JAY

UNIVERSITÉ DE GENÈVE
Section de Mathématiques

FACULTÉ DES SCIENCES
Professeurs E. Hairer et G. Wanner

MÉTHODES DU TYPE RUNGE-KUTTA POUR DES ÉQUATIONS DIFFÉRENTIELLES ALGÉBRIQUES D'INDEX 3 AVEC DES APPLICATIONS AUX SYSTÈMES HAMILTONIENS

THÈSE

présentée à la Faculté des Sciences
de l'Université de Genève
pour obtenir le grade de Docteur ès Sciences,
mention Mathématiques

par

Laurent-Olivier JAY
de
Bernex (Genève)

Thèse N° 2658

Genève
Imprimerie Below The Line Productions S.A.
1994

UNIVERSITÉ DE GENÈVE

FACULTE DES SCIENCES

Doctorat ès sciences
mention mathématiques

Thèse de *Monsieur Laurent Olivier J A Y*

intitulée :

"METHODES DU TYPE RUNGE-KUTTA POUR DES EQUATIONS DIFFERENTIELLES ALGEBRIQUES D'INDEX 3 AVEC DES APPLICATIONS AUX SYSTEMES HAMILTONIENS."

La Faculté des Sciences, sur le préavis de Messieurs E. HAIRER, professeur adjoint et G. WANNER, professeur ordinaire (Section de mathématiques) codirecteurs de thèse, R. JELTSCH, professeur (ETH Zürich - Seminar für angewandte mathematik) et C. LUBICH, professeur (Universität Würzburg - Institut für mathematik), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 24 janvier 1994

Thèse - 2658 -

Le Doyen, Pierre MOESCHLER

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives à la présentation des thèses de doctorat à l'Université de Genève".

Nombre d'exemplaires à livrer par colis séparé à la Faculté : - 10 -

Thus computing is, or at least should be,
intimately bound up with both the source of the problem
and the use that is going to be made of the answers—
it is not a step to be taken in isolation from reality.

Richard Wesley Hamming (1973).

Résumé

Par *équations différentielles algébriques* on entend tout système d'équations de la forme $R(y', y, x) = 0$ où $\partial R/\partial y'$ n'est pas de rang maximum. La formulation de nombreux problèmes ainsi que la modélisation d'une variété de systèmes physiques mènent de façon naturelle à de telles équations.

Cette thèse traite plus spécifiquement d'une classe d'équations différentielles algébriques, dites *semi-explicites d'index* 3 *sous forme de Hessenberg*, dont divers domaines d'application sont les suivants : les systèmes mécaniques et Hamiltoniens munis de contraintes, la dynamique moléculaire, les problèmes de contrôle optimal et la robotique. Les propriétés de méthodes numériques du type *Runge-Kutta* sont étudiées. On fait grand usage de structures arborescentes qui apparaissent dans le développement en série de Taylor des solutions exacte et numérique.

Un résultat principal de cette thèse est une démonstration d'une conjecture relative à la superconvergence des méthodes de Runge-Kutta dites en anglais *stiffly accurate*. Tout aussi intéressante est la découverte d'une classe de méthodes de Runge-Kutta *partitionnées* qui, lorsqu'appliquées à des systèmes Hamiltoniens munis de contraintes, permettent de préserver les contraintes et la structure symplectique du flot.

Abstract

The term *differential-algebraic equations (DAE's)* comprises all systems of equations of the form $R(y', y, x) = 0$ where $\partial R/\partial y'$ is not of full rank. The formulation of numerous problems and the modelling of many physical systems lead in a natural way to such equations.

This thesis deals more specifically with a class of DAE's, called *semi-explicit index 3 DAE's in Hessenberg form*, whose various application domains are given by mechanical and Hamiltonian systems with constraints, molecular dynamics, optimal control problems, and robotics. The properties of numerical methods of *Runge-Kutta* type are studied. We largely make use of tree-structures which appear in the Taylor-expansion of the exact and numerical solutions.

A main result of this thesis is a demonstration of a conjecture related to the superconvergence of *stiffly accurate* Runge-Kutta methods. Also interesting is the discovery of a class of *partitioned* Runge-Kutta methods which, when applied to Hamiltonian systems with constraints, allow to preserve the constraints and the symplectic structure of the flow.

Runge-Kutta Type Methods for Index Three Differential-Algebraic Equations with Applications to Hamiltonian Systems

Laurent-Olivier JAY

Université de Genève, Département de mathématiques, Rue du Lièvre 2-4,
Case postale 240, CH-1211 Genève 24, Switzerland.
e-mail: na.ljay@na-net.ornl.gov

Table of contents

Note.

Chapters I, II, and III are new. Chapters IV and V are based on the texts of [Jay92] and [Jay94] respectively, added with technical details and proofs.

Summary.

The subject of this thesis deals with the *Numerical Analysis of Differential-Algebraic Equations (DAE's)*. DAE's consist in mixed systems of differential and algebraic (nonlinear) equations which cannot be explicitly or implicitly expressed as *Ordinary Differential Equations (ODE's)*. It is well-known that differential equations are a natural framework in which are modeled numerous problems in physics, in chemistry, and in technical applications. In addition to differential equations the models often contain implicit equations, in general purely algebraic (nonlinear) equations, in order to take into account conservation laws, geometrical or kinematic constraints, Kirchoff's laws, etc. DAE's arise typically in the following situations:

- in the motion of mechanical systems;
- in the study of constrained Hamiltonian systems, e.g., in molecular dynamics;
- in electrical circuit analysis;
- in chemical reaction kinetics;
- in the equations arising from the discretization of partial differential equations, such as in fluid dynamics;
- in control theory, e.g., in robotics;
- in the analysis of stiff differential equations.

Although the venerable field of ODE's is traditional in Mathematical Analysis since Newton's time, the systematic treatment of DAE's has really taken wing only in the last decade. DAE's differ in several aspects from ODE's and they present new analytical and numerical difficulties. DAE's can be characterized by the notion of index and DAE's with index strictly greater than 1, called *higher index DAE's*, are ill-posed in the sense that small perturbations may cause arbitrarily large changes in their solutions. The numerical treatment of such DAE's often leads to severe difficulties which can be overcome by reducing the index of the problem by different techniques. Many problems in the above-mentioned fields are formulated or lead to higher index DAE's.

Solving exactly ODE's or DAE's analytically is generally an impossible task. Hence numerical methods have been developed to approximate the solutions to these problems. The first numerical method for the integration of ODE's goes back to Euler's work. Nowadays, with the advent of computer technology, the interests in the modelling, the analysis, the simulation, and the control of various systems have increased enormously. Hence there is a considerable need for efficient and reliable numerical methods and software for DAE's. Many progresses have been made in the theoretical and numerical analysis of DAE's (see the books [BreCamPe89], [GrMä86], [HaLuRo89a], and [HaWa91]). For surveys on DAE's and numerical methods see [Pe89], [Rh91a], [GrHaMä91], [Pe92], [Mä92], and [HaJay93]. In certain situations DAE's can be reduced to ODE's and are therefore solvable by available standard ODE solvers. However, even in this case it may be in fact advantageous to work directly with a DAE. Hence a lot of numerical methods for ODE's have been especially adapted to DAE's. They comprise principally linear multistep methods, one-leg methods, linear implicit methods, Rosenbrock methods, Runge-Kutta methods, and some extrapolation methods. Because of a certain

connection between stiff ODE's and DAE's, stiffly accurate Runge-Kutta methods and backward differentiation formulas are of great interest in the DAE framework.

The main scope of this thesis is to study the application of *partitioned Runge-Kutta (PRK) methods* to *semi-explicit index* 3 *DAE's in Hessenberg form.* The emphasis is on stiffly accurate methods and we restrict ourselves to *initial value problems.* The organization of this thesis is as follows:

- In Chapter I we review some fundamental notions and results related to DAE's and to their numerical treatment. After having described some common basic structures of DAE's we then discuss the important concepts of solvability and index. Next we give some current examples of higher index DAE's. Further we present some available techniques to reduce the index of a problem. We then review some common numerical methods used for the solution of DAE's. Finally a brief overview of the scope of this thesis and the main convergence results is given.

- In Chapter II we give theoretical results related to semi-explicit index 3 DAE's in Hessenberg form. After characterizing the set of consistent values and the index of the problem, we then derive the Taylor expansion of the exact solution by means of a "rooted-tree" type notation. In this setting the theory of B-series due to Hairer and Wanner is extended to such problems giving birth to the new denominated $DA3$-series theory.

- Chapter III deals with the application of (projected) partitioned Runge-Kutta methods to semi-explicit index 3 DAE's in Hessenberg form. Results about the existence and uniqueness, the influence of perturbations, the local error, and the global error of the numerical solution are given. A short discussion on the application of simplified Newton iterations to the arising nonlinear system ends this chapter.

- The next two chapters are similar to Chapter III with the addition of some numerical experiments. In Chapter IV we restrict ourselves to the direct application of pure Runge-Kutta methods to semi-explicit index 3 DAE's in Hessenberg form. A proof of a conjecture related to the superconvergence of stiffly accurate Runge-Kutta methods is given, together with an application of this result to the convergence analysis of these methods for stiff mechanical systems. In Chapter V we mainly deal with the application of a special class of partitioned Runge-Kutta methods to Hamiltonian systems with holonomic constraints. These methods are superconvergent and preserve the symplectic structure of the flow and all underlying constraints as well.

Chapter I. Fundamentals of DAE's and numerical methods.

1. Introduction and notation.

The theoretical and numerical treatment of *differential-algebraic equations (DAE's)* is recent and still an open and active area of research. The systematic development of numerical methods for the solution of DAE's has begun with the original works of Gear [Ge71] and Petzold [Pe82]. In this chapter we review some fundamental notions and results related to DAE's and to their numerical treatment. In the last section we give a brief overview of the scope of this thesis and the main convergence results. In this thesis we restrict ourselves to *initial value problems.*

The following notations are used throughout this thesis. We denote by $\mathbb{R}^n$ the n-dimensional Euclidian space. Let $y = (y_1, \ldots, y_n)^T$ and $z = (z_1, \ldots, z_m)^T$ be two vectors (i.e., elements) of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively then we define the augmented vector $(y, z) \in \mathbb{R}^{n+m}$ by

$$(y, z) := (y^T, z^T)^T = (y_1, \ldots, y_n, z_1, \ldots, z_m)^T \in \mathbb{R}^{n+m} . \tag{1.1}$$

If the mapping

$$\Phi : \begin{cases} \mathbb{R}^{n+m} \to \mathbb{R}^k \\ (y, z) \mapsto \Phi(y, z) = (\Phi_1(y, z), \ldots, \Phi_k(y, z))^T \end{cases} \tag{1.2}$$

is differentiable at (y_0, z_0) then the Jacobian matrix at (y_0, z_0) of Φ with respect to y is given by

$$\Phi_y(y_0, z_0) := \frac{\partial \Phi}{\partial y}(y_0, z_0) = \begin{pmatrix} \frac{\partial \Phi_1}{\partial y_1}(y_0, z_0) & \cdots & \frac{\partial \Phi_1}{\partial y_n}(y_0, z_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial \Phi_k}{\partial y_1}(y_0, z_0) & \cdots & \frac{\partial \Phi_k}{\partial y_n}(y_0, z_0) \end{pmatrix} \tag{1.3}$$

and similarly for $\Phi_z(y_0, z_0)$. Higher derivatives are written as multilinear mappings. For example, for $v = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$ and $w = (w_1, \ldots, w_m)^T \in \mathbb{R}^m$ the ith-component of the expression

$$\Phi_{yz}(y_0, z_0)(v, w) \tag{1.4}$$

corresponds to

$$\sum_{j=1}^{n} \sum_{k=1}^{m} \frac{\partial^2 \Phi_i}{\partial y_j \partial z_k}(y_0, z_0) v_j w_k . \tag{1.5}$$

Finally, for an interval $I \subset \mathbb{R}$ we denote by $C^1(I, \mathbb{R}^n)$ the set of continuously differentiable mappings $u : I \to \mathbb{R}^n$.

2. Structure and solutions of DAE's.

In this section we review some elementary definitions related to the structure and solutions of *differential-algebraic equations (DAE's). Implicit differential equations* are general systems of nonlinear equations

$$\begin{aligned} R_1(x, y_1, \ldots, y_n, y_1', \ldots, y_n') &= 0 \ , \\ &\vdots \\ R_m(x, y_1, \ldots, y_n, y_1', \ldots, y_n') &= 0 \ , \end{aligned} \tag{2.1}$$

or more succinctly

$$R(x, y, y') = 0 \ , \tag{2.1}$$

where $m \geq n$, unless mentioned otherwise x is the one-dimensional variable of integration, $y = (y_1, \ldots, y_n)^T$, and $y' = (y_1', \ldots, y_n')^T = dy/dx$. Throughout the following sections the function R will be assumed to be sufficiently differentiable. Such a system (2.1) is said to be *overdetermined* if $m > n$ and *autonomous* if in the form

$$R(y, y') = 0 \ . \tag{2.2}$$

The non-autonomous system (2.1) can be put in autonomous form by simply appending the equation $x' - 1 = 0$ to (2.1) and by defining $y_{\text{new}} := (x, y)$.

Definition 2.1. A *(classical) solution* u to (2.1) on the interval I is a function $u \in C^1(I, \mathbb{R}^n)$ satisfying (2.1) for all $x \in I$.

If $R_{y'}$ is regular and $R(x_0, y_0, y_0') = 0$ then by the implicit function theorem we can locally rewrite the system (2.1), at least formally, as a system of *ordinary differential equations (ODE's)*

$$y' = f(x, y) \ . \tag{2.3}$$

In this case well-known theorems on the existence and uniqueness of a solution apply (see, e.g., [HaNøWa93, Section I.9] and [But87, Section 112]). However, the theoretical and numerical analysis of such equations is not the purpose of this thesis. In this thesis we are interested in systems (2.1) where $R_{y'}$ is not of maximal rank n (singular if $m = n$) which are designated as *differential-algebraic equations (DAE's)*. Such systems differ in several aspects from the precedent situation and they present new analytical and numerical difficulties. In opposition to ODE's which under reasonable assumptions admit a unique solution for any arbitrary initial value, there is no unified existence and uniqueness theory for general DAE's. For example DAE's may have solutions only for a subset of initial values called *consistent*, may have multiple solutions, may present bifurcations, may possess impasse or turning points, or may have no solution at all. Therefore we are led to deal separately with different types of problems. A common assumption is that the rank of $R_{y'}$ remains constant.

We review some basic types of DAE's. A situation which frequently occurs in practical applications is when the system (2.1) is *linear in the derivative*

$$B(x, y)y' + A(x, y) = 0 \ , \tag{2.4}$$

which is sometimes referred to as *linearly implicit*. Systems of the form

$$\begin{aligned} y' &= f(x,y,z)\ , \\ 0 &= g(x,y,z)\ , \end{aligned} \tag{2.5}$$

or more generally

$$\begin{aligned} F(x,y,z,y') &= 0\ , \\ G(x,y,z) &= 0 \end{aligned} \tag{2.6}$$

where $F_{y'}$ is regular, are called *semi-explicit* systems. The variables y and z are called respectively the *differential* and *algebraic variables*. The general system (2.1) can be rewritten in the semi-explicit form (2.5) by simply considering

$$\begin{aligned} y' &= z\ , \\ 0 &= R(x,y,z)\ . \end{aligned} \tag{2.7}$$

Thus semi-explicit problems can be considered as the generic case. Semi-explicit DAE's in *Hessenberg form of size* $r \geq 2$ are given by

$$\begin{aligned} w^{1\prime} &= F^1(x,w^1,w^2)\ , \\ w^{2\prime} &= F^2(x,w^1,w^2,w^3)\ , \\ &\vdots \\ w^{r-1\prime} &= F^{r-1}(x,w^1,w^2,w^3,\ldots,w^{r-1},w^r)\ , \\ 0 &= F^r(x,w^1) \end{aligned} \tag{2.8}$$

where the product matrix $F^r_{w^1}F^1_{w^2}\cdots F^{r-1}_{w^r}$ is supposed to be regular in a neighbourhood of the solution. Semi-explicit DAE's in *Hessenberg form of size* 1 are given by (2.5) where the matrix g_z is supposed to be regular in a neighbourhood of the solution. There are other important classes of DAE's discussed in the literature (see, e.g., [BreCamPe89, Chapter 2]), namely: linear constant DAE's, linear time varying DAE's, triangular chains, etc. In the other chapters we will deal with autonomous semi-explicit DAE's in Hessenberg form of size 3.

3. Two key concepts: solvability and index.

Two key concepts in the analysis and classification of DAE's are played by the notions of solvability and index. Both of them have evolved over the years and there are many possible refinements in their definition. Here we restrict ourselves to the easiest and most comprehensive definitions.

3.1. Solvability.

We begin with the concept of *solvability* which intuitively means the existence of a family of solutions to the DAE (2.1). Since not all initial values for y admit a solution, we are interested in the cases where the solutions form locally an r-dimensional manifold. The following definition is close to the usual definitions (see [BreCamPe89, p. 16], [PePo92], and [Po93a,c]):

Definition 3.1. [CamGe93]. Let $\Omega \subset \mathbb{R}^{2n+1}$ be a non-empty connected open set. Then the DAE (2.1) is *(locally) geometrically solvable on* Ω if there exist non-empty connected open sets $I \subset \mathbb{R}$ and $\Lambda \subset \mathbb{R}^r$, and a function Φ such that

1. $\Theta : \begin{cases} I \times \Lambda \to I \times \mathbb{R}^n \\ (x, \lambda) \mapsto (x, \Phi(x, \lambda)) \end{cases}$ is a diffeomorphism of $I \times \Lambda$ into $I \times \mathbb{R}^n$.
2. For all $\lambda \in \Lambda$, $\Phi(\cdot, \lambda)$ is a solution to (2.1) on the interval I, i.e.,
$$R(x, \Phi(x, \lambda), \Phi_x(x, \lambda)) = 0 \qquad \forall x \in I \ , \quad \forall \lambda \in \Lambda \ .$$
3. For all $x \in I$ and $\lambda \in \Lambda$ we have $(x, \Phi(x, \lambda), \Phi_x(x, \lambda)) \in \Omega$.
4. If u is a solution of (2.1) on the open interval J such that $(x_0, u(x_0), u'(x_0)) \in \Omega$ for some $x_0 \in I$, then there exists $\lambda \in \Lambda$ such that $u(x) = \Phi(x, \lambda)$ on $I \cap J$.

A value (x_0, u_0) is called *geometrically consistent* if there exists $\lambda_0 \in \Lambda$ such that $u_0 = \Phi(x_0, \lambda_0)$.

Remark 3.1. Bifurcations of solutions may not occur in Ω under these assumptions.

It is obvious that standard ODE's (2.3) and implicit differential equations (2.1) with regular $R_{y'}$ as well are geometrically solvable on $\mathbb{R}^{2n+1}$. It can also be easily shown that semi-explicit DAE's in Hessenberg form of size r are geometrically solvable within a set of consistent values (see Section II.1 for the size 3 case). In view of this definition it turns out that geometric solvability implies the existence of a unique vector field on the solution manifold Ω. The reverse is also true provided the vector field is differentiable (see [Rei92]). This has led several authors (see [Rh84], [Rei90a,b], [Rh91b], [Rei91], [Rei92], [RaRh91], [Gr91], [Mä93], [RaRh94]) to characterize DAE's to which can be translated well-known results on vector fields on manifolds [ArV.92, Chapter 5]. In this thesis we deal with DAE's whose solvability is obvious or is assumed.

The local solvability of the DAE (2.1) can be proved by showing the existence of a *local state-space form* which is a system of r ODE's whose solutions are in one-to-one correspondence with solutions of (2.1) (see Subsection 5.2). For a computational verification of solvability see [CamGr93].

3.2. Index.

The second concept is that of *index* and it provides a convenient way of classification of DAE's. There are many different definitions of the index. We present here the *differential index* and the *perturbation index* which are common in the literature.

Generally a solution to a DAE (2.1) is known to depend on the derivatives of R. Differentiating (2.1) k times with respect to x, we obtain the $(k+1)m$ *derivative array equations* (see [BreCamPe89, p. 32] and [CamGr93])

$$R^k(x, y, y', \ldots, y^{(k+1)}) := \begin{pmatrix} R(x, y, y') \\ R_x(x, y, y') + R_y(x, y, y')y' + R_{y'}(x, y, y')y'' \\ \vdots \\ \frac{d^k}{dx^k} R(x, y, y') \end{pmatrix} = 0 \ . \quad (3.1)$$

Definition 3.2. [GeGupLe85], [BreCamPe89, p. 33], [CamGe93]. The *differential index* $\nu_{d,i}$ *of the variable* y_i is the smallest integer k for which y_i' is uniquely determined by

R^k as a continuous function of (x,y) (supposed to be consistent). The *differential index* ν_d is then defined by $\nu_d := \max_{i=1,\dots,n} \nu_{d,i}$.

Thus the differential index is the minimum number of times that (2.1) must be differentiated with respect to x to determine y' as a continuous function of consistent values (x,y), i.e., to obtain an explicit ODE for y' called the *underlying ODE*.

It is well-known that the higher the index and the more difficulties are encountered when numerically solving DAE's. A convenient measure of the sensitivity of a solution to perturbations in the equations is given by the perturbation index.

Definition 3.3. [HaLuRo89a, Definition 1.1], [HaWa91, Definition VI.5.3]. The ith-component has *perturbation index* $\nu_{p,i}$ along a solution u on a bounded interval I passing through $u(x_0)$ at x_0, if $\nu_{p,i}$ is the smallest integer such that for all functions $\widehat{u}(x)$ having a defect

$$R(x,\widehat{u}(x),\widehat{u}'(x)) = \delta(x) \tag{3.2}$$

there exists on I an estimate

$$|\widehat{u}_i(x)-u_i(x)| \le C_i\left(\|\widehat{u}(x_0)-u(x_0)\|+\sup_{\zeta\in I}\| \int_{x_0}^{\zeta} \delta(\tau)d\tau\|+ \sum_{j=0}^{\nu_{p,i}-1} \sup_{\zeta\in I}\|\delta^{(j)}(\zeta)\| \right) \tag{3.3}$$

whenever the expression on the right-hand side is sufficiently small. Here C_i is a constant which depends only on R and on the length of the interval I. The *perturbation index* ν_p is then defined by $\nu_p := \max_{i=1,\dots,n} \nu_{p,i}$.

Thus the perturbation index is a measure on how strong the problem may be ill-posed. Unlike problems with perturbation index ν_p equal to 0 or 1, problems with perturbation index $\nu_p \ge 2$ are ill-posed in the sense that small perturbations may cause arbitrarily large changes in their solutions. Their numerical treatment often leads to severe difficulties which can be overcome by reducing the index of the problem to 0 or 1 by different techniques (see Section 5).

In view of the above definitions it turns out that standard ODE's (2.3) and implicit differential equations (2.1) with regular $R_{y'}$ as well are of differential and perturbation index 0. It can also be easily shown that semi-explicit DAE's in Hessenberg form of size r are of differential and perturbation index r (see Section II.1 for the size 3 case). For a practical computation of the differential index see [Pan88] (for linear constant DAE's see [BujBo93]).

Both above-defined indices are referred as *standard indices*. We have the relation $\nu_d \le \nu_p$ (see [Ge90]), but these indices may be very different from each other. An example of problems with $\nu_d = 1$ and arbitrary high ν_p is given in [CamGe93]. These indices may also vary in a neighbourhood of a solution or may not exist at all (see [CamGe93]). For these reasons, a second class of indices referred as *uniform indices* has been introduced in [CamGe93] to overcome these troubles. Such indices are defined with respect to a class of perturbations. However, this subject is beyond the scope of this thesis and we refer the reader to [CamGe93].

It must be stressed that in most practical applications the various indices are equal. DAE's with index $\nu \ge 2$ are called *higher index DAE's* (see [HaWa91, Section VI.5]).

As mentioned before such systems are much more difficult to treat than index 0 and 1 DAE's. Nevertheless, they arise in many applications as we will see in the next section.

Notes.

For the important problem of finding consistent initial values in a general setting we refer to [Pan88] and [LePeGe91].

The first result for an existence and uniqueness theory for nonlinear DAE's appears to be for *gradient systems* (see [Ta76])

$$\begin{aligned} y' &= f(y,z) \ , \\ 0 &= \nabla_y h(y,z) = h_y^T(y,z) \ . \end{aligned} \tag{3.4}$$

Most of the research has been focused on solvable DAE's. Only a few researchers have dealt with more general DAE's, e.g., DAE's presenting bifurcations, turning points, or impasse points. For such problems we refer to [Ra89] and [RaRh92a,b].

The first definitions of index were related to linear DAE's by means of the *Weierstrass-Kronecker canonical form* for *matrix pencils* (see, e.g., [BreCamPe89, Section 2.3] and [HaWa91, Section VI.5]).

The first attempt to translate the theory of vector fields on manifolds to a special class of DAE's is due to Rheinboldt [Rh84]. This differential-geometric approach has then been extended by Reich to the class of regular DAE's. A *regular DAE* is a DAE such that a unique vector field can be related to the solutions of the DAE (see [Rei90a,b], [Rei91], and [Rei92]). This notion of *regularity* is closely related to the above-defined geometric solvability. The *geometrical index (the degree)* ν_g of Reich is based on the observation that a family of ν_g embedded manifolds containing the solutions of the DAE can be constructed. For further details on this approach see [Rei90a,b], [Rei91], and [Rei92]. A general existence and uniqueness theory for (2.2) based on constant rank assumptions for the Jacobians R_y and $R_{y'}$ is given in [RaRh91] showing that the geometric approach is conceptually valid. Within this framework an existence and uniqueness theory for the class of *nonsingular DAE's* is given in [RaRh94].

The notion of *index-ν_t-tractability* is a characterization of the problem in terms of suitable projections of solutions. This index is in close relationship with the concept of *transferability* which is nearly equivalent to the assumption of differential index 1. For further details we refer to [GrMä86], [Mä89], and [Mä92].

For a discussion on DAE's with discontinuities see [Ni90] and [PrBeDeSc92]. For DAE's with delays see [AsPe92b].

4. Examples of higher index DAE's.

In this section we give some current examples where higher index DAE's arise. For examples of DAE's with very high index see [Cam93b].

4.1. Mechanical and Hamiltonian systems with constraints.

Since the last decade there has been a growing interest in the modelling and simulation of *mechanical systems (MS's)* (for details see [Hau89], [Sch90], [Sch93], [BreCamPe89, Section 6.2], [HaLuRo89a, pp. 6-7], and [HaWa91, pp. 483-486 & Section VI.9]). Such problems arise in vehicle-systems simulation (see [Po93c]), in aerospace application (see [Bre83] and [Bre86]), in biomechanics (see [Ka93]), in robotics (see [JanBru92]), etc. In the *multibody system* approach, a MS is described in terms of *bodies* and *connections*. The mass is supposed to be entirely concentrated in the rigid and elastic bodies which are coupled by massless connections. Connections like springs and dampers are sources of applied forces acting on the bodies, whereas connections like joints constrain the relative motion of the bodies. The use of computer algebra programs to generate automatically the equations of motion of MS's is nowadays current (see [Sch90] and [Sch93]). These equations can be derived from the *Lagrange-Hamilton principle* as follows: Let $q = (q_1, \ldots, q_n)^T$ be the n *generalized coordinates* of the MS submitted to $m < n$ non-redundant *holonomic (position level) constraints* $g_1(q) = 0, \ldots, g_m(q) = 0$. Here the one-dimensional variable of integration is the time t and a derivative with respect to t is denoted by a dot ($\dot{}$). Then the motion of the MS is described by the solution of the constrained variational problem

$$\min \int_{t_0}^{t_1} L(q(\tau), \dot{q}(\tau)) d\tau \ , \qquad 0 = g(q) \tag{4.1}$$

where $L(q, \dot{q}) = T(q, \dot{q}) - U(q) + \lambda^T g(q)$ is the *Lagrangian*, $T(q, \dot{q}) = \frac{1}{2} \dot{q}^T M(q) \dot{q}$ is the *kinetic energy* with $M(q)$ the symmetrical *generalized mass matrix*, $U(q)$ the *potential energy*, and $\lambda = (\lambda_1, \ldots, \lambda_m)^T$ the vector of *generalized constraint forces* coupled to the system and commonly called the *Lagrange multipliers*. A necessary condition for this variational problem is given by the *Euler-Lagrange equations*

$$\frac{d}{dt} \left(\bar{L}_{\dot{q}}^T(q, \dot{q}) \right) - \bar{L}_q^T(q, \dot{q}) = 0 \tag{4.2}$$

with $\bar{L}(q, \dot{q}, \lambda) = L(q, \dot{q}) - \lambda^T g(q)$

leading to

$$\dot{q} = v \ , \tag{4.3a}$$

$$M(q)\dot{v} = f(q, v) - G^T(q)\lambda \ , \tag{4.3b}$$

$$0 = g(q) \tag{4.3c}$$

where $v = (v_1, \ldots, v_n)^T$ are the n *generalized velocities*,

$$f(q, v) := L_q^T(q, v) - L_{\dot{q}q}^T(q, v)v \tag{4.4}$$

is the vector of *generalized external forces*, and $G(q) := g_q(q)$. The formulation (4.3) is called the *descriptor form* and is referred in classical mechanics as the *Lagrange equations of the first kind*. The importance of this formulation lies in the fact that it

is independent of the choice of the coordinates. In order to determine the differential index of this DAE $(4.3a,b,c)$ we derive twice the algebraic constraints $(4.3c)$ and we obtain the following additional constraints

$$0 = G(q)v \ , \tag{4.3d}$$
$$0 = g_{qq}(q)(v,v) + G(q)\dot{v} \tag{4.3e}$$

which are called respectively the *velocity/acceleration level constraints*. If we consider $(4.3e)$ together with $(4.3b)$ we obtain

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} \dot{v} \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q,v) \\ -g_{qq}(q)(v,v) \end{pmatrix} \tag{4.5}$$

and the following result can be easily shown:

Lemma 4.1. [Hau89, pp. 225-227], [Ka93]. *Under the assumptions*

$$G(q) \quad \text{is of full row rank } m \ , \tag{4.6a}$$
$$w^T M(q) w > 0 \quad \forall w \neq 0 \text{ such that } G(q)w = 0 \tag{4.6b}$$

then the matrix on the left-hand side of (4.5) *is invertible.* □

Hence under these assumptions the differential index of the system $(4.3a,b,c)$ is equal to 3 and the whole system of equations $(4.3a,b,c,d,e)$ forms an overdetermined DAE of differential index 1. If in addition the matrix $M(q)$ is invertible then the system $(4.3a,b,c)$ can be rewritten in Hessenberg form of size 3 and the invertibility of the matrix in the left-hand side of (4.5) is equivalent to the invertibility of the matrix $(GM^{-1}G^T)(q)$.

If the mass matrix $M(q)$ is invertible then an alternative and equivalent formulation to the above Lagrange formulation is given by the *Hamilton formalism* as follows: Let $p=(p_1,\ldots,p_n)^T$ be the n *generalized momenta* of the system defined by

$$p := L_{\dot{q}}^T(q,\dot{q}) \ . \tag{4.7}$$

By a *Legendre transformation* of the Lagrangian $L(q,\dot{q})$ with respect to $\dot{q}$ (given by $H(q,p)=p^T\dot{q}(q,p)-L(q,\dot{q}(q,p))$) we obtain the equivalent Hamiltonian formulation

$$\dot{q} = H_p^T(q,p) \ , \tag{4.8a}$$
$$\dot{p} = -H_q^T(q,p) - G^T(q)\lambda \ , \tag{4.8b}$$
$$0 = g(q) \tag{4.8c}$$

where it can also be shown that $H=T+U$ is the *total energy* of the system. As previously done, differentiating (4.8c) twice we obtain the following additional constraints (omitting the obvious function arguments)

$$0 = GH_p^T \ , \tag{4.8d}$$
$$0 = G_q(H_p^T, H_p^T) + GH_{pq}^T H_p^T - GH_{pp}^T H_q^T - GH_{pp}^T G^T\lambda \ . \tag{4.8e}$$

Remark 4.1. These equations (4.8) are not restricted to the modelling of mechanical systems. For example such systems arise in molecular dynamics too (see [SkBiOk93]).

The following result can be easily shown:

Lemma 4.2. *Under the assumptions*

$$G(q) \quad \text{is of full row rank } m\ , \tag{4.9a}$$
$$H_{pp}(q,p) \quad \text{is a strictly positive definite matrix} \tag{4.9b}$$

then the matrix $(GH_{pp}^{T}G^{T})(q,p)$ *is invertible.* □

Remark 4.2. The second assumption (4.9*b*) means that we have an *optical Hamiltonian system* (see [MK92, p. 140]).

Hence under these assumptions the system (4.8*a*, *b*, *c*) is in Hessenberg form of size 3, therefore of index 3, and the whole system of equations (4.8*a*, *b*, *c*, *d*, *e*) forms an overdetermined DAE of index 1. The main advantage of the Hamiltonian formulation (4.8) over the Lagrangian formulation (4.3) is that it possesses more structures. The main property of Hamiltonian systems is that of *symplecticity*: in the phase space of (q,p) given by the $2(n-m)$-dimensional manifold

$$V = \left\{(q,p) \in \mathbb{R}^n \times \mathbb{R}^n \mid 0 = g(q)\ ,\ 0 = G(q)H_p^T(q,p)\right\} \tag{4.10}$$

the flow generated by (4.8) is *symplectic*, i.e., the differential 2-form

$$\omega^2 = \sum_{k=1}^{n} dq_k \wedge dp_k \qquad \text{is preserved}\ , \tag{4.11}$$

implying that all differential $2d$-forms

$$\underbrace{\omega^2 \wedge \ldots \wedge \omega^2}_{d \text{ times}} \qquad \text{for} \quad d = 1, \ldots, n \tag{4.12}$$

are also conserved ($d = n$ corresponds to the $2n$-form *volume*). This is one way of characterizing symplectic transformations. Another specific feature of such systems is that the Hamiltonian along a solution $(q(t), p(t))$ to (4.8) passing through (q_0, p_0) at t_0 remains invariant, i.e.,

$$H(q(t), p(t)) = H(q_0, p_0) \qquad \text{for all } t. \tag{4.13}$$

Hamiltonian systems also possess numerous other specific properties (see [ArV.89, Part III] and [MK92]).

Notes.

The Lagrange formulation (4.3) of the motion of MS's is not the most general one. The following additional factors may be present in a MS model:

- a MS may be subject to *non-holonomic constraints* $0 = K(q)v+k(q)$, e.g., in order to model the contact of a rolling wheel on a surface [SiFüRen91]. Such constraints can be incorporated into the system (4.3) by addition of supplementary Lagrange multipliers into (4.3*b*);

- a MS may be subject to *flexibility* leading to DAE models with very high indices (see [Cam93b]);

- a MS may be subject to *external excitations* $u(t)$ due to *external (actuator) dynamics* elements like, e.g., dampers and motors (see [SiFüRen91]). These excitations can be modelled by differential equations of the type

$$\dot{u} = d(q, v, \lambda, u) \tag{4.14}$$

  and they enter in the generalized external forces $f = f(q, v, \lambda, u)$ and in the holonomic constraints $g = g(q, u)$;

- the systems (4.3) and (4.8) may be non-autonomous;

- a MS model may present discontinuities or non-differentiabilities due to, e.g., friction forces, impact, discrete time controllers, and tabulated data (see [SiFüRen91], [Ei92], and [AnBoEiSc93]).

4.2. Singular singularly perturbed problems.

Singular perturbed problems form a particular class of stiff differential equations containing a small parameter $0 < \varepsilon \ll 1$ (see [OM74] and [HW91, Chapter VI])

$$y' = f(x, y, z) , \tag{4.15a}$$
$$\varepsilon z' = g(x, y, z) . \tag{4.15b}$$

The analysis of the limit case $\varepsilon = 0$, the *reduced problem*

$$y' = f(x, y, z) , \tag{4.16a}$$
$$0 = g(x, y, z) \tag{4.16b}$$

often gives much insight into the behaviour of solutions to (4.15). Under the stability assumption

$$\langle g_z(x, y, z)w, w\rangle \leq -Const \cdot \|w\|^2 \qquad \text{with } Const > 0 , \tag{4.17}$$

which implies that

$$g_z(x, y, z) \quad \text{is invertible} , \tag{4.18}$$

smooth solutions to (4.15) for suitable initial values are known to possess an ε-expansion

$$\begin{aligned} y(x) &= y^0(x)+\varepsilon y^1(x)+\ldots+\varepsilon^N y^N(x)+\mathcal{O}(\varepsilon^{N+1}) , \\ z(x) &= z^0(x)+\varepsilon z^1(x)+\ldots+\varepsilon^N z^N(x)+\mathcal{O}(\varepsilon^{N+1}) \end{aligned} \tag{4.19}$$

where $(y^0(x), z^0(x))$ is solution of the reduced problem (4.16) (which is an index 1 DAE) and $(y^k(x), z^k(x))$ are solution of index $k+1$ DAE's. However, in certain situations, the matrix g_z may possess zero-eigenvalues implying that the condition (4.18) is not satisfied. In this case we speak of *singular singularly perturbed problems. Stiff mechanical*

systems in which a strong potential $\frac{1}{\varepsilon^2}V(q)$ forces the motion to be close to a manifold fall into this category (see [Lu93]). A precise formulation for unconstrained systems is as follows:

$$\dot{q} = v \,, \tag{4.20a}$$

$$M(q)\dot{v} = f(q,v) - \frac{1}{\varepsilon^2}V_q^T(q) \tag{4.20b}$$

with the assumptions (SMS) (for Stiff Mechanical Systems)

- $M(q)$ is symmetrical and positive definite;
- $V(q)$ attains a (local) minimum on an m-dimensional manifold V;
- in a neighbourhood of V, $V(q)$ is strongly convex along directions non-tangential to V.

Under these hypotheses (SMS) it can be shown that for suitable initial values smooth solutions to (4.20) possess an ε^2-expansion

$$q(t) = q^0(t) + \varepsilon^2 q^1(t) + \ldots + \varepsilon^{2N} q^N(t) + \mathcal{O}(\varepsilon^{2N+2}) \,, \tag{4.21a}$$

$$v(t) = v^0(t) + \varepsilon^2 v^1(t) + \ldots + \varepsilon^{2N} v^N(t) + \mathcal{O}(\varepsilon^{2N+2}) \tag{4.21b}$$

where $(q^0(t), v^0(t))$ is solution of an index 3 problem $(4.3a,b,c)$ with g such that it vanishes on V (and only there) and $G(q) := g_q(q)$ has full row rank m, and $(q^k(t), v^k(t))$ are solution of index $2k+3$ DAE's (see [Lu93]). A typical example of a stiff mechanical system is given by the *stiff spring pendulum* (see [Lu93] and [HaLuRo89a, pp. 10-12]). It consists of a mass point m suspended at a massless spring with a large Hooke's constant $1/\varepsilon^2$, $0 < \varepsilon \ll 1$. Using cartesian coordinates $q = (x,z)^T$, the kinetic energy T and the potential energy U of the system are given respectively by

$$T(\dot{q}) = \frac{m}{2}(\dot{x}^2 + \dot{z}^2), \qquad U(q) = mgz + \frac{1}{2\varepsilon^2}\left(\sqrt{x^2+z^2} - \ell\right)^2 \tag{4.22}$$

where ℓ denotes the rest position of the spring and g the gravitational constant. The Lagrange equations of motion (4.20) are then given by

$$\dot{x} = v_x, \qquad \dot{z} = v_z, \qquad m\dot{v}_x = -\frac{x}{\ell}\lambda, \qquad m\dot{v}_z = -mg - \frac{z}{\ell}\lambda \tag{4.23}$$

where we have defined λ by

$$\frac{\varepsilon^2\lambda}{\ell}\sqrt{x^2+z^2} = \sqrt{x^2+z^2} - \ell \,. \tag{4.24}$$

For the limit case $\varepsilon = 0$ (the pendulum equations), we obtain the constraint

$$0 = \sqrt{x^2+z^2} - \ell \tag{4.25}$$

and λ has now the role of a Lagrange multiplier. It is easy to show that the so obtained differential-algebraic system (4.23)-(4.25) is of the form $(4.3a,b,c)$ and therefore of index 3.

4.3. Control problems.

In *control theory* a process is generally described by a system of differential equations

$$\dot{y} = f(t, y, u) \tag{4.26}$$

where u represents a set of control parameters varying with time t (see [BryHo75]). In fact the process can be itself modelled by a DAE, but we restrict here our discussion to the case (4.26) in order to avoid a possible confusion in the derivation below. The control parameters u are usually chosen so that the solution satisfies some constraints

$$0 = g(t, y, u) \tag{4.27}$$

or/and minimizes some cost functional. In the first situation u is frequently absent from (4.27) and hence the index of (4.26)-(4.27) may be very high. This is often the case in *prescribed path control* problems where the goal is to adjust the control parameters u so that the trajectory follows some prescribed path

$$0 = g(t, y) \ . \tag{4.28}$$

Examples of such problems with very high indices arise in robotics. In *space vehicles simulation* examples of index 2 and 3 problems are given in [Bre83], [Bre86], and [BreCamPe89, Section 6.3], whereas an index 5 problem is described in [Cam93b]. In the second situation a common case is when the cost functional to be minimized is of the form

$$J(u) = \int_{t_0}^{t_f} \varphi(\tau, y(\tau), u(\tau)) d\tau \ . \tag{4.29}$$

In this case for the fixed time, fixed endpoint problem, the variational equations for (4.26)-(4.29) are given by the semi-explicit DAE (see [BryHo75, p. 49], [BreCamPe89, p. 6], and [HaWa91, p. 482])

$$\dot{y} = f(t, y, u) \ , \tag{4.30a}$$

$$\dot{z} = -\varphi_y^T(t, y, u) - f_y^T(t, y, u) z \ , \tag{4.30b}$$

$$0 = \varphi_u^T(t, y, u) + f_u^T(t, y, u) z \ , \tag{4.30c}$$

with $y(t_0) = y_0$ and $z(t_f) = 0$. This is a two-point *DAE boundary value problem* which can also be directly obtained from the *Pontryagin principle* (see [PonBoGaMi62]). If the matrix

$$\varphi_{uu}^T(t, y, u) + f_{uu}^T(t, y, u) z \tag{4.31}$$

is invertible then the differential index of $(4.30a, b, c)$ is equal to 1. However, this condition is not always satisfied. For example, consider a linear control problem

$$\dot{y} = Ay + Bu + f(t) \tag{4.32}$$

with cost functional

$$J(u) = \frac{1}{2} \int_{t_0}^{t_f} y(\tau)^T C y(\tau) + u(\tau)^T D u(\tau) d\tau \tag{4.33}$$

where A, B, C, and D are constant matrices with C and D symmetric and positive semi-definite. In this situation the equations $(4.30a, b, c)$ read

$$\dot{y} = Ay + Bu + f(t) \ , \tag{4.34a}$$

$$\dot{z} = -Cy - A^T z \ , \tag{4.34b}$$

$$0 = Du + B^T z \ . \tag{4.34c}$$

If $D = 0$ and $B^T C B$ is positive definite, then $(4.34a, b, c)$ is in Hessenberg form of size 3, i.e., of index 3. If $B^T C B = 0$ then the index is at least equal to 5. Both problems belong to the class of *cheap control problems* which have been extensively studied since the late seventies (see [Cam80] and [Cam82]).

5. Index transformation techniques.

As mentioned in Subsection 3.2, higher index DAE's are ill-posed in the sense that small perturbations may cause arbitrarily large changes in their solutions (see also [GrMä86], [Han90], [Mä85a,b], and [AsPe93]). Hence different techniques have been proposed to change the index of such problems to 0 or 1 by considering another problem possessing the same solutions as the original problem. In this section we review some current index transformation techniques. For a stability analysis of some of these techniques see [AsPe93].

5.1. Index reduction by differentiation.

A way of reducing the index of higher DAE's is often to differentiate analytically the equations and to do some algebraic manipulations until a DAE of index 0 or 1, or even an ODE is obtained (see [Ge88] and [BreCamPe89, Subsection 2.5.3]). This approach is natural for semi-explicit problems, especially for those in Hessenberg form of size r where after $r-1$ or r differentiations of the algebraic constraints an index 1 DAE or an explicit ODE can be obtained. The advantage of this method is that a great variety of efficient and reliable codes are applicable to the resulting equations. However, this approach has some drawbacks. The first one is that the structure of the original DAE may be lost, e.g., sparsity of the system structure may be destroyed and meaningless variables may be introduced. A second drawback is that in practice it may be very complicated or even impossible to rewrite a DAE into an ODE. A third drawback is that the analytical and stability properties of the resulting equations may be drastically different from the original DAE, creating new difficulties (see [BreCamPe89, Subsection 5.4.1], [FüLe89, Example 2.1], and [AsPe93]). A fourth drawback is that the numerical solution generally no longer satisfies the constraints, thus giving a meaningless solution. A way to circumvent this "drift off" phenomena is (if at all possible) to project the numerical solution back to the solution manifold. This "coordinate" *projection method* is in fact recommended and can improve substantially the stability properties of the numerical scheme. Different projection techniques exist, e.g., by solving certain constrained minimization problems (see [Ei92] and [Ei93]) or by certain natural (orthogonal) projections (see [AsPe91], [AsPe92a], and Subsection III.1.3). Within this framework see also [GrMä86], [Han90], [Lu91b], [Rei92], [Si92], [Si93], [Al93], and [AlÓl93].

In this setting a generalization of the projection method is often to consider the original DAE and the differentiated constraints as a whole, i.e., as an *overdetermined*

DAE (ODAE) (see [FüLe89], [FüLe91], and [PePo92]) of (supposed) index 1 or an *ODE with invariants* (depending on the number of differentiations carried out) (see also [Ge86] and [Sh86]). A numerical method can be applied to a certain formulation (an ODE formulation, an index 0 or 1 formulation, or an high index formulation), and the numerical solution can then be projected back to all underlying constraints. For certain problems and certain numerical methods these projections are in fact quite natural (see [AsPe91], [AsPe92a], and Subsection III.1.3) and may even enter in the definition of the numerical scheme (see Chapter V).

In this context, several authors (see [FüLe89] and [FüLe91]) have also proposed to apply directly a numerical scheme to the ODAE system leading to an overdetermined system of nonlinear equations which generally does not possess a solution. Nevertheless a "pseudo-solution" can be generally defined by application of Gauss-Newton iterations to the overdetermined nonlinear system. In fact this approach can sometimes be equivalently regarded as numerically solving a state-space form (see [EiFüLeRei90], [EiFüYe92], and [PePo92]) or an extended DAE system possessing the same solutions but where additional variables have been introduced (see the next paragraph).

A popular approach for reducing the index of some particular DAE's, especially for constrained mechanical systems, is to introduce new variables in an ODAE system of lower index obtained by differentiation of some constraints such that they do not alter the exact solutions and that a DAE of lower index is obtained. This approach is sometimes referred as a *constraint stabilization* technique. For example the DAE system $(4.3a,b,c)$ is of index 3, whereas the ODAE $(4.3a,b,c,d)$ obtained by one differentiation of the holonomic constraints $(4.3c)$ is of index 2. Now if we introduce additional variables $\mu=(\mu_1,\ldots,\mu_m)^T$ in $(4.3a)$ as follows

$$\dot{q} = v - G^T(q)\mu \ , \tag{5.1a}$$

$$M(q)\dot{v} = f(q,v) - G^T(q)\lambda \ , \tag{5.1b}$$

$$0 = g(q) \ , \tag{5.1c}$$

$$0 = G(q)v \tag{5.1d}$$

then under the assumptions of Lemma 4.1 we obtain an index 2 DAE. This is the famous *GGL formulation* [GeGupLe85]. It can be shown that any solution of the GGL formulation satisfies $\mu=0$. This approach can be generalized to obtain an index 1 DAE system by taking into account the acceleration level constraints $(4.3e)$ (see [FüLe89] and [FüLe91]). In certain cases the numerical solution of a given method can be interpreted as the "pseudo-solution" of the same numerical method applied to an ODAE, hence both approaches may be sometimes regarded as equivalent (see [FüLe89] and [FüLe91]).

5.2. State-space forms.

The basic idea of the *state-space form* is the reduction by a certain parametrization of the DAE to an ODE in a minimal set of independent variables. A precise definition close to the usual definitions is as follows:

Definition 5.1. [PePo92], [Po93a,c]. Let $I\subset\mathbb{R}$ be an open interval and $\Sigma\subset I\times\mathbb{R}^n$ be a $(r+1)$-dimensional manifold. Then the DAE (2.1) has a *local state-space form* at $(x_c,y_c)\in\Sigma$ if there exist an open neighbouring interval $J\subset I$ of x_c, an open neighbourhood $\Gamma\subset\mathbb{R}^n$ of y_c, an open neighbourhood $\Lambda\subset\mathbb{R}^r$ of 0, a diffeomorphism Θ of the form

$$\Theta : \begin{cases} J\times\Lambda \to \Pi := (J\times\Gamma)\cap\Sigma \\ (x,\lambda) \mapsto (x, y=\Phi(x,\lambda)) \ , \end{cases} \tag{5.2}$$

and a Lipschitz continuous mapping

$$\omega : J \times \Lambda \to \mathbb{R}^r \tag{5.3}$$

such that:

1. If $(x_0, y_0) \in \Pi$ and if $\lambda(x)$ is the solution of the IVP

$$\lambda' = \omega(x, \lambda), \quad \lambda(x_0) = \lambda_0 \quad \text{with } (x_0, \lambda_0) = \Theta^{-1}(x_0, y_0) \tag{5.4}$$

then $y(x) := \Phi(x, \lambda(x))$ is a solution of (2.1) satisfying $y(x_0) = y_0$.

2. If $y(x)$ is any solution of (2.1) then for any $x_0 \in J$ there exists $\lambda_0 \in \Lambda$ such that $y(x) = \Phi(x, \lambda(x))$ with $\lambda(x)$ the solution of the IVP (5.4).

The pair (Φ, ω) is called a *local state-space form* of (2.1) at (x_c, y_c). If $J = I$ and $\Pi = \Sigma$ then we have a *global state-space form.*

It is clear that if (2.1) satisfies these above assumptions then it is locally geometrically solvable in the sense of Definition 3.1. For more details we refer to [Po93a,c] and [RaRh94]. There are many different ways to obtain a local state-space form. In general a local state-space form is obtained implicitly, hence a numerical method is applied in an indirect way. In certain cases the solution of a given numerical method based on such constructions can be interpreted as the "pseudo-solution" of the same numerical method applied to an ODAE system (see [Po93a]).

As a first example (see [EiFüLeRei90]) consider the ODE

$$y' = f(x, y) \tag{5.5}$$

with invariants

$$0 = h(x, y) \,. \tag{5.6}$$

Let (x_c, y_c) be fixed and choose $V(x_c, y_c)$ such that

$$\begin{pmatrix} V(x_c, y_c) \\ h_y(x_c, y_c) \end{pmatrix} \tag{5.7}$$

is a square regular matrix. Then by the implicit function theorem we may find an open neighbouring interval J of x_c, an open neighbourhood $\Gamma \subset \mathbb{R}^n$ of y_c, and an open neighbourhood $\Lambda \subset \mathbb{R}^r$ of 0 such that for $x \in J$ fixed and for all $\lambda \in \Lambda$ the system

$$0 = \begin{pmatrix} \lambda - V(x_c, y_c)(y - y_c) \\ h(x, y) \end{pmatrix} \tag{5.8}$$

possesses a unique solution $y = \Phi(x, \lambda) \in \Gamma$. Therefore the inverse mapping Θ (5.2) is well-defined and with the ODE

$$\lambda' = V(x_c, y_c) f(\Theta(x, \lambda)) \tag{5.9}$$

a local state-space form is obtained at $(x_c, y_c) \in J \times \Gamma$. Numerical methods based on such constructions are called *derivative projection methods.* If $V(x_c, y_c)$ spans the null space of $h_y(x_c, y_c)$ then the parametrization is called a *tangent space parametrization.*

Another class of local state-space forms has been proposed for the equations of motion of constrained mechanical systems (4.3). They are based upon two matrices A_1 and A_2 such that

$$\begin{pmatrix} G(q_c) \\ A_1^T \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} G(q_c) \\ A_2^T \end{pmatrix} \tag{5.10}$$

are square regular matrices with q_c satisfying $g(q_c) = 0$. If $A_1^T = A_2^T$ spans the null space of $G(q_c)$ then this choice corresponds to a *tangent space parametrization.* The choice corresponding to a *generalized coordinate partitioning* consists in selecting a permutation $(e_{i_1}, \ldots, e_{i_n})$ of the standard basis of $\mathbb{R}^n$ such that the matrix formed by columns $(i_1, \ldots, i_m)$ of $G(q_c)$ is regular, and in taking $A_1 = A_2 := (e_{i_{m+1}}, \ldots, e_{i_n})$. For more details we refer to [EiFüLeRei90], [PoRh90], [PoRh91], [EiFüYe92], [YeHauPo92], [Po93a,b,c], [RaRh93], and [Ye93].

For higher index linear Hessenberg DAE's of the form

$$\begin{aligned} y^{(m)} &= \sum_{k=0}^{m-1} A_k(x) y^{(k)} + B(x) z + q(x) \ , \\ 0 &= C(x) y + r(x) \end{aligned} \tag{5.11}$$

where the matrices $A_k(x)$, $B(x)$, and $C(x)$ depend smoothly of x, and $C(x)B(x)$ is regular for each x, a special class of state-space forms has been constructed in [AsPe93] allowing a stability analysis of (5.11).

5.3. Regularization techniques.

Another approach for solving DAE's is that of *regularization* (see [Han90], [EiHan91], [Han91], and [AsLi93]). This technique consists in the introduction of small parameters in the equations so that an ODE is obtained and that the initial DAE is obtained when they vanish. In this context the *penalty technique* of Lötstedt [Lö79] and the approach of Knorrenschild [Kno88] can be regarded as regularization techniques. Standard ODE's solvers can then be applied for different small values of these parameters and the solution of the original DAE can then be extrapolated from the different obtained values. However, this approach has not proved yet to be very efficient in practice. There are many difficulties which may arise. For example the small parameters generally introduce stiffness in the equations leading to great numerical difficulties.

A somewhat similar approach referred as a *stabilization* technique is as follows: instead of transforming the DAE into a regularized ODE, new parameters are introduced into a linear combination of the original and differentiated algebraic constraints resulting in an (supposed) index 1 DAE. For constrained mechanical systems this is the well-known *Baumgarte's technique* (see [Bau72]). These parameters are usually adjusted such that a certain constraint manifold is locally attracting (see also [AsChRei94]). However, a general strategy for the determination of these parameters is not possible (see [AsChPeRei93]). A drawback is that the resulting system generally introduce artificial stiffness depending on the choice of the parameters.

6. Numerical methods for DAE's.

Many different numerical methods have been proposed for solving DAE's, ranging from one-step to multistep methods. There are many different ways to solve a DAE numerically. Generally a DAE possesses several equivalent formulations. At a first stage a certain formulation of the problem must be chosen depending on certain criteria of stability, of cost, of real-time constraints, etc. Then a numerical method is chosen depending on these criteria and also on its properties, its reliability, its efficiency, etc. In fact these two stages are often imbricated. All different choices have their own advantages and drawbacks. The most frequent approach to solve DAE's is to apply natural extensions of ODE methods like linear multistep methods, one-leg methods, linear implicit methods, Rosenbrock methods, Runge-Kutta methods, and extrapolation methods. Nevertheless, specialized methods or techniques adapted to certain particular problems exist.

Several techniques to reduce a DAE to an ODE have already been discussed in Subsection 5.1. Once an ODE is obtained, a standard ODE solver can be applied. In this setting a general attempt to solve general unstructured higher index solvable DAE's is due to Campbell (see [Cam89], [Cam93a], and [CamMo93a]). His *least-squares completions* method is based on the construction of an ODE by means of a least-squares solution to the derivative array equations (3.1). The approach of local state-space forms has already been discussed in Subsection 5.2. Once a local state-space form is obtained from a DAE, a standard ODE method can be applied, generally in an indirect way. Since a local parametrization in a minimal set of independent variables is used, such methods present the advantage that they preserve all underlying constraints. Within this framework, a constraint preserving version of the least-squares completions method has been proposed without making use of an explicit knowledge of the constraints [CamMo93b]. In this section we do not discuss again the regularization techniques of Subsection 5.3.

We will now turn our interest to numerical methods applied to pure DAE's. As mentioned in Subsection 5.1 this may present several advantages. The *direct approach* consists in embedding the original DAE into a singular perturbed problem (see (4.15)), to apply formally an ODE method, and to consider the limit $\varepsilon \to 0$. This approach works for certain implicit methods and may provide much insight into the numerical solution of stiff and singular perturbed problems (see Subsection 4.2, [HaLu88], [HaLuRo88], [Ro88b], [HaLuRo89b], [BurPe90], and [HaWa91, Chapter VI]). For implicit *Runge-Kutta (RK)* methods directly applied to (2.1) this approach leads to the following general formulation (see [Pe86] and [BrePe89])

$$\begin{aligned} F(x_n + c_i h, Y_{n,i}, Y'_{n,i}) &= 0 \ , \qquad Y_{n,i} = y_n + h \sum_{j=1}^{s} a_{ij} Y'_{n,j} \ , \\ y_{n+1} &= y_n + h \sum_{i=1}^{s} b_i Y'_{n,i} \end{aligned} \tag{6.1}$$

where the a_{ij}, b_i, and c_i are the coefficients of the method. The RK methods satisfying

$$a_{si} = b_i \qquad \text{for} \quad i = 1, \ldots, s \tag{6.2}$$

are called *stiffly accurate* and they are of great interest in the DAE setting. An important class of stiffly accurate RK methods is given by the Radau IIA methods. The code

RADAU5 of Hairer and Wanner is based on the 3-stage Radau IIA method of order 5. It has been developed for stiff ODE's and DAE's of the form

$$My' = f(x,y) \tag{6.3}$$

where M is a constant square matrix which may be singular in the DAE case (see [HaLuRo89a, Chapter 10] and [HaWa91, Section IV.8]). For implicit *linear multistep methods (LMM's)* directly applied to (2.1) a possible formulation is given by

$$F(x_n+h, y_{n+1}, y'_{n+1}) = 0\ , \qquad \sum_{i=0}^{k} \alpha_i y_{n+1+i-k} = h \sum_{i=0}^{k} \beta_i y'_{n+1+i-k} \tag{6.4}$$

where the α_i and β_i are the coefficients of the method. The popular *backward differentiation formulas (BDF)* are of great interest in the DAE framework. These methods satisfy

$$\beta_i = 0 \qquad \text{for} \quad i = 0,\ldots,k-1 \tag{6.5}$$

and are stable provided $k \leq 6$. Such methods form the basis of the well-known code DASSL of Petzold (see [Pe83] and [BreCamPe89, Chapter 5]) written principally for general DAE's (2.1) of index 0 and 1. For implicit *one-leg methods (OLM's)* directly applied to (2.1) we have

$$F\Big(\sum_{i=0}^{k} \beta_i x_{n+1+i-k}, \sum_{i=0}^{k} \beta_i y_{n+1+i-k}, \frac{1}{h}\sum_{i=0}^{k} \alpha_i y_{n+1+i-k}\Big) = 0\ , \tag{6.6}$$

where the α_i and β_i are the coefficients of the method.

However, in most practical situations it is advantageous to take into account the special structure of the DAE. This approach is called the *indirect approach* and we review in this paragraph some related *specialized methods*. In certain cases the direct and the indirect approach may coincide, e.g, for stiffly accurate Runge-Kutta methods or BDF schemes applied to semi-explicit index 1 problems in Hessenberg form. For semi-explicit index 2 problems in Hessenberg form the class of *half-explicit* methods has been proposed in [HaLuRo89a], [Hi91], [Bra92], [BraHa93a,b], [Os90], and [Os93]. Such methods are explicit in the differential part and implicit in the algebraic part. For constrained mechanical systems special half-explicit extrapolation methods have been proposed in [Lu91a]. For higher index DAE's, the projection methods discussed in Subsection 5.1 may be classified in the indirect approach. For semi-explicit index 2 problems in Hessenberg form the class of *projected RK methods* of Ascher and Petzold fall into this category (see [AsPe91], [AsPe92a], and [Lu91b]) and also the extrapolation method of Lubich (see [Lu89a]). For semi-explicit index 3 problems in Hessenberg form, the *(projected) partitioned RK methods* which form the subject of this thesis also belong to this category, excepted for the unprojected RK methods analyzed in Chapter IV which are related to the direct approach.

It is not the purpose of this thesis to review in detail all known results concerning the application of numerical methods to DAE's. For classical books on the subject we refer to [BreCamPe89], [GrMä86], [HaLuRo89a], and [HaWa91]. We give below the main references, excepted these four mentioned books, according to certain types of DAE's and to certain classes of methods.

1. *Semi-explicit index* 1 *DAE's in Hessenberg form:* RK: [Ro88b], [Ro89]; Rosenbrock: [Ro88a,b], [RenSte89], [RenRoSte89]; extrapolation: [DeHaZu87]; BDF: [LöPe86]; partitioned multistep methods: [Aré93].
2. *Semi-explicit index* 2 *DAE's in Hessenberg form:* RK: [Ro88b], [BrePe89], [Jay93a]; projected RK: [AsPe91], [AsPe92a]; half-explicit methods: [Hi91], [Bra92], [BraHa93a,b]; extrapolation: [Lu89a]; BDF: [Bre83], [LöPe86], [BreEn88], [As89]; partitioned multistep methods: [Aré93].
3. *Semi-explicit index* 3 *DAE's in Hessenberg form:* RK: [Ro88b], [Jay92]; collocation: [Jay93b]; half-explicit and extrapolation methods: [Os90], [Os93]; BDF: [Bre83], [LöPe86], [BreEn88], [Aré93]; generalized BDF: [KeGe91].
4. *Fully implicit index* 1 *DAE's:* RK: [Pe86], [BurPe90]; BDF: [GeGuLe85].
5. *Linear implicit DAE's:* Rosenbrock: [LuRo90]; extrapolation: [Lu89b];
6. *Semi-explicit index* 2 *DAE's:* BDF: [GeGuLe85].
7. *Linear constant DAE's:* RK: [BrePe89]; BDF: [GePe84].
8. *Constrained mechanical systems:* short overviews: [SiFüRen91], [Ka93]; state-space forms: [PoRh90], [PoRh91], [RaRh93], [YeHauPo92], [Ye93]; projection methods: [EiFüLeRei90], [Si92], [Al93], [AlÓl93]; ODAE methods: [FüLe89], [FüLe91], [Ei92], [PePo92]; RK: [Po93b,c]; extrapolation: [Lu91a]; BDF: [GeGuLe85], [Fü88]; projected BDF: [Ei92], [Ei93]; multistep methods: [Fü88], [EiFüYe92], [Po93a,c], [AnBoEiSc93].
9. *Constrained Hamiltonian systems:* Hamiltonian state-space forms: [LeRei94]; partitioned RK: [Rei93], [SkBiOk93], [Jay94], [LeSk94].
10. *Stiff ODE's:* RK: [Ro88b,c], [HaLu88], [HaLuRo88], [BurPe90]; Rosenbrock: [Ro88b], [HaLuRo89b].

7. Scope of this thesis and summary of convergence results.

As seen in Section 4 many important problems are formulated or lead to autonomous *semi-explicit index* 3 *DAE's in Hessenberg form*, i.e., to

$$y' = f(y,z)\ , \qquad z' = k(y,z,u)\ , \qquad 0 = g(y) \tag{7.1a,b,c}$$

where

$$(g_y f_z k_u)(y,z,u) \quad \text{is invertible} \tag{7.2}$$

in a neighbourhood of the exact solution. Hence their numerical treatment is of special interest. The scope of this thesis is to study the application of *(projected) partitioned Runge-Kutta (PRK) methods* to these systems. They include pure *Runge-Kutta methods* as special cases which in turn include *collocation methods* as well (see [Jay93b]).

The analysis of the direct application of collocation methods to the considered problems is the subject of the article [Jay93b]. The main result of this article is a partial proof of the conjecture of [HaLuRo89a, p. 86] giving sharp convergence bounds for stiffly accurate methods, such as the Radau IIA processes. One aim of this thesis is to show a complete proof of this conjecture for the class of stiffly accurate Runge-Kutta methods (see Chapter IV). This result has an application in the convergence analysis of these methods for stiff mechanical systems (see Subsection 4.2 and Section IV.6). Another purpose of this thesis concerns the numerical treatment of Hamiltonian

systems with holonomic constraints (see Subsection 4.1 and Chapter V). A specific class of partitioned Runge-Kutta methods is proposed and analyzed in detail. These methods are superconvergent and preserve the symplectic structure of the flow and all underlying constraints as well.

The choice of partitioned Runge-Kutta methods is natural since they include pure Runge-Kutta methods as special cases and since they are of special interest in the numerical treatment of Hamiltonian systems with holonomic constraints. The major part of this thesis deals with the indirect approach, with the exception of Chapter IV. The main theme is the convergence analysis of the methods under consideration and the main goal is to obtain optimal orders of convergence. We recall that the order of convergence is equal to ν if the global error is bounded by $Const \cdot h^{\nu}$ uniformly on bounded intervals for sufficiently small stepsizes h. We summarize in Table 7.1 and Table 7.2 below the optimal orders of convergence for some important (projected) (partitioned) Runge-Kutta methods when applied to (7.1). Table 7.2 concerns the important special situation when the function k of (7.1) is linear in u. These results follow from Theorem III.5.1, Theorem IV.6.1, Theorem V.4.6, [HaLuRo89a, Theorem 6.4], and [HaJay93], and they are valid for non-constant stepsizes.

Method	stages	order of convergence y	z	u
Lobatto IIIA-IIIB	$s \geq 2$	$2s-2$	$2s-2$	$2s-2$
Radau IIA	$s \geq 2$	$2s-2$	s	$s-1$
projected Radau IIA	$s \geq 2$	$2s-2$	$2s-2$	$2s-2$
Lobatto IIIC	$s \geq 3$	$2s-4$	$s-1$	$s-2$
projected Lobatto IIIC	$s \geq 3$	$2s-4$	$2s-4$	$2s-4$
Gauss	$s \geq 5$	s	$s-2$	$s-4$
projected Gauss	$s \geq 2$	s	s	s
Radau IA	$s \geq 3$	$s-1$	$s-1$	$s-2$
projected Radau IA	$s \geq 3$	$s-1$	$s-1$	$s-1$

Table 7.1. Orders of convergence for the index 3 problem (7.1)-(7.2).

Method	stages	order of convergence y	z	u
Lobatto IIIA-IIIB	$s \geq 2$	$2s-2$	$2s-2$	$2s-2$
Radau IIA	s	$2s-1$	s	$s-1$
projected Radau IIA	s	$2s-1$	$2s-1$	$2s-1$
Lobatto IIIC	$s \geq 2$	$2s-3$	$s-1$	$s-2$
projected Lobatto IIIC	$s \geq 2$	$2s-3$	$2s-3$	$2s-3$
Gauss	$s \geq 3$	s	$s-2$	$s-4$
projected Gauss	s	s	s	s
Radau IA	$s \geq 2$	$s-1$	$s-1$	$s-2$
projected Radau IA	$s \geq 2$	$s-1$	$s-1$	$s-1$

Table 7.2. Orders of convergence for the index 3 problem (7.1)-(7.2) with k linear in u.

Chapter II. Semi-explicit index 3 DAE's in Hessenberg form.

1. Consistency and index.

In this thesis we consider the following autonomous semi-explicit DAE's in Hessenberg form of size 3

$$y' = f(y,z)\ , \qquad z' = k(y,z,u)\ , \qquad 0 = g(y)\ . \tag{1.1a,b,c}$$

The variable of integration will be denoted by x. From now on we suppose that the functions $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $k : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^m$, and $g : \mathbb{R}^n \to \mathbb{R}^p$ are sufficiently differentiable. Differentiating once the constraint (1.1c) we obtain the following additional constraint

$$0 = g_y(y)y' = g_y(y)f(y,z) = \big(g_y f\big)\,(y,z)\ . \tag{1.1d}$$

A second differentiation gives

$$0 = \big(g_{yy}(f,f) + g_y f_y f + g_y f_z k\big)\,(y,z,u)\ . \tag{1.1e}$$

These additional constraints are called *hidden constraints.*

Definition 1.1. The values $(\widetilde{y},\widetilde{z},\widetilde{u})$ are called *consistent* if they satisfy all constraints (1.1c,d,e).

For an initial value problem related to the system (1.1), we see from the above derivation that the initial values cannot be chosen arbitrarily but have to be consistent. We call *the (exact) solution of* (1.1) a solution passing through consistent initial values (y_0,z_0,u_0) at x_0. In order to shorten the notation we will often denote the exact solution of (1.1) at x by $\Psi(x) := (y(x),z(x),u(x))$ and $\Psi_0 := (y_0,z_0,u_0)$.

A third differentiation of (1.1c) gives (omitting the obvious function arguments)

$$\begin{aligned} 0 =& g_{yyy}(f,f,f) + 3g_{yy}(f_y f,f) + 3g_{yy}(f_z k,f) + g_y f_{yy}(f,f) + 2g_y f_{yz}(f,k) \\ &+ g_y f_y f_y f + g_y f_y f_z k + g_y f_{zz}(k,k) + g_y f_z k_y f + g_y f_z k_z k + g_y f_z k_u u'\ . \end{aligned} \tag{1.1f}$$

From now on we suppose that the matrix

$$(g_y f_z k_u)(y,z,u) \qquad \text{is invertible} \tag{1.2}$$

in a neighbourhood of the exact solution. With this hypothesis a differential equation for u' can be obtained from (1.1f)

$$\begin{aligned} u' =& (-g_y f_z k_u)^{-1}\Big(g_{yyy}(f,f,f) + 3g_{yy}(f_y f,f) + 3g_{yy}(f_z k,f) + g_y f_{yy}(f,f) \\ &+ 2g_y f_{yz}(f,k) + g_y f_y f_y f + g_y f_y f_z k + g_y f_{zz}(k,k) + g_y f_z k_y f + g_y f_z k_z k\Big) \end{aligned} \tag{1.3}$$

and the system (1.1a,b,c) is thus of differential index 3. The ODAE (1.1a,b,c,d,e) is therefore of differential index 1. The equations (1.1a,b)-(1.3) constitute the *standard underlying ODE.* For consistent initial values the solvability of the system (1.1) is obvious (see also [Cam91, Theorem 1]):

Theorem 1.1 *For consistent initial values* (y_0, z_0, u_0) *at* x_0 *there exists a unique solution* $(y(x), z(x), u(x))$ *to* (1.1) *on* $\mathbb{R}$.

Proof. This result follows from the ODE theory since for consistent initial values (y_0, z_0, u_0) at x_0, $(y(x), z(x), u(x))$ is a solution to (1.1) if and only if $(y(x), z(x), u(x))$ is a solution to the standard underlying ODE. □

By (1.2) we also have by the implicit function theorem that in a neighbourhood of any fixed values $(\widetilde{y}, \widetilde{z}, \widetilde{u})$ satisfying (1.2) and (1.1*e*), (1.1*e*) defines an implicit function for u, i.e.,

$$u = G(y, z) \ . \tag{1.4}$$

Our aim is now to analyze the perturbation index of the components y, z, and u of the system (1.1*a*, *b*, *c*). For this sake we consider a solution $(y(x), z(x), u(x))$ of (1.1*a*, *b*, *c*) on a bounded interval I passing through (y_0, z_0, u_0) at x_0. We also consider perturbed functions $(\widehat{y}(x), \widehat{z}(x), \widehat{u}(x))$ sufficiently close to $(y(x), z(x), u(x))$ passing through $(\widehat{y}_0, \widehat{z}_0, \widehat{u}_0)$ at x_0 and satisfying

$$\widehat{y}'(x) = f(\widehat{y}(x), \widehat{z}(x)) + \delta(x) \ , \tag{1.5a}$$

$$\widehat{z}'(x) = k(\widehat{y}(x), \widehat{z}(x), \widehat{u}(x)) + \mu(x) \ , \tag{1.5b}$$

$$0 = g(\widehat{y}(x)) + \theta(x) \ . \tag{1.5c}$$

Differentiating twice (1.5*c*) we obtain successively

$$0 = \left(g_y f\right)(\widehat{y}(x), \widehat{z}(x)) + g_y(\widehat{y}(x))\delta(x) + \theta'(x) \tag{1.5d}$$

and (omitting the obvious function arguments)

$$0 = g_{yy}(f, f) + g_y f_y f + g_y f_z k + 2g_{yy}(f, \delta) + g_y f_y \delta + g_{yy}(\delta, \delta) + g_y f_z \mu + g_y \delta' + \theta'' \ . \tag{1.5e}$$

Since (1.2) holds, by the implicit function theorem we obtain the estimate

$$\|\widehat{u}(x) - u(x)\| \le C_1 \Big(\|\widehat{y}(x) - y(x)\| + \|\widehat{z}(x) - z(x)\| + \|\delta(x)\| + \|\delta'(x)\| + \|\mu(x)\| + \|\theta''(x)\| \Big) \tag{1.6}$$

provided the right-hand side is sufficiently small. We now substract (1.1*a*, *b*) from (1.5*a*, *b*), integrate from x_0 to x, use a Lipschitz condition for the functions f and k, and use the estimate (1.6) for $\widehat{u}(x) - u(x)$. We thus get for $e(x) := \|\widehat{y}(x) - y(x)\| + \|\widehat{z}(x) - z(x)\|$

$$e(x) \le e(x_0) + C_2 \left(\int_{x_0}^{x} e(\tau) d\tau + \int_{x_0}^{x} \Big(\|\delta(\tau)\| + \|\delta'(\tau)\| + \|\mu(\tau)\| + \|\theta''(\tau)\| \Big) d\tau \right) \ . \tag{1.7}$$

We finally apply Gronwall Lemma [HaNøWa93, Exercise I.10.2] to obtain

$$\begin{aligned} &\|\widehat{y}(x) - y(x)\| + \|\widehat{z}(x) - z(x)\| \le \\ &\qquad C_3 \left(\|\widehat{y}_0 - y_0\| + \|\widehat{z}_0 - z_0\| + \int_{x_0}^{x} \Big(\|\delta(\tau)\| + \|\delta'(\tau)\| + \|\mu(\tau)\| + \|\theta''(\tau)\| \Big) d\tau \right) \le \\ &\qquad C_4 \left(\|\widehat{y}_0 - y_0\| + \|\widehat{z}_0 - z_0\| + \sup_{\zeta \in I} \Big(\|\delta(\zeta)\| + \|\delta'(\zeta)\| + \|\mu(\zeta)\| + \|\theta''(\zeta)\| \Big) \right) \ . \end{aligned} \tag{1.8}$$

From these estimates the perturbation index of $(1.1a,b,c)$ is also equal to 3. More refined estimates for the components y and z can in fact be obtained ([ArM.92]) by using the techniques of [ArM.93] and [ArM.StrWe93],

$$\|\widehat{y}(x)-y(x)\|+\|\widehat{z}(x)-z(x)\| \leq \tag{1.9}$$
$$C_5\left(\|\widehat{y}_0-y_0\|+\|\widehat{z}_0-z_0\|+\sup_{\zeta\in I}\left(\|\delta(\zeta)\|+\|\int_{x_0}^{\zeta}P_z(\tau)\mu(\tau)d\tau\|+\|\theta(\zeta)\|+\|\theta'(\zeta)\|+D(\zeta)\right)\right)$$

where

$$D(x)=\left(\|\delta(x)\|+\|\delta'(x)\|+\|\mu(x)\|+\|\theta''(x)\|\right)^2 \tag{1.10}$$

and

$$P_z(x):=(I-k_u(g_yf_zk_u)^{-1}g_yf_z)(y(x),z(x),u(x))\ . \tag{1.11}$$

Moreover, if the function k of (1.1) is linear in u we have

$$D(x)=\left(\|\delta(x)\|+\|\delta'(x)\|+\|\mu(x)\|+\|\theta''(x)\|\right)\cdot\left(\|\delta(x)\|+\|\theta'(x)\|\right)\ . \tag{1.12}$$

If in addition the function k is of the form $k(y,z,u)=k_0(y,z)+k_u(y)u$ and the function f is linear in z then we have $D(x)\equiv 0$. Therefore in this last case the perturbation index of the components y and z of $(1.1a,b,c)$ is equal to 2, hence the perturbation index of these components for the ODAE $(1.1a,b,c,d)$ is equal to 1. The index-3-tractability of $(1.1a,b,c)$ can also be shown under lower smoothness assumptions (see [Mä89]). The perturbation index of $(1.1a,b,c,d,e)$ is equal to 1 because it is easy to show that $(1.1a,b,e)$ is of perturbation index 1 (see [HaLuRo89a, pp. 2-3] and [HaWa91, p. 480]).

In the following sections we will develop a theory based on a "tree model" for the Taylor expansion of the exact solution of (1.1). This theory will also be of great help in the study of the numerical methods considered in this thesis for the solution of (1.1).

2. Derivatives of the exact solution.

The aim is now to compute the derivatives of the exact solution of (1.1) under the assumption (1.2). In this section we omit the obvious function arguments. For the first derivatives we have

$$y'=f, \qquad z'=k\ , \tag{2.1a,b}$$

and rewriting (1.3) we get

$$\begin{aligned}u'=&(-g_yf_zk_u)^{-1}g_{yyy}(f,f,f)+3(-g_yf_zk_u)^{-1}g_{yy}(f_yf,f)+3(-g_yf_zk_u)^{-1}g_{yy}(f_zk,f)\\&+(-g_yf_zk_u)^{-1}g_yf_{yy}(f,f)+2(-g_yf_zk_u)^{-1}g_yf_{yz}(f,k)+(-g_yf_zk_u)^{-1}g_yf_yf_yf\\&+(-g_yf_zk_u)^{-1}g_yf_yf_zk+(-g_yf_zk_u)^{-1}g_yf_{zz}(k,k)+(-g_yf_zk_u)^{-1}g_yf_zk_yf\\&+(-g_yf_zk_u)^{-1}g_yf_zk_zk\ .\end{aligned} \tag{2.1c}$$

For the second derivatives of y and z we have

$$y''=f_yy'+f_zz'=f_yf+f_zk \tag{2.2a}$$

and

$$
\begin{aligned}
z'' &= k_y f + k_z k + k_u u' \qquad (2.2b)\\
&= k_y f + k_z k + k_u(-g_y f_z k_u)^{-1} g_{yyy}(f,f,f) + 3k_u(-g_y f_z k_u)^{-1} g_{yy}(f_y f, f)\\
&\quad + 3k_u(-g_y f_z k_u)^{-1} g_{yy}(f_z k, f) + k_u(-g_y f_z k_u)^{-1} g_y f_{yy}(f,f)\\
&\quad + 2k_u(-g_y f_z k_u)^{-1} g_y f_{yz}(f,k) k_u(-g_y f_z k_u)^{-1} g_y f_y f_y f + k_u(-g_y f_z k_u)^{-1} g_y f_y f_z k\\
&\quad + k_u(-g_y f_z k_u)^{-1} g_y f_{zz}(k,k) k_u(-g_y f_z k_u)^{-1} g_y f_z k_y f + k_u(-g_y f_z k_u)^{-1} g_y f_z k_z k\,.
\end{aligned}
$$

We see that the derivatives of y, z, and u can be written as linear combinations of expressions containing only derivatives of f, g, and k. Such expressions are called *elementary differentials* (see the next section for a precise definition). Concerning u'', let us first compute for a constant vector v

$$
\begin{aligned}
&\frac{d}{dx}(-g_y f_z k_u)^{-1} v \qquad (2.3)\\
&\quad = (-g_y f_z k_u)^{-1}\Big(g_{yy}\big(f_z k_u(-g_y f_z k_u)^{-1} v, f\big) + g_y f_{zy}\big(k_u(-g_y f_z k_u)^{-1} v, f\big)\\
&\qquad\qquad + g_y f_{zz}\big(k_u(-g_y f_z k_u)^{-1} v, k\big) + g_y f_z k_{uy}\big((-g_y f_z k_u)^{-1} v, f\big)\\
&\qquad\qquad + g_y f_z k_{uz}\big((-g_y f_z k_u)^{-1} v, k\big) + g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, u'\big)\Big)\\
&\quad = (-g_y f_z k_u)^{-1} g_{yy}\big(f_z k_u(-g_y f_z k_u)^{-1} v, f\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_{zy}\big(k_u(-g_y f_z k_u)^{-1} v, f\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_{zz}\big(k_u(-g_y f_z k_u)^{-1} v, k\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uy}\big((-g_y f_z k_u)^{-1} v, f\big)\\
&\qquad + (-g_y f_z k_u)^{-1}(-g_y f_z k_u)^{-1} g_y f_z k_{uz}\big((-g_y f_z k_u)^{-1} v, k\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_{yyy}(f,f,f)\big)\\
&\qquad + 3(-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_{yy}(f_y f, f)\big)\\
&\qquad + 3(-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_{yy}(f_z k, f)\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_y f_{yy}(f,f)\big)\\
&\qquad + 2(-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_y f_{yz}(f,k)\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_y f_y f_y f\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_y f_y f_z k\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_y f_{zz}(k,k)\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_y f_z k_y f\big)\\
&\qquad + (-g_y f_z k_u)^{-1} g_y f_z k_{uu}\big((-g_y f_z k_u)^{-1} v, (-g_y f_z k_u)^{-1} g_y f_z k_z k\big)
\end{aligned}
$$

which is a consequence of $(M^{-1})' = -M^{-1}M'M^{-1}$ and the chain rule. We see from this long formula that it becomes really impracticable to express u'' as a linear combination of elementary differentials (approximately five pages would be needed). This remark also holds for the higher derivatives of y, z, and u, excepted for y''' which is given (only for the pleasure of it) by

$$
y''' = f_{yy}(f,f) + 2f_{yz}(f,k) + f_y f_y f + f_y f_z k + f_{zz}(k,k) + f_z k_y f + f_z k_z k \qquad (2.4)
$$

$$
\begin{aligned}
&+f_z k_u(-g_y f_z k_u)^{-1} g_{yyy}(f,f,f)+3f_z k_u(-g_y f_z k_u)^{-1} g_{yy}(f_y f,f)\\
&+3f_z k_u(-g_y f_z k_u)^{-1} g_{yy}(f_z k,f)+f_z k_u(-g_y f_z k_u)^{-1} g_y f_{yy}(f,f)\\
&+2f_z k_u(-g_y f_z k_u)^{-1} g_y f_{yz}(f,k)+f_z k_u(-g_y f_z k_u)^{-1} g_y f_y f_y f\\
&+f_z k_u(-g_y f_z k_u)^{-1} g_y f_y f_z k+f_z k_u(-g_y f_z k_u)^{-1} g_y f_{zz}(k,k)\\
&+f_z k_u(-g_y f_z k_u)^{-1} g_y f_z k_y f+f_z k_u(-g_y f_z k_u)^{-1} g_y f_z k_z k\ .
\end{aligned}
$$

As these expressions quickly become very complicated we introduce in the next section a simplified representation of the elementary differentials in terms of specific trees (with three different kinds of vertices). Such a notation will give more insight into the structure of the elementary differentials and will be very useful when studying the order conditions of a (partitioned) Runge-Kutta method applied to (1.1) (see the next chapters). Trees were first introduced in the context of ordinary differential equations by Butcher (see the famous paper [But63]). The extension of this "tree model" to DAE's of index 1 and 2 has already been made by Roche et al. in [Ro88a], [Ro88b], [Ro89], and [HaLuRo89a], and in a different way by Kværnø in [Kv90].

3. Trees and elementary differentials.

For the elementary differentials we make the following graphical identifications:

a) f is identified with a *meagre vertex* and a qth order partial derivative of f with q upwards leaving branches;

b) k is identified with a *cross* and a qth order partial derivative of k with q upwards leaving branches;

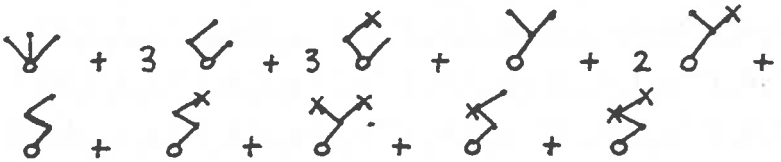
c) the expression $(-g_y f_z k_u)^{-1} g$ is identified with a *fat vertex* and a qth order partial derivative of g therein with q upwards leaving branches;

d) the type of the vertex at the extremity of a branch indicates the variable of derivation: a meagre vertex for y, a cross for z, and a fat vertex for u.

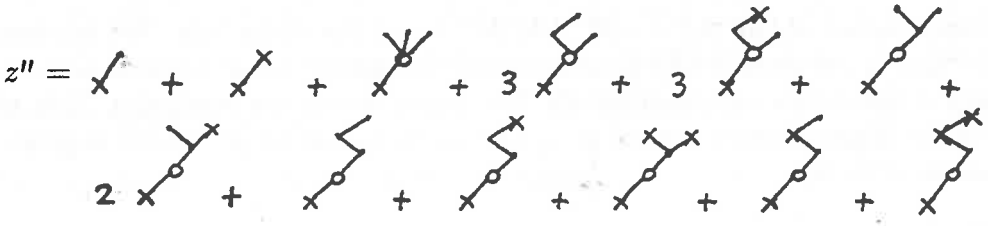
With this representation the corresponding trees for y', z', u', y'', z'', and y''' (see the previous section) are given as follows

$y' =$

$z' =$ 

$u' =$ 

$y'' =$ 

$z'' =$ 

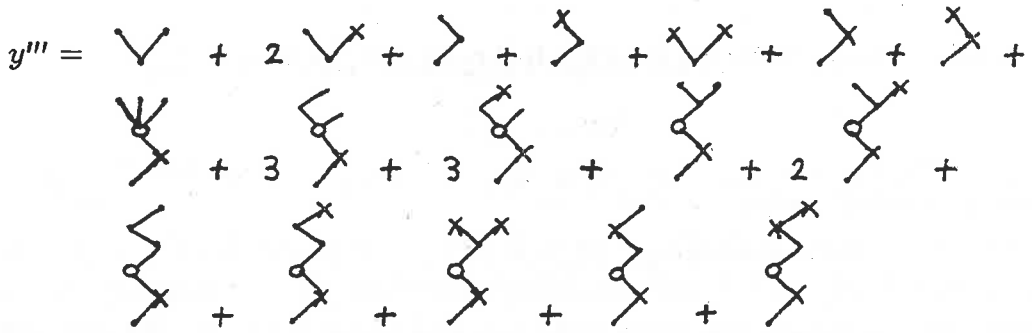


Figure 3.1. Graphical representation of some derivatives of low order.

The lowest vertex of a tree is called its *root*. We can make the following observations:

a) the trees which enter in the derivatives of y have a meagre root;

b) the trees which enter in the derivatives of z have a cross root;

c) the trees which enter in the derivatives of u have a fat root;

d) the vertices over a fat vertex are meagre, because g depends exclusively on y;

e) the vertices under a fat vertex are crosses, because neither f nor g depends on u, only k does;

f) the following trees do not appear in the derivatives of u

Figure 3.2.

Taking the above observations into account, we can now give a recursive definition of the trees corresponding exactly to the expressions appearing in the derivatives of the exact solution of (1.1). We adopt the following conventions:

a) a letter t denotes a tree with a meagre root, the tree consisting of the root only (for $y'=f$) being τ_y;

b) a letter v denotes a tree with a cross root, the tree consisting of the root only (for $z'=k$) being τ_z;

c) a letter u denotes a tree with a fat root.

Definition 3.1. Let $DAT3=DAT3_y \cup DAT3_z \cup DAT3_u$ denote the set of (*differential algebraic index* 3 or $DAT3$-) *trees* defined recursively by

a) $\tau_y \in DAT3_y$, $\tau_z \in DAT3_z$;

b) $[t_1,\ldots,t_m,v_1,\ldots,v_n]_y \in DAT3_y$ if $t_1,\ldots,t_m \in DAT3_y$ and $v_1,\ldots v_n \in DAT3_z$;

c) $[t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p]_z \in DAT3_z$ if $t_1,\ldots,t_m \in DAT3_y$, $v_1,\ldots v_n \in DAT3_z$, and $u_1,\ldots u_p \in DAT3_u$;

d) $[t_1,\ldots,t_m]_u \in DAT3_u$ if $t_1,\ldots,t_m \in DAT3_y$ and if one of the following conditions is satisfied:

- $m \geq 3$;
- $m=2$ and $(t_1,t_2) \neq (\tau_y,\tau_y)$;

- $m=1$ and $t_1 \neq [[u]_z]_y$ with $u \in DAT3_u$, $t_1 \neq \tau_y$, $t_1 \neq [\tau_y]_y$, or $t_1 \neq [\tau_z]_y$.

Remarks 3.1.

1) Here $[t_1,\ldots,t_m,v_1,\ldots,v_n]_y$, $[t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p]_z$, and $[t_1,\ldots,t_m]_u$ represent unordered tuples.
2) From now on we consider $DAT3_y \cup \{\emptyset_y\}$, $DAT3_z \cup \{\emptyset_z\}$, and $DAT3_u \cup \{\emptyset_u\}$ with the *empty trees* $\emptyset_y$, $\emptyset_z$, and $\emptyset_u$ corresponding respectively to the maps id_y, id_z, and id_u defined as $id_y(y,z,u)=y$, $id_z(y,z,u)=z$, and $id_u(y,z,u)=u$. We also define the relations $\tau_y := [\emptyset_y]_y$ and $\tau_z := [\emptyset_z]_z$.
3) The trees of $DAT3_y$, $DAT3_z$, and $DAT3_u$ are characterized respectively by a meagre root, a cross root, and a fat root.

The graphical representation of $DAT3$-trees is as follows:

a) τ_y is represented by a meagre vertex;
b) τ_z is represented by a cross;
c) $t=[t_1,\ldots,t_m,v_1,\ldots,v_n]_y$ is obtained by connecting the roots of $t_1,\ldots,t_m,v_1,\ldots,v_n$ by $m+n$ branches to a new meagre vertex that becomes the root of t;
d) $v=[t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p]_z$ is obtained by connecting the roots of $t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p$ by $m+n+p$ branches to a new cross that becomes the root of v;
e) $u=[t_1,\ldots,t_m]_u \in DAT3_u$ is obtained by connecting the roots of $t_1,\ldots,t_m$ by m branches to a new fat vertex that becomes the root of u.

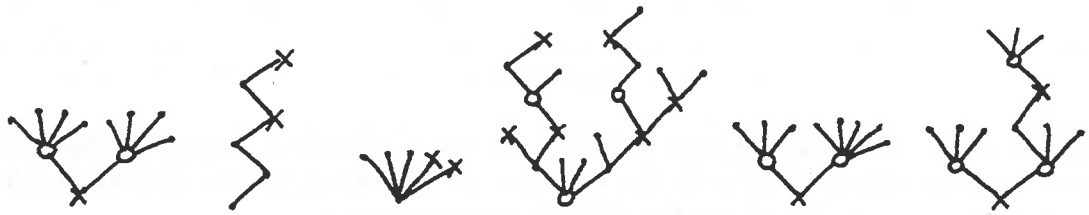


Figure 3.3. Examples of $DAT3$-trees.

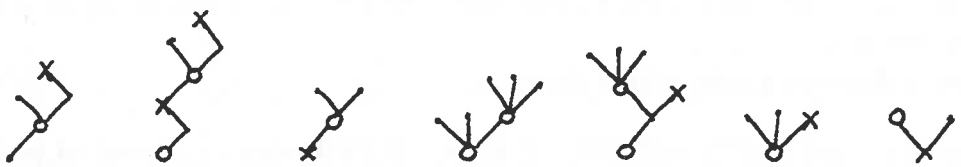


Figure 3.4. Examples of trees which are not $DAT3$-trees.

We are now able to give a precise definition of the elementary differentials which appear in the derivatives of the exact solution of (1.1). By construction they are in one-to-one correspondence with the above-defined $DAT3$-trees.

Definition 3.2. The *elementary differentials* corresponding uniquely to trees in $DAT3 \cup \{\emptyset_y,\emptyset_z,\emptyset_u\}$ are defined recursively as follows:

a) $F(\emptyset_y)=id_y$, $F(\emptyset_z)=id_z$, $F(\emptyset_u)=id_u$, $F(\tau_y)=f$, $F(\tau_z)=k$;

b) $F(t)=\dfrac{\partial^{m+n} f}{\partial y^m \partial z^n}\Big(F(t_1),\ldots,F(t_m),F(v_1),\ldots,F(v_n)\Big)$
if $t=[t_1,\ldots,t_m,v_1,\ldots,v_n]_y \in DAT3_y$;

c) $F(v)=\dfrac{\partial^{m+n+p} k}{\partial y^m \partial z^n \partial u^p}\Big(F(t_1),\ldots,F(t_m),F(v_1),\ldots,F(v_n),F(u_1),\ldots,F(u_p)\Big)$
if $v=[t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p]_z \in DAT3_z$;

d) $F(u)=(-g_y f_z k_u)^{-1}\dfrac{\partial^m g}{\partial y^m}\Big(F(t_1),\ldots,F(t_m)\Big)$ if $u=[t_1,\ldots,t_m]_u \in DAT3_u$.

Remark 3.2. Because of the symmetry of the partial derivatives, each permutation of the subtrees $t_1,\ldots,t_m,v_1,\ldots,v_n$ in a), $t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p$ in b), or $t_1,\ldots,t_m$ in c) leads to the same expression $F(t)$, $F(v)$, or $F(u)$ respectively. Therefore F is well defined.

Examples 3.1.

1. The tree on the left-hand side of Fig. 3.3 corresponds to the following elementary differential
$$k_{uu}\Big((-g_y f_z k_u)^{-1} g_{yyyy}(f,f,f,f),(-g_y f_z k_u)^{-1} g_{yyyy}(f,f,f,f)\Big)\,.$$
2. The tree on the right-hand side of Fig. 3.3 corresponds to the following elementary differential
$$k_{uu}\Big((-g_y f_z k_u)^{-1} g_{yyy}(f,f,f),(-g_y f_z k_u)^{-1} g_{yyy}\big(f_z k_u(-g_y f_z k_u)^{-1} g_{yyy}(f,f,f),f,f\big)\Big)\,.$$

4. Labelled trees and Taylor expansion of the exact solution.

The aim of this section is to derive the Taylor expansion of the exact solution of (1.1). By application of the Leibniz' rule, the chain rule, and (2.3), the differentiation with respect to x of an elementary differential (with the exact solution of (1.1) as arguments) gives rise to a sum of new elementary differentials (see Section 2). In the preceding section, we have established for each elementary differential a one-to-one correspondence with a $DAT3$-tree. In order to know how many times a tree (or, more precisely its corresponding elementary differential) appears in the derivatives of the exact solution of (1.1) we also introduce the concept of *monotonic labelling*. We consider a first completely ordered infinite set of indices $I=\{i<j<k<\ldots\}$ and a second completely ordered set of three subindices $J=\{A<B<C\}$. To each vertex of a tree we associate as described below an index of the set I and in certain cases another subindex of the set J. The first trees which enter in y', z', and u' are labelled in the following way:

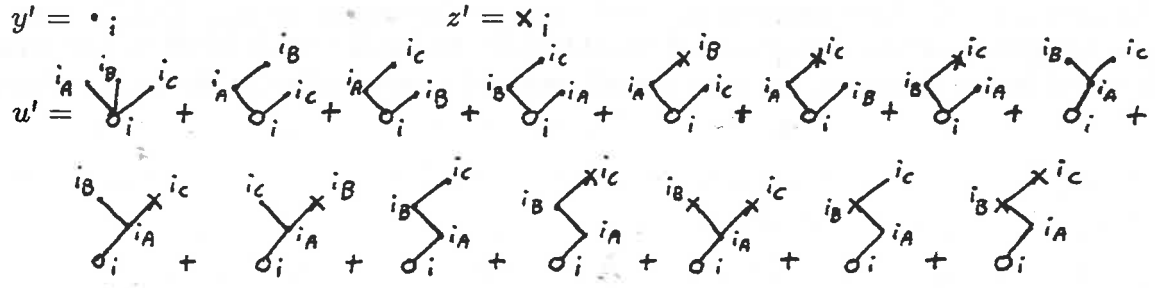


Figure 4.1. Labelled trees of the derivatives of order 1.

Starting from the above labelled trees, the differentiation process can now be easily interpreted by repeated application of the following basic operations on trees:

Differentiation process.

a) a branch with a meagre vertex is linked to a meagre vertex, a cross, or a fat vertex (derivative of f, k, or g with respect to y and addition of the factor $y' = f$); the first index of I which is not present in the original tree is associated to the new meagre vertex;

b) a branch with a cross is linked to a meagre vertex or a cross (derivative of f or k with respect to z and addition of the factor $z' = k$); the first index of I which is not present in the original tree is associated to the new cross;

c) a branch with one of the 15 trees (counted with multiplicity) which enter in u' (see Fig. 3.1) is linked to a cross (derivative of k with respect to u and addition of the factors which enter in u'); the first index of I which is not present in the original tree is associated to all new four vertices, and the three sub-indices A, B, and C of J are associated in a distinct monotonic way to the new non-fat vertices as for u' in Fig. 4.1 (all 15 different labelled trees which enter in u' are grafted);

d) a fat vertex is splitted into two new fat vertices (one above the other) and they are linked via three successive branches containing one new meagre vertex linked over one cross (the new meagre vertex is attached over the lowest new fat vertex and the cross is attached under the upper new fat vertex); the index of the original fat vertex which has been splitted is associated to these new four vertices; the rules a), b), and c) are then applied to the lowest new fat vertex, to the new meagre vertex, and to the new cross (this whole procedure corresponds to the derivative of $(-g_y f_z k_u)^{-1}$ and follows from (2.3)).

These rules describe exactly how to continue the differentiation process of Section 2. All elementary differentials of Definition 3.2 (more precisely all corresponding trees of Definition 3.1) (counted with multiplicity) are therefore obtained in this way. We remark that the above-defined labelling of a $DAT3$-tree is obviously non-decreasing from the root upwards along each branch. The index i is always associated to the root. The labels of I indicate the order of generation of the vertices. The labels of J are introduced to obtain the exact multiplicity of each elementary differential entering in the derivatives of the exact solution of (1.1).

Definition 4.1. A tree $w \in DAT3$ together with a monotonic labelling (obtained by the above differentiation process), is called a *(monotonically labelled) m.l. (differential algebraic index 3) $DAT3$-tree.* The sets of m.l. $DAT3$-trees coming from trees in $DAT3$, $DAT3_y$, $DAT3_z$, and $DAT3_u$ are denoted respectively by $LDAT3$, $LDAT3_y$, $LDAT3_z$, and $LDAT3_u$.

Remark 4.1. We have avoided an overly abstract definition of m.l. $DAT3$-trees. The generalization of the definitions of [HaNøWa93, Sections II.2 & II.15] is not straightforward in our situation, because not all trees with three different kinds of vertices are

$DAT3$-trees.

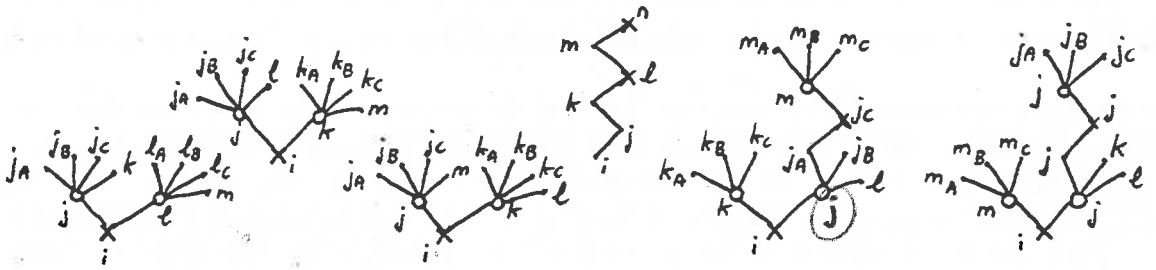


Figure 4.2. Examples of monotonically labelled trees.

Example 4.1. The application of the differentiation process leading to the m.l. tree on the right-hand side of Fig. 4.2 is given as follows

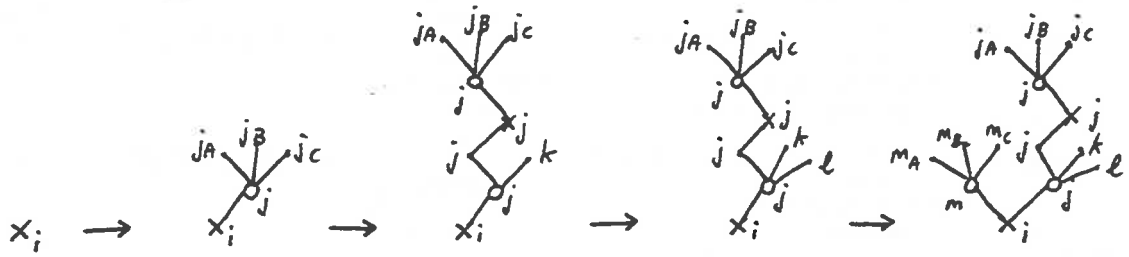


Figure 4.3.

Definition 4.2. For a tree $w \in DAT3$, the integer coefficient $\alpha(w)$ indicates the number of monotonic labellings (obtained by the above differentiation process) and we define $\alpha(\emptyset_y) := 1$, $\alpha(\emptyset_z) := 1$, and $\alpha(\emptyset_u) := 1$.

Equivalently, for $w \in DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\}$, following the differentiation process, we clearly have that $\alpha(w)$ is equal to the number of times that its corresponding elementary differential $F(w)$ appears in the Taylor expansion of the exact solution of (1.1). For a practical computation of these coefficients see Theorem 4.3 below.

In order to characterize the trees which appear in the qth derivative of the exact solution (1.1) we introduce the following definition:

Definition 4.3. The *order* of a tree $w \in (L)DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\}$, denoted by $\varrho(w)$, is the number of meagre vertices plus the number of crosses minus twice the number of fat vertices. A recursive definition is:

a) $\varrho(\emptyset_y) = 0, \quad \varrho(\emptyset_z) = 0, \quad \varrho(\emptyset_u) = 0, \quad \varrho(\tau_y) = 1, \quad \varrho(\tau_z) = 1;$

b) $\varrho(t) = 1 + \varrho(t_1) + \ldots + \varrho(t_m) + \varrho(v_1) + \ldots + \varrho(v_n)$
if $t = [t_1, \ldots, t_m, v_1, \ldots, v_n]_y \in (L)DAT3_y$;

c) $\varrho(v) = 1 + \varrho(t_1) + \ldots + \varrho(t_m) + \varrho(v_1) + \ldots + \varrho(v_n) + \varrho(u_1) + \ldots + \varrho(u_p)$
if $v = [t_1, \ldots, t_m, v_1, \ldots, v_n, u_1, \ldots, u_p]_z \in (L)DAT3_z$;

d) $\varrho(u) = -2 + \varrho(t_1) + \ldots + \varrho(t_m)$ if $u = [t_1, \ldots, t_m]_u \in (L)DAT3_u$.

Examples 4.2.

1. The order of the tree on the left-hand side of Fig 3.3 (or Fig.4.2) is equal to 5.
2. The order of the tree on the right-hand side of Fig 3.3 (or Fig.4.2) is equal to 5.

We see that the trees which enter in the first derivatives of the exact solution are of order 1. We also observe that the new trees generated by the differentiation process and originating from a certain tree are one order higher than this original tree. Therefore every tree which appears in the qth derivative of the exact solution is characterized by an order q. Since the differentiation process of trees generates by definition all elements of $LDAT3$ and each of them exactly once, we have the following result:

Theorem 4.1. *The qth derivatives at x of the exact solution of* (1.1) *denoted by* $\Psi(x)=(y(x),z(x),u(x))$ *are given by*

$$y^{(q)}(x)=\sum_{\substack{t\in LDAT3_y\cup\{\emptyset_y\}\\ \varrho(t)=q}} F(t)(\Psi(x))=\sum_{\substack{t\in DAT3_y\cup\{\emptyset_y\}\\ \varrho(t)=q}} \alpha(t)F(t)(\Psi(x))\ , \tag{4.1a}$$

$$z^{(q)}(x)=\sum_{\substack{v\in LDAT3_z\cup\{\emptyset_z\}\\ \varrho(v)=q}} F(v)(\Psi(x))=\sum_{\substack{v\in DAT3_z\cup\{\emptyset_z\}\\ \varrho(v)=q}} \alpha(v)F(v)(\Psi(x))\ , \tag{4.1b}$$

$$u^{(q)}(x)=\sum_{\substack{u\in LDAT3_u\cup\{\emptyset_u\}\\ \varrho(u)=q}} F(u)(\Psi(x))=\sum_{\substack{u\in DAT3_u\cup\{\emptyset_u\}\\ \varrho(u)=q}} \alpha(u)F(u)(\Psi(x))\ . \tag{4.1c}$$

□

Corollary 4.2. *The Taylor expansions at x of the exact solution of* (1.1) *are given by*

$$y(x+h)=\sum_{t\in LDAT3_y\cup\{\emptyset_y\}} \frac{h^{\varrho(t)}}{\varrho(t)!}F(t)(\Psi(x))=\sum_{t\in DAT3_y\cup\{\emptyset_y\}} \alpha(t)\frac{h^{\varrho(t)}}{\varrho(t)!}F(t)(\Psi(x))\ , \tag{4.2a}$$

$$z(x+h)=\sum_{v\in LDAT3_z\cup\{\emptyset_z\}} \frac{h^{\varrho(v)}}{\varrho(v)!}F(v)(\Psi(x))=\sum_{v\in DAT3_z\cup\{\emptyset_z\}} \alpha(v)\frac{h^{\varrho(v)}}{\varrho(v)!}F(v)(\Psi(x))\ , \tag{4.2b}$$

$$u(x+h)=\sum_{u\in LDAT3_u\cup\{\emptyset_u\}} \frac{h^{\varrho(u)}}{\varrho(u)!}F(u)(\Psi(x))=\sum_{u\in DAT3_u\cup\{\emptyset_u\}} \alpha(u)\frac{h^{\varrho(u)}}{\varrho(u)!}F(u)(\Psi(x))\ . \tag{4.2c}$$

□

The integer coefficients $\alpha(t)$, $\alpha(v)$, and $\alpha(u)$ can be computed recursively as follows:

Theorem 4.3.

a) For the simplest trees we have

$$\alpha(\emptyset_y)=1,\qquad \alpha(\emptyset_z)=1,\qquad \alpha(\emptyset_u)=1,\qquad \alpha(\tau_y)=1,\qquad \alpha(\tau_z)=1\ . \tag{4.3a}$$

b) For a tree $t\in DAT3_y$ of the form

$$t=[\underbrace{t_1,\ldots,t_1}_{m_1},\ldots,\underbrace{t_\mu,\ldots,t_\mu}_{m_\mu},\underbrace{v_1,\ldots,v_1}_{n_1},\ldots,\underbrace{v_\nu,\ldots,v_\nu}_{n_\nu}]_y$$

with μ distinct $t_i \in DAT3_y$ and ν distinct $v_j \in DAT3_z$ we have

$$\alpha(t) = (\varrho(t)-1)! \prod_{i=1}^{\mu} \frac{1}{m_i!} \left(\frac{\alpha(t_i)}{\varrho(t_i)!} \right)^{m_i} \prod_{j=1}^{\nu} \frac{1}{n_j!} \left(\frac{\alpha(v_j)}{\varrho(v_j)!} \right)^{n_j} . \tag{4.3b}$$

c) For a tree $v \in DAT3_z$ of the form

$$v = [\underbrace{t_1, \ldots, t_1}_{m_1}, \ldots, \underbrace{t_\mu, \ldots, t_\mu}_{m_\mu}, \underbrace{v_1, \ldots, v_1}_{n_1}, \ldots, \underbrace{v_\nu, \ldots, v_\nu}_{n_\nu}, \underbrace{u_1, \ldots, u_1}_{p_1}, \ldots, \underbrace{u_\pi, \ldots, u_\pi}_{p_\pi}]_z$$

with μ distinct $t_i \in DAT3_y$, ν distinct $v_j \in DAT3_z$, and π distinct $u_k \in DAT3_u$ we have

$$\alpha(v) = (\varrho(v)-1)! \prod_{i=1}^{\mu} \frac{1}{m_i!} \left(\frac{\alpha(t_i)}{\varrho(t_i)!} \right)^{m_i} \prod_{j=1}^{\nu} \frac{1}{n_j!} \left(\frac{\alpha(v_j)}{\varrho(v_j)!} \right)^{n_j} \prod_{k=1}^{\pi} \frac{1}{p_k!} \left(\frac{\alpha(u_k)}{\varrho(u_k)!} \right)^{p_k} . \tag{4.3c}$$

d) For a tree $u \in DAT3_u$ of the form

$$u = [\underbrace{t_1, \ldots, t_1}_{m_1}, \ldots, \underbrace{t_\mu, \ldots, t_\mu}_{m_\mu}]_u$$

with μ distinct $t_i \in DAT3_y$ we have

$$\alpha(u) = (\varrho(u)+2)! \prod_{i=1}^{\mu} \frac{1}{m_i!} \left(\frac{\alpha(t_i)}{\varrho(t_i)!} \right)^{m_i} . \tag{4.3d}$$

Proof. The proof is simply a generalization to semi-explicit index 3 DAE's in Hessenberg form of the results of [But63, Theorem 3] and [HaWa73, Propositions 1& 2]. For semi-explicit index 1 and 2 DAE's in Hessenberg form, the extension has already been made in [Hi93].

Part a) is trivial. It is sufficient to prove part c) since we have for b)

$$\begin{aligned} \alpha(t) &= \alpha([t_1, \ldots, v_\nu]_y) = \alpha([t_1, \ldots, v_\nu]_z) , \\ \varrho(t) &= \varrho([t_1, \ldots, v_\nu]_y) = \varrho([t_1, \ldots, v_\nu]_z) , \end{aligned} \tag{4.4}$$

and for d)

$$\begin{aligned} \alpha(u) &= \alpha([t_1, \ldots, t_\mu]_u) = \alpha([t_1, \ldots, t_\mu]_z) , \\ \varrho(u) &= \varrho([t_1, \ldots, t_\mu]_u) = \varrho([t_1, \ldots, t_\mu]_z) - 3 . \end{aligned} \tag{4.5}$$

In fact the result c) easily follows from a combinatorial argument (see [HaWa73, Proposition 2])

$$\begin{aligned} \alpha(v) = &\binom{\varrho(v)-1}{\varrho(t_1), \ldots, \varrho(t_\mu), \varrho(v_1), \ldots, \varrho(v_\nu), \varrho(u_1), \ldots, \varrho(u_\pi)} \alpha(t_1) \cdots \alpha(t_\mu) \times \\ &\alpha(v_1) \cdots \alpha(v_\nu) \alpha(u_1) \cdots \alpha(u_\pi) \frac{1}{m_1!} \cdots \frac{1}{m_\mu!} \cdot \frac{1}{n_1!} \cdots \frac{1}{n_\nu!} \cdot \frac{1}{p_1!} \cdots \frac{1}{p_\pi!} . \end{aligned} \tag{4.6}$$

□

Examples 4.3.

1. There are 3 monotonic labellings of the tree on the left-hand side of Fig. 3.3. They are given on the left-hand side of Fig. 4.2.
2. There are 40 monotonic labellings of the tree on the right-hand side of Fig. 3.3. Two of them are given on the right-hand side of Fig. 4.2.

5. DA3-series.

In this section we extend the concept of B-series introduced in the context of ordinary differential equations (see [HaWa74] and [HaNøWa93, Section II.12]) to semi-explicit index 3 DAE's in Hessenberg form. The extension of this theory to DAE's was partially undertaken for semi-explicit index 1 and 2 DAE's in Hessenberg form by Roche et al. in [Ro88a], [Ro88b], [Ro89], and [HaLuRo89a], but without including general results on the composition of such series, as they were in fact not needed for their purposes. The present section is mainly dedicated to the proof of Theorem 5.4 which states that the series under consideration, called *$DA3$-series*, together with an operation of composition related to the natural composition of $DA3$-series form a group. This group structure on $DA3$-series will turn out to be a powerful tool in Section III.4 when we will derive optimal estimates for a certain projection of the local error of the z-component of a (partitioned) Runge-Kutta method (see Theorem III.4.3). For this sake we will need to develop the local error of the z-component at the endpoint of the integration interval and by the use of the composition law of $DA3$-series this expansion will be obtained more easily.

This section is organized in the following way. We first give some basic definitions. We then state an important lemma related to the differentiation of an elementary differential whose argument is a $DA3$-series. The lengthy proof of this lemma requires the introduction of a new type of trees, namely the *composite $LDAT3$-trees*. Next we obtain a central theorem proving that the composition of two $DA3$-series is again a $DA3$-series with coefficients given by a certain composition of the coefficients of the two original $DA3$-series. Hence we finally prove that the $DA3$-series with this composition law form a group. We use the notation of the preceding sections and we illustrate the proofs and results of this section by some examples. The techniques used in this section generalize those used in [HaWa74].

Definition 5.1. Let $\mathbf{a} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ be an arbitrary map. Then for any fixed (non necessarily consistent) $\Psi = (\widetilde{y}, \widetilde{z}, \widetilde{u})$, the series

$$DA3_y(\mathbf{a}, \Psi) := \sum_{t \in DAT3_y \cup \{\emptyset_y\}} \alpha(t) \frac{h^{\varrho(t)}}{\varrho(t)!} \mathbf{a}(t) F(t)(\Psi) \,, \tag{5.1a}$$

$$DA3_z(\mathbf{a}, \Psi) := \sum_{v \in DAT3_z \cup \{\emptyset_z\}} \alpha(v) \frac{h^{\varrho(v)}}{\varrho(v)!} \mathbf{a}(v) F(v)(\Psi) \,, \tag{5.1b}$$

$$DA3_u(\mathbf{a}, \Psi) := \sum_{u \in DAT3_u \cup \{\emptyset_u\}} \alpha(u) \frac{h^{\varrho(u)}}{\varrho(u)!} \mathbf{a}(u) F(u)(\Psi) \,, \tag{5.1c}$$

$$DA3(\mathbf{a}, \Psi) := \big(DA3_y(\mathbf{a}, \Psi), DA3_z(\mathbf{a}, \Psi), DA3_u(\mathbf{a}, \Psi) \big) \tag{5.1d}$$

are called *$DA3_y$-series*, *$DA3_z$-series*, *$DA3_u$-series*, and *$DA3$-series* respectively.

Remark 5.1. Usually one is only interested in truncations of these series. All subsequent results are valid as far that f, g, and k are differentiable.

Examples 5.1.

1. The identity map $DA3(\mathbf{e},\Psi)=\Psi$ is a $DA3$-series with coefficients

$$\mathbf{e}(w)=\begin{cases}1 & \text{if } w\in\{\emptyset_y,\emptyset_z,\emptyset_u\}\ ,\\ 0 & \text{else, i.e., if } w\in DAT3.\end{cases} \tag{5.2}$$

2. The Taylor expansion of the exact solution of (1.1) can be written in the form of a $DA3$-series (see Corollary 4.2) $\Psi(x+\tau h)=DA3(\mathbf{p}_\tau,\Psi(x))$ with coefficients given by

$$\mathbf{p}_\tau(w)=\tau^{\varrho(w)} \quad \text{for } w\in DAT3\cup\{\emptyset_y,\emptyset_z,\emptyset_u\}\ . \tag{5.3a}$$

For $\tau=1$ we define $\mathbf{p}:=\mathbf{p}_1$ and we obtain

$$\mathbf{p}(w)=1 \quad \forall w\in DAT3\cup\{\emptyset_y,\emptyset_z,\emptyset_u\}\ . \tag{5.3b}$$

3. We will see later in Theorem III.3.2 and Theorem III.3.3 that the numerical solution of (1.1) by a (partitioned) Runge-Kutta method can also be written as a $DA3$-series.

In order to distinguish the trees related to the y-, z-, and u-component we introduce the following definition:

Definition 5.2. For a tree $w\in(L)DAT3\cup\{\emptyset_y,\emptyset_z,\emptyset_u\}$, the letter $\sigma(w)$ determines its *type* as follows:

$$\sigma(w)=\begin{cases}y & \text{if } w\in(L)DAT3_y\cup\{\emptyset_y\},\\ z & \text{if } w\in(L)DAT3_z\cup\{\emptyset_z\},\\ u & \text{else, i.e., if } w\in(L)DAT3_u\cup\{\emptyset_u\}.\end{cases} \tag{5.4}$$

Definition 5.3. Let $w\in LDAT3\cup\{\emptyset_y,\emptyset_z,\emptyset_u\}$ be a monotonically labelled tree and $0\le j\le\varrho(w)$ be a fixed integer. Then we denote by $s_j(w)$ the monotonically labelled tree containing only the first j indices of I (with possibly associated subindices from J) in the differentiation process (described in Section 4) which leads to w. We also denote by $d_j(w)$, the *difference set of order* j, the set of labelled subtrees resulting from the removal of the first j indices of I (with possibly associated subindices from J) to w. We define $s_0(w):=\emptyset_{\sigma(w)}$ and for all $j\ge\varrho(w)$ $s_j(w):=w$ and $d_j(w):=\{\emptyset_{\sigma(w)}\}$. For $\emptyset_\sigma\in\{\emptyset_y,\emptyset_z,\emptyset_u\}$ we define for all $j\ge0$ $s_j(\emptyset_\sigma):=\emptyset_\sigma$ and $d_j(\emptyset_\sigma):=\{\emptyset_\sigma\}$.

Remarks 5.2.

1) We trivially have that $d_0(w)=\{w\}$.
2) We easily see see that $d_1(w)=\{w_1,\dots,w_n\}$ if $w\in LDAT3_y\cup LDAT3_z$.

Example 5.2. Consider the m.l. tree, denoted by v, on the right-hand side of Fig. 4.2. The trees $s_j(v)$ for $j=1,\dots,5$ are given in Fig. 4.3 ($s_0(v)=\emptyset_z$) and the difference sets

are as follows

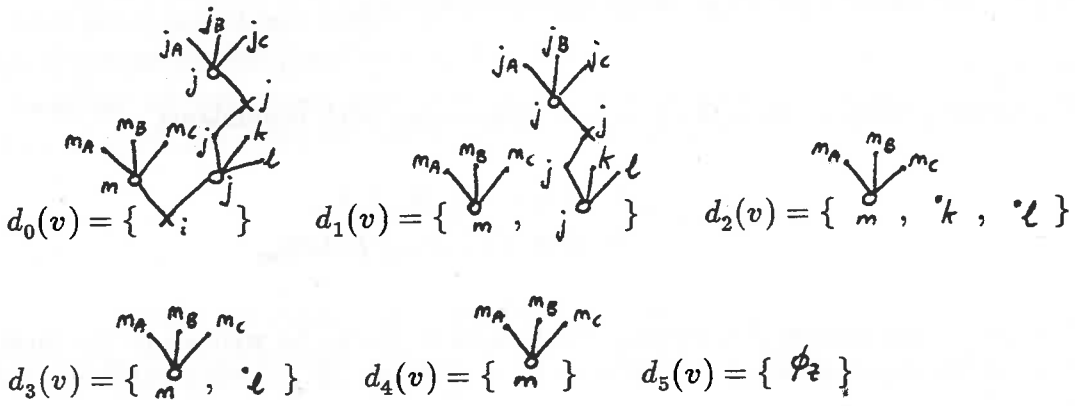


Figure 5.1.

The following lemma is an essential preliminary result which will be used in the proof of Theorem 5.3.

Lemma 5.1. *Let* $\mathbf{a} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ *be an arbitrary map and suppose that* $\mathbf{a}(\emptyset_y)=1$, $\mathbf{a}(\emptyset_z)=1$, *and* $\mathbf{a}(\emptyset_u)=1$. *Then for every tree* $w \in LDAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\}$ *and* $r \geq \varrho(w)$ *we have*

$$\frac{d^{r-\varrho(w)}}{dh^{r-\varrho(w)}}\Big(F(w)(DA3(\mathbf{a},\Psi))\Big)\Big|_{h=0} = \sum_{\substack{s \in LDAT3_{\sigma(w)} \cup \{\emptyset_{\sigma(w)}\} \\ s_{\varrho(w)}(s)=w,\ \varrho(s)=r}} \left(\prod_{\omega \in d_{\varrho(w)}(s)} \mathbf{a}(\omega)\right) F(s)(\Psi) \ . \tag{5.5}$$

Before giving the proof of this lemma we first need some definitions and technical results. Given a function $\Phi(h) = (\Phi_y(h), \Phi_z(h), \Phi_u(h))$ and a certain tree $w \in LDAT3$, our first aim is to differentiate expressions of the type $F(w)(\Phi(h))$ with respect to h. In fact the choice of the labelling of w is unimportant since the elementary differential $F(w)$ is independent of the labelling. However, this choice will allow a more coherent presentation of the derivation below. Hereafter we will sometimes omit the obvious function arguments $\Phi_y(h)$, $\Phi_z(h)$, $\Phi_u(h)$ and a differentiation with respect to h will be symbolized by a prime ($'$).

Consider as a first example the tree $t = [t_1, v_1, v_2]_y$ where $t_1 = \tau_y$, $v_1 = [\tau_y]_z$, and $v_2 = \tau_z$, with the following labelling

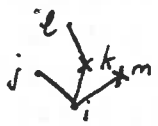


Figure 5.2.

By Definition 3.2 this tree corresponds to

$$F(t)(\Phi(h)) = f_{yzz}(\Phi_y(h), \Phi_z(h))\Big(F(t_1)(\Phi(h)), F(v_1)(\Phi(h)), F(v_2)(\Phi(h))\Big) \ . \tag{5.6}$$

Differentiating this expression once we obtain

$$\frac{d}{dh}F(t)(\Phi(h)) = f_{yyzz}\left(F(t_1), \Phi_y', F(v_1), F(v_2)\right) + f_{yzzz}\left(F(t_1), F(v_1), F(v_2), \Phi_z'\right) +$$

$$f_{yzz}(F(t_1)', F(v_1), F(v_2)) + f_{yzz}(F(t_1), F(v_1)', F(v_2)) + \qquad (5.7)$$
$$f_{yzz}(F(t_1), F(v_1), F(v_2)') \ .$$

The differentiation process can be interpreted by means of an enlarged "tree model" adapted to the present situation. A precise definition will be given below. Each time that a factor is differentiated we graft a double-lined branch with a labelled square vertex to obtain a new monotonically labelled tree. We also distinguish the derivatives of $\Phi_y(h)$, $\Phi_z(h)$, and $\Phi_u(h)$ by the presence respectively of a meagre vertex, a cross, and a fat vertex in a square vertex. Here the expressions entering in (5.7) can be represented graphically by

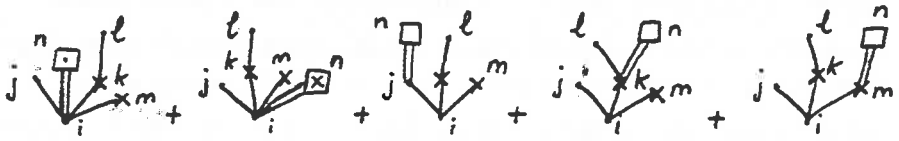
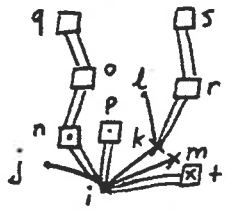
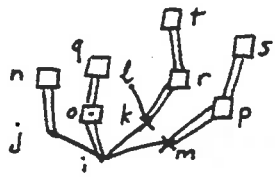
$\frac{d}{dh}F(t)(\Phi(h)) =$ 

Figure 5.3. Graphical representation of (5.7).

Continuing this process we arrive for example to the following trees entering in the 7th derivative of (5.6) (their corresponding expression is mentioned)



$f_{yyyzzz}(F(t_1),\Phi_y''',\Phi_y',F(v_1)'',F(v_2),\Phi_z')$



$f_{yyzz}(F(t_1)',\Phi_y'',F(v_1)'',F(v_2)'')$

Figure 5.4.

To include more general cases, we can give the following precise definition:

Definition 5.4. Starting from a fixed labelled tree $t \in LDAT3_y$, the set of *composite LDAT3-trees* of t, denoted by $CLDAT3(t)$, can be constructed by repeated application of the following rules:

a) a *square vertex* containing a *meagre vertex* is linked via a *double-lined branch* to the meagre root;

b) a *square vertex* containing a *cross* is linked via a *double-lined branch* to the meagre root;

c) a *square vertex* is linked via a *double-lined branch* to a neighbouring non-square vertex of the meagre root if no such connection already exists;

d) a *square vertex* is linked via a *double-lined branch* to a singly-connected square vertex, i.e., to a square vertex with no vertex laying above;

e) after application of the rules a), b), c), or d), the first index of I which is not present in the original tree is associated to the new square vertex.

The corresponding construction for the trees in $LDAT3_z$ is similar:

Definition 5.5. Starting from a fixed labelled tree $v \in LDAT3_z$, the set of *composite*

LDAT3-trees of v, denoted by $CLDAT3(v)$, can be constructed by repeated application of the following rules:

a) a *square vertex* containing a *meagre vertex* is linked via a *double-lined branch* to the cross root;

b) a *square vertex* containing a *cross* is linked via a *double-lined branch* to the cross root;

c) a *square vertex* containing a *fat vertex* is linked via a *double-lined branch* to the cross root;

d) a *square vertex* is linked via a *double-lined branch* to a neighbouring non-square vertex of the cross root if no such connection already exists;

e) a *square vertex* is linked via a *double-lined branch* to a singly-connected square vertex, i.e., to a square vertex with no vertex laying above;

f) after application of the rules a), b), c), d), or e), the first index of I which is not present in the original tree is associated to the new square vertex.

However, for the trees $u \in LDAT3_u$ the situation is slightly more intricate. Consider for example the tree $u = [t_1, t_2]_u$ where $t_1 = \tau_y$ and $t_2 = [\tau_z]_y$, with the following labelling

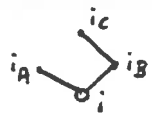


Figure 5.5.

This tree corresponds to

$$F(u)(\Phi(h)) = (-g_y f_z k_u)^{-1}(\Phi(h)) g_{yy}(\Phi_y(h))\Big(F(t_1)(\Phi(h)), F(t_2)(\Phi(h))\Big) . \qquad (5.8)$$

The differentiation of this expression requires the knowledge of the derivatives of the term $(-g_y f_z k_u)^{-1}(\Phi(h))v$ for a constant vector v similarly to (2.3). A first differentiation gives

$$\begin{aligned}\frac{d}{dh}(-g_y f_z k_u)^{-1} v = &(-g_y f_z k_u)^{-1}(g_{yy}(f_z k_u(-g_y f_z k_u)^{-1} v, \Phi_y')) + \\ &(-g_y f_z k_u)^{-1}(g_y f_{zy}(k_u(-g_y f_z k_u)^{-1} v, \Phi_y')) + \\ &(-g_y f_z k_u)^{-1}(g_y f_{zz}(k_u(-g_y f_z k_u)^{-1} v, \Phi_z')) + \\ &(-g_y f_z k_u)^{-1}(g_y f_z k_{uy}((-g_y f_z k_u)^{-1} v, \Phi_y')) + \\ &(-g_y f_z k_u)^{-1}(g_y f_z k_{uz}((-g_y f_z k_u)^{-1} v, \Phi_z')) + \\ &(-g_y f_z k_u)^{-1}(g_y f_z k_{uu}((-g_y f_z k_u)^{-1} v, \Phi_u')) .\end{aligned} \qquad (5.9)$$

Therefore the first derivative of (5.8) is given by

$$\begin{aligned}\frac{d}{dh} F(u)(\Phi(h)) = &\\ &(-g_y f_z k_u)^{-1}(g_{yy}(f_z k_u(-g_y f_z k_u)^{-1} g_{yy}(F(t_1), F(t_2)), \Phi_y')) + \\ &(-g_y f_z k_u)^{-1}(g_y f_{zy}(k_u(-g_y f_z k_u)^{-1} g_{yy}(F(t_1), F(t_2)), \Phi_y')) +\end{aligned} \qquad (5.10)$$

$$
\begin{aligned}
&(-g_y f_z k_u)^{-1}(g_y f_{zz}(k_u(-g_y f_z k_u)^{-1} g_{yy}(F(t_1),F(t_2)),\Phi'_z))+\\
&(-g_y f_z k_u)^{-1}(g_y f_z k_{uy}((-g_y f_z k_u)^{-1} g_{yy}(F(t_1),F(t_2)),\Phi'_y))+\\
&(-g_y f_z k_u)^{-1}(g_y f_z k_{uz}((-g_y f_z k_u)^{-1} g_{yy}(F(t_1),F(t_2)),\Phi'_z))+\\
&(-g_y f_z k_u)^{-1}(g_y f_z k_{uu}((-g_y f_z k_u)^{-1} g_{yy}(F(t_1),F(t_2)),\Phi'_u))+\\
&(-g_y f_z k_u)^{-1} g_{yyy}(F(t_1),F(t_2),\Phi'_y)+\\
&(-g_y f_z k_u)^{-1} g_{yy}(F(t_1)',F(t_2))+\\
&(-g_y f_z k_u)^{-1} g_{yy}(F(t_1),F(t_2)') \ .
\end{aligned}
$$

As previously we adopt a new "tree model". This time we must additionnally take formula (5.9) into account in the representation. For (5.10) we have

$\frac{d}{dh}F(u)(\Phi(h)) =$

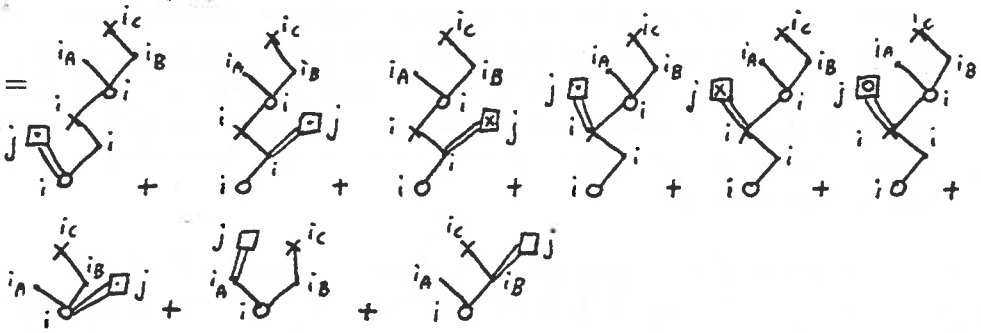
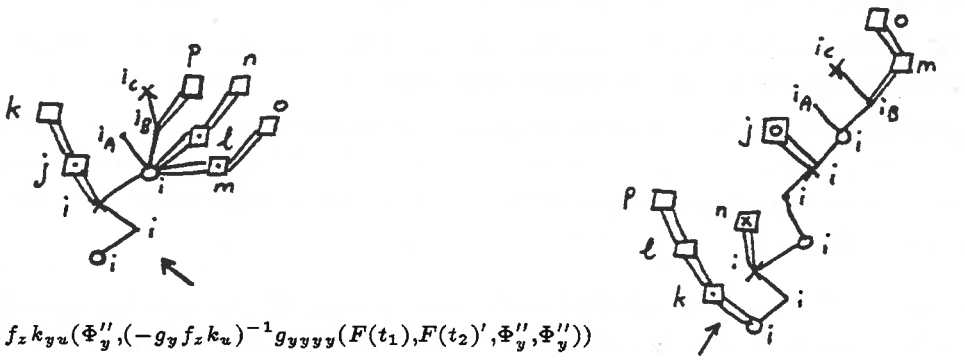


Figure 5.6. Graphical representation of (5.10).

Continuing this process we arrive for example at the following trees which enter in the 7th derivative of (5.8) (their corresponding expression is mentioned)



$(-g_y f_z k_u)^{-1} g_y f_z k_{yu}(\Phi''_y,(-g_y f_z k_u)^{-1} g_{yyyy}(F(t_1),F(t_2)',\Phi''_y,\Phi''_y))$

$(-g_y f_z k_u)^{-1} g_{yy}(\ddot{\Phi}'''_y, f_z k_{yu}(\Phi'_y,(-g_y f_z k_u)^{-1} g_y f_z k_{uu}(\Phi'_u, g_{yy}(F(t_1),F(t_2)''))))$

Figure 5.7.

A refined definition of the above trees is given as follows:

Definition 5.6. Starting from a fixed labelled tree $u \in LDAT3_u$, the set of *composite LDAT3-trees* of u, denoted by $CLDAT3(u)$, can be constructed by the repeated application of the following rules:

a) a *square vertex* containing a *meagre vertex* is linked via a *double-lined branch* to a meagre, cross, or fat vertex having index i (i only, no subindices must be present);

b) a *square vertex* containing a *cross* is linked via a *double-lined branch* to a meagre

or cross vertex having index i (i only, no subindices must be present);

c) a *square vertex* containing a *fat vertex* is linked via a *double-lined branch* to a cross vertex having index i (i only, no subindices must be present);

d) a *square vertex* is linked via a *double-lined branch* to a neighbouring non-square vertex of the upper fat vertex having index i (this vertex can be considered as the original root of the tree u) if no such connection already exists;

e) a *square vertex* is linked via a *double-lined branch* to a singly-connected square vertex;

f) after application of the rules a), b), c), d), or e), the first index of I which is not present in the original tree is associated to the new square vertex;

g) a fat vertex having index i is splitted into two new fat vertices (one above the other) and they are linked via three successive branches containing one new meagre vertex linked over one cross (the new meagre vertex is attached over the lowest new fat vertex and the cross under the upper new fat vertex); the index i is associated to the new four vertices; the rules a), b), and c), together with f) are then applied respectively to the lowest new fat vertex, to the new meagre vertex, and to the new cross (this whole procedure corresponds to the derivative of $(-g_y f_z k_u)^{-1}(\Phi(h))$ and follows from (5.9)).

The identification of the abovedefined composite $LDAT3$-trees with expressions such as those given in Fig. 5.4 and Fig. 5.7 is obvious. For that reason we omit to give a formal definition of this natural correspondence similar to the correspondences developed in Section 3. These trees possess the following features:

a) they are monotonically labelled as the label indicates the order of generation in the differentiation process;

b) the square vertices directly linked to the vertices with label i (i only, no subindices must be present, this concerns only the root for the trees $w \in LDAT3_y \cup LDAT3_z$) contain in addition a meagre vertex, a cross, or a fat vertex;

c) the *distinguished* $DAT3$*-subtrees* constituted exclusively by square vertices and double-lined branches are singly-branched;

d) an arbitrary number of distinguished $DAT3$-subtrees can be connected to the vertices with index i; at most one distinguished $DAT3$-subtree can be connected to the vertices in the neighbouring of the root of the original tree; no such distinguished $DAT3$-subtrees are connected otherwise.

Definition 5.7. For two labelled (sub)trees w and W, we use the notation $w \subset W$ to express the fact that w is a subtree of W having the same root and the same labels.

Example 5.3.

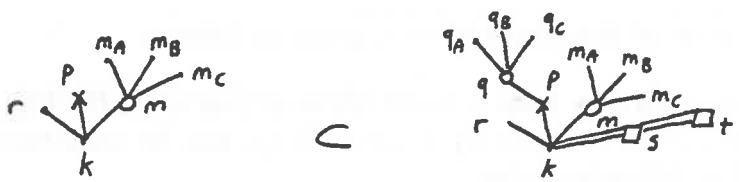


Figure 5.8.

Definition 5.8. For a composite $LDAT3$-tree $\overline{W}$ we denote by $D_1(\overline{W})$, the *difference set of order* 1, the set of composite subtrees resulting from the removing of the index i (i only, no subindices must be present) to $\overline{W}$.

Examples 5.4.

1. The composite subtrees of the difference set of order 1 of the composite tree on the right-hand side of Fig. 5.4 are given by

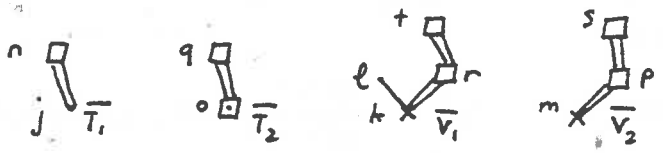


Figure 5.9.

2. The composite subtrees of the difference set of order 1 of the composite tree on the right-hand side of Fig. 5.7 are given by

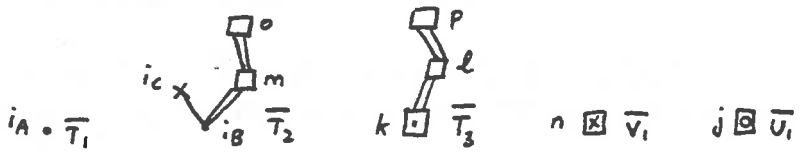


Figure 5.10.

Definition 5.9. The order of a composite (sub)tree $\overline{W}$, denoted by $\varrho(\overline{W})$, is the number of meagre vertices plus the number of crosses plus the number of square vertices minus twice the number of fat vertices. The vertices contained in the square vertices are not counted.

Examples 5.5.

1. The order of the composite trees in Fig. 5.7 is equal to 8.
2. The order of the composite subtree on the right-hand side of Fig. 5.9 is equal to 3.

From all previous definitions we obtain by construction:

Lemma 5.2. *Let* $\Phi(h)=(\Phi_y(h),\Phi_z(h),\Phi_u(h))$ *be given then:*

a) for τ_y *and for all* $r\geq 1$ *we have*

$$\left(F(\tau_y)(\Phi(h))\right)^{(r-1)} = f(\Phi_y(h),\Phi_z(h))^{(r-1)} = \tag{5.11a}$$

$$\sum_{\substack{\overline{T}\in CLDAT3(\tau_y)\\ D_1(\overline{T})=\{\overline{T}_1,\ldots,\overline{T}_M,\overline{V}_1,\ldots,\overline{V}_N\},\ \varrho(\overline{T})=r}} \frac{\partial^{M+N} f}{\partial y^M \partial z^N}\Big(\Phi_y(h)^{(\varrho(\overline{T}_1))},\ldots,\Phi_y(h)^{(\varrho(\overline{T}_M))},\ \Phi_z(h)^{(\varrho(\overline{V}_1))},\ldots,\Phi_z(h)^{(\varrho(\overline{V}_N))}\Big)\ ;$$

b) for τ_z *and for all* $r\geq 1$ *we have*

$$\left(F(\tau_z)(\Phi(h))\right)^{(r-1)} = k(\Phi_y(h),\Phi_z(h),\Phi_u(h))^{(r-1)} = \tag{5.11b}$$

$$\sum_{\substack{\overline{V}\in CLDAT3(\tau_z)\\ D_1(\overline{V})=\{\overline{T}_1,\dots,\overline{T}_M,\overline{V}_1,\dots,\overline{V}_N,\overline{U}_1,\dots,\overline{U}_P\},\ \varrho(\overline{V})=r}} \frac{\partial^{M+N+P}k}{\partial y^M\partial z^N\partial u^P}\Big(\Phi_y(h)^{(\varrho(\overline{T}_1))},\dots,\Phi_y(h)^{(\varrho(\overline{T}_M))},$$
$$\Phi_z(h)^{(\varrho(\overline{V}_1))},\dots,\Phi_z(h)^{(\varrho(\overline{V}_N))},\Phi_u(h)^{(\varrho(\overline{U}_1))},\dots,\Phi_u(h)^{(\varrho(\overline{U}_P))}\Big)\ ;$$

c) for a fixed labelled tree $t\in LDAT3_y$ of the form $t=[t_1,\dots,t_m,v_1,\dots,v_n]_y$ and for all $r\geq\varrho(t)$ we have

$$(F(t)(\Phi(h)))^{(r-\varrho(t))}=\sum_{\substack{\overline{T}\in CLDAT3(t)\\ D_1(\overline{T})=\{\overline{T}_1,\dots,\overline{T}_M,\overline{V}_1,\dots,\overline{V}_N\},\ \varrho(\overline{T})=r,\ M\geq m,\ N\geq n\\ t_1\subset\overline{T}_1,\dots,t_m\subset\overline{T}_m,v_1\subset\overline{V}_1,\dots,v_n\subset\overline{V}_n}} \frac{\partial^{M+N}f}{\partial y^M\partial z^N}\Big((F(t_1)(\Phi(h))^{(\varrho(\overline{T}_1)-\varrho(t_1))},\dots \tag{5.11c}$$
$$\dots,(F(t_M)(\Phi(h))^{(\varrho(\overline{T}_M)-\varrho(t_M))},(F(v_1)(\Phi(h))^{(\varrho(\overline{V}_1)-\varrho(v_1))},\dots$$
$$\dots,(F(v_N)(\Phi(h))^{(\varrho(\overline{V}_N)-\varrho(v_N))}\Big)\ ;$$

where we have defined $t_j:=\emptyset_y$ for $m<j\leq M$ and $v_k:=\emptyset_z$ for $n<k\leq N$ (therefore we have $(F(t_j)(\Phi(h))^{(\varrho(\overline{T}_j)-\varrho(t_j))}=\Phi_y(h)^{(\varrho(\overline{T}_j))}$, $(F(v_k)(\Phi(h))^{(\varrho(\overline{V}_k)-\varrho(v_k))}=\Phi_z(h)^{(\varrho(\overline{V}_k))}$ for these trees).

d) for a fixed labelled tree $v\in LDAT3_z$ of the form $v=[t_1,\dots,t_m,v_1,\dots,v_n,u_1,\dots,u_p]_z$ and for all $r\geq\varrho(v)$ we have

$$(F(v)(\Phi(h)))^{(r-\varrho(v))}=\sum_{\substack{\overline{V}\in CLDAT3(v)\\ D_1(\overline{V})=\{\overline{T}_1,\dots,\overline{T}_M,\overline{V}_1,\dots,\overline{V}_N,\overline{U}_1,\dots,\overline{U}_P\},\ \varrho(\overline{V})=r,\ M\geq m,\ N\geq n,\ P\geq p\\ t_1\subset\overline{T}_1,\dots,t_m\subset\overline{T}_m,v_1\subset\overline{V}_1,\dots,v_n\subset\overline{V}_n,u_1\subset\overline{U}_1,\dots,u_p\subset\overline{U}_p}} \frac{\partial^{M+N+P}k}{\partial y^M\partial z^N\partial u^P}\Big((F(t_1)(\Phi(h))^{(\varrho(\overline{T}_1)-\varrho(t_1))},\dots \tag{5.11d}$$
$$\dots,(F(t_M)(\Phi(h))^{(\varrho(\overline{T}_M)-\varrho(t_M))},(F(v_1)(\Phi(h))^{(\varrho(\overline{V}_1)-\varrho(v_1))},\dots$$
$$\dots,(F(v_N)(\Phi(h))^{(\varrho(\overline{V}_N)-\varrho(v_N))},(F(u_1)(\Phi(h))^{(\varrho(\overline{U}_1)-\varrho(u_1))},\dots$$
$$\dots,(F(u_P)(\Phi(h))^{(\varrho(\overline{U}_P)-\varrho(u_P))}\Big)\ ;$$

where we have defined $t_j:=\emptyset_y$ for $m<j\leq M$, $v_k:=\emptyset_z$ for $n<k\leq N$, and $u_l:=\emptyset_u$ for $n<l\leq P$ (therefore we have $(F(t_j)(\Phi(h))^{(\varrho(\overline{T}_j)-\varrho(t_j))}=\Phi_y(h)^{(\varrho(\overline{T}_j))}$, $(F(v_k)(\Phi(h))^{(\varrho(\overline{V}_k)-\varrho(v_k))}=\Phi_z(h)^{(\varrho(\overline{V}_k))}$, $(F(u_l)(\Phi(h))^{(\varrho(\overline{U}_l)-\varrho(u_l))}=\Phi_u(h)^{(\varrho(\overline{U}_l))}$ for these trees).

e) for a fixed labelled tree $u\in LDAT3_u$ of the form $u=[t_1,\dots,t_m]_u$ and for all $r\geq\varrho(u)$ we have

$$(F(u)(\Phi(h)))^{(r-\varrho(u))}=\sum_{\substack{\overline{U}\in CLDAT3(u)\\ D_1(\overline{U})=\{\overline{T}_1,\dots,\overline{T}_M,\overline{V}_1,\dots,\overline{V}_N,\overline{U}_1,\dots,\overline{U}_P\},\ \varrho(\overline{U})=r,\ M\geq m\\ t_1\subset\overline{T}_1,\dots,t_m\subset\overline{T}_m}} G(\overline{U})\Big((F(t_1)(\Phi(h))^{(\varrho(\overline{T}_1)-\varrho(t_1))},\dots \tag{5.11e}$$
$$\dots,(F(t_M)(\Phi(h))^{(\varrho(\overline{T}_M)-\varrho(t_M))},\Phi_z(h)^{(\varrho(\overline{V}_1))},\dots,\Phi_z(h)^{(\varrho(\overline{V}_N))},$$
$$\Phi_u(h)^{(\varrho(\overline{U}_1))},\dots,\Phi_u(h)^{(\varrho(\overline{U}_P))}\Big)\ ;$$

where we have defined $t_j := \emptyset_y$ *for* $m < j \le M$ *(therefore for these trees we have* $(F(t_j)(\Phi(h))^{(\varrho(\overline{T}_j)-\varrho(t_j))} = \Phi_y(h)^{(\varrho(\overline{T}_j))}$*). $G(\overline{U})$ is a natural expression depending on $\overline{U}$ and whose arguments are given here by* $(F(t_1)(\Phi(h))^{(\varrho(\overline{T}_1)-\varrho(t_1))}$, $\ldots, \Phi_u(h)^{(\varrho(\overline{U}_P))}$ *(see the two examples in Fig 5.7 above).* □

After all these preparations we are now able to prove Lemma 5.1:

Proof of Lemma 5.1. We define $\Phi(h)$ by

$$\Phi(h) = \big(\Phi_y(h), \Phi_z(h), \Phi_u(h)\big) := \big(DA3_y(\mathbf{a}, \Psi), DA3_z(\mathbf{a}, \Psi), DA3_u(\mathbf{a}, \Psi)\big) \ . \quad (5.12)$$

For $\emptyset_\sigma \in \{\emptyset_y, \emptyset_z, \emptyset_u\}$ we have for all $r \ge 0$

$$(F(\emptyset_\sigma)(\Phi(h)))^{(r)}\Big|_{h=0} = \Phi_\sigma(h)^{(r)}\Big|_{h=0} = \sum_{\substack{s \in LDAT3_\sigma \cup \{\emptyset_\sigma\} \\ \varrho(s)=r}} \mathbf{a}(s)F(s)(\Psi) \ . \quad (5.13)$$

For τ_y and all $r \ge 1$, inserting in (5.11a) the derivatives of $\Phi_y(h)$ and $\Phi_z(h)$ at $h=0$ computed above we get

$$\big(F(\tau_y)(\Phi(h))\big)^{(r-1)}\Big|_{h=0} = \quad (5.14)$$

$$\sum_{\substack{\overline{T} \in CLDAT3(\tau_y) \\ D_1(\overline{T}) = \{\overline{T}_1, \ldots, \overline{T}_M, \overline{V}_1, \ldots, \overline{V}_N\}, \ \varrho(\overline{T})=r}} \frac{\partial^{M+N} f}{\partial y^M \partial z^N} \Bigg(\sum_{\substack{T_1 \in LDAT3_y \cup \{\emptyset_y\} \\ \varrho(T_1)=\varrho(\overline{T}_1)}} \mathbf{a}(T_1)F(T_1)(\Psi), \ldots$$

$$\ldots, \sum_{\substack{T_M \in LDAT3_y \cup \{\emptyset_\sigma\} \\ \varrho(T_M)=\varrho(\overline{T}_M)}} \mathbf{a}(T_M)F(T_M)(\Psi), \sum_{\substack{V_1 \in LDAT3_z \cup \{\emptyset_z\} \\ \varrho(V_1)=\varrho(\overline{V}_1)}} \mathbf{a}(V_1)F(V_1)(\Psi), \ldots$$

$$\ldots, \sum_{\substack{V_N \in LDAT3_z \cup \{\emptyset_z\} \\ \varrho(V_N)=\varrho(\overline{V}_N)}} \mathbf{a}(V_N)F(V_N)(\Psi) \Bigg) \ .$$

The main point is that to each tuple $(\overline{T}, T_1, \ldots, T_M, V_1, \ldots, V_N)$ with $T_1, \ldots, T_M \in LDAT3_y$, $V_1, \ldots, V_N \in LDAT3_z$, $D_1(\overline{T}) = \{\overline{T}_1, \ldots, \overline{T}_M, \overline{V}_1, \ldots, \overline{V}_N\}$, $\varrho(T_1) = \varrho(\overline{T}_1)$, $\ldots, \varrho(V_N) = \varrho(\overline{V}_N)$, there corresponds a unique m.l. tree $T = [T_1, \ldots, T_M, V_1, \ldots, V_N]_y \in LDAT3_y$ and conversely. We illustrate this fact on the following example

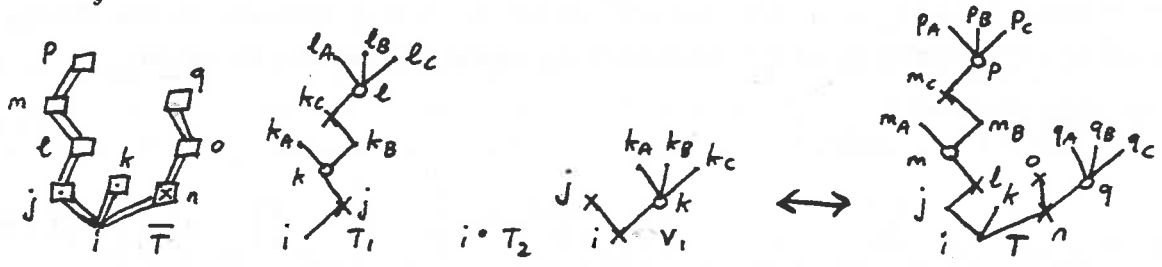


Figure 5.11.

Thus using $LDAT3_y$ as a new summation set and exploiting the multilinearity of the derivatives, we can rewrite (5.14) as

$$\left(F(\tau_y)(\Phi(h))\right)^{(r-1)}\Big|_{h=0} = \sum_{\substack{T=[T_1,\dots,T_M,V_1,\dots,V_N]_y \in LDAT3_y \\ \varrho(T)=r}} \mathrm{a}(T_1)\cdots\mathrm{a}(T_M)\mathrm{a}(V_1)\cdots\mathrm{a}(V_N)F(T)(\Psi) \tag{5.15}$$

and the result easily follows from Remark 5.2.2 which leads to

$$\prod_{\omega\in d_1(T)} \mathrm{a}(\omega) = \mathrm{a}(T_1)\cdots\mathrm{a}(T_M)\mathrm{a}(V_1)\cdots\mathrm{a}(V_N)\,. \tag{5.16}$$

Similarly for τ_z we obtain for all $r\geq 1$

$$\left(F(\tau_z)(\Phi(h))\right)^{(r-1)}\Big|_{h=0} = k(\Phi_y(h),\Phi_z(h),\Phi_u(h))^{(r-1)}\Big|_{h=0} = \sum_{\substack{V=[T_1,\dots,T_M,V_1,\dots,V_N,U_1,\dots,U_P]_z \in LDAT3_z \\ \varrho(V)=r}} \mathrm{a}(T_1)\cdots\mathrm{a}(T_M)\mathrm{a}(V_1)\cdots\mathrm{a}(V_N)\mathrm{a}(U_1)\cdots\mathrm{a}(U_P)F(V)(\Psi)\,. \tag{5.17}$$

We now assume (5.5) by induction hypothesis for all different labellings of the trees $t_1,\dots,t_m\in DAT3_y$, $v_1,\dots,v_n\in DAT3_z$, and $u_1,\dots,u_p\in DAT3_u$. We will show that (5.5) holds for any labelling of the three trees $t=[t_1,\dots,t_m,v_1,\dots,v_n]_y\in DAT3_y$, $v=[t_1,\dots,t_m,v_1,\dots,v_n,u_1,\dots,u_p]_z\in DAT3_z$, and $u=[t_1,\dots,t_m]_u\in DAT3_u$. Let us first consider $t=[t_1,\dots,t_m,v_1,\dots,v_n]_y\in DAT3_y$ with a fixed labelling and $r\geq\varrho(t)$. In formula (5.11c) we can make use of the induction hypothesis for $t_1,\dots,v_n$ and also insert the derivatives of $\Phi_y(h)$, $\Phi_z(h)$, and $\Phi_u(h)$ at $h=0$ computed above in (5.13). The main point is that to each tuple $(\overline{T},T_1,\dots,T_M,V_1,\dots,V_N)$ with $T_1,\dots,T_M\in LDAT3_y$, $V_1,\dots,V_N\in LDAT3_z$, $D_1(\overline{T})=\{\overline{T}_1,\dots,\overline{T}_M,\overline{V}_1,\dots,\overline{V}_N\}$, $\varrho(\overline{T}_1)=\varrho(T_1),\dots,\varrho(\overline{V}_N)=\varrho(V_N)$, and $t_1\subset\overline{T}_1$, $t_1\subset T_1$, …, $v_n\subset\overline{V}_n$, $v_n\subset V_n$, there corresponds a unique m.l. tree $T=[T_1,\dots,T_M,V_1,\dots,V_N]_y\in LDAT3_y$ with $t\subset T$ and conversely. We illustrate this fact on the following example

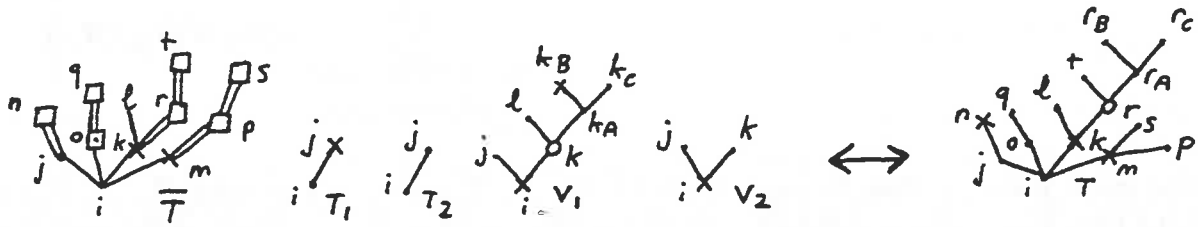


Figure 5.12.

Thus using $LDAT3_y$ as a new summation set in (5.11c) similarly to the change in (5.14)-(5.15) and exploiting the multilinearity of the derivatives, we obtain

$$\left(F(t)(\Phi(h))\right)^{(r-\varrho(t))}\Big|_{h=0} = \sum_{\substack{T=[T_1,\dots,T_M,V_1,\dots,V_N]_y \in LDAT3_y \\ s_{\varrho(t)}(T)=t,\ \varrho(T)=r}} \left(\prod_{\omega\in d_{\varrho(t_1)}(T_1)} \mathrm{a}(\omega)\right)\cdots\left(\prod_{\omega\in d_{\varrho(v_N)}(V_N)} \mathrm{a}(\omega)\right) F(T)(\Psi) \tag{5.18}$$

and as $\varrho(T) \geq 1$ the result easily follows from

$$\prod_{\omega \in d_{\varrho(t)}(T)} \mathbf{a}(z) = \left(\prod_{\omega \in d_{\varrho(t_1)}(T_1)} \mathbf{a}(\omega) \right) \cdots \left(\prod_{\omega \in d_{\varrho(v_N)}(V_N)} \mathbf{a}(\omega) \right) . \tag{5.19}$$

Similarly for any labelling of $v = [t_1, \ldots, t_m, v_1, \ldots, v_n, u_1, \ldots, u_p]_z$ we obtain for all $r \geq \varrho(v)$

$$\left(F(v)(\Phi(h))\right)^{(r-\varrho(v))}\Big|_{h=0} = \tag{5.20}$$

$$\sum_{\substack{V=[T_1,\ldots,T_M,V_1,\ldots,V_N,U_1,\ldots,U_P]_z \in LDAT3_z \\ s_{\varrho(v)}(V)=v,\ \varrho(V)=r}} \left(\prod_{\omega \in d_{\varrho(t_1)}(T_1)} \mathbf{a}(\omega) \right) \cdots \left(\prod_{\omega \in d_{\varrho(u_P)}(U_P)} \mathbf{a}(\omega) \right) F(V)(\Psi) .$$

It remains to show (5.5) for any fixed labelling of $u = [t_1, \ldots, t_m]_u \in DAT3_u$. This is the most difficult part of this proof. In formula (5.11e) we can make use of the induction hypothesis for $t_1, \ldots, t_m$ and insert the derivatives of $\Phi_y(h)$, $\Phi_z(h)$, and $\Phi_u(h)$ at $h = 0$ computed above in (5.13). The main point is that to each tuple $(\overline{U}, T_1, \ldots, T_M, V_1, \ldots, V_N, U_1, \ldots, U_P)$ with $T_1, \ldots, T_M \in LDAT3_y$, $V_1, \ldots, V_N \in LDAT3_z$, $U_1, \ldots, U_P \in LDAT3_u$, $D_1(\overline{U}) = \{\overline{T}_1, \ldots, \overline{T}_M, \overline{V}_1, \ldots, \overline{V}_N, \overline{U}_1, \ldots, \overline{U}_P\}$, $\varrho(\overline{T}_1) = \varrho(T_1), \ldots, \varrho(\overline{U}_P) = \varrho(U_P)$, and $t_1 \subset \overline{T}_1$, $t_1 \subset T_1$, $\ldots$, $t_m \subset \overline{T}_m$, $t_m \subset T_m$, and $\varrho(\overline{T}_1) = \varrho(T_1), \ldots, \varrho(\overline{U}_P) = \varrho(U_P)$, there corresponds a unique m.l. tree $U \in LDAT3_u$ containing u and conversely. We illustrate this fact on the following example

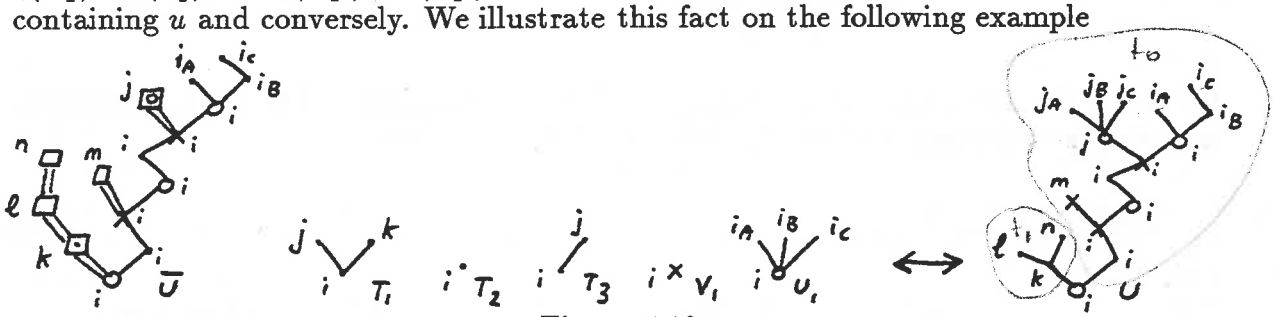


Figure 5.13.

Thus using $LDAT3_u$ as a new summation set in (5.11e) similarly to the change in (5.14)-(5.15) and again exploiting the multilinearity of the derivatives, we obtain the desired result

$$\left(F(u)(\Phi(h))\right)^{(r-\varrho(u))}\Big|_{h=0} = \sum_{\substack{U \in LDAT3_u \\ s_{\varrho(u)}(U)=u,\ \varrho(U)=r}} \left(\prod_{\omega \in d_{\varrho(u)}(U)} \mathbf{a}(\omega) \right) F(U)(\Psi) . \tag{5.21}$$

□

Definition 5.10. Let $\mathbf{a} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ and $\mathbf{b} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ be two arbitrary maps and suppose that $\mathbf{a}_y(\emptyset_y) = 1$, $\mathbf{a}_z(\emptyset_z) = 1$, and $\mathbf{a}_u(\emptyset_u) = 1$. Then we define the composition $\mathbf{a} * \mathbf{b} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ for each tree w by

$$(\mathbf{a} * \mathbf{b})(w) = \frac{1}{\alpha(w)} \sum_{\text{labellings of } w} \left(\sum_{j=0}^{\varrho(w)} \binom{\varrho(w)}{j} \mathbf{b}(s_j(w)) \prod_{\omega \in d_j(w)} \mathbf{a}(\omega) \right) \tag{5.22}$$

where the first summation is over all $\alpha(w)$ different labellings of w.

Here is the main consequence of Lemma 5.1 and the main result of this section:

Theorem 5.3. *As above let* $\mathbf{a} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ *and* $\mathbf{b} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ *be two arbitrary maps and suppose that* $\mathbf{a}(\emptyset_y)=1$, $\mathbf{a}(\emptyset_z)=1$, *and* $\mathbf{a}(\emptyset_u)=1$. *Then the composition of the two corresponding DA3-series is again a DA3-series*

$$DA3(\mathbf{b}, DA3(\mathbf{a}, \Psi)) = DA3(\mathbf{m}, \Psi) \tag{5.23}$$

where the map $\mathbf{m} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ *is given by* $\mathbf{m}=\mathbf{a}*\mathbf{b}$ *of Definition 5.10.*

Proof. In order not to write similar formulas for the y-, z-,and u-component we also denote by σ an arbitrary subscript letter in $\{y, z, u\}$. Putting $\Phi(h) := DA3(\mathbf{a}, \Psi)$, we have

$$DA3_\sigma(\mathbf{b}, DA3(\mathbf{a}, \Psi)) = DA3_\sigma(\mathbf{b}, \Phi(h)) = \sum_{w \in LDAT3_\sigma \cup \{\emptyset_\sigma\}} \frac{h^{\varrho(w)}}{\varrho(w)!} \mathbf{b}(w) F(w)(\Phi(h)) \ . \tag{5.24}$$

By the Leibniz' rule we have for the qth derivative

$$DA3_\sigma(\mathbf{b}, \Phi(h))^{(q)}\Big|_{h=0} = \sum_{\substack{w \in LDAT3_\sigma \cup \{\emptyset_\sigma\} \\ 0 \le \varrho(w) \le q}} \binom{q}{\varrho(w)} \mathbf{b}(w) \left(F(w)(\Phi(h))\right)^{(q-\varrho(w))}\Big|_{h=0} \ . \tag{5.25}$$

We insert the derivatives $\left(F(w)(\Phi(h))\right)^{(q-\varrho(w))}|_{h=0}$ computed in Lemma 5.1, rearrange the terms so obtained, and use the fact that $\Phi(0)=\Psi$ to get

$$\begin{aligned} & DA3_\sigma(\mathbf{b}, \Phi(h))^{(q)}\Big|_{h=0} \qquad (5.26) \\ &= \sum_{\substack{w \in LDAT3_\sigma \cup \{\emptyset_\sigma\} \\ 0 \le \varrho(w) \le q}} \binom{q}{\varrho(w)} \mathbf{b}(w) \sum_{\substack{s \in LDAT3_\sigma \cup \{\emptyset_\sigma\} \\ s_{\varrho(w)}(s)=w,\ \varrho(s)=q}} \left(\prod_{\omega \in d_{\varrho(w)}(s)} \mathbf{a}(\omega) \right) F(s)(\Phi(0)) \\ &= \sum_{\substack{s \in LDAT3_\sigma \cup \{\emptyset_\sigma\} \\ \varrho(s)=q}} \left(\sum_{j=0}^{\varrho(s)} \binom{\varrho(s)}{j} \mathbf{b}(s_j(s)) \prod_{\omega \in d_j(s)} \mathbf{a}(\omega) \right) F(s)(\Psi) \\ &= \sum_{\substack{s \in DAT3_\sigma \cup \{\emptyset_\sigma\} \\ \varrho(s)=q}} \sum_{\text{labellings of } s} \left(\sum_{j=0}^{\varrho(s)} \binom{\varrho(s)}{j} \mathbf{b}(s_j(s)) \prod_{\omega \in d_j(s)} \mathbf{a}(\omega) \right) F(s)(\Psi) \\ &= \sum_{\substack{s \in DAT3_\sigma \cup \{\emptyset_\sigma\} \\ \varrho(s)=q}} \alpha(s)(\mathbf{a}*\mathbf{b})(s) F(s)(\Psi) = DA3_\sigma(\mathbf{a}*\mathbf{b}, \Psi)^{(q)}\Big|_{h=0} \ . \end{aligned}$$

As all derivatives coincide the proof is achieved. □

The main consequence of the preceding theorem is given by the following result:

Theorem 5.4. *The set of mappings*

$$G := \{\mathbf{a} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R};\quad \mathbf{a}(\emptyset_y) = 1, \mathbf{a}(\emptyset_z) = 1, \mathbf{a}(\emptyset_u) = 1\}$$

with the composition law (5.22) *forms a non-commutative group.*

Proof.

a) The associativity of the operation $*$ is a trivial consequence of the preceding Theorem 5.3 as for an arbitrary Ψ we have

$$DA3(\mathbf{c}, DA3(\mathbf{b}, DA3(\mathbf{a}, \Psi))) = DA3(\mathbf{c}, DA3(\mathbf{a} * \mathbf{b}, \Psi)) = DA3((\mathbf{a} * \mathbf{b}) * \mathbf{c}, \Psi)\ , \quad (5.27a)$$
$$DA3(\mathbf{c}, DA3(\mathbf{b}, DA3(\mathbf{a}, \Psi))) = DA3(\mathbf{b} * \mathbf{c}, DA3(\mathbf{a}, \Psi)) = DA3(\mathbf{a} * (\mathbf{b} * \mathbf{c}), \Psi)\ , \quad (5.27b)$$

i.e,

$$DA3((\mathbf{a} * \mathbf{b}) * \mathbf{c}, \Psi) = DA3(\mathbf{a} * (\mathbf{b} * \mathbf{c}), \Psi)\ . \quad (5.28)$$

From the linear independence of the elementary differentials (similarly to [HaNøWa93, Exercise II.2.4]) it follows that

$$((\mathbf{a} * \mathbf{b}) * \mathbf{c})(w) = (\mathbf{a} * (\mathbf{b} * \mathbf{c}))(w) \quad \forall w \in DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\}\ . \quad (5.29)$$

b) The neutral element $\mathbf{e}$ of the group is trivially given by the identity map of Example 5.1.1.

c) Let $\mathbf{a} \in G$ be given. The inverse $\mathbf{a}^{-1}$ of $\mathbf{a}$ for the operation $*$ can be constructed from (5.22) inductively on the order of trees. First for $\emptyset_\sigma \in \{\emptyset_y, \emptyset_z, \emptyset_u\}$ we have $1 = (\mathbf{a} * \mathbf{a}^{-1})(\emptyset_\sigma) = \mathbf{a}(\emptyset_\sigma)\mathbf{a}^{-1}(\emptyset_\sigma)$, therefore $\mathbf{a}^{-1}(\emptyset_\sigma) = 1$ must hold as a consequence of $\mathbf{a}(\emptyset_\sigma) = 1$. Consider now an arbitrary tree $w \in DAT3$. As for any labelling of w we have $\varrho(s_j(w)) \le \varrho(w) - 1$ if $j \le \varrho(w) - 1$, then $\mathbf{a}^{-1}(w)$ can be computed from $0 = \mathbf{e}(w) = (\mathbf{a} * \mathbf{a}^{-1})(w)$ leading to

$$0 = \frac{1}{\alpha(w)} \sum_{\text{labellings of } w} \left(\sum_{j=0}^{\varrho(w)-1} \binom{\varrho(w)}{j} \mathbf{a}^{-1}(s_j(w)) \prod_{\omega \in d_j(w)} \mathbf{a}(\omega) \right) + \mathbf{a}^{-1}(w) \underbrace{\mathbf{a}(\emptyset_{\sigma(w)})}_{=1}\ . \quad (5.30)$$

□

Remark 5.3. The non-commutativity is a trivial consequence of the non-commutativity of the composition of B-series (see [HaNøWa93, Exercise II.12.4]).

Another consequence of Theorem 5.3 is as follows:

Theorem 5.5. *Let* $\mathbf{a} : DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\} \to \mathbb{R}$ *be a DA3-series satisfying* $\mathbf{a}(\emptyset_y) = 1$, $\mathbf{a}(\emptyset_z) = 1$, *and* $\mathbf{a}(\emptyset_u) = 1$. *Consider* $a_y := DAT3_y(\mathbf{a}, \Psi)$, $a_z := DAT3_z(\mathbf{a}, \Psi)$, *and* $a_u := DAT3_u(\mathbf{a}, \Psi)$ *with* $\Psi = (\tilde{y}, \tilde{z}, \tilde{u})$ *consistent. Then we have*

$$hf(a_y, a_z) = DAT3_y(\mathbf{a}', \Psi)\ , \qquad hk(a_y, a_z, a_u) = DAT3_z(\mathbf{a}', \Psi)\ , \quad (5.31)$$

with

$$\begin{aligned}
&\mathbf{a}'(\emptyset_y) = 0\ , \quad \mathbf{a}'(\emptyset_z) = 0\ , \quad \mathbf{a}'(\tau_y) = 1\ , \quad \mathbf{a}'(\tau_z) = 1\ , \\
&\mathbf{a}'(t) = \mathbf{a}'([t_1,\dots,t_m,v_1,\dots,v_n]_y) = \varrho(t)\mathbf{a}(t_1)\cdots\mathbf{a}(t_m)\mathbf{a}(v_1)\cdots\mathbf{a}(v_n)\ , \\
&\mathbf{a}'(v) = \mathbf{a}'([t_1,\dots,t_m,v_1,\dots,v_n,u_1,\dots,v_p]_z) = \\
&\qquad\qquad \varrho(v)\mathbf{a}(t_1)\cdots\mathbf{a}(t_m)\mathbf{a}(v_1)\cdots\mathbf{a}(v_n)\mathbf{a}(u_1)\cdots\mathbf{a}(u_p)\ ,
\end{aligned} \tag{5.32}$$

If in addition $\mathbf{a}(\tau_y)^2 = \mathbf{a}([\tau_y]_y) = \mathbf{a}([\tau_z]_y)$ *is satisfied then we have*

$$\frac{1}{h^2}(-g_y f_z k_u)^{-1}(\widetilde{y},\widetilde{z},\widetilde{u})g(a_y) = DAT3_u(\mathbf{a}',\Psi)\ , \tag{5.33}$$

with $\mathbf{a}'(\emptyset_u) = 0$ *and*

$$\mathbf{a}'(u) = \mathbf{a}'([t_1,\dots,t_m]_u) = \frac{1}{(\varrho(u)+2)(\varrho(u)+1)}\left(\mathbf{a}(t_1)\cdots\mathbf{a}(t_m) - \mathbf{a}([[u]_z]_y)\right)\ . \tag{5.34}$$

Proof. For trees in $DAT3_y \cup \{\emptyset_y\}$, (5.32) is a simple consequence of (5.23) since for $\Psi_0 = (y_0, z_0, u_0)$ we get $hf(y_0,z_0) = DA3_y(\mathbf{b},\Psi_0)$ with $\mathbf{b}(\tau_y) = 1$ and for all other trees $t \in DAT3_y \cup \{\emptyset_y\}$ we have $\mathbf{b}(t) = 0$. Similarly for trees in $DAT3_z \cup \{\emptyset_z\}$, we get $hk(y_0,z_0,u_0) = DA3_z(\mathbf{b},\Psi_0)$ with $\mathbf{b}(\tau_z) = 1$ and for all other trees $v \in DAT3 \cup \{\emptyset_z\}$ we have $\mathbf{b}(v) = 0$. Concerning (5.34) the proof is similar to that of [Ro88a, Theorem (3.4)] and the ideas are identical to those used in the proof of Lemma 5.1. Denoting $\Phi(h) := (-g_y f_z k_u)^{-1}(\widetilde{y},\widetilde{z},\widetilde{u})g(a_y)$ we have

$$\begin{aligned}
\Phi(0) &= \left((-g_y f_z k_u)^{-1} g\right)(\widetilde{y},\widetilde{z},\widetilde{u}) = 0\ , \\
\Phi^{(1)}(0) &= \left((-g_y f_z k_u)^{-1}(\mathbf{a}(\tau_y) g_y f)\right)(\widetilde{y},\widetilde{z},\widetilde{u}) = 0\ , \\
\Phi^{(2)}(0) &= \left((-g_y f_z k_u)^{-1}(\mathbf{a}(\tau_y)^2 g_{yy}(f,f) + \mathbf{a}([\tau_y]_y) g_y f_y f + \mathbf{a}([\tau_z]_y) g_y f_z k)\right)(\widetilde{y},\widetilde{z},\widetilde{u}) = 0\ ,
\end{aligned} \tag{5.35}$$

and the higher derivatives $\Phi^{(k)}(0)$ for $k \geq 3$ are (non-null in general) linear combinations of the elementary differentials of order $k-2$. The factor $1/(\varrho(u)+2)(\varrho(u)+1)$ in (5.34) comes from

$$\frac{1}{h^2}\Phi(h) = \frac{1}{3\cdot 2}\cdot\frac{h}{1!}\Phi^{(3)}(0) + \frac{1}{4\cdot 3}\cdot\frac{h^2}{2!}\Phi^{(4)}(0) + \frac{1}{5\cdot 4}\cdot\frac{h^3}{3!}\Phi^{(5)}(0) + \dots\ .$$

□

Chapter III. Partitioned Runge-Kutta methods for semi-explicit index 3 DAE's in Hessenberg form.

1. PRK methods and related definitions.

We consider the equations (II.1.1a,b,c) and consistent initial values $\Psi_0=(y_0,z_0,u_0)$ at x_0 satisfying (II.1.2). Instead of simply studying the application of Runge-Kutta (RK) methods to this problem, we will consider a more general class of methods exploiting the specific partitioning of the system (II.1.1a,b,c). These methods, called *partitioned Runge-Kutta (PRK) methods*, make use of the conjunction of two Runge-Kutta methods as follows:

Definition 1.1. One step of an *s-stage PRK method* applied to (II.1.1a,b,c) with initial values $\Psi_0=(y_0,z_0,u_0)$ at x_0 reads

$$y_1 = y_0+h\sum_{i=1}^{s} b_i Y_i' \,, \qquad z_1 = z_0+h\sum_{i=1}^{s} \widehat{b}_i Z_i' \tag{1.1a}$$

where

$$Y_i' = f(Y_i,Z_i)\,, \qquad Z_i' = k(Y_i,Z_i,U_i)\,, \qquad 0 = g(Y_i)\,, \tag{1.1b}$$

and the *internal stages* are given by

$$Y_i = y_0+h\sum_{j=1}^{s} a_{ij} Y_j' \,, \qquad Z_i = z_0+h\sum_{j=1}^{s} \widehat{a}_{ij} Z_j' \,. \tag{1.1c}$$

Remarks 1.1.

1) The existence and uniqueness of a solution to these equations is not guaranteed without some assumptions on the coefficients (see Theorem 2.1 below and Theorem V.4.1).
2) If the coefficients of the PRK method satisfy

$$(S): \qquad a_{si} = b_i \qquad \text{for} \quad i=1,\dots,s$$

then we have $y_1=Y_s$ and hence $g(y_1)=g(Y_s)=0$.
3) Various definitions of the numerical u-component u_1 will be given further.

The coefficients of the two RK methods entering in the PRK method (1.1) can be written in two Butcher-tableaux

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \qquad \begin{array}{c|ccc} \widehat{c}_1 & \widehat{a}_{11} & \cdots & \widehat{a}_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{c}_s & \widehat{a}_{s1} & \cdots & \widehat{a}_{ss} \\ \hline & \widehat{b}_1 & \cdots & \widehat{b}_s \end{array} .$$

We always suppose that these two RK methods are based on the same quadrature formula, i.e., that

$$b_i = \widehat{b}_i, \qquad c_i = \widehat{c}_i \qquad \text{for} \quad i = 1, \dots, s \ . \tag{1.2}$$

PRK methods violating these conditions will not be considered (*half-explicit methods* are examples of such methods, see [Os90] and [Os93]). The coefficients c_i enter in the definition of PRK methods for non-autonomous problems. The hypotheses

$$C(1): \qquad \sum_{j=1}^{s} a_{ij} = c_i \ , \quad \text{for} \quad i = 1, \dots, s \ ,$$

$$\widehat{C}(1): \qquad \sum_{j=1}^{s} \widehat{a}_{ij} = c_i \quad \text{for} \quad i = 1, \dots, s \ ,$$

imply some simplifications when deriving the order conditions for high order methods applied to such problems (see also [HaNøWa93, p. 134]): the order conditions reduce to those of the autonomous case considered here. For most PRK methods of interest these assumptions are fulfilled (see Chapter V for an exception: the 2-stage Lobatto IIIA-IIIB method which does not satisfy $\widehat{C}(1)$). PRK methods violating the assumption $C(1)$ will not be considered. Pure *Runge-Kutta methods* applied to (II.1.1a,b,c) satisfy also in addition to (1.2)

$$a_{ij} = \widehat{a}_{ij} \qquad \text{for} \quad i = 1, \dots, s \ , \quad j = 1, \dots, s \ . \tag{1.3}$$

From now on we use the notation $A := (a_{ij})_{i,j=1}^{s}$, $\widehat{A} := (\widehat{a}_{ij})_{i,j=1}^{s}$ for the *RK matrices*, $b := (b_1, \dots, b_s)^T$ for the *weight vector*, and $c := (c_1, \dots, c_s)^T$ for the *node vector*. Under the assumptions (1.2) the two Butcher-tableaux can be rewritten more succinctly as

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \qquad \begin{array}{c|c} c & \widehat{A} \\ \hline & b^T \end{array} \ .$$

We will designate the two couples of RK coefficients by (A, b, c), $(\widehat{A}, b, c)$ and the PRK coefficients by $(A, \widehat{A}, b, c)$. We also denote the product matrix $\overline{A} := A\widehat{A}$. If the matrices A and $\widehat{A}$ are non-singular then we define the inverse matrices $W = (w_{ij})_{i,j=1}^{s} := A^{-1}$, $\widehat{W} = (\widehat{w}_{ij})_{i,j=1}^{s} := \widehat{A}^{-1}$, and $\Omega = (\omega_{ij})_{i,j=1}^{s} := \overline{A}^{-1} = \widehat{W}W$. For a RK method we have $A = \widehat{A}, W = \widehat{W}, \overline{A} = A^2$, and $\Omega = W^2$. We will sometimes use the s-dimensional vectors $\mathbb{1} := (1, \dots, 1)^T$ and $e_s := (0, \dots, 0, 1)^T$.

1.1. The simplifying assumptions.

The construction of high order PRK methods relies heavily on the following conditions, called *simplifying assumptions* (see [But87, Section 34], [HaLuRo89a, pp. 15-16],

and [HaWa91, Section IV.5]):

$$B(p): \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k} \quad \text{for} \quad k = 1, \ldots, p\,;$$

$$C(q): \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{for} \quad i = 1, \ldots, s\,, \quad k = 1, \ldots, q\,;$$

$$\widehat{C}(\widehat{q}): \quad \sum_{j=1}^{s} \widehat{a}_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{for} \quad i = 1, \ldots, s\,, \quad k = 1, \ldots, \widehat{q}\,;$$

$$D(r): \quad \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{b_j}{k}(1 - c_j^k) \quad \text{for} \quad j = 1, \ldots, s\,, \quad k = 1, \ldots, r\,;$$

$$\widehat{D}(\widehat{r}): \quad \sum_{i=1}^{s} b_i c_i^{k-1} \widehat{a}_{ij} = \frac{b_j}{k}(1 - c_j^k) \quad \text{for} \quad j = 1, \ldots, s\,, \quad k = 1, \ldots, \widehat{r}\,;$$

$$C\widehat{C}(Q): \quad \sum_{j=1}^{s} \sum_{l=1}^{s} a_{ij} \widehat{a}_{jl} c_l^{k-2} = \frac{c_i^k}{k(k-1)} \quad \text{for} \quad i = 1, \ldots, s\,, \quad k = 2, \ldots, Q\,;$$

$$D\widehat{D}(R): \quad \sum_{i=1}^{s} \sum_{j=1}^{s} b_i c_i^{k-2} a_{ij} \widehat{a}_{jl} = \frac{b_l}{k} - \frac{b_l c_l}{k-1} + \frac{b_l c_l^k}{k(k-1)} \quad \text{for} \quad l = 1, \ldots, s\,, \quad k = 2, \ldots, R\,;$$

$$(S): \quad a_{si} = b_i \quad \text{for} \quad i = 1, \ldots, s\,.$$

$C\widehat{C}(Q)$ and $D\widehat{D}(R)$ can be rewritten

$$C\widehat{C}(Q): \quad \sum_{l=1}^{s} \overline{a}_{il} c_l^{k-2} = \frac{c_i^k}{k(k-1)} \quad \text{for} \quad i = 1, \ldots, s\,, \quad k = 2, \ldots, Q\,;$$

$$D\widehat{D}(R): \quad \sum_{i=1}^{s} b_i c_i^{k-2} \overline{a}_{il} = \frac{b_l}{k} - \frac{b_l c_l}{k-1} + \frac{b_l c_l^k}{k(k-1)} \quad \text{for} \quad l = 1, \ldots, s\,, \quad k = 2, \ldots, R\,.$$

Remarks 1.2.

1) (S) together with $B(1)$ leads to $c_s = 1$. Moreover, if in addition $C(q)$ and $D(r)$ are satisfied then $B(p)$ holds with $p \geq \max(q, r+1)$.
2) If $C(q)$ and $\widehat{C}(\widehat{q})$ are satisfied then it can be easily shown that $C\widehat{C}(Q)$ holds with $Q \geq \min(q, \widehat{q}+1)$. Similarly if $D(r)$ and $\widehat{D}(\widehat{r})$ are satisfied then $D\widehat{D}(R)$ holds with $R \geq \min(\widehat{r}, r+1)$.
3) If $C\widehat{C}(Q)$ is not satisfied with $Q \geq 2$ then by default we put $Q := 1$. Similarly if $D\widehat{D}(R)$ is not satisfied with $R \geq 2$ then we put $R := 1$.
4) The conditions $C(2)$ and $C\widehat{C}(2)$ are important to show the existence and uniqueness of the PRK solution (see Theorem 2.1 and Theorem V.4.1).
5) If the matrix A is invertible and $C(2)$ is satisfied then $C\widehat{C}(2)$ holds if and only if $\widehat{C}(1)$ holds. This is not true if the matrix A is singular, e.g., the 2-stage Lobatto IIIA-IIIB method satisfies $C(2)$, $C\widehat{C}(2)$, but not $\widehat{C}(1)$ (see Chapter V).
6) For *stiffly accurate* RK methods, i.e., satisfying (S), the condition $D(r)$ plays an important role, not only when estimating the local error, but also for the global error of unprojected methods (see Chapter IV).

If the matrices A and $\widehat{A}$ are invertible then we obtain the following additional simplifying assumptions by multiplication of the above simplifying assumptions with W, $\widehat{W}$, or Ω:

$$IC(q): \quad \sum_{j=1}^{s} w_{ij}c_j^k = kc_i^{k-1} \quad \text{for} \quad i=1,\dots,s\,, \quad k=1,\dots,q\,;$$

$$I\widehat{C}(\widehat{q}): \quad \sum_{j=1}^{s} \widehat{w}_{ij}c_j^k = kc_i^{k-1} \quad \text{for} \quad i=1,\dots,s\,, \quad k=1,\dots,\widehat{q}\,;$$

$$ID(r): \quad \sum_{i=1}^{s} b_ic_i^k w_{ij} = \sum_{i=1}^{s} b_i w_{ij} - kb_jc_j^{k-1} \quad \text{for} \quad j=1,\dots,s\,, \quad k=1,\dots,r\,;$$

$$I\widehat{D}(\widehat{r}): \quad \sum_{i=1}^{s} b_ic_i^k \widehat{w}_{ij} = \sum_{i=1}^{s} b_i \widehat{w}_{ij} - kb_jc_j^{k-1} \quad \text{for} \quad j=1,\dots,s\,, \quad k=1,\dots,\widehat{r}\,;$$

$$IC\widehat{C}(Q): \quad \sum_{j=1}^{s}\sum_{l=1}^{s} \widehat{w}_{ij}w_{jl}c_l^k = k(k-1)c_i^{k-2} \quad \text{for} \quad i=1,\dots,s\,, \quad k=2,\dots,Q\,;$$

$$ID\widehat{D}(R): \quad \sum_{i=1}^{s}\sum_{j=1}^{s} b_ic_i^k\widehat{w}_{ij}w_{jl} = k\sum_{i=1}^{s}\sum_{j=1}^{s} b_ic_i\widehat{w}_{ij}w_{jl} - (k-1)\sum_{i=1}^{s}\sum_{j=1}^{s} b_i\widehat{w}_{ij}w_{jl} + k(k-1)b_lc_l^{k-2} \quad \text{for} \quad l=1,\dots,s\,, \quad k=2,\dots,R\,.$$

$IC\widehat{C}(Q)$ and $ID\widehat{D}(R)$ can be rewritten

$$IC\widehat{C}(Q): \quad \sum_{l=1}^{s} \omega_{il}c_l^k = k(k-1)c_i^{k-2} \quad \text{for} \quad i=1,\dots,s\,, \quad k=2,\dots,Q\,;$$

$$ID\widehat{D}(R): \quad \sum_{i=1}^{s} b_ic_i^k\omega_{il} = k\sum_{i=1}^{s} b_ic_i\omega_{il} - (k-1)\sum_{i=1}^{s} b_i\omega_{il} + k(k-1)b_lc_l^{k-2} \quad \text{for} \quad l=1,\dots,s\,, \quad k=2,\dots,R\,.$$

We recall that we are not interested in methods violating (1.2). The above simplifying assumptions will mainly be used to obtain optimal estimates concerning the local error of PRK methods, especially for those satisfying the condition (S) which are of special interest (see also Chapter IV and Chapter V). We give in Table 1.1 some properties of common examples of s-stage implicit Runge-Kutta methods.

Method	simplifying assumptions				classical order	other properties		
Gauss	$B(2s)$	$C(s)$	$D(s)$		$2s$			
Radau IA	$B(2s-1)$	$C(s-1)$	$D(s)$		$2s-1$	$c_1=0$		
Radau IIA	$B(2s-1)$	$C(s)$	$D(s-1)$	(S)	$2s-1$		$c_s=1$	
Lobatto IIIA	$B(2s-2)$	$C(s)$	$D(s-2)$	(S)	$2s-2$	$c_1=0$	$c_s=1$	$a_{1j}=0$
Lobatto IIIB	$B(2s-2)$	$C(s-2)$	$D(s)$		$2s-2$	$c_1=0$	$c_s=1$	$a_{is}=0$
Lobatto IIIC	$B(2s-2)$	$C(s-1)$	$D(s-1)$	(S)	$2s-2$	$c_1=0$	$c_s=1$	

Table 1.1. Fully implicit Runge-Kutta methods.

1.2. Computation for the u-component.

Several definitions of the numerical u-component u_1 are conceivable. Here we only describe the most natural choices. The first possibility is to choose u_1 such that (II.1.1e) is satisfied (see the next subsection), i.e., u_1 is the solution of

$$0 = \left(g_{yy}(f,f)+g_y f_y f+g_y f_z k\right)(y_1, z_1, u_1)\ . \tag{1.4}$$

For PRK methods which satisfy $c_s=1$ a second fairly good choice is often given by

$$u_1 := U_s\ . \tag{1.5}$$

Otherwise, as a third possibility, one may also formally define

$$U_i = u_0+h\sum_{j=1}^{s} a_{ij}U'_j, \qquad u_1 = u_0+h\sum_{i=1}^{s} b_i U'_i\ . \tag{1.6}$$

For PRK methods satisfying the assumption (S) this leads to (1.5). If the RK matrix A is invertible we get

$$u_1 = u_0+\sum_{i,j=1}^{s} b_i w_{ij}(U_j-u_0) = R(\infty)u_0+\sum_{i,j=1}^{s} b_i w_{ij}U_j \tag{1.7}$$

where $R(\infty)=1-b^TA^{-1}\mathbb{1}$ is the stability function at infinity of the RK method (A,b,c). Similarly a fourth possibility is to formally define

$$U_i = u_0+h\sum_{j=1}^{s} \widehat{a}_{ij}U'_j, \qquad u_1 = u_0+h\sum_{i=1}^{s} b_i U'_i \tag{1.8}$$

and if

$$(\widehat{S}): \qquad \widehat{a}_{si} = b_i \qquad \text{for} \quad i=1,\dots,s$$

we again obtain (1.5). If the RK matrix $\widehat{A}$ is invertible we have

$$u_1 = u_0+\sum_{i,j=1}^{s} b_i \widehat{w}_{ij}(U_j-u_0) = \widehat{R}(\infty)u_0+\sum_{i,j=1}^{s} b_i \widehat{w}_{ij}U_j \tag{1.9}$$

where $\widehat{R}(\infty)=1-b^T\widehat{A}^{-1}\mathbb{1}$ is the stability function at infinity of the RK method $(\widehat{A},b,c)$.

Remark 1.3. A very accurate value for u_1 is often unnecessary (see Remark 2.1.2). For a cheap computation one may discard the first possibility (1.4) and prefer the other ones, especially if one wants to avoid the computation of extra derivatives such as g_{yy}.

1.3. Construction of consistent values.

In general arbitrary initial values, as well as the numerical solution of a PRK method, do not satisfy the constraints (II.1.1c,d,e). They can be projected onto these constraints with the help of the procedures described below (see [HaWa91, Section VI.7,

p. 513] [AsPe91], [AsPe92a], and [Lu91b] for similar ideas in the context of semi-explicit index 2 problems in Hessenberg form).

We consider here the values (η, ζ, ν) satisfying

$$(g_y f_z k_u)(y, z, u) \text{ is invertible in a neighbourhood of } (\eta, \zeta, \nu) \tag{1.10}$$

and with $g(\eta)$, $(g_y f)(\eta, \zeta)$, and $(g_{yy}(f,f)+g_y f_y f+g_y f_z k)(\eta, \zeta, \nu)$ sufficiently small. We denote hereafter the corresponding projected values by $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$.

A first projection procedure.

Firstly we define $\widetilde{\eta}, \lambda$ as the solution of the system

$$0 = \widetilde{\eta} - \eta - (f_z k_u)(\eta, \zeta, \nu)\lambda \,, \qquad 0 = g(\widetilde{\eta}) \,. \tag{1.11a}$$

Then we define $\widetilde{\zeta}, \mu$ as the solution of the system

$$0 = \widetilde{\zeta} - \zeta - k_u(\eta, \zeta, \nu)\mu \,, \qquad 0 = (g_y f)(\widetilde{\eta}, \widetilde{\zeta}) \,. \tag{1.11b}$$

Finally we define $\widetilde{\nu}$ as the solution of the equation

$$0 = \left(g_{yy}(f,f)+g_y f_y f+g_y f_z k\right)(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu}) \,. \tag{1.11c}$$

If one wants to avoid an accurate computation of the derivatives f_z and k_u when projecting η and ζ, an alternative possibility is as follows:

A second projection procedure.

We define $\widetilde{\eta}, \lambda, \widetilde{\zeta}, \mu$, and $\widetilde{\nu}$ as the solutions of the systems

$$0 = \widetilde{\eta} - \eta - f(\eta, \zeta + k(\eta, \zeta, \nu+\lambda) - k(\eta, \zeta, \nu)) + f(\eta, \zeta) \,, \qquad 0 = g(\widetilde{\eta}) \,, \tag{1.12a}$$

$$0 = \widetilde{\zeta} - \zeta - k(\eta, \zeta, \nu+\mu) + k(\eta, \zeta, \nu) \,, \qquad 0 = (g_y f)(\widetilde{\eta}, \widetilde{\zeta}) \,, \tag{1.12b}$$

$$0 = \left(g_{yy}(f,f)+g_y f_y f+g_y f_z k\right)(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu}) \,. \tag{1.12c}$$

Another possibility whose formulation is more theoretical than practical is as follows:

A third projection procedure (see [HaLuRo89a, p. 103]).

We define $\widetilde{\eta}, \lambda, \widetilde{\zeta}, \mu$, and $\widetilde{\nu}$ as the solutions of the systems

$$0 = P_y(\eta, \zeta, \nu)(\widetilde{\eta} - \eta) \,, \qquad 0 = g(\widetilde{\eta}) \,, \tag{1.13a}$$

$$0 = P_z(\eta, \zeta, \nu)(\widetilde{\zeta} - \zeta) \,, \qquad 0 = (g_y f)(\widetilde{\eta}, \widetilde{\zeta}) \,, \tag{1.13b}$$

$$0 = \left(g_{yy}(f,f)+g_y f_y f+g_y f_z k\right)(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu}) \tag{1.13c}$$

where P_y, Q_y, P_z, and Q_z are *projectors* defined under the condition (II.1.2) by

$$\begin{aligned} S &:= k_u(g_y f_z k_u)^{-1} g_y \,, \\ Q_y &:= f_z S \,, \qquad P_y := I - Q_y \,, \qquad Q_z := S f_z \,, \qquad P_z := I - Q_z \,. \end{aligned} \tag{1.14}$$

Remark 1.4. In fact this projection procedure is equivalent to the first projection procedure, because it is easy to show that

$$0 = P_y(\eta,\zeta,\nu)\delta_y \quad \Longleftrightarrow \quad \exists\, \lambda \text{ such that } \delta_y = (f_z k_u)(\eta,\zeta,\nu)\lambda \,, \tag{1.15}$$

$$0 = P_z(\eta,\zeta,\nu)\delta_z \quad \Longleftrightarrow \quad \exists\, \mu \text{ such that } \delta_z = k_u(\eta,\zeta,\nu)\mu \,. \tag{1.16}$$

The existence and uniqueness of a solution to the above systems follows, e.g., from the theorem of Newton-Kantorovich (see [OrRh70, p. 421]). For example the Jacobian of (1.11a) evaluated at $\widetilde{\eta}=\eta$, $\lambda=0$ is given by

$$\begin{pmatrix} I & -(f_z k_u)(\eta,\zeta,\nu) \\ g_y(\eta) & 0 \end{pmatrix} \tag{1.17}$$

and is inversible by (1.10). Newton-type iterations with starting values $\widetilde{\eta}^{(0)}=\eta$, $\lambda^{(0)}=0$ will converge to the solution.

In order to construct consistent values $(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$ independently of ν, we can first define $\widehat{\nu}$ as the solution of

$$0 = \left(g_{yy}(f,f) + g_y f_y f + g_y f_z k\right)(\eta,\zeta,\widehat{\nu}) \tag{1.18}$$

and then use this value in the projection procedures.

One may also effect the above projections by replacing the arguments η,ζ,ν by $\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu}$. For example, instead of (1.11) one may define

$$0 = \widetilde{\eta} - \eta - (f_z k_u)(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})\lambda \,, \qquad 0 = g(\widetilde{\eta}) \,, \tag{1.19a}$$

$$0 = \widetilde{\zeta} - \zeta - k_u(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})\mu \,, \qquad 0 = (g_y f)(\widetilde{\eta},\widetilde{\zeta}) \,, \tag{1.19b}$$

$$0 = \left(g_{yy}(f,f) + g_y f_y f + g_y f_z k\right)(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu}) \,. \tag{1.19c}$$

For PRK methods the above projections are recommended, because they allow to stabilize the numerical solution as regards the influence of perturbations (see Section 2, Section IV.3, and Section V.4). If the assumption (S) holds then the constraint $g(y)=0$ is satisfied (see Remark 1.1.2: $\widetilde{\eta}=\eta=y_1$, $\lambda=0$ in (1.11a), (1.12a), or (1.13a)).

2. Existence, uniqueness of the PRK solution, and influence of perturbations.

This section is devoted to the analysis of the solution of the nonlinear system (1.1) with $\Psi_0 = (y_0, z_0, u_0)$ replaced by approximate h-dependent starting values $(\eta, \zeta, \nu) = (\eta(h), \zeta(h), \nu(h))$. An important result is given by Theorem 2.4 which will be useful in Section 5 to the study of the error propagation. We first investigate the existence and uniqueness of the PRK solution.

2.1. Existence and uniqueness.

Theorem 2.1. *Let us suppose that the PRK coefficients satisfy $C(2)$, $C\widehat{C}(2)$, and that the matrix $\overline{A} = A\widehat{A}$ is invertible. If we assume that*

$$g(\eta) = \mathcal{O}(h^\tau) \ , \quad \tau \geq 3 \ , \tag{2.1a}$$

$$(g_y f)(\eta, \zeta) = \mathcal{O}(h^\kappa) \ , \quad \kappa \geq 2 \ , \tag{2.1b}$$

$$(g_{yy}(f,f) + g_y f_y f + g_y f_z k)(\eta, \zeta, \nu) = \mathcal{O}(h) \ , \tag{2.1c}$$

$$(g_y f_z k_u)(y, z, u) \text{ is invertible in a neighbourhood of } (\eta, \zeta, \nu) \ , \tag{2.1d}$$

then for $h \leq h_0$ there exists a locally unique solution to

$$Y_i = \eta + h \sum_{j=1}^{s} a_{ij} f(Y_j, Z_j) \ , \qquad Z_i = \zeta + h \sum_{j=1}^{s} \widehat{a}_{ij} k(Y_j, Z_j, U_j) \ , \tag{2.2a,b}$$

$$0 = g(Y_i) \tag{2.2c}$$

for $i = 1, \ldots, s$, which satisfies

$$Y_i - \eta = \mathcal{O}(h) \ , \qquad Z_i - \zeta = \mathcal{O}(h) \ , \qquad U_i - \nu = \mathcal{O}(h) \ . \tag{2.3}$$

Remarks 2.1.

1) If the function k of (II.1.1) is linear in u then the assumptions $C(2)$, $C\widehat{C}(2)$, and (2.1c) can be omitted. In this situation, $\tau \geq 2$ and $\kappa \geq 1$ are sufficient. However, if $C(2)$ or $C\widehat{C}(2)$ is not satisfied, $\tau = 2$ or $\kappa = 1$, we only have the estimate $U_i - \nu = \mathcal{O}(1)$.
2) The value of ν in (2.1c) only prescribes the solution of (2.2) to be close to the manifold defined by (II.1.1.e). However, (Y_i, Z_i, U_i) are clearly independent of ν.
3) If the function k of (II.1.1) is not linear in u then $C(2)$ and the invertibility of the matrix A show the necessity of having $s \geq 2$.

Proof. A way of proving this theorem is by homotopy (see [HaLuRo89a, Theorem 6.1]). Here we present a short proof inspired by [HaWa91, Theorem VI.7.1].

We first develop $0 = g(Y_i)$ and $f(Y_i, Z_i)$ into Taylor-series

$$0 = g(Y_i) = g(\eta) + g_y(\eta)(Y_i - \eta) + \int_0^1 (1-\tau)\, g_{yy}\Big(\eta + \tau(Y_i - \eta)\Big) d\tau \cdot (Y_i - \eta \ , Y_i - \eta) \ , \tag{2.4}$$

$$f(Y_i, Z_i) = f(\eta, \zeta) + \int_0^1 f_y\Big(\eta + \tau(Y_i - \eta), \zeta + \tau(Z_i - \zeta)\Big) d\tau \cdot (Y_i - \eta) + \tag{2.5}$$

$$\int_0^1 f_z\Big(\eta + \tau(Y_i - \eta), \zeta + \tau(Z_i - \zeta)\Big) d\tau \cdot (Z_i - \zeta) \ .$$

Replacing the factors $Y_i-\eta$ and $Z_i-\zeta$ with the help of (2.2a,b), we obtain

$$0 = g(Y_i) = g(\eta)+h\sum_{j=1}^{s} a_{ij} g_y(\eta) f(Y_j,Z_j)+ \tag{2.6}$$
$$\frac{h^2}{2}\sum_{j,k=1}^{s} a_{ij}a_{ik}\int_0^1 g_{yy}\Big(\eta+\tau(Y_i-\eta)\Big)d\tau\cdot\Big(f(Y_j,Z_j)\,,f(Y_k,Z_k)\Big)\,,$$

$$f(Y_i,Z_i) = f(\eta,\zeta)+h\sum_{j=1}^{s} a_{ij}\int_0^1 f_y\Big(\eta+\tau(Y_i-\eta),\zeta+\tau(Z_i-\zeta)\Big)d\tau\cdot f(Y_j,Z_j)+ \tag{2.7}$$
$$h\sum_{j=1}^{s}\widehat{a}_{ij}\int_0^1 f_z\Big(\eta+\tau(Y_i-\eta),\zeta+\tau(Z_i-\zeta)\Big)d\tau\cdot k(Y_j,Z_j,U_j)\,.$$

Inserting (2.7) into the second term of (2.6) and dividing the result by h^2 we get the system

$$0 = Y_i-\eta-h\sum_{j=1}^{s} a_{ij} f(Y_j,Z_j)\,, \tag{2.8a}$$

$$0 = Z_i-\zeta-h\sum_{j=1}^{s}\widehat{a}_{ij} k(Y_j,Z_j,U_j)\,, \tag{2.8b}$$

$$0 = \frac{1}{h^2}g(\eta)+\frac{1}{h}\sum_{j=1}^{s} a_{ij}(g_y f)(\eta,\zeta)+ \tag{2.8c}$$
$$\sum_{j,k=1}^{s} a_{ij}a_{jk}g_y(\eta)\int_0^1 f_y\Big(\eta+\tau(Y_j-\eta),\zeta+\tau(Z_j-\zeta)\Big)d\tau\cdot f(Y_k,Z_k)+$$
$$\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk}g_y(\eta)\int_0^1 f_z\Big(\eta+\tau(Y_j-\eta),\zeta+\tau(Z_j-\zeta)\Big)d\tau\cdot k(Y_k,Z_k,U_k)+$$
$$\frac{1}{2}\sum_{j,k=1}^{s} a_{ij}a_{ik}\int_0^1 g_{yy}\Big(\eta+\tau(Y_i-\eta)\Big)d\tau\cdot\Big(f(Y_j,Z_j)\,,f(Y_k,Z_k)\Big)\,.$$

Taking (2.1a,b) into account, for $h=0$ the right-hand side of (2.8c) with values $Y_i=\eta(0)$, $Z_i=\zeta(0)$, and $U_i=\nu(0)$ reads

$$\frac{1}{2}\sum_{j,k=1}^{s} a_{ij}a_{ik}(g_{yy}(f,f))(\eta(0),\zeta(0))+\sum_{j,k=1}^{s} a_{ij}a_{jk}(g_y f_y f)(\eta(0),\zeta(0))+ \tag{2.9}$$
$$\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk}(g_y f_z k)(\eta(0),\zeta(0),\nu(0))\,.$$

By $C(2)$ and $C\widehat{C}(2)$ this expression is equal to

$$\frac{c_i^2}{2}\left(g_{yy}(f,f)+g_y f_y f+g_y f_z k\right)(\eta(0),\zeta(0),\nu(0)) \tag{2.10}$$

which vanishes by (2.1c). Hence the values $Y_i = \eta(0)$, $Z_i = \zeta(0)$, and $U_i = \nu(0)$ satisfy (2.8) at $h=0$. Further, the derivative of (2.8) with respect to (Y_i, Z_i, U_i) is of the form

$$\begin{pmatrix} I+\mathcal{O}(h) & \mathcal{O}(h) & 0 \\ \mathcal{O}(h) & I+\mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \mathcal{O}(1) & A\hat{A} \otimes (g_y f_z k_u)(\eta,\zeta,\nu)+\mathcal{O}(h) \end{pmatrix} \tag{2.11}$$

which is invertible for $h \le h_0$. Therefore the implicit function theorem (see [OrRh70, p. 128]) yields the existence of a locally unique solution to (2.8) and hence to (2.2).

If the function k of (II.1.1) is linear in u then we can directly define the values U_i at $h=0$ such that the right-hand side of (2.8c) vanishes. □

2.2. Influence of perturbations.

The next result is concerned with the influence of perturbations to (2.2). In practice these perturbations may come from, e.g., round-off errors and errors in the iterative solution of the nonlinear system (2.2).

Theorem 2.2. *Let (Y_i, Z_i, U_i) be given by (2.2) and let us consider perturbed values $(\widehat{Y}_i, \widehat{Z}_i, \widehat{U}_i)$ satisfying*

$$\widehat{Y}_i = \widehat{\eta}+h\sum_{j=1}^{s} a_{ij} f(\widehat{Y}_j, \widehat{Z}_j)+h\delta_i\,, \qquad \widehat{Z}_i = \widehat{\zeta}+h\sum_{j=1}^{s} \hat{a}_{ij} k(\widehat{Y}_j, \widehat{Z}_j, \widehat{U}_j)+h\mu_i\,, \tag{2.12a,b}$$

$$0 = g(\widehat{Y}_i)+\theta_i \tag{2.12c}$$

for $i = 1, \dots, s$. In addition to the assumptions of Theorem 2.1 let us suppose that

hyp. qui pourraient être relaxée si on ne considère que (I.14a,c) par exemple

$$\begin{cases} \Delta\eta = \mathcal{O}(h^3)\,, & \Delta\zeta = \mathcal{O}(h^2)\,, & \widehat{U}_i - \nu = \mathcal{O}(h)\,, \\ \delta_i = \mathcal{O}(h^2)\,, & \mu_i = \mathcal{O}(h)\,, & \theta_i = \mathcal{O}(h^3)\,. \end{cases} \tag{2.13}$$

Then we have for $h \le h_0$ the estimates

$$\Delta Y_i = P_y\Delta\eta+hc_i f_z P_z\Delta\zeta+\mathcal{O}\Big(h\|\Delta\eta\|+h^2\|\Delta\zeta\|+\frac{1}{h^2}\|Q_y\Delta\eta\|^2+\|Q_z\Delta\zeta\|^2 \tag{2.14a}$$
$$+h\|\delta\|+h^2\|\mu\|+\|\theta\|\Big)\,,$$

$$\Delta Z_i = -\frac{1}{h}\sigma_i \cdot SQ_y\Delta\eta+P_z\Delta\zeta+\mathcal{O}\Big(\|\Delta\eta\|+h\|\Delta\zeta\|+\frac{1}{h^3}\|Q_y\Delta\eta\|^2+\frac{1}{h}\|Q_z\Delta\zeta\|^2 \tag{2.14b}$$
$$+\|\delta\|+h\|\mu\|+\frac{1}{h}\|\theta\|\Big)\,,$$

$$P_{z,i}\Delta Z_i = P_z\Delta\zeta+\mathcal{O}\Big(\|Q_y\Delta\eta\|+h\|P_y\Delta\eta\|+h\|\Delta\zeta\|+\frac{1}{h^3}\|Q_y\Delta\eta\|^2+\frac{1}{h}\|Q_z\Delta\zeta\|^2 \tag{2.14c}$$
$$+h\|\delta\|+h\|\mu\|+\|\theta\|\Big)\,,$$

$$\Delta U_i = \mathcal{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|+\frac{1}{h}\|P_y\Delta\eta\|+\frac{1}{h}\|Q_z\Delta\zeta\|+\|P_z\Delta\zeta\|+\frac{1}{h}\|\delta\|+\|\mu\|+\frac{1}{h^2}\|\theta\|\Big)\,, \tag{2.14d}$$

where $\sigma_i = e_i^T A^{-1}\mathbb{1}$ *with* $e_i = (0,\dots,0,1,0,\dots,0)^T$ *(the s-dimensional vector with all components equal to 0 excepted the ith which is equal to 1),* $\delta = (\delta_1,\dots,\delta_s)^T$, $\|\delta\| = \max_i \|\delta_i\|$, *and similarly for* μ *and* θ. P_y, Q_y, P_z, *and* Q_z *are the projectors defined under the condition (II.1.2) by* (1.14).

Remarks 2.2.

1) We have used the notation $\Delta\eta = \widehat{\eta} - \eta$, $\Delta\zeta = \widehat{\zeta} - \zeta$, $Y = (Y_1,\dots,Y_s)^T$, $\Delta Y = \widehat{Y} - Y$, $\|\Delta Y\| = \max_i \|\Delta Y_i\|$, and similarly for the z- and u-components.
2) The missing arguments for f_z, S, P_y, Q_z, etc., are (η,ζ,ν) or $(\eta,\zeta,G(\eta,\zeta))$ with G as described in (II.1.4). Those of $P_{z,i}$ are $(Y_i,Z_i,G(Y_i,Z_i))$ or (Y_i,Z_i,U_i).
3) The conditions (2.13) ensure that all $\mathcal{O}$-terms in the proof below are small.
4) If $g(\widehat{\eta}) = 0 = g(\eta)$ then $Q_y\Delta\eta = \mathcal{O}(\|\Delta\eta\|^2)$. Consequently, this term may be neglected and the hypothesis $\Delta\eta = \mathcal{O}(h^3)$ can be relaxed to $\mathcal{O}(h^2)$. If we have $(g_y f)(\widehat{\eta},\widehat{\zeta}) = 0 = (g_y f)(\eta,\zeta)$ then similarly $\Delta\zeta = \mathcal{O}(h)$ suffices.
5) If the function k of (II.1.1) is linear in u then the terms $\|Q_y\Delta\eta\|^2$ and $\|Q_z\Delta\zeta\|^2$ in (2.13a,b,c) are multiplied by one additional factor h. In this case $\Delta\eta = \mathcal{O}(h^2)$, $\Delta\zeta = \mathcal{O}(h)$, $\widehat{U}_i - \nu = \mathcal{O}(1)$, $\delta_i = \mathcal{O}(h)$, $\mu_i = \mathcal{O}(1)$, and $\theta_i = \mathcal{O}(h^2)$ are sufficient, but then we only have the estimate $\Delta U_i = \mathcal{O}(1)$.
6) The constants implied by the $\mathcal{O}$-terms in (2.14) depend on bounds for certain derivatives of f, g, and k, but not on the constants entering in the $\mathcal{O}$-terms in (2.1a,b) and (2.13), if h is sufficiently small.
7) It can be observed that the terms $\|\theta\|/h^2$, $\|\theta\|/h$, $\|\delta\|/h$, $\|\delta\|$, and $\|\mu\|$ are not present in (2.14a,c).

Proof. Subtracting (2.2) from (2.12) we obtain by linearization

$$\Delta Y_i = \Delta\eta + h\sum_{j=1}^{s} a_{ij} f_y(Y_j,Z_j)\Delta Y_j + h\sum_{j=1}^{s} a_{ij} f_z(Y_j,Z_j)\Delta Z_j + h\delta_i + \mathcal{O}(h\|\Delta Y\|^2 + h\|\Delta Z\|^2)\,, \tag{2.15a}$$

$$\Delta Z_i = \Delta\zeta + h\sum_{j=1}^{s} \widehat{a}_{ij} k_y(Y_j,Z_j,U_j)\Delta Y_j + h\sum_{j=1}^{s} \widehat{a}_{ij} k_z(Y_j,Z_j,U_j)\Delta Z_j + h\sum_{j=1}^{s} \widehat{a}_{ij} k_u(Y_j,Z_j,U_j)\Delta U_j + h\mu_i + \mathcal{O}(h\|\Delta Y\|^2 + h\|\Delta Z\|^2 + h\|\Delta U\|^2)\,, \tag{2.15b}$$

$$0 = g_y(Y_i)\Delta Y_i + \theta_i + \mathcal{O}(\|\Delta Y_i\|^2)\,, \tag{2.15c}$$

which can be rewritten, using tensor notation,

$$\Delta Y = \mathbb{1}\otimes\Delta\eta + h(A\otimes I)\{f_y\}\Delta Y + (A\otimes I)\{f_z\}h\Delta Z + h\delta + \mathcal{O}(h\|\Delta Y\|^2 + h\|\Delta Z\|^2)\,, \tag{2.16a}$$

$$h\Delta Z = \mathbb{1}\otimes h\Delta\zeta + h^2(\widehat{A}\otimes I)\{k_y\}\Delta Y + h(\widehat{A}\otimes I)\{k_z\}h\Delta Z + (\widehat{A}\otimes I)\{k_u\}h^2\Delta U + h^2\mu + \mathcal{O}(h^2\|\Delta Y\|^2 + h^2\|\Delta Z\|^2 + h^2\|\Delta U\|^2)\,, \tag{2.16b}$$

$$0 = \{g_y\}\Delta Y + \theta + \mathcal{O}(\|\Delta Y\|^2) \tag{2.16c}$$

where

$$\{g_y\} := \text{blockdiag}\left(g_y(Y_1),\dots,g_y(Y_s)\right) , \tag{2.17a}$$

$$\{f_z\} := \text{blockdiag}\left(f_z(Y_1,Z_1),\dots,f_z(Y_s,Z_s)\right) , \tag{2.17b}$$

$$\{k_u\} := \text{blockdiag}\left(k_u(Y_1,Z_1,U_1),\dots,k_u(Y_s,Z_s,U_s)\right) , \tag{2.17c}$$

and similarly for $\{f_y\}$, $\{k_y\}$ and $\{k_z\}$. Inserting (2.16a) into (2.16c) and (2.16b) into the resulting formula leads to

$$\begin{aligned} -\{g_y\}(A\otimes I)\{f_z\}(\hat{A}\otimes I)\{k_u\}h^2\Delta U = \qquad\qquad (2.18)\\ \{g_y\}\Big[\mathbb{1}\otimes\Delta\eta+(A\otimes I)\{f_z\}(\mathbb{1}\otimes h\Delta\zeta)+h(A\otimes I)\{f_y\}\Delta Y+\\ h^2(A\otimes I)\{f_z\}(\hat{A}\otimes I)\{k_y\}\Delta Y+h(A\otimes I)\{f_z\}(\hat{A}\otimes I)\{k_z\}h\Delta Z+\\ (A\otimes I)\{f_z\}h^2\mu+h\delta\Big]+\theta+\mathcal{O}(\|\Delta Y\|^2+h\|\Delta Z\|^2+h^2\|\Delta U\|^2) . \end{aligned}$$

In accordance with (2.3) we have

$$g_y(Y_i)\,a_{ij}f_z(Y_j,Z_j)\hat{a}_{jk}k_u(Y_k,Z_k,U_k) = a_{ij}\hat{a}_{jk}(g_yf_zk_u)(\eta,\zeta,\nu)+\mathcal{O}(h) , \tag{2.19}$$

thus, the left matrix of (2.19) can be written as

$$\{g_y\}(A\otimes I)\{f_z\}(\hat{A}\otimes I)\{k_u\} = A\hat{A}\otimes(g_yf_zk_u)(\eta,\zeta,\nu)+\mathcal{O}(h) \tag{2.20}$$

and is invertible for h sufficiently small. Putting

$$\begin{aligned} &\tilde{G}_y := \{g_y\} , \quad \tilde{F}_z := (A\otimes I)\{f_z\}(A\otimes I)^{-1} , \quad \tilde{K}_u := (A\hat{A}\otimes I)\{k_u\}(A\hat{A}\otimes I)^{-1} ,\\ &\tilde{S} := \tilde{K}_u(\tilde{G}_y\tilde{F}_z\tilde{K}_u)^{-1}\tilde{G}_y , \qquad\qquad (2.21)\\ &\tilde{Q}_y := \tilde{F}_z\tilde{S} , \quad \tilde{P}_y := I-\tilde{Q}_y , \quad \tilde{Q}_z := \tilde{S}\tilde{F}_z , \quad \tilde{P}_z := I-\tilde{Q}_z , \end{aligned}$$

we remark for example that $\tilde{G}_y = \tilde{G}_y\tilde{Q}_y$, $\tilde{S} = \tilde{S}\tilde{Q}_y$, $\tilde{G}_y\tilde{F}_z = \tilde{G}_y\tilde{F}_z\tilde{Q}_z$, and $\tilde{F}_z\tilde{P}_z = \tilde{P}_y\tilde{F}_z$. Hence, from (2.18) and (2.16a,b) we obtain

$$\begin{aligned} h^2\Delta U = -(A\hat{A}\otimes I)^{-1}(\tilde{G}_y\tilde{F}_z\tilde{K}_u)^{-1}\tilde{G}_y\Big[\mathbb{1}\otimes\Delta\eta+\tilde{F}_z(A\mathbb{1}\otimes h\Delta\zeta)+h(A\otimes I)\{f_y\}\Delta Y+\\ h^2\tilde{F}_z(A\hat{A}\otimes I)\{k_y\}\Delta Y+h\tilde{F}_z(A\hat{A}\otimes I)\{k_z\}h\Delta Z\Big]+ \qquad (2.22a)\\ \mathcal{O}(\|\Delta Y\|^2+h\|\Delta Z\|^2+h^2\|\Delta U\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|) , \end{aligned}$$

$$\begin{aligned} h\Delta Z = (A\otimes I)^{-1}\Big[-\tilde{S}(\mathbb{1}\otimes\Delta\eta)+\tilde{P}_z(A\mathbb{1}\otimes h\Delta\zeta)-h\tilde{S}(A\otimes I)\{f_y\}\Delta Y+ \qquad (2.22b)\\ h^2\tilde{P}_z(A\hat{A}\otimes I)\{k_y\}\Delta Y+h\tilde{P}_z(A\hat{A}\otimes I)\{k_z\}h\Delta Z\Big]+\\ \mathcal{O}(\|\Delta Y\|^2+h\|\Delta Z\|^2+h^2\|\Delta U\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|) , \end{aligned}$$

$$\begin{aligned} \Delta Y = \tilde{P}_y(\mathbb{1}\otimes\Delta\eta)+\tilde{F}_z\tilde{P}_z(A\mathbb{1}\otimes h\Delta\zeta)+h\tilde{P}_y(A\otimes I)\{f_y\}\Delta Y+ \qquad (2.22c)\\ h^2\tilde{P}_y\tilde{F}_z(A\hat{A}\otimes I)\{k_y\}\Delta Y+h\tilde{P}_y\tilde{F}_z(A\hat{A}\otimes I)\{k_z\}h\Delta Z+\\ \mathcal{O}(\|\Delta Y\|^2+h\|\Delta Z\|^2+h^2\|\Delta U\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|) . \end{aligned}$$

The $\mathcal{O}$-terms in (2.22a,b,c) lead to the estimate $\mathcal{O}(\|Q_y\Delta\eta\|^2/h^2+\|P_y\Delta\eta\|^2+\|Q_z\Delta\zeta\|^2+h\|P_z\Delta\zeta\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|)$. If the function k of (II.1.1) is linear in u, they can be replaced by $\mathcal{O}(\|\Delta Y\|^2+h\|\Delta Z\|^2+h^2\|\Delta Y\|\cdot\|\Delta U\|+h^2\|\Delta Z\|\cdot\|\Delta U\|+h\|\delta\|+h^2\|\mu\|+\|\theta\|)$ yielding $\mathcal{O}(\|Q_y\Delta\eta\|^2/h+\|P_y\Delta\eta\|^2+h\|Q_z\Delta\zeta\|^2+h\|P_z\Delta\zeta\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|)$. The term $P_y\Delta\eta$ entering in $P_{z,i}\Delta Z_i$ leads to the estimate $\mathcal{O}(h\|P_y\Delta\eta\|)$, because of $(I\otimes P_{z,i})\widetilde{S}(\mathbb{1}\otimes P_y)=\mathcal{O}(h^2)$ which is a consequence of $P_z k_u\equiv 0$ and $g_y P_y\equiv 0$. The estimate (2.14c) for the perturbations δ and θ simply follows from (2.14a,b,d), (2.22b), and $P_z k_u\equiv 0$. □

We are now interested in the influence of perturbations to the numerical solution.

Theorem 2.3. *In addition to the assumptions of Theorem 2.2, including those of Theorem 2.1, let y_1, z_1 and $\widehat{y}_1, \widehat{z}_1$ be given by*

$$y_1 = \eta + h\sum_{i=1}^{s} b_i f(Y_i, Z_i)\ , \qquad \widehat{y}_1 = \widehat{\eta} + h\sum_{i=1}^{s} b_i f(\widehat{Y}_i, \widehat{Z}_i) + h\delta_{s+1}\ , \tag{2.23a,b}$$

$$z_1 = \zeta + h\sum_{i=1}^{s} b_i k(Y_i, Z_i, U_i)\ , \qquad \widehat{z}_1 = \widehat{\zeta} + h\sum_{i=1}^{s} b_i k(\widehat{Y}_i, \widehat{Z}_i, \widehat{U}_i) + h\mu_{s+1} \tag{2.23c,d}$$

with

$$\delta_{s+1} = \mathcal{O}(h^2)\ , \qquad \mu_{s+1} = \mathcal{O}(h)\ . \tag{2.24}$$

Then we have the estimates

$$\Delta y_1 = P_y\Delta\eta + R(\infty)Q_y\Delta\eta + hf_z P_z\Delta\zeta + \mathcal{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\delta_{s+1}\| + h^2\|\mu\| + \|\theta\|\Big)\ , \tag{2.25a}$$

$$P_y^1\Delta y_1 = P_y\Delta\eta + hf_z P_z\Delta\zeta + \mathcal{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\delta_{s+1}\| + h^2\|\mu\| + \|\theta\|\Big)\ , \tag{2.25b}$$

$$\Delta z_1 = -\frac{1}{h}\sigma\cdot SQ_y\Delta\eta + P_z\Delta\zeta + \widehat{R}(\infty)Q_z\Delta\zeta + \mathcal{O}\Big(\|\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + \|\delta\| + h\|\mu\| + h\|\mu_{s+1}\| + \frac{1}{h}\|\theta\|\Big)\ , \tag{2.25c}$$

$$P_z^1\Delta z_1 = P_z\Delta\zeta + \mathcal{O}\Big(\|Q_y\Delta\eta\| + h\|P_y\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\mu\| + h\|\mu_{s+1}\| + \|\theta\|\Big)\ , \tag{2.25d}$$

where $\Delta y_1 = \widehat{y}_1 - y_1$, $\Delta z_1 = \widehat{z}_1 - z_1$, $R(\infty) = 1 - b^T A^{-1}\mathbb{1}$, $\widehat{R}(\infty) = 1 - b^T\widehat{A}^{-1}\mathbb{1}$, and $\sigma = b^T\widehat{A}^{-1}A^{-1}\mathbb{1}$. The arguments of P_y^1 and P_z^1 are given by (y_1, z_1, u_1) with u_1 defined by one of the alternatives described in Subsection 1.2.

Remark 2.3. Similar remarks to those in Remarks 2.2 hold.

Proof. The results of this theorem are a simple consequence of

$$\Delta y_1 = (1 - b^T A^{-1}\mathbb{1})\Delta\eta + (b^T A^{-1}\otimes I)\Delta Y + h\delta_{s+1}\ , \tag{2.26a}$$

$$\Delta z_1 = (1 - b^T\widehat{A}^{-1}\mathbb{1})\Delta\zeta + (b^T\widehat{A}^{-1}\otimes I)\Delta Z + h\mu_{s+1}\ , \tag{2.26b}$$

and (2.14a,b). □

Finally we want to show that better estimates hold for the projected values.

Theorem 2.4. *Additionnally to the assumptions of Theorem 2.3, including those of Theorem 2.2 and Theorem 2.1, let us assume that $B(2)$ holds and let us consider projected values $\widetilde{y}_1, \widetilde{z}_1, \widehat{\widetilde{y}}_1, \widehat{\widetilde{z}}_1, \widetilde{u}_1, \widehat{\widetilde{u}}_1$ satisfying*

$$0 = \widetilde{y}_1 - y_1 - (f_z k_u)(y_1, z_1, u_1)\lambda\,, \qquad 0 = g(\widetilde{y}_1)\,, \tag{2.27a}$$

$$0 = \widehat{\widetilde{y}}_1 - \widehat{y}_1 - (f_z k_u)(\widehat{y}_1, \widehat{z}_1, \widehat{u}_1)\widehat{\lambda}\,, \qquad 0 = g(\widehat{\widetilde{y}}_1) + \theta_{s+1}\,, \tag{2.27b}$$

$$0 = \widetilde{z}_1 - z_1 - k_u(y_1, z_1, u_1)\mu\,, \qquad 0 = (g_y f)(\widetilde{y}_1, \widetilde{z}_1)\,, \tag{2.27c}$$

$$0 = \widehat{\widetilde{z}}_1 - \widehat{z}_1 - k_u(\widehat{y}_1, \widehat{z}_1, \widehat{u}_1)\widehat{\mu}\,, \qquad 0 = (g_y f)(\widehat{\widetilde{y}}_1, \widehat{\widetilde{z}}_1) + \theta'_{s+1}\,, \tag{2.27d}$$

$$0 = \big(g_{yy}(f,f) + g_y f_y f + g_y f_z k\big)(\widetilde{y}_1, \widetilde{z}_1, \widetilde{u}_1)\,, \tag{2.27e}$$

$$0 = \big(g_{yy}(f,f) + g_y f_y f + g_y f_z k\big)(\widehat{\widetilde{y}}_1, \widehat{\widetilde{z}}_1, \widehat{\widetilde{u}}_1) + \theta''_{s+1} \tag{2.27f}$$

with

$$\theta_{s+1} = \mathcal{O}(h^3)\,, \qquad \theta'_{s+1} = \mathcal{O}(h^2)\,, \qquad \theta''_{s+1} = \mathcal{O}(h)\,. \tag{2.28}$$

Then we have the estimates

$$\Delta\widetilde{y}_1 = P_y\Delta\eta + h f_z P_z\Delta\zeta + \mathcal{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\delta_{s+1}\| + h^2\|\mu\| + \|\theta\| + \|\theta_{s+1}\|\Big)\,, \tag{2.29a}$$

$$\Delta\widetilde{z}_1 = P_z\Delta\zeta + \mathcal{O}\Big(\|Q_y\Delta\eta\| + h\|P_y\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\delta_{s+1}\| + h\|\mu\| + h\|\mu_{s+1}\| + \|\theta\| + \|\theta_{s+1}\| + \|\theta'_{s+1}\|\Big)\,, \tag{2.29b}$$

$$\Delta\widetilde{u}_1 = \mathcal{O}\Big(\|\Delta\eta\| + \|P_z\Delta\zeta\| + h\|Q_z\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\delta_{s+1}\| + h\|\mu\| + h\|\mu_{s+1}\| + \|\theta\| + \|\theta_{s+1}\| + \|\theta'_{s+1}\| + \|\theta''_{s+1}\|\Big) \tag{2.29c}$$

where $\Delta\widetilde{y}_1 = \widehat{\widetilde{y}}_1 - \widetilde{y}_1$, $\Delta\widetilde{z}_1 = \widehat{\widetilde{z}}_1 - \widetilde{z}_1$, $\Delta\widetilde{u}_1 = \widehat{\widetilde{u}}_1 - \widetilde{u}_1$, and $\widehat{u}_1$ is defined similarly to u_1.

Remarks 2.4.

1) A crucial observation is that the terms $\|\theta\|/h^2$, $\|\theta\|/h$, $\|\delta\|/h$, $\|\delta\|$, and $\|\mu\|$ are not present. The effect of the projections is to stabilize the numerical solution regarding the influence of perturbations. The problem (II.1.1) is not ill-posed, because the constraints (II.1.1d,e) are taken into account.
2) The above projections (2.27) correspond to the first projection procedure (1.11) described in Subsection 1.3 which is also equivalent to the third projection procedure (1.13) (see Remark 1.4). Concerning the second projection procedure (1.12), identical results hold (see the comment at the end of the proof below).
3) If the function k of (II.1.1) is not linear in u and if $\Delta\nu = \widehat{\nu} - \nu$ enters in the estimate of Δu_1, then the $\mathcal{O}$-terms $\mathcal{O}(h^3\|\Delta\nu\|)$ and $\mathcal{O}(h^2\|\Delta\nu\|)$ must to be added to (2.29a) and (2.29b,c) respectively. This happens for example if for the computation of the u-component the alternative (1.7) has been chosen with $R(\infty) \neq 0$.

4) Identical results hold for $\Delta\widetilde{z}_1$ and $\Delta\widetilde{u}_1$ even if the y-component is not projected onto $g(y)=0$.

5) If the function k of (II.1.1) is linear in u then the assumption $B(2)$ can be omitted.

6) Similar remarks to those in Remarks 2.2 hold.

Proof. We begin with the estimate (2.28a). Analogously to the proof of Theorem 2.2 it is easy to show that

$$\begin{aligned}\frac{1}{h^2}g(y_1) = &\frac{1}{h^2}g(\eta)+\frac{1}{h}\sum_{j=1}^{s} b_j(g_yf)(\eta,\zeta)+ \\ &\sum_{j,k=1}^{s} b_j a_{jk} g_y(\eta)\int_0^1 f_y\Big(\eta+\tau(Y_j-\eta),\zeta+\tau(Z_j-\zeta)\Big)d\tau\cdot f(Y_k,Z_k)+ \\ &\sum_{j,k=1}^{s} b_j \widehat{a}_{jk} g_y(\eta)\int_0^1 f_z\Big(\eta+\tau(Y_j-\eta),\zeta+\tau(Z_j-\zeta)\Big)d\tau\cdot k(Y_k,Z_k,U_k)+ \\ &\frac{1}{2}\sum_{j,k=1}^{s} b_j b_k\int_0^1 g_{yy}\Big(\eta+\tau(y_1-\eta)\Big)d\tau\cdot\Big(f(Y_j,Z_j)\ ,f(Y_k,Z_k)\Big)\ .\end{aligned} \tag{2.30}$$

With the help of $C(1)$, $\widehat{C}(1)$, and $B(2)$ we get $g(y_1)=\mathcal{O}(h^3)$. We also have $g(\widehat{y}_1)=\mathcal{O}(h^3)$. From (2.27$b$) we can express $\widehat{\lambda}$ by

$$\widehat{\lambda}=\left(\int_0^1 g_y\Big(\widehat{y}_1+\tau f_z k_u(\widehat{y}_1,\widehat{z}_1,\widehat{u}_1)\widehat{\lambda}\Big)d\tau\cdot f_z k_u(\widehat{y}_1,\widehat{z}_1,\widehat{u}_1)\right)^{-1}\Big(g(\widetilde{y}_1)-g(\widehat{y}_1)\Big)\ . \tag{2.31}$$

Since $g(\widetilde{y}_1)=\mathcal{O}(h^3)$ too we get the estimate $\widehat{\lambda}=\mathcal{O}(h^3)$. We now use similar techniques to those given in the proof of Theorem 2.2. Subtracting (2.27b) from (2.27a) we obtain by linearization, using also $\widetilde{y}_1-y_1=\mathcal{O}(h)$,

$$\Delta\widetilde{y}_1 = \Delta y_1+(f_zk_u)(y_1,z_1,u_1)\Delta\lambda+\mathcal{O}(h^3\|\Delta y_1\|+h^3\|\Delta z_1\|+h^3\|\Delta u_1\|)\ , \tag{2.32a}$$

$$0 = g_y(y_1)\Delta\widetilde{y}_1+\theta_{s+1}+\mathcal{O}(h\|\Delta\widetilde{y}_1\|+\|\Delta\widetilde{y}_1\|^2)\ . \tag{2.32b}$$

Inserting (2.32a) into (2.32b) gives

$$\begin{aligned}\Delta\lambda = &\left((-g_yf_zk_u)^{-1}g_y\right)(y_1,z_1,u_1)\Delta y_1+ \\ &\mathcal{O}(h^3\|\Delta y_1\|+h^3\|\Delta z_1\|+h^3\|\Delta u_1\|+\|\theta_{s+1}\|+h\|\Delta\widetilde{y}_1\|+\|\Delta\widetilde{y}_1\|^2)\ ,\end{aligned} \tag{2.33a}$$

$$\Delta\widetilde{y}_1 = P_y^1\Delta y_1+\mathcal{O}(h^3\|\Delta y_1\|+h^3\|\Delta z_1\|+h^3\|\Delta u_1\|+\|\theta_{s+1}\|+h\|\Delta\widetilde{y}_1\|+\|\Delta\widetilde{y}_1\|^2)\ . \tag{2.33b}$$

Depending on the definition used for the computation of the u-component, we can estimate Δu_1 either by $\mathcal{O}(\|\Delta y_1\|+\|\Delta z_1\|)$ or by $\mathcal{O}(\|\Delta U\|)$ (and possibly $\mathcal{O}(\|\Delta\nu\|)$). Hence from (2.25$a,b,c$) we get the estimate (2.29a).

The formula (2.29b) can be proved in a similar way. The estimates $\widetilde{z}_1-\widehat{z}_1=\mathcal{O}(h^2)$, $\widehat{\mu}=\mathcal{O}(h^2)$, and $\widetilde{z}_1-z_1=\mathcal{O}(h)$ lead to

$$\begin{aligned}\Delta\mu = &\left((-g_yf_zk_u)^{-1}g_yf_z\right)(y_1,z_1,u_1)\Delta z_1+ \\ &\mathcal{O}(h^2\|\Delta y_1\|+h^2\|\Delta z_1\|+h^2\|\Delta u_1\|+\|\Delta\widetilde{y}_1\|+\|\theta'_{s+1}\|+h\|\Delta\widetilde{z}_1\|+\|\Delta\widetilde{z}_1\|^2)\ ,\end{aligned} \tag{2.34a}$$

$$\begin{aligned}\Delta\widetilde{z}_1 = &P_z^1\Delta z_1+ \\ &\mathcal{O}(h^2\|\Delta y_1\|+h^2\|\Delta z_1\|+h^2\|\Delta u_1\|+\|\Delta\widetilde{y}_1\|+\|\theta'_{s+1}\|+h\|\Delta\widetilde{z}_1\|+\|\Delta\widetilde{z}_1\|^2)\ .\end{aligned} \tag{2.34b}$$

From (2.25a,c,d) and the estimates for $\Delta\widetilde{y}_1$ and Δu_1 we get the result (2.29b).

The estimate (2.29c) simply follows from (2.29a,b).

For the second projection procedure described in Section 1 we can prove identical results (2.29) by rewriting

$$f(y_1, z_1+k(y_1,z_1,u_1+\lambda)-k(y_1,z_1,u_1))-f(y_1,z_1) = \tag{2.35a}$$
$$\int_0^1 f_z\left(y_1, z_1+\kappa\Big(\int_0^1 k_u(y_1,z_1,u_1+\tau\lambda)d\tau\cdot\lambda\Big)\right)d\kappa\cdot\left(\int_0^1 k_u(y_1,z_1,u_1+\tau\lambda)d\tau\cdot\lambda\right),$$

$$k(y_1,z_1,u_1+\mu)-k(y_1,z_1,u_1) = \int_0^1 k_u(y_1,z_1,u_1+\tau\mu)d\tau\cdot\mu \tag{2.35b}$$

and similarly for the "hat"-values. □

3. Taylor expansion of the PRK solution.

We consider one step of the PRK method (1.1) (always under the assumption (1.2)) with consistent initial values $\Psi_0=(y_0,z_0,u_0)$ at x_0 satisfying (II.1.2). The main objective of this section is to calculate the Taylor h-expansion of the numerical solution y_1, z_1 and to derive a result analogous to that obtained in Corollary II.4.2 for the exact solution of (II.1.1).

The forthcoming analysis follows that of Chapter II concerning the exact solution of (II.1.1) and is similar to those given in [HaWa91, Section VI.8] and [HaLuRo89a, Section 5] for semi-explicit index 2 DAE's. In Definition 1.1 we substitute hY_i', hZ_i' by k_i, ℓ_i in order to simplify the calculations below. Hence we obtain the new formulation

$$y_1 = y_0+\sum_{i=1}^s b_i k_i\,, \qquad z_1 = z_0+\sum_{i=1}^s b_i\ell_i\,, \tag{3.1a}$$

$$k_i = hf(Y_i,Z_i)\,, \qquad \ell_i = hk(Y_i,Z_i,U_i)\,, \qquad 0 = g(Y_i)\,, \tag{3.1b}$$

$$Y_i = y_0+\sum_{j=1}^s a_{ij}k_j\,, \qquad Z_i = z_0+\sum_{j=1}^s \widehat{a}_{ij}\ell_j\,. \tag{3.1c}$$

Our aim is to compute the derivatives at $h=0$ of $y_1, z_1, k_i, \ell_i, Y_i, Z_i$, and U_i considered as functions of h. Therefore the differentiations below are effected with respect to h. From now on we suppose that this nonlinear system (3.1) possesses a unique solution with (Y_i,Z_i,U_i) laying in a neighbourhood of (y_0,z_0,u_0) (see Theorem 2.1 and Theorem V.4.1). Hence for $h=0$ we have

$$Y_i(0) = y_0, \qquad Z_i(0) = z_0, \qquad U_i(0) = u_0\,. \tag{3.2}$$

For $q\geq 1$ (3.1c) yields

$$Y_i^{(q)} = \sum_{j=1}^s a_{ij}k_j^{(q)}, \qquad Z_i^{(q)} = \sum_{j=1}^s \widehat{a}_{ij}\ell_j^{(q)} \tag{3.3}$$

and for $h=0$ we get from (3.1b)

$$k_i^{(q)}(0) = q\Big(f(Y_i,Z_i)\Big)^{(q-1)}\Big|_{h=0}\,, \qquad \ell_i^{(q)}(0) = q\Big(k(Y_i,Z_i,U_i)\Big)^{(q-1)}\Big|_{h=0}\,. \tag{3.4}$$

The first two derivatives of $f(Y_i,Z_i)$ and the first derivative of $k(Y_i,Z_i,U_i)$ are given by

$$f(Y_i,Z_i)^{(1)} = f_y(Y_i,Z_i)Y_i^{(1)} + f_z(Y_i,Z_i)Z_i^{(1)}\,, \tag{3.5a}$$

$$f(Y_i,Z_i)^{(2)} = f_{yy}(Y_i,Z_i)(Y_i^{(1)},Y_i^{(1)}) + 2f_{yz}(Y_i,Z_i)(Y_i^{(1)},Z_i^{(1)}) + f_y(Y_i,Z_i)Y_i^{(2)} + f_{zz}(Y_i,Z_i)(Z_i^{(1)},Z_i^{(1)}) + f_z(Y_i,Z_i)Z_i^{(2)}\,, \tag{3.5b}$$

$$k(Y_i,Z_i,U_i)^{(1)} = k_y(Y_i,Z_i,U_i)Y_i^{(1)} + k_z(Y_i,Z_i,U_i)Z_i^{(1)} + k_u(Y_i,Z_i,U_i)U_i^{(1)}\,. \tag{3.6}$$

From these above results we obtain successively for $h=0$

$$k_i^{(1)}(0) = (f)_0\,, \tag{3.7a}$$

$$Y_i^{(1)}(0) = \sum_{j=1}^{s} a_{ij}(f)_0\,, \tag{3.8a}$$

$$\ell_i^{(1)}(0) = (k)_0\,, \tag{3.9a}$$

$$Z_i^{(1)}(0) = \sum_{j=1}^{s} \hat{a}_{ij}(k)_0\,, \tag{3.10a}$$

$$k_i^{(2)}(0) = 2\sum_{j=1}^{s} a_{ij}(f_y f)_0 + 2\sum_{j=1}^{s} \hat{a}_{ij}(f_z k)_0\,, \tag{3.7b}$$

$$Y_i^{(2)}(0) = 2\sum_{j,k=1}^{s} a_{ij}a_{jk}(f_y f)_0 + 2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk}(f_z k)_0\,, \tag{3.8b}$$

$$\ell_i^{(2)}(0) = 2\sum_{j=1}^{s} a_{ij}(k_y f)_0 + 2\sum_{j=1}^{s} \hat{a}_{ij}(k_z k)_0 + 2(k_u)_0 U_i^{(1)}(0)\,, \tag{3.9b}$$

$$Z_i^{(2)}(0) = 2\sum_{j,k=1}^{s} \hat{a}_{ij}a_{jk}(k_y f)_0 + 2\sum_{j,k=1}^{s} \hat{a}_{ij}\hat{a}_{jk}(k_z k)_0 + 2\sum_{j=1}^{s} \hat{a}_{ij}(k_u)_0 U_j^{(1)}(0)\,, \tag{3.10b}$$

$$\begin{aligned} k_i^{(3)}(0) = {} & 3\sum_{j,k=1}^{s} a_{ij}a_{ik}(f_{yy}(f,f))_0 + 3\sum_{j,k=1}^{s} a_{ij}\hat{a}_{ik}2(f_{yz}(f,k))_0 + \\ & 3\cdot 2\sum_{j,k=1}^{s} a_{ij}a_{jk}(f_y f_y f)_0 + 3\cdot 2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk}(f_y f_z k)_0 + \\ & 3\sum_{j,k=1}^{s} \hat{a}_{ij}\hat{a}_{ik}(f_{zz}(k,k))_0 + 3\cdot 2\sum_{j,k=1}^{s} \hat{a}_{ij}a_{jk}(f_z k_y f)_0 + \\ & 3\cdot 2\sum_{j,k=1}^{s} \hat{a}_{ij}\hat{a}_{jk}(f_z k_y k)_0 + 3\cdot 2\sum_{j=1}^{s} \hat{a}_{ij}(f_z k_u)_0 U_j^{(1)}(0)\,, \end{aligned} \tag{3.7c}$$

$$Y_i^{(3)}(0) = 3\sum_{j,k,l=1}^{s} a_{ij}a_{jk}a_{jl}(f_{yy}(f,f))_0 + 3\sum_{j,k,l=1}^{s} a_{ij}a_{jk}\hat{a}_{jl}2(f_{yz}(f,k))_0 + \tag{3.8c}$$

$$3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}a_{jk}a_{kl}(f_y f_y f)_0 + 3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}a_{jk}\hat{a}_{kl}(f_y f_z k)_0 +$$
$$3\sum_{j,k,l=1}^{s} a_{ij}\hat{a}_{jk}\hat{a}_{jl}(f_{zz}(k,k))_0 + 3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}\hat{a}_{jk}a_{kl}(f_z k_y f)_0 +$$
$$3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}\hat{a}_{jk}\hat{a}_{kl}(f_z k_y k)_0 + 3\cdot 2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk}(f_z k_u)_0 U_k^{(1)}(0)$$

where the self-evident subscript 0 indicates that the arguments are given by (y_0, z_0, u_0).

From the third equation of (3.1b) we have for all $q \geq 1$

$$0 = g(Y_i)^{(q-1)} \tag{3.11}$$

and for $q = 1, 2, 3$ we get

$$0 = g(Y_i)\,, \tag{3.12a}$$
$$0 = g_y(Y_i)Y_i^{(1)}\,, \tag{3.12b}$$
$$0 = g_{yy}(Y_i)(Y_i^{(1)}, Y_i^{(1)}) + g_y(Y_i)Y_i^{(2)}\,. \tag{3.12c}$$

For $h = 0$ inserting (3.2) and (3.8a, b) into (3.12) yields

$$0 = (g)_0\,, \tag{3.13a}$$
$$0 = \sum_{j=1}^{s} a_{ij}(g_y f)_0\,, \tag{3.13b}$$
$$0 = \sum_{j,k=1}^{s} a_{ij}a_{ik}(g_{yy}(f,f))_0 + 2\sum_{j,k=1}^{s} a_{ij}a_{jk}(g_y f_y f)_0 + 2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk}(g_y f_z k)_0\,. \tag{3.13c}$$

From the consistency of the initial values, the equations (3.13a, b) are satisfied for every choice of the PRK coefficients. However, concerning (3.13c), using $(g_{yy}(f,f))_0 + (g_y f_y f)_0 + (g_y f_z k)_0 = 0$ we obtain

$$0 = \left(2\sum_{j,k=1}^{s} a_{ij}a_{jk} - \sum_{j,k=1}^{s} a_{ij}a_{ik}\right)(g_y f_y f)_0 + \left(2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk} - \sum_{j,k=1}^{s} a_{ij}a_{ik}\right)(g_y f_z k)_0\,, \tag{3.14a}$$

$$0 = \left(\sum_{j,k=1}^{s} a_{ij}a_{ik} - 2\sum_{j,k=1}^{s} a_{ij}a_{jk}\right)(g_{yy}(f,f))_0 + \left(2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk} - 2\sum_{j,k=1}^{s} a_{ij}a_{jk}\right)(g_y f_z k)_0\,, \tag{3.14b}$$

$$0 = \left(\sum_{j,k=1}^{s} a_{ij}a_{ik} - 2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk}\right)(g_{yy}(f,f))_0 + \left(2\sum_{j,k=1}^{s} a_{ij}a_{jk} - 2\sum_{j,k=1}^{s} a_{ij}\hat{a}_{jk}\right)(g_y f_y f)_0\,. \tag{3.14c}$$

From the linear independency of the expressions $(g_{yy}(f,f))_0$, $(g_y f_y f)_0$, and $(g_y f_z k)_0$ which do not vanish in general, we must necessarily have that (recall that we want (3.1) to be satisfied)

$$\frac{1}{2}\sum_{j,k=1}^{s} a_{ij}a_{ik} = \sum_{j,k=1}^{s} a_{ij}a_{jk} = \sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk} \qquad \text{for} \quad i=1,\dots,s\,. \tag{3.15}$$

Hence if the simplifying assumption $C(1)$ is satisfied, then we see the necessity of $C(2)$ and $C\widehat{C}(2)$ to be satisfied too. These conditions are exactly those used in Theorem 2.1 and Theorem V.4.1 to show the existence and uniqueness of the PRK solution.

Now we want to compute the values $U_i^{(1)}(0)$. For $q=4$ (3.11) reads

$$0 = g(Y_i)^{(3)} = g_{yyy}(Y_i)(Y_i^{(1)},Y_i^{(1)},Y_i^{(1)})+3g_{yy}(Y_i)(Y_i^{(1)},Y_i^{(2)})+g_y(Y_i)Y_i^{(3)}\,. \tag{3.16}$$

Inserting the derivatives of Y_i for $h=0$ given in (3.8) yields

$$\begin{aligned}
0 = &\sum_{j,k,l=1}^{s} a_{ij}a_{ik}a_{il}(g_{yyy}(f,f,f))_0+2\sum_{j,k,l=1}^{s} a_{ij}a_{ik}a_{kl}3(g_{yy}(f,f_y f))_0+ \qquad (3.17)\\
&2\sum_{j,k,l=1}^{s} a_{ij}a_{ik}\widehat{a}_{kl}3(g_{yy}(f,f_z k))_0+3\sum_{j,k,l=1}^{s} a_{ij}a_{jk}a_{jl}(g_y f_{yy}(f,f))_0+\\
&3\sum_{j,k,l=1}^{s} a_{ij}a_{jk}\widehat{a}_{jl}2(g_y f_{yz}(f,k))_0+3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}a_{jk}a_{kl}(g_y f_y f_y f)_0+\\
&3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}a_{jk}\widehat{a}_{kl}(g_y f_y f_z k)_0+3\sum_{j,k,l=1}^{s} a_{ij}\widehat{a}_{jk}\widehat{a}_{jl}(g_y f_{zz}(k,k))_0+\\
&3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}\widehat{a}_{jk}a_{kl}(g_y f_z k_y f)_0+3\cdot 2\sum_{j,k,l=1}^{s} a_{ij}\widehat{a}_{jk}\widehat{a}_{kl}(g_y f_z k_y k)_0+\\
&3\cdot 2\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk}(g_y f_z k_u)_0 U_k^{(1)}(0)\,.
\end{aligned}$$

By hypothesis the matrix $(g_y f_z k_u)_0$ is non-singular. Thus if the matrix $\overline{A} = A\widehat{A}$ is invertible then we can obtain the values $U_i^{(1)}(0)$ from the previous formula and we get

$$\begin{aligned}
U_i^{(1)}(0) = &\frac{1}{3\cdot 2}\sum_{j,k,l=1}^{s} \omega_{ij}a_{jk}a_{jl}a_{jm}((-g_y f_z k_u)^{-1}g_{yyy}(f,f,f))_0+ \qquad (3.18)\\
&\frac{1}{3\cdot 2}\sum_{j,k,l=1}^{s} \omega_{ij}a_{jk}a_{jl}a_{lm}3((-g_y f_z k_u)^{-1}g_{yy}(f,f_y f))_0+\\
&\frac{1}{3\cdot 2}\sum_{j,k,l=1}^{s} \omega_{ij}a_{jk}a_{jl}\widehat{a}_{lm}3((-g_y f_z k_u)^{-1}g_{yy}(f,f_z k))_0+\\
&\frac{1}{3\cdot 2}3\sum_{j,k,l=1}^{s} \omega_{ij}a_{jk}a_{kl}a_{km}((-g_y f_z k_u)^{-1}g_y f_{yy}(f,f))_0+
\end{aligned}$$

$$\frac{1}{3\cdot 2}3\sum_{j,k,l=1}^{s}\omega_{ij}a_{jk}a_{kl}\widehat{a}_{km}2((-g_yf_zk_u)^{-1}g_yf_{yz}(f,k))_0+$$

$$\frac{1}{3\cdot 2}3\cdot 2\sum_{j,k,l=1}^{s}\omega_{ij}a_{jk}a_{kl}a_{lm}((-g_yf_zk_u)^{-1}g_yf_yf_yf)_0+$$

$$\frac{1}{3\cdot 2}3\cdot 2\sum_{j,k,l=1}^{s}\omega_{ij}a_{jk}a_{kl}\widehat{a}_{lm}((-g_yf_zk_u)^{-1}g_yf_yf_zk)_0+$$

$$\frac{1}{3\cdot 2}3\sum_{j,k,l=1}^{s}\omega_{ij}a_{jk}\widehat{a}_{kl}\widehat{a}_{km}((-g_yf_zk_u)^{-1}g_yf_{zz}(k,k))_0+$$

$$\frac{1}{3\cdot 2}3\cdot 2\sum_{j,k,l=1}^{s}\omega_{ij}a_{jk}\widehat{a}_{kl}a_{lm}((-g_yf_zk_u)^{-1}g_yf_zk_yf)_0+$$

$$\frac{1}{3\cdot 2}3\cdot 2\sum_{j,k,l=1}^{s}\omega_{ij}a_{jk}\widehat{a}_{kl}\widehat{a}_{lm}((-g_yf_zk_u)^{-1}g_yf_zk_yk)_0\ .$$

This result can then be inserted into (3.10*b*) and (3.8*c*) to give the exact expressions of $Z_i^{(2)}(0)$ and $Y_i^{(3)}(0)$ (these easy calculations are left to the reader).

Before giving general formulas for the derivatives of k_i, ℓ_i, Y_i, Z_i, and U_i at $h=0$, we first need the following definitions:

Definition 3.1. For each tree in $(L)DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\}$ we define recursively the rational numbers γ by

a) $\gamma(\emptyset_y)=1,\ \gamma(\emptyset_z)=1,\ \gamma(\emptyset_u)=1,\ \gamma(\tau_y)=1,\ \gamma(\tau_z)=1$;

b) $\gamma(t)=\varrho(t)\gamma(t_1)\dots\gamma(t_m)\gamma(v_1)\dots\gamma(v_n)$ if $t=[t_1,\dots,t_m,v_1,\dots,v_n]_y\in(L)DAT3_y$;

c) $\gamma(v)=\varrho(v)\gamma(t_1)\dots\gamma(t_m)\gamma(v_1)\dots\gamma(v_n)\gamma(u_1)\dots\gamma(u_p)$
if $v=[t_1,\dots,t_m,v_1,\dots,v_n,u_1,\dots,u_p]_z\in(L)DAT3_z$;

d) $\gamma(u)=\dfrac{1}{(\varrho(u)+2)(\varrho(u)+1)}\gamma(t_1)\dots\gamma(t_m)$ if $u=[t_1,\dots,t_m]_u\in(L)DAT3_u$.

Definition 3.2. For each tree in $(L)DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\}$ we define recursively the expressions Φ_i depending on the coefficients of the matrices $A, \widehat{A}$, and Ω ($\overline{A}=A\widehat{A}$ is supposed to be invertible) by

a) $\Phi_i(\emptyset_y)=1,\ \Phi_i(\emptyset_z)=1,\ \Phi_i(\emptyset_u)=1,\ \Phi_i(\tau_y)=1,\ \Phi_i(\tau_z)=1$;

b) $\Phi_i(t)=\displaystyle\sum_{\substack{\mu_1,\dots,\mu_m\\ \nu_1,\dots,\nu_n}} a_{i\mu_1}\dots a_{i\mu_m}\widehat{a}_{i\nu_1}\dots\widehat{a}_{i\nu_n}\Phi_{\mu_1}(t_1)\dots\Phi_{\mu_m}(t_m)\Phi_{\nu_1}(v_1)\dots\Phi_{\nu_n}(v_n)$
if $t=[t_1,\dots,t_m,v_1,\dots,v_n]_y\in(L)DAT3_y$;

c) $$\Phi_i(v) = \sum_{\substack{\mu_1,\dots,\mu_m \\ \nu_1,\dots,\nu_n}} a_{i\mu_1}\dots a_{i\mu_m}\widehat{a}_{i\nu_1}\dots\widehat{a}_{i\nu_n}\Phi_{\mu_1}(t_1)\dots\Phi_{\mu_m}(t_m)\times \Phi_{\nu_1}(v_1)\dots\Phi_{\nu_n}(v_n)\Phi_i(u_1)\dots\Phi_i(u_p)$$
$$\text{if } v=[t_1,\dots,t_m,v_1,\dots,v_n,u_1,\dots,u_p]_z\in(L)DAT3_z\ ;$$

d) $$\Phi_i(u) = \sum_{j,\mu_1,\dots\mu_m} \omega_{ij}a_{j\mu_1}\dots a_{j\mu_m}\Phi_{\mu_1}(t_1)\dots\Phi_{\mu_m}(t_m)$$
$$\text{if } u=[t_1,\dots,t_m]_u\in(L)DAT3_u\ .$$

The summation indices in b), c), and d) take their values in $\{1,\dots,s\}$.

We are now able to formulate a general result:

Theorem 3.1. *Under the assumptions of Theorem 2.1 (with consistent values $\Psi_0=(y_0,z_0,u_0)$ at x_0), the qth derivatives for $q\geq 1$ at $h=0$ of k_i,ℓ_i,Y_i,Z_i, and U_i satisfy*

$$k_i^{(q)}(0) = \sum_{\substack{t\in LDAT3_y \\ \varrho(t)=q}} \gamma(t)\Phi_i(t)F(t)(\Psi_0) = \sum_{\substack{t\in DAT3_y \\ \varrho(t)=q}} \alpha(t)\gamma(t)\Phi_i(t)F(t)(\Psi_0)\ , \tag{3.19a}$$

$$\ell_i^{(q)}(0) = \sum_{\substack{v\in LDAT3_z \\ \varrho(v)=q}} \gamma(v)\Phi_i(v)F(v)(\Psi_0) = \sum_{\substack{v\in DAT3_z \\ \varrho(v)=q}} \alpha(v)\gamma(v)\Phi_i(v)F(v)(\Psi_0)\ , \tag{3.19b}$$

$$Y_i^{(q)}(0) = \sum_{\substack{t\in LDAT3_y \\ \varrho(t)=q}} \gamma(t)\sum_{i=1}^{s} a_{ij}\Phi_j(t)F(t)(\Psi_0) = \sum_{\substack{t\in DAT3_y \\ \varrho(t)=q}} \alpha(t)\gamma(t)\sum_{i=1}^{s} a_{ij}\Phi_j(t)F(t)(\Psi_0)\ , \tag{3.19c}$$

$$Z_i^{(q)}(0) = \sum_{\substack{v\in LDAT3_z \\ \varrho(v)=q}} \gamma(v)\sum_{i=1}^{s} \widehat{a}_{ij}\Phi_j(v)F(v)(\Psi_0) = \sum_{\substack{v\in DAT3_z \\ \varrho(v)=q}} \alpha(v)\gamma(v)\sum_{i=1}^{s} \widehat{a}_{ij}\Phi_j(v)F(v)(\Psi_0)\ , \tag{3.19d}$$

$$U_i^{(q)}(0) = \sum_{\substack{u\in LDAT3_u \\ \varrho(u)=q}} \gamma(u)\Phi_i(u)F(u)(\Psi_0) = \sum_{\substack{u\in DAT3_u \\ \varrho(u)=q}} \alpha(u)\gamma(u)\Phi_i(u)F(u)(\Psi_0) \tag{3.19e}$$

where the coefficients α are those of Definition II.4.2.

Proof. Two proofs are given. The first one is similar to that of [HaWa91, Theorem VI.8.6] and is long. The second one only makes use of $DA3$-series and is short.

A first proof. The case $q=1$ follows from (3.7a), (3.8a), (3.9a), (3.10a), and (3.18).

For the general case one has to make use of formulas (3.3), (3.4), and (3.11). The total derivatives of $f(Y_i,Z_i)$, $k(Y_i,Z_i,U_i)$, and $g(Y_i)$ can be computed by Faà di Bruno's formula (see [HaNøWa93, Lemma II.2.8]) leading to

$$\Big(f(Y_i,Z_i)\Big)^{(q-1)} = \sum_{SLDAT3_{y,q}} \frac{\partial^{m+n}f(Y_i,Z_i)}{\partial y^m\partial z^n}\Big(Y_i^{(\mu_1)},\dots,Y_i^{(\mu_m)},Z_i^{(\nu_1)},\dots,Z_i^{(\nu_n)}\Big) \tag{3.20a}$$

with $\mu_1+\ldots+\mu_m+\nu_1+\ldots+\nu_n=q-1$,

$$\Big(k(Y_i,Z_i,U_i)\Big)^{(q-1)} = \sum_{SLDAT3_{z,q}} \frac{\partial^{m+n+p}k(Y_i,Z_i,U_i)}{\partial y^m\partial z^n\partial u^p}\Big(Y_i^{(\mu_1)},\ldots,Y_i^{(\mu_m)},Z_i^{(\nu_1)},\ldots,Z_i^{(\nu_n)},U_i^{(\kappa_1)},\ldots,U_i^{(\kappa_p)}\Big) \tag{3.20b}$$

with $\mu_1+\ldots+\mu_m+\nu_1+\ldots+\nu_n+\kappa_1+\ldots+\kappa_p=q-1$, and

$$0=\Big(g(Y_i)\Big)^{(q-1)} = \sum_{SLDAT3_{u,q}} \frac{\partial^m g(Y_i)}{\partial y^m}\Big(Y_i^{(\mu_1)},\ldots,Y_i^{(\mu_m)}\Big) \tag{3.20c}$$

with $\mu_1+\ldots+\mu_m=q-1$. The summations in (3.20a,b,c) are over sets of *special labelled DAT3-trees* of order q. These trees are completely similar to those of LS_q described in [HaNøWa93, Definition II.2.7] and they can be characterized more precisely as follows:

a) they are made of triangular vertices; the order of a (sub)tree is given by the number of such vertices;

b) they have no ramifications excepted possibly at the root;

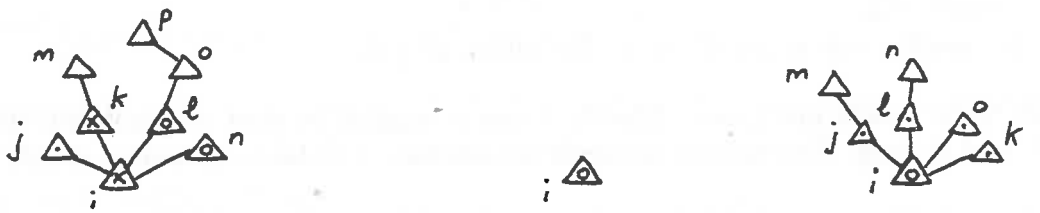
c) they are monotonically labelled trees, e.g., with the labels of $I=\{i<j<k<\ldots\}$;

d) a tree t is in $SLDAT3_{y,q}$ if it is of order q, if its root contains a meagre vertex, and if the vertices directly linked to the root contain a meagre vertex or a cross (no fat vertices); for a tree $t=[t_1,\ldots,t_m,v_1,\ldots,v_n]_y\in SLDAT3_{y,q}$ the integers μ_i and ν_j in (3.20a) are the orders of t_i and v_j respectively;

e) a tree v is in $SLDAT3_{z,q}$ if it is of order q, if its root contains a cross, and if the vertices directly linked to the root contain a meagre vertex, a cross, or a fat vertex; for a tree $v=[t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p]_z\in SLDAT3_{z,q}$ the integers μ_i, ν_j, and κ_k in (3.20b) are the orders of t_i, v_j, and u_k respectively;

f) a tree u is in $SLDAT3_{u,q}$ if it is of order q, if its root contains a fat vertex, and if the vertices directly linked to the root contain a meagre vertex (neither crosses, nor fat vertices); for a tree $t=[t_1,\ldots,t_m]_u\in SLDAT3_{u,q}$ the integer μ_i in (3.20c) is the order of t_i.

We give below three examples of special labelled DAT3-trees (their corresponding expression is mentioned)



$\frac{\partial^4 k}{\partial y\partial z\partial u^2}(Y_i,Z_i,U_i)(Y_i^{(1)},Z_i^{(2)},U_i^{(3)},U_i^{(1)})$ $\qquad$ $g(Y_i)$ $\qquad$ $\frac{\partial^4 g}{\partial y^4}(Y_i)(Y_i^{(2)},Y_i^{(2)},Y_i^{(1)},Y_i^{(1)})$

Figure 3.1.

With these preparations we are now able to pursue the proof. The formulas (3.3) can be inserted for $h=0$ into (3.20a,b,c) giving for $q\geq 1$ with the help of (3.4)

$$k_i^{(q)}(0) = q\sum_{SLDAT3_{y,q}}\left(\frac{\partial^{m+n}f}{\partial y^m\partial z^n}\right)_0\left(\sum_{j=1}^s a_{ij}k_j^{(\mu_1)}(0),\ldots,\sum_{j=1}^s a_{ij}k_j^{(\mu_m)}(0),\right. \tag{3.21a}$$

$$\sum_{j=1}^{s} \widehat{a}_{ij}\ell_j^{(\nu_1)}(0),\ldots,\sum_{j=1}^{s} \widehat{a}_{ij}\ell_j^{(\nu_n)}(0)\Big)\,,$$

$$\ell_i^{(q)}(0) = q \sum_{SLDAT3_{z,q}} \left(\frac{\partial^{m+n+p}k}{\partial y^m \partial z^n \partial u^p}\right)_0 \Big(\sum_{j=1}^{s} a_{ij}k_j^{(\mu_1)}(0),\ldots,\sum_{j=1}^{s} a_{ij}k_j^{(\mu_m)}(0), \tag{3.21b}$$

$$\sum_{j=1}^{s} \widehat{a}_{ij}\ell_j^{(\nu_1)}(0),\ldots,\sum_{j=1}^{s} \widehat{a}_{ij}\ell_j^{(\nu_n)}(0),$$

$$U_i^{(\kappa_1)}(0),\ldots,U_i^{(\kappa_p)}(0)\Big)\,,$$

$$0 = \sum_{SLDAT3_{u,q}} \left(\frac{\partial^m g}{\partial y^m}\right)_0 \Big(\sum_{j=1}^{s} a_{ij}k_j^{(\mu_1)}(0),\ldots,\sum_{j=1}^{s} a_{ij}k_j^{(\mu_m)}(0)\Big)\,. \tag{3.21c}$$

We rewrite (3.21c) in the form

$$0 = (g_y)_0 \sum_{j=1}^{s} a_{ij}k_j^{(q-1)}(0) + \sum_{\substack{SLDAT3_{u,q}\\ m\neq 1}} \left(\frac{\partial^m g}{\partial y^m}\right)_0 \Big(\sum_{j=1}^{s} a_{ij}k_j^{(\mu_1)}(0),\ldots,\sum_{j=1}^{s} a_{ij}k_j^{(\mu_m)}(0)\Big)\,. \tag{3.21c}$$

In this formula we are not interested in the cases $q=1,2,3$ which have been analyzed before (see (3.13a,b,c)). Inserting (3.21a) (with q replaced by $q-1$) into the first term of (3.21c) gives

$$\begin{aligned} 0 =& (q-1)\sum_{j=1}^{s} a_{ij}\cdot(g_y)_0 \sum_{SLDAT3_{y,q-1}} \left(\frac{\partial^{m+n}f}{\partial y^m \partial z^n}\right)_0 \Big(\sum_{k=1}^{s} a_{jk}k_k^{(\mu_1)}(0),\ldots,\sum_{k=1}^{s} \widehat{a}_{jk}\ell_k^{(\nu_1)}(0),\ldots\Big) \\ &+ \sum_{\substack{SLDAT3_{u,q}\\ m\neq 1}} \left(\frac{\partial^m g}{\partial y^m}\right)_0 \Big(\sum_{j=1}^{s} a_{ij}k_j^{(\mu_1)}(0),\ldots\Big) \end{aligned} \tag{3.22}$$

which can be rewritten

$$\begin{aligned} 0 =& (q-1)\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk}\cdot(g_y f_z)_0 \ell_k^{(q-2)}(0) + \\ & (q-1)\sum_{j=1}^{s} a_{ij}\cdot(g_y)_0 \sum_{\substack{SLDAT3_{y,q-1}\\ (m,n)\neq(0,1)}} \left(\frac{\partial^{m+n}f}{\partial y^m \partial z^n}\right)_0 \Big(\sum_{k=1}^{s} a_{jk}k_k^{(\mu_1)}(0),\ldots,\sum_{k=1}^{s} \widehat{a}_{jk}\ell_k^{(\nu_1)}(0),\ldots\Big) + \\ & \sum_{\substack{SLDAT3_{u,q}\\ m\neq 1}} \left(\frac{\partial^m g}{\partial y^m}\right)_0 \Big(\sum_{j=1}^{s} a_{ij}k_j^{(\mu_1)}(0),\ldots\Big)\,. \end{aligned} \tag{3.23}$$

Another insertion of (3.21b) (with q replaced by $q-2$) into the first term of (3.23) yields

$$0 = (q-1)(q-2)\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk}\cdot(g_y f_z)_0 \times \tag{3.24}$$

$$\sum_{SLDAT3_{z,q-2}} \left(\frac{\partial^{m+n+p}k}{\partial y^m \partial z^n \partial u^p}\right)_0 \left(\sum_{l=1}^{s} a_{kl} k_l^{(\mu_1)}(0), \ldots, \sum_{l=1}^{s} \widehat{a}_{kl} \ell_l^{(\nu_1)}(0), \ldots, U_k^{(\kappa_1)}(0), \ldots\right) +$$

$$(q-1)\sum_{j=1}^{s} a_{ij} \cdot (g_y)_0 \sum_{\substack{SLDAT3_{y,q-1} \\ (m,n)\neq(0,1)}} \left(\frac{\partial^{m+n} f}{\partial y^m \partial z^n}\right)_0 \left(\sum_{k=1}^{s} a_{jk} k_k^{(\mu_1)}(0), \ldots, \sum_{k=1}^{s} \widehat{a}_{jk} \ell_k^{(\nu_1)}(0), \ldots\right) +$$

$$\sum_{\substack{SLDAT3_{u,q} \\ m\neq 1}} \left(\frac{\partial^m g}{\partial y^m}\right)_0 \left(\sum_{j=1}^{s} a_{ij} k_j^{(\mu_1)}(0), \ldots\right)$$

which can be rewritten

$$0 = (q-1)(q-2)\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk} \cdot (g_y f_z k_u)_0 U_k^{(q-3)}(0) + \tag{3.25}$$

$$(q-1)(q-2)\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk} \cdot (g_y f_z)_0 \times$$

$$\sum_{\substack{SLDAT3_{z,q-2} \\ (m,n,p)\neq(0,0,1)}} \left(\frac{\partial^{m+n+p}k}{\partial y^m \partial z^n \partial u^p}\right)_0 \left(\sum_{l=1}^{s} a_{kl} k_l^{(\mu_1)}(0), \ldots, \sum_{l=1}^{s} \widehat{a}_{kl} \ell_l^{(\nu_1)}(0), \ldots, U_k^{(\kappa_1)}(0), \ldots\right) +$$

$$(q-1)\sum_{j=1}^{s} a_{ij} \cdot (g_y)_0 \sum_{\substack{SLDAT3_{y,q-1} \\ (m,n)\neq(0,1)}} \left(\frac{\partial^{m+n} f}{\partial y^m \partial z^n}\right)_0 \left(\sum_{k=1}^{s} a_{jk} k_k^{(\mu_1)}(0), \ldots, \sum_{k=1}^{s} \widehat{a}_{jk} \ell_k^{(\nu_1)}(0), \ldots\right) +$$

$$\sum_{\substack{SLDAT3_{u,q} \\ m\neq 1}} \left(\frac{\partial^m g}{\partial y^m}\right)_0 \left(\sum_{j=1}^{s} a_{ij} k_j^{(\mu_1)}(0), \ldots\right) .$$

As the matrices $(g_y f_z k_u)_0$ and $\overline{A} = A\widehat{A}$ are supposed to be regular then we can extract the values $U_i^{(q-3)}(0)$ for $q \geq 4$ from this formula, yielding

$$U_i^{(q-3)}(0) = \tag{3.26}$$

$$\frac{1}{(q-1)(q-2)}(q-1)(q-2)\sum_{j,k,l=1}^{s} \omega_{ij} a_{jk} \widehat{a}_{kl} \cdot \left((-g_y f_z k_u)^{-1} g_y f_z\right)_0 \times$$

$$\sum_{\substack{SLDAT3_{z,q-2} \\ (m,n,p)\neq(0,0,1)}} \left(\frac{\partial^{m+n+p}k}{\partial y^m \partial z^n \partial u^p}\right)_0 \left(\sum_{m=1}^{s} a_{lm} k_m^{(\mu_1)}(0), \ldots, \sum_{m=1}^{s} \widehat{a}_{lm} \ell_m^{(\nu_1)}(0), \ldots, U_l^{(\kappa_1)}(0), \ldots\right) +$$

$$\frac{1}{(q-1)(q-2)}(q-1)\sum_{j=1}^{s} \omega_{ij} a_{jk} \cdot \left((-g_y f_z k_u)^{-1} g_y\right)_0 \times$$

$$\sum_{\substack{SLDAT3_{y,q-1} \\ (m,n)\neq(0,1)}} \left(\frac{\partial^{m+n} f}{\partial y^m \partial z^n}\right)_0 \left(\sum_{l=1}^{s} a_{kl} k_l^{(\mu_1)}(0), \ldots, \sum_{l=1}^{s} \widehat{a}_{kl} \ell_l^{(\nu_1)}(0), \ldots\right) +$$

$$\frac{1}{(q-1)(q-2)}\sum_{j=1}^{s} \omega_{ij} (-g_y f_z k_u)_0^{-1} \sum_{\substack{SLDAT3_{u,q} \\ m\neq 1}} \left(\frac{\partial^m g}{\partial y^m}\right)_0 \left(\sum_{k=1}^{s} a_{jk} k_k^{(\mu_1)}(0), \ldots\right) .$$

From the formulas (3.3), (3.4), (3.21a,b), and (3.26) we obtain, by induction on q and exploiting the multilinearity of the derivatives, the following results

$$k_i^{(q)}(0) = \sum_{\substack{t\in DAT3_y \\ \varrho(t)=q}} \beta(t)\gamma(t)\Phi_i(t)F(t)(\Psi_0)\ , \tag{3.27a}$$

$$\ell_i^{(q)}(0) = \sum_{\substack{v\in DAT3_z \\ \varrho(v)=q}} \beta(v)\gamma(v)\Phi_i(v)F(v)(\Psi_0)\ , \tag{3.27b}$$

$$Y_i^{(q)}(0) = \sum_{\substack{t\in DAT3_y \\ \varrho(t)=q}} \beta(t)\gamma(t)\sum_{j=1}^{s} a_{ij}\Phi_j(t)F(t)(\Psi_0)\ , \tag{3.27c}$$

$$Z_i^{(q)}(0) = \sum_{\substack{v\in DAT3_z \\ \varrho(v)=q}} \beta(v)\gamma(v)\sum_{j=1}^{s} \widehat{a}_{ij}\Phi_j(v)F(v)(\Psi_0)\ , \tag{3.27d}$$

$$U_i^{(q)}(0) = \sum_{\substack{u\in DAT3_u \\ \varrho(u)=q}} \beta(u)\gamma(u)\Phi_i(u)F(u)(\Psi_0) \tag{3.27e}$$

where the integer coefficients β count the number of times that each elementary differential appears. It only remains to show that these coefficients are equal to the coefficients α of Definition II.4.2. The important fact is that these coefficients β are independent of the choice of the PRK method. The trick is now as follows: we consider a RK method $(a_{ij}=\widehat{a}_{ij})$ with an invertible matrix A, e.g., a Gauss method, satisfying $C(\eta)$ with η arbitrary high (we do not use the letter q in order to avoid a possible confusion). From [HaLuRo89a, Lemma 6.3] (see also Theorem 4.2 part a)) the internal stages of this method are known to satisfy (we point out that the order conditions described in the next section cannot be used at this point)

$$Y_i{-}y(x_0{+}c_ih) = \mathcal{O}(h^{\eta+1}),\ Z_i{-}z(x_0{+}c_ih) = \mathcal{O}(h^{\eta}),\ U_i{-}u(x_0{+}c_ih) = \mathcal{O}(h^{\eta-1}) \tag{3.28}$$

where $(y(x),z(x),u(x))$ is the exact solution of (II.1.1) passing through (y_0,z_0,u_0) at x_0. On the other hand one can easily show with the help of $C(\eta)$, as in the proof of Theorem 4.2 below, that the coefficients of this method satisfy the following relations

$$\gamma(t)\sum_{i=1}^{s} a_{ij}\Phi_j(t) = c_i^{\varrho(t)} \qquad \text{for all } t\in DAT3_y \text{ such that } \varrho(t)\le\eta\ , \tag{3.29a}$$

$$\gamma(v)\sum_{i=1}^{s} a_{ij}\Phi_j(v) = c_i^{\varrho(v)} \qquad \text{for all } v\in DAT3_z \text{ such that } \varrho(v)\le\eta{-}1\ , \tag{3.29b}$$

$$\gamma(u)\Phi_i(u) = c_i^{\varrho(u)} \qquad \text{for all } u\in DAT3_u \text{ such that } \varrho(u)\le\eta{-}2\ . \tag{3.29c}$$

Hence if $\eta\ge q{+}2$ holds for this method then we get from (3.27c,d,e)

$$Y_i^{(q)}(0) = c_i^q \sum_{\substack{t\in DAT3_y \\ \varrho(t)=q}} \beta(t)F(t)(\Psi_0)\ , \tag{3.30a}$$

$$Z_i^{(q)}(0) = c_i^q \sum_{\substack{v\in DAT3_z \\ \varrho(v)=q}} \beta(v)F(v)(\Psi_0)\ , \tag{3.30b}$$

$$U_i^{(q)}(0) = c_i^q \sum_{\substack{u \in DAT3_u \\ \varrho(u)=q}} \beta(u) F(u)(\Psi_0) \,. \tag{3.30c}$$

Concerning the derivatives of the exact solution, from Theorem II.4.1 we have

$$y(x_0+c_ih)^{(q)}\Big|_{h=0} = c_i^q \sum_{\substack{t \in DAT3_y \\ \varrho(t)=q}} \alpha(t) F(t)(\Psi_0) \,, \tag{3.31a}$$

$$z(x_0+c_ih)^{(q)}\Big|_{h=0} = c_i^q \sum_{\substack{v \in DAT3_z \\ \varrho(v)=q}} \alpha(v) F(v)(\Psi_0) \,, \tag{3.31b}$$

$$u(x_0+c_ih)^{(q)}\Big|_{h=0} = c_i^q \sum_{\substack{u \in DAT3_u \\ \varrho(u)=q}} \alpha(u) F(u)(\Psi_0) \,. \tag{3.31c}$$

Using the estimates (3.29) for $\eta \geq q+2$, we deduce that the above derivatives (3.30) and (3.31) must coincide. Thus from the linear independency of the elementary differentials (similarly to [HaNøWa93, Exercise II.2.4]) we get the desired result $\beta \equiv \alpha$. An alternative way of showing this result is by application of similar arguments used in the proof of Lemma II.5.1. We do not write all the details. For example in (3.21a), to each tuple $(\overline{T}, T_1, \ldots, T_M, V_1, \ldots, V_N)$ with $\overline{T} \in SLDAT3_y$, $T_1, \ldots, T_M \in LDAT3_y$, $V_1, \ldots, V_N \in LDAT3_z$, $D_1(\overline{T}) = \{\overline{T}_1, \ldots, \overline{T}_M, \overline{V}_1, \ldots, \overline{V}_N\}$, $\varrho(T_1) = \varrho(\overline{T}_1), \ldots, \varrho(V_N) = \varrho(\overline{V}_N)$, there corresponds a unique m.l. tree $T = [T_1, \ldots, T_M, V_1, \ldots, V_N]_y \in LDAT3_y$ and conversely. We illustrate this fact on the following example

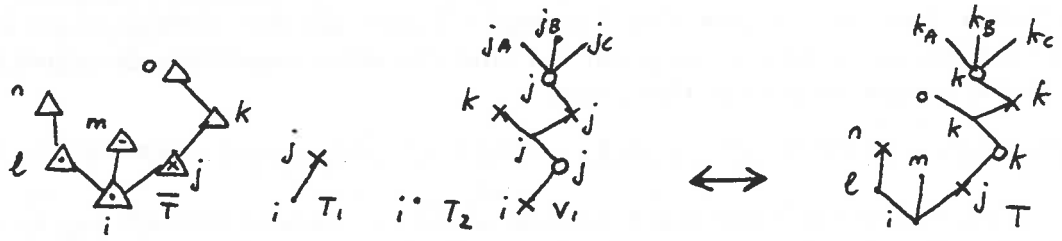


Figure 3.2.

Each m.l. tree appears exactly once, thus we get the desired result.

A second proof. By Theorem II.5.5 the expressions k_i, Y_i are $DA3_y$-series, ℓ_i, Z_i are $DA3_z$-series, and U_i are $DA3_u$-series

$$\begin{aligned} k_i &= DA3_y(\mathbf{k}_i, \Psi_0) \,, \qquad Y_i = DA3_y(\mathbf{Y}_i, \Psi_0) \,, \\ \ell_i &= DA3_z(\mathbf{L}_i, \Psi_0) \,, \qquad Z_i = DA3_z(\mathbf{Z}_i, \Psi_0) \,, \qquad U_i = DA3_u(\mathbf{U}_i, \Psi_0) \,, \end{aligned} \tag{3.32}$$

with coefficients satisfying

$$\mathbf{k}_i(\emptyset_y) = 0 \,, \quad \mathbf{Y}_i(\emptyset_y) = 1 \,, \quad \mathbf{L}_i(\emptyset_z) = 0 \,, \quad \mathbf{Z}_i(\emptyset_z) = 1 \,, \quad \mathbf{U}_i(\emptyset_u) = 1 \,, \tag{3.33}$$

and

$$\mathbf{k}_i(t) = \varrho(t) \mathbf{Y}_i(t_1) \cdots \mathbf{Y}_i(t_m) \mathbf{Z}_i(v_1) \cdots \mathbf{Z}_i(v_n) \,,$$

$$\begin{aligned}
\mathbf{Y}_i(t) &= \sum_{j=1}^{s} a_{ij}\mathbf{k}_j(t) \ , \\
\mathbf{L}_i(v) &= \varrho(v)\mathbf{Y}_i(t_1)\cdots\mathbf{Y}_i(t_m)\mathbf{Z}_i(v_1)\cdots\mathbf{Z}_i(v_n)\mathbf{U}_i(u_1)\cdots\mathbf{U}_i(u_p) \ , \qquad (3.34) \\
\mathbf{Z}_i(v) &= \sum_{j=1}^{s} \widehat{a}_{ij}\mathbf{L}_j(v) \ , \\
0 &= \mathbf{Y}_i(t_1)\cdots\mathbf{Y}_i(t_m) - (\varrho(u)+2)(\varrho(u)+1)\sum_{j,k=1}^{s} a_{ij}\widehat{a}_{jk}\mathbf{U}_k(u) \ ,
\end{aligned}$$

for $t=[t_1,\ldots,t_m,v_1,\ldots,v_n]_y \in DAT3_y$, $v=[t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p]_z \in DAT3_z$, and $u=[t_1,\ldots,t_m]_u \in DAT3_u$. □

For the numerical solution y_1, z_1, an easy application of the preceding theorem yields:

Theorem 3.2. *Under the assumptions of Theorem 2.1 (with consistent values $\Psi_0 = (y_0, z_0, u_0)$ at x_0), the qth derivatives for $q \geq 1$ at $h=0$ of the numerical solution y_1, z_1 satisfy*

$$y_1^{(q)}(0) = \sum_{\substack{t\in LDAT3_y \\ \varrho(t)=q}} \gamma(t)\sum_{i=1}^{s} b_i\Phi_i(t)F(t)(\Psi_0) = \sum_{\substack{t\in DAT3_y \\ \varrho(t)=q}} \alpha(t)\gamma(t)\sum_{i=1}^{s} b_i\Phi_i(t)F(t)(\Psi_0) \ , \qquad (3.35a)$$

$$z_1^{(q)}(0) = \sum_{\substack{v\in LDAT3_z \\ \varrho(v)=q}} \gamma(v)\sum_{i=1}^{s} b_i\Phi_i(v)F(v)(\Psi_0) = \sum_{\substack{v\in DAT3_z \\ \varrho(v)=q}} \alpha(v)\gamma(v)\sum_{i=1}^{s} b_i\Phi_i(v)F(v)(\Psi_0) \ . \qquad (3.35b)$$

Hence the Taylor expansions at x_0 are given by

$$\begin{aligned}
y_1 &= y_0 + \sum_{t\in LDAT3_y} \frac{h^{\varrho(t)}}{\varrho(t)!}\gamma(t)\sum_{i=1}^{s} b_i\Phi_i(t)F(t)(\Psi_0) = \qquad (3.36a) \\
&\qquad y_0 + \sum_{t\in DAT3_y} \alpha(t)\frac{h^{\varrho(t)}}{\varrho(t)!}\gamma(t)\sum_{i=1}^{s} b_i\Phi_i(t)F(t)(\Psi_0) \ , \\
z_1 &= z_0 + \sum_{v\in LDAT3_z} \frac{h^{\varrho(v)}}{\varrho(v)!}\gamma(v)\sum_{i=1}^{s} b_i\Phi_i(v)F(v)(\Psi_0) = \qquad (3.36b) \\
&\qquad z_0 + \sum_{v\in DAT3_z} \alpha(v)\frac{h^{\varrho(v)}}{\varrho(v)!}\gamma(v)\sum_{i=1}^{s} b_i\Phi_i(v)F(v)(\Psi_0) \ .
\end{aligned}$$

Proof. From (3.1a) we get for $q \geq 1$

$$y_1^{(q)}(0) = \sum_{i=1}^{s} b_i k_i^{(q)}(0) \ , \qquad z_1^{(q)}(0) = \sum_{i=1}^{s} b_i \ell_i^{(q)}(0) \ . \qquad (3.37)$$

Thus the statement of this theorem is an immediate consequence of (3.19a,b). □

In the same way, for the u-component we easily obtain:

Theorem 3.3. *Under the assumptions of Theorem 2.1 (with consistent values* $\Psi_0 = (y_0, z_0, u_0)$ *at* x_0*), the* q*th derivatives for* $q \geq 1$ *at* $h=0$ *of* u_1 *given by the choice* (1.7) *satisfy*

$$u_1^{(q)}(0) = \sum_{\substack{u \in LDAT3_u \\ \varrho(u)=q}} \gamma(u) \sum_{i,j=1}^{s} b_i w_{ij} \Phi_j(u) F(u)(\Psi_0) = \sum_{\substack{u \in DAT3_u \\ \varrho(u)=q}} \alpha(u)\gamma(u) \sum_{i,j=1}^{s} b_i w_{ij} \Phi_j(u) F(u)(\Psi_0) \; . \tag{3.38}$$

Hence the Taylor expansion at x_0 *is given by*

$$u_1 = u_0 + \sum_{u \in LDAT3_u} \frac{h^{\varrho(u)}}{\varrho(u)!} \gamma(u) \sum_{i,j=1}^{s} b_i w_{ij} \Phi_j(u) F(u)(\Psi_0) = u_0 + \sum_{u \in DAT3_u} \alpha(u) \frac{h^{\varrho(u)}}{\varrho(u)!} \gamma(u) \sum_{i,j=1}^{s} b_i w_{ij} \Phi_j(u) F(u)(\Psi_0) \; . \tag{3.39}$$

For the choice (1.9) *the coefficients* w_{ij} *have to be replaced by* $\widehat{w}_{ij}$. □

4. Local error and order conditions.

In this section we first give necessary and sufficient conditions for a PRK method to attain a certain order in the local error. With the help of the simplifying assumptions introduced in Section 1, we then show optimal estimates for the local error of methods satisfying the simplifying assumption (S). In this section we suppose that the matrix $\overline{A} = A\widehat{A}$ is invertible and that the nonlinear system (1.1) possesses a unique solution.

Definition 4.1. The *local error* of one step of a PRK method (1.1) with consistent initial values $\Psi_0 = (y_0, z_0, u_0)$ at x_0 is given by

$$\delta y_h(x_0) = y_1 - y(x_0+h) \; , \quad \delta z_h(x_0) = z_1 - z(x_0+h) \; , \quad \delta u_h(x_0) = u_1 - u(x_0+h) \tag{4.1}$$

where $\Psi(x) = (y(x), z(x), u(x))$ is the exact solution of (II.1.1).

A comparison of Theorem 3.2 with Corollary II.4.2 gives:

Theorem 4.1. *For the PRK method* (1.1) *with an invertible matrix* $\overline{A} = A\widehat{A}$ *we have*

$$\delta y_h(x_0) = \mathcal{O}(h^{\mu+1}) \iff \sum_{i=1}^{s} b_i \Phi_i(t) = \frac{1}{\gamma(t)} \tag{4.2a}$$

for all $t \in DAT3_y$ *such that* $\varrho(t) \leq \mu$,

$$\delta z_h(x_0) = \mathcal{O}(h^{\nu+1}) \iff \sum_{i=1}^{s} b_i \Phi_i(v) = \frac{1}{\gamma(v)} \tag{4.2b}$$

for all $v \in DAT3_z$ *such that* $\varrho(v) \leq \nu$

where the coefficients γ and Φ_i are given in Definition 3.1 and Definition 3.2 respectively. For the definition of u_1 given by (1.7) *we have*

$$\delta u_h(x_0) = \mathcal{O}(h^{\kappa+1}) \quad \Longleftrightarrow \quad \sum_{i,j=1}^{s} b_i w_{ij} \Phi_j(u) = \frac{1}{\gamma(u)} \tag{4.2c}$$

for all $u \in DAT3_u$ such that $\varrho(u) \leq \kappa$.

For the choice (1.9) *the coefficients w_{ij} have to be replaced by $\widehat{w}_{ij}$.* □

Applying repeatedly the definition of Φ_i in Definition 3.2 we get the following algorithm:

Formation of the left-hand side of the order condition for a given tree.

To each vertex we associate one distinct summation index, excepted for a fat root to which three distinct indices are associated. The left-hand side of the order condition is then a sum over all present indices of a product with factors:

1) b_i if "i" is associated to a meagre root or a cross root;
2) $b_i w_{ij} \omega_{jk}$ if "i, j, k" are associated to a fat root (for the choice (1.7));
3) a_{ij} if "j" is associated to a meagre vertex laying directly above "i";
4) $\widehat{a}_{ij}$ if "j" is associated to a cross laying directly above "i";
5) ω_{ij} if "j" is associated to a fat vertex laying directly above "i".

For the choice (1.9) the factor w_{ij} in 2) has to be replaced by $\widehat{w}_{ij}$. If the condition $C(1)$ is satisfied then for a *terminal* meagre vertex, the factor a_{ij} in 3) can be changed into c_i. Similarly if the condition $\widehat{C}(1)$ is satisfied then for a *terminal* cross the factor $\widehat{a}_{ij}$ in 4) can be changed into c_i.

The coefficients γ of Definition 3.1 entering in the right-hand side of the order conditions can be computed in a similar way.

Example 4.1. We suppose that $C(1)$ is satisfied. The order condition of the tree on the right-hand side of Fig. II.3.3 is then given by

$$\sum_{i,j,k,l,m,n=1}^{s} b_i \omega_{ij} c_j^3 \omega_{ik} c_k^2 a_{kl} \widehat{a}_{lm} \omega_{mn} c_n^3 = \left(5 \cdot \frac{1}{3 \cdot 2} \cdot \frac{1}{5 \cdot 4} \cdot 3 \cdot 2 \cdot \frac{1}{3 \cdot 2}\right)^{-1} = 24 \, .$$

As $\Omega = (A\widehat{A})^{-1} = \widehat{W}W$ this expression can be reduced to

$$\sum_{i,j,k=1}^{s} b_i \omega_{ij} c_j^3 \omega_{ik} c_k^5 = 24 \, .$$

This simplified order condition is also exactly that corresponding to the tree situated on the left of the considered tree in Fig. II.3.3.

The simplification shown in the above example can be applied systematically leading to the following result:

Lemma 4.2. *If a DAT3-tree possesses a strict subtree of one of the following forms*

$[[u]_z]_y$ *with* $u \in DAT3_u$, $\quad [[t]_u]_z$ *with* $t \in DAT3_y$, $\quad [[v]_y]_u$ *with* $v \in DAT3_z$

then its order condition is equivalent to that of a tree of the same order, but with fewer fat vertices. □

Consequently trees satisfying the above assumption need not to be considered for the construction of PRK methods.

For the important special case where the function k of (II.1.1) is linear in u, i.e.,

$$k(y,z,u) = k_0(y,z) + k_u(y,z)u \ , \tag{4.3}$$

the order conditions of trees possessing a subtree of the form

$$[t_1,\ldots,t_m,v_1,\ldots,v_n,u_1,\ldots,u_p]_z \qquad \text{with } p \geq 2,$$

need not to be considered too, because their corresponding elementary differential vanishes identically (since $k_{uu} \equiv 0$). An important class of problems satisfying (4.3) are constrained mechanical and Hamiltonian systems in index 3 formulation (see Subsection I.4.1).

The simplifying assumptions introduced in Section 1 are of great importance in the study of the local error. They allow to reduce the order condition of a given tree to other order conditions corresponding to "simpler" trees. From now on we adopt the conventions that a *square vertex* stands for a meagre vertex, a cross, or a fat vertex, and a *triangular vertex* for a meagre vertex or a cross.

Hereafter φ_{ij} and φ_i stand for arbitrary expressions. From the simplifying assumption $C(q)$, we have

$$\sum b_i \varphi_{ij} a_{jl} c_l^{k-1} = \frac{1}{k} \sum b_i \varphi_{ij} c_j^k \qquad \text{for} \quad k = 1,\ldots,q \ . \tag{4.4}$$

Hence in Fig. 4.1 the order condition of the tree on the left-hand side is implied by that of the tree on the right-hand side.

Figure 4.1. Reduction by $C(q)$.

Similarly we have from $\widehat{C}(\widehat{q})$

$$\sum b_i \varphi_{ij} \widehat{a}_{jl} c_l^{k-1} = \frac{1}{k} \sum b_i \varphi_{ij} c_j^k \qquad \text{for} \quad k = 1,\ldots,\widehat{q} \ . \tag{4.5}$$

Figure 4.2. Reduction by $\widehat{C}(\widehat{q})$.

From $IC\widehat{C}(Q)$ we get

$$\sum b_i\varphi_{ij}\omega_{jl}c_l^k = k(k-1)\sum b_i\varphi_{ij}c_j^{k-2} \qquad \text{for} \quad k=2,\dots,Q\,. \tag{4.6}$$

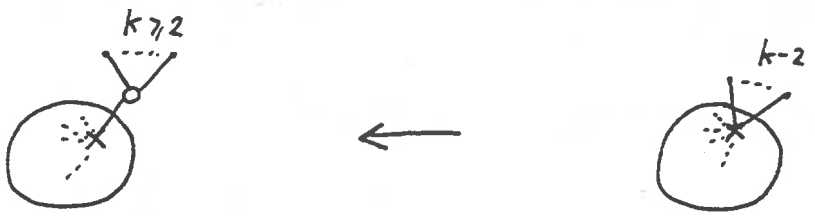


Figure 4.3. Reduction by $IC\widehat{C}(Q)$.

For the simplifying assumption $D(r)$ we have

$$\sum b_i c_i^{k-1}a_{ij}\varphi_j = \frac{1}{k}\Big(\sum b_j\varphi_j - \sum b_j c_j^k\varphi_j\Big) \qquad \text{for} \quad k=1,\dots,r\,. \tag{4.7}$$

Hence in Fig. 4.4 the order condition of the tree on the left-hand side is implied by those of the two trees on the right-hand side.

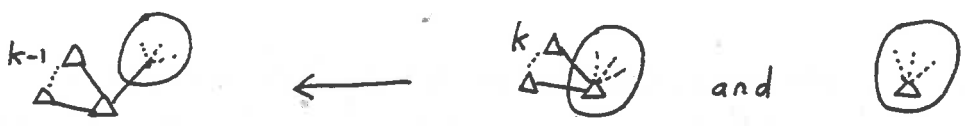


Figure 4.4. Reduction by $D(r)$.

Similarly we have from $\widehat{D}(\widehat{r})$

$$\sum b_i c_i^{k-1}\widehat{a}_{ij}\varphi_j = \frac{1}{k}\Big(\sum b_j\varphi_j - \sum b_j c_j^k\varphi_j\Big) \qquad \text{for} \quad k=1,\dots,\widehat{r}\,. \tag{4.8}$$

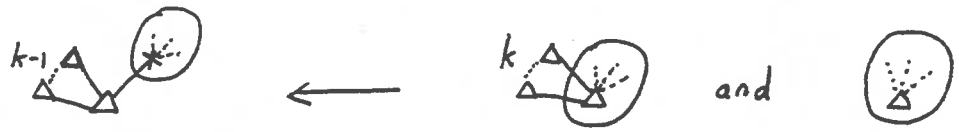


Figure 4.5. Reduction by $\widehat{D}(\widehat{r})$.

If $I\widehat{D}(1)$ (or equivalently $\widehat{D}(1)$) and (S) are satisfied then $ID\widehat{D}(R)$ can be rewritten

$$ID\widehat{D}(R): \qquad \sum_{i=1}^{s} b_i c_i^k \omega_{il} = \sum_{i=1}^{s} b_i \omega_{il} + k(k-1) b_l c_l^{k-2} - k\delta_{sl}$$

$$\text{for} \quad l = 1, \ldots, s\,, \quad k = 2, \ldots, R\,,$$

where δ_{sl} is the lth-component of the vector $e_s = (0, \ldots, 0, 1)^T$, i.e., $\delta_{sl} = 0$ if $l \neq s$, else $\delta_{ss} = 1$. From $ID\widehat{D}(R)$-$\widehat{D}(\widehat{r})$-(S) $(\widehat{r} \geq 1)$ we get the relation

$$\sum b_i c_i^{m-1} \widehat{a}_{ij} c_j^n \omega_{jl} \varphi_l = \frac{1}{m}\Big(m\varphi_s + n(n-1) \sum b_l c_l^{n-2} \varphi_l - (m+n)(m+n-1) \sum b_l c_l^{m+n-2} \varphi_l \Big) \tag{4.9}$$

$$\text{for } 1 \leq m \leq \widehat{r} \text{ and } m+n \leq R.$$

For $m+n=1$ and $n \leq 1$, the terms $(m+n)(m+n-1) \sum b_l c_l^{m+n-2} \varphi_l$ and $n(n-1) \sum b_l c_l^{n-2} \varphi_l$ in (4.9) have to be respectively removed. Hence the order condition of the tree on the left-hand side is implied by those of the trees on the right-hand side.

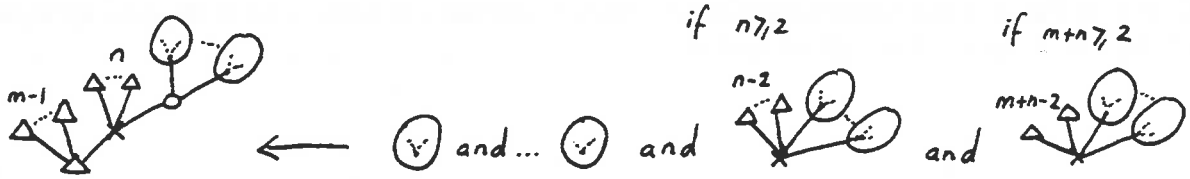


Figure 4.6. Reduction by $ID\widehat{D}(R)$-$\widehat{D}(\widehat{r})$-(S) $(\widehat{r} \geq 1)$.

We clearly see that the three above reductions allow to reduce the order condition corresponding to a certain tree to other order conditions corresponding to trees with a smaller height which is defined as follows:

Definition 4.2. The *height* of a tree $w \in (L)DAT3 \cup \{\emptyset_y, \emptyset_z, \emptyset_u\}$, denoted by $H(w)$, is defined recursively as follows:

a) $H(\emptyset_y) = 0, \quad H(\emptyset_z) = 0, \quad H(\emptyset_u) = 0, \quad H(\tau_y) = 1, \quad H(\tau_z) = 1;$

b) $H(t) = 1 + \max(H(t_1), \ldots, H(t_m), H(v_1), \ldots, H(v_n))$

if $t=[t_1,\dots,t_m,v_1,\dots,v_n]_y \in (L)DAT3_y$;

c) $H(v)=1+\max(H(t_1),\dots,H(t_m),H(v_1),\dots,H(v_n),H(u_1),\dots,H(u_p))$
if $v=[t_1,\dots,t_m,v_1,\dots,v_n,u_1,\dots,u_p]_z \in (L)DAT3_z$;

d) $H(u)=1+\max(H(t_1),\dots,H(t_m))$ if $u=[t_1,\dots,t_m]_u \in (L)DAT3_u$.

Examples 4.2.

1. The height of the tree on the left-hand side of Fig. II.3.3 (or Fig.4.2) is equal to 6.
2. That of the tree on the right-hand side of Fig. II.3.3 (or Fig.4.2) is equal to 3.

Here is the main result of this section:

Theorem 4.3.

a) Let us suppose that the PRK method satisfies $B(p)$, $C(q)$, $\widehat{C}(\widehat{q})$, with $q \geq 2$ and $\widehat{q} \geq 1$, and that the matrix $\overline{A}=A\widehat{A}$ is invertible. Then we have

$$\delta y_h(x_0) = \mathcal{O}(h^{\min(p,q,\widehat{q}+1)+1}) , \quad P_y(x_0+h)\delta y_h(x_0) = \mathcal{O}(h^{\min(p,q+1,\widehat{q}+1)+1}) , \tag{4.10a}$$

$$\delta z_h(x_0) = \mathcal{O}(h^{\min(p,q-1,\widehat{q})+1}) , \quad P_z(x_0+h)\delta z_h(x_0) = \mathcal{O}(h^{\min(p,q,\widehat{q}+1)+1}) , \tag{4.10b}$$

$$\delta u_h(x_0) = \mathcal{O}(h^{\min(p,q-2,\widehat{q}-1)+1}) \quad \textit{(for the choice (1.7) or (1.9))} \tag{4.10c}$$

where $P_y(x)$, $P_z(x)$ are the projectors (1.14) evaluated at $\Psi(x)=(y(x),z(x),u(x))$, the exact solution of (II.1.1) at x.

b) Moreover, if in addition $C\widehat{C}(Q)$, $D(r)$, $\widehat{D}(\widehat{r})$, $D\widehat{D}(R)$, and (S) hold, then we obtain

$$\delta y_h(x_0) = \mathcal{O}(h^{k+1}), \quad \textit{with} \tag{4.11a}$$
$$k = \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q-1, Q+\widehat{r}, Q+R) ,$$

$$P_z(x_0+h)\delta z_h(x_0) = \mathcal{O}(h^{\ell+1}), \quad \textit{with} \tag{4.11b}$$
$$\ell = \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q-2, Q+\widehat{r}, Q+R) .$$

Remark 4.1. If the function k of (II.1.1) is linear in u then the assumptions $q \geq 2$ and $\widehat{q} \geq 1$ can be omitted. The estimates (4.11) change to

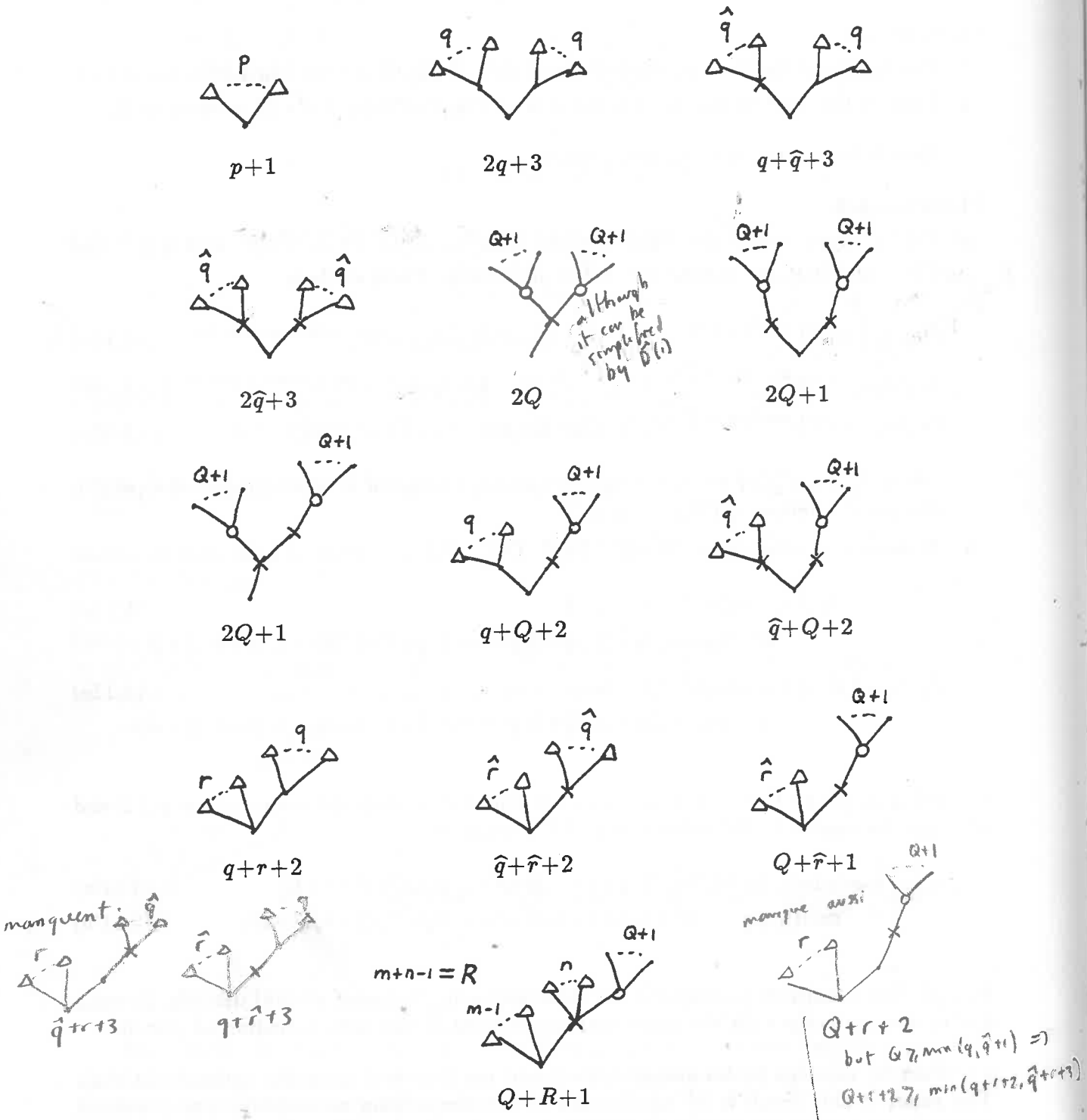
$$k = \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q, Q+\widehat{r}, Q+R) , \tag{4.11'a}$$
$$\ell = \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q-1, Q+\widehat{r}, Q+R) . \tag{4.11'b}$$

Proof. The results of part a) can be obtained as in the proof of [HaLuRo89a, Lemma 6.3] or alternatively with the same techniques used in the demonstration of part b).

Part b) remains to be demonstrated and we first deal with the estimate (4.11*a*). The proof of this result is by application of the simplifying assumptions (as described above) to the order conditions of $DAT3_y$-trees of order less than k:

- We first apply repeatedly the three reductions given by $C(q)$, $\widehat{C}(\widehat{q})$, and $IC\widehat{C}(Q)$;
- We then apply repeatedly the reduction given by $ID\widehat{D}(R)$-$\widehat{D}(\widehat{r})$-(S) $(\widehat{r} \geq 1)$;
- We finally apply repeatedly the two reductions given by $D(r)$ and $\widehat{D}(\widehat{r})$.

The value of k follows from the "first worse" order conditions which cannot be reduced with simplifying assumptions. These order conditions correspond to the "first worse" trees which are of the following forms (their corresponding order is mentioned)



$p+1$ $\quad 2q+3$ $\quad q+\hat{q}+3$

$2\hat{q}+3$ $\quad 2Q$ $\quad 2Q+1$

$2Q+1$ $\quad q+Q+2$ $\quad \hat{q}+Q+2$

$q+r+2$ $\quad \hat{q}+\hat{r}+2$ $\quad Q+\hat{r}+1$

$\hat{q}+r+3$ $\quad q+\hat{r}+3$ $\quad Q+R+1$

$Q+r+2$
but $Q \geq \min(q, \hat{q}+1) \Rightarrow$
$Q+r+2 \geq \min(q+r+2, \hat{q}+r+3)$

Figure 4.7. The "first worse" trees for the y-component.

An easy calculation shows that $q+\hat{q}+3 \geq \min(2q+3, 2\hat{q}+3)$, $q+Q+2 \geq \min(2q+3, 2Q)$, and $\hat{q}+Q+2 \geq \min(2\hat{q}+3, 2Q)$. After the above reductions, all order conditions that remain to be satisfied are those corresponding to the "bushy" trees (of order less than

k). They are given by $\sum_{i=1}^{s} b_i c_i^{j-1} = 1/j$ (with $j \le k$) and by $B(p)$ they are obviously satisfied. This achieves the proof for the estimate (4.11a).

However, the second estimate (4.11b) is much more difficult to prove. The results on $DA3$-series of Section II.5 will be used in a crucial way. The first idea is to develop the local error of the z-component, not at x_0, but at x_0+h. We get the $DA3_z$-series

$$\begin{aligned} z_1 - z(x_0+h) &= DA3_z(\mathbf{a}, \Psi(x_0+h)) - z(x_0+h) \\ &= \sum_{v \in DAT3_z} \alpha(v) \frac{h^{\varrho(v)}}{\varrho(v)!} \mathbf{a}(v) F(v)\big(\Psi(x_0+h)\big) \ . \end{aligned} \tag{4.12}$$

The coefficients $\mathbf{a}(v)$ remain to be found. From Theorem 3.2 the Taylor expansion of z_1 can be written in a $DA3_z$-series

$$z_1 = DA3_z(\mathbf{\Phi}, \Psi(x_0)) \tag{4.13}$$

where $\mathbf{\Phi}(v) = \gamma(v)\Phi(v)$ with $\Phi(\emptyset_z) = 1$ and for a tree $v \in DAT3_z$ $\Phi(v) = \sum_{i=1}^{s} b_i \Phi_i(v)$. On the other hand with the help of Example II.5.1.2 and formula (II.5.23) we have

$$z_1 = DA3_z(\mathbf{a}, \Psi(x_0+h)) = DA3_z(\mathbf{a}, DA3_z(\mathbf{p}, \Psi(x_0))) = DA3_z(\mathbf{p} * \mathbf{a}, \Psi(x_0)) \ . \tag{4.14}$$

Hence the relation $\mathbf{p} * \mathbf{a} = \mathbf{\Phi}$ holds. Since by Theorem II.5.4 the $DA3$-series with the composition law given by $*$ of Definition II.5.10 form a group, we get $\mathbf{a} = \mathbf{p}^{-1} * \mathbf{\Phi}$. As $\mathbf{p}^{-1}$ is simply given by $\mathbf{p}_{-1}$ of Example II.5.1.2, we obtain

$$\mathbf{a}(v) = \mathbf{p}_{-1} * \mathbf{\Phi}(v) = \frac{1}{\alpha(v)} \sum_{\text{labellings of } v} \left(\sum_{j=0}^{\varrho(v)} \binom{\varrho(v)}{j} \Phi(s_j(v)) \prod_{\omega \in d_j(v)} \mathbf{p}_{-1}(\omega) \right) \tag{4.15}$$

Because of

$$\prod_{\omega \in d_j(v)} \mathbf{p}_{-1}(\omega) = \prod_{\omega \in d_j(v)} (-1)^{\varrho(\omega)} = (-1)^{\sum_{\omega \in d_j(v)} \varrho(\omega)} = (-1)^{\varrho(v)-j} = (-1)^{\varrho(v)+j} \tag{4.16}$$

and

$$0 = (1-1)^{\varrho(v)} = \sum_{j=0}^{\varrho(v)} (-1)^j \binom{\varrho(v)}{j} \tag{4.17}$$

we arrive at

$$\mathbf{a}(v) = \frac{(-1)^{\varrho(v)}}{\alpha(v)} \sum_{\text{labellings of } v} \left(\sum_{j=0}^{\varrho(v)} (-1)^j \binom{\varrho(v)}{j} \Big(\gamma\big(s_j(v)\big) \Phi\big(s_j(v)\big) - 1 \Big) \right) \ . \tag{4.18}$$

The main idea is to show that for all trees $v \in DAT3_z$ which are not of the form $[u]_z$ with $u \in DAT3_u$, i.e., for which $F(v) \not\equiv k_u F(u)$, we have $\mathbf{a}(v) = 0$ if $\varrho(v) \le \ell$ (with ℓ given by (4.11b)). This will give the desired result since the remaining trees are of higher order or satisfy $v = [u]_z$ with $u \in DAT3_u$ implying that $P_z F(v) \equiv P_z k_u F(u) \equiv 0$.

Now we consider a tree $v \in DAT3_z$, $v \ne [u]_z$ with $u \in DAT3_u$, satisfying $\varrho(v) \le \ell$. We can suppose that $\varrho(v) \ge \min(p, q, \widehat{q}+1)+1$, because we already know by (4.10b)

that $\mathrm{a}(v)=0$ for all trees v of order $\varrho(v)\le\min(p,q,\widehat{q}+1)$. We first recursively simplify the terms $\gamma(s_i(v))\Phi(s_i(v))-1$ in (4.18) by repeated application of the three reductions given by $C(q)$, $\widehat{C}(\widehat{q})$, and $IC\widehat{C}(Q)$. The order conditions reduced by $ID\widehat{D}(R)$-$\widehat{D}(\widehat{r})$-(S) $(\widehat{r}\ge1)$, $D(r)$, and $\widehat{D}(\widehat{r})$ to those of the bushy trees can also be eliminated. But a linear combination of various terms of the form

$$\sum_{l=0}^{\varrho(v)-m}\mu_v(v_l)(-1)^{\varrho(v_l)}\binom{\varrho(v)}{\varrho(v_l)}\Big(\gamma(v_l)\Phi(v_l)-1\Big) \tag{4.19}$$

remains, where $m=\varrho(v_0)=\varrho(u)+1$, $\Phi(v_l)=\sum_{i=1}^{s}b_ic_i^l\Phi_i(u)$ with $u\in DAT3_u$ satisfying $\varrho(u)\ge\min(Q-1,q,\widehat{q}+1)$, $\varrho(v_l)=m+l$, and $\gamma(v_l)=\gamma(v_0)(m+l)/m$. In general several trees u exist which are not reducible by $C(q)$, $\widehat{C}(\widehat{q})$, and $IC\widehat{C}(Q)$. The coefficients $\mu_v(v_l)$ count the number of times that its multiplicand in (4.19) appears in the sum (4.18) after reduction by $C(q)$, $\widehat{C}(\widehat{q})$, and $IC\widehat{C}(Q)$. For u fixed we have the relations $\mu_v(v_l)=\binom{m+l-1}{l}\cdot\mu_v(v_0)$ which are related to the number of labellings of v. The term corresponding to $l=0$ can be isolated in (4.19) giving

$$\begin{aligned}\sum_{l=1}^{\varrho(v)-m}\Big[\mu_v(v_l)(-1)^{m+l}\binom{\varrho(v)}{m+l}\Big(\gamma(v_l)\Phi(v_l)-1\Big)\Big]+\\ \mu_v(v_0)(-1)^m\binom{\varrho(v)}{m}\Big(\gamma(v_0)\Phi(v_0)-1\Big)\,.\end{aligned} \tag{4.20}$$

Since we have the relation (see (4.17))

$$1=-\sum_{l=1}^{\varrho(v)-m}(-1)^l\binom{\varrho(v)-m}{l}\,, \tag{4.21}$$

(4.20) can be rewritten

$$\begin{aligned}\mu_v(v_0)\sum_{l=1}^{\varrho(v)-m}\Big[(-1)^{m+l}\binom{m+l-1}{l}\binom{\varrho(v)}{m+l}\Big(\frac{m+l}{m}\gamma(v_0)\Phi(v_l)-1\Big)-\\ (-1)^{m+l}\binom{\varrho(v)-m}{l}\binom{\varrho(v)}{m}\Big(\gamma(v_0)\Phi(v_0)-1\Big)\Big]\,.\end{aligned} \tag{4.22}$$

It is easy to show that

$$\binom{m+l-1}{l}\binom{\varrho(v)}{m+l}\frac{m+l}{m}=\binom{\varrho(v)-m}{l}\binom{\varrho(v)}{m}=\binom{\varrho(v)}{l,\ \varrho(v)-m-l,\ m}\,, \tag{4.23}$$

hence (4.20) can finally be rewritten

$$\mu_v(v_0)\sum_{l=1}^{\varrho(v)-m}(-1)^{m+l}\binom{\varrho(v)}{l,\ \varrho(v)-m-l,\ m}\Big(\gamma(v_0)\big(\Phi(v_l)-\Phi(v_0)\big)+\frac{l}{m+l}\Big)\,. \tag{4.24}$$

Now we will show that the expressions $\gamma(v_0)\big(\Phi(v_l)-\Phi(v_0)\big)+l/(m+l)$ in (4.24) vanish. With the help of $ID\widehat{D}(R)$-$I\widehat{D}(1)$-(S) we obtain for $l\geq 2$

$$\begin{aligned}
&\gamma(v_0)\big(\Phi(v_l)-\Phi(v_0)\big)+\frac{l}{m+l}\\
&\quad=\gamma(v_0)\left(\sum_{i,j=1}^{s} b_i c_i^l \omega_{ij}\varphi_j-\sum_{i,j=1}^{s} b_i\omega_{ij}\varphi_j\right)+\frac{l}{m+l}\\
&\quad=\gamma(v_0)\left(\sum_{i,j=1}^{s} b_i\omega_{ij}\varphi_j+l(l-1)\sum_{j=1}^{s} b_j c_j^{l-2}\varphi_j-l\varphi_s-\sum_{i=1}^{s} b_i\omega_{ij}\varphi_j\right)+\frac{l}{m+l}\\
&\quad=\gamma(v_0)\left(l(l-1)\sum_{j=1}^{s} b_j c_j^{l-2}\varphi_j-l\varphi_s\right)+\frac{l}{m+l}
\end{aligned}\tag{4.25}$$

where $\varphi_j=\sum_{j_1,\ldots,j_\pi=1}^{s} a_{jj_1}\Phi_{j_1}(t_1)\cdots a_{jj_\pi}\Phi_{j_\pi}(t_\pi)$ if u is of the form $[t_1,\ldots,t_\pi]_u$. For $l=1$ we have

$$\gamma(v_0)\big(\Phi(v_1)-\Phi(v_0)\big)+\frac{1}{m+1}=-\gamma(v_0)\varphi_s+\frac{1}{m+1}\tag{4.26}$$

Since $l+m(\leq\ell)\leq k$ we have (for $l\geq 2$)

$$\sum_{j=1}^{s} b_j c_j^{l-2}\varphi_j=\frac{1}{(m+l)(m+1)\gamma(v_0)}\tag{4.27}$$

and by (S)

$$\varphi_s=\frac{1}{(m+1)\gamma(v_0)}\,,\tag{4.28}$$

because these relations correspond to order conditions of $DAT3_y$-trees of order less than k which are satisfied as shown before. Inserting these relations into (4.25) gives

$$\gamma(v_0)\big(\Phi(v_l)-\Phi(v_0)\big)+\frac{l}{m+l}=0\qquad\text{for } l=1,\ldots,\varrho(v)-m,\tag{4.29}$$

which finally implies that $\mathrm{a}(v)=0$. The value of ℓ in (4.11b) follows from the "first worse" order conditions which cannot be reduced with simplifying assumptions. These order conditions correspond to the "first worse" trees which are of the following forms

(their corresponding order is mentioned)

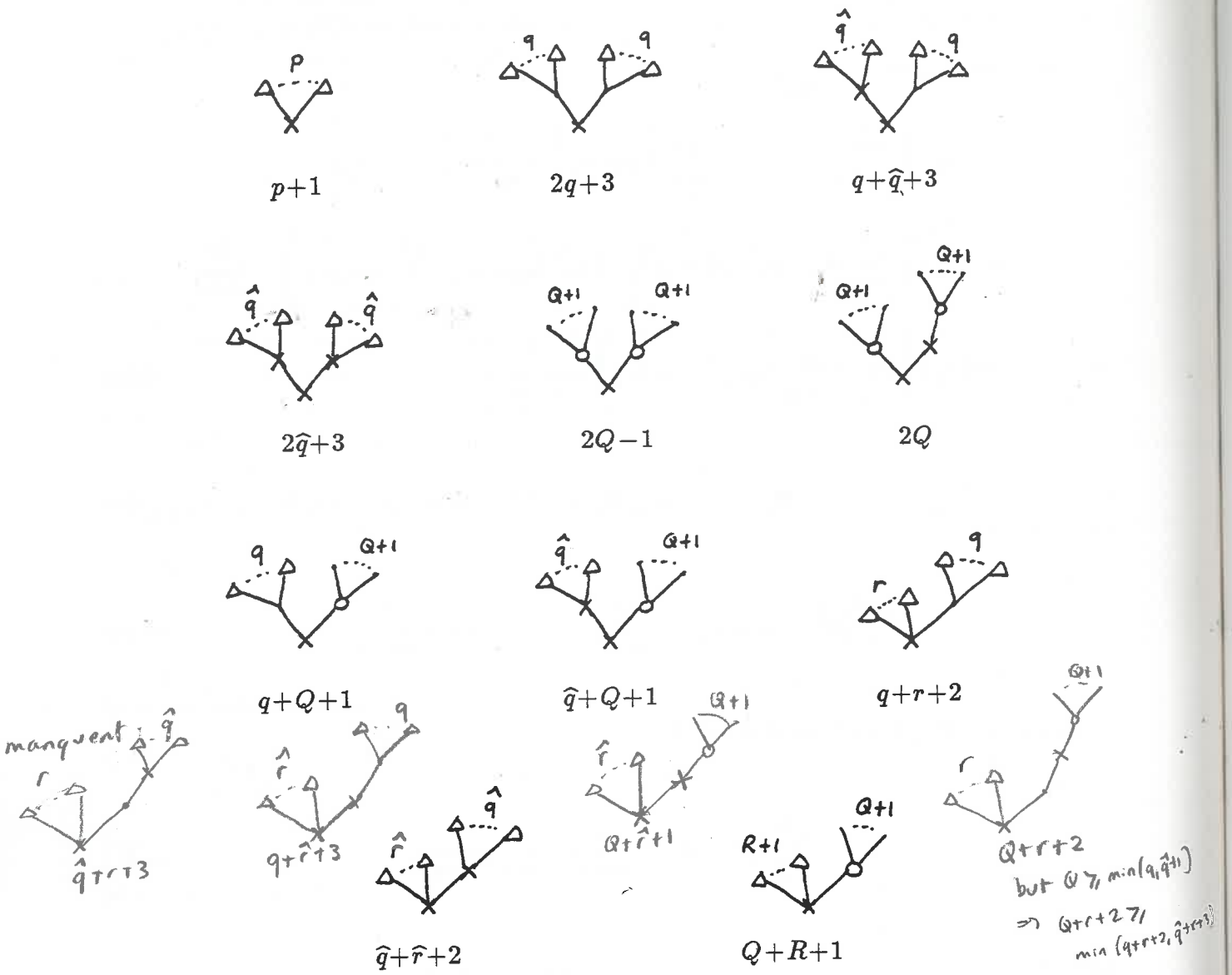


Figure 4.8. The "first worse" trees for the z-component.

Because of

$$q+Q+1 = q+1+Q-1/2+1/2 \geq 2\min(q+1, Q-1/2)+1/2 = \min(2q+2, 2Q-1)+1/2\ , \quad (4.30)$$

we get $q+Q+1 \geq \min(2q+3, 2Q)$. We also have $\hat{q}+Q+1 \geq \min(2\hat{q}+3, 2Q)$ and as previously $q+\hat{q}+3 \geq \min(2q+3, 2\hat{q}+3)$ is verified. This achieves the proof of the estimate (4.11b). □

Example 4.3. We illustrate the proof of $\mathbf{a}(v) = 0$ for the tree v on the right-hand side of Fig. II.3.3. This tree possesses 40 monotonic labellings (see Example II.4.3.2) and a tedious calculation shows that

$$\sum_{\text{labellings of } v} \left(\sum_{j=0}^{\varrho(v)} (-1)^j \binom{\varrho(v)}{j} \left(\gamma(s_j(v)) \Phi(\bar{s}_j(v)) - 1 \right) \right) = \quad (4.31)$$

$$40\binom{5}{0}(1\cdot 1-1)-40\binom{5}{1}\left(1\sum b_i-1\right)+22\binom{5}{2}\left(\frac{1}{3}\sum b_i\omega_{ij}c_j^3-1\right)+$$
$$18\binom{5}{2}\left(\frac{2}{3}\sum b_i\omega_{ij}c_ja_{jk}c_k-1\right)-8\binom{5}{3}\left(\frac{1}{12}\sum b_i\omega_{ij}c_j^3\omega_{ik}c_k^3-1\right)-$$
$$12\binom{5}{3}\left(\frac{1}{2}\sum b_i\omega_{ij}c_j^2a_{jk}c_k-1\right)-8\binom{5}{3}\left(\frac{1}{4}\sum b_i\omega_{ij}c_ja_{jk}\widehat{a}_{kl}\omega_{lm}c_m^3-1\right)-$$
$$12\binom{5}{3}\left(\frac{1}{6}\sum b_i\omega_{ij}c_j^3\omega_{ik}c_ka_{kl}c_l-1\right)+18\binom{5}{4}\left(\frac{1}{9}\sum b_i\omega_{ij}c_j^3\omega_{ik}c_k^2a_{kl}c_l-1\right)+$$
$$12\binom{5}{4}\left(\frac{1}{18}\sum b_i\omega_{ij}c_j^3\omega_{ik}c_ka_{kl}\widehat{a}_{lm}\omega_{mn}c_n^3-1\right)+$$
$$10\binom{5}{4}\left(\frac{1}{5}\sum b_i\omega_{ij}c_j^2a_{jk}\widehat{a}_{kl}\omega_{lm}c_m^3-1\right)-$$
$$40\binom{5}{5}\left(\frac{1}{24}\sum b_i\omega_{ij}c_j^3\omega_{ik}c_k^2a_{kl}\widehat{a}_{lm}\omega_{mn}c_n^3-1\right).$$

If we suppose that $C(3)$, $\widehat{C}(2)$, and $IC\widehat{C}(4)$ are satisfied, this expression reduces to

$$-40\binom{5}{1}\left(1\sum b_i-1\right)+22\binom{5}{2}\left(2\sum b_ic_i-1\right)+18\binom{5}{2}\left(2\sum b_ic_i-1\right)-$$
$$8\binom{5}{3}\left(3\sum b_ic_i^2-1\right)-12\binom{5}{3}\left(3\sum b_ic_i^2-1\right)-8\binom{5}{3}\left(3\sum b_ic_i^2-1\right)-$$
$$12\binom{5}{3}\left(3\sum b_ic_i^2-1\right)+18\binom{5}{4}\left(4\sum b_ic_i^3-1\right)+12\binom{5}{4}\left(4\sum b_ic_i^3-1\right)+$$
$$10\binom{5}{4}\left(\frac{1}{5}\sum b_i\omega_{ij}c_j^5-1\right)-40\binom{5}{5}\left(\frac{1}{4}\sum b_ic_i\omega_{ij}c_j^5-1\right). \tag{4.32}$$

If we also assume that $B(4)$ holds, it only remains the last two terms which can be rewritten as in (4.24)

$$-10\frac{5!}{1!\;0!\;4!}\left(\frac{1}{5}\left(\sum b_ic_i\omega_{ij}c_j^5-\sum b_i\omega_{ij}c_j^5\right)+\frac{1}{4+1}\right). \tag{4.33}$$

If we suppose that $I\widehat{D}(1)$ and (S) are satisfied then we finally obtain 0.

The results of the preceding theorem part b) show that the PRK methods satisfying the simplifying assumption (S) are of great interest. For such methods the numerical solution y_1 lies on the manifold $g(y)=0$ and the local error of the z- and u-components can be greatly improved if the numerical solution is projected onto the manifolds $(g_yf)(y,z)=0$ and $(g_{yy}(f,f)+g_yf_yf+g_yf_zk)(y,z,u)=0$ (see Subsection 1.3). These projections are even recommended (see Remark 2.4.1).

Theorem 4.4.

a) Let us suppose that the assumptions of Theorem 4.3 part a) are satisfied, then the local error of the projected values $\widetilde{y}_1,\widetilde{z}_1,\widetilde{u}_1$ given by (2.27a, c, e) satisfies

$$\widetilde{\delta y_h}(x_0):=\widetilde{y}_1-y(x_0+h)=\mathcal{O}(h^{\mu+1}),\quad \widetilde{\delta z_h}(x_0):=\widetilde{z}_1-z(x_0+h)=\mathcal{O}(h^{\nu+1}),$$
$$\widetilde{\delta u_h}(x_0):=\widetilde{u}_1-u(x_0+h)=\mathcal{O}(h^{\nu+1}) \tag{4.34}$$

where

$$\mu = \min(p, q+1, \hat{q}+1), \qquad \nu = \min(p, q, \hat{q}+1) \ . \tag{4.35}$$

b) Moreover, if in addition the assumptions of Theorem 4.3 part b) are satisfied, then we obtain

$$\begin{aligned} \mu &= \min(p, 2q+2, 2\hat{q}+2, q+r+1, \hat{q}+\hat{r}+1, 2Q-1, Q+\hat{r}, Q+R) \ , \\ \nu &= \min(p, 2q-2, 2\hat{q}, q+r+1, \hat{q}+\hat{r}+1, Q+\hat{r}, Q+R) \ . \end{aligned} \tag{4.36}$$

Remarks 4.2.

1) Identical results hold for the second projection procedure (1.12) (see Remark 2.4.2).
2) If the function k of (II.1.1) is linear in u then the values of μ and ν in (4.36) change to

$$\begin{aligned} \mu &= \min(p, 2q+2, 2\hat{q}+2, q+r+1, \hat{q}+\hat{r}+1, 2Q, Q+\hat{r}, Q+R) \ , \\ \nu &= \min(p, 2q-1, 2\hat{q}+1, q+r+1, \hat{q}+\hat{r}+1, Q+\hat{r}, Q+R) \ . \end{aligned} \tag{4.36'}$$

Proof. The results of part a) are a direct consequence of (4.10).

Concerning part b), similarly to the proof of [HaWa91, Theorem VI.7.2], we split $\widetilde{\delta z_h}(x_0)$ according to

$$\widetilde{\delta z_h}(x_0) = P_z^1 \widetilde{\delta z_h}(x_0) + Q_z^1 \widetilde{\delta z_h}(x_0) \tag{4.37}$$

where P_z^1 and Q_z^1 are evaluated at (y_1, z_1, u_1). From (2.27c) we obtain

$$\begin{aligned} P_z^1 \widetilde{\delta z_h}(x_0) = P_z^1 \delta z_h(x_0) = & P_z(x_0+h)\delta z_h(x_0) + \\ & \mathcal{O}\Big(\|\delta z_h(x_0)\| \cdot (\|\delta y_h(x_0)\| + \|\delta z_h(x_0)\| + \|\delta u_h(x_0)\|) \Big) \ . \end{aligned} \tag{4.38}$$

We also have

$$\begin{aligned} 0 &= (g_y f)(y_1, \widetilde{z}_1) - (g_y f)(y(x_0+h), z(x_0+h)) \\ &= (g_y f_z)(y_1, z_1)\widetilde{\delta z_h}(x_0) + \mathcal{O}\Big(\|\delta y_h(x_0)\| + \|\delta z_h(x_0)\| \cdot \|\widetilde{\delta z_h}(x_0)\| + \|\widetilde{\delta z_h}(x_0)\|^2 \Big) \end{aligned} \tag{4.39}$$

which yields

$$Q_z^1 \widetilde{\delta z_h}(x_0) = \mathcal{O}\left(\|\delta y_h(x_0)\| + \|\delta z_h(x_0)\| \cdot \|\widetilde{\delta z_h}(x_0)\| + \|\widetilde{\delta z_h}(x_0)\|^2 \right) \ . \tag{4.40}$$

The above results lead to

$$\begin{aligned} \widetilde{\delta z_h}(x_0) = & P_z(x_0+h)\delta z_h(x_0) + \mathcal{O}\Big(\|\delta y_h(x_0)\| + \|\delta z_h(x_0)\|^2 + \|\delta u_h(x_0)\| \cdot \|\delta z_h(x_0)\| + \\ & \|\delta z_h(x_0)\| \cdot \|\widetilde{\delta z_h}(x_0)\| + \|\widetilde{\delta z_h}(x_0)\|^2 \Big) \ . \end{aligned} \tag{4.41}$$

Hence with the help of the estimates given in Theorem 4.3, we obtain the desired result for the z-component. For the u-component the result simply follows from

$$\widetilde{\delta u_h}(x_0) = \mathcal{O}\Big(\|\delta y_h(x_0)\| + \|\widetilde{\delta z_h}(x_0)\| \Big) \ . \tag{4.42}$$

□

5. Convergence of projected PRK methods.

In this section we give estimates for the global error of *projected PRK methods*, i.e., of PRK methods for which a projection procedure of Subsection 1.3 is applied to all components after every step. The following theorem is one of the main results of this chapter.

Theorem 5.1. *Consider the differential-algebraic system* (II.1.1) *with consistent initial values* $\Psi_0=(y_0,z_0,u_0)$ *at* x_0, *a PRK method* (1.1) *satisfying the hypotheses of Theorem 4.4, and a projection procedure of Subsection 1.3. Then for* $x_n-x_0=nh\le Const$, *the global error of the projected values* $\widetilde{y}_n,\widetilde{z}_n,\widetilde{u}_n$ *after* n *steps satisfies*

$$\widetilde{y}_n-y(x_n)=\mathcal{O}(h^\nu),\qquad \widetilde{z}_n-z(x_n)=\mathcal{O}(h^\nu),\qquad \widetilde{u}_n-u(x_n)=\mathcal{O}(h^\nu) \tag{5.1}$$

where the value of ν *is given in Theorem 4.4.*

Remarks 5.1.

1) The theorem remains valid in the case of variable stepsizes with $h=\max_i h_i$.
2) For the y- and z-components, (5.1) holds even if the numerical u-component is not defined with the help of $(g_{yy}(f,f)+g_yf_yf+g_yf_zk)(y,z,u)=0$ (see Remark 2.1.2). In this case the estimate (5.1) for the u-component can be recovered if such a projection is performed for the last step.
3) In contrast with [HaLuRo89a, Theorem 6.4], $|R(\infty)|\le 1$ and $|\widehat{R}(\infty)|\le 1$ are not supposed.

Proof. We denote two neighbouring projected PRK solutions by $\{\overline{y}_n,\overline{z}_n\}$, $\{\widehat{y}_n,\widehat{z}_n\}$ and their difference by $\Delta y_n=\overline{y}_n-\widehat{y}_n$, $\Delta z_n=\overline{z}_n-\widehat{z}_n$. We suppose for the moment that

$$\|\widehat{y}_n-y(x_n)\|\le C_0h^2\ ,\ \|\widehat{z}_n-z(x_n)\|\le C_0h\ ,\ \|\Delta y_n\|\le C_1h^3\ ,\ \|\Delta z_n\|\le C_1h^2 \tag{5.2}$$

(this will be justified below). Because of $g(\overline{y}_n)=0=g(\widehat{y}_n)$ and $(g_yf)(\overline{y}_n,\overline{z}_n)=0=(g_yf)(\widehat{y}_n,\widehat{z}_n)$, Remark 2.2.4 holds, implying that

$$(Q_y)_n\Delta y_n=\mathcal{O}(\|\Delta y_n\|^2)=\mathcal{O}(h^2\|(P_y)_n\Delta y_n\|)\ , \tag{5.3a}$$

$$(Q_z)_n\Delta z_n=\mathcal{O}(\|\Delta y_n\|+\|\Delta z_n\|^2)=\mathcal{O}(\|(P_y)_n\Delta y_n\|+h^2\|(P_z)_n\Delta z_n\|)\ . \tag{5.3b}$$

Theorem 2.4 can be applied with $\delta=0$, $\delta_{s+1}=0$, $\mu=0$, $\mu_{s+1}=0$, $\theta=0$, $\theta_{s+1}=0$, $\theta'_{s+1}=0$, and $\theta''_{s+1}=0$ yielding

$$(P_y)_{n+1}\Delta y_{n+1}=(P_y)_n\Delta y_n+\mathcal{O}\left(h\|(P_y)_n\Delta y_n\|+h\|(P_z)_n\Delta z_n\|\right), \tag{5.3c}$$

$$(P_z)_{n+1}\Delta z_{n+1}=(P_z)_n\Delta z_n+\mathcal{O}\left(h\|(P_y)_n\Delta y_n\|+h\|(P_z)_n\Delta z_n\|\right)\ . \tag{5.3d}$$

In (5.3) $(P_y)_n$, $(Q_y)_n$, $(P_z)_n$, and $(Q_z)_n$ are evaluated at $(\widehat{y}_n,\widehat{z}_n,\widehat{u}_n)$. These estimates (5.3) lead to

$$\|\Delta y_n\|\le C\left(\|(P_y)_0\Delta y_0\|+\|(P_z)_0\Delta z_0\|\right)\ , \tag{5.4a}$$

$$\|\Delta z_n\|\le C\left(\|(P_y)_0\Delta y_0\|+\|(P_z)_0\Delta z_0\|\right)\ . \tag{5.4b}$$

Hence the result (5.1) follows from standard techniques (see [HaLuRo89a, Fig. 4.1, p. 36] or [HaNøWa93, Fig. II.3.2, p. 160]). The assumption (5.2) is justified by induction on n provided the constants C_0 and C_1 are chosen sufficiently large and h is sufficiently small. □

6. Solution of the nonlinear system by simplified Newton iterations.

At each step of the application of a PRK method we have to solve a nonlinear system of the form (2.2a,b,c) which can be rewritten as follows

$$0 = H^y(Y,Z) = Y - \mathbb{1}\otimes\eta - h(A\otimes I)F(Y,Z)\,, \tag{6.1a}$$

$$0 = H^z(Y,Z,U) = Z - \mathbb{1}\otimes\zeta - h(\widehat{A}\otimes I)K(Y,Z,U)\,, \tag{6.1b}$$

$$0 = H^u(Y) = -hG(Y) \tag{6.1c}$$

where

$$\mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},\quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_s \end{pmatrix},\quad Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_s \end{pmatrix},\quad U = \begin{pmatrix} U_1 \\ \vdots \\ U_s \end{pmatrix},\quad F(Y,Z) = \begin{pmatrix} f(Y_1,Z_1) \\ \vdots \\ f(Y_s,Z_s) \end{pmatrix},$$

$$K(Y,Z,U) = \begin{pmatrix} k(Y_1,Z_1,U_1) \\ \vdots \\ k(Y_s,Z_s,U_s) \end{pmatrix},\qquad G(Y) = \begin{pmatrix} g(Y_1) \\ \vdots \\ g(Y_s) \end{pmatrix}. \tag{6.2}$$

With the simplified Jacobian matrix

$$\begin{pmatrix} I & -hA\otimes f_z(\eta,\zeta) & O \\ O & I & -h\widehat{A}\otimes k_u(\eta,\zeta,\nu) \\ -hI\otimes g_y(\eta) & O & O \end{pmatrix}, \tag{6.3}$$

the simplified Newton iterations read

$$Y^{(k+1)} = Y^{(k)} + \Delta Y^{(k)},\qquad Z^{(k+1)} = Z^{(k)} + \Delta Z^{(k)},\qquad U^{(k+1)} = U^{(k)} + \Delta U^{(k)} \tag{6.4}$$

where, after application of a block-Gauss elimination, we have

$$\Delta U^{(k)} = \Big((A\widehat{A})^{-1}\otimes(g_y f_z k_u)^{-1}(\eta,\zeta,\nu)\Big)\Big(H^u(Y^{(k)}) + h(I\otimes g_y(\eta))H^y(Y^{(k)},Z^{(k)}) \tag{6.5a}$$

$$+ h^2(A\otimes(g_y f_z)(\eta,\zeta))H^z(Y^{(k)},Z^{(k)},U^{(k)})\Big)\,,$$

$$\Delta Z^{(k)} = -H^z(Y^{(k)},Z^{(k)},U^{(k)}) + h(\widehat{A}\otimes k_u(\eta,\zeta,\nu))\Delta U^{(k)}\,, \tag{6.5b}$$

$$\Delta Y^{(k)} = -H^y(Y^{(k)},Z^{(k)}) + h(A\otimes f_z(\eta,\zeta))\Delta Z^{(k)}\,. \tag{6.5c}$$

With starting values $Y_i^{(0)} = \eta + hc_i f(\eta,\zeta) + \mathcal{O}(h^2)$, $Z_i^{(0)} = \zeta + \mathcal{O}(h)$, and $U_i^{(0)} = \nu + \mathcal{O}(h)$, every simplified Newton iteration improves the approximation by a factor h in the norm $\|y\| + h\|z\| + h^2\|u\|$. For more details on this subject we refer to [HaLuRo89a, Section 7] (for index 1 problems see [Kv92]).

other possibility:

$$\Delta Z^{(k)} = -H^z(Y^{(k)}, Z^{(k)}, U^{(k)} + \Delta U^{(k)})$$

$$\Delta Y^{(k)} = -H^y(Y^{(k)}, Z^{(k)} + \Delta Z^{(k)})$$

Chapter IV. Convergence of Runge-Kutta methods for semi-explicit index 3 DAE's in Hessenberg form.

1. Introduction.

This chapter presents optimal convergence results for stiffly accurate RK methods when applied (direct approach, see Section I.6) to semi-explicit index 3 DAE's in Hessenberg form (see Chapter II). For solving such problems an index reduction is usually possible by differentiating the constraints, although some difficulties may occur (see Subsection I.5.1). However, for multibody systems containing very stiff springs, i.e., whose Hooke's constant $1/\varepsilon^2$ is very large, the numerical solution behaves like that for the limit problem ($\varepsilon \to 0$) which is of index 3 (see [HaLuRo89a, pp. 10-12], [Lu93], and Subsection I.4.2). In this situation an index reduction is not applicable and the convergence behaviour for the index 3 case (direct approach) must be studied. This remark remains valid for the equations of motion of very stiff mechanical systems in which a large potential forces the motion to be close to a manifold (see Subsection I.4.2).

Non-optimal orders of convergence of RK methods for semi-explicit index 3 DAE's in Hessenberg form have been demonstrated in [HaLuRo89a, Section 6] and sharper estimates have been numerically observed and hypothesized (see [HaLuRo89a, pp. 18-19 & 86]). The main result of this chapter (Theorem 6.1 below) is a proof of the conjecture of [HaLuRo89a, p. 86], giving sharp convergence bounds for stiffly accurate RK methods, such as the Lobatto IIIC and Radau IIA schemes. This result has an application in the convergence analysis of these methods when applied to stiff mechanical systems. Furthermore it extends the results of [Jay93b] for collocation methods to general RK methods, but with completely different techniques.

In this chapter we again consider semi-explicit index 3 DAE's in Hessenberg form (see Chapter II)

$$y' = f(y,z)\ , \qquad z' = k(y,z,u)\ , \qquad 0 = g(y) \tag{1.1a,b,c}$$

where the initial values (y_0, z_0, u_0) at x_0 are assumed to be *consistent*, i.e., they satisfy

$$0 = g(y)\ , \tag{1.1c}$$

$$0 = \left(g_y f\right)(y,z)\ , \tag{1.1d}$$

$$0 = \left(g_{yy}(f,f) + g_y f_y f + g_y f_z k\right)(y,z,u)\ . \tag{1.1e}$$

We suppose that

$$(g_y f_z k_u)(y,z,u) \quad \text{is invertible} \tag{1.2}$$

in a vicinity of the exact solution (*index* 3 assumption).

The application (direct approach) of Runge-Kutta methods to (1.1a,b,c) is presented in Section 2. Existence and uniqueness of the RK solution, and influence of perturbations are studied in Section 3. Section 4 deals with the calculation of expressions encountered in the preceding section and involving the RK coefficients. Estimates of the local error and of certain associated projections are then given in Section 5. With the help of the results contained in the previous sections, a global convergence theorem is presented in Section 6 proving the conjecture of [HaLuRo89a, p. 86]. An application of this result to the convergence analysis of certain RK methods for stiff mechanical systems is given. Finally, Section 7 includes some numerical experiments illustrating the theoretical results.

2. RK methods for semi-explicit index 3 DAE's in Hessenberg form.

Definition 2.1. One step of an *s-stage Runge-Kutta (RK)* method applied to (1.1a,b,c) (direct approach) reads (see Section I.6, [BrePe89, p. 75], and [HaLuRo89a, p. 71])

$$y_1 = y_0 + h\sum_{i=1}^{s} b_i Y_i' , \qquad z_1 = z_0 + h\sum_{i=1}^{s} b_i Z_i' , \qquad u_1 = u_0 + h\sum_{i=1}^{s} b_i U_i' \tag{2.1a}$$

where

$$Y_i' = f(Y_i, Z_i) , \qquad Z_i' = k(Y_i, Z_i, U_i) , \qquad 0 = g(Y_i) , \tag{2.1b}$$

and the *internal stages* are given by

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} Y_j' , \qquad Z_i = z_0 + h\sum_{j=1}^{s} a_{ij} Z_j' , \qquad U_i = u_0 + h\sum_{j=1}^{s} a_{ij} U_j' . \tag{2.1c}$$

We are interested in RK methods satisfying the hypotheses

(I) : the RK matrix A is invertible ;

(S) : the method is *stiffly accurate*, i.e., $a_{si} = b_i$ for $i = 1, \ldots, s$.

Remarks 2.1. The following results can be easily proved.

1) (S) implies that $y_1 = Y_s$, $g(y_1) = g(Y_s) = 0$, $z_1 = Z_s$, and $u_1 = U_s$ in (2.1).
2) (I) and (S) imply that $R(\infty) = 0$ where R is the stability function of the RK method (see [HaWa91, Proposition IV.3.8]).

In the following sections we will use the notations $C := \mathrm{diag}(c_1, \ldots, c_s)$ and $\mathbb{1} := (1, \ldots, 1)^T$ (with s components).

3. Existence, uniqueness of the RK solution, and influence of perturbations.

This section is devoted to the analysis of the solution of the nonlinear system (2.1) with (y_0, z_0, u_0) replaced by approximate h-dependent starting values $(\eta, \zeta, \nu) = (\eta(h), \zeta(h), \nu(h))$. An important result is given by Theorem 3.4 which will be useful in Section 6 to the study of the error propagation. We first investigate the existence and uniqueness of the RK solution.

Theorem 3.1. [HaLuRo89a, Theorem 6.1]. *Let us suppose that*

$$g(\eta) = \mathcal{O}(h^\tau) , \quad \tau \geq 3 , \tag{3.1a}$$

$$(g_y f)(\eta, \zeta) = \mathcal{O}(h^\kappa) , \quad \kappa \geq 2 , \tag{3.1b}$$

$$(g_{yy}(f,f) + g_y f_y f + g_y f_z k)(\eta, \zeta, \nu) = \mathcal{O}(h) , \tag{3.1c}$$

$$(g_y f_z k_u)(y, z, u) \text{ is invertible in a neighbourhood of } (\eta, \zeta, \nu) , \tag{3.1d}$$

and that $C(2)$ and (I) are fulfilled. Then for $h \le h_0$ there exists a locally unique solution to

$$Y_i = \eta + h\sum_{j=1}^{s} a_{ij} f(Y_j, Z_j)\ , \qquad Z_i = \zeta + h\sum_{j=1}^{s} a_{ij} k(Y_j, Z_j, U_j)\ , \tag{3.2a,b}$$

$$0 = g(Y_i) \tag{3.2c}$$

for $i = 1, \dots, s$, which satisfies

$$Y_i - \eta = \mathcal{O}(h)\ , \qquad Z_i - \zeta = \mathcal{O}(h)\ , \qquad U_i - \nu = \mathcal{O}(h)\ . \tag{3.3}$$

Remarks 3.1.

1) If the function k of (1.1) is linear in u then the assumptions $C(2)$ and (3.1c) can be omitted. In this situation, $\tau \ge 2$ and $\kappa \ge 1$ are sufficient. However, if $C(2)$ is not satisfied, $\tau = 2$ or $\kappa = 1$, we only have the estimate $U_i - \nu = \mathcal{O}(1)$ (for more details see Theorem III.2.1, Lemma 3.3, and [HaLuRo89a, p. 74]).
2) The value of ν in (3.1c) only prescribes the solution of (3.2) to be close to the manifold defined by (1.1e). However, (Y_i, Z_i, U_i) are clearly independent of ν.
3) If the function k of (1.1) is not linear in u then $C(2)$ and (I) show the necessity of having $s \ge 2$.

Proof. By a simple application of Theorem III.2.1 with $\widehat{A} = A$. An alternative proof is given in [HaLuRo89a, Theorem 6.1]. □

The next result, a more complete and precise formulation of [HaLuRo89a, Theorem 6.2], is concerned with the influence of perturbations to (3.2). This is a simple application of Theorem III.2.2 with $\widehat{A} = A$.

Theorem 3.2. *Let (Y_i, Z_i, U_i) be given by (3.2) and let us consider perturbed values $(\widehat{Y}_i, \widehat{Z}_i, \widehat{U}_i)$ satisfying*

$$\widehat{Y}_i = \widehat{\eta} + h\sum_{j=1}^{s} a_{ij} f(\widehat{Y}_j, \widehat{Z}_j) + h\delta_i\ , \qquad \widehat{Z}_i = \widehat{\zeta} + h\sum_{j=1}^{s} a_{ij} k(\widehat{Y}_j, \widehat{Z}_j, \widehat{U}_j) + h\mu_i\ , \tag{3.4a,b}$$

$$0 = g(\widehat{Y}_i) + \theta_i \tag{3.4c}$$

for $i = 1, \dots, s$. In addition to the assumptions of Theorem 3.1 let us suppose that

$$\begin{aligned} \Delta\eta &= \mathcal{O}(h^3)\ , & \Delta\zeta &= \mathcal{O}(h^2)\ , & \widehat{U}_i - \nu &= \mathcal{O}(h)\ , \\ \delta_i &= \mathcal{O}(h^2)\ , & \mu_i &= \mathcal{O}(h)\ , & \theta_i &= \mathcal{O}(h^3)\ . \end{aligned} \tag{3.5}$$

Then we have for $h \le h_0$ the estimates

$$\Delta Y_i = P_y \Delta\eta + hc_i f_z P_z \Delta\zeta + \mathcal{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 \tag{3.6a}$$

$$+ h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big)\ ,$$

$$\Delta Z_i = -\frac{1}{h}\sigma_i \cdot SQ_y\Delta\eta + P_z\Delta\zeta + \mathcal{O}\Big(\|\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + \|\delta\| + h\|\mu\| + \frac{1}{h}\|\theta\|\Big)\,, \tag{3.6b}$$

$$P_{z,i}\Delta Z_i = P_z\Delta\zeta + \mathcal{O}\Big(\|Q_y\Delta\eta\| + h\|P_y\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\mu\| + \|\theta\|\Big)\,, \tag{3.6c}$$

$$\Delta U_i = \mathcal{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\| + \frac{1}{h}\|P_y\Delta\eta\| + \frac{1}{h}\|Q_z\Delta\zeta\| + \|P_z\Delta\zeta\| + \frac{1}{h}\|\delta\| + \|\mu\| + \frac{1}{h^2}\|\theta\|\Big) \tag{3.6d}$$

where $\sigma_i = e_i^T A^{-1}\mathbb{1}$ *with* $e_i = (0,\dots,0,1,0,\dots,0)^T$ *(the* s*-dimensional vector with all components equal to 0 excepted the* i*th which is equal to 1),* $\delta = (\delta_1,\dots,\delta_s)^T$, $\|\delta\| = \max_i \|\delta_i\|$, *and similarly for* μ *and* θ. P_y, Q_y, P_z, *and* Q_z *are projectors defined under the condition* (1.2) *by*

$$\begin{aligned} S &:= k_u(g_y f_z k_u)^{-1} g_y\,, \\ Q_y &:= f_z S\,, \qquad P_y := I - Q_y\,, \qquad Q_z := Sf_z\,, \qquad P_z := I - Q_z\,. \end{aligned} \tag{3.7}$$

□

Our next aim in Theorem 3.4 is to show that the estimates (3.6a, c) can be improved for $i = s$. The following lemma will be useful in the proof of this theorem.

Lemma 3.3. *Besides the hypotheses of Theorem 3.1, let us suppose further that* $C(q)$ *holds. Then the solution* (Y_i, Z_i, U_i) *of* (3.2) *satisfies*

$$Y_i = \widetilde{\eta} + \sum_{m=1}^{\lambda} \frac{c_i^m h^m}{m!} D_m Y(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu}) + \mathcal{O}(h^{\lambda+1})\,, \tag{3.8a}$$

$$Z_i = \widetilde{\zeta} + \sum_{n=1}^{\gamma} \frac{c_i^n h^n}{n!} D_n Z(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu}) + \mathcal{O}(h^{\gamma+1})\,, \tag{3.8b}$$

$$U_i = \widetilde{\nu} + \sum_{p=1}^{\mu} \frac{c_i^p h^p}{p!} D_p U(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu}) + \mathcal{O}(h^{\mu+1}) \tag{3.8c}$$

where $\lambda = \min(\tau, \kappa+1, q)$, $\gamma = \min(\tau-2, \kappa, q-1)$, $\mu = \min(\tau-3, \kappa-2, q-2)$, *and* $D_m Y$, $D_n Z$, $D_p U$ *are functions composed with derivatives of* f, g, *and* k. $(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$ *are consistent values close to* (η,ζ,ν) *but constructed independently of* ν. *They are uniquely determined by the equations* (1.1c,d,e), $P_y(\eta,\zeta,\nu^*)(\widetilde{\eta}-\eta) = 0$, *and* $P_z(\eta,\zeta,\nu^*)(\widetilde{\zeta}-\zeta) = 0$ *with* $\nu^* := G(\eta,\zeta)$.

Proof. We find $Q_y(\eta,\zeta,\nu^*)(\widetilde{\eta}-\eta) = \mathcal{O}(h^\tau)$ and $Q_z(\eta,\zeta,\nu^*)(\widetilde{\zeta}-\zeta) = \mathcal{O}(h^{\min(\tau,\kappa)})$. We define $(y(x), z(x), u(x))$ the solution of (1.1) with initial values $y(x_0) = \widetilde{\eta}$, $z(x_0) = \widetilde{\zeta}$, and $u(x_0) = \widetilde{\nu}$. The exact solution at $x_0 + c_i h$ satisfies (3.4) with $\theta_i = 0$ and

$$\delta_i = \frac{h^q}{q!} y^{(q+1)}(x_0)\Big(\frac{c_i^{q+1}}{q+1} - \sum_{j=1}^{s} a_{ij} c_j^q\Big) + \mathcal{O}(h^{q+1}) = \mathcal{O}(h^q)\,, \tag{3.9a}$$

$$\mu_i = \frac{h^q}{q!} z^{(q+1)}(x_0)\Big(\frac{c_i^{q+1}}{q+1} - \sum_{j=1}^{s} a_{ij} c_j^q\Big) + \mathcal{O}(h^{q+1}) = \mathcal{O}(h^q)\,. \tag{3.9b}$$

The difference from the numerical solution (3.2) can thus be estimated with Theorem 3.2, yielding

$$\|Y_i - y(x_0+c_ih)\| = \mathcal{O}(h^{\min(\tau+1,\kappa+2,q+1)}) , \tag{3.10a}$$

$$\|Z_i - z(x_0+c_ih)\| = \mathcal{O}(h^{\min(\tau-1,\kappa+1,q)}) , \tag{3.10b}$$

$$\|U_i - u(x_0+c_ih)\| = \mathcal{O}(h^{\min(\tau-2,\kappa-1,q-1)}) . \tag{3.10c}$$

□

Here is the main result of this section.

Theorem 3.4. *In addition to the assumptions of Theorem 3.2 and Lemma 3.3, including those of Theorem 3.1, let us suppose that $B(1)$, $D(r)$, and (S) hold. Then we have*

$$\Delta Y_s = P_y\Delta\eta + hf_zP_z\Delta\zeta + \mathcal{O}\Big(h\|\Delta\eta\| + h^2\|P_z\Delta\zeta\| + h^{m+2}\|Q_z\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big) , \tag{3.11a}$$

$$hP_{z,s}\Delta Z_s = hP_z\Delta\zeta + \mathcal{O}\Big(h\|Q_y\Delta\eta\| + h^2\|P_y\Delta\eta\| + h^2\|P_z\Delta\zeta\| + h^{n+2}\|Q_z\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h^2\|\delta\| + h^2\|\mu\| + h\|\theta\|\Big) , \tag{3.11b}$$

$$hQ_{z,s}\Delta Z_s = -\sigma\cdot SQ_y\Delta\eta + \mathcal{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big) \tag{3.11c}$$

where $\sigma = b^TA^{-2}\mathbb{1}$, $m = \min(\tau-3, \kappa-2, q-2, \max(r-1, 0))$, $n = \min(\tau-3, \kappa-2, q-2, r)$. P_y, Q_y, P_z, Q_z, S, and f_z are evaluated at (η, ζ, ν^*) with ν^* defined as in Lemma 3.3. The arguments of $P_{z,s}$ and $Q_{z,s}$ are $(Y_s, Z_s, G(Y_s, Z_s))$ or (Y_s, Z_s, U_s).

Remarks 3.2.

1) If the function k of (1.1) is linear in u, then $m = \min(\tau-2, \kappa-1, q-1, \max(r-1, 0))$, $n = \min(\tau-2, \kappa-1, q-1, r)$. Remark III.2.2.5 also holds here.
2) The important results consist in the splitting of $\Delta\zeta$ according to the projections P_z and Q_z, and in the h-exponents in front of $\|Q_z\Delta\zeta\|$ in (3.11*a*, *b*).
3) It must be stressed that m and n satisfy $0 \le m \le n \le m+1$.
4) In the proof the missing arguments for f_z, P_y, P_z, etc., are $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$ defined in Lemma 3.3. At the end of the proof a final estimate shows that they can be replaced by (η, ζ, ν^*).
5) Remark III.2.2.6 is also valid here for almost all constants entering in the $\mathcal{O}$-expressions of (3.11). The exceptions are the constants implied by the $\mathcal{O}$-terms in (3.11*a*, *b*) $\mathcal{O}(h^{m+2}\|Q_z\Delta\zeta\|)$ and $\mathcal{O}(h^{n+2}\|Q_z\Delta\zeta\|)$ which depend on those of (3.1*a*, *b*) if m or $n \ge 1$. Nevertheless, this will not affect the proof of Theorem 6.1 (see Section 6) where Theorem 3.4 will be applied.

Proof. We resume the proof of Theorem 3.2 (see Theorem III.2.2) with the help of Lemma 3.3, using the same notations and definitions, except for

$$G_y := \{g_y\} , \quad F_z := (A\otimes I)\{f_z\}(A\otimes I)^{-1} , \quad K_u := (A\otimes I)^2\{k_u\}(A\otimes I)^{-2} ,$$

$$S_A := K_u(G_yF_zK_u)^{-1}G_y\ , \qquad (3.12)$$
$$Q_{y,A} := F_zS_A\ ,\quad P_{y,A} := I-Q_{y,A}\ ,\quad Q_{z,A} := S_AF_z\ ,\quad P_{z,A} := I-Q_{z,A}\ .$$

Because of

$$\Delta Y_s = (e_s^T\otimes I)\Delta Y\ ,\ P_{z,s}\Delta Z_s = (e_s^T\otimes P_{z,s})\Delta Z\ ,\ Q_{z,s}\Delta Z_s = (e_s^T\otimes Q_{z,s})\Delta Z\ , \quad (3.13)$$

(3.11) is a simple consequence of (3.6a,b,c) with the exception of the h-exponents in front of $\|Q_z\Delta\zeta\|$ in (3.11a,b) which remain to be shown. They will be computed with similar techniques used in the proof of [Jay93a, Theorem 4.4].

The formulas coming from (III.2.22b,c) read here

$$\begin{aligned} h\Delta Z = (A\otimes I)^{-1}\Big[&-S_A(\mathbb{1}\otimes\Delta\eta)+P_{z,A}(A\mathbb{1}\otimes h\Delta\zeta)-hS_A(A\otimes I)\{f_y\}\Delta Y \\ &+h^2P_{z,A}(A\otimes I)^2\{k_y\}\Delta Y+hP_{z,A}(A\otimes I)^2\{k_z\}h\Delta Z\Big] \\ &+\mathcal{O}(\|\Delta Y\|^2+h\|\Delta Z\|^2+h^2\|\Delta U\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|)\ , \end{aligned} \qquad (3.14a)$$

$$\begin{aligned} \Delta Y = &P_{y,A}(\mathbb{1}\otimes\Delta\eta)+F_zP_{z,A}(A\mathbb{1}\otimes h\Delta\zeta)+hP_{y,A}(A\otimes I)\{f_y\}\Delta Y \\ &+h^2P_{y,A}F_z(A\otimes I)^2\{k_y\}\Delta Y+hP_{y,A}F_z(A\otimes I)^2\{k_z\}h\Delta Z \\ &+\mathcal{O}(\|\Delta Y\|^2+h\|\Delta Z\|^2+h^2\|\Delta U\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|) \end{aligned} \qquad (3.14b)$$

and they can be rewritten

$$\begin{aligned} &\begin{pmatrix} I-hS-h^2T \end{pmatrix}\begin{pmatrix} \Delta Y \\ h\Delta Z \end{pmatrix} = V \\ &+\mathcal{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|^2+\|P_y\Delta\eta\|^2+\|Q_z\Delta\zeta\|^2+h\|P_z\Delta\zeta\|^2+h\|\delta\|+h^2\|\mu\|+\|\theta\|\Big) \end{aligned} \qquad (3.15)$$

where the matrices S, T, and the vector V are given by

$$S = \begin{pmatrix} P_{y,A}(A\otimes I)\{f_y\} & P_{y,A}F_z(A\otimes I)^2\{k_z\} \\ -(A\otimes I)^{-1}S_A(A\otimes I)\{f_y\} & (A\otimes I)^{-1}P_{z,A}(A\otimes I)^2\{k_z\} \end{pmatrix}\ , \qquad (3.16a)$$

$$T = \begin{pmatrix} P_{y,A}F_z(A\otimes I)^2\{k_y\} & O \\ (A\otimes I)^{-1}P_{z,A}(A\otimes I)^2\{k_y\} & O \end{pmatrix}\ , \qquad (3.16b)$$

$$V = \begin{pmatrix} P_{y,A}(\mathbb{1}\otimes\Delta\eta)+F_zP_{z,A}(A\mathbb{1}\otimes h\Delta\zeta) \\ (A\otimes I)^{-1}\Big[-S_A(\mathbb{1}\otimes\Delta\eta)+P_{z,A}(A\mathbb{1}\otimes h\Delta\zeta)\Big] \end{pmatrix}\ . \qquad (3.16c)$$

We put $K_{u,0} := I\otimes k_u(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$ corresponding to K_u with (Y_i,Z_i,U_i) replaced by $(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$. By the use of $\Delta\zeta = P_z\Delta\zeta+Q_z\Delta\zeta$, $Q_y = SQ_y$, $Q_z^2 = Q_z$, and $P_{z,A}K_u = 0$, V can be estimated by

$$V = W+\begin{pmatrix} \mathbb{1}\otimes P_y\Delta\eta+A\mathbb{1}\otimes hf_zP_z\Delta\zeta+\mathcal{O}(h\|\Delta\eta\|+h^2\|P_z\Delta\zeta\|) \\ -A^{-1}\mathbb{1}\otimes Q_y\Delta\eta+\mathbb{1}\otimes hP_z\Delta\zeta+\mathcal{O}(h\|\Delta\eta\|+h^2\|P_z\Delta\zeta\|) \end{pmatrix} \qquad (3.17a)$$

where we have isolated in W the terms including $Q_z\Delta\zeta$

$$W = \begin{pmatrix} -F_z P_{z,A}(K_u - K_{u,0})(A\mathbb{1} \otimes h(g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta) \\ -(A \otimes I)^{-1} P_{z,A}(K_u - K_{u,0})(A\mathbb{1} \otimes h(g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta) \end{pmatrix} . \tag{3.17b}$$

Computing the inverse of the left matrix in (3.15) by means of the series of Von Neumann we arrive at

$$\begin{pmatrix} \Delta Y \\ h\Delta Z \end{pmatrix} = \sum_{\varrho=0}^{n} (hS + h^2 T)^{\varrho} W \tag{3.18}$$

$$+ \begin{pmatrix} \mathbb{1} \otimes P_y \Delta\eta + A\mathbb{1} \otimes h f_z P_z \Delta\zeta + \mathcal{O}(h\|\Delta\eta\| + h^2\|P_z\Delta\zeta\| + h^{n+2}\|Q_z\Delta\zeta\|) \\ -A^{-1}\mathbb{1} \otimes Q_y\Delta\eta + \mathbb{1} \otimes h P_z \Delta\zeta + \mathcal{O}(h\|\Delta\eta\| + h^2\|P_z\Delta\zeta\| + h^{n+2}\|Q_z\Delta\zeta\|) \end{pmatrix}$$

$$+ \mathcal{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|P_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|P_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big) .$$

Our next aim is to develop at the point $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$ the expression involving S, T, and W in (3.18) into h-powers. By the use of Lemma 3.3, the expressions G_y, F_z, K_u, $\{f_y\}$, $\{k_y\}$, and $\{k_z\}$ can be expanded, e.g.,

$$F_z = \sum_{k=0}^{\gamma} h^k A C^k A^{-1} \otimes B_k + \mathcal{O}(h^{\gamma+1}) = I \otimes f_z + hAC A^{-1} \otimes \big(f_{yz}(f,\cdot) + f_{zz}(f,\cdot)\big) + \dots \tag{3.19}$$

where the B_k are functions compound with derivatives of f, g and k and evaluated at $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$. In order to develop $(G_y F_z K_u)^{-1}$ we first consider

$$G_y F_z K_u = I \otimes (g_y f_z k_u) + \sum_{0<i+j+k\le\omega} h^{i+j+k} (C^i A C^j A^{-1} A^2 C^k A^{-2}) \otimes D_{ijk} + \mathcal{O}(h^{\omega+1}) \tag{3.20}$$

where $\omega = \mu$ ($\omega = \gamma$ if k is linear in u because $k_u(y,z,u)$ is independent of u) and the D_{ijk} are other expressions like the B_k evaluated at $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$. By using once again the series of Von Neumann we obtain

$$(G_y F_z K_u)^{-1} = I \otimes (g_y f_z k_u)^{-1} \tag{3.21}$$

$$+ \sum_{0<|\alpha|+|\beta|+|\delta|\le\omega} h^{|\alpha|+|\beta|+|\delta|} \left(\prod_{i=1}^{\omega} C^{\alpha_i} A C^{\beta_i} A^{-1} A^2 C^{\delta_i} A^{-2} \right) \otimes E_{\alpha\beta\delta} + \mathcal{O}(h^{\omega+1})$$

where the $E_{\alpha\beta\delta}$ are other expressions like the B_k evaluated at $(\widetilde{\eta}, \widetilde{\zeta}, \widetilde{\nu})$, $\alpha = (\alpha_1, \dots, \alpha_\omega)$, $\beta = (\beta_1, \dots, \beta_\omega)$, and $\delta = (\delta_1, \dots, \delta_\omega)$ are multi-indices in $\mathbb{N}^\omega$. The norm of a multi-index $\kappa = (\kappa_1, \dots, \kappa_\omega) \in \mathbb{N}^\omega$ is defined by $|\kappa| := \sum_{i=1}^{\omega} \kappa_i$. Hence, we are now able to develop the sum containing $Q_z\Delta\zeta$ in (3.18).

In order to show (3.11a), we carefully examine only two representative terms, as the others can be treated similarly. We first consider

$$H := -(e_s^T \otimes I) F_z P_{z,A}(K_u - K_{u,0})(A\mathbb{1} \otimes h(g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta) \tag{3.22}$$

which is simply a constituent of W and it appears in ΔY_s when $\varrho = 0$ in the sum of (3.18). By expanding H into h-powers we arrive at

$$H = h \sum_{\substack{1 \le |\sigma| \le m \\ \sigma := (\alpha,\beta,\delta,\nu)}} h^{|\sigma|} C_\sigma \cdot K_\sigma Q_z \Delta\zeta + \mathcal{O}(h^{m+2} \|Q_z \Delta\zeta\|) \tag{3.23}$$

where $\alpha, \beta, \delta, \nu$ are multi-indices, the K_σ are of the same type as the D_{ijk}, and the coefficients C_σ are given by

$$C_\sigma = e_s^T A C^{\nu_1} A^{-1} A^2 C^{\nu_2} A^{-2} \left(\prod_{i=1}^{\omega} C^{\alpha_i} A C^{\beta_i} A^{-1} A^2 C^{\delta_i} A^{-2} \right) \times \tag{3.24}$$
$$C^{\nu_3} A C^{\nu_4} A^{-1} \underbrace{A^2 C^{\nu_5} A^{-2} A\mathbb{1}}_{A \cdot A C^{\nu_5} A^{-1} \mathbb{1}}$$

with $\nu_5 > 0$. If $D(r)$ is satisfied with $r \ge 1$ then in accordance with Theorem 4.1 below, these coefficients vanish for $|\sigma| + 1 = |\alpha| + |\beta| + |\delta| + |\nu| + 1 \le r$. For $r = 0$, we have $H = \mathcal{O}(h^2 \|Q_z \Delta\zeta\|)$ as a consequence of $K_u - K_{u,0} = \mathcal{O}(h)$. We thus get $H = \mathcal{O}(h^{m+2} \|Q_z \Delta\zeta\|)$.

Assuming that $m \ge 3$, we now consider a second expression involving $Q_z \Delta\zeta$ which enters in the computation of ΔY_s, coming from $h^2 S^2 W$ in (3.18),

$$J := -h^2 (e_s^T \otimes I) F_z P_{z,A} (A \otimes I)^2 \{k_z\} (A \otimes I)^{-1} P_{z,A} (A \otimes I)^2 \{k_z\} \times \tag{3.25}$$
$$(A \otimes I)^{-1} P_{z,A} (K_u - K_{u,0}) (A\mathbb{1} \otimes h (g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta) \,.$$

As seen above we get

$$J = h^3 \sum_{\substack{1 \le |\sigma| \le m-2 \\ \sigma := (\alpha,\beta,\delta,\varepsilon,\theta,\kappa,\lambda,\varsigma,\upsilon,\nu)}} h^{|\sigma|} D_\sigma \cdot L_\sigma Q_z \Delta\zeta + \mathcal{O}(h^{m+2} \|Q_z \Delta\zeta\|) \tag{3.26}$$

where the L_σ are other expressions like the D_{ijk} and evaluated at $(\tilde\eta, \tilde\zeta, \tilde\nu)$. The coefficients D_σ are given by

$$D_\sigma = e_s^T A C^{\nu_1} A^{-1} \times \tag{3.27}$$
$$A^2 C^{\nu_2} A^{-2} \left(\prod_{i=1}^{\omega} C^{\alpha_i} A C^{\beta_i} A^{-1} A^2 C^{\delta_i} A^{-2} \right) C^{\nu_3} A C^{\nu_4} A^{-1} \underbrace{A^2 C^{\nu_5} A^{-1}}_{A \cdot A C^{\nu_5} A^{-1}} \times$$
$$A^2 C^{\nu_6} A^{-2} \left(\prod_{i=1}^{\omega} C^{\varepsilon_i} A C^{\theta_i} A^{-1} A^2 C^{\kappa_i} A^{-2} \right) C^{\nu_7} A C^{\nu_8} A^{-1} \underbrace{A^2 C^{\nu_9} A^{-1}}_{A \cdot A C^{\nu_9} A^{-1}} \times$$
$$A^2 C^{\nu_{10}} A^{-2} \left(\prod_{i=1}^{\omega} C^{\lambda_i} A C^{\varsigma_i} A^{-1} A^2 C^{\upsilon_i} A^{-2} \right) C^{\nu_{11}} A C^{\nu_{12}} A^{-1} \underbrace{A^2 C^{\nu_{13}} A^{-2} A\mathbb{1}}_{A \cdot A C^{\nu_{13}} A^{-1} \mathbb{1}}$$

with $\nu_{13} > 0$, and according to Theorem 4.1 they vanish if $|\sigma| + 3 \le r$. Therefore, J can be estimated by $\mathcal{O}(h^{m+2} \|Q_z \Delta\zeta\|)$ too. All other remaining terms can be treated in a similar way, so that (3.11a) results.

The h-exponent in front of $Q_z\Delta\zeta$ in (3.11b) remains to be proved. We use the same techniques as above to estimate ΔY_s. The main difference is that we expand some specific terms in the matrix products involving S and T (3.18) into h-powers not at the point $(\widetilde{\eta},\widetilde{\zeta},\widetilde{\nu})$ but at $(Y_s,Z_s,G(Y_s,Z_s))$. Such an expansion concerns only the first factors of these matrix products which are equal to K_u. From Lemma 3.3 we easily get

$$Y_i = Y_s - \sum_{m=1}^{\lambda} \frac{(1-c_i^m)h^m}{m!} D_m Y(Y_s,Z_s,G(Y_s,Z_s)) + \mathcal{O}(h^{\lambda+1}) , \tag{3.28a}$$

$$Z_i = Z_s - \sum_{n=1}^{\gamma} \frac{(1-c_i^n)h^n}{n!} D_n Z(Y_s,Z_s,G(Y_s,Z_s)) + \mathcal{O}(h^{\gamma+1}) , \tag{3.28b}$$

$$U_i = G(Y_s,Z_s) - \sum_{p=1}^{\mu} \frac{(1-c_i^p)h^p}{p!} D_p U(Y_s,Z_s,G(Y_s,Z_s)) + \mathcal{O}(h^{\mu+1}) . \tag{3.28c}$$

We rewrite an expression entering in the vector W (see (3.17b)) in the form

$$P_{z,A}(K_u - K_{u,0})(A\otimes I) = P_{z,A}(A\otimes I)(A\otimes I)^{-1}(K_u - K_{u,0})(A\otimes I) . \tag{3.29}$$

The above expression $(A\otimes I)^{-1}(K_u - K_{u,0})(A\otimes I)$ can be expanded leading to

$$(A\otimes I)^{-1}(K_u - K_{u,0})(A\otimes I) = \sum_{k=1}^{\omega} h^k A C^k A^{-1} \otimes K_{k,s} \tag{3.30}$$

where the $K_{k,s}$ are other expressions like the B_k. To end the proof, the arguments are similar to those used when estimating ΔY_s. The only problem could arise from coefficients of the form $e_s^T A^{-1}(I-C^k)\cdots\mathbb{1}_s$ or $e_s^T A^{-1}C^k\cdots\mathbb{1}_s$ with $k\geq 1$. But these coefficients do not appear, because of the premultiplication with $P_{z,s}$ and the fact that $P_z(Y_s,Z_s,G(Y_s,Z_s))K_{u,s}\equiv 0$ with $K_{u,s} := K_u(Y_s,Z_s,G(Y_s,Z_s))$. Without such a premultiplication these coefficients appear in ΔZ_s coming from, e.g., the expansion of

$$(e_s^T\otimes I)(A^{-1}\otimes I)(I\otimes K_{u,s})(G_y F_z K_u)^{-1}\cdots(\mathbb{1}\otimes h(g_y f_z k_u)^{-1} g_y f_z Q_z \Delta\zeta) . \tag{3.31}$$

□

4. Properties of the RK coefficients.

The following theorem deals with the valuation of expressions encountered in the demonstration of Theorem 3.4.

Theorem 4.1. *Let us assume that the conditions $B(1)$, $D(r)$, (I), and (S) hold. For a fixed $\varrho\in\mathbb{N}\backslash\{0\}$, let us consider a multi-index $\nu=(\nu_1,\ldots,\nu_\varrho)$ satisfying $\nu_i\geq 1$ and let $\alpha\geq 0$. If $|\nu|\leq r$ then we have*

$$e_s^T C^\alpha \left(\prod_{i=1}^{\varrho} M_i\right)\mathbb{1} = e_s^T C^\alpha M_1\cdots M_\varrho\mathbb{1} = 0 \tag{4.1}$$

where the matrices M_i are of the form

$$A^{\nu_i}, \qquad A^{\sigma_i}C^{\nu_i}A^{-\sigma_i}, \qquad A^{\sigma_i}(I-C^{\nu_i})A^{-\sigma_i-1} \tag{4.2}$$

with $\sigma_i \in \{0,1,2\}$ and it is supposed that $M_\varrho = AC^{\nu_\varrho}A^{-1}$.

Remark. In the proof we adopt the convention that if a null factor multiplies a term of the form b^TC^{-m} with $m \geq 1$, then this expression has to be omitted. For example (4.3) with $k=0$ reads $b^TA^{-1} = e_s^T - 0 \cdot b^TC^{-1} = e_s^T$.

Proof. In this proof k and l denote two non-negative integers. Assuming that $k \leq r$, the multiplication of (S) and $D(r)$ with A^{-1} leads to

$$b^TC^kA^{-1} = e_s^T - kb^TC^{k-1} \; . \tag{4.3}$$

(S), $D(r)$, and (4.3) together imply that

$$b^TC^kAC^lA^{-2} = e_s^T + \frac{l(l-1)}{k+1}b^TC^{l-2} - \frac{(k+l+1)(k+l)}{k+1}b^TC^{k+l-1} \tag{4.4}$$

provided $k+l \leq r-1$. Similarly, if $k+l \leq r$ we get

$$b^TC^k(I-C^l)A^{-2} = le_s^T + k(k-1)b^TC^{k-2} - (k+l)(k+l-1)b^TC^{k+l-2} \tag{4.5}$$

and for $k+l \leq r-1$ we have

$$\begin{aligned} b^TC^kA(I-C^l)A^{-3} =& 2le_s^T + \frac{l(l-1)(l-2)}{k+1}b^TC^{l-3} \\ &+ k(k-1)b^TC^{k-2} - \frac{(k+l+1)(k+l)(k+l-1)}{k+1}b^TC^{k+l-2} \; . \end{aligned} \tag{4.6}$$

A repeated application of (S), $D(r)$, (4.3)-(4.6) to (4.1) shows that this expression is a linear combination of terms $b^TC^kA^{-1}\mathbb{1}$ with $1 \leq k \leq r$. They all vanish because of

$$b^TC^kA^{-1}\mathbb{1} = e_s^T\mathbb{1} - kb^TC^{k-1}\mathbb{1} = 1 - k\frac{1}{k} = 0 \tag{4.7}$$

which is a consequence of (4.3) and $B(r)$ (Remark III.1.2.1 applies). □

5. Local error.

We consider one step of a RK method (2.1) with consistent initial values (y_0, z_0, u_0) at x_0 and we want to give estimates for the local error.

Theorem 5.1.

a) Let us suppose that the RK method satisfies (I), B(p), and C(q) with $q \geq 2$. Then we have

$$\delta y_h(x_0) = \mathcal{O}(h^{\min(p,q)+1}), \qquad P_y(x_0+h)\delta y_h(x_0) = \mathcal{O}(h^{\min(p,q+1)+1}), \tag{5.1a}$$

$$\delta z_h(x_0) = \mathcal{O}(h^q), \qquad P_z(x_0+h)\delta z_h(x_0) = \mathcal{O}(h^{\min(p,q)+1}), \tag{5.1b}$$

$$\delta u_h(x_0) = \mathcal{O}(h^{\min(p,q-2)+1}) \tag{5.1c}$$

where $P_y(x)$, $P_z(x)$ are the projectors (3.7) evaluated at $(y(x), z(x), u(x))$, the exact solution of (1.1) at x.

b) Moreover, if in addition D(r) and (S) hold, then we obtain

$$\delta y_h(x_0) = \mathcal{O}(h^{k+1}), \quad \text{with } k = \min(p, 2q-1, q+r), \tag{5.2}$$

$$P_z(x_0+h)\delta z_h(x_0) = \mathcal{O}(h^{\ell+1}), \quad \text{with } \ell = \min(p, 2q-2, q+r). \tag{5.3}$$

Remark. If the function k of (1.1) is linear in u, then, in (5.2) and (5.3) we have

$$k = \min(p, 2q, q+r), \qquad \ell = \min(p, 2q-1, q+r), \tag{5.2$'$)–(5.3$'$)}$$

and the condition $q \geq 2$ can be omitted.

Proof. By a simple application of Theorem III.4.3 part b) with $q = \widehat{q} = Q$ and $r = \widehat{r} = R$. □

6. Convergence results.

We present here the main result of this chapter and we partly follow [HaLuRo89a, pp. 78-82]. Theorem 6.1 proves the conjecture, based on numerical experiments, stated in [HaLuRo89a, p. 86].

Theorem 6.1. *Let us consider the differential-algebraic system (1.1a, b, c) of index 3 with consistent initial values (y_0, z_0, u_0) at x_0 and the RK method (2.1). Let us assume that the RK coefficients satisfy B(p), C(q) with $q \geq 2$, D(r), (I), and (S). Then for $x_n - x_0 = nh \leq$ Const, the global error satisfies*

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^{\min(p,2q-2,q+r)}), \quad z_n - z(x_n) = \mathcal{O}(h^q), \\ P_z(x_n)(z_n - z(x_n)) &= \mathcal{O}(h^{\min(p,2q-2,q+r)}), \quad u_n - u(x_n) = \mathcal{O}(h^{q-1}). \end{aligned} \tag{6.1}$$

Remarks 6.1.

1) If in addition the function k of (1.1) is linear in u, we get

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^{\min(p,2q-1,q+r)}) \,, \\ P_z(x_n)\big(z_n - z(x_n)\big) &= \mathcal{O}(h^{\min(p,2q-1,q+r)}) \,, \end{aligned} \tag{6.1'}$$

and the assumption $C(2)$ can be omitted. In the proof in this case we have $m = \min(q-1, \max(r-1,0))$, $n = \min(q-1, r)$, k and ℓ are given by (5.2')-(5.3'), and the terms $\|(Q_z)_n \Delta z_n\|^2$ in (6.2*b*, *c*, *d*) have to be replaced by $h\|(Q_z)_n \Delta z_n\|^2$. In the situation of $q=1$, the proof given in [HaLuRo89a, Theorem 6.4] requires slight modifications.

2) The theorem remains valid in the case of variable stepsizes with $h = \max_i h_i$.

Proof. In a first step, we can show that global convergence of order q for the (y,z)-components occurs (see [HaLuRo89a, Theorem 6.4]).

In the second step, we use once again the techniques of the previous step. We denote two neighbouring RK solutions by $\{\widetilde{y}_n, \widetilde{z}_n\}$, $\{\widehat{y}_n, \widehat{z}_n\}$ and their difference by $\Delta y_n = \widetilde{y}_n - \widehat{y}_n$, $\Delta z_n = \widetilde{z}_n - \widehat{z}_n$. We assume that $\Delta y_n = \mathcal{O}(h^{q+1})$ and $\Delta z_n = \mathcal{O}(h^{q+1})$ (see [HaLuRo89a, Formula (6.28)]). Because of $g(\widetilde{y}_n) = 0 = g(\widehat{y}_n)$ Remark III.2.2.4 holds, implying that

$$(Q_y)_n \Delta y_n = \mathcal{O}(\|\Delta y_n\|^2) = \mathcal{O}(h^{q+1}\|(P_y)_n \Delta y_n\|) \,. \tag{6.2a}$$

By the use of the results of the first step, Theorem 3.4 can be applied with $\delta = 0$, $\mu = 0$, and $\theta = 0$, yielding

$$\begin{aligned} (P_y)_{n+1}\Delta y_{n+1} &= (P_y)_n \Delta y_n + h(f_z)_n (P_z)_n \Delta z_n \qquad (6.2b) \\ &\quad + \mathcal{O}\left(h\|(P_y)_n \Delta y_n\| + h^2\|(P_z)_n \Delta z_n\| + h^{m+2}\|(Q_z)_n \Delta z_n\| + \|(Q_z)_n \Delta z_n\|^2\right) , \\ h(P_z)_{n+1}\Delta z_{n+1} &= h(P_z)_n \Delta z_n \qquad (6.2c) \\ &\quad + \mathcal{O}\left(h^2\|(P_y)_n \Delta y_n\| + h^2\|(P_z)_n \Delta z_n\| + h^{n+2}\|(Q_z)_n \Delta z_n\| + \|(Q_z)_n \Delta z_n\|^2\right) , \\ h(Q_z)_{n+1}\Delta z_{n+1} &= \mathcal{O}\left(h\|(P_y)_n \Delta y_n\| + h^2\|\Delta z_n\| + \|(Q_z)_n \Delta z_n\|^2\right) \qquad (6.2d) \end{aligned}$$

where $m = \min(q-2, \max(r-1,0))$, $n = \min(q-2, r)$ and $(P_y)_n$, $(Q_y)_n$, $(P_z)_n$, $(Q_z)_n$, $(f_z)_n$ are evaluated at $(\widehat{y}_n, \widehat{z}_n, \widehat{u}_n^*)$. We define $\widehat{u}_n^* := G(\widehat{y}_n, \widehat{z}_n)$ with G as described in (II.1.4). This choice of $\widehat{u}_n^*$ does not influence the values $(\widehat{y}_n, \widehat{z}_n)$ (see Remark 3.1.2) and simplifies the proof. The estimates (6.2) lead to (by induction or similarly to the proof of [Os93, Theorem 3.3])

$$\|\Delta y_n\| \le C\left(\|(P_y)_0 \Delta y_0\| + \|(P_z)_0 \Delta z_0\| + h^{n+1}\|(Q_z)_0 \Delta z_0\|\right) \,, \tag{6.3a}$$

$$h\|(P_z)_n \Delta z_n\| \le C\left(h\|(P_y)_0 \Delta y_0\| + h\|(P_z)_0 \Delta z_0\| + h^{n+2}\|(Q_z)_0 \Delta z_0\|\right) \,, \tag{6.3b}$$

$$h\|(Q_z)_n \Delta z_n\| \le C\left(h\|(P_y)_0 \Delta y_0\| + h\|(P_z)_0 \Delta z_0\| + h^2\|(Q_z)_0 \Delta z_0\|\right) \,. \tag{6.3c}$$

Hence it follows from standard techniques (see [HaLuRo89a, Fig. 4.1, p. 36] or [HaNø-Wa93, Fig. II.3.2, p. 160]) that

$$y_n - y(x_n) = \mathcal{O}(h^{\min(k,\ell,q+n)}) \,, \quad P_z(x_n)\big(z_n - z(x_n)\big) = \mathcal{O}(h^{\min(k,\ell,q+n)}) \tag{6.4}$$

where k and ℓ are given in (5.2)-(5.3). □

The estimates (6.1)-(6.1') gives us more insight into the structure of the global error for the z-component. If the numerical solution is projected onto the manifold

$(g_y f)(y,z) = 0$, then the accuracy of the z-component can be improved. This can be done similarly as described in Subsection III.1.3, but here theoretically ~~only~~ merely at the end of the integration process. We thus obtain:

Corollary 6.2. *Under the assumptions of Theorem 6.1, let $\widehat{z}_n, \mu_n$ be the solution of*

$$\widehat{z}_n = z_n + k_u(y_n, z_n, u_n)\mu_n, \qquad 0 = (g_y f)(y_n, \widehat{z}_n) \tag{6.5}$$

where (y_n, z_n, u_n) is the numerical solution (2.1) after n steps. Then we get

$$\widehat{z}_n - z(x_n) = \begin{cases} \mathcal{O}(h^{\min(p,2q-1,q+r)}) & \text{if } k \text{ is linear in } u\ , \\ \mathcal{O}(h^{\min(p,2q-2,q+r)}) & \text{else}\ . \end{cases} \tag{6.6}$$

Moreover, if we define $\widehat{u}_n$ as the solution together with y_n and $\widehat{z}_n$ of (1.1e), we have

$$\widehat{u}_n - u(x_n) = \begin{cases} \mathcal{O}(h^{\min(p,2q-1,q+r)}) & \text{if } k \text{ is linear in } u\ , \\ \mathcal{O}(h^{\min(p,2q-2,q+r)}) & \text{else}\ . \end{cases} \tag{6.7}$$

□

The problem (1.1) is not ill-posed if all constraints (1.1c, d, e) are taken into account and not only (1.1~~d~~) since it is of index 1 (see Section II.1). It is in fact preferable to effect the projection (6.5) after every step, because the numerical solution is stabilized as concerns the influence of perturbations (see formulas (3.6c) and (3.11b)). ~~However~~, Nevertheless a very accurate approximation of the u-component of the solution may be unnecessary (see Remark 3.1.2), therefore the projection onto the manifold (1.1e) can be omitted. This remark is important if one wants to avoid the computation of extra derivatives such as g_{yy}. For stiffly accurate RK methods a fairly good choice is often given by $u_1 := U_s$.

Corollary 6.3. *For the s-stage ($s \geq 3$) Lobatto IIIC method applied to the index 3 system (1.1a, b, c), the global error satisfies*

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^{2s-4})\ , & P_z(x_n)\big(z_n - z(x_n)\big) &= \mathcal{O}(h^{2s-4})\ , \\ z_n - z(x_n) &= \mathcal{O}(h^{s-1})\ , & \widehat{z}_n - z(x_n) &= \mathcal{O}(h^{2s-4})\ , \\ u_n - u(x_n) &= \mathcal{O}(h^{s-2})\ , & \widehat{u}_n - u(x_n) &= \mathcal{O}(h^{2s-4})\ . \end{aligned} \tag{6.8}$$

Moreover, if k is linear in u we have ($s \geq 2$)

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^{2s-3}), & P_z(x_n)\big(z_n - z(x_n)\big) &= \mathcal{O}(h^{2s-3}), \\ \widehat{z}_n - z(x_n) &= \mathcal{O}(h^{2s-3}), & \widehat{u}_n - u(x_n) &= \mathcal{O}(h^{2s-3})\ . \end{aligned} \tag{6.8′}$$

Proof. The proof is obtained by putting $p = 2s-2$, $q = s-1$, $r = s-1$ in (6.1)-(6.1′), (6.6), and (6.7). □

The next result provides an alternative proof of [Jay93b, Corollary 2.3] demonstrated with completely different techniques.

Corollary 6.4. *For the s-stage ($s \geq 2$) Radau IIA method applied to the index 3 system* (1.1a,b,c), *the global error satisfies*

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^{2s-2}) , & P_z(x_n)\big(z_n - z(x_n)\big) &= \mathcal{O}(h^{2s-2}) , \\ z_n - z(x_n) &= \mathcal{O}(h^{s}) , & \widehat{z}_n - z(x_n) &= \mathcal{O}(h^{2s-2}) , \\ u_n - u(x_n) &= \mathcal{O}(h^{s-1}) , & \widehat{u}_n - u(x_n) &= \mathcal{O}(h^{2s-2}) . \end{aligned} \tag{6.9}$$

Moreover, if k is linear in u we have ($s \geq 1$)

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^{2s-1}) , & P_z(x_n)\big(z_n - z(x_n)\big) &= \mathcal{O}(h^{2s-1}) , \\ \widehat{z}_n - z(x_n) &= \mathcal{O}(h^{2s-1}) , & \widehat{u}_n - u(x_n) &= \mathcal{O}(h^{2s+1}) . \end{aligned} \tag{6.9'}$$

Proof. The proof is obtained by putting $p = 2s-1$, $q = s$, $r = s-1$ in (6.1)-(6.1'), (6.6), and (6.7). □

Remark 6.2. For a constant-stepsize application of the implicit Euler method ($s = 1$) with k linear in u, it can be shown that $u_n - u(x_n) = \mathcal{O}(h)$ for $n \geq 2$ (see [BreEn88] and [HaLuRo89a, p. 90]).

An application of Theorem 6.2 concerns the convergence analysis of Runge-Kutta methods when applied to stiff mechanical systems (see Subsection I.4.2). The following theorem is due to Lubich:

Theorem 6.5. [Lu93, Theorem 3.1]. *Let us consider a stiff mechanical system* (I.4.20) *satisfying the assumptions (SMS) (see Subsection I.4.2) and a RK method satisfying the hypotheses (I), C(q), $|R(\infty)| < 1$, and*

$$A \text{ has no eigenvalues on the imaginary axis and } |R(i\omega)| < 1 \ \forall \omega \in \mathbb{R} \setminus \{0\} . \tag{6.10}$$

Let us suppose that the exact solution $(q^\varepsilon(t), v^\varepsilon(t))$ of (I.4.20) *with initial values $(q_0^\varepsilon, v_0^\varepsilon)$ is smooth. Then for $0 < \varepsilon \leq h \leq h_0$, with h_0 sufficiently small but independent of ε, there exists a unique RK solution of the stiff system* (I.4.20), *whose error satisfies*

$$q_n^\varepsilon - q^\varepsilon(t_n) = q_n^0 - q^0(t_n) + \mathcal{O}(\varepsilon^2 h^{q-2}) , \quad v_n^\varepsilon - v^\varepsilon(t_n) = v_n^0 - v^0(t_n) + \mathcal{O}(\varepsilon^2 h^{q-2}) \tag{6.11}$$

uniformly for $t_0 \leq t_n \leq T$. Here (q_n^0, v_n^0) and $(q^0(t), v^0(t))$ denote the RK and exact solution respectively of an index 3 DAE (I.4.3a,b,c), *where the initial values (q_0^0, v_0^0) are the coefficients of ε^0 in the ε^2-expansion of $(q_0^\varepsilon, v_0^\varepsilon)$. Therefore, for RK methods satisfying in addition the hypotheses of Theorem 6.1 (k is linear in u) we have*

$$q_n^\varepsilon - q^\varepsilon(t_n) = \mathcal{O}(h^{\min(p, 2q-1, q+r)} + \varepsilon^2 h^{q-2}) , \quad v_n^\varepsilon - v^\varepsilon(t_n) = \mathcal{O}(h^q + \varepsilon^2 h^{q-2}) . \tag{6.12}$$

□

7. Numerical experiments.

The global convergence results of Section 6 have been confirmed by numerical tests. As a first example, consider the following index 3 problem

$$\begin{aligned} y_1' &= 2y_1y_2z_1z_2\ , & y_2' &= -y_1y_2z_2^2\ , \\ z_1' &= (y_1y_2+z_1z_2)u\ , & z_2' &= -y_1y_2^2z_2^2u\ , \\ 0 &= y_1y_2^2-1\ , \end{aligned} \tag{7.1}$$

which is of the form (1.1a,b,c) with k linear in u. For the consistent initial values $y_0=(1,1)^T$, $z_0=(1,1)^T$, and $u_0=1$ the exact solution is given by

$$y_1(x)=z_1(x)=e^{2x}, \qquad y_2(x)=z_2(x)=e^{-x}, \qquad u(x)=e^x\ . \tag{7.2}$$

In Fig. 7.1 the global errors at $x_{\text{end}}=0.1$ for the Lobatto IIIC methods ($s=2,3,4,5,6$) applied to (7.1) are plotted as functions of h (the stepsizes have been chosen alternatively as $h/3$ and $2h/3$). Since we have used logarithmic scales, the curves appear as straight lines of slope k whenever the leading term of the error is $\mathcal{O}(h^k)$. This behaviour is indicated in the figures.

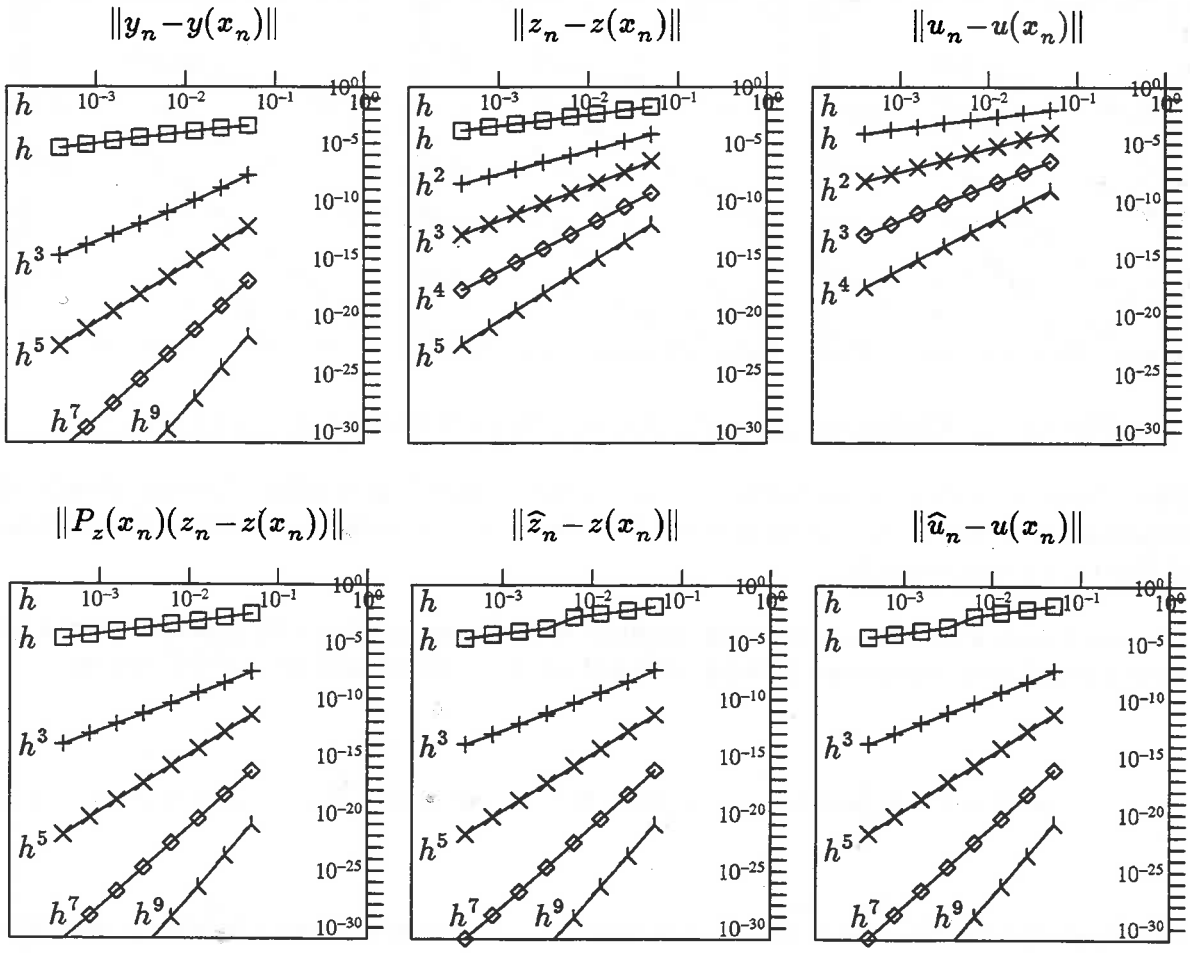


Figure 7.1. Global errors of the Lobatto IIIC methods ($s=2$:□;3:+; 4:×; 5:◇; 6:⋀).

The second example is a slight modification of problem (7.1) where the equation for z_2' has been replaced by $z_2'=-y_1y_2^2z_2^3u^2$, so that k becomes nonlinear in u. With the

same consistent initial values defined previously, the exact solution is identical to (7.2). The Radau IIA methods ($s=2,3,4,5$) have been applied to this modified problem and in Fig. 7.2 the global errors at $x_{\text{end}}=0.1$ are plotted.

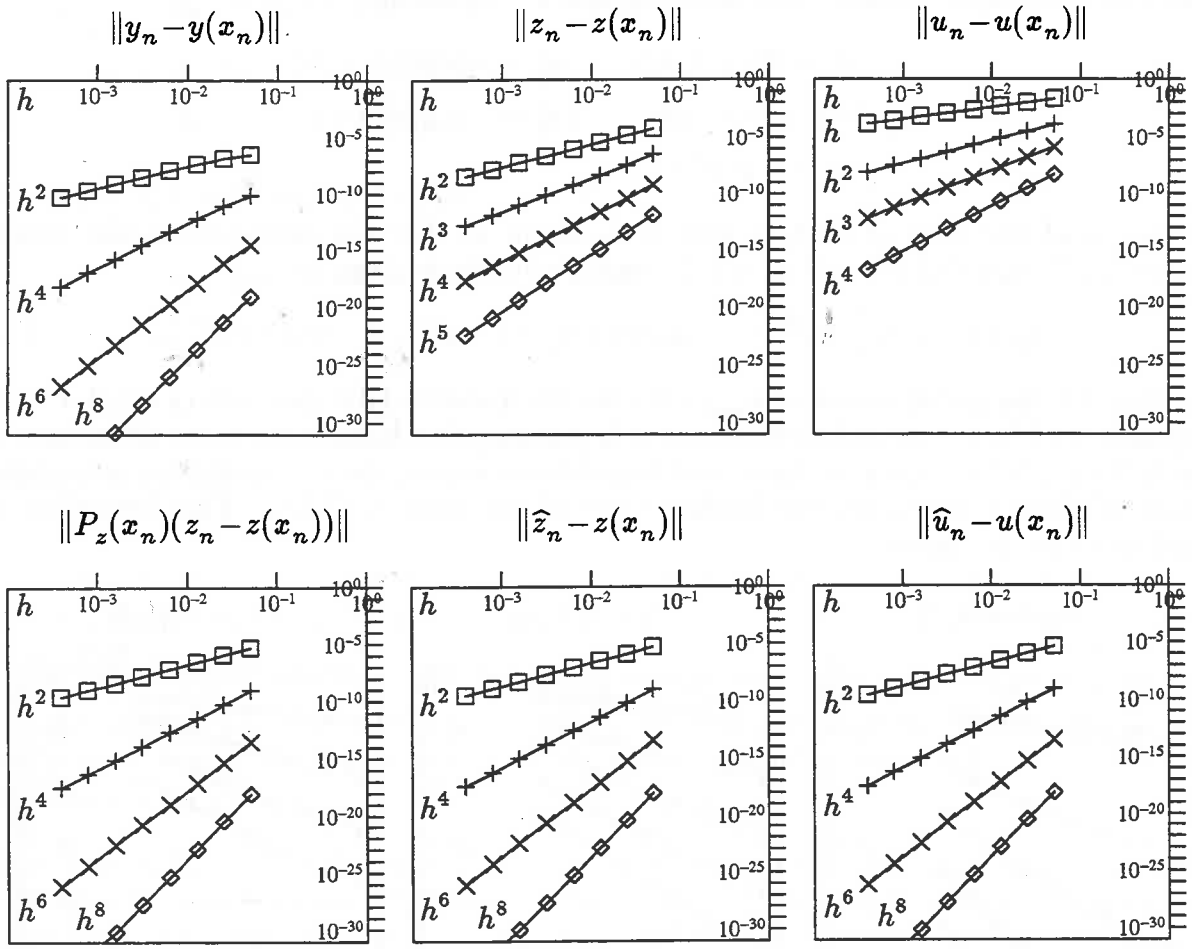


Figure 7.2. Global errors of the Radau IIA methods ($s=2$:□;3:+; 4:×; 5:◇).

The observed orders of convergence match the theoretical results, showing clearly the importance of the projections described in Section 6 in order to improve the accuracy of the (z,u)-components.

As a last experiment we have applied the 3-stage Radau IIA method to the stiff spring pendulum equations (I.4.23)-(I.4.24) with $\varepsilon=0.001$ and the initial values

$$x(0)=1-3\varepsilon^4+\mathcal{O}(\varepsilon^8)\,,\quad z(0)=0\,,\quad v_x(0)=\mathcal{O}(\varepsilon^8)\,,\quad v_z(0)=0 \tag{7.3}$$

chosen such that the exact solution is smooth, i.e., does not include highly oscillatory terms (see [HaLuRo89a, p. 119] and [HaJay93]). This allows the use of stepsizes which are significantly larger than ε. Fig. 7.3 shows the global errors as a function of h. We see that the errors behave like $\mathcal{O}(h^2),\mathcal{O}(h^3),\mathcal{O}(h^5)$ for the components λ, v_z, and z respectively. The errors for x,v_x behave similarly to those for z,v_z and are not plotted. This experiment confirms the results predicted by Theorem 6.5.

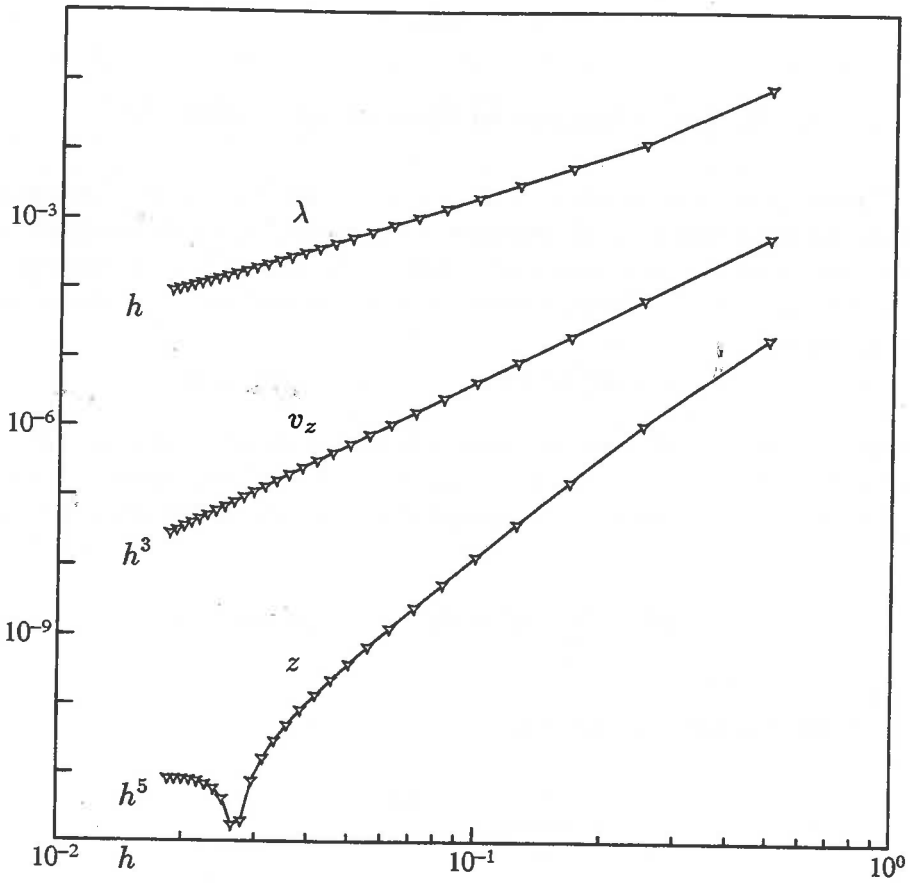


Figure 7.3. Global errors of the 3-stage Radau IIA method applied to (I.4.23)-(I.4.24).

Chapter V. Symplectic partitioned Runge-Kutta methods for constrained Hamiltonian systems.

1. Introduction to Hamiltonian systems.

Hamiltonian problems arise in a lot of applications where dissipative forces can be neglected, such as mechanical systems, astronomy, electrodynamics, molecular dynamics, plasma physics, fluid dynamics, etc. The *Hamiltonian system* of differential equations associated to the *Hamiltonian* $H(q,p)$ (a real function supposed sufficiently smooth) is given by

$$\dot{q} = H_p^T(q,p) \ , \qquad \dot{p} = -H_q^T(q,p) \tag{1.1}$$

where $q=(q^1,\dots,q^n)^T \in \mathbb{R}^n$ are the *generalized coordinates* and $p=(p^1,\dots,p^n)^T \in \mathbb{R}^n$ the *generalized momenta*. The flow generated in the phase space $\mathbb{R}^n \times \mathbb{R}^n$ of (q,p) by these equations (1.1) is known to be *symplectic*, i.e., the differential 2-form

$$\omega^2 = \sum_{k=1}^{n} dq^k \wedge dp^k \qquad \text{is preserved,} \tag{1.2}$$

implying that all differential $2d$-forms

$$\underbrace{\omega^2 \wedge \dots \wedge \omega^2}_{d \text{ times}} \qquad \text{for} \quad d = 1,\dots,n \tag{1.3}$$

are also conserved ($d=n$ corresponds to the $2n$-form *volume*). Another specific feature of such systems is that the Hamiltonian along a solution $(q(t),p(t))$ to (1.1) passing through (q_0,p_0) at t_0 remains invariant, i.e.,

$$H(q(t),p(t)) = H(q_0,p_0) \qquad \text{for all } t. \tag{1.4}$$

Hamiltonian systems also possess numerous other specific properties (see [ArV.89, Part III] and [MK92]). Unfortunately, most of numerical methods applied to (1.1) do not maintain the above two properties (1.2) and (1.4). Various authors ([SS88], [La88], [Sur89], [Yo90], [Sun92], [Sun93], [AbSS93] among others) have identified or constructed *symplectic schemes*, i.e., methods maintaining (1.2). For an overview on symplectic integrators we refer to [SS92] and [HaNøWa93, Section II.16].

In this chapter we consider Hamiltonian systems with holonomic constraints (see Subsection I.4.1). Such problems form a particular class of semi-explicit index 3 DAE's in Hessenberg form. We present a very efficient class of PRK methods for the solution of these problems. It consists of the couples of s-stage Lobatto IIIA and Lobatto IIIB methods. These methods combine three attractive properties:

- symplecticity, such as seen with the RK methods Gauss and Lobatto IIIS (see [Sun92], [Sun93], and [Cha90]);
- the fact that the numerical solution can be naturally projected onto the manifolds where the exact solution lies (see [AsPe91], [AsPe92b], [HaWa91, Section VI.7], [LeSk94], and Subsection III.1.3 for similar ideas), without loss of symplecticity;

- superconvergence, a property shared by stiffly accurate (projected) (P)RK methods (see Chapters III and IV).

The importance of symplecticity in numerical integration, especially for long-time computations, is nowadays underlined by a sort of "backward analysis" by interpreting the numerical solution as the *exact* solution of a nearby *perturbed* Hamiltonian system (see [SS92]). In [Ha94] Hairer proved recently the general result that *all* symplectic methods whose numerical solution is a (partitioned) P-series possess this property. An extension of this result to Hamiltonian systems with holonomic constraints and the numerical methods treated in this chapter is likely. This fact is corroborated by the numerical observations of Section 6.

A typical example of a constrained Hamiltonian system is given by the pendulum equations. Using the cartesian coordinates $q=(x,z)^T$ for the description of the position of the pendulum, the holonomic constraint on the length ℓ of the rod of the pendulum is

$$0 = \sqrt{x^2+z^2}-\ell \ . \tag{1.5}$$

The kinetic energy T and the potential energy U of the system are given respectively by

$$T(\dot q) = \frac{m}{2}(\dot x^2+\dot z^2), \qquad U(q) = mgz \tag{1.6}$$

where g is the gravitational constant. The Lagrangian of the system is $L(q,\dot q)=T(\dot q)-U(q)$ and the generalized momenta are $p=(p_x,p_z)^T=L_{\dot q}^T(q,\dot q)$ leading to

$$p_x = m\dot x, \qquad p_z = m\dot z \ . \tag{1.7}$$

The Hamiltonian $H=T+U$ can be expressed by

$$H(q,p) = \frac{1}{2m}\left(p_x^2+p_z^2\right)+mgz \ , \tag{1.8}$$

and the Hamilton equations of motion become

$$\dot x = \frac{p_x}{m}, \qquad \dot z = \frac{p_z}{m}, \qquad \dot p_x = -\frac{x}{\ell}\lambda, \qquad \dot p_z = -mg-\frac{z}{\ell}\lambda \ . \tag{1.9}$$

One differentiation of (1.5) implies that

$$0 = xp_x+zp_z \ , \tag{1.10}$$

and another one permits to obtain

$$\lambda = \frac{1}{\ell}\left(\frac{1}{m}(p_x^2+p_z^2)-mgz\right) \ . \tag{1.11}$$

Thus the differential-algebraic system (1.9)-(1.5) is of differential index 3.

This chapter is organized in six sections. In Section 2 we give some basic definitions and results related to symplectic PRK methods. Section 3 deals with Hamiltonian systems with holonomic constraints and the application of PRK methods. Section 4 concentrates on the application of a class of PRK methods to semi-explicit index 3 DAE's in Hessenberg form. Existence and uniqueness of the PRK solution, influence of perturbations, estimates of the local error, and global convergence are studied. Convergence results for the specific family of Lobatto IIIA-IIIB schemes are then stated in Section 5. Finally, Section 6 includes some numerical experiments illustrating the theoretical results.

2. Symplectic PRK methods for Hamiltonian systems.

Hamiltonian systems (1.1) are intrinsically splitted in two parts, therefore the use of *partitioned Runge-Kutta (PRK) methods* is very natural.

Definition 2.1. One step of an *s-stage PRK method* applied to (1.1), with stepsize h and initial values (q_0, p_0) at t_0 reads

$$q_1 = q_0 + h\sum_{i=1}^{s} b_i k_i \ , \qquad p_1 = p_0 + h\sum_{i=1}^{s} \widehat{b}_i \ell_i \tag{2.1a}$$

where

$$k_i = H_p^T(Q_i, P_i) \ , \qquad \ell_i = -H_q^T(Q_i, P_i) \ , \tag{2.1b}$$

and the *internal stages* are given by

$$Q_i = q_0 + h\sum_{j=1}^{s} a_{ij} k_j \ , \qquad P_i = p_0 + h\sum_{j=1}^{s} \widehat{a}_{ij} \ell_j \ . \tag{2.1c}$$

For a PRK method the symplecticity condition (1.2) is expressed by

$$\sum_{k=1}^{n} dq_1^k \wedge dp_1^k = \sum_{k=1}^{n} dq_0^k \wedge dp_0^k \ , \tag{2.2}$$

and symplectic PRK schemes can be characterized as follows:

Theorem 2.1. *If the coefficients of an s-stage PRK method* (2.1) *satisfy*

$$b_i = \widehat{b}_i \qquad \text{for} \quad i = 1, \dots, s \ , \tag{2.3a}$$

$$b_i \widehat{a}_{ij} + \widehat{b}_j a_{ji} - b_i \widehat{b}_j = 0 \qquad \text{for} \quad i = 1, \dots, s \ , \quad j = 1, \dots, s \ , \tag{2.3b}$$

then the PRK method is symplectic.

If the PRK method is irreducible, then the conditions (2.3) *are also necessary for symplecticity.* □

Remarks 2.1.

1) For *separable Hamiltonians* $H(q,p) = T(p) + U(q)$, the first condition (2.3*a*) can be omitted (see [AbSS93]).
2) For a proof of the sufficiency of the conditions (2.3) see [SS88], [La88], [Sur89], and [Sun92].
3) For *irreducible* PRK methods, i.e., methods without equivalent stages (see [But87, Section 383], [Ha94, Section 4] or the definition of *S-irreducibility* in [HaWa91, p. 200]), a way of showing the necessity of (2.3) is to extend the characterization of canonical B-series of [CalSS93] to (partitioned) P-series (see [Ha94, Lemma 11]) and to apply the proof of [Ha94, Theorem 5]. For separable Hamiltonians an alternative proof of the necessity of (2.3*b*) is given in [AbSS93].

Definition 2.2. The *local error* of a PRK method (2.1) is defined by

$$\delta q_h(t_0) = q_1 - q(t_0+h) , \qquad \delta p_h(t_0) = p_1 - p(t_0+h) \tag{2.4}$$

where $(q(t), p(t))$ is the exact solution of (1.1) passing through (q_0, p_0) at t_0.

Considering the elegant W-transformation of Hairer and Wanner (see [HaWa91, Section IV.5]), it is possible to construct high order symplectic PRK methods starting from known RK methods as follows:

Theorem 2.2. [Sun92]. *Suppose that a RK method with coefficients a_{ij}, $b_i \neq 0$, and distinct c_i, satisfies the simplifying assumptions $B(p)$, $C(q)$, and $D(r)$. Then the PRK method* (2.1) *with coefficients $\widehat{b}_i := b_i$, $\widehat{c}_i := c_i$, and $\widehat{a}_{ij} := b_j(1 - a_{ji}/b_i)$ is symplectic and satisfies*

$$\delta q_h(t_0) = \mathcal{O}(h^{\eta+1}) , \qquad \delta p_h(t_0) = \mathcal{O}(h^{\eta+1}) \tag{2.5}$$

with an order $\eta = \min(p, 2q+2, 2r+2, q+r+1)$. □

Remarks 2.2.

1) With the help of the W-transformation it can be shown that the RK method $(\widehat{A}, b, c)$ satisfies $\widehat{C}(r)$ and $\widehat{D}(q)$ (see [HaWa91, Section IV.5] and [Sun92]).
2) The simplifying assumptions $C(1)$ and $D(1)$ (which is equivalent to $\widehat{C}(1)$ by the symplecticity conditions (2.3)) ensure here that $c_i = \sum_{j=1}^{s} a_{ij}$ and $c_i = \sum_{j=1}^{s} \widehat{a}_{ij}$ respectively. This implies some simplifications when deriving the order conditions of PRK methods applied to non-autonomous problems: in this case the order conditions reduce to those of the autonomous case (see also [HaNøWa93, p. 134]). An example of a RK method violating one of these assumptions is given by the 2-stage Lobatto IIIA method, namely the *trapezoidal rule*, which satisfies $B(2)$, $C(2)$, but not $D(1)$. Another example consists in its symplectically associated method, the 2-stage Lobatto IIIB method, which satisfies $B(2)$, $D(2)$, but not $C(1)$.
3) The symplecticity conditions (2.3), acting as simplifying assumptions, introduce a reduction of the number of order conditions (see [AbSS93]).

Examples of symplectic PRK methods are given in [Sun92]. In this chapter we focus our attention on PRK methods adapted to the situation where holonomic constraints are encountered. In this context the couples of s-stage Lobatto IIIA methods for (A, b, c) and Lobatto IIIB methods for $(\widehat{A}, b, c)$ turn out to be of main interest. These PRK methods satisfy the simplifying assumptions $B(2s-2)$, $C(s)$, $D(s-2)$, $\widehat{C}(s-2)$, and $\widehat{D}(s)$. Concerning the coefficients of these methods, the weights c_i of Lobatto quadratures are given by $c_1 = 0$, $c_s = 1$, and the remaining c_i for $i = 2, \ldots, s-1$ are the roots of the polynomial of degree $s-2$ $P_{s-2}^{(1,1)}(2x-1)$ where

$$P_{s-2}^{(1,1)}(y) = Const \cdot \frac{1}{(y^2-1)} \frac{d^{s-2}}{dy^{s-2}} \left((y^2-1)^{s-1} \right) \tag{2.6}$$

is a Jacobi polynomial. The coefficients $b_j = a_{sj}$ and a_{ij} can be computed for example by the use of $C(s)$, and the coefficients $\widehat{a}_{ij}$ as in Theorem 2.2 or with the help of $\widehat{D}(s)$. The Butcher-tableaux of the 2- and 3-stage Lobatto IIIA-IIIB methods are given below

in Tables 2.1 and 2.2 respectively. For separable Hamiltonians the 2-stage method can be applied explicitly.

0	0	0
1	1/2	1/2
	1/2	1/2

0	1/2	0
1	1/2	0
	1/2	1/2

Table 2.1. Coefficients of the 2-stage Lobatto IIIA-IIIB method of order 2.

0	0	0	0
1/2	5/24	1/3	−1/24
1	1/6	2/3	1/6
	1/6	2/3	1/6

0	1/6	−1/6	0
1/2	1/6	1/3	0
1	1/6	5/6	0
	1/6	2/3	1/6

Table 2.2. Coefficients of the 3-stage Lobatto IIIA-IIIB method of order 4.

The coefficients of the 4-stage Lobatto IIIA-IIIB method can be found in [HaWa91, p. 80]. We list also below the weights c_i of the 11-stage Lobatto method of order 20

$$
\begin{aligned}
&c_1 = 0, \quad c_2 = \frac{1}{2} - \delta_2, \quad c_3 = \frac{1}{2} - \delta_3, \quad c_4 = \frac{1}{2} - \delta_4, \quad c_5 = \frac{1}{2} - \delta_5, \quad c_6 = \frac{1}{2}, \\
&c_7 = \frac{1}{2} + \delta_5, \quad c_8 = \frac{1}{2} + \delta_4, \quad c_9 = \frac{1}{2} + \delta_3, \quad c_{10} = \frac{1}{2} + \delta_2, \quad c_{11} = 1, \\
&\delta_2 = \frac{1}{2}\sqrt{\frac{1}{2}(-u + a_1) + a_3}, \qquad \delta_3 = \frac{1}{2}\sqrt{\frac{1}{2}(-u - a_1) + a_3}, \\
&\delta_4 = \frac{1}{2}\sqrt{\frac{1}{2}(u + a_2) + a_3}, \qquad \delta_5 = \frac{1}{2}\sqrt{\frac{1}{2}(u - a_2) + a_3}, \\
&a_1 = \sqrt{v - 4\alpha}, \quad a_2 = \sqrt{v - 4\beta}, \quad a_3 = \frac{9}{19}, \quad \alpha = d_1 - d_2, \quad \beta = d_1 + d_2, \\
&d_1 = \frac{1}{2}(e_1 + v), \quad d_2 = \frac{1}{2}\sqrt{(e_1 + v)^2 - 4e_2}, \quad e_1 = -\frac{1080}{6137}, \quad e_2 = \frac{96912}{28800941}, \\
&u = \sqrt{v}, \quad v = \frac{32}{323c} \cdot \cos\left(\frac{1}{3}\arccos\left(\frac{3}{2c}\right)\right) + \frac{720}{6137}, \quad c = \sqrt{\frac{13}{42}}.
\end{aligned}
\tag{2.7}
$$

Due to their symmetry, Lobatto schemes are often used for the solution of boundary value problems (see [As85] and [AsMaRu88]). The analysis of the application of Lobatto IIIA methods to semi-explicit index 2 DAE's in Hessenberg form is given in [Jay93a].

3. Hamiltonian systems with holonomic constraints and PRK methods.

Mechanical systems where dissipative forces can be neglected are Hamiltonian systems (see Subsection I.4.1). Their Hamiltonian is of the form $H=T+U$ where T represents the *kinetic energy* and U the *potential energy*. Usually the equations of motion are not written with an Hamiltonian formalism, but in an Euler-Lagrange formulation (see also [HaWa91, Sections VI.5 & VI.9]). Our aim here is to study Hamiltonian systems with $m<n$ *holonomic* constraints $g^i(q)=0$ $(i=1,\dots,m)$ (see Subsection I.4.1, [Cho92, Chapter 2], and [ArV.89, Section 17]). A Lagrange-type variational principle exists if the constraints are holonomic (see [Cho92, Subsection 4.2.2]). Applying this principle to $\widehat{H}(q,p)=H(q,p)+\lambda^T g(q)$, we arrive at

$$\dot q = H_p^T(q,p)\ , \qquad \dot p = -H_q^T(q,p) - G^T(q)\lambda\ , \qquad 0 = g(q) \qquad (3.1a,b,c)$$

where $G(q) := g_q(q)$. The variables λ^i $(i=1,\dots,m)$ are the *Lagrange multipliers*. Differentiating $(3.1c)$ twice similarly to the example given in Section 1, we obtain the following additional constraints (omitting the obvious function arguments)

$$0 = GH_p^T\ , \qquad 0 = G_q(H_p^T, H_p^T) + GH_{pq}^T H_p^T - GH_{pp}^T H_q^T - GH_{pp}^T G^T\lambda\ . \qquad (3.1d,e)$$

Initial values for the problem $(3.1a,b,c)$ have to be *consistent*, i.e., they must satisfy $(3.1c,d,e)$. From now on we suppose that G is of full row rank m and that we have an *optical Hamiltonian system* (see [MK92, p. 140]), meaning that H_{pp} is a strictly positive definite matrix. From these hypotheses it follows that the matrix $GH_{pp}^T G^T$ is invertible (see Lemma I.4.2), hence we get from $(3.1e)$

$$\lambda(q,p) = \left(GH_{pp}^T G^T\right)^{-1}\left(G_q(H_p^T, H_p^T) + GH_{pq}^T H_p^T - GH_{pp}^T H_q^T\right)(q,p)\ , \qquad (3.1f)$$

thus the original system $(3.1a,b,c)$ is a semi-explicit index 3 DAE in Hessenberg form. This explicit relation $(3.1f)$ for λ introduced in $(3.1b)$ defines the *standard underlying ODE* $(3.1a,b)$ which is not an Hamiltonian system in general. All equations $(3.1a,b,c,d,e)$ form an ODAE of index 1 (see Section II.1). A standard analysis shows that on the $2(n-m)$-dimensional manifold

$$V = \left\{(q,p)\in\mathbb{R}^n\times\mathbb{R}^n \mid 0 = g(q)\ ,\ 0 = G(q)H_p^T(q,p)\right\}\ , \qquad (3.2)$$

the flow generated by the equations (3.1) is symplectic (see also [HaNøWa93, Section I.14]).

Disregarding the property (1.4), the ideal properties for a numerical method would be to be symplectic, to have a numerical solution remaining on the manifold V, and to have a high order of convergence occuring with a minimal computational work. The Gauss methods applied to $(3.1a,b,c)$ are symplectic, but they have the disadvantages that the numerical solution does not satisfy the constraints $(3.1c,d,e)$ and that a poor (or even no) convergence occurs (see [HaJay93]). Even if projections are effected they are not superconvergent and the symplecticity property is destroyed. To our knowledge the *Rattle algorithm*, a method of order 2 due to Andersen (see [And83]) and based on the *Verlet method*, is the only known symplectic method preserving the constraints which has been proposed in the literature for separable Hamiltonians of the form $H(q,p)=\frac{1}{2}p^T M^{-1}p+U(q)$ with M a constant positive definite matrix (see [LeSk94] and

[SkBiOk93]). For such Hamiltonians two different approaches have also been derived in [LeRei94]. The first one is the reduction of $(3.1a,b,c)$ to an Hamiltonian state-space form via a parametrization of the constraints $(3.1c)$. The second one is the construction of an unconstrained Hamiltonian system which preserves the constraint manifold V and whose flow reduces to the flow of (3.1) along this manifold. We also mention the investigation by Reich on constrained Hamiltonian systems in [Rei93] where high order symplectic methods preserving the constraints are constructed by composition of a first order method (see also [Yo90]). For short-time computations, a non-symplectic alternative is to integrate the standard underlying ODE and to project frequently the numerical solution onto the manifolds $(3.1c,d,e)$ (see also [AsPe93]).

Here we turn our interest to PRK methods (see Section III.1).

Definition 3.1. The application of an *s-stage PRK method* to the equations $(3.1a,b,c)$ reads

$$q_1 = q_0 + h\sum_{i=1}^{s} b_i k_i \ , \qquad p_1 = p_0 + h\sum_{i=1}^{s} \widehat{b}_i \ell_i \tag{3.3a}$$

where

$$k_i = H_p^T(Q_i,P_i) \ , \qquad \ell_i = -H_q^T(Q_i,P_i) - G^T(Q_i)\Lambda_i \ , \qquad 0 = g(Q_i) \ , \tag{3.3b}$$

and the *internal stages* are given by

$$Q_i = q_0 + h\sum_{j=1}^{s} a_{ij} k_j \ , \qquad P_i = p_0 + h\sum_{j=1}^{s} \widehat{a}_{ij} \ell_j \ , \tag{3.3c}$$

Remark. The existence and uniqueness of a solution to these equations is not guaranteed without some assumptions on the coefficients (see Theorem 4.1 below and Theorem III.2.1).

Let us start by supposing that we have a locally unique solution to this system (3.3). Our aim now is to check the symplecticity condition (2.2) for PRK methods with coefficients satisfying (2.3). Without any surprise we have the following result:

Theorem 3.1. *If the coefficients of the PRK method* (3.3) *satisfy* (2.3) *and if* (q_1,p_1) *are uniquely determined then the numerical flow* $(q_0,p_0)\mapsto(q_1,p_1)$ *is symplectic.*

Proof. This proof is inspired by the calculations of [SS88], [SS92], and [HaNøWa93, Theorem II.16.6]. We neglect the equations $0=g(Q_i)$ for the moment. Using $(3.3a)$ and the bilinearity of the wedge product "$\wedge$" we compute for $k=1,\ldots,n$

$$dq_1^k \wedge dp_1^k - dq_0^k \wedge dp_0^k = h\sum_{i=1}^{s} b_i dk_i^k \wedge dp_0^k + h\sum_{j=1}^{s} \widehat{b}_j dq_0^k \wedge d\ell_j^k + h^2 \sum_{i,j=1}^{s} b_i\widehat{b}_j dk_i^k \wedge d\ell_j^k \ . \tag{3.4}$$

We then replace the differentials dq_0^k, dp_0^k with the help of $(3.3c)$ and we obtain

$$dq_1^k \wedge dp_1^k - dq_0^k \wedge dp_0^k = \tag{3.5}$$
$$h\sum_{i=1}^{s} b_i dk_i^k \wedge dP_i^k + h\sum_{j=1}^{s} \widehat{b}_j dQ_j^k \wedge d\ell_j^k - h^2 \sum_{i,j=1}^{s} (b_i\widehat{a}_{ij} + \widehat{b}_j a_{ji} - b_i\widehat{b}_j) dk_i^k \wedge d\ell_j^k \ .$$

An easy calculation shows that

$$\sum_{k=1}^{n} dk_i^k \wedge dP_i^k + \sum_{k=1}^{n} dQ_i^k \wedge d\ell_i^k = -\sum_{l=1}^{m}\sum_{k=1}^{n} \frac{\partial g^l}{\partial q^k}(Q_i) dQ_i^k \wedge d\Lambda_i^l \,. \tag{3.6}$$

An alternative way to understand this formula is as follows: if the variable λ would be constant then as for unconstrained Hamiltonian systems (see [HaNøWa93, Formula II.16.18]) the expression on the left-hand side of (3.6) would vanish; hence only the terms involving $d\Lambda_i^l$ have to be considered. From the hypotheses (2.3) and the formulas (3.5) and (3.6) we get

$$\sum_{k=1}^{n} dq_1^k \wedge dp_1^k - \sum_{k=1}^{n} dq_0^k \wedge dp_0^k = -h \sum_{i=1}^{s} \widehat{b}_i \sum_{l=1}^{m}\sum_{k=1}^{n} \frac{\partial g^l}{\partial q^k}(Q_i) dQ_i^k \wedge d\Lambda_i^l \,. \tag{3.7}$$

Now by the use of $g(Q_i)=0$, we have $G(Q_i)dQ_i=0$, i.e.,

$$\sum_{k=1}^{n} \frac{\partial g^l}{\partial q^k}(Q_i) dQ_i^k = 0 \tag{3.8}$$

which finally gives the desired result. □

This result is another motivation to consider the constraints (3.1c) of index 3 and not those of reduced indices (3.1d, e) in Definition 3.1.

RK methods are special cases of PRK methods with coefficients satisfying $\widehat{a}_{ij} = a_{ij}$, $\widehat{b}_i = b_i$, and $\widehat{c}_i = c_i$. In [Jay93b] and Chapter IV the convergence behaviour of respectively collocation and RK methods applied to semi-explicit index 3 DAE's in Hessenberg form has been analyzed in detail, confirming the conjecture of [HaLuRo89a, p. 86]. Compared to other methods requiring an equivalent work, *stiffly accurate* RK methods, i.e., methods which satisfy

$$a_{sj} = b_j \qquad \text{for} \quad j = 1, \ldots, s \tag{3.9}$$

are tuned to give highly accurate results when applied to DAE's. Unfortunately this later assumption and the symplecticity condition $b_i a_{ij} + b_j a_{ji} - b_i b_j = 0$ lead to $b_s = 0$ and $a_{is} = 0$ for i satisfying $b_i \neq 0$. Therefore we have the following result:

Theorem 3.2. *No symplectic and stiffly accurate RK schemes exist.* □

This negative result is another motivation for the consideration of PRK methods. For RK methods with an invertible RK matrix A, one can also easily show that (3.9) implies that $R(\infty) = 0$ where R is the stability function of the method, whereas symplectic schemes must satisfy $|R(\infty)| = 1$.

For PRK methods the stiff accuracy condition (3.9) implies that $q_1 = Q_s$ and $g(q_1) = g(Q_s) = 0$. For symplectic PRK methods (see (2.3)) satisfying $b_i \neq 0$ this condition (3.9) implies that

$$\widehat{a}_{js} = 0 \qquad \text{for} \quad j = 1, \ldots, s \,, \tag{3.10}$$

and conversely. Consequently, we restrict now our analysis to PRK methods satisfying (3.10). Under this assumption, Λ_s does not influence the solution of the following nonlinear system originating from (3.3b, c)

$$\begin{aligned} Q_i &= q_0 + h\sum_{j=1}^{s} a_{ij} H_p^T(Q_j, P_j)\ , \\ P_i &= p_0 - h\sum_{j=1}^{s-1} \widehat{a}_{ij}\Big(H_q^T(Q_j, P_j) + G^T(Q_j)\Lambda_j\Big)\ , \qquad 0 = g(Q_i)\ . \end{aligned} \tag{3.11}$$

However, Λ_s enters into the definition of p_1 in (3.3a). It is therefore natural to use this extra freedom by choosing Λ_s such that (q_1, p_1) satisfy (3.1d), i.e., p_1 and Λ_s are solution of

$$\begin{aligned} p_1 &= p_0 - h\sum_{i=1}^{s-1} \widehat{b}_i\Big(H_q^T(Q_i, P_i) + G^T(Q_i)\Lambda_i\Big) - h\widehat{b}_s\Big(H_q^T(Q_s, P_s) + G^T(Q_s)\Lambda_s\Big)\ , \\ 0 &= G(q_1)H_p^T(q_1, p_1)\ . \end{aligned} \tag{3.3d}$$

For constrained Hamiltonian systems this projection onto the manifold (3.1d) does not destroy the symplecticity property shown in Theorem 3.1 (see also [LeSk94]). The system (3.3d) defines implicitly p_1 and Λ_s in a unique way (see Theorem 4.1 part b) below).

Several definitions of the numerical Lagrange multiplier λ_1 are conceivable. One possibility is to define λ_1 such that (q_1, p_1, λ_1) satisfy (3.1e), i.e., λ_1 is given by

$$\lambda_1 = \big(GH_{pp}^T G^T\big)^{-1}\big(G_q(H_p^T, H_p^T) + GH_{pq}^T H_p^T - GH_{pp}^T H_q^T\big)(q_1, p_1)\ . \tag{3.3e}$$

However, a very accurate value for λ_1 may be unnecessary. This remark is important if one wants to avoid the computation of extra derivatives such as G_q. For PRK methods which satisfy $c_s = \widehat{c}_s = 1$ a fairly good choice is often given by $\lambda_1 := \Lambda_s$.

Because of the singularity of the matrix $\widehat{A}$ due to (3.10), the nonlinear system (3.11) does not possess a solution in general. This remark applies specifically to the cases where the coefficients (A, b, c) are those of the Radau IIA and Lobatto IIIC methods, and $(\widehat{A}, b, c)$ those of their dual symplectic methods. As it seems obviously necessary to have as many unknowns as equations in (3.11), the only supplementary reasonable assumption to make on the PRK coefficients is

$$a_{1j} = 0 \qquad \text{for} \quad j = 1, \dots, s\ . \tag{3.12}$$

Thus we get $Q_1 = q_0$ and $g(Q_1) = g(q_0) = 0$ is automatically satisfied if q_0 is consistent. For symplectic PRK methods (see (2.3)) satisfying $b_i \neq 0$, this assumption implies that

$$\widehat{a}_{j1} = b_1 \qquad \text{for} \quad j = 1, \dots, s\ , \tag{3.13}$$

and conversely. Under the assumptions (3.10) and (3.12) the existence and local uniqueness of the PRK solution (q_1, p_1, λ_1) (see Theorem 4.1 below) can be shown provided h is sufficiently small, $\widehat{b}_s \neq 0$, and $\overline{A}_0 := A_0\widehat{A}_0$ is invertible where

$$A_0 = \begin{pmatrix} a_{21} & \dots & a_{2s} \\ \vdots & \ddots & \vdots \\ a_{s1} & \dots & a_{ss} \end{pmatrix}, \qquad \widehat{A}_0 = \begin{pmatrix} \widehat{a}_{11} & \dots & \widehat{a}_{1,s-1} \\ \vdots & \ddots & \vdots \\ \widehat{a}_{s1} & \dots & \widehat{a}_{s,s-1} \end{pmatrix}. \tag{3.14}$$

In this situation, for an efficient solution of the nonlinear system (3.11) (the unknowns are $Q_2,\dots,Q_s,P_1,\dots,P_s,\Lambda_1,\dots,\Lambda_{s-1}$), simplified Newton iterations with the following approximate Jacobian matrix

$$\begin{pmatrix} I_{(s-1)n} & -hA_0\otimes H_{pp}^T(q_0,p_0) & O \\ O & I_{sn} & -h\widehat{A}_0\otimes G^T(q_0) \\ I_{s-1}\otimes G(q_0) & O & O \end{pmatrix} \tag{3.15}$$

lead to very simple iterations (see Section III.6 and also [HaLuRo89a, Section 7]). Only the decomposition of the matrix $(GH_{pp}^TG^T)(q_0,p_0)$ is needed, and at each iteration $s-1$ independent linear systems of dimension m must be solved, this remark being important for a parallel implementation. Due to the invertibility of $\overline{A}_0=A_0\widehat{A}_0$ and of $(GH_{pp}^TG^T)(q_0,p_0)$, the matrix given in (3.15) is invertible. Another important remark is that high order methods allow the use of larger stepsizes than low order methods. However, the higher the order of the method, the more Newton-type iterations are necessary to preserve this order, and the larger the number s of involved internal stages is required. Therefore a trade-off between a high order and a low number s of internal stages must be made for an efficient implementation.

We mention that a particular class of PRK methods is given by the *half-explicit methods (HEM's)* of Ostermann (see [Os93]), whose coefficients satisfy

$$\begin{aligned} &a_{j1}=a_{1j}=a_{2j}=0\ ,\quad a_{sj}=b_j\ ,\quad \widehat{a}_{s-1,j}=\widehat{b}_j \qquad \text{for}\quad j=1,\dots,s\ , \\ &a_{ij}=\widehat{a}_{ij}=0\quad \text{if}\quad i\le j \qquad \text{for}\quad i=1,\dots,s\ ,\quad j=1,\dots,s\ . \end{aligned} \tag{3.16}$$

However, no second order HEM constructed in [Os93] is symplectic. The exception is the first order HEM already presented in [Os90] and [HaLuRo89a, p. 90] which is symplectic if the Hamiltonian is separable.

4. A class of PRK methods for semi-explicit index 3 DAE's in Hessenberg form.

Instead of dealing with the equations (3.1) of Hamiltonian systems with holonomic constraints, we consider the more general autonomous semi-explicit index 3 DAE in Hessenberg form (see Chapter II)

$$y'=f(y,z)\ ,\qquad z'=k(y,z,u)\ ,\qquad 0=g(y)\ . \tag{4.1a,b,c}$$

For Hamiltonian systems with holonomic constraints we have $(y,z,u)=(q,p,\lambda)$, $f(y,z)=H_p^T(q,p)$, $k(y,z,u)=-H_q^T(q,p)-G^T(q)\lambda$ (linear in $u=\lambda$), and $g(y)=g(q)$. Differentiating (5.1c) twice we obtain additionally

$$0=\big(g_yf\big)(y,z)\ ,\qquad 0=\big(g_{yy}(f,f)+g_yf_yf+g_yf_zk\big)(y,z,u)\ . \tag{4.1d,e}$$

We suppose that

$$(g_yf_zk_u)(y,z,u)\qquad \text{is invertible} \tag{4.2}$$

in a vicinity of the exact solution (*index* 3 assumption).

The analysis of the application to such systems of the class of PRK methods (see Definition III.1.1) with invertible matrix $\overline{A} = A\widehat{A}$ is given in Chapter III. From the discussion of the preceding section we only consider here the class of PRK methods with coefficients (A, b, c)-$(\widehat{A}, \widehat{b}, \widehat{c})$ satisfying the following hypotheses

$$H: \quad \begin{array}{l} a_{1j} = 0,\ a_{sj} = b_j,\ \widehat{a}_{js} = 0,\ b_j = \widehat{b}_j,\ c_j = \widehat{c}_j \qquad \text{for} \quad j = 1, \ldots, s\ , \\ \overline{A}_0 := A_0 \widehat{A}_0 \text{ is invertible},\ b_s \neq 0\ , \\ B(p),\ C(q),\ D(r),\ \widehat{C}(\widehat{q}),\ \widehat{D}(\widehat{r}),\ C\widehat{C}(Q),\ D\widehat{D}(R)\ . \end{array}$$

Definition 4.1. One step of an *s-stage PRK method* applied to $(4.1a, b, c)$ with initial values (y_0, z_0, u_0) at x_0 reads

$$y_1 = y_0 + \sum_{i=1}^{s} b_i k_i\ , \qquad z_1 = z_0 + \sum_{i=1}^{s} \widehat{b}_i \ell_i \tag{4.3a}$$

where

$$k_i = hf(Y_i, Z_i)\ , \qquad \ell_i = hk(Y_i, Z_i, U_i)\ , \qquad 0 = g(Y_i)\ , \tag{4.3b}$$

and the *internal stages* are given by

$$Y_i = y_0 + \sum_{j=1}^{s} a_{ij} k_j\ , \qquad Z_i = z_0 + \sum_{j=1}^{s} \widehat{a}_{ij} \ell_j\ . \tag{4.3c}$$

From $a_{1j} = 0$ we get in $Y_1 = y_0$ and $g(Y_1) = g(y_0) = 0$ is automatically satisfied if y_0 is consistent. From $a_{sj} = b_j$ we get $y_1 = Y_s$ and $g(y_1) = g(Y_s) = 0$. From $\widehat{a}_{js} = 0$, U_s does not influence the solution of $(4.3b, c)$, but it enters in the definition of z_1 in $(4.3a)$. As in $(3.3d)$ it is therefore natural to use this extra freedom by choosing U_s such that (y_1, z_1) satisfy $(5.1d)$, i.e., z_1 and U_s are solution of

$$\begin{aligned} z_1 &= z_0 + h \sum_{i=1}^{s-1} \widehat{b}_i k(Y_i, Z_i, U_i) + h\widehat{b}_s k(Y_s, Z_s, U_s)\ , \\ 0 &= (g_y f)(y_1, z_1)\ . \end{aligned} \tag{4.3d}$$

Several definitions of the numerical u-component u_1 are conceivable. One possibility is to define u_1 such that (y_1, z_1, u_1) satisfy $(4.1e)$, i.e., u_1 is the solution of

$$0 = \left(g_{yy}(f, f) + g_y f_y f + g_y f_z k\right)(y_1, z_1, u_1)\ . \tag{4.3e}$$

However, a very accurate value for u_1 may be unnecessary (see Remark 4.1.3 below). This remark is important if one wants to avoid the computation of extra derivatives such as g_{yy}. For PRK methods which satisfy $c_s = \widehat{c}_s = 1$ a fairly good choice is often given by $u_1 := U_s$.

We investigate now the existence and uniqueness of the PRK solution under the assumptions given by H and where the initial values (y_0, z_0, u_0) have been replaced

by approximate h-dependent starting values $(\eta,\zeta,\nu)=(\eta(h),\zeta(h),\nu(h))$. We do not suppose the symplecticity conditions (2.3a,b) for the moment.

Theorem 4.1. (Existence and local uniqueness).

a) Let us suppose that

$$g(\eta)=0\ , \tag{4.4a}$$
$$(g_y f)(\eta,\zeta)=\mathcal{O}(h^2)\ , \tag{4.4b}$$
$$(g_{yy}(f,f)+g_y f_y f+g_y f_z k)(\eta,\zeta,\nu)=\mathcal{O}(h)\ , \tag{4.4c}$$
$$(g_y f_z k_u)(y,z,u) \text{ is invertible in a neighbourhood of } (\eta,\zeta,\nu)\ , \tag{4.4d}$$

and that the hypotheses given by H are satisfied with $q\geq 2$ and $Q\geq 2$. Then for $h\leq h_0$ there exists a locally unique solution to

$$Y_1=\eta\ ,\qquad Y_i=\eta+h\sum_{j=1}^{s}a_{ij}f(Y_j,Z_j)\qquad \text{for}\quad i=2,\ldots,s\ , \tag{4.5a}$$

$$Z_i=\zeta+h\sum_{j=1}^{s-1}\widehat{a}_{ij}k(Y_j,Z_j,U_j)\qquad \text{for}\quad i=1,\ldots,s\ , \tag{4.5b}$$

$$0=g(Y_i)\qquad \text{for}\quad i=1,\ldots,s \tag{4.5c}$$

which satisfies

$$Y_i-\eta=\mathcal{O}(h)\qquad \text{for}\quad i=2,\ldots,s\ , \tag{4.6a}$$
$$Z_i-\zeta=\mathcal{O}(h)\qquad \text{for}\quad i=1,\ldots,s\ , \tag{4.6b}$$
$$U_i-\nu=\mathcal{O}(h)\qquad \text{for}\quad i=1,\ldots,s-1\ . \tag{4.6c}$$

b) Moreover, for

$$y_1=\eta+h\sum_{i=1}^{s}b_i f(Y_i,Z_i)=Y_s\ , \tag{4.7a}$$

$$z_1=\zeta+h\sum_{i=1}^{s-1}\widehat{b}_i k(Y_i,Z_i,U_i)+h\widehat{b}_s k(Y_s,Z_s,U_s)\ , \tag{4.7b}$$

$$0=(g_y f)(y_1,z_1)\ , \tag{4.7c}$$

$$0=\left(g_{yy}(f,f)+g_y f_y f+g_y f_z k\right)(y_1,z_1,u_1)\ , \tag{4.7d}$$

we also have local uniqueness with y_1, z_1, U_s, and u_1 satisfying

$$y_1-\eta=\mathcal{O}(h)\ ,\quad z_1-\zeta=\mathcal{O}(h)\ ,\quad U_s-\nu=\mathcal{O}(h)\ ,\quad u_1-\nu=\mathcal{O}(h)\ . \tag{4.8}$$

Remarks 4.1.

1) A similar proof to that of [HaLuRo89a, Theorem 6.1, pp. 72-74] is possible.

2) If the function k of (4.1) is linear in u then instead of (4.4b,c) it is sufficient to have

$$(g_y f)(\eta,\zeta) = \mathcal{O}(h)\ , \tag{4.4b'}$$

$$(g_{yy}(f,f)+g_y f_y f+g_y f_z k)(\eta,\zeta,\nu) = \mathcal{O}(1)\ , \tag{4.4c'}$$

and the conditions $q\geq 2$ and $Q\geq 2$ can be omitted. However, for U_i and u_1 we only have the estimates

$$U_i-\nu = \mathcal{O}(1)\ , \qquad u_1-\nu = \mathcal{O}(1)\ . \tag{4.9}$$

3) The value of ν in (4.4c) only prescribes the solution of (4.5) and (4.7) to be close to the manifold defined by (4.1e). However, (Y_i, Z_i, U_i) and (y_1, z_1, u_1) are clearly independent of ν.

4) The invertibility of $\overline{A}_0 = A_0\widehat{A}_0$ is essential for the existence of a solution to the nonlinear system (4.5). Because of (3.11) and (3.13) the matrix $\overline{A} := A\widehat{A}$ satisfies $\overline{a}_{is} = 0$ for $i = 1,\dots,s$ and $\overline{a}_{1j} = 0$ for $j = 1,\dots,s$. The condition (4.4d) and the invertibility of the matrix $\overline{A}_0 = A_0\widehat{A}_0$ ensure the invertibility of the Jacobian of the system (4.5) (see formula (4.10) below).

Proof. Part a) can be proved completely similarly to Theorem III.2.1. We get a nonlinear system identical to (III.2.8) where the unknowns are $Y_2,\dots,Y_s, Z_1,\dots,Z_s$, $U_1,\dots,U_{s-1}$. The only modification is that the Jacobian of this system is given by

$$\begin{pmatrix} I+\mathcal{O}(h) & \mathcal{O}(h) & 0 \\ \mathcal{O}(h) & I+\mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \mathcal{O}(1) & A_0\widehat{A}_0 \otimes (g_y f_z k_u)(\eta,\zeta,\nu)+\mathcal{O}(h) \end{pmatrix}\ . \tag{4.10}$$

Concerning part b), z_1, U_s are the unknowns of the equations (4.7b,c). Using again similar techniques to those of part a) we get the nonlinear system

$$0 = z_1-\zeta-h\sum_{i=1}^{s-1}\widehat{b}_i k(Y_i,Z_i,U_i)-h\widehat{b}_s k(Y_s,Z_s,U_s)\ , \tag{4.11a}$$

$$\begin{aligned} 0 = {} & \frac{1}{h}(g_y f)(\eta,\zeta)+ \\ & \sum_{i=1}^{s} b_i g_y(\eta)\int_0^1 f_y\Big(\eta+\tau(y_1-\eta),\zeta+\tau(z_1-\zeta)\Big)d\tau\cdot f(Y_i,Z_i)+ \\ & \sum_{i=1}^{s-1}\widehat{b}_i g_y(\eta)\int_0^1 f_z\Big(\eta+\tau(y_1-\eta),\zeta+\tau(z_1-\zeta)\Big)d\tau\cdot k(Y_i,Z_i,U_i)+ \\ & \widehat{b}_s g_y(\eta)\int_0^1 f_z\Big(\eta+\tau(y_1-\eta),\zeta+\tau(z_1-\zeta)\Big)d\tau\cdot k(Y_s,Z_s,U_s)+ \\ & \sum_{i=1}^{s} b_i\int_0^1 g_{yy}\Big(\eta+\tau(y_1-\eta)\Big)d\tau\cdot\Big(f(Y_i,Z_i)\ ,f(y_1,z_1)\Big)\ . \end{aligned} \tag{4.11b}$$

Because of $b_i = \widehat{b}_i$, (4.4b,c), (4.6), and (4.7a), for $h=0$ the values $z_1 = \zeta(0)$ and $U_s = \nu(0)$ satisfy (4.11). The Jacobian of the system (4.11a,b) with respect to (z_1, U_s) is of the form

$$\begin{pmatrix} I & \mathcal{O}(h) \\ \mathcal{O}(1) & b_s\cdot(g_y f_z k_u)(\eta,\zeta,\nu)+\mathcal{O}(h) \end{pmatrix} \tag{4.12}$$

which is invertible for $h \le h_0$. Concerning u_1, from the preceding results (4.8) related to y_1 and z_1 and because of (4.4c), for $h = 0$ the value $u_1 = \nu(0)$ satisfies (4.7d). The derivative of this equation (4.7d) with respect to u_1 is of the form

$$(g_y f_z k_u)(\eta, \zeta, \nu) + \mathcal{O}(h) \tag{4.13}$$

which is invertible for $h \le h_0$. □

The next result is concerned with the influence of perturbations to (4.5)-(4.7).

Theorem 4.2. (Influence of perturbations). *Let us suppose that the hypotheses of Theorem 4.1 are fulfilled. Let (Y_i, Z_i, U_i) and (y_1, z_1, u_1) be given by (4.5) and (4.7), and let us consider perturbed values $(\widehat{Y}_i, \widehat{Z}_i, \widehat{U}_i)$ and $(\widehat{y}_1, \widehat{z}_1, \widehat{u}_1)$ satisfying*

$$\widehat{Y}_1 = \widehat{\eta} + h\delta_1 , \qquad \widehat{Y}_i = \widehat{\eta} + h\sum_{j=1}^{s} a_{ij} f(\widehat{Y}_j, \widehat{Z}_j) + h\delta_i \quad \text{for} \quad i = 2, \dots, s , \tag{4.14a}$$

$$\widehat{Z}_i = \widehat{\zeta} + h\sum_{j=1}^{s-1} \widehat{a}_{ij} k(\widehat{Y}_j, \widehat{Z}_j, \widehat{U}_j) + h\mu_i \quad \text{for} \quad i = 1, \dots, s , \tag{4.14b}$$

$$0 = g(\widehat{Y}_i) + \theta_i \quad \text{for} \quad i = 1, \dots, s , \tag{4.14c}$$

$$\widehat{y}_1 = \widehat{Y}_s , \tag{4.15a}$$

$$\widehat{z}_1 = \widehat{\zeta} + h\sum_{i=1}^{s-1} \widehat{b}_i k(\widehat{Y}_i, \widehat{Z}_i, \widehat{U}_i) + h\widehat{b}_s k(\widehat{Y}_s, \widehat{Z}_s, \widehat{U}_s) + h\mu_{s+1} , \tag{4.15b}$$

$$0 = (g_y f)(\widehat{y}_1, \widehat{z}_1) + \theta'_{s+1} , \tag{4.15c}$$

$$0 = \big(g_{yy}(f,f) + g_y f_y f + g_y f_z k\big)(\widehat{y}_1, \widehat{z}_1, \widehat{u}_1) + \theta''_{s+1} \tag{4.15d}$$

where we have supposed that

$$\begin{aligned} &\Delta\eta = \mathcal{O}(h^3) , \quad \Delta\zeta = \mathcal{O}(h^2) , \quad \widehat{U}_i - \nu = \mathcal{O}(h) , \quad \widehat{u}_1 - \nu = \mathcal{O}(h) , \\ &\delta_i = \mathcal{O}(h^2) , \quad \mu_i = \mathcal{O}(h) , \quad \theta_i = \mathcal{O}(h^3) , \quad \theta'_{s+1} = \mathcal{O}(h^2) , \quad \theta''_{s+1} = \mathcal{O}(h) . \end{aligned} \tag{4.16}$$

Then we have for $h \le h_0$ the estimates

$$\Delta Y_i = P_y\Delta\eta + hc_i f_z P_z \Delta\zeta + \mathcal{O}\Big(h\|\Delta\eta\| + h^2\|\Delta\zeta\| + \frac{1}{h^2}\|Q_y\Delta\eta\|^2 + \|Q_z\Delta\zeta\|^2 + h\|\delta\| + h^2\|\mu\| + \|\theta\|\Big) , \tag{4.17a}$$

$$\Delta Z_i = P_z\Delta\zeta + \mathcal{O}\Big(\frac{1}{h}\|Q_y\Delta\eta\| + \|P_y\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + \|\delta\| + h\|\mu\| + \frac{1}{h}\|\theta\|\Big) , \tag{4.17b}$$

$$P_{z,i}\Delta Z_i = P_z\Delta\zeta + \mathcal{O}\Big(\|Q_y\Delta\eta\| + h\|P_y\Delta\eta\| + h\|\Delta\zeta\| + \frac{1}{h^3}\|Q_y\Delta\eta\|^2 + \frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\| + h\|\mu\| + \|\theta\|\Big) , \tag{4.17c}$$

$$\Delta U_i = \mathcal{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|+\frac{1}{h}\|P_y\Delta\eta\|+\frac{1}{h}\|Q_z\Delta\zeta\|+\|P_z\Delta\zeta\|+\frac{1}{h}\|\delta\|+\|\mu\|+\frac{1}{h^2}\|\theta\|\Big) \quad (4.17d)$$

$$\Delta U_s = \mathcal{O}\Big(\frac{1}{h^2}\|Q_y\Delta\eta\|+\frac{1}{h}\|P_y\Delta\eta\|+\frac{1}{h}\|Q_z\Delta\zeta\|+\|P_z\Delta\zeta\|+\frac{1}{h}\|\delta\|+\|\mu\|+\|\mu_{s+1}\| + \frac{1}{h^2}\|\theta\|+\frac{1}{h}\|\theta'_{s+1}\|\Big) , \quad (4.17e)$$

$$\Delta y_1 = P_y\Delta\eta+hf_zP_z\Delta\zeta+\mathcal{O}\Big(h\|\Delta\eta\|+h^2\|\Delta\zeta\|+\frac{1}{h^2}\|Q_y\Delta\eta\|^2+\|Q_z\Delta\zeta\|^2 + h\|\delta\|+h^2\|\mu\|+\|\theta\|\Big) , \quad (4.17f)$$

$$\Delta z_1 = P_z\Delta\zeta+\mathcal{O}\Big(\|Q_y\Delta\eta\|+h\|P_y\Delta\eta\|+h\|\Delta\zeta\|+\frac{1}{h^3}\|Q_y\Delta\eta\|^2+\frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\|+h\|\mu\|+h\|\mu_{s+1}\|+\|\theta\|+\|\theta'_{s+1}\|\Big) , \quad (4.17g)$$

$$\Delta u_1 = \mathcal{O}\Big(\|\Delta\eta\|+\|P_z\Delta\zeta\|+h\|Q_z\Delta\zeta\|+\frac{1}{h^3}\|Q_y\Delta\eta\|^2+\frac{1}{h}\|Q_z\Delta\zeta\|^2 + h\|\delta\|+h\|\mu\|+h\|\mu_{s+1}\|+\|\theta\|+\|\theta'_{s+1}\|+\|\theta''_{s+1}\|\Big) \quad (4.17h)$$

where $\delta = (\delta_1,\dots,\delta_s)^T$, $\|\delta\| = \max_i\|\delta_i\|$, *and similarly for* μ *and* θ. P_y, Q_y, P_z, *and* Q_z *are projectors defined under the condition* (4.2) *by*

$$\begin{aligned} S &:= k_u(g_yf_zk_u)^{-1}g_y , \\ Q_y &:= f_zS , \qquad P_y := I-Q_y , \qquad Q_z := Sf_z , \qquad P_z := I-Q_z . \end{aligned} \quad (4.18)$$

□

Remarks 4.2.

1) We have used the notation $\Delta\eta=\widehat{\eta}-\eta$, $\Delta Y_i=\widehat{Y}_i-Y_i$, $\Delta y_1=\widehat{y}_1-y_1$, and similarly for the z- and u-components.
2) The missing arguments for f_z, S, P_y, Q_z, etc., are (η,ζ,ν) or $(\eta,\zeta,G(\eta,\zeta))$ with G as described in (II.1.4). Those of $P_{z,i}$ are $(Y_i,Z_i,G(Y_i,Z_i))$ or (Y_i,Z_i,U_i).
3) The conditions (4.16) ensure that all $\mathcal{O}$-terms in the proof below are small.
4) If $g(\widehat{\eta}) = 0 = g(\eta)$ then $Q_y\Delta\eta = \mathcal{O}(\|\Delta\eta\|^2)$. Consequently, this term may be neglected and the hypothesis $\Delta\eta = \mathcal{O}(h^3)$ can be relaxed to $\mathcal{O}(h^2)$. If we have $(g_yf)(\widehat{\eta},\widehat{\zeta})=0=(g_yf)(\eta,\zeta)$ then similarly $\Delta\zeta=\mathcal{O}(h)$ suffices.
5) If the function k of (4.1) is linear in u, then the terms $\|Q_y\Delta\eta\|^2$ and $\|Q_z\Delta\zeta\|^2$ in (4.17) are multiplied by one additional factor h. In this case, it is sufficient to have

$$\begin{aligned} &\Delta\eta = \mathcal{O}(h^2) , \quad \Delta\zeta = \mathcal{O}(h) , \quad \widehat{U}_i-\nu = \mathcal{O}(1) , \quad \widehat{u}_1-\nu = \mathcal{O}(1) , \\ &\delta_i = \mathcal{O}(h) , \quad \mu_i = \mathcal{O}(1) , \quad \theta_i = \mathcal{O}(h^2) , \quad \theta'_{s+1} = \mathcal{O}(h) , \quad \theta''_{s+1} = \mathcal{O}(1) , \end{aligned} \quad (4.19)$$

but then we only have

$$\Delta U_i = \mathcal{O}(1) , \qquad \Delta u_1 = \mathcal{O}(1) . \quad (4.20)$$

6) The constants implied by the $\mathcal{O}$-terms in (4.17) depend on bounds for certain derivatives of f, g, and k, but not on the constants entering in the $\mathcal{O}$-terms in (4.4*b*) and (4.16) or (4.19), if h is sufficiently small.

7) A crucial observation is that the terms $\|\theta\|/h^2$, $\|\theta\|/h$, $\|\delta\|/h$, $\|\delta\|$, and $\|\mu\|$ are not present in (4.17a,c,f,g,h). The effect of the projections (4.3d,e) is to stabilize the numerical solution as regards the influence of perturbations. Since the constraints (4.1d,e) are taken into account, the ODAE (4.1) is not ill-posed as it is of index 1 (see Section II.1).

Proof. The results (4.17a,b,c,d,f) can be shown completely similarly to Theorem III.2.2 and Theorem IV.3.2.

The estimates (4.17e,g) can be proved with analogous techniques. Subtracting (4.7b) from (4.15b) we obtain by linearization

$$\Delta z_1 = \Delta\zeta + h\sum_{i=1}^{s-1}\widehat{b}_i k_u(Y_i,Z_i,U_i)\Delta U_i + h\widehat{b}_s k_u(Y_s,Z_s,U_s)\Delta U_s + h\mu_{s+1} + \qquad (4.21a)$$
$$\mathcal{O}\left(h\|\Delta Y\| + h\|\Delta Z\| + h\|\Delta U\|^2\right),$$
$$0 = (g_y f_z)(y_1,z_1)\Delta z_1 + \theta'_{s+1} + \mathcal{O}\left(\|\Delta y_1\| + \|\Delta z_1\|^2\right). \qquad (4.21b)$$

Inserting (4.21a) into (4.21b), using $\widehat{b}_s \neq 0$ and the invertibility of the matrix

$$(g_y f_z)(y_1,z_1)k_u(Y_s,Z_s,U_s) \qquad (4.22)$$

for h sufficiently small, we get (4.17e). From (4.21b) we have

$$\Delta z_1 = P_z^1 \Delta z_1 + \mathcal{O}(\|\Delta y_1\| + \|\Delta z_1\|^2 + \|\theta'_{s+1}\|) \qquad (4.23)$$

where the arguments of P_z^1 are given by (y_1,z_1,u_1) with $u_1 = G(y_1,z_1)$ given by (II.1.4). The estimate (4.17g) is obtained with the help of (4.17a,b,d,e,f), (4.21a), and $P_z k_u \equiv 0$.

The estimate (4.17h) simply follows from (4.17f,g) because of

$$\Delta u_1 = \mathcal{O}(\|\Delta y_1\| + \|\Delta z_1\| + \|\Delta u_1\|^2 + \|\theta''_{s+1}\|). \qquad (4.24)$$

□

Now we consider one step of a PRK method (4.3) with consistent initial values $(y_0; z_0, u_0)$ at x_0 and we want to give estimates for the local error.

Theorem 4.3. (Local error). *Let us suppose that the hypotheses given by H are satisfied with $q \geq 2$ and $Q \geq 2$. Then for consistent initial values (y_0,z_0,u_0) at x_0 we have*

$$\delta y_h(x_0) = \mathcal{O}(h^{k+1}), \quad \delta z_h(x_0) = \mathcal{O}(h^{\ell+1}), \quad \delta u_h(x_0) = \mathcal{O}(h^{\ell+1}) \qquad (4.25)$$

where

$$k = \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q-1, Q+\widehat{r}, Q+R),$$
$$\ell = \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q-2, Q+\widehat{r}, Q+R). \qquad (4.26)$$

Remarks 4.3.

1. If the function k of (4.1) is linear in u then the assumptions $q \geq 2$ and $Q \geq 2$ can be omitted. Instead of (4.26) we get

$$\begin{aligned} k &= \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q, Q+\widehat{r}, Q+R) \,, \\ \ell &= \min(p, 2q+2, 2\widehat{q}+2, q+r+1, \widehat{q}+\widehat{r}+1, 2Q-1, Q+\widehat{r}, Q+R) \,. \end{aligned} \tag{4.26'}$$

2. If the RK method (A, b, c) is symmetrical, then by the symplecticity conditions (2.3a, b) it can be easily shown that $(\widehat{A}, b, c)$ is also symmetrical. Therefore in this situation the PRK method (4.3) is symmetrical. Hence it follows that the orders for the local error must be even.

Before giving the proof of this theorem we first need similar definitions and results to those given in Section III.3. The most difficult part is to estimate the local error for the z-component. We only outline the main differences and the slight modifications to bring in the definitions and results of Chapter III which are needed for our purposes. We do not rewrite all nearly identical formulas.

We denote the product matrix $\overline{A}_0 := A_0 \widehat{A}_0$ with $A_0, \widehat{A}_0$ given by (3.15) and we define the new "inverse matrix"

$$\Omega = (\omega_{ij})_{i,j=1}^{s} := \begin{pmatrix} 0 & & \\ \vdots & & \overline{A}_0^{-1} \\ 0 & \cdots & 0 \end{pmatrix} . \tag{4.27}$$

Following the derivation of Section III.3, we also obtain the results of Theorem III.3.1 and Theorem III.3.2, excepted those related to $\ell_s = hk(Y_s, Z_s, U_s)$, z_1, and U_s which will be determined with the help of (4.7b, c). The coefficients Φ_i of Definition III.3.2 remain valid, excepted for $\Phi_s(v)$ with $v \in LDAT3_z$ and $\Phi_s(u)$ with $u \in LDAT3_u$ which enter in the derivatives of ℓ_s, z_1, and U_s. Nevertheless, in Definition III.3.2 the summation indices i and j of a_{ij} take their values in $\{2, \ldots, s\}$ and $\{1, \ldots, s\}$ respectively, those of $\widehat{a}_{ij}$ in $\{1, \ldots, s\}$ and $\{1, \ldots, s-1\}$ respectively, and those of ω_{ij} in $\{1, \ldots, s-1\}$ and $\{2, \ldots, s\}$ respectively.

Concerning ℓ_s, z_1, and U_s we have the following result:

Lemma 4.4. *Under the assumptions of Theorem 4.1 (with consistent values $\Psi_0 = (y_0, z_0, u_0)$ at x_0), the qth derivatives for $q \geq 1$ at $h = 0$ of $\ell_s = hk(Y_s, Z_s, U_s)$, z_1, and U_s satisfy*

$$\ell_s^{(q)}(0) = \sum_{\substack{v \in LDAT3_z \\ \varrho(v)=q}} \gamma(v) \Phi_s(v) F(v)(\Psi_0) \,, \tag{4.28a}$$

$$z_1^{(q)}(0) = \sum_{\substack{v \in LDAT3_z \\ \varrho(v)=q}} \gamma(v) \sum_{i=1}^{s} b_i \Phi_i(v) F(v)(\Psi_0) \,, \tag{4.28b}$$

$$\sum_{i=1}^{s} b_i U_i^{(q)}(0) = \sum_{\substack{u \in LDAT3_u \\ \varrho(u)=q}} \gamma(u) \sum_{i=1}^{s} b_i \Phi_i(u) F(u)(\Psi_0) \tag{4.28c}$$

where the coefficients $\Phi_s(v)$ with $v \in LDAT3_z$ and $\Phi_s(u)$ with $u \in LDAT3_u$, depending on the monotonic labelling of v and u respectively, satisfy

$$\Phi_s(v) = \sum_{\substack{\mu_1,\dots,\mu_m \\ \nu_1,\dots,\nu_n}} a_{s\mu_1}\dots a_{s\mu_m}\hat{a}_{s\nu_1}\dots\hat{a}_{s\nu_n}\Phi_{\mu_1}(t_1)\dots\Phi_{\mu_m}(t_m)\times \Phi_{\nu_1}(v_1)\dots\Phi_{\nu_n}(v_n)\Phi_s(u_1)\dots\Phi_s(u_p) \tag{4.29a}$$

$$\text{if } v=[t_1,\dots,t_m,v_1,\dots,v_n,u_1,\dots,u_p]_z \in LDAT3_z \ ,$$

$$\sum_{i=1}^{s} b_i\Phi_i(u) = \Upsilon(u)\sum_{\substack{M_1,\dots,M_M \\ N_1,\dots,N_N \\ \mu_1,\dots,\mu_m}} b_{M_1}\dots b_{M_M}b_{N_1}\dots b_{N_N}b_{\mu_1}\dots b_{\mu_m}\Phi_{M_1}(T_1)\dots\Phi_{M_M}(T_M)\times \Phi_{N_1}(V_1)\dots\Phi_{N_N}(V_N)\Phi_{\mu_1}(t_1)\dots\Phi_{\mu_m}(t_m) \tag{4.29b}$$

$$\text{if } u=[t_0,t_1,\dots,t_m]_u \in LDAT3_u \text{ with } t_0=[T_1,\dots,T_M,V_1,\dots,V_N]_y \in LDAT3_y$$

and t_0 is the subtree of u among $\{t_0,t_1,\dots,t_m\}$ having the smallest index attached to the root. The coefficients $\Upsilon(u)$ depending on the monotonic labelling of u (only by the position of the smallest index among the meagre vertices attached to the root) are given by

$$\Upsilon(u) = \frac{1}{(\varrho(u)+1)\gamma(u)}\gamma(T_1)\dots\gamma(T_M)\gamma(V_1)\dots\gamma(V_N)\gamma(t_1)\dots\gamma(t_m) \ . \tag{4.30}$$

Remark 4.4. A similar proof to the second one given for Theorem III.3.1 is not directly possible, because we do not obtain strict DA3-series. The theory of DA3-series should require slight modifications to be applicable in our new situation.

Example 4.1. The coefficient $\Upsilon(u)$ of the labelled tree u in Fig. II.5.13 is given by

$$\Upsilon(u) = \frac{1}{7\cdot\frac{1}{8\cdot 7}\cdot 3\cdot 5\cdot\gamma(T_1)}\cdot\gamma(T_1)^2 = \frac{8}{15} \ .$$

Proof. From $0=(g_yf)(y_1,z_1)$ we have for $q\geq 2$

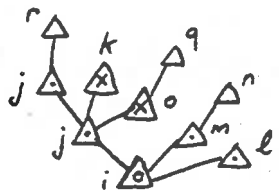
$$0 = \Big((g_yf)(y_1,z_1)\Big)^{(q-2)} \tag{4.31}$$

$$= \sum_{VSLDAT3_{u,q}} \frac{\partial^{1+m}g(y_1)}{\partial y^{1+m}}\left(\frac{\partial^{M+N}f(y_1,z_1)}{\partial y^M\partial z^N}\left(y_1^{(\mu_1^2)},\dots,y_1^{(\mu_M^2)},z_1^{(\nu_1)},\dots,z_1^{(\nu_N)}\right),\right.$$

$$\left. y_1^{(\mu_1^1)},\dots,y_1^{(\mu_m^1)}\right)$$

with $\mu_1^1+\dots+\mu_m^1+\mu_1^2+\dots+\mu_M^2+\nu_1+\dots+\nu_N=q-2$. The trees of $VSLDAT3_{u,q}$ are similar to those described in the proof of Theorem III.3.1. We give below one example

of a such *very special labelled $DAT3_u$-tree* (its corresponding expression is mentioned)



$$\frac{\partial^3 g}{\partial y^3}(y_1)\left(\frac{\partial^3 f}{\partial y \partial z^2}(y_1,z_1)(y_1^{(2)},z_1^{(1)},z_1^{(2)}),y_1^{(2)},y_1^{(1)}\right)$$

Figure 4.1.

From (4.3a) and $b_i=\widehat{b}_i$ we get for $q\geq 1$

$$y_1^{(q)}=\sum_{i=1}^{s} b_i k_i^{(q)}\,, \qquad z_1^{(q)}=\sum_{i=1}^{s} b_i \ell_i^{(q)}\,. \tag{4.32}$$

These formulas can be inserted for $h=0$ into (4.31) leading to

$$\begin{aligned} 0=&(g_y f_z)_0 \sum_{i=1}^{s} b_i \ell_i^{(q-2)}(0)+ \\ &\sum_{\substack{VSLDAT3_{u,q}\\(m,M,N)\neq(0,0,1)}} \left(\frac{\partial^{1+m} g}{\partial y^{1+m}}\right)_0 \left(\left(\frac{\partial^{M+N} f}{\partial y^M \partial z^N}\right)_0 \left(\sum_{i=1}^{s} b_i k_i^{(\mu_1^2)}(0),\ldots,\sum_{i=1}^{s} b_i \ell_i^{(\nu_1)}(0),\ldots\right),\right. \\ &\left.\sum_{i=1}^{s} b_i k_i^{(\mu_1^1)}(0),\ldots\right)\,. \end{aligned} \tag{4.33}$$

Similarly to the proof of Theorem III.3.1 we are not interested in the cases $q=2,3$. Formula (III.3.21b) is also valid for $\ell_s^{(q)}(0)$. Inserting this formula (with q replaced by $q-2$) into the first term of (4.33) yields

$$\begin{aligned} 0=&(q-2)\sum_{i=1}^{s} b_i\cdot(g_y f_z k_u)_0 U_i^{(q-3)}(0)+ \\ &(q-2)\sum_{i=1}^{s} b_i\cdot(g_y f_z)_0\times \\ &\sum_{\substack{SLDAT3_{z,q-2}\\(m,n,p)\neq(0,0,1)}} \left(\frac{\partial^{m+n+p} k}{\partial y^m \partial z^n \partial u^p}\right)_0 \cdot\left(\sum_{l=1}^{s} a_{il} k_l^{(\mu_1)}(0),\ldots,\sum_{l=1}^{s} \widehat{a}_{il} \ell_l^{(\nu_1)}(0),\ldots,U_i^{(\kappa_1)}(0),\ldots\right)+ \\ &\sum_{\substack{VSLDAT3_{u,q}\\(m,M,N)\neq(0,0,1)}} \left(\frac{\partial^{1+m} g}{\partial y^{1+m}}\right)_0 \left(\left(\frac{\partial^{M+N} f}{\partial y^M \partial z^N}\right)_0 \left(\sum_{i=1}^{s} b_i k_i^{(\mu_1^2)}(0),\ldots,\sum_{i=1}^{s} b_i \ell_i^{(\nu_1)}(0),\ldots\right),\right. \\ &\left.\sum_{i=1}^{s} b_i k_i^{(\mu_1^1)}(0),\ldots\right) \end{aligned} \tag{4.34}$$

Since the matrix $(g_y f_z k_u)_0$ is supposed to be regular, then we can extract the expression $\sum_{i=1}^{s} b_i U_i^{(q-3)}(0)$ for $q \geq 4$ from this formula, giving

$$\sum_{i=1}^{s} b_i U_i^{(q-3)}(0) = \tag{4.35}$$

$$\frac{1}{(q-2)}(q-2)\sum_{i=1}^{s} b_i \cdot ((-g_y f_z k_u)^{-1} g_y f_z)_0 \times$$

$$\sum_{\substack{SLDAT3_{z,q-2} \\ (m,n,p)\neq(0,0,1)}} \left(\frac{\partial^{m+n+p} k}{\partial y^m \partial z^n \partial u^p}\right)_0 \left(\sum_{l=1}^{s} a_{il} k_l^{(\mu_1)}(0), \ldots, \sum_{l=1}^{s} \widehat{a}_{il} \ell_l^{(\nu_1)}(0), \ldots, U_i^{(\kappa_1)}(0), \ldots\right) +$$

$$\frac{1}{(q-2)}(-g_y f_z k_u)_0^{-1} \sum_{\substack{VSLDAT3_{u,q} \\ (m,M,N)\neq(0,0,1)}} \left(\frac{\partial^{1+m} g}{\partial y^{1+m}}\right)_0 \left(\left(\frac{\partial^{M+N} f}{\partial y^M \partial z^N}\right)_0 \left(\sum_{i=1}^{s} b_i k_i^{(\mu_1^2)}(0), \ldots,\right.\right.$$

$$\left.\left.\sum_{i=1}^{s} b_i \ell_i^{(\nu_1)}(0), \ldots\right), \sum_{i=1}^{s} b_i k_i^{(\mu_1^1)}(0), \ldots\right) .$$

From the formulas (III.3.19), (4.32), (4.35), by induction on q, and exploiting the multilinearity of the derivatives, we obtain the desired result. Similar arguments to those given in the first proof of Theorem III.3.1 can be used. For example to each expression appearing in the right-hand side of (4.33) there corresponds a unique m.l. tree and conversely. We illustrate this fact on the following example

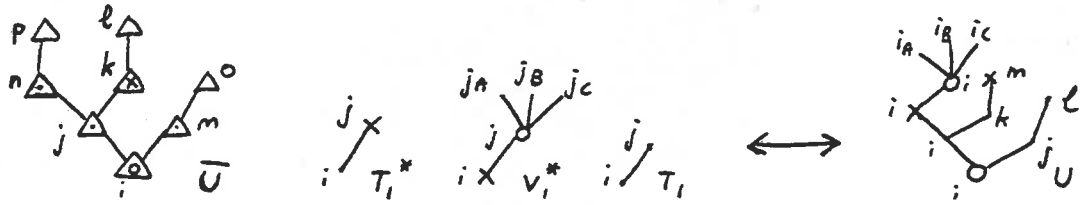


Figure 4.2.

Since each m.l. tree appears exactly once the relations (4.28a, b, c) are established. □

For the numerical solution y_1, z_1, we then easily obtain:

Theorem 4.5. *Under the assumptions of Theorem 4.1 (with consistent values $\Psi_0 = (y_0, z_0, u_0)$ at x_0), the Taylor expansions at x_0 of the numerical solution y_1, z_1 are given by*

$$y_1 = y_0 + \sum_{t\in LDAT3_y} \frac{h^{\varrho(t)}}{\varrho(t)!}\gamma(t)\sum_{i=1}^{s} b_i \Phi_i(t) F(t)(\Psi_0) = \tag{4.36a}$$

$$y_0 + \sum_{t\in DAT3_y} \alpha(t)\frac{h^{\varrho(t)}}{\varrho(t)!}\gamma(t)\sum_{i=1}^{s} b_i \Phi_i(t) F(t)(\Psi_0) ,$$

$$z_1 = z_0 + \sum_{v\in LDAT3_z} \frac{h^{\varrho(v)}}{\varrho(v)!}\gamma(v)\sum_{i=1}^{s} b_i \Phi_i(v) F(v)(\Psi_0) . \tag{4.36b}$$

□

The order conditions for the y-component are given by (III.4.2a). Those for the z-component are given by (III.4.2b) where trees of $LDAT3_z$ must be considered, i.e., a different order condition occurs for each monotonic labelling of a $DAT3_z$-tree. This fact is due to the results of Lemma 4.4.

In order to estimate the local error under the assumptions of Theorem 5.3, the reductions by $C(q)$, $\widehat{C}(\widehat{q})$, $IC\widehat{C}(Q)$, $D(r)$, $\widehat{D}(\widehat{r})$, and $ID\widehat{D}(R)$-$\widehat{D}(\widehat{r})$-(S) $(\widehat{r}\geq 1)$ described in Section III.4 can also be applied in our situation. We are now able to prove Theorem 4.3:

Outline of the proof of Theorem 4.3. The estimate (4.25)-(4.26)-(4.26′) for the y-component can be obtained with the same techniques used in the proof of Theorem III.4.3 to show (III.4.11a)-(III.4.11′a).

We outline the main points of the proof for the z-component. The techniques used to prove (4.25)-(4.26)-(4.26′) are again similar. We do not need the results on the composition of $DA3$-series of Section II.5. If for a tree $u\in LDAT3_u$ the right-hand side of (4.29b) can be reduced, i.e., if all expressions $\sum_{i=1}^{s} b_i\Phi_i(w)$ are replaced by $1/\gamma(w)$, then we get

$$\sum_{i=1}^{s} b_i\Phi_i(u) = \frac{1}{\gamma(u)(\varrho(u)+1)}\ . \tag{4.37}$$

After application of the reductions $C(q)$, $\widehat{C}(\widehat{q})$, etc., there remain trees of the form

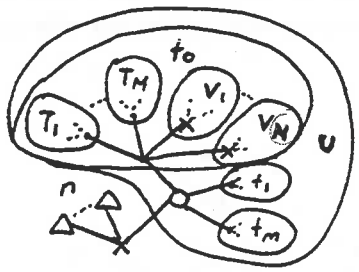


Figure 4.3.

where $u=[t_0,t_1,\ldots,t_m]_u\in LDAT3_u$ with $t_0=[T_1,\ldots,T_M,V_1,\ldots,V_N]_y\in LDAT3_y$. From the simplifying assumptions $ID\widehat{D}(R)$-$\widehat{D}(1)$-(S) we get the following supplementary "reduction"

$$\sum b_i(c_i^k-1)\omega_{ij}\varphi_j = k(k-1)\sum b_i c_i^{k-2}\varphi_i - k\varphi_s \qquad \text{for} \quad k=1,\ldots,R\ . \tag{4.38}$$

For $k=1$ the term $k(k-1)\sum b_i c_i^{k-2}\varphi_i$ has to be removed. This relation implies that the order condition of the tree in Fig. 4.3 with $n\leq R$ can be reduced to those of $DAT3_y$-trees or to those of $LDAT3_z$-trees with a smaller height and order, because of

$$\begin{aligned}
\sum b_i c_i^n\Phi_i(u) &= \sum b_i(c_i^n-1)\Phi_i(u)+\sum b_i\Phi_i(u)\\
&= \sum b_i(c_i^n-1)\omega_{ij}a_{j\mu_0}\Phi_{\mu_0}(t_0)\ldots a_{j\mu_m}\Phi_{\mu_m}(t_m)+\sum b_i\Phi_i(u)\\
&= n(n-1)\sum b_i c_i^{n-2}\omega_{ij}a_{j\mu_0}\Phi_{\mu_0}(t_0)\ldots a_{j\mu_m}\Phi_{\mu_m}(t_m)- \qquad (4.39)\\
&\quad \sum a_{s\mu_0}\Phi_{\mu_0}(t_0)\ldots a_{s\mu_m}\Phi_{\mu_m}(t_m)+\\
&\quad \Upsilon(u)\sum_{\substack{M_1,\ldots,M_M\\ N_1,\ldots,N_N\\ \mu_1,\ldots,\mu_m}} b_{M_1}\Phi_{M_1}(T_1)\cdots b_{M_M}\Phi_{M_M}(T_M)b_{N_1}\Phi_{N_1}(V_1)\cdots\\
&\qquad\qquad \cdots b_{N_N}\Phi_{N_N}(V_N)b_{\mu_1}\Phi_{\mu_1}(t_1)\cdots b_{\mu_m}\Phi_{\mu_m}(t_m)\ .
\end{aligned}$$

The estimate (4.25)-(4.26)-(4.26′) for the z-component then follows from the "first worse" order conditions which correspond to the same trees given in Fig. III.4.8.

For the u-component (u_1 defined by (4.3e)), the result trivially follows from the results for the y- and z-component as a consequence of

$$\delta u_h(t_0) = \mathcal{O}\left(\|\delta y_h(t_0)\| + \|\delta z_h(t_0)\|\right) . \tag{4.40}$$

□

We are now able to give a global convergence result.

Theorem 4.6. (Global error). *Consider the differential-algebraic system* (4.1) *with consistent initial values* (y_0, z_0, u_0) *at* x_0 *and the PRK method* (4.3) *with coefficients satisfying the hypotheses of Theorem 4.4. Then for* $x_n - x_0 = nh \leq Const$, *we have*

$$y_n - y(x_n) = \mathcal{O}(h^\ell) \ , \ z_n - z(x_n) = \mathcal{O}(h^\ell) \ , \ u_n - u(x_n) = \mathcal{O}(h^\ell) \tag{4.41}$$

where ℓ *is the value defined in Theorem 4.3.*

Remark 4.5. This theorem remains valid in the case of variable stepsizes with $h = \max_i h_i$.

Proof. The result (4.41) is a direct consequence of Theorem 4.2 and Theorem 4.3. We denote two neighbouring PRK solutions by $\{\widetilde{y}_n, \widetilde{z}_n\}$, $\{\widehat{y}_n, \widehat{z}_n\}$ and their difference by $\Delta y_n = \widetilde{y}_n - \widehat{y}_n$, $\Delta z_n = \widetilde{z}_n - \widehat{z}_n$. We suppose for the moment that

$$\|\widehat{y}_n - y(x_n)\| \leq C_0 h^2 \ , \ \|\widehat{z}_n - z(x_n)\| \leq C_0 h \ , \ \|\Delta y_n\| \leq C_1 h^3 \ , \ \|\Delta z_n\| \leq C_1 h^2 \tag{4.42}$$

(this will be justified below). Because of $g(\widetilde{y}_n) = 0 = g(\widehat{y}_n)$ and $(g_y f)(\widetilde{y}_n, \widetilde{z}_n) = 0 = (g_y f)(\widehat{y}_n, \widehat{z}_n)$, Remark 4.2.4 holds, implying that

$$(Q_y)_n \Delta y_n = \mathcal{O}(\|\Delta y_n\|^2) = \mathcal{O}(h^2 \|(P_y)_n \Delta y_n\|) \ , \tag{4.43a}$$

$$(Q_z)_n \Delta z_n = \mathcal{O}(\|\Delta y_n\| + \|\Delta z_n\|^2) = \mathcal{O}(\|(P_y)_n \Delta y_n\| + h^2 \|(P_z)_n \Delta z_n\|) . \tag{4.43b}$$

Theorem 4.2 can be applied with $\delta = 0$, $\mu = 0$, $\mu_{s+1} = 0$, $\theta = 0$, $\theta'_{s+1} = 0$, and $\theta''_{s+1} = 0$ yielding

$$(P_y)_{n+1} \Delta y_{n+1} = (P_y)_n \Delta y_n + \mathcal{O}\left(h\|(P_y)_n \Delta y_n\| + h\|(P_z)_n \Delta z_n\|\right) \ , \tag{4.43c}$$

$$(P_z)_{n+1} \Delta z_{n+1} = (P_z)_n \Delta z_n + \mathcal{O}\left(h\|(P_y)_n \Delta y_n\| + h\|(P_z)_n \Delta z_n\|\right) . \tag{4.43d}$$

In (4.43) $(P_y)_n$, $(Q_y)_n$, $(P_z)_n$, and $(Q_z)_n$ are evaluated at $(\widehat{y}_n, \widehat{z}_n, \widehat{u}_n)$. The estimates (4.43) lead to

$$\|\Delta y_n\| \leq C\left(\|(P_y)_0 \Delta y_0\| + \|(P_z)_0 \Delta z_0\|\right) \ , \tag{4.44a}$$

$$\|\Delta z_n\| \leq C\left(\|(P_y)_0 \Delta y_0\| + \|(P_z)_0 \Delta z_0\|\right) . \tag{4.44b}$$

Hence the result (4.41) follows from standard techniques (see [HaLuRo89a, Fig. 4.1, p. 36] or [HaNøWa93, Fig. II.3.2, p. 160]). The assumption (4.42) is justified by induction on n provided the constants C_0 and C_1 are chosen sufficiently large and h is sufficiently small. □

5. High order symplectic PRK methods for constrained Hamiltonian systems.

In this section we turn our interest to PRK methods satisfying the hypotheses given by H (see the previous section) and the symplecticity conditions (2.3a,b).

A direct application of Theorem 4.6 to symplectic PRK methods leads to the following result (put $\widehat{q}:=r$ and $\widehat{r}:=q$ in (4.26)-(4.26′), see Remark 2.2.1):

Corollary 5.1. *Consider the differential-algebraic system* (4.1) *with consistent initial values* (y_0, z_0, u_0) *at* x_0 *and the PRK method* (4.3) *with coefficients satisfying the hypotheses given by Theorem 4.3 and Theorem 2.2. Then for* $x_n - x_0 = nh \leq Const$, *we have*

$$y_n - y(x_n) = \mathcal{O}(h^\nu)\ , \qquad z_n - z(x_n) = \mathcal{O}(h^\nu)\ , \qquad u_n - u(t_n) = \mathcal{O}(h^\nu) \tag{5.1}$$

where $\nu = \min(p, 2q+2, 2r+2, q+r+1, 2Q-2)$. *Moreover if the function* k *of* (4.1) *is linear in* u *then convergence of order* $\nu = \min(p, 2q+2, 2r+2, q+r+1, 2Q-1)$ *occurs.* □

For the application of symplectic PRK methods to Hamiltonian systems with holonomic constraints, the convergence behaviour is now a direct consequence of Corollary 5.1.

Corollary 5.2. *Consider the Hamiltonian system with holonomic constraints* (3.1) *with consistent initial values* (q_0, p_0, λ_0) *at* t_0 *and the PRK method* (3.3) *with coefficients satisfying the hypotheses given by* H *and Theorem 2.2. Then for* $t_n - t_0 = nh \leq Const$, *we have*

$$q_n - q(t_n) = \mathcal{O}(h^\nu)\ , \qquad p_n - p(t_n) = \mathcal{O}(h^\nu)\ , \qquad \lambda_n - \lambda(t_n) = \mathcal{O}(h^\nu) \tag{5.2}$$

with $\nu = \min(p, 2q+2, 2r+2, q+r+1, 2Q-1)$. □

The couples of s-stage Lobatto IIIA-IIIB methods (see Section 2) satisfy the simplifying assumptions $C(s)$, $D(s-2)$, $\widehat{C}(s-2)$, and $\widehat{D}(s)$. Hence $C\widehat{C}(s-1)$ and $D\widehat{D}(s-1)$ must hold. In fact they also satisfy $C\widehat{C}(s)$ and $D\widehat{D}(s)$ and this is the subject of the following lemma:

Lemma 5.3. *Suppose that* $a_{1j} = 0$ *and* $a_{sj} = b_j$ *hold for* $j = 1, \ldots, s$, *and that the hypotheses of Theorem 2.2 are fulfilled with* $p = 2s-2$, $q = s-1$, *and* $r = s-2$. *Then* $C\widehat{C}(s)$ *and* $D\widehat{D}(s)$ *are satisfied.*

Proof.

Proof of $C\widehat{C}(s)$. Because of $Q \geq s-1$ it is sufficient to show that the coefficients

$$\delta_i := \sum_{j=1}^{s} \sum_{k=1}^{s} a_{ij} \widehat{a}_{jk} c_k^{s-2} - \frac{c_i^s}{s(s-1)} \tag{5.3}$$

vanish for $i=1,\dots,s$. From $a_{1j}=0$ and $c_1=0$ we have $\delta_1=0$. Using $a_{sj}=b_j$ and $c_s=1$ we get

$$\delta_s = \sum_{k=1}^{s}\sum_{j=1}^{s} b_j \hat{a}_{jk} c_k^{s-2} - \frac{1}{s(s-1)} \overset{\widehat{D}(1)}{=} \sum_{k=1}^{s} b_k(1-c_k)c_k^{s-2} - \frac{1}{s(s-1)} \tag{5.4}$$
$$\overset{B(s)}{=} \frac{1}{s-1} - \frac{1}{s} - \frac{1}{s(s-1)} = 0 .$$

We will next show that the sums

$$S_m := \sum_{i=1}^{s} b_i c_i^{m-1} \delta_i \tag{5.5}$$

vanish for $m=1,\dots,s-2$. This will give the desired result $\delta_i=0$ for $i=2,\dots,s-1$. By the symplecticity condition $a_{ij}=b_j(1-\hat{a}_{ji}/b_i)$ we get $S_m=A_m-B_m-C_m$ where

$$A_m = \sum_{i=1}^{s} b_i c_i^{m-1} \sum_{k=1}^{s}\sum_{j=1}^{s} b_j \hat{a}_{jk} c_k^{s-2} , \tag{5.6}$$

$$B_m = \sum_{j=1}^{s} b_j \sum_{i=1}^{s} \hat{a}_{ji} c_i^{m-1} \sum_{k=1}^{s} \hat{a}_{jk} c_k^{s-2} , \tag{5.7}$$

$$C_m = \frac{1}{s(s-1)} \sum_{i=1}^{s} b_i c_i^{m+s-1} . \tag{5.8}$$

Each term can be computed separately

$$A_m \overset{\widehat{D}(1)}{=} \sum_{i=1}^{s} b_i c_i^{m-1} \sum_{k=1}^{s} b_k(1-c_k)c_k^{s-2} \overset{B(s)}{=} \frac{1}{m}\left(\frac{1}{s-1}-\frac{1}{s}\right) = \frac{1}{ms(s-1)} , \tag{5.9}$$

$$B_m \overset{\widehat{C}(s-2)}{=} \frac{1}{m}\sum_{k=1}^{s}\sum_{j=1}^{s} b_j c_j^m \hat{a}_{jk} c_k^{s-2} \overset{\widehat{D}(s-1)}{=} \frac{1}{m(m+1)} \sum_{k=1}^{s} b_k(1-c_k^{m+1})c_k^{s-2}$$
$$\overset{B(2s-2)}{=} \frac{1}{m(m+1)}\left(\frac{1}{s-1}-\frac{1}{m+s}\right) = \frac{1}{m(m+s)(s-1)} , \tag{5.10}$$

$$C_m \overset{B(2s-2)}{=} \frac{1}{(m+s)s(s-1)} . \tag{5.11}$$

From these results we easily get $S_m=0$.

Proof of $D\widehat{D}(s)$. Because of $R\geq s-1$ it is sufficient to show that the coefficients

$$\mu_k := \sum_{i=1}^{s}\sum_{j=1}^{s} b_i c_i^{s-2} a_{ij} \hat{a}_{jk} - b_k\left(\frac{1}{s}-\frac{c_k}{s-1}+\frac{c_k^s}{s(s-1)}\right) \tag{5.12}$$

vanish for $k=1,\dots,s$. From $\hat{a}_{js}=0$ (see (3.11)) and $c_s=1$ we have $\mu_s=0$. Using $\hat{a}_{j1}=b_1$ (see (3.14)) and $c_1=0$ we get

$$\mu_1 = b_1\left(\sum_{i=1}^{s} b_i c_i^{s-1} - \frac{1}{s}\right) \overset{B(s)}{=} 0 . \tag{5.13}$$

We will next show that the sums

$$T_m := \sum_{k=1}^{s} \mu_k c_k^{m-1} \tag{5.14}$$

vanish for $m = 1, \ldots, s-2$. This will give the desired result $\mu_k = 0$ for $k = 2, \ldots, s-1$. By the symplecticity condition $\widehat{a}_{jk} = b_k(1 - a_{kj}/b_j)$ we get $T_m = D_m - E_m - F_m$ where

$$D_m = \sum_{i=1}^{s} b_i c_i^{s-2} \sum_{j=1}^{s} a_{ij} \sum_{k=1}^{s} b_k c_k^{m-1} , \tag{5.15}$$

$$E_m = \sum_{i=1}^{s} b_i c_i^{s-2} \sum_{j=1}^{s} \frac{a_{ij}}{b_j} \sum_{k=1}^{s} b_k c_k^{m-1} a_{kj} , \tag{5.16}$$

$$F_m = \frac{1}{s} \sum_{k=1}^{s} b_k c_k^{m-1} - \frac{1}{s-1} \sum_{k=1}^{s} b_k c_k^{m} + \frac{1}{s(s-1)} \sum_{k=1}^{s} b_k c_k^{m+s-1} . \tag{5.17}$$

Each term can be computed separately

$$D_m \overset{C(1)}{=} \sum_{i=1}^{s} b_i c_i^{s-1} \sum_{k=1}^{s} b_k c_k^{m-1} \overset{B(s)}{=} \frac{1}{ms} , \tag{5.18}$$

$$E_m \overset{D(s-2)}{=} \frac{1}{m} \sum_{i=1}^{s} b_i c_i^{s-2} \sum_{j=1}^{s} a_{ij}(1 - c_j^m) \overset{C(s-1)}{=} \frac{1}{m} \sum_{i=1}^{s} b_i c_i^{s-1} - \frac{1}{m(m+1)} \sum_{i=1}^{s} b_i c_i^{m+s-1}$$
$$\overset{B(2s-2)}{=} \frac{1}{ms} - \frac{1}{m(m+1)(m+s)} , \tag{5.19}$$

$$F_m \overset{B(2s-2)}{=} \frac{1}{ms} - \frac{1}{(m+1)(s-1)} + \frac{1}{s(s-1)(m+s)} . \tag{5.20}$$

From these results we easily get $T_m = 0$. □

From Corollary 5.1, Corollary 5.2, and Lemma 5.3 we have now the following convergence results:

Corollary 5.4. *For the couples of s-stage Lobatto IIIA-IIIB methods applied to the system* (4.1) *(see* (4.3)*) with consistent initial values* (y_0, z_0, u_0) *at* x_0*, the global error satisfies for* $x_n - x_0 = nh \leq Const$

$$y_n - y(x_n) = \mathcal{O}(h^{2s-2}) , \quad z_n - z(x_n) = \mathcal{O}(h^{2s-2}) , \quad u_n - u(x_n) = \mathcal{O}(h^{2s-2}) . \tag{5.21}$$

Proof. These methods satisfy the hypotheses given by H with $p = 2s-2$, $q = \widehat{r} = s$, $r = \widehat{q} = s-2$, and $Q = R = s$. The invertibility of the matrix $\overline{A}_0 = A_0 \widehat{A}_0$ simply follows from $C\widehat{C}(s)$. □

Corollary 5.5. *For the symplectic couples of s-stage Lobatto IIIA-IIIB methods applied to the constrained Hamiltonian system* (3.1) *(see* (3.3)*) with consistent initial values* (q_0, p_0, λ_0) *at* t_0*, the global error satisfies for* $t_n - t_0 = nh \leq Const$

$$q_n - q(t_n) = \mathcal{O}(h^{2s-2}) , \quad p_n - p(t_n) = \mathcal{O}(h^{2s-2}) , \quad \lambda_n - \lambda(t_n) = \mathcal{O}(h^{2s-2}) . \tag{5.22}$$

Proof. The symplecticity of Lobatto IIIA-IIIB methods has been proved in [Sun92b]. The estimates (5.22) are an immediate consequence of (5.21). □

For separable Hamiltonian systems the 2-stage Lobatto IIIA-IIIB method is half-explicit and is equivalent to the *Rattle algorithm* proposed in [And83] (see also [LeSk94]).

Because of the presence of the "explicit" stage P_s in (2.1) and (3.3) for symplectic PRK methods satisfying $a_{sj} = b_j$, the Lobatto IIIA-IIIB methods are not appropriate when solving stiff Hamiltonian systems, e.g., Hamiltonian systems containing a strong potential of the form

$$\frac{1}{\varepsilon^2} V(q) , \qquad 0 < \varepsilon \ll 1 . \tag{5.23}$$

This has been numerically observed when trying to solve the stiff spring pendulum equations (I.4.23)-(I.4.24) (see [Lu93] and [HaLuRo89a, pp. 10-12]) with Lobatto IIIA-IIIB methods.

For the long-time integration of Hamiltonian systems, a constant-stepsize application of symplectic methods performs generally better than variable-stepsize algorithms if the time-scale does not vary greatly along the solution (see [CalSS92]). The reason lies in a "backward analysis" argument (see [Ha94]). For constant stepsizes and symplectic methods, the numerical solution can be interpreted as the exact solution of a nearby perturbed Hamiltonian system. It is likely that this result can be extended to Hamiltonian systems with holonomic constraints. We also point out that the construction of an embedded PRK scheme is not crucial for a constant-stepsize implementation if an approximation to the global error of the method is not needed.

6. Numerical experiments.

We first notice that for the solution of the nonlinear systems (3.12) or (4.5), the s-stage Lobatto IIIA-IIIB method requires a computational work approximately equivalent to that arising for the $(s-1)$-stage RK methods Radau IIA and Gauss. Hence these methods are comparable.

Example 1: we consider the motion of a particle of mass m and electric charge e, moving on a sphere of radius R under the action of forces due to an electric field $(0,0,E)^T$ and to a magnetic field $(0,0,B)^T$ (see [Cho92, Problem 7.16]). We use the cartesian coordinates $q = (x,y,z)^T$ for the description of the position of the particle. The holonomic constraint is expressed by

$$0 = \sqrt{x^2+y^2+z^2} - R . \tag{6.1}$$

The Lagrangian of the system is given by

$$L(q,\dot q) = \frac{m}{2}(\dot x^2+\dot y^2+\dot z^2) + m\omega(x\dot y - y\dot x) + eEz \tag{6.2}$$

where $\omega = eB/(2mc)$ and c is the speed of light. The generalized momenta are $p = (p_x,p_y,p_z)^T = L_{\dot q}^T(q,\dot q)$ leading to

$$p_x = m\dot x - m\omega y, \qquad p_y = m\dot y + m\omega x, \qquad p_z = m\dot z . \tag{6.3}$$

The Hamiltonian $H = p^T\dot{q} - L$ is given by

$$H(q,p) = \frac{1}{2m}\left((p_x + m\omega y)^2 + (p_y - m\omega x)^2 + p_z^2\right) - eEz \tag{6.4}$$

and is therefore non-separable. The Hamilton equations of motion are

$$\begin{aligned} \dot{x} &= \frac{p_x}{m} + \omega y, \qquad \dot{y} = \frac{p_y}{m} - \omega x, \qquad \dot{z} = \frac{p_z}{m}, \\ \dot{p}_x &= \omega p_y - m\omega^2 x - \frac{x}{R}\lambda, \qquad \dot{p}_y = -\omega p_x - m\omega^2 y - \frac{y}{R}\lambda, \qquad \dot{p}_z = eE - \frac{z}{R}\lambda . \end{aligned} \tag{6.5}$$

One differentiation of (6.1) implies that

$$0 = xp_x + yp_y + zp_z , \tag{6.6}$$

and another one permits to obtain

$$\lambda = \frac{1}{R}\left(\frac{1}{m}(p_x^2 + p_y^2 + p_z^2) - m\omega^2(x^2 + y^2) + zeE\right) . \tag{6.7}$$

We have applied 5000 steps of the 3-stage Lobatto IIIA-IIIB method of order 4 with stepsize $h = 0.12$,

$$m = 1, \qquad \omega = 1, \qquad R = 1, \qquad eE = 1 , \tag{6.8}$$

and consistent initial values

$$\begin{aligned} x(0) &= 0.2, \qquad y(0) = 0.2, \qquad z(0) = \sqrt{0.92}, \\ p_x(0) &= 1, \qquad p_y(0) = -1, \qquad p_z(0) = 0 . \end{aligned} \tag{6.9}$$

We have plotted in Fig. 6.1 the phase portraits (x, p_x) and (z, p_z).

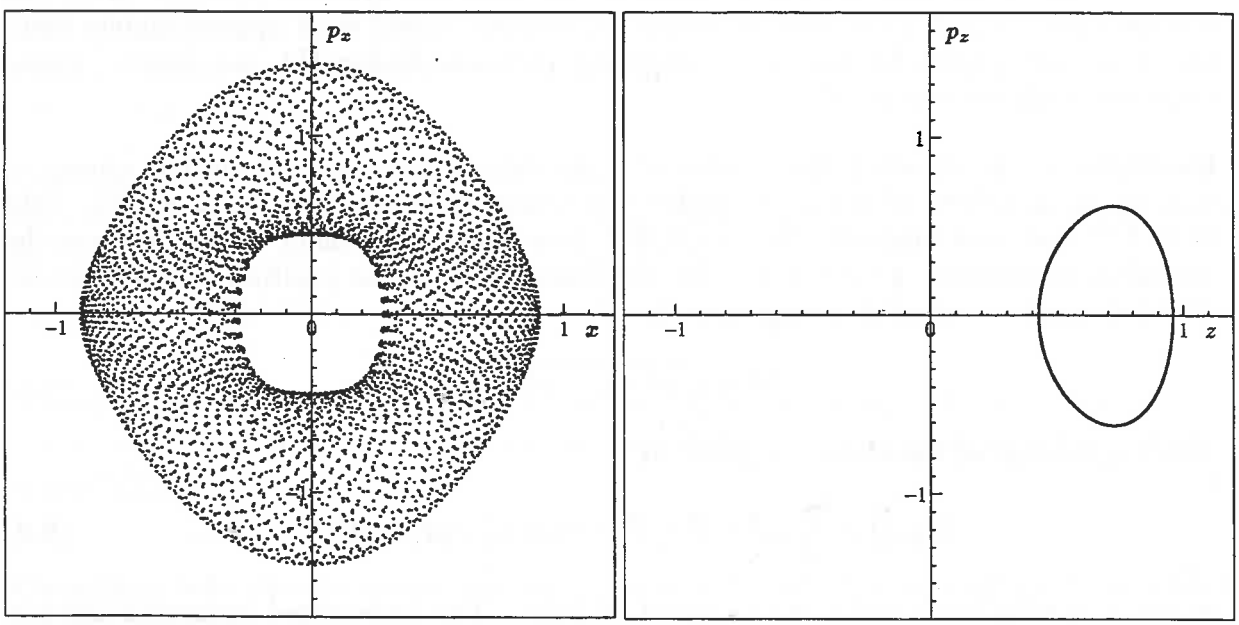


Figure 6.1. The phase portraits (x, p_x) and (z, p_z) of the 3-stage Lobatto IIIA-IIIB method applied to Example 1 with stepsize $h = 0.12$.

In Fig. 6.2 we have drawn the first 500 steps of the numerical Hamiltonian, whose value for the exact solution is $H = 1.2^2 - \sqrt{0.92} \approx 0.4808336955$. The remaining 4500 steps show the same periodic behaviour. If the scale of Fig. 6.4 would be used here then the numerical Hamiltonian would appear nearly equal to the exact value.

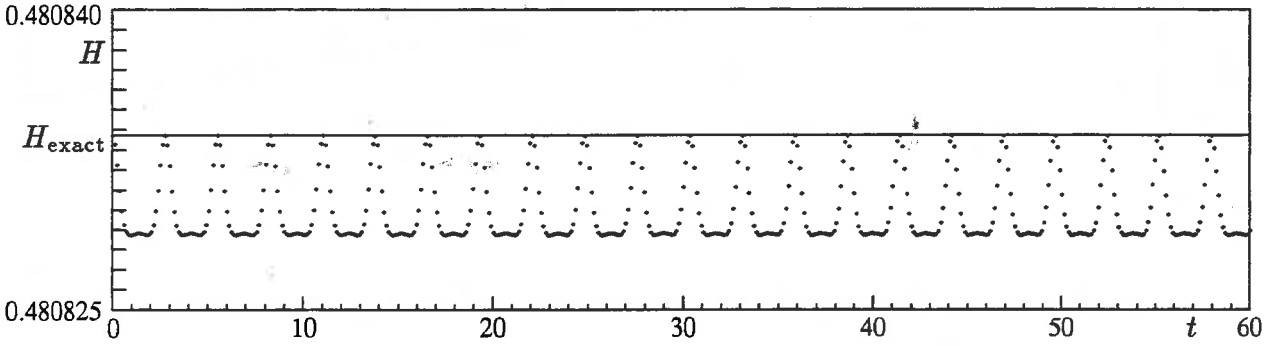


Figure 6.2. The numerical Hamiltonian of the 3-stage Lobatto IIIA-IIIB method applied to Example 1 with stepsize $h=0.12$.

As a comparison we have applied the 2-stage Radau IIA method with the same stepsize $h=0.12$. The numerical results are given in Fig. 6.3 and 6.4.

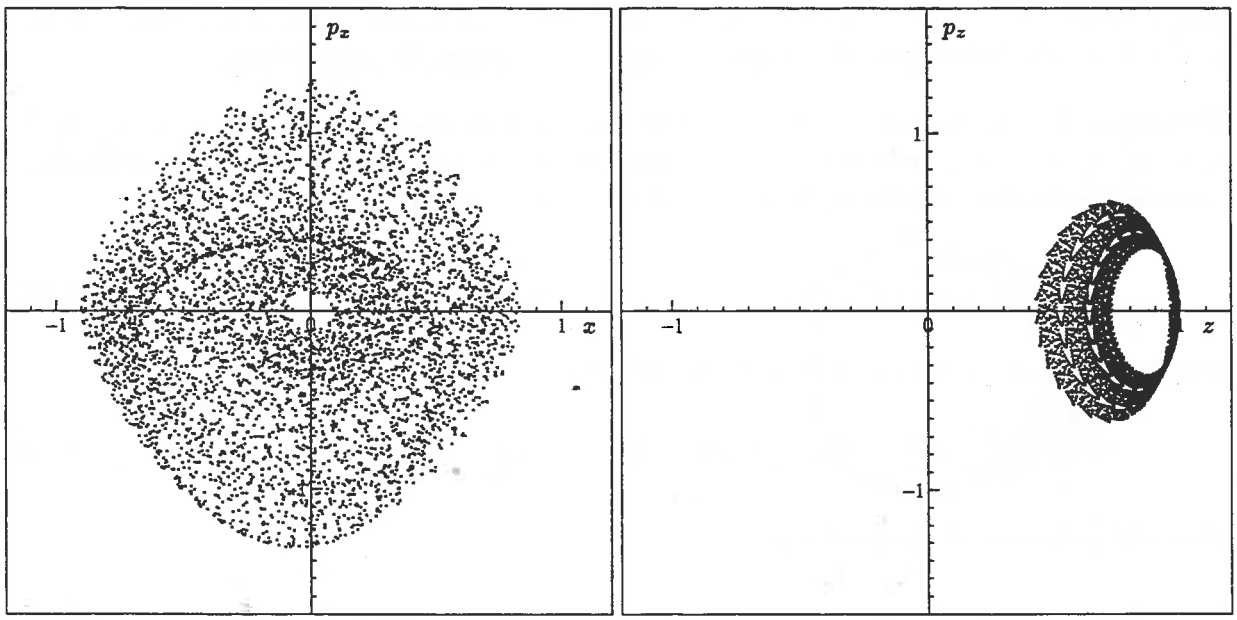


Figure 6.3. The phase portraits (x, p_x) and (z, p_z) of the 2-stage Radau IIA method applied to Example 1 with stepsize $h=0.12$.

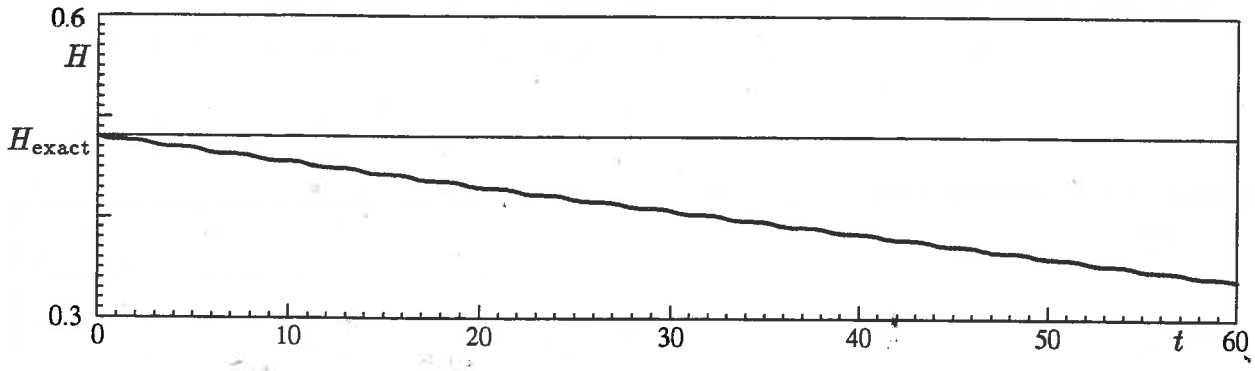


Figure 6.4. The numerical Hamiltonian of the 2-stage Radau IIA method applied to Example 1 with stepsize $h=0.12$.

Since the numerical solution of the Radau IIA method does not satisfy all underlying constraints, we have also applied this method with projections onto these constraints after every step. Although the theoretical order of convergence is improved compared to the unprojected method, the numerical results did not exhibit any visible difference with regards to Fig. 6.3 and Fig. 6.4. For that reason the corresponding figures are not plotted.

We observe that for the Lobatto IIIA-IIIB method the numerical Hamiltonian remains in tolerable bounds, but it drifts away from the exact value (roughly linearly with time) for the unprojected and projected Radau IIA methods. This is a demonstration of the different behaviour of symplectic and non-symplectic integrators.

Example 2: the double pendulum. We use the cartesian coordinates $q_1=(x_1,z_1)^T$, $q_2=(x_2,z_2)^T$ for the description of the position of each pendulum. The two holonomic constraints on the lengths ℓ_1 and ℓ_2 of the two pendula are

$$0=\sqrt{x_1^2+z_1^2}-\ell_1, \qquad 0=\sqrt{(x_2-x_1)^2+(z_2-z_1)^2}-\ell_2\ . \tag{6.10}$$

The Lagrangian of the system is $L=T-U$ where

$$T(\dot q)=\frac{m_1}{2}\left(\dot x_1^2+\dot z_1^2\right)+\frac{m_2}{2}\left(\dot x_2^2+\dot z_2^2\right), \qquad U(q)=m_1gz_1+m_2gz_2\ , \tag{6.11}$$

thus the generalized momenta are

$$p_{x_1}=m_1\dot x_1, \qquad p_{z_1}=m_1\dot z_1, \qquad p_{x_2}=m_2\dot x_2, \qquad p_{z_2}=m_2\dot z_2\ . \tag{6.12}$$

The Hamiltonian $H=T+U$ is given by

$$H(q,p)=\frac{1}{2m_1}\left(p_{x_1}^2+p_{z_1}^2\right)+\frac{1}{2m_2}\left(p_{x_2}^2+p_{z_2}^2\right)+m_1gz_1+m_2gz_2 \tag{6.13}$$

and is separable. The Hamilton equations of motion are

$$\begin{aligned}
\dot{x}_1 &= \frac{p_{x_1}}{m_1}, \qquad \dot{z}_1 = \frac{p_{z_1}}{m_1}, \qquad \dot{x}_2 = \frac{p_{x_2}}{m_2}, \qquad \dot{z}_2 = \frac{p_{z_2}}{m_2}\,, \\
\dot{p}_{x_1} &= -\frac{x_1}{\ell_1}\lambda_1 + \frac{(x_2-x_1)}{\ell_2}\lambda_2, \qquad \dot{p}_{z_1} = -m_1 g - \frac{z_1}{\ell_1}\lambda_1 + \frac{(z_2-z_1)}{\ell_2}\lambda_2\,, \\
\dot{p}_{x_2} &= -\frac{(x_2-x_1)}{\ell_2}\lambda_2, \qquad \dot{p}_{z_1} = -m_2 g - \frac{(z_2-z_1)}{\ell_2}\lambda_2\,.
\end{aligned} \tag{6.14}$$

One differentiation of (6.10) implies that

$$0 = x_1 p_{x_1} + z_1 p_{z_1}\,, \qquad 0 = (x_2-x_1)\left(\frac{p_{x_2}}{m_2} - \frac{p_{x_1}}{m_1}\right) + (z_2-z_1)\left(\frac{p_{z_2}}{m_2} - \frac{p_{z_1}}{m_1}\right)\,, \tag{6.15}$$

and another one leads to

$$\begin{aligned}
0 &= \frac{p_{x_1}^2 + p_{z_1}^2}{m_1} - \ell_1\lambda_1 + \frac{\lambda_2}{2\ell_2}(x_2^2 + z_2^2 - \ell_1^2 - \ell_2^2) - m_1 g z_1\,, \\
0 &= \left(\frac{p_{x_2}}{m_2} - \frac{p_{x_1}}{m_1}\right)^2 + \left(\frac{p_{z_2}}{m_2} - \frac{p_{z_1}}{m_1}\right)^2 + \frac{\lambda_1}{2m_1\ell_1}(x_2^2 + z_2^2 - \ell_1^2 - \ell_2^2) - \lambda_2\ell_2\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\,.
\end{aligned} \tag{6.16}$$

We have applied 5000 steps of the 3-stage Lobatto IIIA-IIIB method of order 4 with stepsize $h=0.12$,

$$m_1 = 1 = m_2, \qquad \ell_1 = 1 = \ell_2, \qquad g = 1\,, \tag{6.17}$$

and consistent initial values

$$\begin{aligned}
x_1(0) &= 0.5, \qquad z_1(0) = -\sqrt{0.75}, \qquad x_2(0) = 0, \qquad z_2(0) = -2\sqrt{0.75}, \\
p_{x_1}(0) &= 0, \qquad p_{z_1}(0) = 0\,, \qquad p_{x_2}(0) = 0, \qquad p_{z_2}(0) = 0\,.
\end{aligned} \tag{6.18}$$

We have plotted in Fig. 6.5 the phase portraits (x_1, p_{x_1}), (z_2, p_{z_2}), and in Fig. 6.6 the first 500 steps of the numerical Hamiltonian whose value for the exact solution is $H = -3\sqrt{0.75} \approx -2.5980762113$.

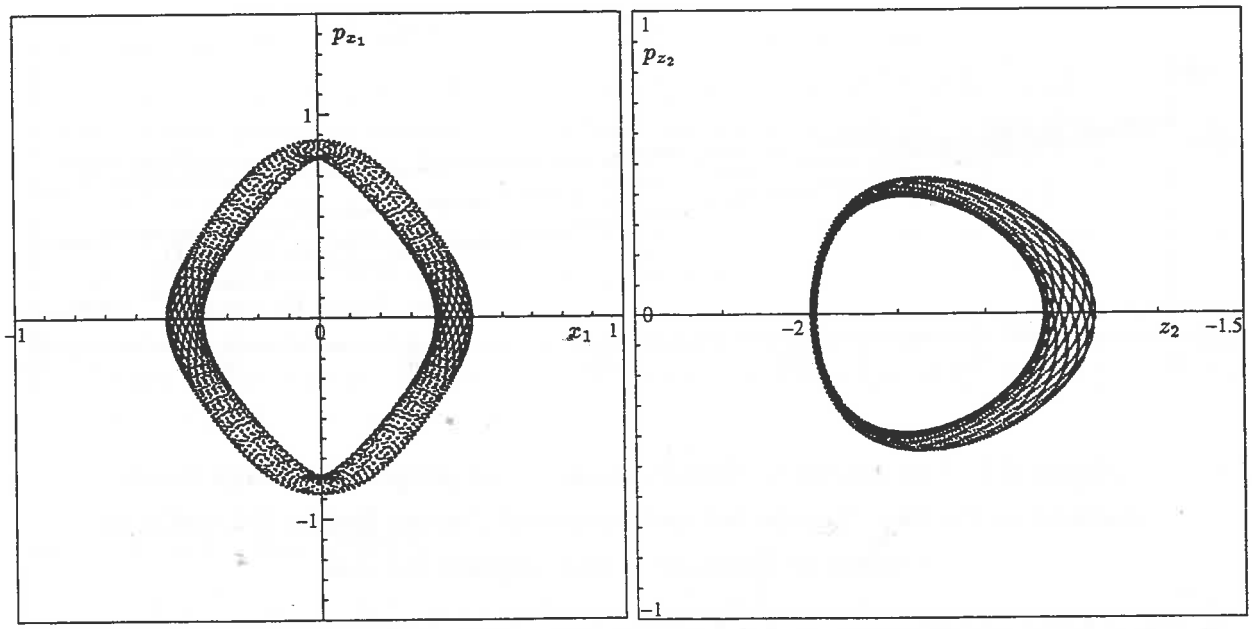


Figure 6.5. The phase portraits (x_1, p_{x_1}) and (z_2, p_{z_2}) of the 3-stage Lobatto IIIA-IIIB method applied to Example 2 with stepsize $h=0.12$.

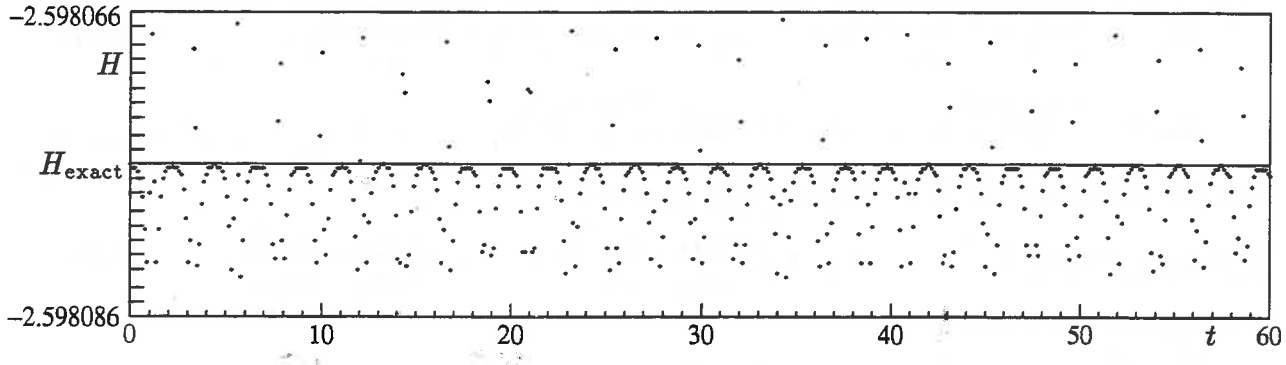


Figure 6.6. The numerical Hamiltonian of the 3-stage Lobatto IIIA-IIIB method applied to Example 2 with stepsize $h=0.12$.

As a comparison we have applied the projected 2-stage Gauss method and the unprojected and projected 2-stage Radau IIA methods to this problem with the same stepsize $h=0.12$. Their numerical Hamiltonian is plotted in Fig. 6.7. We point out that the unprojected 2-stage Gauss method generally diverges when applied to Hamiltonian systems with holonomic constraints (see [HaJay93]). This has been numerically observed for this problem. Although the unprojected Gauss methods are symplectic, the projected Gauss methods are not, and we clearly see here that the numerical Hamiltonian drifts off the exact value. However, this drift is less drastic here than for the unprojected Radau IIA method which in turn is less severe than for the projected Radau IIA method.

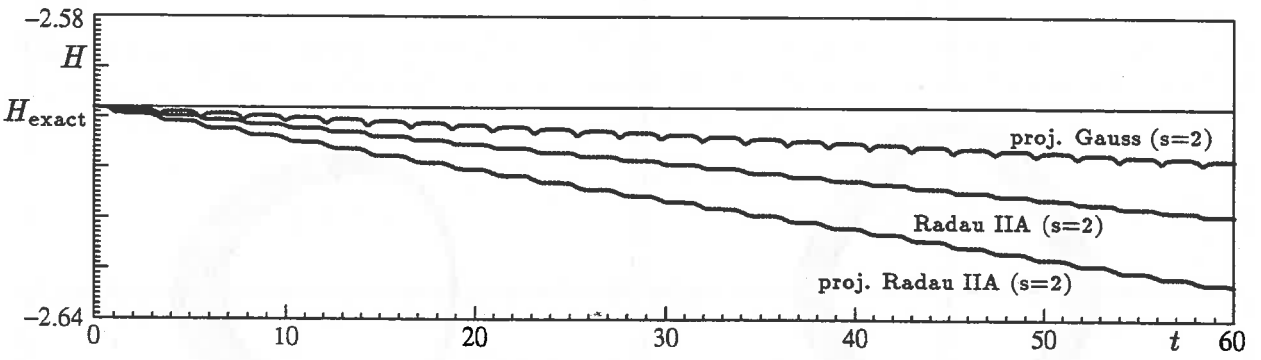


Figure 6.7. The numerical Hamiltonians of the projected 2-stage Gauss method and of the unprojected and projected 2-stage Radau IIA methods applied to Example 2 with stepsize $h=0.12$.

In Fig. 6.8 the global errors at $t=5$ of the four above-mentioned methods have been plotted as functions of h. Since we have used logarithmic scales, the curves appear as straight lines of slope k whenever the leading term of the error is $\mathcal{O}(h^k)$. This behaviour is indicated in the figures.

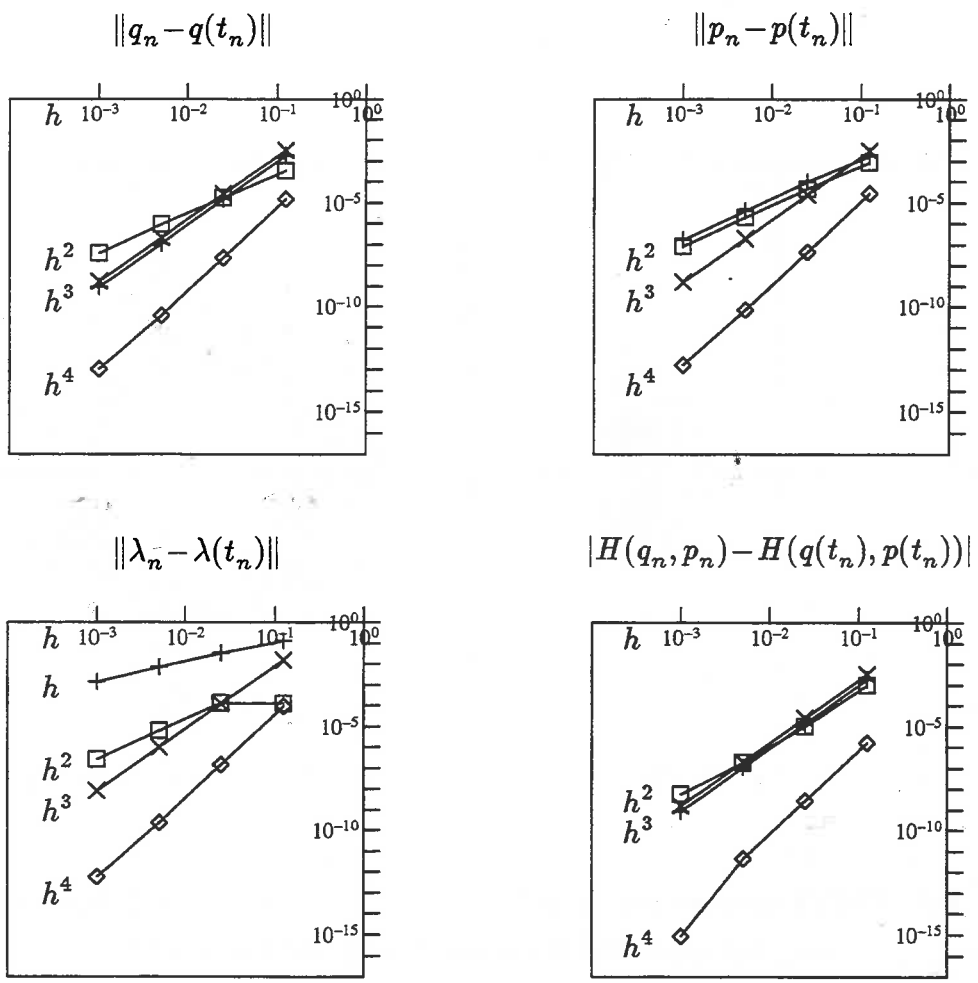


Figure 6.8. Global errors at $t=5$ of four methods applied to Example 2.
(projected 2-stage Gauss:□;projected 2-stage Radau IIA:×;
2-stage Radau IIA:+; 3-stage Lobatto IIIA-IIIB:◇).

The order of convergence of the projected s-stage Gauss method is s and that of the projected s-stage Radau IIA method is $2s-1$ (see Theorem III.5.1). For the unprojected s-stage Radau IIA method the order of convergence is $2s-1$ for the q-component, s for the p-component, and $s-1$ for the λ-component (see Corollary IV.6.2 and [J93b]). The predicted orders are confirmed in Fig. 6.8 and this clearly shows the superiority of the Lobatto IIIA-IIIB schemes also in terms of accuracy.

Example 3: the pendulum, whose equations are given in Section 1. We have applied 5000 steps of the 2-stage Lobatto IIIA-IIIB method with stepsize $h=0.3$ to the ODAE (1.9)-(1.5)-(1.10)-(1.11) with

$$m=1, \qquad \ell=1, \qquad g=1\,, \tag{6.19}$$

and consistent initial values

$$x(0)=0.9, \qquad z(0)=-\sqrt{0.19}, \qquad p_x(0)=0, \qquad p_z(0)=0\,. \tag{6.20}$$

We have plotted in Fig. 6.9 the phase portraits (x,p_x), (z,p_z), and in Fig. 6.10 the first 500 steps of the numerical Hamiltonian whose value for the exact solution is $H=-\sqrt{0.19}=-0.4358898943$.

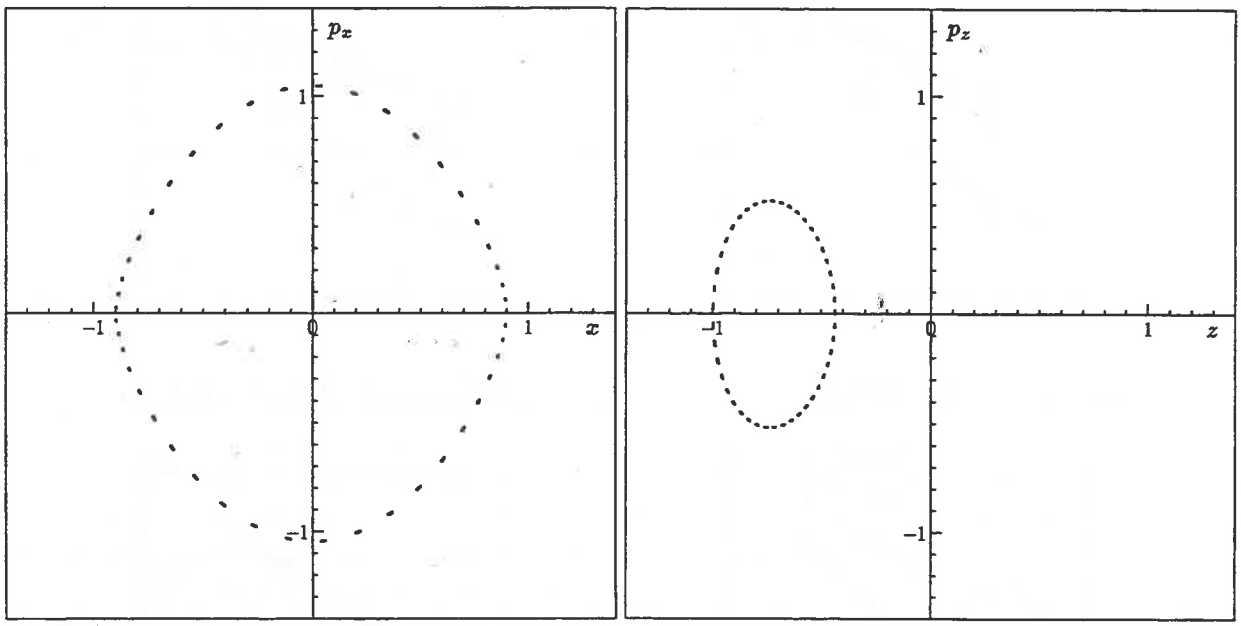


Figure 6.9. The phase portraits (x, p_x) and (z, p_z) of the 2-stage Lobatto IIIA-IIIB method applied to Example 3 with stepsize $h = 0.3$.

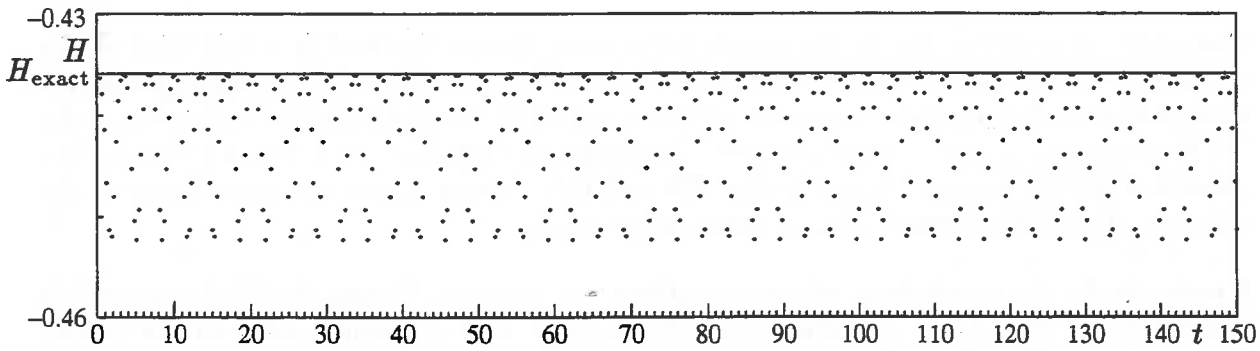


Figure 6.10. The numerical Hamiltonian of the 2-stage Lobatto IIIA-IIIB method applied to Example 3 with stepsize $h = 0.3$.

As a comparison we have applied on the index 3 problem (1.9)-(1.5) 5000 steps of the 2-stage Radau IIA method with stepsize $h = 0.3$. The numerical results are given in Fig. 6.11 and 6.12.

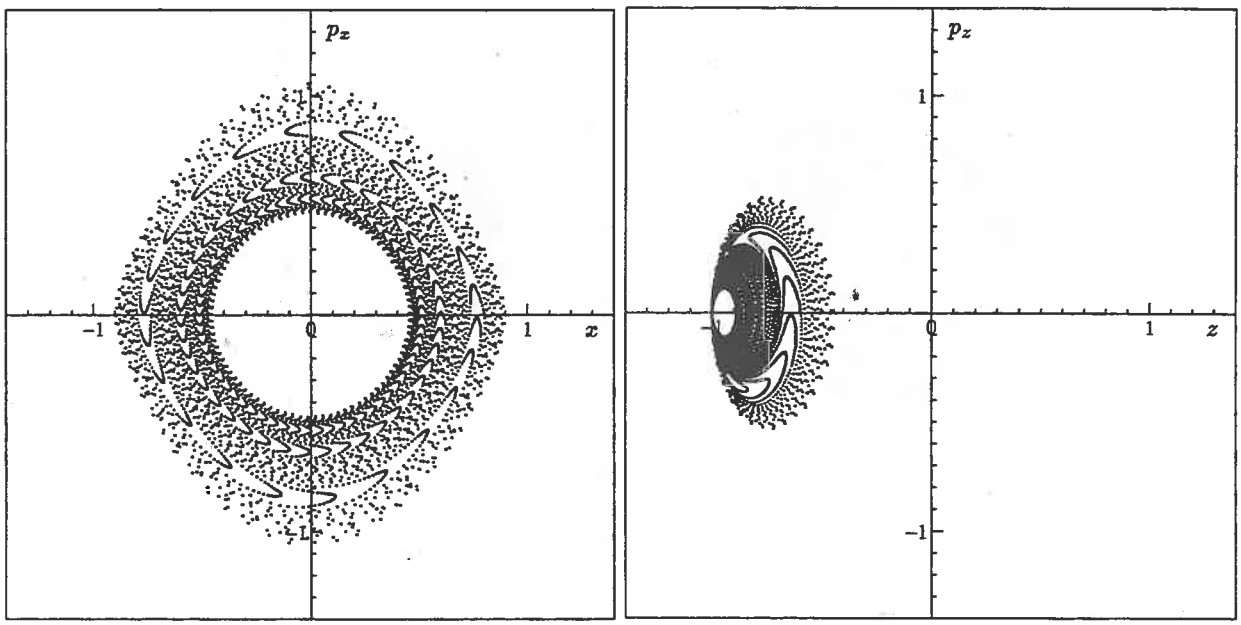


Figure 6.11. The phase portraits (x, p_x) and (z, p_z) of the 2-stage Radau IIA method applied to Example 3 with stepsize $h = 0.3$.

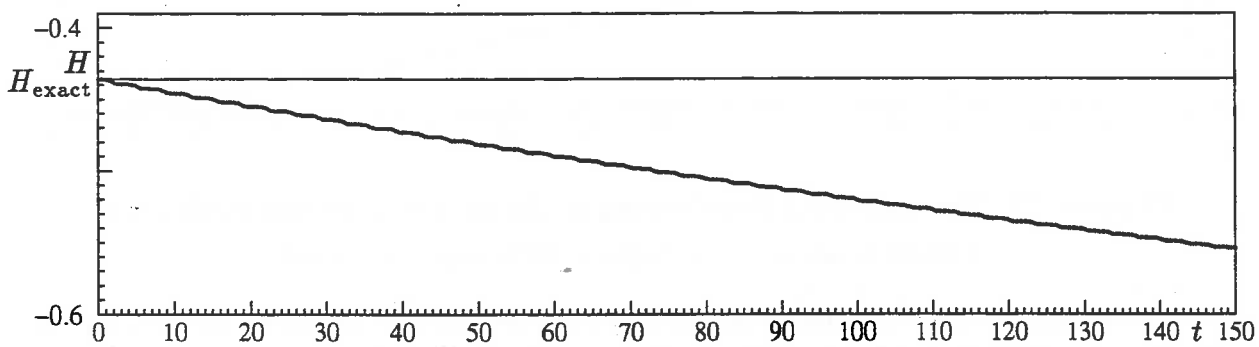


Figure 6.12. The numerical Hamiltonian of the 2-stage Radau IIA method applied to Example 3 with stepsize $h = 0.3$.

As in the preceding example we have also applied the Radau IIA method with projections onto the constraints (1.10) and (1.11) after every step. The numerical results plotted in Fig. 6.13 and 6.14 are worse for this problem.

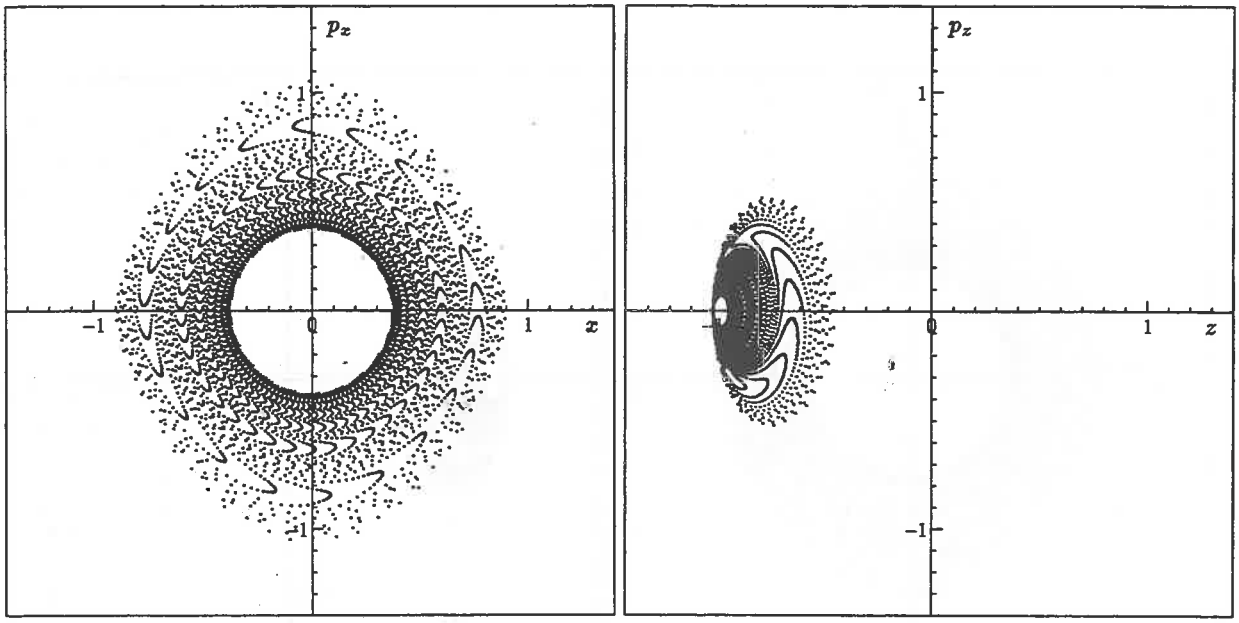


Figure 6.13. The phase portraits (x, p_x) and (z, p_z) of the projected 2-stage Radau IIA method applied to Example 3 with stepsize $h = 0.3$.

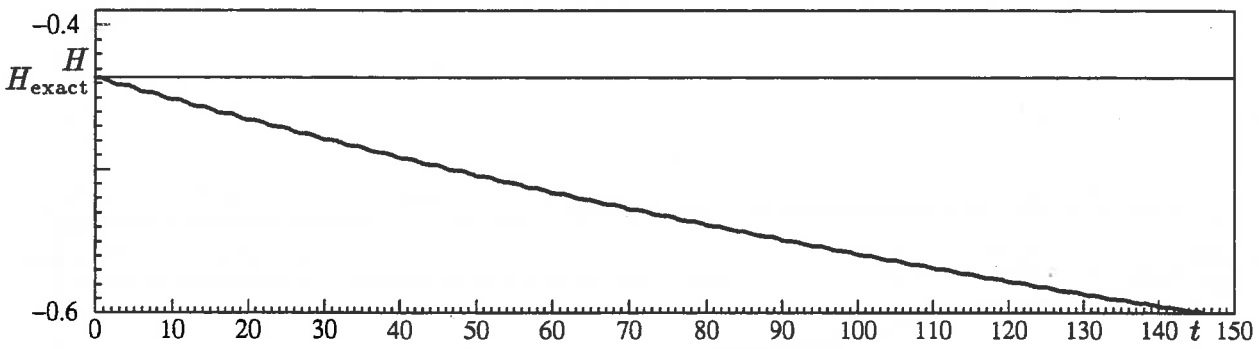


Figure 6.14. The numerical Hamiltonian of the projected 2-stage Radau IIA method applied to Example 3 with stepsize $h = 0.3$.

We observe again that the numerical Hamiltonian remains in tolerable bounds for the Lobatto IIIA-IIIB method, but drifts away from the exact value (roughly linearly with time) for the unprojected and projected Radau IIA methods.

Example 4: For our last experiment we have applied the Lobatto IIIA-IIIB methods to the following index 3 problem (a non-Hamiltonian system)

$$\begin{aligned} y_1' &= 2y_1y_2z_1z_2\ , & y_2' &= -y_1y_2z_2^2\ , \\ z_1' &= (y_1y_2 + z_1z_2)u\ , & z_2' &= -y_1y_2^2z_2^3u^2\ , \\ 0 &= y_1y_2^2 - 1\ , \end{aligned} \tag{6.21}$$

which is of the form (5.1a,b,c) with k nonlinear in u. For the consistent initial values $y(0)=(1,1)^T$, $z(0)=(1,1)^T$, and $u(0)=1$ the exact solution is given by

$$y_1(x) = z_1(x) = e^{2x}, \qquad y_2(x) = z_2(x) = e^{-x}, \qquad u(x) = e^x \ . \tag{6.22}$$

In Fig. 6.15 the global errors at $x_{\text{end}} = 0.1$ for the Lobatto IIIA-IIIB methods ($s = 2,3,4,5,6$) applied to (6.21) have been plotted as functions of h. Since we have used logarithmic scales, the curves appear as straight lines of slope k whenever the leading term of the error is $\mathcal{O}(h^k)$. This behaviour is indicated in the figures.

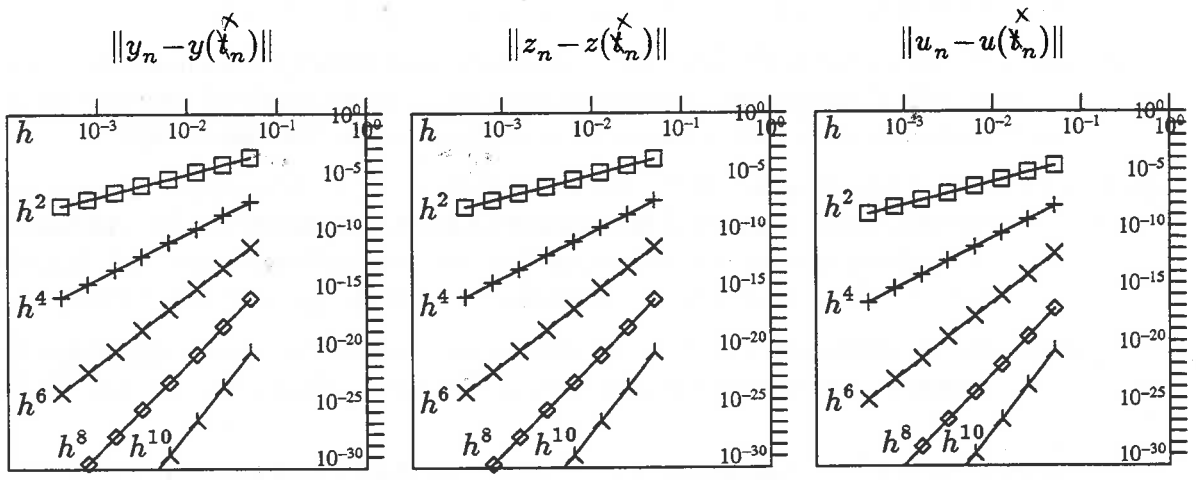


Figure 6.15. Global errors of the Lobatto IIIA-IIIB methods applied to Example 4. ($s=2$:□;3:+; 4:×; 5:◇; 6:⅄).

This is a numerical confirmation of Corollary 5.4.

References.

[AbSS93]: L. Abia & J.M. Sanz-Serna : *Partitioned Runge-Kutta methods for separable Hamiltonian problems.* Math. Comput., Vol. 60, pp. 617-634, (1993).

[Al93]: T. Alishenas : *Zur numerischen Behandlung, Stabilisierung durch Projektion und Modellierung mechanischer Systeme mit Nebenbedingungen und Invarianten.* Ph.D. thesis, Royal Inst. of Tech., Stockholm, Sweden, (1993).

[AlÓl93]: T. Alishenas & Ö. Ólafsson : *Modeling and velocity stabilization of constrained mechanical systems with comparative study of two test problems.* Report, Royal Inst. of Tech., Stockholm, Sweden, (1993).

[AnBoEiSc93]: Th. Andrzejewski, H.G. Bock, E. Eich & R. v. Schwerin : *Recent advances in the numerical integration of multibody systems.* In: *Advanced multibody system dynamics, simulation and software tools.* W. Schiehlen ed., Kluwer Academic Publishers, London, pp. 127-151, (1993).

[And83]: H. C. Andersen : *Rattle: a velocity version of the Shake algorithm for molecular dynamics calculations.* J. Comput. Phys., Vol. 52, pp. 24-34, (1983).

[Aré93]: C. Arévalo : *Matching the structure of DAEs and multistep methods.* Ph.D. thesis, Lund Univ., Sweden, (1993).

[ArV.89]: V.I. Arnold : *Mathematical methods of classical mechanics.* Graduate Texts in Mathematics, Vol. 60, Springer-Verlag, New York, Second Edition, (1989).

[ArV.92]: V.I. Arnold : *Ordinary differential equations.* Springer-Textbook, Springer-Verlag, Berlin, Third Edition, (1992).

[ArM.92]: M. Arnold : Personal communication, (1992).

[ArM.93]: M. Arnold : *Stability of numerical methods for differential-algebraic equations of higher index.* Appl. Numer. Math., Vol. 13, pp. 5-15, (1993).

[ArM.StrWe93]: M. Arnold, K. Strehmel & R. Weiner : *Small perturbations in differential-algebraic systems of index* 2. Preprint 93/1, Univ. Rostock, Germany, (1993).

[As85]: U.M. Ascher : *Two families of symmetric difference schemes for singular perturbation problems.* In: *Numerical boundary values ODEs.* Proceedings of an International Workshop, Vancouver, B.C., Canada, July 10-13, 1984, U.M. Ascher & R.D. Russel eds., Progress in Scientific Computing, Birkhäuser, Vol. 5, pp. 173-191, (1985).

[As89]: U.M. Ascher : *On numerical differential algebraic problems with application to semiconductor device simulation.* SIAM J. Numer. Anal., Vol. 26, pp. 517-538, (1989).

[AsChPeRei93]: U.M. Ascher, H. Chin, L.R. Petzold & S. Reich : *Stabilization of constrained mechanical systems with DAEs and invariant manifolds.* Technical Report CICSR-TR93-02, Univ. of British Columbia, Vancouver, B.C., Canada, (1993).

[AsChRei94]: U.M. Ascher, H. Chin & S. Reich : *Stabilization of DAEs and invariant manifolds.* Numer. Math., (1994). To appear.

[AsLi93]: U.M. Ascher & P. Lin : *Sequential regularization methods for higher index DAEs with constraint singularities: I. Linear index-2 case.* Technical Report 93-24, Univ. of British Columbia, Vancouver, B.C., Canada, (1993).

[AsMaRu88]: U.M. Ascher, R.M.M. Mattheij & R.D. Russel : *Numerical solution of boundary value problems for ordinary differential equations.* Prentice Hall Series in Computational Mathematics, Englewood Cliffs, New Jersey, (1988).

[AsPe91]: U.M. Ascher & L.R. Petzold : *Projected implicit Runge-Kutta methods for differential-algebraic equations.* SIAM J. Numer. Anal., Vol. 28, pp. 1097-1120, (1991).

[AsPe92a]: U.M. Ascher & L.R. Petzold : *Projected collocation methods for higher-order higher-index differential-algebraic equations.* J. Comput. Appl. Math., Vol. 43, pp. 243-259, (1992).

[AsPe92b]: U.M. Ascher & L.R. Petzold : *The numerical solution of delay-differential-algebraic equations of retarded and neutral type.* Technical Report 92-19, Univ. of British Columbia, Vancouver, B.C., Canada, (1992).

[AsPe93]: U.M. Ascher & L.R. Petzold : *Stability of computational methods for constrained dynamics systems.* SIAM J. Sci. Stat. Comput., Vol. 14, pp. 95-120, (1993).

[Bau72]: J. Baumgarte : *Stabilization of constraints and integrals of motion in dynamical systems.* Comput. Math. Appl. Mech. Engrg., Vol. 1, pp. 1-16, (1972).

[Bra92]: V. Brasey : *A half-explicit Runge-Kutta method of order 5 for solving constrained mechanical systems.* Computing, Vol. 48, pp. 191-201, (1992).

[BraHa93a]: V. Brasey & E. Hairer : *Half-explicit Runge-Kutta methods for differential-algebraic systems of index 2.* SIAM J. Numer. Anal., Vol. 30, pp. 538-552, (1993).

[BraHa93b]: V. Brasey & E. Hairer : *Symmetrized half-explicit methods for constrained mechanical systems.* Appl. Numer. Math., Vol. 13, pp. 23-31, (1993).

[Bre83]: K.E. Brenan : *Stability and convergence of difference approximations for higher index differential-algebraic systems with applications in trajectory control.* Ph.D. thesis, Univ. of California at Los Angeles, CA, U.S.A., (1983).

[Bre86]: K.E. Brenan : *Numerical solution of trajectory prescribed path control problems by the backward difference formulas.* IEEE, Trans. Aut. Control, Vol. AC-31, pp. 266-269, (1986).

[BreCamPe89]: K.E. Brenan, S.L. Campbell & L.R. Petzold : *Numerical solution of initial-value problems in differential-algebraic equations.* North-Holland, New York, (1989).

[BreEn88]: K.E. Brenan & B.E. Engquist : *Backward differentiation approxima-*

tions of nonlinear differential/algebraic systems. Math. Comput., Vol. 51, pp. 659-676 & Supplement pp. S7-S16, (1988).

[BrePe89]: K.E. Brenan & L.R. Petzold : *The numerical solution of higher index differential/algebraic equations by implicit Runge-Kutta methods.* SIAM J. Numer. Anal., Vol. 26, pp. 976-996, (1989).

[BryHo75]: A.E. Bryson & Y.-C. Ho : *Applied optimal control.* Revised printing. Hemisphere Publ. Corp., New York, (1975).

[BujBo93]: P. Bujakiewicz & P.P.J. van den Bosch : *Determination of perturbation index of a DAE with maximum weighted matching algorithm.* Report, Delft Univ. of Tech., The Netherlands, (1993).

[BurPe90]: K. Burrage & L.R. Petzold : *On order reduction for Runge-Kutta methods applied to differential/algebraic systems and to stiff systems of ODEs.* SIAM J. Numer. Anal., Vol. 27, pp. 447-456, (1990).

[But63]: J.C. Butcher : *Coefficients for the study of Runge-Kutta integration processes.* J. Austral. Math. Soc., Vol. 3, pp. 185-201, (1963).

[But87]: J.C. Butcher : *The numerical analysis of ordinary differential equations. Runge-Kutta and general linear methods.* John Wiley & Sons, Chichester, (1987).

[CalSS92]: M.P. Calvo & J.M. Sanz-Serna : *Variable steps for symplectic integrators.* In: *Numerical analysis 1991.* D.F. Griffiths & G.A. Watson eds., Pitman Research Notes in Math. Series, No. 260, pp. 34-48, London, (1992).

[CalSS93]: M.P. Calvo & J.M. Sanz-Serna : *Canonical B-series.* Report/2, Univ. of Valladolid, Spain, (1993). To be published in Numer. Math.

[Cam80]: S.L. Campbell : *Singular systems of differential equations I.* Pitman Research Notes in Math. Series, No. 40, London, (1980).

[Cam82]: S.L. Campbell : *Singular systems of differential equations II.* Pitman Research Notes in Math. Series, No. 61, London, (1982).

[Cam89]: S.L. Campbell : *A computational method for general higher index nonlinear singular systems of differential equations.* In: *Numerical and applied mathematics.* IMACS Trans. Scient. Comput., Paris, France, 1988, C. Brezinski ed., J.C. Baltzer AG, IMACS Ann. Comput. Appl. Maths., Vol. 1.2, pp. 555-560, (1989).

[Cam91]: S.L. Campbell : *Least squares completions of nonlinear index three Hessenberg DAEs.* Proceedings of 13th IMACS World Congress on Scientific Computation, Dublin, Ireland, 1991, pp. 1145-1148, (1991).

[Cam93a]: S.L. Campbell : *Least squares completions for nonlinear differential algebraic equations.* Numer. Math., Vol. 65, pp. 77-94, (1993).

[Cam93b]: S.L. Campbell : *High index differential algebraic equations.* Report, North Carolina State Univ., NC, U.S.A., (1993).

[CamGe93]: S.L. Campbell & C.W. Gear : *The index of general nonlinear DAEs.* Report, North Carolina State Univ., NC, U.S.A., (1993).

[CamGr93]: S.L. Campbell & E. Griepentrog : *Solvability of general differential algebraic equations.* Report, North Carolina State Univ., NC, U.S.A.,

(1993).

[CamMo93a]: S.L. Campbell & E. Moore : *Progress on a general numerical method for nonlinear higher index DAEs II.* Circ. Syst. Sign. Proc., (1993). To appear.

[CamMo93b]: S.L. Campbell & E. Moore : *Constraint preserving integrators for general nonlinear higher index DAEs.* Report, North Carolina State Univ., NC, U.S.A., (1993).

[Cha90]: R.P.K. Chan : *On symmetric Runge-Kutta methods of high order.* Computing, Vol. 45, pp. 301-309, (1990).

[Cho92]: P. Choquard : *Mécanique analytique.* Vol. 1, Cahiers mathématiques de l'Ecole Polytechnique Fédérale de Lausanne, Presses Polytechniques et Universitaires Romandes, Switzerland, (1992).

[DeHaZu87]: P. Deuflhard, E. Hairer & J. Zugck : *One-step and extrapolation methods for differential-algebraic systems.* Numer. Math., Vol. 51, pp. 501-516, (1987).

[Ei92]: E. Eich : *Projizierende Mehrschrittverfahren zur numerischen Lösung von Bewegungsgleichungen technischer Mehrkörpersysteme mit Zwangsbedingungen und Unstetigkeiten.* Ph.D. thesis, VDI Fortschrittsberichte, Reihe 18, Nr. 109, VDI-Verlag, Düsseldorf, (1992).

[Ei93]: E. Eich : *Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints.* SIAM J. Numer. Anal., Vol. 30, pp. 1467-1482, (1993).

[EiFüLeRei90]: E. Eich, C. Führer, B.J. Leimkuhler & S. Reich : *Stabilization and projection methods for multibody dynamics.* Research Report A281, Helsinki Univ. of Tech., Finland, (1990).

[EiFüYe92]: E. Eich, C. Führer & J. Yen : *On the error control for multistep methods applied to ODEs with invariants and DAEs in multibody dynamics.* Technical Report R-160, Univ. of Iowa, Iowa City, IA, U.S.A., (1992).

[EiHan91]: E. Eich & M. Hanke : *Regularization methods for constrained mechanical multibody systems.* Preprint 91-8, Humboldt-Univ. of Berlin, Germany, (1991).

[Fü88]: C. Führer : *Differential-algebraische-Gleichungssysteme in mechanischen Mehrkörpersystemen. Theorie, numerische Ansätze und Anwendungen.* Ph.D. thesis, Tech. Univ. Munich, Germany, (1988).

[FüLe89]: C. Führer & B.J. Leimkuhler : *Formulation and numerical solution of the equations of constrained mechanical motion.* Research Report DFVLR-FB 89-08, Oberpaffenhofen, Germany, (1989).

[FüLe91]: C. Führer & B.J. Leimkuhler : *Numerical solution of differential-algebraic equations for constrained mechanical motion.* Numer. Math., Vol. 59, pp. 55-69, (1991).

[Ge71]: C.W. Gear : *Simultaneous numerical solution of differential-algebraic equations.* IEEE Trans. Circuit Theory, Vol. CT-18, pp. 89-95, (1971).

[Ge86]: C.W. Gear : *Maintaining solution invariants in the numerical solution of ODEs.* SIAM J. Sci. Stat. Comput., Vol. 7, pp. 734-743, (1986).

[Ge88]: C.W. Gear : *Differential-algebraic equation index transformation.* SIAM J. Sci. Stat. Comput., Vol. 9, pp. 39-47, (1988).

[Ge90]: C.W. Gear : *Differential algebraic equations, indices, and integral algebraic equations.* SIAM J. Numer. Anal., Vol. 27, pp. 1527-1534, (1990).

[GeGupLe85]: C.W. Gear, G.K. Gupta & B.J. Leimkuhler : *Automatic integration of Euler-Lagrange equations with constraints.* J. Comput. Appl. Math., Vols. 12-13, pp. 77-90, (1985).

[GePe84]: C.W. Gear & L.R. Petzold : *ODE methods for the solution of differential/algebraic systems.* SIAM J. Numer. Anal., Vol. 21, pp. 716-728, (1984).

[Gr91]: E. Griepentrog : *Index reduction methods for differential-algebraic equations.* Preprint 91-12, Humboldt-Univ. of Berlin, Germany, (1991).

[GrHaMä91]: E. Griepentrog, M. Hanke & R. März : *Toward a better understanding of differential algebraic equations (Introductory survey).* Preprint 91-13, Humboldt-Univ. of Berlin, Germany, (1991).

[GrMä86]: E. Griepentrog & R. März : *Differential-algebraic equations and their numerical treatment.* Teubner-Texte zur Mathematik, Leipzig, Germany, Band 88, (1986).

[Ha94]: E. Hairer : *Backward analysis of numerical integrators and symplectic methods.* In: *Annals of numerical mathematics. Scientific computation and differential equations.* Proceedings of the SCADE'93 conference, Auckland, New-Zealand, January 1993, K. Burrage, C. Baker, P. v.d. Houwen, Z. Jackiewicz & P.Sharp eds., J.C. Baltzer, Amsterdam, Vol. 1, (1994). To appear.

[HaJay93]: E. Hairer & L. Jay : *Implicit Runge-Kutta methods for higher index differential-algebraic systems.* In: *Contributions in numerical mathematics.* WSSIAA, Vol. 2, pp. 213-224, (1993).

[HaLu88]: E. Hairer & Ch. Lubich : *On extrapolation methods for stiff and differential algebraic equations.* In: *Numerical treatment of differential equations.* Proceedings of the Fourth Seminar "Numdiff-4", Halle, Germany, 1987, K. Strehmel ed., Teubner-Texte zur Mathematik, Band 104, pp. 130-137, (1988).

[HaLuRo88]: E. Hairer, Ch. Lubich & M. Roche : *Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations.* BIT, Vol. 28, pp. 678-700, (1988).

[HaLuRo89a]: E. Hairer, Ch. Lubich & M. Roche : *The numerical solution of differential-algebraic systems by Runge-Kutta methods.* Lecture Notes in Mathematics, Vol. 1409, Springer-Verlag, (1989).

[HaLuRo89b]: E. Hairer, Ch. Lubich & M. Roche : *Error of Rosenbrock methods for stiff problems studied via differential algebraic equations.* BIT, Vol. 29, pp. 77-90, (1989).

[HaNøWa93]: E. Hairer, S.P. Nørsett & G. Wanner : *Solving ordinary differential equations I. Nonstiff problems.* Second Revised Edition, Computational Mathematics, Vol. 8, Springer-Verlag, Berlin, (1993).

[HaWa73]: E. Hairer & G. Wanner : *Multistep-multistage-multiderivative methods*

for ordinary differential equations. Computing, Vol. 11, pp. 287-303, (1973).

[HaWa74]: E. Hairer & G. Wanner : *On the Butcher group and general multi-value methods.* Computing, Vol. 13, pp. 1-15, (1974).

[HaWa91]: E. Hairer & G. Wanner : *Solving ordinary differential equations II. Stiff and differential-algebraic problems.* Computational Mathematics, Vol. 14, Springer-Verlag, Berlin, (1991).

[Ham86]: R. W. Hamming : *Numerical methods for scientists and engineers.* Second Reviewed Edition, Dover books on engineering, Dover Publications, New York, (1986).

[Han90]: M. Hanke : *Regularization methods for higher index differential-algebraic equations.* Preprint 268, Humboldt-Univ. of Berlin, Germany, (1990).

[Han91]: M. Hanke : *On the asymptotic representation of a regularization approach to nonlinear semiexplicit higher-index differential-algebraic equations.* IMA J. Appl. Math., Vol. 46, pp. 225-245, (1991).

[Hau89]: E.J. Haug : *Computer aided kinematics and dynamics of mechanical systems. Volume I: Basic methods.* Allyn and Bacon, Boston, U.S.A., (1989).

[Hi91]: I. Higueras-Sanz : *Métodos Runge-Kutta explícitos para la integración numérica de ecuaciones diferenciales algebraicas.* Ph.D. thesis, Univ. of Zaragoza, Spain, (1991).

[Hi93]: I. Higueras-Sanz : *Coefficients of the Taylor expansion for the solution of differential-algebraic systems.* Appl. Numer. Math., Vol. 12, pp. 497-501, (1993).

[JanBru92]: K.P. Jankowski & H. Van Brussel : *An approach to discrete inverse dynamics control of flexible-joint robots.* IEEE, Trans. Robot. Autom., Vol. RA-8, pp. 651-658, (1992).

[Jay92]: L. Jay : *Convergence of Runge-Kutta methods for differential-algebraic systems of index 3.* Report, Univ. of Geneva, Switzerland, (1992). Submitted to Appl. Numer. Math.

[Jay93a]: L. Jay : *Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2.* BIT, Vol. 33, pp. 137-150, (1993).

[Jay93b]: L. Jay : *Collocation methods for differential-algebraic equations of index 3.* Numer. Math., Vol. 65, pp. 407-421, (1993).

[Jay94]: L. Jay : *Symplectic partitioned Runge-Kutta methods for constrained Hamiltonian systems.* Accepted for publication in SIAM J. Numer. Anal., (1994). To appear.

[Ka93]: P. Kaps : *Numerische Lösung der Bewegungsgleichungen für mechanische Systeme mit Zwangsbedingungen.* Preprint, Univ. of Innsbruck, Austria, (1993).

[KeGe91]: J.B. Keiper & C.W. Gear : *The analysis of generalized backwards difference formula methods applied to Hessenberg form differential-algebraic equations.* SIAM J. Numer. Anal., Vol. 28, pp. 833-858, (1991).

[Kno88]: M. Knorrenschild : *Regularisierung von differentiell-algebraischen Systemen - theoretische und numerische Aspekte.* Ph.D. thesis, Rhein.-Westfälische Techn. Hochschule, Aachen, Germany, (1988).

[Kv90]: A. Kværnø : *Runge-Kutta methods applied to fully implicit differential-algebraic equations of index* 1. Math. Comput., Vol. 54, pp. 583-625, (1990).

[Kv92]: A. Kværnø : *The order of Runge-Kutta methods applied to semi-explicit DAEs of index* 1, *using Newton-type iterations to compute the internal stage values.* Preprint numerics 2, Univ. of Trondheim, Norway, (1992).

[La88]: F.M. Lasagni : *Canonical Runge-Kutta methods.* ZAMP, Vol. 39, pp. 952-953, (1988).

[LePeGe91]: B.J. Leimkuhler, L.R. Petzold & C.W. Gear : *Approximation methods for the consistent initialization of differential-algebraic equations.* SIAM J. Numer. Anal., Vol. 28, pp. 205-226, (1991).

[LeRei94]: B.J. Leimkuhler & S. Reich : *Symplectic integration of constrained Hamiltonian systems.* Math. Comput., (1994). To appear.

[LeSk94]: B.J. Leimkuhler & R.D. Skeel : *Symplectic numerical integrators in constrained Hamiltonian systems.* J. Comput. Phys., (1994). To appear.

[Lö79]: P. Lötstedt : *On a penalty function method for the simulation of mechanical systems subject to unilateral constraints.* Technical Report, Royal Inst. of Tech., Stockholm, Sweden, (1979).

[LöPe86]: P. Lötstedt & L.R. Petzold : *Numerical solution of nonlinear differential equations with algebraic constraints I: convergence results for backward differentiation formulas.* Math. Comput., Vol. 46, pp. 491-516, (1986).

[Lu89a]: Ch. Lubich : *h^2-extrapolation methods for differential-algebraic systems of index* 2. Impact Comput. Sci. Eng., Vol. 1, pp. 260-268, (1989).

[Lu89b]: Ch. Lubich : *Linearly implicit extrapolation methods for differential-algebraic systems.* Numer. Math., Vol. 55, pp. 197-211, (1989).

[Lu91a]: Ch. Lubich : *Extrapolation integrators for constrained multibody systems.* Impact Comput. Sci. Eng., Vol. 3, pp. 213-234, (1991).

[Lu91b]: Ch. Lubich : *On projected Runge-Kutta methods for differential-algebraic equations.* BIT, Vol. 31, pp. 545-550, (1991).

[Lu93]: Ch. Lubich : *Integration of stiff mechanical systems by Runge-Kutta methods.* ZAMP, Vol. 44, pp. 1022-1053, (1993).

[LuRo90]: Ch. Lubich & M. Roche : *Rosenbrock methods for differential-algebraic systems with solution-dependent singular matrix multiplying the derivative.* Report, Univ. of Geneva, Switzerland, (1989). Computing, Vol. 43, pp. 325-342, (1990).

[Mä85a]: R. März : *On initial value problems in differential-algebraic equations and their numerical treatment.* Computing, Vol. 35, pp. 13-37, (1985).

[Mä85b]: R. März : *On well-posedness and ill-posedness in case of differential-algebraic equations.* Preprint 92, Humboldt-Univ. of Berlin, Germany, (1985).

[Mä89]: R. März : *Some new results concerning index-3 differential-algebraic equations.* J. Math. Anal. Appl., Vol. 140, pp. 177-199, (1989).

[Mä92]: R. März : *Numerical methods for differential algebraic equations.* Acta Numerica, Vol. 1, pp. 141-198, (1992).

[Mä93]: R. März : *Progress in handling differential algebraic equations.* Preprint 93-3, Proceedings of the SCADE93 conference in Auckland, (1993). To appear.

[MK92]: R.S. MacKay : *Some aspects of the dynamics and numerics of Hamiltonian systems.* In: *The dynamics of numerics and the numerics of dynamics.* D.S. Broomhead & A. Iserles eds., Clarendon Press, Oxford, (1992).

[Ni90]: H.-D. Niepage : *On the numerical solution of differential-algebraic equations with discontinuities.* In: *Numerical treatment of differential equations.* Proceedings of the Fifth Seminar "Numdiff-5", Halle, Germany, 1989, K. Strehmel ed., Teubner-Texte zur Mathematik, Band 121, pp. 108-116, (1990).

[OM74] R.E. O'Malley : *Introduction to singular perturbations.* Academic Press, Applied Mathematics and Mechanics, Vol. 14, New York, (1974).

[OrRh70] J.M. Ortega & W.C. Rheihboldt : *Iterative solution of nonlinear equations in several variables.* Academic Press, Computer Science and Applied Mathematics, New York, (1970).

[Os90]: A. Ostermann : *A half-explicit extrapolation method for differential-algebraic systems of index 3.* IMA J. Numer. Anal., Vol. 10, pp. 171-180, (1990).

[Os93]: A. Ostermann : *A class of half-explicit Runge-Kutta methods for differential-algebraic systems of index 3.* Appl. Numer. Math., Vol. 13, pp. 165-179, (1993).

[Pan88]: C.C. Pantelides : *The consistent initialization of differential-algebraic systems.* SIAM J. Sci. Stat. Comput., Vol. 9, pp. 213-231, (1988).

[Pe82]: L.R. Petzold : *Differential-algebraic equations are not ODE's.* SIAM J. Sci. Stat. Comput., Vol. 3, pp. 367-384, (1982).

[Pe83]: L.R. Petzold : *A description of DASSL: a differential/algebraic system solver.* In: *Scientific computing.* Proceedings of 10th IMACS World Congress, Montreal, Canada, 1982, R.S. Stepleman et al. eds., North-Holland, Amsterdam, Vol. 1, pp. 65-68, (1983).

[Pe86]: L.R. Petzold : *Order results for implicit Runge-Kutta methods applied to differential/algebraic systems.* SIAM J. Numer. Anal., Vol. 23, pp. 837-852, (1986).

[Pe89]: L.R. Petzold : *Recent developments in the numerical solution of differential/algebraic systems.* Comput. Meth. Appl. Mech. Eng., Vol. 75, pp. 77-89, (1989).

[Pe92]: L.R. Petzold : *Numerical methods for differential-algebraic equations - current status and future directions.* In: *Computational ordinary differential equations.* Proceedings of a Conference on Computational Ordinary Differential Equations, Univ. of London, England, July 1989,

J.R. Cash & I. Gladwell eds., Clarendon Press, Oxford, pp. 259-273, (1992).

[PePo92]: L.R. Petzold & F.A. Potra : *ODAE methods for the numerical solution of Euler-Lagrange equations.* Appl. Numer. Math., Vol. 10, pp. 397-413, (1992).

[PonBoGaMi62]: L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze & E.F. Mishchenko : *The mathematical theory of optimal processes.* In Russian, Fitmatgiz Moskow, (1961). Engl. Transl., John Wiley & Sons, New York, (1962).

[Po93a]: F.A. Potra : *Implementation of linear multistep methods for solving constrained equations of motion.* SIAM J. Numer. Anal., Vol. 30, pp. 774-789, (1993).

[Po93b]: F.A. Potra : *Runge-Kutta integrators for multibody dynamics.* Report, Univ. of Iowa, Iowa City, IA, U.S.A., (1993).

[Po93c]: F.A. Potra : *Numerical methods for differential-algebraic equations with application to real-time simulation of mechanical systems.* Report, Univ. of Iowa, Iowa City, IA, U.S.A., (1993).

[PoRh90]: F.A. Potra & W.C. Rheinboldt : *Differential-geometric techniques for solving differential-algebraic equations.* In: *Real-time integration methods for mechanical system simulation.* R.C. Deyo & E.J. Haug eds., NATO ASI Series, Vol. F 69, Computer and Systems Sciences, Springer-Verlag, Berlin, pp. 155-191, (1990).

[PoRh91]: F.A. Potra & W.C. Rheinboldt : *On the numerical solution of Euler-Lagrange equations.* J. Mech. Struct. Mach., Vol. 19, pp. 1-18, (1991).

[PrBeDeSc92]: A.J. Preston, M. Berzins, P.M. Dew & L.E. Scales : *Towards efficient DAE solvers for the solution of dynamic simulation problems.* In: *Computational ordinary differential equations.* Proceedings of a conference on Computational Ordinary Differential Equations, Univ. of London, England, July 1989, J.R. Cash & I. Gladwell eds., Clarendon Press, Oxford, pp. 299-308, (1992).

[Ra89]: P.J. Rabier : *Implicit differential equations near a singular point.* J. Math. Anal. Appl., Vol. 144, pp. 425-449, (1989).

[RaRh91]: P.J. Rabier & W.C. Rheinboldt : *A general existence and uniqueness theory for implicit differential-algebraic equations.* J. Diff. Int. Eq., Vol. 4, pp. 563-582, (1991).

[RaRh92a]: P.J. Rabier & W.C. Rheinboldt : *On impasse points of quasilinear differential-algebraic equations.* Technical Report ICMA-92-171, Univ. of Pittsburgh, PA, U.S.A., (1992).

[RaRh92b]: P.J. Rabier & W.C. Rheinboldt : *On the computation of impasse points of quasilinear differential-algebraic equations.* Technical Report ICMA-92-172, Univ. of Pittsburgh, PA, U.S.A., (1992).

[RaRh93]: P.J. Rabier & W.C. Rheinboldt : *On the numerical solution of the Euler-Lagrange equations.* Technical Report ICMA-93-177, Univ. of Pittsburgh, PA, U.S.A., (1993).

[RaRh94]: P.J. Rabier & W.C. Rheinboldt : *A geometric treatment of implicit*

differential-algebraic equations. J. Diff. Eq., (1994). To appear.

[Rei90a]: S. Reich : *On a geometrical interpretation of differential algebraic equations.* In: *Numerical treatment of differential equations.* Proceedings of the Fifth Seminar "Numdiff-5", Halle, Germany, 1989, K. Strehmel ed., Teubner-Texte zur Mathematik, Band 121, pp. 134-139, (1990).

[Rei90b]: S. Reich : *On a geometrical interpretation of differential-algebraic equations.* Circ. Syst. Sign. Proc., Vol. 9, pp. 367-382, (1990).

[Rei91]: S. Reich : *On an existence and uniqueness theory for nonlinear differential-algebraic equations.* Circ. Syst. Sign. Proc., Vol. 10, pp. 343-359 & Erratum, Vol. 11, p. 281, (1991).

[Rei92]: S. Reich : *Existence and uniqueness results for nonlinear differential-algebraic equations.* In: *Berlin seminar on differential-algebraic equations.* Research monography 92-1, Humboldt-Univ. of Berlin, Germany, E. Griepentrog, M. Hanke & R. März eds., pp. 61-81, (1992).

[Rei93]: S. Reich : *Symplectic integration of constrained Hamiltonian systems by Runge-Kutta methods.* Report, (1993). To be published.

[RenRoSte89]: P. Rentrop, M. Roche & G. Steinebach : *The application of Rosenbrock-Wanner type methods with stepsize control in differential-algebraic equations.* Numer. Math., Vol. 55, pp. 545-563, (1989).

[RenSte89]: P. Rentrop & G. Steinebach : *The application of Runge-Kutta type methods in vehicle dynamic.* In: Proc. Workshop *"Road-vehicle systems and related mathematics II"*, Turin, Italy, June 1987, H. Neunzert ed., Teubner-Verlag, Stuttgart, pp. 143-161, (1989).

[Rh84]: W.C. Rheinboldt : *Differential-algebraic systems as differential equations on manifolds.* Math. Comput., Vol. 43, pp. 473-482, (1984).

[Rh91a]: W.C. Rheinboldt : *The theory and numerics of differential-algebraic equations.* In: *Advances in numerical analysis.* Oxford Sci. Publ., Oxford Univ. Press, New York, Vol. 1, pp. 237-275, (1991).

[Rh91b]: W.C. Rheinboldt : *On the existence and uniqueness of solutions of nonlinear semi-implicit differential-algebraic equations.* Nonlin. Anal., Th., Meths. & Appls., Vol. 16, pp. 647-661, (1991).

[Ro88a]: M. Roche : *Rosenbrock methods for differential algebraic equations.* Numer. Math., Vol. 52, pp. 45-63, (1988).

[Ro88b]: M. Roche : *Runge-Kutta and Rosenbrock methods for differential-algebraic equations and stiff ODE's.* Ph.D. thesis, Univ. of Geneva, Switzerland, (1988).

[Ro88c]: M. Roche : *Differential algebraic systems and very stiff equations.* In: *Numerical treatment of differential equations.* Proceedings of the Fourth Seminar "Numdiff-4", Halle, Germany, 1987, K. Strehmel ed., Teubner-Texte zur Mathematik, Band 104, pp. 130-137, (1988).

[Ro89]: M. Roche : *Implicit Runge-Kutta for differential algebraic equations.* SIAM J. Numer. Anal., Vol. 26, pp. 963-975, (1989).

[SS88]: J.M. Sanz-Serna : *Runge-Kutta schemes for Hamiltonian systems.* BIT, Vol. 28, pp. 877-883, (1988).

[SS92]: J.M. Sanz-Serna : *Symplectic integrators for Hamiltonian problems: an overview.* Acta Numerica, Vol. 1, pp. 243-286, (1992).

[SSAb91]: J.M. Sanz-Serna & L. Abia : *Order conditions for canonical Runge-Kutta schemes.* SIAM J. Numer. Anal., Vol. 28, pp. 1081-1096, (1991).

[Sch90]: W. Schiehlen (ed.) : *Multibody systems handbook.* Springer-Verlag, Berlin, (1990).

[Sch93]: W. Schiehlen (ed.) : *Advanced multibody system dynamics, simulation and software tools.* Kluwer Academic Publishers, London, (1993).

[Sh86]: L.F. Shampine : *Conservation laws and the numerical solution of ODEs.* Comp. & Maths. with Appls., Vol. 12B, pp. 1287-1296, (1986).

[Si92]: B. Simeon : *Numerical integration of multibody systems by a projection technique.* Research Report TUM-M9201, Tech. Univ. Munich, Germany, (1992).

[Si93]: B. Simeon : *An extended descriptor form for the numerical integration of multibody systems.* Appl. Numer. Math., Vol. 13, pp. 209-221, (1993).

[SiFüRen91]: B. Simeon, C. Führer & P. Rentrop : *Introduction to differential-algebraic equations in vehicle system dynamics.* Surv. Math. Ind., Vol. 1, pp. 1-37, (1991).

[SkBiOk93]: R.D. Skeel, J.J. Biesiadecki & D. Okunbor : *Symplectic integration for macromolecular dynamics.* Proceedings of the International Conference Computation of Differential Equations and Dynamical Systems, World Scientific Publishing Co., (1993). To appear.

[Sun92]: Sun Geng : *Symplectic partitioned Runge-Kutta methods.* Report, (1992). To be published in J. Comput. Math.

[Sun93]: Sun Geng : *Construction of high order symplectic Runge-Kutta methods.* J. Comput. Math., Vol. 11, pp. 250-260, (1993).

[Sur89]: Y.B. Suris : *On canonicality of mappings resulting from Runge-Kutta-type methods during the integration of the system* $\ddot{x} = -\partial U/\partial x$. Zh. Vychisl. Mat. i Mat. Fiz., Vol. 29, pp. 202-211, (1989), in Russian; same as U.S.S.R. Comput. Maths. Phys., Vol. 29, pp. 138-144, (1989).

[Ta76]: F. Takens : *Constrained equations; a study of implicit differential equations and their discontinuous solutions.* In: *Structural stability, the theory of catastrophes, and applications in the sciences.* Battelle Seattle Research Center, U.S.A., 1975, P. Hilton ed., Lecture Notes in Mathematics, Vol. 525, Springer-Verlag, Berlin, pp. 143-234, (1976).

[YeHauPo92]: J. Yen, E.J. Haug & F.A. Potra : *Numerical methods for constrained equations of motion in mechanical systems dynamics.* Ph.D. thesis, Technical Report R-92, Univ. of Iowa, Iowa City, IA, U.S.A., (1992).

[Ye93]: J. Yen : *Constrained equations of motion in multibody dynamics as ODEs on manifolds.* SIAM J. Numer. Anal., Vol. 30, pp. 553-568, (1993).

[Yo90]: H. Yoshida : *Construction of higher order symplectic integrators.* Phys. Lett. A, Vol. 150, pp. 262-268, (1990).

Résumé de la thèse en français.

Introduction.

Le sujet de cette thèse traite de *l'Analyse Numérique des Equations Différentielles Algébriques (EDA)*. Les EDA consistent en des systèmes mixtes d'équations différentielles et d'équations algébriques (i.e., non linéaires) qui ne peuvent être exprimés sous forme *d'Equations Différentielles Ordinaires (EDO)*. Ce sujet qui est actuellement en plein essor, est essentiel pour les calculs scientifiques en physique, en chimie et dans les applications techniques. Les équations différentielles représentent un cadre mathématique naturel dans lequel se modélisent de nombreux problèmes dans les domaines susmentionnés. Bien souvent, en plus d'équations différentielles les modèles contiennent des équations implicites, en général purement algébriques (non linéaires), afin de tenir compte par exemple de lois de conservation, de contraintes géométriques et cinématiques, des lois de Kirchoff, etc. Des exemples typiques où de tels systèmes d'EDA surviennent sont les suivants :

- en dynamique des systèmes mécaniques;
- dans l'étude des systèmes Hamiltoniens munis de contraintes, par exemple en dynamique moléculaire;
- en analyse des circuits électriques;
- en cinétique des réactions chimiques;
- dans les équations provenant de la discrétisation d'équations aux dérivées partielles, par exemple en mécanique des fluides;
- en théorie du contrôle, par exemple en robotique;
- dans l'analyse des EDO raides (en anglais "stiff").

De façon plus précise on entend par EDA tout système d'equations de la forme

$$\begin{aligned} R_1(x, y_1, \ldots, y_n, y_1', \ldots, y_n') &= 0 \ , \\ &\vdots \\ R_m(x, y_1, \ldots, y_n, y_1', \ldots, y_n') &= 0 \ , \end{aligned} \tag{1}$$

ou plus succinctement

$$R(x, y, y') = 0 \ , \tag{1'}$$

et où $m \geq n$, x est la variable (unidimensionelle) d'intégration, $y = (y_1, \ldots, y_n)^T$, $y' = (y_1', \ldots, y_n')^T = dy/dx$ et $R_{y'}$ n'est pas de rang maximal n (singulière si $m = n$).

Quoique le domaine des EDO est traditionnel en Analyse Mathématique depuis l'époque de Newton, le traitement systématique des EDA n'a réellement pris son essor que depuis la dernière décennie. Les EDA diffèrent par plusieurs aspects des EDO et elles présentent de nouvelles difficultés tant sur le plan analytique que numérique.

Contrairement aux EDO qui sous des hypothèses raisonnables admettent une solution unique pour de quelconques valeurs initiales, il n'existe pas de théorie générale d'existence et d'unicité des solutions pour les EDA. Par exemple les EDA peuvent ne posséder des solutions que pour un sous-ensemble de valeurs initiales dites *consistantes*, peuvent avoir plusieurs solutions possibles, peuvent présenter des bifurcations, peuvent posséder des points d'impasse ou de rebroussement, ou peuvent carrément ne pas avoir de solution du tout. C'est la raison pour laquelle chaque type de problème nécessite une analyse particulière.

Les EDA peuvent être caractérisées par la notion *d'index* dont diverses définitions existent. *L'index de perturbation* est une mesure convenable de la sensibilité d'une solution à des perturbations dans les équations.

Définition 1. La $i^{\text{ème}}$ composante a un *index de perturbation* $\nu_{p,i}$ le long d'une solution u sur un intervalle borné I passant par $u(x_0)$ en x_0, si $\nu_{p,i}$ est le plus petit entier tel que pour toutes les fonctions $\widehat{u}(x)$ ayant un résidu

$$R(x, \widehat{u}(x), \widehat{u}'(x)) = \delta(x) \tag{2}$$

il existe sur I une estimation de la forme

$$|\widehat{u}_i(x) - u_i(x)| \leq C_i \left(\|\widehat{u}(x_0) - u(x_0)\| + \sup_{\zeta \in I} \| \int_{x_0}^{\zeta} \delta(\tau) d\tau \| + \sum_{j=0}^{\nu_{p,i}-1} \sup_{\zeta \in I} \|\delta^{(j)}(\zeta)\| \right) \tag{3}$$

pourvu que l'expression de droite de (3) soit suffisamment petite. Ici C_i est une constante qui ne dépend que de R et de la longueur de l'intervalle I. *L'index de perturbation* ν_p est défini par $\nu_p := \max_{i=1,\ldots,n} \nu_{p,i}$.

Contrairement aux EDA d'index $\nu_p = 0$ (tels que les EDO) ou 1, les EDA d'index $\nu_p \geq 2$, appelées *EDA d'index élevé*, sont des problèmes mal posés dans le sens où de petites perturbations peuvent être la source de changements importants des solutions. Le traitement numérique de tels problèmes mène souvent à de sérieuses difficultés qui peuvent néanmoins être surmontées en réduisant l'index du problème à 0 ou 1 à l'aide de différentes techniques. De nombreux problèmes courants dans les domaines susmentionnés sont formulés ou mènent à des EDA d'index élevé (voir plus loin).

Résoudre des EDO ou des EDA de façon analytique est une tâche généralement impossible. Ainsi des méthodes numériques ont été développées afin d'obtenir des approximations aux solutions de ces problèmes. De nos jours, avec l'avénement de la technologie des ordinateurs, les intérêts dans la modélisation, dans l'analyse, dans la simulation, et dans le contrôle de nombreux systèmes ont énormément augmenté, d'où un besoin accru en méthodes et logiciels sûrs et efficaces pour les EDA. Beaucoup de progrès ont été faits dans l'analyse théorique et numérique des EDA (voir les livres [BreCamPe89], [GrMä86], [HaLuRo89a] et [HaWa91]). Pour un rapide survol sur les EDA et les méthodes numériques voir [Pe89], [Rh91a], [GrHaMä91], [Pe92], [Mä92] et [HaJay93]. Dans certaines situations des EDA peuvent être réduites en EDO étant ainsi directement résolubles par des "résolveurs d'EDO" standards. Néanmoins, même dans ce cas il peut être en fait avantageux de travailler directement avec des EDA. Le développement systématique de méthodes numériques pour la résolution d'EDA ne date que d'un peu plus d'une dizaine d'années. La recherche dans ce domaine a débuté

par les travaux originaux de Gear [Ge71] et principalement de Petzold [Pe82]. De nombreuses méthodes numériques pour les EDO ont été spécialement adaptées pour les EDA. Elles comprennent principalement les méthodes linéaires multipas, les méthodes dites en anglais "one-leg", les méthodes linéairement implicites, les méthodes de Rosenbrock, les méthodes de Runge-Kutta et quelques méthodes d'extrapolation. A cause d'une certaine connexion entre les EDO raides (en anglais "stiff") et les EDA, les méthodes de Runge-Kutta dites en anglais "stiffly accurate" et les méthodes de différentiation rétrograde (en anglais "backward differentiation formula") sont d'un grand intérêt.

EDA semi-explicites d'index 3 sous forme de Hessenberg.

Cette thèse traite plus spécifiquement d'une classe d'EDA, dites *semi-explicites d'index 3 sous forme de Hessenberg* qui se formulent

$$y' = f(y,z)\ , \qquad z' = k(y,z,u)\ , \qquad 0 = g(y) \tag{4a,b,c}$$

et où

$$(g_y f_z k_u)(y,z,u) \qquad \text{est inversible} \tag{5}$$

dans un voisinage de la solution exacte. Nous nous restreignons à des problèmes à valeurs initiales. En dérivant deux fois la contrainte (4c) nous obtenons successivement

$$0 = \left(g_y f\right)(y,z)\ , \tag{4d}$$

$$0 = \left(g_{yy}(f,f) + g_y f_y f + g_y f_z k\right)(y,z,u)\ . \tag{4e}$$

Il s'ensuit que des valeurs initiales au problème (4a,b,c) ne peuvent pas être choisies arbitrairement, mais doivent satisfaire toutes les contraintes (4c,d,e). De telles valeurs sont dites *consistantes*. On peut vérifier que l'index de tels problèmes (4a,b,c) est bien égal à 3. A dessein nous considérons une solution $(y(x),z(x),u(x))$ de (4) sur un intervalle borné I passant par des valeurs initiales consistantes (y_0,z_0,u_0) en x_0. Nous considérons aussi des fonctions perturbées $(\widehat{y}(x),\widehat{z}(x),\widehat{u}(x))$ suffisamment proches de $(y(x),z(x),u(x))$ passant par $(\widehat{y}_0,\widehat{z}_0,\widehat{u}_0)$ en x_0 et satisfaisant

$$\begin{aligned} \widehat{y}'(x) &= f(\widehat{y}(x),\widehat{z}(x)) + \delta(x)\ , \\ \widehat{z}'(x) &= k(\widehat{y}(x),\widehat{z}(x),\widehat{u}(x)) + \mu(x)\ , \\ 0 &= g(\widehat{y}(x)) + \theta(x)\ . \end{aligned} \tag{6}$$

On montre alors l'estimation suivante

$$\begin{aligned} &\|\widehat{y}(x)-y(x)\| + \|\widehat{z}(x)-z(x)\| + \|\widehat{u}(x)-u(x)\| \le \\ &\qquad C\left(\|\widehat{y}_0-y_0\| + \|\widehat{z}_0-z_0\| + \sup_{\zeta\in I}\left(\|\delta(\zeta)\| + \|\delta'(\zeta)\| + \|\mu(\zeta)\| + \|\theta''(\zeta)\|\right)\right)\ . \end{aligned} \tag{7}$$

Divers exemples typiques où de telles équations surviennent sont donnés ci-dessous. D'autres exemples existent en théorie du contrôle.

Les systèmes mécaniques munis de contraintes.

Les équations de systèmes mécaniques munis de contraintes se déduisent du principe de Lagrange-Hamilton. Si $q = (q_1, \ldots, q_n)^T$ sont les n coordonnées généralisées d'un système mécanique soumis à m contraintes holonômes $g_1(q) = 0, \ldots, g_m(q) = 0$, alors les equations de mouvement sont données par

$$\begin{aligned} \dot{q} &= v \ , \\ M(q)\dot{v} &= f(q,v) - G^T(q)\lambda \ , \\ 0 &= g(q) \end{aligned} \tag{8}$$

où $G(q) := g_q(q)$ et en général $(GM^{-1}G^T)(q)$ est inversible.

Les systèmes Hamiltoniens munis de contraintes.

Les équations de systèmes Hamiltoniens munis de contraintes se déduisent du formalisme Hamiltonien. Si $q = (q_1, \ldots, q_n)^T$ sont les n coordonnées généralisées et $p = (p_1, \ldots, p_n)^T$ les moments généralisés d'un système Hamiltonien d'Hamiltonien $H(q,p)$ soumis à m contraintes holonômes $g_1(q) = 0, \ldots, g_m(q) = 0$, alors les équations de la dynamique du système sont données par

$$\begin{aligned} \dot{q} &= H_p^T(q,p) \ , \\ \dot{p} &= -H_q^T(q,p) - G^T(q)\lambda \ , \\ 0 &= g(q) \end{aligned} \tag{9}$$

où $G(q) := g_q(q)$ et en général $(GH_{pp}^T G^T)(q,p)$ est inversible. Le flot généré dans l'espace de phase de dimension $2(n-m)$

$$V = \left\{(q,p) \in \mathbb{R}^n \times \mathbb{R}^n \mid 0 = g(q) \ , \ 0 = G(q)H_p^T(q,p)\right\} \tag{10}$$

par le système (9) possède la propriété de *symplecticité*, c'est-à-dire que la forme différentielle

$$\omega^2 = \sum_{k=1}^{n} dq_k \wedge dp_k \qquad \text{est préservée.} \tag{11}$$

Des exemples de systèmes Hamiltoniens munis de contraintes sont donnés par les systèmes mécaniques ainsi qu'en dynamique moléculaire.

Les systèmes mécaniques raides (en anglais "stiff mechanical systems").

Les problèmes à perturbation singulière et singuliers (en anglais "singular singularly perturbed problems") forment une classe particulière d'EDO raides, contenant un petit paramètre $0 < \varepsilon \ll 1$

$$\begin{aligned} y' &= f(x,y,z) \ , \\ \varepsilon z' &= g(x,y,z) \ . \end{aligned} \tag{12}$$

Pour ces problèmes la matrice g_z est supposée posséder des valeurs propres nulles et vérifier

$$\langle g_z(x,y,z)w, w\rangle \leq -Const \cdot \|w\|^2 \qquad \text{avec} \quad Const \geq 0 \ . \tag{13}$$

L'analyse du cas limite $\varepsilon = 0$, le *problème réduit*, fournit une certaine information concernant le comportement des solutions de tels systèmes. Les systèmes mécaniques raides pour lesquels un fort potentiel $\frac{1}{\varepsilon^2}V(q)$ force la solution à être proche d'une certaine variété tombent dans cette catégorie de problèmes. Une formulation précise est

$$\begin{aligned} \dot{q} &= v \ , \\ M(q)\dot{v} &= f(q,v) - \frac{1}{\varepsilon^2}V_q^T(q) \end{aligned} \tag{14}$$

avec les hypothèses

- $M(q)$ est symétrique et définie positive;
- $V(q)$ atteint un minimum (local) sur une variété V de dimension m;
- dans un voisinage de V, $V(q)$ est fortement convexe le long de directions non tangentes à V.

Sous ces hypothèses on peut montrer que pour des valeurs initiales bien choisies les solutions lisses de (14) possèdent un développement en ε^2

$$\begin{aligned} q(t) &= q^0(t) + \varepsilon^2 q^1(t) + \ldots + \varepsilon^{2N} q^N(t) + \mathcal{O}(\varepsilon^{2N+2}) \ , \\ v(t) &= v^0(t) + \varepsilon^2 v^1(t) + \ldots + \varepsilon^{2N} v^N(t) + \mathcal{O}(\varepsilon^{2N+2}) \end{aligned} \tag{15}$$

où $(q^0(t), v^0(t))$ est la solution d'un problème de la forme (8) avec g s'annulant sur V (et seulement sur V), et $(q^k(t), v^k(t))$ sont solutions d'EDA d'index $2k+3$.

Développement en série de Taylor de la solution exacte.

L'analyse de méthodes numériques appliquées aux EDA du type (4) nécessite tout d'abord de connaître le développement en série de Taylor de la solution exacte. Celui-ci peut être obtenu à l'aide de structures arborescentes.

Théorème 1. *Le développement en série de Taylor de la solution exacte de* (4) *est donné par*

$$\begin{aligned} y(x+h) &= y(x) + \sum_{t \in DAT3_y} \alpha(t) \frac{h^{\varrho(t)}}{\varrho(t)!} F(t)(y(x), z(x), u(x)) \ , \\ z(x+h) &= z(x) + \sum_{v \in DAT3_z} \alpha(v) \frac{h^{\varrho(v)}}{\varrho(v)!} F(v)(y(x), z(x), u(x)) \ , \\ u(x+h) &= u(x) + \sum_{u \in DAT3_u} \alpha(u) \frac{h^{\varrho(u)}}{\varrho(u)!} F(u)(y(x), z(x), u(x)) \end{aligned} \tag{16}$$

où $DAT3_y$, $DAT3_z$ et $DAT3_u$ sont des ensembles d'arbres.

Dans ce contexte la théorie des B-séries d'Hairer et Wanner (voir [HaWa74] et [HaNøWa93, Section II.12]) est étendue aux EDA considérées (4), donnant naissance à ce que l'on appelle une théorie des *DA3-séries*.

Méthodes de Runge-Kutta partitionnées.

Le but de cette thèse est principalement d'étudier l'application de *méthodes de Runge-Kutta partitionnées (RKP)* aux EDA semi-explicites d'index 3 sous forme de Hessenberg $(4a, b, c)$.

Définition 2. Un pas d'une *méthode RKP à s étages* appliquée à $(4a, b, c)$ avec valeurs initiales (y_0, z_0, u_0) en x_0 est donné par

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i Y_i' , \qquad z_1 = z_0 + h \sum_{i=1}^{s} \widehat{b}_i Z_i'$$

où

$$Y_i' = f(Y_i, Z_i) , \qquad Z_i' = k(Y_i, Z_i, U_i) , \qquad 0 = g(Y_i) \tag{17}$$

et où les *étages internes* sont donnés par

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} Y_j' , \qquad Z_i = z_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} Z_j' .$$

Plusieurs définitions pour la composante numérique u sont possibles (omises ici). Nous nous intéressons plus particulièrement aux méthodes dites en anglais "stiffly accurate" qui satisfont

$$(S): \qquad a_{si} = b_i \qquad \text{pour} \quad i = 1, \ldots, s .$$

Nous supposons que la méthode RKP est basée sur une seule formule de quadrature, c'est-à-dire

$$b_i = \widehat{b}_i, \qquad c_i = \widehat{c}_i \qquad \text{pour} \quad i = 1, \ldots, s . \tag{18}$$

La construction de méthodes RKP d'ordre élevé est intimement reliée aux conditions suivantes, dites *conditions simplificatrices*

$$B(p): \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k} \qquad \text{pour} \quad k = 1, \ldots, p ;$$

$$C(q): \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \qquad \text{pour} \quad i = 1, \ldots, s , \quad k = 1, \ldots, q ;$$

$$\widehat{C}(\widehat{q}): \quad \sum_{j=1}^{s} \widehat{a}_{ij} c_j^{k-1} = \frac{c_i^k}{k} \qquad \text{pour} \quad i = 1, \ldots, s , \quad k = 1, \ldots, \widehat{q} ;$$

$$D(r): \quad \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{b_j}{k}\left(1 - c_j^k\right) \qquad \text{pour} \quad j = 1, \ldots, s , \quad k = 1, \ldots, r ;$$

$$\widehat{D}(\widehat{r}): \quad \sum_{i=1}^{s} b_i c_i^{k-1} \widehat{a}_{ij} = \frac{b_j}{k}\left(1 - c_j^k\right) \qquad \text{pour} \quad j = 1, \ldots, s , \quad k = 1, \ldots, \widehat{r} ;$$

$$C\widehat{C}(Q): \quad \sum_{j=1}^{s} \sum_{l=1}^{s} a_{ij} \widehat{a}_{jl} c_l^{k-2} = \frac{c_i^k}{k(k-1)} \qquad \text{pour} \quad i = 1, \ldots, s , \quad k = 2, \ldots, Q ;$$

$$D\widehat{D}(R): \quad \sum_{i=1}^{s} \sum_{j=1}^{s} b_i c_i^{k-2} a_{ij} \widehat{a}_{jl} = \frac{b_l}{k} - \frac{b_l c_l}{k-1} + \frac{b_l c_l^k}{k(k-1)} \qquad \text{pour} \quad l = 1, \ldots, s , \quad k = 2, \ldots, R ;$$

$$(S): \quad a_{si} = b_i \qquad \text{pour} \quad i = 1, \ldots, s .$$

Sous les conditions $C(2)$ et $C\widehat{C}(2)$ on peut démontrer l'existence et l'unicité de la solution numérique pourvu que la matrice $\overline{A} := A\widehat{A}$ soit inversible où $A := (a_{ij})_{i,j=1}^{s}$ et $\widehat{A} := (\widehat{a}_{ij})_{i,j=1}^{s}$. En général il est préférable de projeter la solution numérique obtenue sur les contraintes $(4c, d, e)$, comme suit

$$\begin{aligned} \widetilde{y}_1 &= y_1 - (f_z k_u)(y_1, z_1, u_1)\lambda\,, \qquad 0 = g(\widetilde{y}_1)\,, \\ \widetilde{z}_1 &= z_1 - k_u(y_1, z_1, u_1)\mu\,, \qquad 0 = (g_y f)(\widetilde{y}_1, \widetilde{z}_1)\,, \\ 0 &= \big(g_{yy}(f,f) + g_y f_y f + g_y f_z k\big)(\widetilde{y}_1, \widetilde{z}_1, \widetilde{u}_1)\,. \end{aligned} \tag{19}$$

Pour le développement en série de Taylor de la solution numérique une formule similaire à celle donnée au Théorème 1 peut être obtenue.

Théorème 2. *Sous les hypothèses d'existence et d'unicité de la solution numérique et pour des valeurs initiales (y_0, z_0, u_0) consistantes, le développement en série de Taylor de la solution numérique* (17) *est donné par*

$$\begin{aligned} y_1 &= y_0 + \sum_{t \in DAT3_y} \alpha(t) \frac{h^{\varrho(t)}}{\varrho(t)!} \gamma(t) \sum_{i=1}^{s} b_i \Phi_i(t) F(t)(y_0, z_0, u_0)\,, \\ z_1 &= z_0 + \sum_{v \in DAT3_z} \alpha(v) \frac{h^{\varrho(v)}}{\varrho(v)!} \gamma(v) \sum_{i=1}^{s} b_i \Phi_i(v) F(v)(y_0, z_0, u_0)\,. \end{aligned} \tag{20}$$

Pour la composante u_1 un développement similaire peut être obtenu dépendant de la définition choisie. Ainsi en comparant les résultats des Théorèmes 1 et 2 on obtient ainsi aisément les conditions d'ordre des méthodes RKP. A l'aide des conditions simplificatrices on peut alors montrer des estimations optimales concernant l'erreur locale de telles méthodes. Finalement des estimations concernant l'erreur globale peuvent être obtenues (voir Tables 1 et 2 plus loin). Pour des méthodes de Runge-Kutta pures, c'est-à-dire non partitionnées et non projetées, une démonstration d'une conjecture (voir [HaLuRo89a]) relative à la superconvergence des méthodes "stiffly accurate" est aussi donnée. C'est l'un des résultats principaux de cette thèse et il est le suivant.

Théorème 3. *Considérons le système $(4a, b, c)$ d'index 3 avec des valeurs initiales consistantes et une méthode de Runge-Kutta pure. Supposons que les coefficients de la méthode satisfont $B(p)$, $C(q)$ avec $q \geq 2$, $D(r)$, (S) et que la matrice A soit inversible. Alors pour $x_n - x_0 = nh \leq$ Const, l'erreur globale satisfait*

$$\begin{aligned} y_n - y(x_n) &= \mathcal{O}(h^{\min(p, 2q-2, q+r)}), \quad z_n - z(x_n) = \mathcal{O}(h^q), \\ P_z(x_n)\big(z_n - z(x_n)\big) &= \mathcal{O}(h^{\min(p, 2q-2, q+r)}), \quad u_n - u(x_n) = \mathcal{O}(h^{q-1}) \end{aligned} \tag{21}$$

où $P_z(x) := (I - k_u(g_y f_z k_u)^{-1} g_y f_z)(y(x), z(x), u(x))$.

Ce résultat a une application dans l'analyse de convergence des méthodes de Runge-Kutta appliquées à des systèmes mécaniques raides (en anglais "stiff mechanical systems") (voir [Lu93]).

Une classe importante de méthodes de Runge-Kutta partitionnées, pour lesquelles la matrice $\overline{A}=A\widehat{A}$ n'est pas inversible, est donnée par les conditions suivantes

$$
\begin{aligned}
& a_{1j}=0,\ a_{sj}=b_j,\ \widehat{a}_{js}=0,\ b_j=\widehat{b}_j,\ c_j=\widehat{c}_j \qquad \text{pour} \quad j=1,\dots,s\ , \\
H:\quad & rang(A\widehat{A})=s-1,\ b_s\neq 0\ , \\
& B(p),\ C(q),\ D(r),\ \widehat{C}(\widehat{q}),\ \widehat{D}(\widehat{r}),\ C\widehat{C}(Q),\ D\widehat{D}(R)\ .
\end{aligned}
$$

Les méthodes satisfaisant de plus les relations

$$
\begin{aligned}
& b_i=\widehat{b}_i \qquad \text{pour} \quad i=1,\dots,s\ , \\
& b_i\widehat{a}_{ij}+\widehat{b}_j a_{ji}-b_i\widehat{b}_j=0 \qquad \text{pour} \quad i=1,\dots,s\ , \quad j=1,\dots,s
\end{aligned}
\tag{22}
$$

sont d'une importance spéciale pour les systèmes Hamiltoniens munis de contraintes (9). De telles méthodes peuvent s'appliquer de façon à ce que toutes les contraintes soient satisfaites et ceci de la manière suivante

$$
\begin{aligned}
& Q_i=q_0+h\sum_{j=1}^{s}a_{ij}H_p^T(Q_j,P_j)\ , \quad P_i=p_0-h\sum_{j=1}^{s-1}\widehat{a}_{ij}\Big(H_q^T(Q_j,P_j)+G^T(Q_j)\Lambda_j\Big)\ , \\
& 0=g(Q_i)\ , \qquad q_1=Q_s\ , \\
& p_1=p_0-h\sum_{i=1}^{s-1}b_i\Big(H_q^T(Q_i,P_i)+G^T(Q_i)\Lambda_i\Big)-hb_s\Big(H_q^T(Q_s,P_s)+G^T(Q_s)\Lambda_s\Big)\ , \\
& 0=G(q_1)H_p^T(q_1,p_1)\ , \\
& \lambda_1=\left(GH_{pp}^TG^T\right)^{-1}\left(G_q(H_p^T,H_p^T)+GH_{pq}^TH_p^T-GH_{pp}^TH_q^T\right)(q_1,p_1)\ .
\end{aligned}
\tag{23}
$$

De plus de telles méthodes préservent aussi la symplecticité du flot, plus précisément on peut montrer qu'elles vérifient

$$
\sum_{k=1}^{n}dq_{1,k}\wedge dp_{1,k}=\sum_{k=1}^{n}dq_{0,k}\wedge dp_{0,k}\ . \tag{24}
$$

Comme précédemment on peut montrer l'existence et l'unicité de la solution numérique et l'on peut aussi obtenir des résultats optimaux concernant l'erreur locale et l'erreur globale de telles méthodes.

Finalement on résume les ordres de convergence pour différentes méthodes de Runge-Kutta (partitionnées) (et projetées) dans les Tables 1 et 2. On rappelle que l'ordre de convergence d'une méthode est égal à ν si l'erreur globale est uniformément bornée par $Const\cdot h^{\nu}$ sur des intervalles bornés et pour des pas h suffisamment petits. La Table 2 concerne la situation éminemment importante où la fonction k est linéaire en u. Tous ces résultats sont valides pour des pas non constants avec $h=\max_i h_i$.

Méthode	étages	ordre de convergence y	z	u
Lobatto IIIA-IIIB	$s \geq 2$	$2s-2$	$2s-2$	$2s-2$
Radau IIA	$s \geq 2$	$2s-2$	s	$s-1$
Radau IIA projeté	$s \geq 2$	$2s-2$	$2s-2$	$2s-2$
Lobatto IIIC	$s \geq 3$	$2s-4$	$s-1$	$s-2$
Lobatto IIIC projeté	$s \geq 3$	$2s-4$	$2s-4$	$2s-4$
Gauss	$s \geq 5$	s	$s-2$	$s-4$
Gauss projeté	$s \geq 2$	s	s	s
Radau IA	$s \geq 3$	$s-1$	$s-1$	$s-2$
Radau IA projeté	$s \geq 3$	$s-1$	$s-1$	$s-1$

Table 1. Ordres de convergence pour le problème (4)-(5).

Méthode	étages	ordre de convergence y	z	u
Lobatto IIIA-IIIB	$s \geq 2$	$2s-2$	$2s-2$	$2s-2$
Radau IIA	s	$2s-1$	s	$s-1$
Radau IIA projeté	s	$2s-1$	$2s-1$	$2s-1$
Lobatto IIIC	$s \geq 2$	$2s-3$	$s-1$	$s-2$
Lobatto IIIC projeté	$s \geq 2$	$2s-3$	$2s-3$	$2s-3$
Gauss	$s \geq 3$	s	$s-2$	$s-4$
Gauss projeté	s	s	s	s
Radau IA	$s \geq 2$	$s-1$	$s-1$	$s-2$
Radau IA projeté	$s \geq 2$	$s-1$	$s-1$	$s-1$

Table 2. Ordres de convergence pour le problème (4)-(5) avec k linéaire en u.

Organisation générale de la thèse.

Le but principal de cette thèse est d'étudier l'application de *méthodes de Runge-Kutta (projetées) partitionnées* à des EDA semi-explicites d'index 3 sous forme de Hessenberg $(4a, b, c)$ (voir (17)). Nous nous intéressons principalement aux méthodes dites en anglais *"stiffly accurate"* et nous nous restreignons à des problèmes à valeurs initiales. L'organisation générale de la thèse se présente ainsi :

- Au Chapitre I nous passons en revue quelques notions et résultats fondamentaux relatifs aux EDA et à leur traitement numérique. Après avoir décrit certains types d'EDA, nous discutons ensuite des concepts importants de solubilité et d'index. Puis nous donnons quelques exemples courants d'EDA ayant un index élevé et nous présentons quelques techniques possibles afin de reduire l'index d'un problème. Nous passons ensuite en revue quelques méthodes numériques utilisées pour la résolution des EDA. Finalement un bref survol des buts de la thèse et des principaux résultats de convergence est donné.

- Au Chapitre II nous donnons des résultats théoriques relatifs aux EDA semi-explicites d'index 3 sous forme de Hessenberg. Après avoir caractérisé l'ensemble des valeurs consistantes et l'index du problème, nous dérivons ensuite le développement en série de Taylor de la solution exacte à l'aide de structures arborescentes. Dans ce contexte la théorie des B-séries d'Hairer et Wanner est étendue aux EDA considérées, donnant naissance à ce que l'on appelle une théorie des $DA3$-séries.

- Le Chapitre III traite de l'application de méthodes de Runge-Kutta (projetées) partitionnées aux EDA semi-explicites d'index 3 sous forme de Hessenberg. Nous donnons des résultats concernant l'existence et l'unicité de la solution numérique, de l'influence de perturbations, de son erreur locale et de son erreur globale. Une courte discussion sur l'application d'itérations de Newton simplifiées au système d'équations non linéaires induit par la méthode numérique clôt ce chapitre.

- Les deux chapitres suivants sont similaires au Chapitre III additionnés de quelques exemples numériques. Au Chapitre IV nous nous restreignons à l'application directe de méthodes de Runge-Kutta pures aux EDA semi-explicites d'index 3 sous forme de Hessenberg. Une preuve d'une conjecture relative à la superconvergence de méthodes de Runge-Kutta dites en anglais "stiffly accurate" est donnée, ainsi qu'une application de ce résultat à l'analyse de convergence de ces méthodes aux systèmes mécaniques raides (en anglais "stiff mechanical systems"). Au Chapitre V nous traitons principalement de l'application d'une classe spéciale de méthodes de Runge-Kutta partitionnées aux systèmes Hamiltoniens munis de contraintes holonômes. Ces méthodes sont superconvergentes et préservent la structure symplectique du flot ainsi que toutes les contraintes sous-jacentes.