathrm{e}^{L_{\mathbf{\Phi}}(t_{r+1} - \tau)} \, \mathrm{d}\tau$$

and get

$$\|\mathbf{e}_{r+1}\| \le \mathrm{e}^{L_{\mathbf{\Phi}}(t_{r+1} - t_r)} \|\mathbf{e}_r\| + \int_{t_r}^{t_{r+1}} \mathrm{e}^{L_{\mathbf{\Phi}}(t_{r+1} - \tau)} \, \mathrm{d}\tau \cdot \frac{1}{h_r} \|\mathbf{le}_r\| \,. \tag{111}$$

(c) Estimate (111) is a special case of the more general expression

$$\|\mathbf{e}_n\| \le \mathrm{e}^{L_{\mathbf{\Phi}}(t_n - t_r)} \|\mathbf{e}_r\| + \int_{t_r}^{t_n} \mathrm{e}^{L_{\mathbf{\Phi}}(t_n - \tau)} \, \mathrm{d}\tau \cdot \max_{r \le l < n} \frac{1}{h_l} \|\mathbf{le}_l\| \,, \tag{112}$$

($r = 0, 1, \dots, n-1$), that may be considered as a time discrete counterpart to (105). To prove the error bound (112) by induction, we observe that (111) is estimate (112) with $r = n - 1$. For the induction step, we suppose that (112) is satisfied for $r + 1$:

$$\|\mathbf{e}_n\| \le \mathrm{e}^{L_{\mathbf{\Phi}}(t_n - t_{r+1})} \|\mathbf{e}_{r+1}\| + \int_{t_{r+1}}^{t_n} \mathrm{e}^{L_{\mathbf{\Phi}}(t_n - \tau)} \, \mathrm{d}\tau \cdot \max_{r+1 \le l < n} \frac{1}{h_l} \|\mathbf{le}_l\| \,.$$

Inserting in this expression the upper bound (111) for $\|\mathbf{e}_{r+1}\|$, we get estimate (112) since

$$\mathrm{e}^{L_{\mathbf{\Phi}}(t_n - t_{r+1})} \mathrm{e}^{L_{\mathbf{\Phi}}(t_{r+1} - \tau)} = \mathrm{e}^{L_{\mathbf{\Phi}}(t_n - \tau)}$$

for any $\tau \in [t_r, t_{r+1}]$.

(d) To complete the proof, we use the identity (106) and see that (112) with $r = 0$ proves the global error bound (109). □

Abstracting from the specific setting in Theorem 4.11, we may consider more general one-step error recursions and the resulting error bounds. For simplicity, we restrict this analysis to constant time step sizes h. In that case, we may substitute the term $\|\mathbf{le}_r\|$ in (110) by hM with an appropriate constant $M \ge 0$ and get a one-step recursion

$$u_{n+1} \leq (1 + Lh)u_n + hM\,, \tag{113}$$

($n \geq 0$), that implies

$$u_n \leq \mathrm{e}^{L(t_n - t_0)} u_0 + \frac{\mathrm{e}^{L(t_n - t_0)} - 1}{L} M \tag{114}$$

with $u_n := \|\mathbf{e}_n\|$, $L := L_{\mathbf{\Phi}} > 0$ and $t_n := t_0 + nh$, see (109). The convergence analysis of Theorem 4.11 may be generalized straightforwardly to more complex error recursions:

Lemma 4.12 *Consider sequences* $(v_n)_{n \geq 0}$, $(w_n)_{n \geq 0}$ *of non-negative numbers that satisfy*

$$v_{n+1} \leq (1 + Lh)v_n + Lh\kappa^n e_0 + hM\,, \tag{115a}$$

$$w_{n+1} \leq (\kappa + Lh)w_n + Lh\kappa^n e_0 + M \tag{115b}$$

with a positive constant L and non-negative constants $\kappa \in [0, 1)$, *M and* e_0. *All these constants are supposed to be independent of* $h > 0$ *and* $n \geq 0$.

Using the notation $t_n := t_0 + nh$, *we get for all* $n \geq 0$ *the estimate*

$$v_n \leq \mathrm{e}^{L(t_n - t_0)} \big(v_0 + h\, \frac{Le_0}{1 - \kappa} \big) + \frac{\mathrm{e}^{L(t_n - t_0)} - 1}{L} M\,. \tag{116a}$$

For the sequence $(w_n)_{n \geq 0}$, *an estimate*

$$w_n \leq (\kappa + Lh)^n w_0 + h\, \frac{Le_0}{1 - \kappa} + \frac{M}{1 - (\kappa + Lh)} \tag{116b}$$

may be shown for all $n \geq 0$ *and all* $h \in (0, h_0]$ *with* $h_0 > 0$ *denoting a constant such that* $\kappa + Lh_0 < 1$.

Proof Following part (b) of the proof of Theorem 4.11, we rewrite the one-step error recursions (115) in a form that is appropriate for recursive application:

$$\begin{aligned} v_{r+1} &\leq \mathrm{e}^{L(t_{r+1} - t_r)} \big(v_r + Lh\kappa^r e_0 \big) + \int_{t_r}^{t_{r+1}} \mathrm{e}^{L(t_{r+1} - \tau)}\, \mathrm{d}\tau \cdot M\,, \\ w_{r+1} &\leq (\kappa + Lh)w_r + Lh\kappa^r e_0 + M\,. \end{aligned}$$

Then, the error bounds

$$v_n \leq \mathrm{e}^{L(t_n - t_r)} \big(v_r + h \sum_{l=r}^{n-1} \kappa^l \cdot Le_0 \big) + \int_{t_r}^{t_n} \mathrm{e}^{L(t_n - \tau)}\, \mathrm{d}\tau \cdot M\,, \tag{117a}$$

$$w_n \leq (\kappa + Lh)^{n-r} w_r + h \sum_{l=r}^{n-1} \kappa^l \cdot Le_0 + \sum_{l=r}^{n-1} (\kappa + Lh)^{n-(l+1)} \cdot M\,, \tag{117b}$$

($r = 0, 1, \dots, n-1$), follow (similar to part (c) of the proof of Theorem 4.11) by induction starting at $r = n-1$. In the induction step, we have to take into account that

$$\mathrm{e}^{L(t_n - t_r)} \kappa^r + \mathrm{e}^{L(t_n - t_{r+1})} \sum_{l=r+1}^{n-1} \kappa^l \le \mathrm{e}^{L(t_n - t_r)} \sum_{l=r}^{n-1} \kappa^l \,.$$

and $(\kappa + Lh)^{n-(r+1)} < 1$ for any $h \in (0, h_0]$. Error bounds (117) with $r = 0$ prove the lemma since $\kappa \in [0, 1)$ and $\kappa + Lh \in [0, 1)$ imply

$$\sum_{l=r}^{n-1} \kappa^l \le \sum_{l=0}^{\infty} \kappa^l = \frac{1}{1-\kappa} \,, \quad \sum_{l=r}^{n-1} (\kappa + Lh)^{n-(l+1)} \le \frac{1}{1-(\kappa + Lh)}$$

and the integral term in (117a) may be evaluated in closed form, see (106). □

Lemma 4.13 *Let* $(\mathbf{E}_n)_{n \ge 0}$ *be a sequence of vectors that satisfy*

$$\|\mathbf{E}_{n+1} - \mathbf{T}\mathbf{E}_n\| \le L_0 (h \|\mathbf{E}_n\| + h \|\mathbf{E}_{n+1}\|) + h M_0 \tag{118}$$

with a matrix $\mathbf{T}$ *and positive constants* L_0, M_0 *that are independent of* $h > 0$ *and* $n \ge 0$. *If there is a norm* $\|.\|_\varrho$ *such that* $\kappa_\varrho := \|\mathbf{T}\|_\varrho \le 1$ *then (118) implies for time step sizes* $h \in (0, h_0]$ *a one-step recursion*

$$\|\mathbf{E}_{n+1} - \mathbf{T}^{n+1}\mathbf{E}_0\|_\varrho \le (\kappa_\varrho + \tilde{L}_0 h) \|\mathbf{E}_n - \mathbf{T}^n \mathbf{E}_0\|_\varrho + \tilde{L}_0 h \kappa_\varrho^n \|\mathbf{E}_0\|_\varrho + h \tilde{M}_0 \tag{119}$$

and error bounds

$$\|\mathbf{E}_n\| \le \|\mathbf{T}^n \mathbf{E}_0\| + C_0 \|\mathbf{E}_n - \mathbf{T}^n \mathbf{E}_0\|_\varrho \,, \tag{120a}$$

$$\|\mathbf{E}_n\| \le C_0 (\|\mathbf{E}_0\|_\varrho + \|\mathbf{E}_n - \mathbf{T}^n \mathbf{E}_0\|_\varrho) \tag{120b}$$

with appropriate constants h_0, $\tilde{L}_0$, $\tilde{M}_0$ *and* C_0 *that are supposed to be positive. They depend on the norm* $\|.\|$ *and on the constants* L_0, M_0 *in (118).*

Proof (a) Since all norms in a finite-dimensional vector space are equivalent, there are positive constants $\underline{c}$, $\overline{c}$ with

$$\underline{c} \|\mathbf{E}\|_\varrho \le \|\mathbf{E}\| \le \overline{c} \|\mathbf{E}\|_\varrho \tag{121}$$

for any vector $\mathbf{E}$. Therefore, estimate (118) implies

$$\|\mathbf{E}_{n+1} - \mathbf{T}\mathbf{E}_n\|_\varrho \le \hat{L}_0 (h \|\mathbf{E}_n\|_\varrho + h \|\mathbf{E}_{n+1}\|_\varrho) + h \hat{M}_0 \tag{122}$$

with $\hat{L}_0 := \overline{c} L_0 / \underline{c}$, $\hat{M}_0 := M_0 / \underline{c}$.

(b) For the proof of estimate (119), we use the triangle inequality and get

$$\|\mathbf{E}_{n+1} - \mathbf{T}^{n+1}\mathbf{E}_0\|_\varrho \le \|\mathbf{E}_{n+1} - \mathbf{T}\mathbf{E}_n\|_\varrho + \|\mathbf{T}(\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0)\|_\varrho \,.$$

The term $\|\mathbf{T}(\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0)\|_\varrho$ is bounded by $\kappa_\varrho\|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho$ with $\kappa_\varrho = \|\mathbf{T}\|_\varrho \le 1$. We obtain

$$\|\mathbf{E}_{n+1} - \mathbf{T}^{n+1}\mathbf{E}_0\|_\varrho \le \kappa_\varrho\|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho + \|\mathbf{E}_{n+1} - \mathbf{T}\mathbf{E}_n\|_\varrho$$

and may substitute $\|\mathbf{E}_{n+1} - \mathbf{T}\mathbf{E}_n\|_\varrho$ by the upper bound (122) taking into account that

$$\|\mathbf{E}_n\|_\varrho \le \|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho + \|\mathbf{T}\|_\varrho^n\,\|\mathbf{E}_0\| = \|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho + \kappa_\varrho^n\|\mathbf{E}_0\|_\varrho \,.$$

The resulting inequality

$$\begin{aligned}&(1-\hat{L}_0 h)\|\mathbf{E}_{n+1} - \mathbf{T}^{n+1}\mathbf{E}_0\|_\varrho\\ &\qquad \le (\kappa_\varrho + \hat{L}_0 h)\|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho + 2\hat{L}_0 h\kappa_\varrho^n\|\mathbf{E}_0\|_\varrho + h\hat{M}_0\end{aligned}$$

is multiplied by $1/(1-\hat{L}_0 h)$ to get an upper bound for $\|\mathbf{E}_{n+1} - \mathbf{T}^{n+1}\mathbf{E}_0\|_\varrho$. If we suppose that $h \in (0, h_0]$ with $h_0 := 1/(2\hat{L}_0)$ then $1 - \hat{L}_0 h \ge 1/2$ and we may use the inequalities $(\kappa_\varrho + x)/(1-x) \le \kappa_\varrho + 4x$ and $1/(1-x) \le 2$ that are valid for all $x \in [0, 1/2]$. To complete the proof of (119), we set $\tilde{L}_0 := 4\hat{L}_0$ and $\tilde{M}_0 := 2\hat{M}_0$.

(c) Because of $\|\mathbf{E}_n\| \le \|\mathbf{T}^n\mathbf{E}_0\| + \|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|$, error bound (120a) with $C_0 := \bar{c}$ follows from the equivalence of norms $\|.\|$ and $\|.\|_\varrho$, see (121). With this definition of C_0, we have furthermore $\|\mathbf{E}_n\| \le C_0\|\mathbf{E}_n\|_\varrho$ and (120b) results from $\|\mathbf{E}_n\|_\varrho \le \|\mathbf{T}^n\mathbf{E}_0\|_\varrho + \|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho$ with $\|\mathbf{T}^n\|_\varrho \le \kappa_\varrho^n \le 1$. □

Corollary 4.14 *If the assumptions of Lemma 4.13 are satisfied with $\kappa_\varrho = \|\mathbf{T}\|_\varrho = 1$ then estimates (119) and (120b) imply*

$$\|\mathbf{E}_n\| \le \tilde{C}_0\big(\mathrm{e}^{\tilde{L}_0(t_n-t_0)}\|\mathbf{E}_0\| + \frac{\mathrm{e}^{\tilde{L}_0(t_n-t_0)}-1}{\tilde{L}_0}\tilde{M}_0\big) \tag{123}$$

with $t_n := t_0 + nh$, ($n \ge 0$), and a constant $\tilde{C}_0 > 0$ that depends on C_0 and the norm $\|.\|$.

Proof For $\kappa_\varrho = 1$, estimate (119) gets the form (113) with the notations $u_n := \|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho$, $L := \tilde{L}_0$ and $M := \tilde{L}_0\|\mathbf{E}_0\|_\varrho + \tilde{M}_0$. Inserting these expressions in error bound (114), we get

$$\|\mathbf{E}_n - \mathbf{T}^n\mathbf{E}_0\|_\varrho \le (\mathrm{e}^{\tilde{L}_0(t_n-t_0)} - 1)\|\mathbf{E}_0\|_\varrho + \frac{\mathrm{e}^{\tilde{L}_0(t_n-t_0)}-1}{\tilde{L}_0}\tilde{M}_0$$

since $u_0 = \|\mathbf{E}_0 - \mathbf{T}^0\mathbf{E}_0\|_\varrho = 0$. Therefore, the assertion of the corollary follows directly from (120b) if constant $\tilde{C}_0$ is set to $\tilde{C}_0 := C_0/\min\{1, \underline{c}\}$ such that $C_0\|\mathbf{E}_0\|_\varrho \le \tilde{C}_0\|\mathbf{E}_0\|$ and $C_0\tilde{M}_0 \le \tilde{C}_0\tilde{M}_0$, see (121). □

Remark 4.15 (a) For constant time step sizes $h_n = h = \text{const}$, the convergence result in Theorem 4.11 is a special case of the error analysis in Lemma 4.13 and Corollary 4.14 with $\mathbf{E}_n = \mathbf{e}_n$, $\mathbf{T} = \mathbf{I}$, $\tilde{C}_0 = 1$, $\tilde{L}_0 = L_{\mathbf{\Phi}}$ and $M = \max_l \|\mathbf{le}_l\|/h$.

(b) In ODE time integration, the error estimate of Corollary 4.14 is used to prove the convergence of linear multi-step methods by an equivalent one-step formulation, see (Hairer et al. 1993, Sect. III.4). For a k-step method, vector $\mathbf{E}_n$ is composed of global errors $\mathbf{e}_{n-j}$ at k consecutive grid points $t_{n-(k-1)}, \dots, t_{n-1}, t_n$ and matrix $\mathbf{T}$ has a Kronecker product structure $\mathbf{T} = \mathbf{A} \otimes \mathbf{I}$ with a companion matrix $\mathbf{A} \in \mathbb{R}^{k\times k}$ that satisfies $\|\mathbf{A}\|_\varrho = 1$ in a suitable norm $\|.\|_\varrho$ if the method is zero-stable. For a more detailed discussion of this convergence analysis, the interested reader is referred to the above cited reference.

(c) For matrices $\mathbf{T}$ with spectral radius $\varrho(\mathbf{T}) = 1$, the transformation to Jordan canonical form may be used to construct a norm $\|.\|_\varrho$ with $\|\mathbf{T}\|_\varrho = 1$ provided that all Jordan blocks corresponding to eigenvalues $\lambda_i[\mathbf{T}]$ with $|\lambda_i[\mathbf{T}]| = 1$ are of dimension 1×1, see (Hairer et al. 1993, Lemma III.4.4).

With appropriate matrices $\mathbf{T}$ of norm $\|\mathbf{T}\|_\varrho = 1$, Lemma 4.13 and Corollary 4.14 provide a unified framework for the error analysis of one-step and multi-step methods in ODE time integration. Corollary 4.14 may be generalized to the technically more challenging DAE case that is characterized by a coupled error propagation in differential and algebraic solution components. The error analysis employs two different error propagation matrices satisfying $\|\mathbf{T_y}\|_{\mathbf{y},\varrho} = 1$ and $\|\mathbf{T_z}\|_{\mathbf{z},\varrho} < 1$, respectively. It is inspired by the classical convergence analysis of one-step methods for index-1 DAEs in (Deuflhard et al. 1987), see also (Arnold et al. 2015, Lemma 7).

Theorem 4.16 *Let $(\mathbf{E}_n^\mathbf{y})_{n\ge 0}$ and $(\mathbf{E}_n^\mathbf{z})_{n\ge 0}$ be sequences of vectors that satisfy*

$$\|\mathbf{E}_{n+1}^\mathbf{y} - \mathbf{T_y}\mathbf{E}_n^\mathbf{y}\| \le L_0 h(\|\mathbf{E}_n^\mathbf{y}\| + \|\mathbf{E}_{n+1}^\mathbf{y}\| + \|\mathbf{E}_n^\mathbf{z}\| + \|\mathbf{E}_{n+1}^\mathbf{z}\|) + hM_0, \tag{124a}$$

$$\|\mathbf{E}_{n+1}^\mathbf{z} - \mathbf{T_z}\mathbf{E}_n^\mathbf{z}\| \le L_0(\|\mathbf{E}_n^\mathbf{y}\| + \|\mathbf{E}_{n+1}^\mathbf{y}\| + h\|\mathbf{E}_n^\mathbf{z}\| + h\|\mathbf{E}_{n+1}^\mathbf{z}\|) + M_0 \tag{124b}$$

with matrices $\mathbf{T_y}$, $\mathbf{T_z}$ and positive constants L_0, M_0 that are independent of $h > 0$ and $n \ge 0$. If there are norms $\|.\|_{\mathbf{y},\varrho}$, $\|.\|_{\mathbf{z},\varrho}$ such that $\|\mathbf{T_y}\|_{\mathbf{y},\varrho} = 1$ and $\|\mathbf{T_z}\|_{\mathbf{z},\varrho} < 1$ then (124) implies for time step sizes $h \in (0, h_0]$ error bounds

$$\|\mathbf{E}_n^\mathbf{y}\| \le \mathrm{e}^{\bar{L}_0(t_n-t_0)}(\|\mathbf{E}_0^\mathbf{y}\| + \bar{C}_0 h\|\mathbf{E}_0^\mathbf{z}\|) + \frac{\mathrm{e}^{\bar{L}_0(t_n-t_0)} - 1}{\bar{L}_0}\bar{M}_0\,, \tag{125a}$$

$$\|\mathbf{E}_n^\mathbf{z} - \mathbf{T}_\mathbf{z}^n\mathbf{E}_0^\mathbf{z}\| \le \bar{C}_0\mathrm{e}^{\bar{L}_0(t_n-t_0)}(\|\mathbf{E}_0^\mathbf{y}\| + h\|\mathbf{E}_0^\mathbf{z}\| + \bar{M}_0) \tag{125b}$$

with $t_n := t_0 + nh$, ($n \ge 0$). The constants h_0, $\bar{C}_0$, $\bar{L}_0$ and $\bar{M}_0$ are supposed to be positive. They depend on constants L_0, M_0 in (124) and may depend furthermore on the vector norms $\|.\| = \|.\|_\mathbf{y}$ and $\|.\| = \|.\|_\mathbf{z}$ for $\mathbf{E}_n^\mathbf{y}$ and $\mathbf{E}_n^\mathbf{z}$.

Proof (a) Using the same arguments as in parts (a) and (c) of the proof of Lemma 4.13, we may verify that the assertion of the Theorem (with appropriate norm dependent constants $\bar{C}_0$, $\bar{L}_0$ and $\bar{M}_0$) is valid for *any* pair of norms $(\|.\|_{\mathbf{y}}, \|.\|_{\mathbf{z}})$ if it is valid for one specific pair $(\|.\|_{\mathbf{y},*}, \|.\|_{\mathbf{z},*})$. To simplify the notation, we will therefore restrict the error analysis to a pair of norms with $\kappa_{\mathbf{y}} := \|\mathbf{T}_{\mathbf{y}}\|_{\mathbf{y}} = 1$ and $\kappa_{\mathbf{z}} := \|\mathbf{T}_{\mathbf{z}}\|_{\mathbf{z}} < 1$ and will furthermore omit the indices $\mathbf{y}$ and $\mathbf{z}$ at the norm symbol $\|.\|$.

(b) Similar to Lemma 4.13 and Corollary 4.14, the coupled error propagation is studied in terms of sequences $(u_n)_{n\geq 0}$, $(w_n)_{n\geq 0}$ with

$$u_n := \|\mathbf{E}_n^{\mathbf{y}} - \mathbf{T}_{\mathbf{y}}^n \mathbf{E}_0^{\mathbf{y}}\|\,, \quad w_n := \|\mathbf{E}_n^{\mathbf{z}} - \mathbf{T}_{\mathbf{z}}^n \mathbf{E}_0^{\mathbf{z}}\|\,. \tag{126}$$

For a one-step error recursion, we look for error bounds like (119) for u_{n+1} and w_{n+1}. As in Lemma 4.13, we get from assumptions (124) the estimates

$$u_{n+1} \leq (1 + \tilde{L}_0 h) u_n + \tilde{L}_0 h w_n + \tilde{L}_0 h \kappa_{\mathbf{z}}^n \|\mathbf{E}_0^{\mathbf{z}}\| + h(\tilde{M}_0 + \tilde{L}_0 \|\mathbf{E}_0^{\mathbf{y}}\|)\,, \tag{127a}$$

$$w_{n+1} \leq \tilde{L}_0 u_n + (\kappa_{\mathbf{z}} + \tilde{L}_0 h) w_n + \tilde{L}_0 h \kappa_{\mathbf{z}}^n \|\mathbf{E}_0^{\mathbf{z}}\| + \tilde{M}_0 + \tilde{L}_0 \|\mathbf{E}_0^{\mathbf{y}}\| \tag{127b}$$

with appropriate positive constants $\tilde{L}_0$ and $\tilde{M}_0$. Here, we have taken into account that $\kappa_{\mathbf{y}} = \|\mathbf{T}_{\mathbf{y}}\| = 1$ and $\kappa_{\mathbf{z}} = \|\mathbf{T}_{\mathbf{z}}\| < 1$ and restricted the analysis to $h \in (0, h_0]$ with a sufficiently small constant $h_0 > 0$.

(c) The recursive application of error bounds (127) shows that the coupled error propagation in differential and algebraic solution components may be studied analysing powers of the 2×2 error amplification matrix

$$\mathbf{W}(h) := \begin{pmatrix} 1 + \tilde{L}_0 h & \tilde{L}_0 h \\ \tilde{L}_0 & \kappa_{\mathbf{z}} + \tilde{L}_0 h \end{pmatrix},$$

see (Deuflhard et al. 1987, Lemma 2). The eigenvalue analysis for matrix $\mathbf{W}(h)$ yields an eigenvalue $\lambda(h) = \kappa_{\mathbf{z}} + \mathcal{O}(h)$. Because of $\kappa_{\mathbf{z}} < 1$, this eigenvalue satisfies $\lambda(h) < 1$ for all sufficiently small time step sizes $h > 0$. The corresponding eigenvector

$$\boldsymbol{\zeta}(h) := \begin{pmatrix} -L_v h \\ 1 \end{pmatrix} \text{ with } L_v := \frac{\tilde{L}_0}{1 + \tilde{L}_0 h - \lambda(h)} = \frac{\tilde{L}_0}{1 - \kappa_{\mathbf{z}}} + \mathcal{O}(h) \tag{128}$$

is used to transform $\mathbf{W}(h)$ to lower triangular form: We define the transformation matrix

$$\mathbf{V}(h) := [\,\mathbf{e}_1 \;\; \boldsymbol{\zeta}(h)\,] = \begin{pmatrix} 1 & -L_v h \\ 0 & 1 \end{pmatrix} \text{ with } \mathbf{V}^{-1}(h) = \begin{pmatrix} 1 & L_v h \\ 0 & 1 \end{pmatrix} \tag{129}$$

and observe that the second column vector of $\mathbf{W}(h)\mathbf{V}(h)$ is a multiple of the second column vector of $\mathbf{V}(h)$ since $\mathbf{W}(h)\boldsymbol{\zeta}(h) = \lambda(h)\boldsymbol{\zeta}(h)$. Therefore, the scalar product of the first row vector of $\mathbf{V}^{-1}(h)$ and the second column vector of $\mathbf{W}(h)\mathbf{V}(h)$, i.e., the upper right element of $\mathbf{V}^{-1}(h)\mathbf{W}(h)\mathbf{V}(h)$, vanishes. Straightforward computations yield

$$\mathbf{V}^{-1}(h)\mathbf{W}(h)\mathbf{V}(h) = \begin{pmatrix} 1+\tilde{L}_0(L_v+1)h & 0 \\ \tilde{L}_0 & \kappa_{\mathbf{z}} + \tilde{L}_0(1-L_v)h \end{pmatrix} \tag{130}$$

and

$$\begin{aligned} v_{n+1} &\le (1+\tilde{L}_0(L_v+1)h)v_n + && (131a)\\ &\quad + \tilde{L}_0(L_v h+1)h\kappa_{\mathbf{z}}^n \|\mathbf{E}_0^{\mathbf{z}}\| + (L_v+1)h(\tilde{M}_0 + \tilde{L}_0\|\mathbf{E}_0^{\mathbf{y}}\|)\,, \\ w_{n+1} &\le \tilde{L}_0 v_n + (\kappa_{\mathbf{z}} + \tilde{L}_0(1-L_v)h)w_n + && (131b)\\ &\quad + \tilde{L}_0 h\kappa_{\mathbf{z}}^n\|\mathbf{E}_0^{\mathbf{z}}\| + \tilde{M}_0 + \tilde{L}_0\|\mathbf{E}_0^{\mathbf{y}}\| \end{aligned}$$

with a sequence $(v_n)_{n\ge 0}$ of non-negative numbers v_n that are defined by

$$\begin{pmatrix} v_n \\ w_n \end{pmatrix} = \mathbf{V}^{-1}(h)\begin{pmatrix} u_n \\ w_n \end{pmatrix},$$

see (127), (130) and (131). Note, that all matrix elements of $\mathbf{V}^{-1}(h)$ are non-negative which is an essential assumption for the transformation from (127) to (131).

(d) The right-hand side of (131a) depends nonlinearly on h because $L_v = L_v(h)$. If we substitute L_v for sufficiently small time step sizes $h > 0$ by the upper bound $\tilde{L}_v := 2\tilde{L}_0/(1-\kappa_{\mathbf{z}})$, see (128), then Lemma 4.12 may be applied with constants $L := \tilde{L}_0(\tilde{L}_v \max\{1,h_0\}+1)$, $\kappa := \kappa_{\mathbf{z}} < 1$, $e_0 := \|\mathbf{E}_0^{\mathbf{z}}\|$ and $M := (\tilde{L}_v+1)\tilde{M}_0 + L\|\mathbf{E}_0^{\mathbf{y}}\|$. Inequality (116a) yields the error bound

$$v_n \le \mathrm{err}_n - \|\mathbf{E}_0^{\mathbf{y}}\| \tag{132}$$

with

$$\mathrm{err}_n := \mathrm{e}^{L(t_n-t_0)}(\|\mathbf{E}_0^{\mathbf{y}}\| + \frac{hL}{1-\kappa}\|\mathbf{E}_0^{\mathbf{z}}\|) + \frac{\mathrm{e}^{L(t_n-t_0)}-1}{L}(\tilde{L}_v+1)\tilde{M}_0 \tag{133}$$

because $v_0 = u_0 + L_v h w_0 = 0$, see (126). Inequality (132) proves the global error bound (125a) since $u_n = v_n - L_v h w_n \le v_n$ and

$$\|\mathbf{E}_n^{\mathbf{y}}\| \le \|\mathbf{T}_{\mathbf{y}}^n\mathbf{E}_0^{\mathbf{y}}\| + \|\mathbf{E}_n^{\mathbf{y}} - \mathbf{T}_{\mathbf{y}}^n\mathbf{E}_0^{\mathbf{y}}\| \le \|\mathbf{T}_{\mathbf{y}}\|^n\|\mathbf{E}_0^{\mathbf{y}}\| + u_n \le \|\mathbf{E}_0^{\mathbf{y}}\| + v_n \le \mathrm{err}_n\,.$$

For the proof of error bound (125b), we substitute in (131b) the variable v_n by its upper bound (132) and get

$$w_{n+1} \le (\kappa + Lh)w_n + Lh\kappa^n e_0 + \tilde{M}_0 + \tilde{L}_0\,\mathrm{err}_n$$

since $\tilde{L}_0(1-L_v) \le \tilde{L}_0 \le L$. For all $r \le n$, the term $\tilde{M}_0 + \tilde{L}_0 \,\mathrm{err}_r$ is bounded by $\tilde{M}_0 + \tilde{L}_0 \,\mathrm{err}_n$ because $(\mathrm{err}_n)_{n\ge 0}$ is monotonically increasing. Therefore, Lemma 4.12 with

$$M := \tilde{M}_0 + \tilde{L}_0 \,\mathrm{err}_n \le \tilde{L}_0 \mathrm{e}^{L(t_n-t_0)}\big(\|\mathbf{E}_0^{\mathbf{y}}\| + \frac{hL}{1-\kappa}\|\mathbf{E}_0^{\mathbf{z}}\|\big) + \mathrm{e}^{L(t_n-t_0)}\tilde{M}_0\,,$$

see (133), yields

$$w_n \le C_0^{\mathbf{z}}\big(h\|\mathbf{E}_0^{\mathbf{z}}\| + \mathrm{e}^{L(t_n-t_0)}(\|\mathbf{E}_0^{\mathbf{y}}\| + \frac{hL}{1-\kappa}\|\mathbf{E}_0^{\mathbf{z}}\| + \tilde{M}_0)\big)$$

with an appropriate constant $C_0^{\mathbf{z}} > 0$, see (116b). Error bound (125b) follows straightforwardly from $\|\mathbf{E}_n^{\mathbf{z}} - \mathbf{T}_{\mathbf{z}}^n \mathbf{E}_0^{\mathbf{z}}\| = w_n$, see (126). □

4.3 Convergence of Lie Group Time Integration Methods

For the application of Theorem 4.16 to the one-step error recursion (93) we have to verify the assumptions on error propagation matrices $\mathbf{T}_{\mathbf{y}}$ and $\mathbf{T}_{\mathbf{z}}$. Because of $\mathbf{T}_{\mathbf{y}} = \mathbf{I}_{2k}$, we get $\|\mathbf{T}_{\mathbf{y}}\|_2 = 1$. For proving $\|\mathbf{T}_{\mathbf{z}}\|_{\mathbf{z},\varrho} < 1$ in a suitable norm $\|.\|_{\mathbf{z},\varrho}$, we analyse the spectral radius $\rho(\mathbf{T}_{\mathbf{z}})$:

Lemma 4.17 *(a) For algorithmic parameters α_m, α_f, β, γ that satisfy the order condition (41) and the stability conditions*

$$\alpha_m < \alpha_f < 0.5\,, \qquad \gamma < 2\beta\,, \tag{134}$$

the spectral radii of matrices $\mathbf{T}_{\mathbf{z}}$ in (95) and (100) are bounded by $\rho(\mathbf{T}_{\mathbf{z}}) < 1$.

(b) For the "optimal" parameters of Chung and Hulbert (1993), see (42), the stability conditions (134) are satisfied for any $\rho_\infty \in [0,1)$.

Proof (a) The block-diagonal structure of matrix $\mathbf{T}_{\mathbf{z}} \in \mathbb{R}^{m+3k}$ in (95) implies that its characteristic polynomial is given by

$$\det(\zeta\mathbf{I} - \mathbf{T}_{\mathbf{z}}) = \Big(\zeta + \frac{\alpha_m}{1-\alpha_m}\Big)^k \Big(\det \mathbf{T}_+^{-1} \det(\zeta\mathbf{T}_+ - \mathbf{T}_0)\Big)^m.$$

Straightforward computations show that matrix $\mathbf{T}_{\mathbf{z}}$ has an eigenvalue $\zeta_m := -\alpha_m/(1-\alpha_m)$ of multiplicity k, an eigenvalue $\zeta_f := -\alpha_f/(1-\alpha_f)$ of multiplicity m and eigenvalues $\zeta_{1,2}$ that are given by the roots of the quadratic polynomial $\sigma(\zeta) := a\zeta^2 + b\zeta + c$ with

$$a := \beta\,, \quad b := 0.5 + \gamma - 2\beta\,, \quad c := 1 - a - b\,, \tag{135}$$

see also (Arnold and Brüls 2007, Lemma 1). The stability conditions (134) imply $|\zeta_m| < 1$, $|\zeta_f| < 1$ and $\gamma = 0.5 + \alpha_f - \alpha_m > 0.5$.

Therefore, the coefficients a, b, c in (135) satisfy $a = \beta > 0$, $b > 1 - 2\beta = 1 - 2a$ and $c = 1 - a - b < a$. Since $c/a < 1$ and $\zeta_1\zeta_2 = c/a$ (Vieta's theorem), we get $|\zeta_1|^2 = |\zeta_2|^2 = \zeta_1\zeta_2 = c/a < 1$ whenever $\sigma(\zeta) = 0$ has a pair of conjugate complex roots ζ_1, ζ_2.

If both roots of σ are real then the discriminant

$$b^2 - 4ac = b^2 - 4a(1 - a - b) = (2a + b)^2 - 4a$$

has to be non-negative. Hence,

$$\sqrt{b^2 - 4ac} < \sqrt{(2a + b)^2} = 2a + b \tag{136a}$$

since $a > 0$ and $2a + b = 0.5 + \gamma > 1 \geq 0$, see (135). On the other hand, stability condition $\gamma < 2\beta$ results in $b < 0.5$ and

$$(2a + b)^2 - 4a = (2a - b)^2 + 8a(b - 0.5) < (2a - b)^2 ,$$

i.e.,

$$\sqrt{b^2 - 4ac} = \sqrt{(2a + b)^2 - 4a} < \sqrt{(2a - b)^2} = 2a - b \tag{136b}$$

since $2a - b = 2(2\beta - \gamma) + (\gamma - 0.5) > 0$. Estimates (136) show that the roots $\zeta_{1,2} = (-b \pm \sqrt{b^2 - 4ac})/2a$ of σ satisfy $-1 < \zeta_i < 1$, $(\, i = 1, 2\,)$. This completes the proof of $\rho(\mathbf{T_z}) < 1$ for matrix $\mathbf{T_z}$ being defined in (95).

Substituting the quadratic polynomial $\sigma(\zeta)$ by $\sigma(\zeta) := \zeta + (1 - \gamma)/\gamma$, we may extend this analysis straightforwardly to the matrix $\mathbf{T_z}$ in (100).

(b) With $\rho_\infty \in [0, 1)$, the algorithmic parameters α_m, α_f in (42) satisfy $\alpha_m < \alpha_f < 0.5$ and $\gamma = 0.5 + \alpha_f - \alpha_m > 0.5$. For the second stability condition in (134), we observe that (42) implies $2\beta - \gamma = (\gamma - 0.5)^2/2 > 0$. □

Theorem 4.18 *Let the order condition (41) and the stability conditions (134) be fulfilled and suppose that the starting values* q_0, $\mathbf{v}_0$, $\dot{\mathbf{v}}_0$, $\mathbf{a}_0$ *and* $\boldsymbol{\lambda}_0$ *satisfy*

$$\|\mathbf{e}_0^q\| + \|\mathbf{e}_0^\mathbf{v}\| + h\|\mathbf{e}_0^\mathbf{Pa}\| = \mathcal{O}(h^2)\,, \quad \|\mathbf{e}_0^{\dot{\mathbf{v}}}\| + \|\mathbf{e}_0^\mathbf{Ba}\| = \mathcal{O}(h^{1+\delta})\,, \tag{137a}$$

$$\|\mathbf{M}(q_0)\dot{\mathbf{v}}_0 + \mathbf{g}(q_0, \mathbf{v}_0, t_0) + \mathbf{B}^\top(q_0)\boldsymbol{\lambda}_0\| = \mathcal{O}(h^{1+\delta}) \tag{137b}$$

with a non-negative constant $\delta \in [0, 1]$. *Then, there are positive constants* C_0, $\tilde{L}$, h_0 *being independent of* n *and* h *such that we have for all* $h \in (0, h_0]$ *and all* $n \geq 0$ *with* $t_0 + nh \leq t_{\text{end}} - h$*:*

(a) a global error bound

$$\|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| \le C_0 \mathrm{e}^{\tilde{L}(t_n-t_0)} h^2 , \tag{138a}$$

$$\|\mathbf{e}_n^{\boldsymbol{\lambda}}\| \le C_0 (\|(\mathbf{T}_+^{-1}\mathbf{T}_0)^n\| h^{1+\delta} + \mathrm{e}^{\tilde{L}(t_n-t_0)} h^2) \tag{138b}$$

for the index-3 integrator (37) provided that the starting values q_0, $\mathbf{v}_0$ satisfy the additional assumption

$$\|\mathbf{e}_0^q\| + \|\mathbf{e}_0^{\mathbf{Bv}} + \frac{1}{h}\mathbf{B}(q(t_0))\mathbf{l}_0^q\| = \mathcal{O}(h^{2+\delta}) \tag{139}$$

and

(b) a global error bound

$$\|\mathbf{e}_n^q\| + \|\mathbf{e}_n^{\mathbf{v}}\| + \|\boldsymbol{\eta}_n\| \le C_0 \mathrm{e}^{\tilde{L}(t_n-t_0)} h^2 , \tag{140a}$$

$$\|\mathbf{e}_n^{\boldsymbol{\lambda}}\| \le C_0 (\|(\bar{\mathbf{T}}_+^{-1}\bar{\mathbf{T}}_0)^n\| h^{1+\delta} + \mathrm{e}^{\tilde{L}(t_n-t_0)} h^2) \tag{140b}$$

for the stabilized index-2 integrator (56).

Proof These error estimates are a straightforward consequence of Theorem 4.16 and Lemma 4.17 since error recursion (93) with matrices $\mathbf{T}_{\mathbf{y}}$ and $\mathbf{T}_{\mathbf{z}}$ being defined in (95), (100) and $\varepsilon_n = \mathcal{O}(1)(\|\mathbf{E}_n^{\mathbf{y}}\| + h\|\mathbf{E}_n^{\mathbf{z}}\|)$ imply (124). Furthermore, assumptions (137) and (139) result in $\|\mathbf{E}_0^{\mathbf{y}}\| = \mathcal{O}(h^2)$, $\|\mathbf{E}_0^{\mathbf{z}}\| = \mathcal{O}(h)$ and $\|\mathbf{E}_0^{\mathbf{r}}\| = \mathcal{O}(h^{1+\delta})$. Finally, the upper bound for $\|\boldsymbol{\eta}_n\|$ in (140a) is obtained from (98). □

Lemma 4.17 and Theorem 4.18 show that transient errors of size $\mathcal{O}(h^{1+\delta})$ are damped out by numerical dissipation if the generalized-α methods (37) and (56) have algorithmic parameters according to (42) with $\rho_\infty < 1$. For starting values $q_0 = q(t_0), \mathbf{v}_0 = \mathbf{v}(t_0), \dot{\mathbf{v}}_0 = \dot{\mathbf{v}}(t_0)$ and $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}(t_0)$ being defined by consistent initial values $q(t_0), \mathbf{v}(t_0), \dot{\mathbf{v}}(t_0), \boldsymbol{\lambda}(t_0)$, assumptions (137) and (139) are satisfied with $\delta \ge 0$ if $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0) + \mathcal{O}(h)$. Beyond the transient phase, we observe second-order convergence in all solution components, see Fig. 8.

For the heavy top benchmark problem in configuration space $G = \mathrm{SE}(3)$, we may even prove that there is no order reduction at all in generalized-α Lie group time integration:

Example 4.19 (a) For consistent initial values $q(t_0)$, $\mathbf{v}(t_0)$, $\dot{\mathbf{v}}(t_0)$ and $\boldsymbol{\lambda}(t_0)$, the starting values $q_0 = q(t_0), \mathbf{v}_0 = \mathbf{v}(t_0), \dot{\mathbf{v}}_0 = \dot{\mathbf{v}}(t_0), \mathbf{a}_0 = \dot{\mathbf{v}}(t_0), \boldsymbol{\lambda}_0 = \boldsymbol{\lambda}(t_0)$ satisfy assumption (137) with $\delta = 1$ if $\mathbf{B}\big(q(t_0)\big)\ddot{\mathbf{v}}(t_0) = \mathbf{0}$ since Taylor expansion of $\dot{\mathbf{v}}(t_0 + \Delta_\alpha h)$ at $h = 0$ shows in that case that $\|\mathbf{e}_0^{\mathbf{Ba}}\| = \|\mathbf{B}\big(q(t_0)\big)\big(\dot{\mathbf{v}}(t_0 + \Delta_\alpha h) - \mathbf{a}_0\big)\| = \mathcal{O}(h^2)$.

(b) Condition $\mathbf{B}\big(q(t_0)\big)\ddot{\mathbf{v}}(t_0) = \mathbf{0}$ in part (a) of this example is satisfied for the equations of motion (22) of the heavy top benchmark in configuration space $G = \mathrm{SE}(3)$ since $\mathbf{B}\big(q(t)\big) \equiv \mathbf{B}^{\mathbf{X}} := \big(-\widetilde{\mathbf{X}} \quad -\mathbf{I}_3\big)$ along any solution curve $q(t)$ in the constraint manifold $\mathfrak{M} := \{\, q \,:\, \boldsymbol{\Phi}(q) = \mathbf{0} \,\}$, see Lemma 3.5, and the hidden constraints (16),

(18) are given by $\mathbf{0} = \mathbf{B}^{\mathbf{X}}\mathbf{v}(t) = \mathbf{B}^{\mathbf{X}}\dot{\mathbf{v}}(t)$ implying $\mathbf{B}\big(q(t)\big)\ddot{\mathbf{v}}(t) = \mathbf{0}$. Therefore, Theorem 4.18(b) proves second-order convergence of the stabilized index-2 integrator (56) for this benchmark problem. These theoretical investigations are illustrated by the numerical test results in the right plot of Fig. 11.

(c) The equations of motion (22) of the heavy top benchmark in configuration space $G = \mathrm{SE}(3)$ fulfill the assumptions of Lemma 3.5. Therefore, the generalized-α integrator (37) defines a numerical solution that satisfies the hidden constraints (16) at the level of velocity coordinates. I.e., integrators (37) and (56) define identical numerical solutions for this benchmark problem and we get $\boldsymbol{\eta}_n = \mathbf{0}$. The numerical test results in the right plots of Figs. 6 and 11 illustrate this coincidence.

For a more direct proof of the corresponding second-order convergence result for integrator (37), we may verify that for this benchmark problem assumption (139) in Theorem 4.18(a) is satisfied with $\delta = 1$: Taking into account $\mathbf{B}\big(q(t_n)\big)\ddot{\mathbf{v}}(t_n) = \mathbf{0}$ and the structure of the leading error term in $\mathbf{l}_n^q$, we get $\mathbf{B}\big(q(t_n)\big)\mathbf{l}_n^q = \mathcal{O}(h^4)$ if $\mathbf{B}\big(q(t)\big)\widehat{\mathbf{v}}(t)\dot{\mathbf{v}}(t) \equiv \mathbf{0}$, see Lemma 4.2. Here, we have substituted the term $[\widetilde{\mathbf{v}}, \widetilde{\dot{\mathbf{v}}}] \in \mathfrak{se}(3)$ in (69) by its equivalent $\widehat{\mathbf{v}}\dot{\mathbf{v}} \in \mathbb{R}^6$ with $\widehat{\mathbf{v}} \in \mathbb{R}^{6\times 6}$ being defined in (34), see also (29). For consistent velocity vectors $\mathbf{v}$, the skew symmetric matrix $\widetilde{\mathbf{U}}$ in (34) may be expressed in terms of $\widetilde{\mathbf{X}}$ and $\widetilde{\boldsymbol{\Omega}}$ since $\mathbf{B}^{\mathbf{X}}\mathbf{v} = \mathbf{0}$ implies $\mathbf{U} = -\widetilde{\mathbf{X}}\boldsymbol{\Omega} = \widetilde{\boldsymbol{\Omega}}\mathbf{X} = \widehat{\boldsymbol{\Omega}}\mathbf{X}$, i.e., $\widetilde{\mathbf{U}} = [\widetilde{\boldsymbol{\Omega}}, \widetilde{\mathbf{X}}] = \widetilde{\boldsymbol{\Omega}}\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\Omega}}$, see (29). The identity $\widetilde{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}$ is valid for any $\boldsymbol{\Omega} \in \mathbb{R}^3$, see Remark 2.8(b). We get

$$\mathbf{B}\big(q(t)\big)\widehat{\mathbf{v}}(t) = \mathbf{B}^{\mathbf{X}}\begin{pmatrix} \widetilde{\boldsymbol{\Omega}} & \mathbf{0} \\ \widetilde{\mathbf{U}} & \widetilde{\boldsymbol{\Omega}} \end{pmatrix} = \begin{pmatrix} -\widetilde{\mathbf{X}} & -\mathbf{I}_3 \end{pmatrix}\begin{pmatrix} \widetilde{\boldsymbol{\Omega}} & \mathbf{0} \\ \widetilde{\boldsymbol{\Omega}}\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\Omega}} & \widetilde{\boldsymbol{\Omega}} \end{pmatrix} = \widetilde{\boldsymbol{\Omega}}\mathbf{B}^{\mathbf{X}}$$

and therefore also $\mathbf{B}\big(q(t)\big)\widehat{\mathbf{v}}(t)\dot{\mathbf{v}}(t) \equiv \mathbf{0}$ since $\mathbf{B}^{\mathbf{X}}\dot{\mathbf{v}}(t) \equiv \mathbf{0}$, see (18). Hence, $\mathbf{B}\big(q(t_n)\big)$ $\mathbf{l}_n^q = \mathcal{O}(h^4)$ and assumptions (139) are satisfied for this benchmark problem with $\delta = 1$ if the starting values in the index-3 integrator (37) are set to $q_0 = q(t_0)$, $\mathbf{v}_0 = \mathbf{v}(t_0)$.

Example 4.19 illustrates that the trivial initialization $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0)$ results for certain problem classes in transient error terms of size $\mathcal{O}(h^{1+\delta})$ with $\delta = 1$ such that second-order convergence is already observed in the transient phase. In general, however, this trivial initialization yields transient errors of size $\mathcal{O}(h)$ since $\|\mathbf{e}_0^{\mathbf{Ba}}\| = \mathcal{O}(h)$ if $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0)$ and $\mathbf{B}\big(q(t_0)\big)\ddot{\mathbf{v}}(t_0) \neq \mathbf{0}$. These first order error terms have been observed numerically for the heavy top benchmark problem in configuration space $G = \mathrm{SO}(3) \times \mathbb{R}^3$ in Figs. 6, 7 and 13.

More sophisticated initializations of sequence $(\mathbf{a}_n)_{n\geq 0}$ in HHT-α and generalized-α time integration have been discussed, e.g., in (Jay and Negrut 2007) and (Arnold et al. 2015). We follow the latter approach and set

$$\mathbf{a}_0 := \dot{\mathbf{v}}(t_0) + \boldsymbol{\Delta}_0^{\mathbf{a}} \quad \text{with} \quad \boldsymbol{\Delta}_0^{\mathbf{a}} := \Delta_\alpha h\,\frac{\dot{\mathbf{v}}_{sh} - \dot{\mathbf{v}}_{-sh}}{2sh}\,, \tag{141}$$

vectors $\dot{\mathbf{v}}_{\pm sh} = \dot{\mathbf{v}}(t_0 \pm sh) + \mathcal{O}(h^2)$ and a (small) parameter $s \in (0, 1]$ that may be set, e.g., to $s := 1/10$. For the computation of $\boldsymbol{\Delta}_0^{\mathbf{a}}$, we have to evaluate the equations

Table 1 Initialization of the stabilized index-2 integrator (56)

Data	Consistent initial values $q(t_0)$, $\mathbf{v}(t_0)$; parameter $s \in (0, 1]$
Result	Modified starting values q_0, $\mathbf{v}_0$, $\dot{\mathbf{v}}_0$, $\mathbf{a}_0$, $\boldsymbol{\lambda}_0$ of integrator (56)
Step 1	Set starting values q_0, $\mathbf{v}_0$ to the consistent initial values: $q_0 := q(t_0)$, $\mathbf{v}_0 := \mathbf{v}(t_0)$
Step 2	Solve system (19) with $t = t_0$, $q = q_0$, $\mathbf{v} = \mathbf{v}_0$ to get consistent starting values $\dot{\mathbf{v}}_0$ and $\boldsymbol{\lambda}_0$
Step 3	Get $\dot{\mathbf{v}}_{sh}$ from system (19) with $t = t_0 + sh$ and $q = q_0 \circ \exp\bigl(sh\mathbf{v}_0 + s^2h^2\dot{\mathbf{v}}_0/2\bigr)$, $\mathbf{v} = \mathbf{v}_0 + sh\dot{\mathbf{v}}_0$
Step 4	Get $\dot{\mathbf{v}}_{-sh}$ from system (19) with $t = t_0 - sh$ and $q = q_0 \circ \exp\bigl(-sh\mathbf{v}_0 + s^2h^2\dot{\mathbf{v}}_0/2\bigr)$, $\mathbf{v} = \mathbf{v}_0 - sh\dot{\mathbf{v}}_0$
Step 5	Compute starting value $\mathbf{a}_0 := \dot{\mathbf{v}}_0 + \Delta_\alpha h\,(\dot{\mathbf{v}}_{sh} - \dot{\mathbf{v}}_{-sh})/(2sh)$

of motion at $t_0 + sh$ and at $t_0 - sh$. Then, vectors $\dot{\mathbf{v}}_{sh}$ and $\dot{\mathbf{v}}_{-sh}$ may be obtained from block-structured systems of linear equations (19), see the numerical algorithm in Table 1 for a more detailed discussion of this initialization phase.

Starting values $\mathbf{a}_0$ according to (141) satisfy assumption (137) with $\delta = 1$ since $\dot{\mathbf{v}}(t_0) + \Delta_\alpha h(\dot{\mathbf{v}}_{sh} - \dot{\mathbf{v}}_{-sh})/2sh = \dot{\mathbf{v}}(t_0 + \Delta_\alpha h) + \mathcal{O}(h^2)$. Hence, Theorem 4.18(b) proves second-order convergence of the stabilized index-2 integrator (56) for all solution components. This convergence result may be verified by a numerical test for the heavy top benchmark problem in configuration space $G = \mathrm{SO}(3) \times \mathbb{R}^3$: Fig. 16 shows for time step size $h = 1.0 \times 10^{-3}$ the global error $e_n^{\lambda_1}/\|\boldsymbol{\lambda}_n\|$ of the stabilized index-2 integrator (56) in time interval $[0, 0.1]$. The test results in the left plot are already known from the left plot of Fig. 13. They show the transient oscillating first-order error term being characteristic of the trivial initialization $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0)$. The test results in the right plot illustrate that this first-order error term disappears if we use the modified starting value $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0) + \boldsymbol{\Delta}_0^{\mathbf{a}} \approx \dot{\mathbf{v}}(t_0 + \Delta_\alpha h)$.

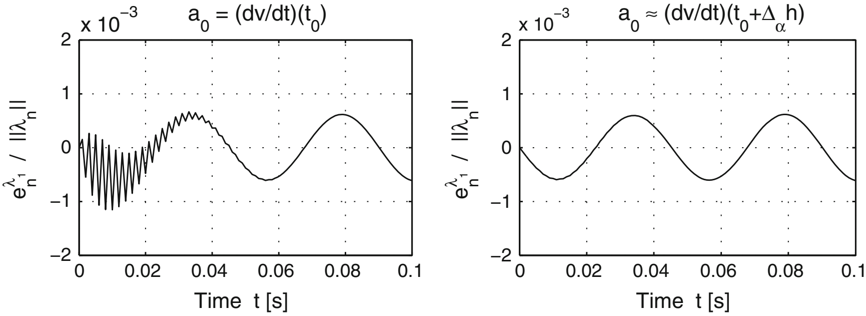


Fig. 16 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, starting values $q_0 = q(t_0)$, $\mathbf{v}_0 = \mathbf{v}(t_0)$, stabilized index-2 formulation, $G = \mathrm{SO}(3) \times \mathbb{R}^3$): Global error $e_n^{\lambda_1}/\|\boldsymbol{\lambda}_n\|$. *Left plot* $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0)$, *right plot* $\mathbf{a}_0 \approx \dot{\mathbf{v}}(t_0 + \Delta_\alpha h)$

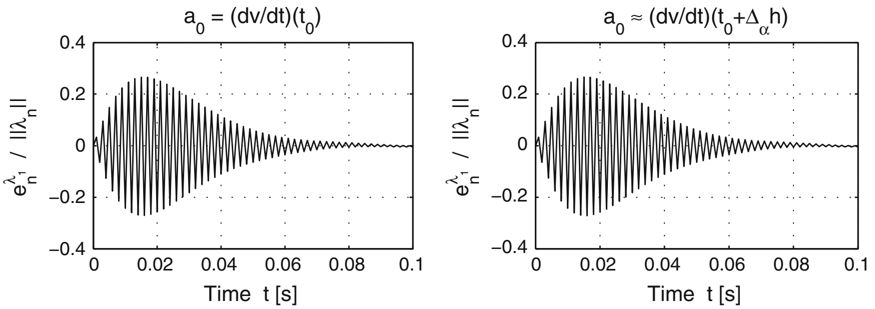


Fig. 17 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, starting values $q_0 = q(t_0)$, $\mathbf{v}_0 = \mathbf{v}(t_0)$, index-3 formulation, $G = \mathrm{SO}(3) \times \mathbb{R}^3$): Global error $e_n^{\lambda_1}/\|\boldsymbol{\lambda}_n\|$. *Left plot* $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0)$, *right plot* $\mathbf{a}_0 \approx \dot{\mathbf{v}}(t_0 + \Delta_\alpha h)$

Note, that this modification of starting value $\mathbf{a}_0$ does not contribute significantly to the result accuracy of the index-3 integrator (37) since the additional assumption (139) in part (a) of Theorem 4.18 is (as before) only satisfied with $\delta = 0$. The resulting large first-order error term in $\lambda_{n,1}$ is (up to plot accuracy) not affected by modified starting values $\mathbf{a}_0$, see Fig. 17.

This first-order error term is well known from the convergence analysis for the linear test problem in Sect. 3.2. In Theorem 3.1(b), we proposed a systematic perturbation of starting values v_0 to get second-order convergence, see (52). In the Lie group setting, these modified starting values are given by

$$\mathbf{v}_0 = \mathbf{v}(t_0) + [\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}]\big(q(t_0)\big)\,\mathbf{l}_0^q/h + \mathcal{O}(h^3)\,.$$

In a practical implementation, we restrict ourselves to the leading error term in $\mathbf{l}_0^q$, see (69), and use again a difference approximation of $\ddot{\mathbf{v}}(t_0)$, see (141). The modified starting values are given by $\mathbf{v}_0 = \mathbf{v}(t_0) + \boldsymbol{\Delta}_0^\mathbf{v}$ with

$$\boldsymbol{\Delta}_0^\mathbf{v} := h^2\,[\mathbf{M}^{-1}\mathbf{B}^\top(\mathbf{B}\mathbf{M}^{-1}\mathbf{B}^\top)^{-1}\mathbf{B}]\big(q(t_0)\big)\cdot \\ \cdot\Big(C_q\,\frac{\dot{\mathbf{v}}_{sh} - \dot{\mathbf{v}}_{-sh}}{2sh} + \frac{1}{12}\widehat{\mathbf{v}}(t_0)\dot{\mathbf{v}}(t_0)\Big)\,. \tag{142}$$

They may be computed efficiently by the numerical algorithm in Table 2. The numerical test results for two different time step sizes in Fig. 18 illustrate that the modified starting values eliminate the first-order error term. The maximum amplitude of $e_n^{\lambda_1}/\|\boldsymbol{\lambda}_n\|$ is reduced by a factor of 4 if the time step size is reduced from $h = 1.0 \times 10^{-3}$ to $h = 5.0 \times 10^{-4}$.

The perturbation of size $\mathcal{O}(h^2)$ in (142) results in starting values $q_0 = q(t_0)$ and $\mathbf{v}_0 = \mathbf{v}(t_0) + \boldsymbol{\Delta}_0^\mathbf{v}$ that satisfy assumption (139) in Theorem 4.18(a) with $\delta = 1$. In general, these starting values are *not* consistent with the hidden constraints (16) at the level of velocity coordinates but introduce systematically a residual of size $\mathbf{B}(q_0)\mathbf{v}_0 = \mathcal{O}(h^2)$ at $t = t_0$. The numerical test results in Figs. 19 and 20 show that this

Table 2 Initialization of the index-3 integrator (37)

Data	Consistent initial values $q(t_0)$, $\mathbf{v}(t_0)$; parameter $s \in (0, 1]$
Result	Modified starting values q_0, $\mathbf{v}_0$, $\dot{\mathbf{v}}_0$, $\mathbf{a}_0$, $\boldsymbol{\lambda}_0$ of integrator (37)
Step 1	Set starting value q_0 to the consistent initial value: $q_0 := q(t_0)$
Step 2	Solve system (19) with $t = t_0$, $q = q(t_0)$, $\mathbf{v} = \mathbf{v}(t_0)$ to get consistent starting values $\dot{\mathbf{v}}_0$ and $\boldsymbol{\lambda}_0$
Step 3	Get $\dot{\mathbf{v}}_{sh}$ from system (19) with $t = t_0 + sh$ and $q = q(t_0) \circ \exp\big(sh\mathbf{v}(t_0) + s^2h^2\dot{\mathbf{v}}_0/2\big)$, $\mathbf{v} = \mathbf{v}(t_0) + sh\dot{\mathbf{v}}_0$
Step 4	Get $\dot{\mathbf{v}}_{-sh}$ from system (19) with $t = t_0 - sh$ and $q = q(t_0) \circ \exp\big(-sh\mathbf{v}(t_0) + s^2h^2\dot{\mathbf{v}}_0/2\big)$, $\mathbf{v} = \mathbf{v}(t_0) - sh\dot{\mathbf{v}}_0$
Step 5	Compute starting value $\mathbf{a}_0 := \dot{\mathbf{v}}_0 + \Delta_\alpha h\,(\dot{\mathbf{v}}_{sh} - \dot{\mathbf{v}}_{-sh})/(2sh)$
Step 6	Get $\boldsymbol{\Delta}_0^{\mathbf{v}} := \mathbf{x}_{\dot{\mathbf{v}}}$ from the system of linear equations (20) with $\mathbf{r}_{\dot{\mathbf{v}}} = \mathbf{0}_k\,,\quad \mathbf{r}_{\boldsymbol{\lambda}} = h^2\mathbf{B}(q_0)\big(C_q\dfrac{\dot{\mathbf{v}}_{sh} - \dot{\mathbf{v}}_{-sh}}{2sh} + \dfrac{1}{12}\widehat{\mathbf{v}}(t_0)\dot{\mathbf{v}}(t_0)\big)$ and matrices $\mathbf{M} = \mathbf{M}(q_0)$, $\mathbf{B} = \mathbf{B}(q_0)$
Step 7	Set starting value $\mathbf{v}_0$ to $\mathbf{v}_0 := \mathbf{v}(t_0) + \boldsymbol{\Delta}_0^{\mathbf{v}}$, see (142)

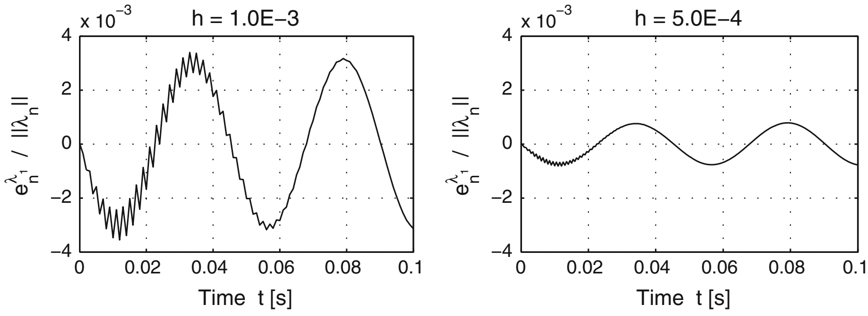


Fig. 18 Heavy top benchmark (index-3 formulation, starting values $q_0 = q(t_0)$, $\mathbf{v}_0 = \mathbf{v}(t_0) + \boldsymbol{\Delta}_0^{\mathbf{v}}$, $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0) + \boldsymbol{\Delta}_0^{\mathbf{a}}$, $G = \mathrm{SO}(3) \times \mathbb{R}^3$): Global error $e_n^{\lambda_1}/\|\boldsymbol{\lambda}_n\|$. *Left plot* $h = 1.0 \times 10^{-3}$, *right plot* $h = 5.0 \times 10^{-4}$

non-vanishing initial constraint residual helps to avoid the oscillating second-order term in the constraint residuals $\mathbf{B}(q_n)\mathbf{v}_n$ as well as the corresponding oscillating first-order error term in the Lagrange multipliers $\boldsymbol{\lambda}_n$: In the left plots of Figs. 19 and 20, we see the simulation data for (classical) starting values $\mathbf{v}_0 = \mathbf{v}(t_0)$, $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0)$ that are already known from the numerical tests in Sect. 3.3 (left plots of Figs. 10 and 7). The test results in the right plots of Figs. 19 and 20 show that the transient oscillating terms disappear up to plot accuracy for the modified starting values $\mathbf{v}_0 = \mathbf{v}(t_0) + \boldsymbol{\Delta}_0^{\mathbf{v}} = \mathbf{v}(t_0) + \mathcal{O}(h^2)$ and $\mathbf{a}_0 = \dot{\mathbf{v}}(t_0) + \boldsymbol{\Delta}_0^{\mathbf{a}} = \dot{\mathbf{v}}(t_0 + \Delta_\alpha h) + \mathcal{O}(h^2)$.

The algorithm in Table 2 spends moderate numerical effort to get (modified) starting values q_0, $\mathbf{v}_0$, $\dot{\mathbf{v}}_0$, $\mathbf{a}_0$ and $\boldsymbol{\lambda}_0$ for the generalized-α Lie group integrator (37) that satisfy assumptions (137) and (139) in the convergence theorem with $\delta = 1$. The error bounds (138) in Theorem 4.18(a) prove second-order convergence in all solu-

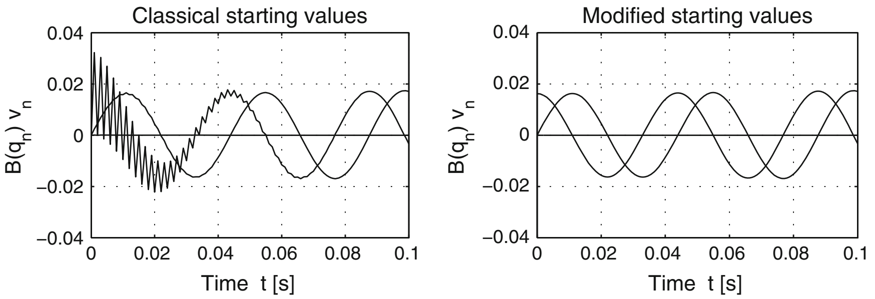


Fig. 19 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, index-3 formulation, $G = \mathrm{SO}(3) \times \mathbb{R}^3$): Residuals in hidden constraints (16). *Left plot* classical starting values $\mathbf{v}_0$, $\mathbf{a}_0$, *right plot* modified starting values $\mathbf{v}_0$, $\mathbf{a}_0$

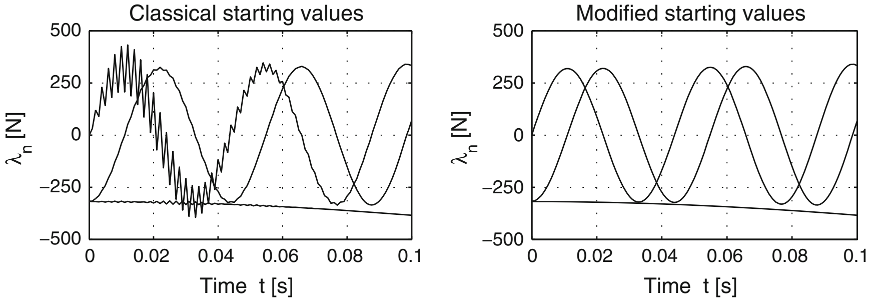


Fig. 20 Heavy top benchmark ($h = 1.0 \times 10^{-3}$, index-3 formulation, $G = \mathrm{SO}(3) \times \mathbb{R}^3$): Numerical solution $\boldsymbol{\lambda}_n$. *Left plot* classical starting values $\mathbf{v}_0$, $\mathbf{a}_0$, *right plot* modified starting values $\mathbf{v}_0$, $\mathbf{a}_0$

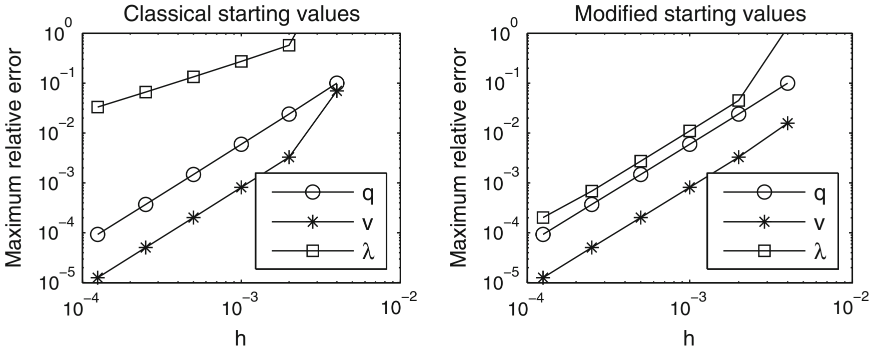


Fig. 21 Heavy top benchmark (index-3 formulation, $G = \mathrm{SO}(3) \times \mathbb{R}^3$): Global error of integrator (37) versus h for $t \in [0, 1]$. *Left plot* classical starting values $\mathbf{v}_0$, $\mathbf{a}_0$, *right plot* modified starting values $\mathbf{v}_0$, $\mathbf{a}_0$

tion components. The right plot of Fig. 21 shows numerical test results for the heavy top benchmark problem that are in perfect agreement with this asymptotic error analysis for small time step sizes h.

5 Summary

The generalized-α method is a Newmark-type method and one of the standard time integration methods in structural dynamics. The method is second-order accurate for unconstrained systems in linear spaces and has a free algorithmic parameter that allows to control the amount of numerical dissipation for high-frequency solution components. Following a Lie algebra approach, the method may be applied as well to mechanical systems that have a nonlinear configuration space with Lie group structure. In each time step, the increment of the configuration variables is parametrized by an element of the corresponding Lie algebra that may be obtained numerically by a classical Newton–Raphson iteration in linear spaces.

The Lie algebra approach is used as well in the asymptotic error analysis for the application to constrained systems that are typical of multibody dynamics. Newmark-type time integration methods of second-order accuracy are known to suffer from "overshooting", i.e., from an oscillating transient error term in the application to a scalar linear test equation with high-frequency solutions. For constrained systems, these large transient errors may result in order reduction unless the starting values of the generalized-α method are perturbed by an appropriate second-order correction term. Second-order convergence of the algorithm with perturbed starting values is proved analytically studying a coupled error propagation in differential and algebraic solution components that takes into account a quadratic approximation of hidden constraints at the level of acceleration coordinates.

The order reduction phenomenon may be avoided by an analytical index reduction before time discretization. The Lie algebra approach allows to modify the increment of configuration variables such that the numerical solution satisfies in each time step the original holonomic constraints at the level of position coordinates as well as the corresponding hidden constraints at the level of velocity coordinates (stabilized index-2 formulation). With an appropriate initialization of the acceleration like variables $\mathbf{a}_n$ in the generalized-α method, this stabilized index-2 Lie group DAE integrator is second-order accurate for any starting values being consistent with original and hidden constraints in the equations of motion.

All results of the convergence analysis have been verified in detail by numerical tests for a heavy top benchmark problem in Lie groups $\mathrm{SO}(3) \times \mathbb{R}^3$ and $\mathrm{SE}(3)$, respectively. The theoretical investigations are limited to fixed time step sizes but will be extended to variable step size implementations with error control in future work. In that case, the acceleration like variables $\mathbf{a}_n$ need to be updated whenever the time step size is changed at $t = t_n$. Furthermore, the velocity vector $\mathbf{v}_n$ has to be perturbed by an appropriate second-order correction term unless the generalized-α Lie group DAE integrator is applied to the index-reduced stabilized index-2 formulation of the equations of motion.

Acknowledgments The work of this author was supported by the German Minister of Education and Research, BMBF grant 05M13NHB: "Modelling and structure preserving discretization of inelastic components in system simulation".

References

Arnold, M., & Brüls, O. (2007). Convergence of the generalized-α scheme for constrained mechanical systems. *Multibody System Dynamics*, *18*, 185–202. doi:10.1007/s11044-007-9084-0.

Arnold, M., Brüls, O., & Cardona, A. (2011a). Convergence analysis of generalized-α Lie group integrators for constrained systems. In J. C. Samin & P. Fisette (Eds.), *Proceedings of Multibody Dynamics 2011 (ECCOMAS Thematic Conference)*, Brussels, Belgium.

Arnold, M., Brüls, O., & Cardona, A. (2011b). Improved stability and transient behaviour of generalized-α time integrators for constrained flexible systems. In *Fifth International Conference on Advanced COmputational Methods in ENgineering (ACOMEN 2011)*, Liège, Belgium, 14–17 November 2011.

Arnold, M., Burgermeister, B., Führer, C., Hippmann, G., & Rill, G. (2011c). Numerical methods in vehicle system dynamics: State of the art and current developments. *Vehicle System Dynamics*, *49*, 1159–1207. doi:10.1080/00423114.2011.582953.

Arnold, M., Cardona, A., & Brüls, O. (2014). Order reduction in time integration caused by velocity projection. In *Proceedings of the 3rd Joint International Conference on Multibody System Dynamics and the 7th Asian Conference on Multibody Dynamics*, 30 June–3 July 2014. *BEXCO*. Korea: Busan.

Arnold, M., Brüls, O., & Cardona, A. (2015). Error analysis of generalized-α Lie group time integration methods for constrained mechanical systems. *Numerische Mathematik*, *129*, 149–179. doi:10.1007/s00211-014-0633-1.

Betsch, P., & Leyendecker, S. (2006). The discrete null space method for the energy consistent integration of constrained mechanical systems. Part II: Multibody dynamics. *International Journal for Numerical Methods in Engineering*, *67*, 499–552. doi:10.1002/nme.1639.

Betsch, P., & Siebert, R. (2009). Rigid body dynamics in terms of quaternions: Hamiltonian formulation and conserving numerical integration. *International Journal for Numerical Methods in Engineering*, *79*, 444–473. doi:10.1002/nme.2586.

Bottasso, C. L., & Borri, M. (1998). Integrating finite rotations. *Computer Methods in Applied Mechanics and Engineering*, *164*, 307–331. doi:10.1016/S0045-7825(98)00031-0.

Bottasso, C. L., Bauchau, O. A., & Cardona, A. (2007). Time-step-size-independent conditioning and sensitivity to perturbations in the numerical solution of index three differential algebraic equations. *SIAM Journal on Scientific Computing*, *29*, 397–414. doi:10.1137/050638503.

Brenan, K. E., Campbell, S. L., & Petzold, L. R. (1996). *Numerical solution of initial-value problems in differential-algebraic equations* (2nd ed.). Philadelphia: SIAM.

Brüls, O., & Arnold, M. (2008). The generalized-α scheme as a linear multistep integrator: Towards a general mechatronic simulator. *Journal of Computational and Nonlinear Dynamics*, *3*(4), 041007. doi:10.1115/1.2960475.

Brüls, O., & Cardona, A. (2010). On the use of Lie group time integrators in multibody dynamics. *Journal of Computational and Nonlinear Dynamics*, *5*, 031002. doi:10.1115/1.4001370.

Brüls, O., & Golinval, J. C. (2006). The generalized-α method in mechatronic applications. *ZAMM—Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, *86*, 748–758. doi:10.1002/zamm.200610283.

Brüls, O., Arnold, M., & Cardona, A. (2011). Two Lie group formulations for dynamic multibody systems with large rotations. In *Proceedings of IDETC/MSNDC 2011, ASME 2011 International Design Engineering Technical Conferences*, Washington, USA.

Brüls, O., Cardona, A., & Arnold, M. (2012). Lie group generalized-α time integration of constrained flexible multibody systems. *Mechanism and Machine Theory*, *48*, 121–137. doi:10.1016/j.mechmachtheory.2011.07.017.

Cardona, A., & Géradin, M. (1989). Time integration of the equations of motion in mechanism analysis. *Computers and Structures*, *33*, 801–820. doi:10.1016/0045-7949(89)90255-1.

Cardona, A., & Géradin, M. (1994). Numerical integration of second order differential-algebraic systems in flexible mechanism dynamics. In M. F. O. Seabra Pereira & J. A. C. Ambrósio (Eds.), *Computer-Aided Analysis of Rigid and Flexible Mechanical Systems*, *NATO ASI Series* (Vol. E–268). Dordrecht: Kluwer Academic Publishers. doi:10.1007/978-94-011-1166-9_16.

Celledoni, E., & Owren, B. (2003). Lie group methods for rigid body dynamics and time integration on manifolds. *Computer Methods in Applied Mechanics and Engineering*, *192*, 421–438. doi:10.1016/S0045-7825(02)00520-0.

Chung, J., & Hulbert, G. (1993). A time integration algorithm for structural dynamics with improved numerical dissipation: The generalized-α method. *ASME Journal of Applied Mechanics*, *60*, 371–375. doi:10.1115/1.2900803.

Crouch, P. E., & Grossman, R. (1993). Numerical integration of ordinary differential equations on manifolds. *Journal of Nonlinear Science*, *3*, 1–33. doi:10.1007/BF02429858.

Deuflhard, P., Hairer, E., & Zugck, J. (1987). One-step and extrapolation methods for differential-algebraic systems. *Numerische Mathematik*, *51*, 501–516. doi:10.1007/BF01400352.

Erlicher, S., Bonaventura, L., & Bursi, O. (2002). The analysis of the generalized-α method for non-linear dynamic problems. *Computational Mechanics*, *28*, 83–104. doi:10.1007/s00466-001-0273-z.

García de Jalón, J., & Bayo, E. (1994). *Kinematic and dynamic simulation of multibody systems: The real-time challenge*. New York: Springer-Verlag.

García de Jalón, J., & Gutiérrez-López, M. D. (2013). Multibody dynamics with redundant constraints and singular mass matrix: Existence, uniqueness, and determination of solutions for accelerations and constraint forces. *Multibody System Dynamics*, *30*, 311–341. doi:10.1007/s11044-013-9358-7.

Gear, C. W., Leimkuhler, B., & Gupta, G. K. (1985). Automatic integration of Euler-Lagrange equations with constraints. *Journal of Computational and Applied Mathematics*, *12&13*, 77–90. doi:10.1016/0377-0427(85)90008-1.

Géradin, M., & Cardona, A. (2001). *Flexible multibody dynamics: A finite element approach*. Chichester: Wiley.

Géradin, M., & Cardona, A. (1989). Kinematics and dynamics of rigid and flexible mechanisms using finite elements and quaternion algebra. *Computational Mechanics*, *4*, 115–135. doi:10.1007/BF00282414.

Golub, G. H., & van Loan, Ch. F. (1996). *Matrix computations* (3rd ed.). Baltimore London: The Johns Hopkins University Press.

Hairer, E., & Wanner, G. (1996). *Solving ordinary differential equations. II. Stiff and differential-algebraic problems* (2nd ed.). Berlin, Heidelberg, New York: Springer-Verlag.

Hairer, E., Nørsett, S. P., & Wanner, G. (1993). *Solving ordinary differential equations. I. Nonstiff problems* (2nd ed.). Berlin, Heidelberg, New York: Springer-Verlag.

Hairer, E., Lubich, Ch., & Wanner, G. (2006). *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations* (2nd ed.). Berlin, Heidelberg, New York: Springer-Verlag.

Hilber, H. M., & Hughes, T. J. R. (1978). Collocation, dissipation and 'overshoot' for time integration schemes in structural dynamics. *Earthquake Engineering and Structural Dynamics*, *6*, 99–117. doi:10.1002/eqe.4290060111.

Hilber, H. M., Hughes, T. J. R., & Taylor, R. L. (1977). Improved numerical dissipation for time integration algorithms in structural dynamics. *Earthquake Engineering and Structural Dynamics*, *5*, 283–292. doi:10.1002/eqe.4290050306.

Iserles, A., Munthe-Kaas, H. Z., Nørsett, S., & Zanna, A. (2000). Lie-group methods. *Acta Numerica*, *9*, 215–365.

Jay, L. O., & Negrut, D. (2007). Extensions of the HHT-method to differential-algebraic equations in mechanics. *Electronic Transactions on Numerical Analysis*, *26*, 190–208.

Jay, L. O., & Negrut, D. (2008). A second order extension of the generalized-α method for constrained systems in mechanics. In C. Bottasso (Ed.), *Multibody Dynamics. Computational Methods and Applications*, *Computational Methods in Applied Sciences* (Vol. 12, pp. 143–158). Dordrecht: Springer. doi:10.1007/978-1-4020-8829-2_8.

Kelley, C. T. (1995). *Iterative methods for linear and nonlinear equations*. Philadelphia: SIAM.

Kobilarov, M., Crane, K., & Desbrun, M. (2009). Lie group integrators for animation and control of vehicles. *ACM Transactions on Graphics*, *28*(2, Article 16), 1–14. doi:10.1145/1516522.1516527.

Lubich, Ch. (1993). Integration of stiff mechanical systems by Runge-Kutta methods. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, *44*, 1022–1053. doi:10.1007/BF00942763.

Lunk, C., & Simeon, B. (2006). Solving constrained mechanical systems by the family of Newmark and α-methods. *Zeitschrift für Angewandte Mathematik und Mechanik*, *86*, 772–784. doi:10.1002/zamm.200610285.

Müller, A. (2010). Approximation of finite rigid body motions from velocity fields. *ZAMM—Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, *90*, 514–521. doi:10.1002/zamm.200900383.

Müller, A., & Terze, Z. (2014a). On the choice of configuration space for numerical Lie group integration of constrained rigid body systems. *Journal of Computational and Applied Mathematics*, *262*, 3–13. doi:10.1016/j.cam.2013.10.039.

Müller, A., & Terze, Z. (2014b). The significance of the configuration space Lie group for the constraint satisfaction in numerical time integration of multibody systems. *Mechanism and Machine Theory*, *82*, 173–202. doi:10.1016/j.mechmachtheory.2014.06.014.

Munthe-Kaas, H. (1995). Lie-Butcher theory for Runge-Kutta methods. *BIT Numerical Mathematics*, *35*, 572–587. doi:10.1007/BF01739828.

Munthe-Kaas, H. (1998). Runge-Kutta methods on Lie groups. *BIT Numerical Mathematics*, *38*, 92–111. doi:10.1007/BF02510919.

Negrut, D., Rampalli, R., Ottarsson, G., & Sajdak, A. (2005). On the use of the HHT method in the context of index 3 differential algebraic equations of multi-body dynamics. In J. M. Goicolea, J. Cuadrado & J. C. García Orden (Eds.), *Proceedings of Multibody Dynamics 2005 (ECCOMAS Thematic Conference)*, Madrid, Spain.

Orel, B. (2010). Accumulation of global error in Lie group methods for linear ordinary differential equations. *Electronic Transactions on Numerical Analysis*, *37*, 252–262.

Petzold, L. R., & Lötstedt, P. (1986). Numerical solution of nonlinear differential equations with algebraic constraints II: Practical implications. *SIAM Journal on Scientific and Statistical Computing*, *7*, 720–733. doi:10.1137/0907049.

Quarteroni, A., Sacco, R., & Saleri, F. (2000). *Numerical mathematics*. New York: Springer.

Sanborn, G. G., Choi, J., & Choi, J. H. (2014). Review of RecurDyn integration methods. In Proceedings of the 3rd Joint International Conference on Multibody System Dynamics and the 7th Asian Conference on Multibody Dynamics, 30 June–3 July (2014). *BEXCO*. Korea: Busan.

Simo, J. C., & Vu-Quoc, L. (1988). On the dynamics in space of rods undergoing large motions—A geometrically exact approach. *Computer Methods in Applied Mechanics and Engineering*, *66*, 125–161. doi:10.1016/0045-7825(88)90073-4.

Sonneville, V., Cardona, A., & Brüls, O. (2014). Geometrically exact beam finite element formulated on the special Euclidean group. *Computer Methods in Applied Mechanics and Engineering*, *268*, 451–474. doi:10.1016/j.cma.2013.10.008.

Walter, W. (1998). *Ordinary Differential Equations*. Number 182 in Graduate Texts in Mathematics. Berlin: Springer.

Wensch, J. (2001). Extrapolation methods in Lie groups. *Numerische Mathematik*, *89*, 591–604. doi:10.1007/s211-001-8017-5.