Spring 2011

# Runge-Kutta type methods for differential-algebraic equations in mechanics

Scott Joseph Small
*University of Iowa*

# RUNGE-KUTTA TYPE METHODS FOR DIFFERENTIAL-ALGEBRAIC EQUATIONS IN MECHANICS

by

Scott Joseph Small

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy degree in
Applied Mathematical and Computational Sciences
in the Graduate College of
The University of Iowa

May 2011

Thesis Supervisor: Professor Laurent O. Jay

## ABSTRACT

Differential-algebraic equations (DAEs) consist of mixed systems of ordinary differential equations (ODEs) coupled with linear or nonlinear equations. Such systems may be viewed as ODEs with integral curves lying in a manifold. DAEs appear frequently in applications such as classical mechanics and electrical circuits. This thesis concentrates on systems of index 2, originally index 3, and mixed index 2 and 3.

Fast and efficient numerical solvers for DAEs are highly desirable for finding solutions. We focus primarily on the class of Gauss-Lobatto SPARK methods. However, we also introduce an extension to methods proposed by Murua for solving index 2 systems to systems of mixed index 2 and 3. An analysis of these methods is also presented in this thesis. We examine the existence and uniqueness of the proposed numerical solutions, the influence of perturbations, and the local error and global convergence of the methods.

When applied to index 2 DAEs, SPARK methods are shown to be equivalent to a class of collocation type methods. When applied to originally index 3 and mixed index 2 and 3 DAEs, they are equivalent to a class of discontinuous collocation methods. Using these equivalences, $(s, s)$–Gauss-Lobatto SPARK methods can be shown to be superconvergent of order $2s$.

Symplectic SPARK methods applied to Hamiltonian systems with holonomic constraints preserve well the total energy of the system. This follows from a backward error analysis approach. SPARK methods and our proposed EMPRK methods are shown to be Lagrange-d'Alembert integrators.

This thesis also presents some numerical results for Gauss-Lobatto SPARK and EMPRK methods. A few problems from mechanics are considered.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

# RUNGE-KUTTA TYPE METHODS FOR DIFFERENTIAL-ALGEBRAIC EQUATIONS IN MECHANICS

by

Scott Joseph Small

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy degree in
Applied Mathematical and Computational Sciences
in the Graduate College of
The University of Iowa

May 2011

Thesis Supervisor: Professor Laurent O. Jay

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

---

PH.D. THESIS

---

This is to certify that the Ph.D. thesis of

Scott Joseph Small

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy degree
in Applied Mathematical and Computational Sciences
at the May 2011 graduation.

Thesis Committee: _____
Laurent O. Jay, Thesis Supervisor

_____
Rodica Curtu

_____
Luca Dieci

_____
Weimin Han

_____
David Stewart

## ACKNOWLEDGMENTS

I would like to thank many associates and colleagues for helping me complete my studies at the University of Iowa.

First and foremost, I would like to thank my advisor, Professor Laurent Jay. His insight and guidance have been of great help during my time in Iowa. His advice will benefit me the rest of my life.

I would also like to thank the members of my defense committee: Professor Rodica Curtu, Professor Luca Dieci, Professor Weimin Han, and Professor David Stewart, for taking the time to review my work. Their time and input have been greatly appreciated.

I thank Professor Kendall Atkinson for all his advice. I give my gratitude to Professor Keith Stroyan for showing me his approaches to teaching and to working with students.

I must also thank my many colleagues for their support and friendships during my time at Iowa. In particular, I thank Joseph Eichholz and Stephen Welch for keeping me sane all these years. I wish all my colleagues the best. I also give my thanks to Karen Staats. May her bag of jolly ranchers always remain full.

Finally, I give my thanks to my family, who were always there to support me in mind and spirit.

## ABSTRACT

Differential-algebraic equations (DAEs) consist of mixed systems of ordinary differential equations (ODEs) coupled with linear or nonlinear equations. Such systems may be viewed as ODEs with integral curves lying in a manifold. DAEs appear frequently in applications such as classical mechanics and electrical circuits. This thesis concentrates on systems of index 2, originally index 3, and mixed index 2 and 3.

Fast and efficient numerical solvers for DAEs are highly desirable for finding solutions. We focus primarily on the class of Gauss-Lobatto SPARK methods. However, we also introduce an extension to methods proposed by Murua for solving index 2 systems to systems of mixed index 2 and 3. An analysis of these methods is also presented in this thesis. We examine the existence and uniqueness of the proposed numerical solutions, the influence of perturbations, and the local error and global convergence of the methods.

When applied to index 2 DAEs, SPARK methods are shown to be equivalent to a class of collocation type methods. When applied to originally index 3 and mixed index 2 and 3 DAEs, they are equivalent to a class of discontinuous collocation methods. Using these equivalences, $(s, s)$–Gauss-Lobatto SPARK methods can be shown to be superconvergent of order $2s$.

Symplectic SPARK methods applied to Hamiltonian systems with holonomic constraints preserve well the total energy of the system. This follows from a backward error analysis approach. SPARK methods and our proposed EMPRK methods are shown to be Lagrange-d'Alembert integrators.

This thesis also presents some numerical results for Gauss-Lobatto SPARK and EMPRK methods. A few problems from mechanics are considered.

## TABLE OF CONTENTS

## LIST OF TABLES

Table

## LIST OF FIGURES

Figure

# CHAPTER 1
# INTRODUCTION

## 1.1 Introduction to DAEs

Differential equations are used in real-world applications as models to problems arising in the physical sciences. Solving these equations analytically is generally out of the question; analytic solutions may be difficult or impossible to obtain. Therefore, fast and efficient numerical solvers are desirable.

When a system of ordinary differential equations (ODEs) is coupled with linear/nonlinear equations, i.e. equations with no derivatives, the resulting system of equations are called *differential-algebraic equations (DAEs)*. The linear/nonlinear equations constrain the flow of the differential equations to a manifold. DAEs arise in many applications, including classical mechanics, control systems, electrical circuits, and partial differential equations (PDEs). See, for example, [1], [2], [8], [9], [10], [20].

A first order DAE is a system of equations of the form

$$F(t, x, \dot{x}) = 0 \tag{1.1}$$

where $t \in \mathbb{R}$ is the independent variable (generally referred to as the "time" variable), $x(t) \in \mathbb{R}^{n_x}$ is the unknown function, and $\dot{x}(t) := \frac{dx}{dt}(t)$. The function

$$F : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^m$$

is assumed to be differentiable. If the Jacobian matrix $\frac{\partial F}{\partial \dot{x}}$ is invertible, than the system (1.1) can be expressed as a system of ODEs. The system (1.1) is a very general form of DAEs. We consider in this thesis only initial value problems, i.e., systems of the form (1.1) subject to the additional initial condition $x(t_0) = x_0$ for some initial time $t_0 \in \mathbb{R}$ and value $x_0 \in \mathbb{R}^{n_x}$.

One of the simplest examples of DAEs is the system

$$\dot{y}_1 = m y_3 \tag{1.2a}$$

$$\dot{y}_2 = m y_4 \tag{1.2b}$$

$$\dot{y}_3 = -2 y_1 \lambda \tag{1.2c}$$

$$\dot{y}_4 = m\gamma - 2 y_2 \lambda \tag{1.2d}$$

$$0 = \frac{1}{2}\left(y_1^2 + y_2^2 - \ell^2\right), \tag{1.2e}$$

where the unknown vector is given by

$$x(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \\ \lambda(t) \end{bmatrix} \in \mathbb{R}^5.$$

This system describes the motion of a pendulum in Cartesian coordinates. The variables $y_1$ and $y_2$ are the position coordinates of the pendulum, and $y_3$ and $y_4$ are the velocity coordinates. The constants $m$, $\gamma$, and $\ell$ are the mass of the pendulum, acceleration due to gravity, and length of the pendulum, respectively. The variables $y_1$, $y_2$, $y_3$, $y_4$ are referred to as the *differential variables*, while $\lambda$ is referred to as the *algebraic variable*. The condition (1.2e) is called the *constraint* or *algebraic equation*. Taking the derivative of the constraint (1.2e) gives

$$0 = y_1 \dot{y}_1 + y_2 \dot{y}_2 = y_1 y_3 + y_2 y_4. \tag{1.3}$$

Taking the derivative again results in

$$0 = m y_3^2 - 2 y_1^2 \lambda + m y_4^2 + m\gamma y_2 - 2 y_2^2 \lambda. \tag{1.4}$$

This allows us to solve for $\lambda$ in terms of $y_1$, $y_2$, $y_3$, and $y_4$. Substituting this into

(1.2), we arrive at the *underlying ODE*

$$\dot{y}_1 = my_3 \tag{1.5a}$$

$$\dot{y}_2 = my_4 \tag{1.5b}$$

$$\dot{y}_3 = -m\frac{y_1}{\ell^2}(y_3^2 + y_4^2 + \gamma y_2) \tag{1.5c}$$

$$\dot{y}_4 = m\gamma - m\frac{y_2}{\ell^2}(y_3^2 + y_4^2 + \gamma y_2). \tag{1.5d}$$

There are three classes of systems of DAEs that will be considered in depth in this thesis. The first class is of the form

$$\dot{y} = f(t, y, \psi) \tag{1.6a}$$

$$0 = k(t, y). \tag{1.6b}$$

We assume that the matrix $k_y(t, y)f_\psi(t, y, \psi)$ is invertible. This system is referred to as an *index 2 DAEs in Hessenberg form*. Here and for the entirety of this thesis, we consider the differentiation index, i.e., it is the total number of derivatives required of the constraints to turn the system into a system of ODEs. The class of *index 3 DAEs in Hessenberg form* is given by

$$\dot{y} = v(t, y, z) \tag{1.7a}$$

$$\dot{z} = f(t, y, z, \lambda) \tag{1.7b}$$

$$0 = g(t, y), \tag{1.7c}$$

where the matrix $g_y(t, y)v_z(t, y, z)f_\lambda(t, y, z, \lambda)$ is assumed invertible. The invertibility of the matrices for both (1.6) and (1.7) allow the systems of DAEs to be expressed as an underlying system of ODEs. This allows for the existence and uniqueness of a solution to initial value problems, provided the initial conditions are consistent. Notice that a system of the form

$$\dot{y} = v(t, y, z)$$

$$\dot{z} = f(t, y, z, \lambda)$$
$$0 = g_t(t, y) + g_y(t, y)v(t, y, z),$$

appears to be of index 2. However, the constraints may be expressed as the time derivative of the condition $0 = g(t, y)$, which is index 3. Because the solution of (1.7) satisfies (1.7c), it must also satisfy the total derivative of the constraints as well. This is sometimes called the *hidden constraints*. Throughout this thesis, for index 3 problems, we include the hidden constraints into the system:

$$\dot{y} = v(t, y, z) \tag{1.8a}$$
$$\dot{z} = f(t, y, z, \lambda) \tag{1.8b}$$
$$0 = g(t, y), \tag{1.8c}$$
$$0 = g_t(t, y) + g_y(t, y)v(t, y, z). \tag{1.8d}$$

We refer to this overdetermined system of DAEs as being *originally index 3*.

The final general class of overdetermined DAEs is

$$\dot{y} = v(t, y, z) \tag{1.9a}$$
$$\dot{z} = f(t, y, z, \lambda, \psi) \tag{1.9b}$$
$$0 = g(t, y) \tag{1.9c}$$
$$0 = g_t(t, y) + g_y(t, y)v(t, y, z) \tag{1.9d}$$
$$0 = k(t, y, z). \tag{1.9e}$$

The matrices

$$g_y(t, y)v_z(t, y, z)f_\lambda(t, y, \lambda, \psi) \tag{1.10a}$$
$$\begin{bmatrix} g_y(t, y)v_z(t, y, z)f_\lambda(t, y, z, \lambda, \psi) & g_y(t, y)v_z(t, y, z)f_\psi(t, y, z, \lambda, \psi) \\ k_z(t, y, z)f_\lambda(t, y, z, \lambda, \psi) & k_z(t, y, z)f_\psi(t, y, z, \lambda, \psi) \end{bmatrix} \tag{1.10b}$$

are assumed invertible. The invertibility of the matrices (1.10) allows (1.9) to be

expressed as a system of ODEs. This type of system we refer to as *mixed index 2 and index 3 DAEs*. The solution to each of these systems lies in the manifold determined by the constraints.

For each of the systems (1.6), (1.7), and (1.9), we assume that the initial conditions are *consistent*, i.e., that the initial conditions satisfy the constraints. More information regarding DAEs in general can be found in [2], [7], [8], [10], [22].

## 1.2 Summary of Background Material

In this section, we summarize important background material for this thesis. More detailed information can be found at the sources cited.

### 1.2.1 Runge-Kutta Type Methods for Systems of DAEs

Many numerical methods have been developed for DAEs. Historically, DAEs were reduced to underlying ODEs and solved with a standard Runge-Kutta (RK) or multistep method. Methods where the constraints are differentiated to obtain lower index constraints are called *index reduction methods*. However, solving the underlying ODEs can have negative consequences, see [10], [22], and [23]. Some of these include:

- Index reduction methods can be very computational. Solving for the algebraic variables requires computing the inverse of a matrix, such as (1.10b). With standard numerical ODE methods, this inverse must be computed at every time step. Although the appropriate matrices are assumed invertible, they could be ill-conditioned.

- For index 2, originally index 3, and mixed index 2 and 3 DAEs, the constraints are invariants to the flow of the system. However, a numerical solution to the underlying ODE generally does not satisfy the original constraints. This can be true despite consistent initial conditions.

More sophisticated methods have since arisen, allowing DAEs to be solved directly without the need for finding the underlying ODEs.

RK methods for solving ODEs are well known and have enjoyed much success. Given an initial value problem

$$\dot{y} = f(t, y) \tag{1.11a}$$

$$y(t_0) = y_0, \tag{1.11b}$$

an *s-stage Runge-Kutta method* with step size $h$ is a system of equations of the form

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j), \quad i = 1, \ldots, s \tag{1.12a}$$

$$y_1 = y_0 + h \sum_{j=1}^{s} b_j f(t_0 + c_j h, Y_j). \tag{1.12b}$$

The value $y_1$ is used as an approximate solution to $y(t_1)$ for $t_1 := t_0 + h$. The *internal stages* $Y_i$ are generally determined implicitly by the nonlinear system of equations (1.12a). The real-valued coefficients $a_{ij}$, $b_j$, $c_j$ can be seen as the nodes and weights of quadrature formulas when applied to the integrated form of (1.11). Some noteworthy choices for the coefficients are based upon the Gauss, Radau, and Lobatto quadratures. See [8], [9], [10] for more information regarding RK methods for ODEs.

RK methods have also been extended to DAEs. The most straightforward approach is to simply require the internal stages to satisfy the constraints. An example of the so called *standard RK methods* for index 2 DAEs (1.6) is given by

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j, \Psi_j), \quad i = 1, \ldots, s \tag{1.13a}$$

$$y_1 = y_0 + h \sum_{j=1}^{s} b_j f(t_0 + c_j h, Y_j, \Psi_j) \tag{1.13b}$$

$$0 = k(t_0 + c_i h, Y_i), \quad i = 1, \ldots, s. \tag{1.13c}$$

This method can be modified to handle originally index 3 and mixed index 2 and 3 DAEs. An analysis of these methods can be found in [7], [8], and [10]. The numerical solution of the standard RK methods for DAEs do not, in general, satisfy the constraints. In other words, we do not necessarily have

$$0 = k(t_1, y_1).$$

One approach to overcome this difficulty is to consider *stiffly accurate methods*. These are methods in which the coefficients are chosen to satisfy $a_{ij} = b_i$ and $c_j = 1$ for all $i = 1, \ldots, s$ and some $j$, thereby forcing $y_1 = Y_j$ and hence $0 = k(t_1, y_1)$ through (1.13c). For orginially index 3 and mixed index 2 and 3 DAEs, the standard methods, however, do not in general satisfy the hidden constraints

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1). \tag{1.14}$$

Other approaches have been developed beyond the standard RK methods to insure that the constraints are satisfied by the numerical solution. Two notable examples are Murua's partitioned Runge-Kutta (MPRK) methods proposed in (see [21]) and specialized partitioned additive Runge-Kutta (SPARK) methods proposed by Jay (see [16], [15]). MPRK methods introduce additional internal stages to handle the constraints for index 2 DAEs. SPARK methods also introduce some additional internal stages for originally index 3 DAEs, as well as a linear combination of the constraints to handle index 2 DAEs. These methods will be discussed in much detail throughout this thesis.

### 1.2.2 Numerical Methods in Classical Mechanics

Classical mechanics is a very well known area. Many sources with information concerning classical mechanics exist; see [1], [19], [20], [26]. The area is related to

variational calculus, see [4]. The Lagrangian equations

$$\dot{q} = v \tag{1.15a}$$

$$\frac{d}{dt}\nabla_v L(t,q,v) = \nabla_q L(t,q,v) \tag{1.15b}$$

and the Hamiltonian equations

$$\dot{q} = \nabla_p H(t,q,p) \tag{1.16a}$$

$$\dot{p} = -\nabla_q H(t,q,p) \tag{1.16b}$$

describe the evolution of an unconstrained system. The values $q(t)$, $v(t)$, and $p(t)$ represent, respectively, the generalized coordinates, velocities, and momenta of the system. In (1.16), $p$ and $v$ are related by the definition

$$p := \nabla_v L(t,q,v)$$

and the *Hamiltonian* $H$ and the *Lagrangian* $L$ are related by the Legendre transformation of $L$

$$H(t,q,p) = v^T p - L(t,q,v).$$

The systems (1.15) and (1.16) are derived from Hamilton's variational principle.

**Definition 1.2.1.** *(Hamilton's Principle) The evolution of a system $q(t)$ with fixed endpoints $q(t_0) = q_0$ and $q(t_N) = q_N$ is the minimizer of the action integral*

$$A(q(t)) := \int_{t_0}^{t_N} L(t,q(t),v(t))dt. \tag{1.17}$$

For mechanical systems, the Hamiltonian $H$ is given for example as

$$H(t,q,p) = T(t,q,p) + U(t,q),$$

with $T$ the kinetic energy and $U$ the potential energy. The Hamiltonian thus represents the total energy of the system. If $H$ is independent of time, i.e. $H(t,q,p) = H(q,p)$, then the flow of (1.16) has $H$ as a first integral.

Forced Lagrangian systems take the form

$$\dot{q} = v \tag{1.18a}$$

$$\frac{d}{dt}\nabla_v L(t,q,v) = \nabla_q L(t,q,v) + f_L(t,q,v). \tag{1.18b}$$

Forced Lagrangian systems are derived from the forced Lagrange-d'Alembert principle

$$0 = \delta A(q)(\delta q) + \int_{t_0}^{t_N} f_L(t,q(t),\dot{q}(t))^T \delta q(t) dt, \tag{1.19}$$

where $\delta A(q)$ is the first variation of the action integral. This principle can be found in [1] and [20].

Numerical methods can be formed from analogous discrete principles by discretizing the action integral. These methods are referred to as *variational integrators*. Information about variational integrators can be found in [1], [8], [18], [20]. The discrete Hamilton's principle requires extremizing the discrete action

$$\sum_{i=0}^{N-1} L_d(q_i, q_{i+1}) \tag{1.20}$$

where the *discrete Lagrangian* $L_d$ is some approximation to the action integral

$$L_d(q_i, q_{i+1}) \approx \int_{t_i}^{t_{i+1}} L(t,q(t),v(t)) dt. \tag{1.21}$$

The value $q_i$ is an approximation at a time $t_i$. As an example for a discrete Lagrangian, a simple quadrature method can be applied to approximate the action integral. It can be seen that extremizing (1.20) is equivalent to the *discrete Euler-Lagrange* equations

$$D_2 L_d(q_{i-1}, q_i) + D_1 L_d(q_i, q_{i+1}) = 0, \quad i = 1, \ldots, N-1. \tag{1.22}$$

This results in a system of equations for each $q_i$.

Problems arising in mechanics often include restrictions on the positions and velocities of the system. Such considerations require the use of DAEs as opposed to

usual ODEs. Examples of this would include the swinging of a pendulum in Cartesian coordinates, the motion of an ice skate, or the rolling of a ball. Constraints on the position of a system are referred to as *holonomic constraints*, while restrictions on the velocities of a system are referred to as *nonholonomic constraints*, provided they cannot be integrated as holonomic constraints. In the language of DAEs, holonomic constraints correspond to index 3 constraints, and nonholonomic to index 2 constraints. Mechanical constraints independent of time are called *scleronomic*, while constraints dependent upon time are called *rheonomic*.

The Lagrange-d'Alembert principle gives a generalization of Hamilton's principle for systems with nonholonomic constraints. We assume *ideal* nonholonomic constraints, i.e., $k(t,q,v) = K(t,q)v$.

**Definition 1.2.2.** *(Lagrange-d'Alembert Principle) The evolution of a system $q(t)$ with fixed endpoints $q(t_0) = q_0$ and $q(t_N) = q_N$ satisfies*

$$\delta A(q)(\delta q) = 0, \tag{1.23}$$

*where the virtual displacements $\delta q(t)$ satisfy*

$$0 = K(t, q(t))\delta q(t). \tag{1.24}$$

However, as proposed in [11], the forced Lagrange-d'Alembert principle (1.19) can also be used, assuming that the matrix

$$k_v(t,q,v)L_{vv}(t,q,v)^{-1}k_v(t,q,v)^T$$

is invertible and by taking

$$f_L(t,q,v) = -k_v(t,q,v)^T\psi(t,q,v).$$

This is because the multiplier $\psi$ can be solved for in terms of $q$ and $v$. We must

also require that

$$0 = k(t, q, v).$$

More generally, under ideal holonomic and nonholonomic constraints, the forced Lagrange-d'Alembert principle can be used by taking

$$f_L(t, q, v) = -g_q(t, q)^T \lambda(t, q, v) - k_v(t, q, v)^T \psi(t, q, v)$$

For constrained systems, the Lagrange-d'Alembert principle gives rise to Lagrange's equations of motion:

$$\dot{q} = v \tag{1.25a}$$

$$\frac{d}{dt}\nabla_v L(t, q, v) = \nabla_q L(t, q, v) - g_q(t, q)^T \lambda - k_v(t, q, v)^T \psi \tag{1.25b}$$

$$0 = g(t, q) \tag{1.25c}$$

$$0 = g_t(t, q) + g_q(t, q)v \tag{1.25d}$$

$$0 = k(t, q, v). \tag{1.25e}$$

Hamilton's equations can be expressed similarly as

$$\dot{q} = \nabla_p H(t, q, p) \tag{1.26a}$$

$$\dot{p} = -\nabla_q H(t, q, p) - g_q(t, q)^T \lambda - k_p(t, q, p)^T \psi \tag{1.26b}$$

$$0 = g(t, q) \tag{1.26c}$$

$$0 = g_t(t, q) + g_q(t, q)\nabla_p H(t, q, p) \tag{1.26d}$$

$$0 = k(t, q, p). \tag{1.26e}$$

### 1.2.3 Symplectic Transformations

A symplectic transformation is a mapping that preserves a 2-form. For instance, consider a parallelogram in $\mathbb{R}^{2n}$ given by two vectors

$$\alpha = \begin{pmatrix} \alpha^q \\ \alpha^p \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta^q \\ \beta^p \end{pmatrix}$$

with $\alpha^q$, $\alpha^p$, $\beta^q$, $\beta^p$ in $\mathbb{R}^n$. Then the sum of the oriented areas by projecting onto the coordinate planes $(q_i, p_i)$ can be given by

$$\omega(\alpha, \beta) := \sum_{i=1}^{n} \det \begin{pmatrix} \alpha_i^q & \beta_i^q \\ \alpha_i^p & \beta_i^p \end{pmatrix} = \sum_{i=1}^{n} (\alpha_i^q \beta_i^p - \alpha_i^p \beta_i^q).$$

A differentiable map $\varphi$ is symplectic if it preserves the 2-form $\omega$ in the sense that

$$\omega \left( \frac{\partial \varphi}{\partial (q,p)}(q,p)\alpha, \frac{\partial \varphi}{\partial (q,p)}(q,p)\beta \right) = \omega(\alpha, \beta).$$

More information about symplectic transformations can be found in [8], [9], [19]. The flow of Hamiltonian systems, either unconstrained or with holonomic constraints, is a symplectic mapping. In addition, one step methods $y_1 = \Phi_h(y_0)$ applied to these Hamiltonian systems can also be symplectic. For unconstrained problems, Runge-Kutta methods with coefficients satisfying

$$b_i a_{ij} + b_j a_{ji} = b_i b_j, \quad i, j = 1, \ldots, s$$

are symplectic (see [8], [9]). Lobatto IIIA-IIIB RK methods for originally index 3 problems can be shown to be symplectic (see [10], [13]). The Gauss-Lobatto SPARK methods is another class of symplectic methods for originally index 3 systems (see [16]).

Related to symplectic mappings and Hamiltonian systems is the so called

Hamilton-Jacobi equation

$$0 = \frac{\partial S}{\partial t}(t, x, y) - H(x + \nabla_y S(t, x, y), y). \tag{1.27}$$

This is a partial differential equation, where the function

$$S : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

is the unknown, and

$$H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

is given and sufficiently differentiable. It can be shown (see for example, [6] or [8, Theorem VI.5.6]) that with the added initial condition

$$0 = S(0, x, y),$$

the mapping $\varphi_t(q, p) := (Q(t), P(t))$ defined implicitly by

$$p = P(t) + \nabla_q S(t, q, P(t)) \tag{1.28a}$$

$$Q(t) = q + \nabla_p S(t, q, P(t)), \tag{1.28b}$$

with $S$ the solution to (1.27), is the flow of the Hamiltonian system with Hamiltonian function $H$. The function $S$ in (1.28) is an example of a generating function of type I (see [8]). From the theory of generating functions, the flow map $\varphi_t(q, p)$ is symplectic. This provides a proof that the flow of a Hamiltonian system is symplectic.

### 1.2.4 Backward Error Analysis

Backward error analysis for DAEs is performed to demonstrate properties of a numerical solution. In the context of ODEs, backward error analysis has been discussed extensively. See, for example, [5], [8], [9], [25]. Consider an autonomous

system of ODEs

$$\dot{y} = f(y) \in \mathbb{R}^n \tag{1.29}$$

and a numerical solution with a discrete flow $\Phi_h : \mathbb{R}^n \to \mathbb{R}^n$. The goal is to find another system of ODEs

$$\dot{\widetilde{y}} = \widetilde{f}(\widetilde{y}, h) \in \mathbb{R}^n \tag{1.30}$$

where the numerical solution to (1.29) is the solution to (1.30) for a fixed stepsize $h$. The system (1.30) is the so called *modified ODEs*. This can be done by expanding in a series the solution $\widetilde{y}(t+h)$ of the modified equation and the numerical solution of (1.29) and matching like powers of $h$.

If system (1.29) is a Hamiltonian system, it has been shown in, for example, [5], [8], [9], that symplectic numerical methods produce a modified equation that is also a Hamiltonian system. This explains the excellent long term energy preservation observed by such methods.

An analogous approach for DAEs can also be performed. Given a differential equation $\dot{y} = f(y)$ on a manifold $\mathcal{M}$ and a numerical solution, the modified equation can be shown to remain in $\mathcal{M}$. See [8, Theorem IX.5.1]. For Hamiltonian systems with holonomic constraints, Hairer shows in [6] that a certain class of symplectic methods produces a modified Hamiltonian system.

## 1.3 Thesis Overview

This thesis contains original results regarding the numerical solutions of index 2, originally index 3, and mixed index 2 and 3 DAEs by Runge-Kutta type methods. We focus primarily on SPARK methods. However, we also consider an extension to the PRK methods proposed in [21]. This thesis consists of nine chapters concerning various aspects of these methods.

Chapter 2 concerns theoretical aspects of SPARK methods applied to index 2 DAEs. Analysis of these methods is presented in [15] using trees. We present an

alternative proof deriving the local error for the Gauss SPARK methods by showing their equivalence to a class of collocation type methods. The methods employed in this proof are useful for the local error of the SPARK methods for mixed index 2 and 3 problems used later on.

Chapter 3 presents theoretical results for SPARK methods applied to originally index 3 DAEs. Although the theorems of this chapter were originally presented in [16], the derivation of the local order relies upon a technical detail that was found to be incorrect. We correct this oversight by showing that the Gauss-Lobatto SPARK methods are equivalent to a class of discontinuous collocation methods and are able to confirm the original results.

Chapter 4 shows that the symplecticity of SPARK methods for originally index 3 DAEs gives rise to good energy preservation of Hamiltonian systems by using backward error analysis techniques.

Chapter 5 combines the results for SPARK methods for mixed index 2 and 3 problems. Gauss-Lobatto SPARK methods for DAEs with mixed index 2 and 3 constraints are presented and analyzed. The results of this chapter generalize the results of Chapters 2 and 3.

Chapter 6 presents an extension for the MPRK methods for systems with mixed index 2 and 3 constraints. Some analysis of the extended method is also presented. The analysis is similar to that of the SPARK methods. However, because of the way in which the constraints are handled, the technical details are different.

Chapter 7 concerns Lagrange-d'Alembert integrators. SPARK, MPRK, and our proposed extended MPRK (or EMPRK) methods with Gauss-Lobatto coefficients are shown to be Lagrange-d'Alembert integrators when applied to problems in mechanics.

Chapter 8 gives numerical results concerning the SPARK methods and the EMPRK methods. Problems arising from classical mechanics are considered.

Finally, Chapter 9 gives the conclusion of this thesis. Future work is also presented.

# CHAPTER 2
# SPARK METHODS FOR INDEX 2 DAES

## 2.1 Introduction

In this chapter, we focus on SPARK methods for index 2 problems. The local error for the Gauss methods was determined by Jay in [15]. An alternative proof of this result will provide insight for the derivation of the local order of the Gauss-Lobatto SPARK methods for mixed index 2 and 3 problems.

We consider problems of the form

$$\dot{y} = f(t, y, \psi) \tag{2.1a}$$

$$0 = k(t, y) \tag{2.1b}$$

where $f : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_k} \to \mathbb{R}^{n_y}$ and $k : \mathbb{R} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_k}$. We also assume that the matrix

$$k_y(t, y) f_\psi(t, y, \psi) \tag{2.2}$$

is invertible. The system

$$\frac{d}{dt} p(t, y) = f(t, y, \psi) \tag{2.3a}$$

$$0 = k(t, y) \tag{2.3b}$$

is a generalization of (2.1). Both Hamiltonian and Lagrangian systems can be expressed in this form. To insure existence and uniqueness of a solution, the following matrices are assumed invertible:

$$p_y(t, y) \tag{2.4a}$$

$$k_y(t, y) p_y(t, y)^{-1} f_\psi(t, y, \psi). \tag{2.4b}$$

Under these assumptions, differentiating the left side of (2.3a) gives

$$\dot{y} = p_y(t,y)^{-1}\left(f(t,y,\psi) - p_t(t,y)\right). \tag{2.5}$$

Taking the derivative of (2.3b), and substituting in (2.5), we arrive at

$$0 = k_t(t,y) + k_y(t,y)p_y(t,y)^{-1}(f(t,y,\psi) - p_t(t,y))$$

which determines uniquely the term $\psi$ by (2.4b) and the implicit function theorem.

Following [16], we define the new variables

$$p := p(t,y).$$

By (2.4a) we can express $y$ as a function of $t$ and $p$. Defining

$$F(t,p,\psi) := f(t,y(t,p),\psi), \quad K(t,p) := k(t,y(t,p)),$$

the system (2.3) can be expressed as

$$\dot{p} = F(t,p,\psi) \tag{2.6a}$$

$$0 = K(t,p). \tag{2.6b}$$

Thus, the system (2.3) can be equivalently expressed in the form (2.1). For the analysis presented in this chapter, we consider systems with $p(t,y) = y$, but the results are also valid in the more general case (2.3).

## 2.2 SPARK Methods

We introduce here SPARK methods applied to problems with index 2 constraints.

**Definition 2.2.1.** *One step of a s-stage specialized partitioned additive Runge-Kutta (SPARK) method applied to the system (2.1) with stepsize h starting at* $y_0$ *at*

*time $t_0$ is given by the solution of the nonlinear equations*

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j, \Psi_j), \quad i = 1, \ldots, s \tag{2.7a}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j f(t_0 + c_j h, Y_j, \Psi_j) \tag{2.7b}$$

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j) + \omega_{i,s+1} k(t_1, y_1), \quad i = 1, \ldots, s, \tag{2.7c}$$

*with $t_1 := t_0 + h$. The coefficients $\omega_{ij}$ are from a matrix $\widetilde{\Omega}_0$ with*

$$\widetilde{\Omega}_0 := \begin{bmatrix} 0_s^T & 1 \\ b^T & 0 \\ b^T C & 0 \\ \vdots & \vdots \\ b^T C^{s-2} & 0 \end{bmatrix} \in \mathbb{R}^{s\times(s+1)}.$$

Because of the definition of the matrix $\widetilde{\Omega}_0$, the constraints can be equivalently expressed as

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j), \quad i = 2, \ldots, s$$
$$0 = k(t_1, y_1).$$

We will also make use of the coefficient matrix

$$\alpha := \begin{bmatrix} A \\ b^T \end{bmatrix} \in \mathbb{R}^{(s+1)\times s}.$$

The SPARK methods (2.7) will be assumed to satisfy

$$\sum_{j=1}^{s} b_j = 1 \tag{2.8a}$$

$$A \in \mathbb{R}^{s\times s} \text{ is invertible} \tag{2.8b}$$

$$M := \begin{bmatrix} b^T \\ b^T - b^T A \\ b^T - 2b^T CA \\ \vdots \\ b^T - (s-1)b^T C^{s-2} A \end{bmatrix} \in \mathbb{R}^{s \times s} \text{ is invertible.} \tag{2.8c}$$

With these invertibility assumptions, the coefficients can be shown to satisfy a useful property.

**Lemma 2.2.2.** *[15] Assume the coefficients A, b, and c for the SPARK method (2.7) satisfy (2.8). Then the matrix $\widetilde{\Omega}_0 \alpha$ is invertible.*

*Proof.* This lemma is easily seen to be true by observing that

$$M = D\widetilde{\Omega}_0 \alpha, \tag{2.9}$$

for $D$ defined as

$$D := \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & -(s-1) \end{bmatrix}. \tag{2.10}$$

Because the matrix $M$ is assumed invertible, and the triangular matrix $D$ is invertible, the product $\widetilde{\Omega}_0 \alpha$ is easily seen to be invertible. □

### 2.2.1 Gauss SPARK Methods

An important example of a class of SPARK methods are the $s$-Gauss SPARK methods. The coefficients $a_{ij}$, $b_i$, and $c_i$ are from the $s$-stage Gauss RK methods. These coefficients satisfy the properties

$$B(2s): \quad \sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \quad k = 1, \dots, 2s \tag{2.11}$$

$$C(s): \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_j^k}{k}, \quad i = 1, \ldots, s, \quad k = 1, \ldots, s \tag{2.12}$$

$$D(s): \quad \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{b_j}{k}(1 - c_j^k), \qquad j = 1, \ldots, s, \ k = 1, \ldots, s. \tag{2.13}$$

The condition (2.11) is from the $s$-stage Gaussian quadrature and the condition (2.12) is from the Gauss RK coefficients. The matrix $A$ is known to be invertible for the Gauss methods. A proof of the invertibility of the matrix $M$ is now presented, although the theorem is used in [15].

**Theorem 2.2.3.** *[15] If the SPARK methods applied to problem (2.1) satisfy the condition $D(s-1)$, the matrix*

$$M = \begin{bmatrix} b^T \\ b^T - b^T A \\ b^T - 2b^T C A \\ \vdots \\ b^T - (s-1) b^T C^{s-2} A \end{bmatrix}$$

*is invertible.*

*Proof.* The condition $D(s-1)$ can be written as $k b^T C^{k-1} A = b^T - b^T C^k$ for $k = 1, \ldots, s-1$, or as $b^T C^k = b^T - k b^T C^{k-1}$. Using this, the matrix $M$ can be expressed as

$$M = \begin{bmatrix} b^T \\ b^T C \\ b^T C^2 \\ \vdots \\ b^T C^{s-1} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ c_1 & c_2 & \ldots & c_s \\ c_1^2 & c_2^2 & \ldots & c_s^2 \\ \vdots & \vdots & & \vdots \\ c_1^{s-1} & c_2^{s-1} & \ldots & c_s^{s-1} \end{bmatrix} \begin{bmatrix} b_1 & & & & O \\ & b_2 & & & \\ & & b_3 & & \\ & & & \ddots & \\ O & & & & b_s \end{bmatrix}.$$

This shows that $M$ can be expressed as the product of an invertible Vandermonde matrix and an invertible diagonal matrix. Thus, $M$ must be invertible. □

Because the Gauss methods satisfy $D(s)$, Theorem 2.2.3 implies the invertibility of $M$ for the Gauss coefficients.

## 2.3 Existence, Uniqueness, and Influence of Perturbations

The existence of a unique solution to the system of equations (2.7) is given in [15]. We state it here for completeness. In this section we use the notation $y_0 = y_0(h)$ and $\psi_0 = \psi_0(h)$.

**Theorem 2.3.1.** *Suppose that $y_0$ and $\psi_0$ satisfy*

$$k(t_0, y_0) = o(h), \quad k_y(t_0, y_0) f(t_0, y_0, \psi_0) = o(1) \tag{2.14}$$

*and that $k_y(t, y) f_\psi(t, y, \psi)$ is invertible. Then for $|h| \leq h_0$, there exists a locally unique solution to (2.7) that satisfies*

$$Y_i - y_0 = \mathcal{O}(h), \quad y_1 - y_0 = \mathcal{O}(h), \quad \Psi_i - \psi_0 = \mathcal{O}(h), \quad i = 1, \ldots, s. \tag{2.15}$$

We now consider the influence of perturbations to the solution of the SPARK method (2.7). We consider the perturbed system

$$\widehat{Y}_i = \widehat{y}_0 + h \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j) + h\delta_i^y, \quad i = 1, \ldots, s \tag{2.16a}$$

$$\widehat{y}_1 = \widehat{y}_0 + h \sum_{j=1}^{s} b_j f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j) + h\delta_{s+1}^y \tag{2.16b}$$

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, \widehat{Y}_j) + \omega_{i,s+1} k(t_1, \widehat{y}_1) + \delta_i^\psi, \quad i = 1, \ldots, s. \tag{2.16c}$$

We examine the influence of the perturbations $\delta^y := [\delta_1^y, \delta_2^y, \ldots, \delta_s^y, \delta_{s+1}^y]^T$ and $\delta^\psi := [\delta_1^\psi, \delta_2^\psi, \ldots, \delta_s^\psi]^T$ on the numerical solution. For simplicity, we introduce the notations

$$Y := [Y_1^T, Y_2^T, \ldots, Y_s^T]^T \qquad \widehat{Y} := [\widehat{Y}_1^T, \widehat{Y}_2^T, \ldots, \widehat{Y}_s^T]^T$$

$$\Psi := [\Psi_1^T, \Psi_2^T, \ldots, \Psi_s^T]^T \qquad \widehat{\Psi} := [\widehat{\Psi}_1^T, \widehat{\Psi}_2^T, \ldots, \widehat{\Psi}_s^T]^T$$

$$\Delta Y_i := \widehat{Y}_i - Y_i \qquad \Delta\Psi_i := \widehat{\Psi}_i - \Psi_i \qquad \Delta\widetilde{Y} := [\widehat{Y}^T, \widehat{y}_1^T]^T - [Y^T, y_1^T]^T$$

$$\Delta y_0 := \widehat{y}_0 - y_0 \qquad \Delta y_1 := \widehat{y}_1 - y_1.$$

We also define $||Y|| := \max_i\{||Y_i||\}$, $||\Psi|| := \max_i\{||\Psi_i||\}$, etc.

**Theorem 2.3.2.** *Suppose* $(t_0, y_0, \psi_0)$ *satisfy*

$$k(t_0, y_0) = o(h)$$
$$k_y(t_0, y_0) f(t_0, y_0, \psi_0) = o(1),$$

*and* $k_y(t, y) f_\psi(t, y, \psi)$ *is invertible around* $(t_0, y_0, \psi_0)$. *Let* $Y_i$, $\Psi_i$, $y_1$ *satisfy (2.7), and* $\widehat{Y}_i$, $\widehat{\Psi}_i$, $\widehat{y}_1$ *satisfy (2.16). Assume also that*

$$\Delta y_0 = \mathcal{O}(h^2), \quad \widehat{\Psi}_i - \psi_0 = \mathcal{O}(h), \quad \delta^y = \mathcal{O}(h), \quad \delta^\psi = \mathcal{O}(h^2). \tag{2.17}$$

*Then for* $|h| \le h_0$, *we have*

$$\Delta Y_i = \Delta y_0 + \mathcal{O}(||k_y(t_0, y_0)\Delta y_0|| + h||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||) \tag{2.18a}$$
$$\Delta y_1 = \Delta y_0 + \mathcal{O}(||k_y(t_0, y_0)\Delta y_0|| + h||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||) \tag{2.18b}$$
$$h\Delta\Psi_i = \mathcal{O}(||k_y(t_0, y_0)\Delta y_0|| + h||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||), \tag{2.18c}$$

*for* $i = 1, \ldots, s$.

*Proof.* A similar proof can be found in [12]. By subtracting the formulas (2.7) from (2.16), we get

$$\begin{aligned}\Delta Y_i = \Delta y_0 + h\sum_{j=1}^{s} a_{ij}(f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j) - f(t_0 + c_j h, Y_j, \Psi_j)) \\ + h\delta_i^y, \quad i = 1, \ldots, s,\end{aligned} \tag{2.19a}$$
$$\Delta y_1 = \Delta y_0 + h\sum_{j=1}^{s} b_j(f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j) - f(t_0 + c_j h, Y_j, \Psi_j)) + h\delta_{s+1}^y \tag{2.19b}$$
$$\begin{aligned}0 = \sum_{j=1}^{s} \omega_{ij}(k(t_0 + c_j h, \widehat{Y}_j) - k(t_0 + c_j h, Y_j)) \\ + \omega_{i,s+1}(k(t_1, \widehat{y}_1) - k(t_1, y_1)) + \delta_i^\psi, \quad i = 1, \ldots, s.\end{aligned} \tag{2.19c}$$

Expanding the terms $f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j)$ and $k(t_0 + c_j h, \widehat{Y}_j)$ in a Taylor series around

$(Y_j, \Psi_j)$ gives

$$\begin{aligned} f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j) &= f(t_0 + c_j h, Y_j, \Psi_j) + f_y(t_0 + c_j h, Y_j, \Psi_j)\Delta Y_j \\ &\quad + f_\psi(t_0 + c_j h, Y_j, \Psi_j)\Delta\Psi_j + \mathcal{O}\left(||\Delta Y||^2 + ||\Delta\Psi||^2\right) \\ k(t_0 + c_j h, \widehat{Y}_j) &= k(t_0 + c_j h, Y_j) + k_y(t_0 + c_j h, Y_j)\Delta Y_j + \mathcal{O}\left(||\Delta Y||^2\right). \end{aligned}$$

In the above, we eliminate the cross term $||\Delta Y|| \cdot ||\Delta\Psi||$ from the big-oh, as $||\Delta Y|| \cdot ||\Delta\Psi|| \leq \frac{1}{2}(||\Delta Y||^2 + ||\Delta\Psi||^2)$. Inserting these two equations into (2.19) gives

$$\begin{aligned} \Delta Y_i &= \Delta y_0 + h\sum_{j=1}^{s} a_{ij}(f_y(t_0 + c_j h, Y_j, \Psi_j)\Delta Y_j + f_\psi(t_0 + c_j h, Y_j, \Psi_j)\Delta\Psi_j) \\ &\quad + h\delta_i^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta\Psi||^2\right) \\ \Delta y_1 &= \Delta y_0 + h\sum_{j=1}^{s} b_j(f_y(t_0 + c_j h, Y_j, \Psi_j)\Delta Y_j + f_\psi(t_0 + c_j h, Y_j, \Psi_j)\Delta\Psi_j) \\ &\quad + h\delta_{s+1}^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta\Psi||^2\right) \\ 0 &= \sum_{j=1}^{s} \omega_{ij} k_y(t_0 + c_j h, Y_j)\Delta Y_j + \omega_{i,s+1} k_y(t_1, y_1)\Delta y_1 \\ &\quad + \delta_i^\psi + \mathcal{O}(||\Delta Y||^2 + ||\Delta y_1||^2). \end{aligned}$$

This can be rewritten by using tensor notation. Doing so gives

$$\begin{aligned} \Delta\widetilde{Y} &= \mathbb{1}_{s+1} \otimes \Delta y_0 + h(\alpha \otimes I_{n_y})\{f_y\}\Delta Y + (\alpha \otimes I_{n_y})\{f_\psi\}(h\Delta\Psi) \\ &\quad + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta\Psi||^2 + h||\delta^y||\right) \end{aligned} \tag{2.20}$$

$$0 = (\widetilde{\Omega}_0 \otimes I_{n_k})\widetilde{\{k_y\}}\Delta\widetilde{Y} + \mathcal{O}\left(||\Delta\widetilde{Y}||^2 + ||\delta^\psi||\right). \tag{2.21}$$

Here we use the notations

$$\{f_y\} := \text{blockdiag}(f_y(t_0 + c_1 h, Y_1, \Psi_1), \ldots, f_y(t_0 + c_s h, Y_s, \Psi_s)) \in \mathbb{R}^{sn_y \times sn_y}$$

$$\{f_\psi\} := \text{blockdiag}(f_\psi(t_0 + c_1 h, Y_1, \Psi_1), \ldots, f_\psi(t_0 + c_s h, Y_s, \Psi_s)) \in \mathbb{R}^{sn_y \times sn_k}$$

$$\widetilde{\{k_y\}} := \text{blockdiag}(k_y(t_0 + c_1 h, Y_1), \ldots, k_y(t_0 + c_s h, Y_s), k_y(t_1, y_1)) \in \mathbb{R}^{(s+1)n_k \times (s+1)n_y}$$

Substituting (2.20) for $\Delta\widetilde{Y}$ in (2.21), and rearranging, results in

$$\begin{aligned}-(\widetilde{\Omega}_0 \otimes I_{n_k})\widetilde{\{k_y\}}(\alpha \otimes I_{n_y})\{f_\psi\}(h\Delta\Psi) = \\ (\widetilde{\Omega}_0 \otimes I_{n_k})\widetilde{\{k_y\}}\left(\mathbb{1}_{s+1} \otimes \Delta y_0 + h(\alpha \otimes I_{n_y})\{f_y\}\Delta Y\right) \\ + \mathcal{O}\left(||\Delta\widetilde{Y}||^2 + h||\Delta\Psi||^2 + h||\delta^y|| + ||\delta^\psi||\right).\end{aligned} \tag{2.22}$$

This equation can be solved for $h\Delta\Psi$. To see this, observe that

$$\begin{aligned}k_y(t_0 + c_i h, Y_i) &= k_y(t_0 + (c_i h), y_0 + (Y_i - y_0)) \\ &= k_y(t_0, y_0) + \mathcal{O}(h) \\ f_\psi(t_0 + c_l h, Y_l, \Psi_l) &= f_\psi(t_0 + (c_l h), y_0 + (Y_l - y_0), \psi_0 + (\Psi_l - \psi_0)) \\ &= f_\psi(t_0, y_0, \psi_0) + \mathcal{O}(h).\end{aligned}$$

Therefore, the matrix on the left-hand side of equation (2.22) can be expressed as

$$\begin{aligned}&(\widetilde{\Omega}_0 \otimes I_{n_k})\widetilde{\{k_y\}}(\alpha \otimes I_{n_y})\{f_\psi\} \\ &\qquad = \left[\sum_{i=1}^{s+1} \omega_{ji}\alpha_{il}k_y(t_0 + c_i h, Y_i)f_\psi(t_0 + c_l h, Y_l, \Psi_l)\right]_{\substack{j=1,\ldots,s \\ l=1,\ldots,s}} \\ &\qquad = \left[\sum_{i=1}^{s+1} \omega_{ji}\alpha_{il}k_y(t_0, y_0)f_\psi(t_0, y_0, \psi_0) + \mathcal{O}(h)\right]_{\substack{j=1,\ldots,s \\ l=1,\ldots,s}} \\ &\qquad = (\widetilde{\Omega}_0\alpha) \otimes (k_y(t_0, y_0)f_\psi(t_0, y_0, \psi_0)) + \mathcal{O}(h)\end{aligned}$$

is invertible, for sufficiently small $h$. In the last step above, we use the fact that $\widetilde{\Omega}_0\alpha$ is invertible by Lemma 2.2.2. To simplify notation, let

$$D := \left(-(\widetilde{\Omega}_0 \otimes I_{n_k})\widetilde{\{k_y\}}(\alpha \otimes I_{n_y})\{f_\psi\}\right)^{-1}.$$

We therefore get

$$\begin{aligned}h\Delta\Psi = D(\widetilde{\Omega}_0 \otimes I_{n_k})\widetilde{\{k_y\}}\left[\mathbb{1}_{s+1} \otimes \Delta y_0 + h(\alpha \otimes I_{n_y})\{f_y\}\Delta Y\right] \\ + \mathcal{O}\left(||\Delta\widetilde{Y}||^2 + h||\Delta\Psi||^2 + h||\delta^y|| + ||\delta^\psi||\right).\end{aligned} \tag{2.23}$$

Substituting (2.23) into (2.20) gives

$$\begin{aligned}\Delta\widetilde{Y} = P\left(\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\alpha\otimes I_{n_y})\{f_y\}\Delta Y\right)\\ + \mathcal{O}\left(||\Delta\widetilde{Y}||^2 + h||\Delta\Psi||^2 + h||\delta^y|| + ||\delta^\psi||\right)\end{aligned} \tag{2.24}$$

with the definition

$$P := I_{(s+1)n_y} - (\alpha\otimes I_{n_y})\{f_\psi\}D(\widetilde{\Omega}_0\otimes I_{n_k})\widetilde{\{k_y\}}.$$

Because of the expansions

$$\begin{aligned}k_y(t_0+c_ih, Y_i)\Delta y_0 &= k_y(t_0+(c_ih), y_0+(Y_i-y_0))\Delta y_0\\ &= k_y(t_0,y_0)\Delta y_0 + \mathcal{O}(h||\Delta y_0||)\\ k_y(t_1,y_1)\Delta y_0 &= k_y(t_0,y_0)\Delta y_0 + \mathcal{O}(h||\Delta y_0||),\end{aligned}$$

the final results follow from (2.23) and (2.24). □

Each of the coefficients on big-oh terms in the result of Theorem 2.3.2 depend only upon the derivatives of the functions $f$ and $k$, not upon any of the constants from the hypothesis. With some additional assumptions, the bounds of this perturbation theorem can be improved.

**Corollary 2.3.3.** *If, in addition to the conditions of Theorem 2.3.2, we assume*

$$k(t_0,y_0) = 0 = k(t_0,\widehat{y}_0),$$

*then we have the bounds*

$$\Delta Y_i = \Delta y_0 + \mathcal{O}(h||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||), \quad i=1,\ldots,s \tag{2.25a}$$

$$\Delta y_1 = \Delta y_0 + \mathcal{O}(h||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||) \tag{2.25b}$$

$$h\Delta\Psi_i = \mathcal{O}(h||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||), \quad i=1,\ldots,s. \tag{2.25c}$$

*With the added assumption of this corollary, we may relax the assumption* $\Delta y_0 = \mathcal{O}(h^2)$ *to* $\Delta y_0 = \mathcal{O}(h)$ *in Theorem 2.3.2.*

*Proof.* Because the constraints are satisfied at $y_0$, the terms $k_y(t_0, y_0)\Delta y_0$ in the result of Theorem 2.3.2 can be removed, as

$$0 = k(t_0, \widehat{y}_0) - k(t_0, y_0) = k_y(t_0, y_0)\Delta y_0 + \mathcal{O}(||\Delta y_0||^2),$$

and

$$k_y(t_0, y_0)\Delta y_0 = \mathcal{O}(||\Delta y_0||^2) = \mathcal{O}(h||\Delta y_0||).$$

This is because of the assumption $\Delta y_0 = \mathcal{O}(h)$. This gives the equations (2.25). □

## 2.4 Collocation Type Methods

We give a presentation of a type of collocation method for the index 2 problem (2.1). The SPARK methods (2.7) are equivalent to these methods. Similar proofs can be found in [9], [10], and [16].

**Definition 2.4.1.** *Let $c_1, \ldots, c_s$ be distinct real numbers. We define $Y(t)$ and $\Psi(t)$ to be the $s$ and $s-1$ degree polynomials, respectively, that satisfy*

$$Y(t_0) = y_0 \tag{2.26a}$$

$$\dot{Y}(t_0 + c_i h) = f(t_0 + c_i h, Y(t_0 + c_i h), \Psi(t_0 + c_i h)) \tag{2.26b}$$

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y(t_0 + c_j h)) + \omega_{i,s+1} k(t_1, Y(t_1)) \tag{2.26c}$$

*for $i = 1, \ldots, s$. We refer to the polynomials $Y$ and $\Psi$ as* collocation polynomials. *The value of $Y(t_1)$ is used as an approximation to the exact solution $y(t)$ of (2.1) at time $t_1 := t_0 + h$.*

**Theorem 2.4.2.** *The collocation polynomials $Y(t)$ and $\Psi(t)$ defined by (2.26) are equivalent to an $s$-stage SPARK method for index 2 problems. The coefficients are given by*

$$a_{ij} = \int_0^{c_i} \ell_j(\tau) d\tau, \qquad b_j = \int_0^1 \ell_j(\tau) d\tau, \qquad i, j = 1, \ldots, s, \tag{2.27}$$

*where the $\ell_j(\tau)$ are the Lagrange polynomials given by*

$$\ell_j(\tau) = \prod_{\substack{k=1 \\ k\neq j}}^{s} \left( \frac{\tau - c_k}{c_j - c_k} \right). \tag{2.28}$$

*Proof.* Using Lagrangian interpolation, we write

$$\dot{Y}(t_0 + \tau h) = \sum_{j=1}^{s} \ell_j(\tau) f(t_0 + c_j h, Y(t_0 + c_j h), \Psi(t_0 + c_j h)). \tag{2.29}$$

Next, using the Fundamental Theorem of Calculus gives

$$Y(t_0 + c_i h) = y_0 + h \int_0^{c_i} \dot{Y}(t_0 + \tau h) d\tau. \tag{2.30}$$

Inserting (2.29) gives

$$Y(t_0 + c_i h) = y_0 + h \sum_{j=1}^{s} \dot{Y}(t_0 + c_j h) \int_0^{c_i} \ell_j(\tau) d\tau. \tag{2.31}$$

Take $a_{ij} := \int_0^{c_i} \ell_j(\tau) d\tau$. If we define $Y_i := Y(t_0 + c_i h)$ and $\Psi_i := \Psi(t_0 + c_i h)$ here, then from (2.26b),

$$\begin{aligned} \dot{Y}(t_0 + c_i h) &= f(t_0 + c_i h, Y(t_0 + c_i h), \Psi(t_0 + c_i h)) \\ &= f(t_0 + c_i h, Y_i, \Psi_i). \end{aligned}$$

Note that (2.31) can thus be viewed as the internal stages of a SPARK method. Similarly, we write

$$Y(t_1) = y_0 + h \int_0^1 \dot{Y}(t_0 + \tau h) d\tau. \tag{2.32}$$

Again inserting (2.29) gives

$$\begin{aligned} Y(t_1) &= y_0 + h \sum_{j=1}^{s} \dot{Y}(t_0 + c_j h) \int_0^1 \ell_j(\tau) d\tau \\ &= y_0 + h \sum_{j=1}^{s} f(t_0 + c_j h, Y_j, \Psi_j) \int_0^1 \ell_j(\tau) d\tau. \end{aligned} \tag{2.33}$$

Taking $b_j := \int_0^1 \ell_j(\tau) d\tau$ and $y_1 := Y(t_1)$, (2.33) becomes the numerical solution

given by a SPARK method. Lastly, the values $Y_i$ and $y_1$ satisfy the linear combination of a SPARK method for the constraint directly from their definitions and from (2.26c). □

Many SPARK methods for index 2 problems, including the Gauss SPARK methods, can be expressed as such collocation type methods with reasonable hypotheses. This fact is useful for determining the order of SPARK methods. We take advantage of this later. For now, we give the equivalence of SPARK methods and collocation type methods in the following theorem.

**Theorem 2.4.3.** *An s-stage SPARK method (2.7) applied to an index 2 problem (2.1) with distinct $c_i$ coefficients ($i = 1, \ldots, s$) is a collocation type method (2.26) iff $C(s)$ and $B(s)$ hold.*

*Proof.* Given the coefficients $c_i$, the conditions $C(s)$ and $B(s)$ determine the $a_{ij}$ and $b_i$ coefficients uniquely. These two conditions can be expressed as

$$\sum_{j=1}^{s} a_{ij} p(c_j) = \int_0^{c_i} p(t)dt, \quad i = 1, \ldots, s \tag{2.34a}$$

$$\sum_{j=1}^{s} b_j p(c_j) = \int_0^{1} p(t)dt, \quad i = 1, \ldots, s \tag{2.34b}$$

for all polynomials $p$ with degree less than or equal to $s-1$. However, the coefficients $a_{ij}$ and $b_i$ defined in Theorem 2.4.2 satisfy these relations, as then (2.34) are just the Lagrange interpolation formulas.

Thus, if an $s$-stage SPARK method is a collocation type method, it satisfies (2.34) and thus satisfies $C(s)$ and $B(s)$. If the SPARK method satisfies $C(s)$ and $B(s)$, then $a_{ij}$ and $b_i$ are unique, with (2.34) satisfied by the coefficients of Theorem 2.4.2. This completes the proof. □

We present now a lemma regarding the local error of the internal stages of a collocation type or SPARK method, assuming Gauss coefficients. This lemma will be useful for showing the effectiveness of the derivatives of collocation methods, as

well as for a proof of local convergence.

**Lemma 2.4.4.** *Suppose the internal stages $Y_i$ and $\Psi_i$ are as defined in (2.7) with Gauss coefficients, for $i = 1, \ldots, s$. Let $y(t), \psi(t)$ be the exact solutions to (2.1). Then we have the bounds*

$$Y_i - y(t_0 + c_i h) = \mathcal{O}(h^{s+1}), \qquad \Psi_i - \psi(t_0 + c_i h) = \mathcal{O}(h^s). \tag{2.35}$$

*Proof.* We apply Corollary 2.3.3 using the exact solution for the perturbed values. So take

$$\widehat{Y}_i = y(t_0 + c_i h), \quad \widehat{y}_1 = y(t_1), \quad \widehat{\Psi}_i = \psi(t_0 + c_i h), \quad \widehat{y}_0 = y(t_0). \tag{2.36}$$

Since $k(t_0 + c_i h, y(t_0 + c_i h)) = k(t_1, y(t_1)) = 0$, the constraint (2.16c) gives that $\delta_i^{\psi} = 0$ for all $i = 1, \ldots, s$. Using a Taylor series expansion around $h = 0$, the value $\widehat{Y}_i = y(t_0 + c_i h)$ can be expressed as

$$y(t_0 + c_i h) = \sum_{k=0}^{s} \frac{h^k}{k!} c_i^k y^{(k)}(t_0) + \mathcal{O}(h^{s+1}). \tag{2.37}$$

Thus (2.16a) gives, for $i = 1, \ldots, s$,

$$\begin{aligned}
\delta_i^y &= \frac{1}{h}(\widehat{Y}_i - \widehat{y}_0) - \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{k!} c_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} \sum_{k=0}^{s} a_{ij} \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \left[ \frac{h^{k-1}}{k!} c_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} a_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{c_i^k}{k} - \sum_{j=1}^{s} a_{ij} c_j^{k-1} \right] + \mathcal{O}(h^{s+1}) \\
&= \frac{h^s}{s!} y^{(s+1)}(t_0) \left[ \frac{c_i^{s+1}}{s+1} - \sum_{j=1}^{s} a_{ij} c_j^s \right] + \mathcal{O}(h^{s+1}) \\
&= \mathcal{O}(h^s).
\end{aligned}$$

We have made use of the fact that the Gauss coefficients satisfy the property $C(s)$.

For $\delta^y_{s+1}$, a similar derivation can be made, using the fact that the Gauss coefficients also satisfy $B(2s)$, i.e., $\sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}$ for $k = 1, \ldots, 2s$. This, along with (2.16b), gives

$$\begin{aligned}
\delta^y_{s+1} &= \frac{1}{h}(\widehat{y}_1 - \widehat{y}_0) - \sum_{j=1}^{s} b_j f(t_0 + c_j h, \widehat{Y}_j, \widehat{\Psi}_j) \\
&= \sum_{k=1}^{2s+1} \frac{h^{k-1}}{k!} y^{(k)}(t_0) - \sum_{j=1}^{2s} \sum_{k=0}^{s} b_j \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{2s+1}) \\
&= \sum_{k=1}^{2s+1} \left[ \frac{h^{k-1}}{k!} y^{(k)}(t_0) - \sum_{j=1}^{s} b_j \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{2s+1}) \\
&= \sum_{k=1}^{2s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{1}{k} - \sum_{j=1}^{s} b_j c_j^{k-1} \right] + \mathcal{O}(h^{2s+1}) \\
&= \frac{h^{2s}}{(2s)!} y^{(2s+1)}(t_0) \left[ \frac{1}{2s+1} - \sum_{j=1}^{s} b_j c_j^{2s} \right] + \mathcal{O}(h^{2s+1}) \\
&= \mathcal{O}(h^{2s}).
\end{aligned}$$

Thus, applying Corollary 2.3.2 gives

$$\begin{aligned}
\widehat{Y}_i - Y_i &= y(t_0 + c_i h) - Y_i \\
&= \mathcal{O}(||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||) \\
&= \mathcal{O}(h||\delta^y||) \\
&= \mathcal{O}(h^{s+1}) \\
\widehat{\Psi}_i - \Psi_i &= \psi(t_0 + c_i h) - \Psi_i \\
&= \frac{1}{h}\mathcal{O}(h||\Delta y_0|| + h||\delta^y|| + ||\delta^\psi||) \\
&= \mathcal{O}(||\delta^y||) \\
&= \mathcal{O}(h^s).
\end{aligned}$$

□

The next theorem gives the quality of the derivatives of the approximations by collocation type methods. The proof of this theorem is similar to that of [9, Theorem II.7.10] and [10, Theorem VII.4.8].

**Theorem 2.4.5.** *Let $y(t), \psi(t)$ be the exact solutions to the problem (2.1). The collocation type polynomials $Y(t)$ and $\Psi(t)$ defined by (2.26) with Gauss coefficients $c_i$ satisfy for $k = 0, \ldots, s$ and $t \in [t_0, t_1]$*

$$||Y^{(k)}(t) - y^{(k)}(t)|| \leq Ch^{s+1-k} \tag{2.38a}$$

$$||\Psi^{(k)}(t) - \psi^{(k)}(t)|| \leq Ch^{s-k}. \tag{2.38b}$$

*Proof.* As in the proof of Theorem 2.4.2, we use $Y_i := Y(t_0 + c_i h)$ and $\Psi_i := \Psi(t_0 + c_i h)$. We can write the collocation polynomials as

$$Y(t_0 + \tau h) = y_0 \ell_0(\tau) + \sum_{i=1}^{s} \ell_i(\tau) Y_i \tag{2.39a}$$

$$\Psi(t_0 + \tau h) = \sum_{i=1}^{s} \ell_i(\tau) \Psi_i \tag{2.39b}$$

with Lagrange polynomials $\ell_i$ defined as

$$\ell_0(\tau) = \prod_{j=1}^{s} \left( \frac{\tau - c_j}{-c_j} \right), \qquad \ell_i(\tau) = \frac{\tau}{c_i} \prod_{\substack{j=1 \\ j \neq i}}^{s} \left( \frac{\tau - c_j}{c_i - c_j} \right).$$

Applying the Lagrange interpolation formula to the exact solution gives

$$y(t_0 + \tau h) = \ell_0(\tau) y_0 + \sum_{i=1}^{s} \ell_i(\tau) y(t_0 + c_i h) + \mathcal{O}(h^{s+1}) \tag{2.40a}$$

$$\psi(t_0 + \tau h) = \sum_{i=1}^{s} \ell_i(\tau) \psi(t_0 + c_i h) + \mathcal{O}(h^s). \tag{2.40b}$$

The functions $\alpha(\tau) := y(t_0 + \tau h) - y_0 \ell_0(\tau) - \sum_{i=1}^{s} \ell_i(\tau) y(t_0 + c_i h)$ and $\beta(\tau) := \psi(t_0 + \tau h) - \sum_{i=1}^{s} \ell_i(\tau) \psi(t_0 + c_i h)$ have at least $s + 1$ and $s$ zeros, respectively, at each $c_i$, $i = 1, \ldots, s$ (and at 0 for $\alpha(\tau)$). Thus, applying Rolle's Theorem to each interval $(c_i, c_{i+1})$ for $i = 1, \ldots, s - 1$, and to $(0, c_1)$, we see that

$$\alpha^{(k)}(\tau) = h^k y^{(k)}(t_0 + \tau h) - \left( \ell_0^{(k)}(\tau) y_0 + \sum_{i=1}^{s} \ell_i^{(k)}(\tau) y(t_0 + c_i h) \right) \tag{2.41a}$$

$$\beta^{(k)}(\tau) = h^k \psi^{(k)}(t_0 + \tau h) - \left( \sum_{i=1}^{s} \ell_i^{(k)}(\tau) \psi(t_0 + c_i h) \right) \tag{2.41b}$$

have $s+1-k$ zeros and $s-k$ zeros, respectively. The terms in brackets in (2.41) can thus be viewed as interpolation polynomials of degree $s-k$ and $s-k-1$, respectively, for $h^k y^{(k)}(t_0+\tau h)$ and $h^k \psi^{(k)}(t_0+\tau h)$. The interpolation error for each is thus

$$\alpha^{(k)}(\tau) = \mathcal{O}(h^{s+1}), \qquad \beta^{(k)}(\tau) = \mathcal{O}(h^s).$$

Therefore, subtracting (2.40) from (2.39) and taking $k$ derivatives with respect to $s$, we arrive at

$$h^k \left(Y^{(k)}(t_0+\tau h) - y^{(k)}(t_0+\tau h)\right) = \sum_{i=1}^{s} (Y_i - y(t_0+c_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1})$$

$$h^k \left(\Psi^{(k)}(t_0+\tau h) - \psi^{(k)}(t_0+\tau h)\right) = \sum_{i=1}^{s} (\Psi_i - \psi(t_0+c_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s}).$$

Invoking Lemma 2.4.4, and dividing by $h^k$, we arrive at

$$Y^{(k)}(t_0+\tau h) - y^{(k)}(t_0+\tau h) = \mathcal{O}(h^{s+1-k})$$

$$\Psi^{(k)}(t_0+\tau h) - \psi^{(k)}(t_0+\tau h) = \mathcal{O}(h^{s-k}).$$

This gives the desired result. □

## 2.5 Local Error Analysis

A derivation of the local error for the SPARK method (2.7) is given in [15] using a Taylor series approach. We present here an alternative proof using collocation type methods instead. The proof for the Gauss-Lobatto SPARK methods for mixed index constraints will utilize similar techniques. The proof presented here for Gauss coefficients is similar to that found in [16] for local error analysis. However, the treatment of the constraints by the SPARK methods for index 2 problems does alter the proof.

**Theorem 2.5.1.** *For the Gauss SPARK methods (2.7) with a consistent initial value $y_0$ at time $t_0$, i.e., $k(t_0, y_0) = 0$, $k_y(t,y) f_\psi(t,y,\psi)$ invertible, and $|h| \le h_0$, the*

*local error is of order* $2s$*, i.e.,*

$$y_1 - y(t_1) = \mathcal{O}(h^{2s+1}). \tag{2.42}$$

*Proof.* For the proof, we utilize the collocation polynomials $Y(t)$ and $\Psi(t)$ defined in (2.26). We define the defects $\delta(t)$ and $\theta(t)$ by

$$\dot{Y}(t) = f(t, Y(t), \Psi(t)) + \delta(t) \tag{2.43a}$$

$$0 = k(t, Y(t)) + \theta(t). \tag{2.43b}$$

From the definition of $Y$ and $\Psi$, we have that $\delta(t_0 + c_i h) = 0$ for $i = 1, \ldots, s$ and $\theta(t_0) = \theta(t_1) = 0$ since $k(t_0, Y(t_0)) = k(t_0, y_0) = 0$ and $k(t_1, Y(t_1)) = k(t_1, y_1) = 0$ from the definition of the SPARK method. Note that in general, $k(t_0 + c_i h, Y(t_0 + c_i h)) \neq 0$, and therefore $\theta(t_0 + c_i h) \neq 0$.

The derivative of the constraint (2.43b) gives

$$0 = k_y(t, Y(t))\dot{Y}(t) + \dot{\theta}(t) = k_y(t, Y(t)) \left(f(t, Y(t), \Psi(t)) + \delta(t)\right) + \dot{\theta}(t). \tag{2.44}$$

Because $k_y f_\psi$ is invertible, the implicit function theorem allows us to write

$$\Psi(t) := \Upsilon \left(t, Y(t), \delta(t), \dot{\theta}(t)\right). \tag{2.45}$$

Using this in (2.43a), we get the ODE

$$\dot{Y}(t) = f \left(t, y(t), \Upsilon \left(t, Y(t), \delta(t), \dot{\theta}(t)\right)\right) + \delta(t). \tag{2.46}$$

Note that for the exact solution to the problem (2.1), we have

$$\dot{y}(t) = f(t, y(t), \Upsilon(t, y(t), 0, 0)), \tag{2.47}$$

as the exact solution satisfies (2.43a) and (2.43b) with $\delta(t) = \theta(t) = 0$.

We now apply the Gröbner-Alekseev formula ([9, Theorem I.14.5]). To do so,

we calculate the defect $d(t) := \dot{Y}(t) - \dot{y}(t)$ as

$$\begin{aligned} d(t) &= f\left(t, Y(t), \Upsilon\left(t, Y(t), \delta(t), \dot{\theta}(t)\right)\right) + \delta(t) - f\left(t, Y(t), \Upsilon(t, Y(t), 0, 0)\right) \\ &= \Phi(t,1) - \Phi(t,0) + \delta(t) \\ &= \int_0^1 \frac{\partial \Phi}{\partial \tau}(t,\tau)d\tau + \delta(t) \end{aligned}$$

for $\Phi(t,\tau) := f\left(t, Y(t), \Upsilon\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)\right)$. However, $\frac{\partial \Phi}{\partial \tau}$ can be expressed as

$$\begin{aligned} \frac{\partial \Phi}{\partial \tau} = f_\psi &\left(t, Y(t), \Upsilon\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)\right) \cdot \\ &\left[\Upsilon_\delta\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)\delta(t) + \Upsilon_{\dot{\theta}}\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)\dot{\theta}(t)\right]. \end{aligned} \tag{2.48}$$

Note here that $d(t)$ can be expressed in the form

$$d(t) = Q_1(t, Y(t), \delta(t), \dot{\theta}(t))\delta(t) + Q_2(t, Y(t), \delta(t), \dot{\theta}(t))\dot{\theta}(t) \tag{2.49}$$

for

$$\begin{aligned} Q_1(t, Y(t), &\delta(t), \dot{\theta}(t)) = I_{n_y} + \\ &\int_0^1 f_\psi\left(t, Y(t), \Upsilon\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)\right)\Upsilon_\delta\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)d\tau \\ Q_2(t, Y(t), &\delta(t), \dot{\theta}(t)) = \\ &\int_0^1 f_\psi\left(t, Y(t), \Upsilon\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)\right)\Upsilon_{\dot{\theta}}\left(t, Y(t), \tau\delta(t), \tau\dot{\theta}(t)\right)d\tau \end{aligned}$$

So, by the Gröbner-Alekseev formula, we get

$$\begin{aligned} Y(t) - y(t) &= \int_{t_0}^t \frac{\partial y}{\partial y_0}(t, s, Y(s))d(s)ds \\ &= \int_{t_0}^t S_1(t,s)\delta(s) + S_2(t,s)\dot{\theta}(s)ds, \end{aligned}$$

where we have introduced

$$\begin{aligned} S_1(t,s) &:= \frac{\partial y}{\partial y_0}(t, s, Y(s))Q_1(s, Y(s), \delta(s), \dot{\theta}(s)) \\ S_2(t,s) &:= \frac{\partial y}{\partial y_0}(t, s, Y(s))Q_2(s, Y(s), \delta(s), \dot{\theta}(s)). \end{aligned}$$

Integrating by parts then gives

$$Y(t) - y(t) = \int_{t_0}^{t} S_1(t,s)\delta(s)ds + S_2(t,s)\theta(s)\Big|_{s=t_0}^{s=t} - \int_{t_0}^{t} \frac{\partial S_2}{\partial s}(t,s)\theta(s)ds. \quad (2.50)$$

But since $\theta(t_0) = \theta(t_1) = 0$, and substituting $t = t_1$, this becomes

$$\begin{aligned} Y(t_1) - y(t_1) &= y_1 - y(t_1) \\ &= \int_{t_0}^{t_1} S_1(t_1,s)\delta(s) - \frac{\partial S_2}{\partial s}(t_1,s)\theta(s)ds \\ &=: \int_{t_0}^{t_1} \sigma(s)ds. \end{aligned}$$

To evaluate this integral, we apply the Gaussian quadrature formula with $s$ nodes. This results in

$$y_1 - y(t_1) = h\sum_{j=1}^{s} b_j\sigma(t_0 + c_j h) + err(\sigma),$$

with the quadrature error given by

$$||err(\sigma)|| \leq Ch^{2s+1} \max_{t\in[t_0,t_1]} ||\sigma^{(2s+1)}(t)||.$$

The $(2s+1)$-st derivative of $\sigma$ contains derivatives of $f$, $k$, $\theta$, and $\delta$. Because of Theorem 2.4.5 and the smoothness of $\theta$ and $\delta$, each is uniformly bounded on $[t_0, t_1]$ for $h \leq h_0$. Thus $||err(\sigma)|| = \mathcal{O}(h^{2s+1})$. We therefore have

$$y_1 - y(t_1) = h\sum_{j=1}^{s} b_j\sigma(t_0 + c_j h) + \mathcal{O}(h^{2s+1}). \quad (2.51)$$

However, as noted earlier in the proof, $\delta(t_0 + c_j h) = 0$. In addition,

$$\begin{aligned} \theta(t_0 + c_j h) &= -k(t_0 + c_j h, Y_j) \\ &= -k(t_0 + c_j h, y(t_0 + c_j h)) + \mathcal{O}\left(||Y_j - y(t_0 + c_j h)||\right) \\ &= \mathcal{O}\left(||Y_j - y(t_0 + c_j h)||\right) \\ &= \mathcal{O}\left(h^{s+1}\right). \end{aligned}$$

The final equality above is the result from Lemma 2.4.4. Using these results, and

by expanding $\frac{\partial S_2}{\partial s}$ in a Taylor series, we get

$$\begin{aligned}
h\sum_{j=1}^{s} b_j \sigma(t_0 + c_j h) &= -h\sum_{j=1}^{s} b_j \frac{\partial S_2}{\partial s}(t_1, t_0 + c_j h)\theta(t_0 + c_j h) \\
&= -h\sum_{j=1}^{s} b_j \sum_{i=0}^{s-1} \frac{h^i}{i!} c_j^i \frac{\partial^{i+1} S_2}{\partial s^{i+1}}(t_1, t_0)\theta(t_0 + c_j h) \\
&\qquad + \mathcal{O}\left(h^{s+1}\right)\theta(t_0 + c_j h) \\
&= -\sum_{i=1}^{s}\sum_{j=1}^{s} b_j c_j^{i-1} \frac{h^i}{(i-1)!} \frac{\partial^i S_2}{\partial^i s}(t_1, t_0)\theta(t_0 + c_j h) \\
&\qquad + \mathcal{O}\left(h^{2s+2}\right) \\
&= \sum_{i=1}^{s}\sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j) \frac{h^i}{(i-1)!} \frac{\partial^i S_2}{\partial^i s}(t_1, t_0)\theta(t_0 + c_j h) \\
&\qquad + \mathcal{O}\left(h^{2s+2}\right) \\
&= \mathcal{O}\left(h^{2s+2}\right).
\end{aligned}$$

Thus, by (2.51), we have that $y_1 - y(t_1) = \mathcal{O}\left(h^{2s+1}\right)$. □

# CHAPTER 3
# SPARK METHODS FOR ORIGINALLY INDEX 3 DAES

## 3.1 Introduction

This chapter focuses on SPARK methods for solving originally index 3 DAEs. We consider the overdetermined system of differential-algebraic equations

$$\dot{y} = v(t, y, z) \tag{3.1a}$$

$$\dot{z} = f(t, y, z) + r(t, y, \lambda) \tag{3.1b}$$

$$0 = g(t, y) \tag{3.1c}$$

$$0 = g_t(t, y) + g_y(t, y)v(t, y, z) \tag{3.1d}$$

where $y(t) \in \mathbb{R}^{n_y}$, $z(t) \in \mathbb{R}^{n_z}$, and $\lambda(t) \in \mathbb{R}^{n_g}$, and the functions

$$v : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_y}$$

$$f : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_z}$$

$$r : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_g} \to \mathbb{R}^{n_z}$$

$$g : \mathbb{R} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_g}.$$

We make the assumption that the matrix

$$g_y(t, y)v_z(t, y, z)r_\lambda(t, y, \lambda) \tag{3.2}$$

is invertible. The system

$$\frac{d}{dt}q(t, y) = v(t, y, z) \tag{3.3a}$$

$$\frac{d}{dt}p(t, y, z) = f(t, y, z) + r(t, y, \lambda) \tag{3.3b}$$

$$0 = g(t, y) \tag{3.3c}$$

$$0 = g_t(t, y) + g_y(t, y)\dot{y} \tag{3.3d}$$

is a generalization of (3.1). Both Hamiltonian and Lagrangian systems with holonomic constraints can be expressed in this form, with $q(t,y) = y$, $p(t,y,z) = z$ for Hamiltonian systems and $q(t,y) = y$, $p(t,y,z) = \nabla_z L(t,y,z)$ for Lagrangian systems. To insure existence and uniqueness of a solution, the following matrices are assumed to be invertible:

$$q_y(t,y) \tag{3.4a}$$

$$p_z(t,y,z) \tag{3.4b}$$

$$g_y(t,y)q_y(t,y)^{-1}v_z(t,y,z)p_z(t,y,z)^{-1}r_\lambda(t,y,\lambda). \tag{3.4c}$$

Under these assumptions, differentiating the left-hand sides of (3.3a) and (3.3b) gives

$$\dot{y} = q_y(t,y)^{-1}(v(t,y,z) - q_t(t,y)) \tag{3.5a}$$

$$\begin{aligned}\dot{z} = p_z(t,y,z)^{-1}\big(&f(t,y,z) + r(t,y,\lambda) - p_t(t,y,z)\\ &- p_y(t,y,z)q_y(t,y)^{-1}(v(t,y,z) - q_t(t,y))\big).\end{aligned} \tag{3.5b}$$

Taking the derivative of (3.3d), and substituting in (3.5), we arrive at

$$g_y(t,y)q_y(t,y)^{-1}v_z(t,y,z)p_z(t,y,z)^{-1}r(t,y,\lambda) + A(t,y,z) = 0$$

with the function $A(t,y,z)$ in terms of sums and products of derivatives of the functions $q$, $p$, $q_y^{-1}$, $p_z^{-1}$, $v$, $r$, and $f$. This determines uniquely the term $\lambda$ by (3.4c) and the implicit function theorem.

Following [16], we define the new variables $q$, $p$ satisfying the relations

$$q = q(t,y), \quad p = p(t,y,z).$$

By (3.4a) we can express $y$ and $z$ as functions of $t$, $q$, and $p$. Defining

$$V(t,q,p) := v(t,y(t,q,p),z(t,q,p)), \quad F(t,q,p) := f(t,y(t,q,p),z(t,q,p)),$$

$$R(t,q,\lambda) := r(t,y(t,q),\lambda), \qquad G(t,q) := g(t,y(t,q)),$$

the system (3.3) can be expressed as

$$\dot{q} = V(t, q, p) \tag{3.6a}$$

$$\dot{p} = F(t, q, p) + R(t, q, \lambda) \tag{3.6b}$$

$$0 = G(t, q) \tag{3.6c}$$

$$0 = G_t(t, q) + G_q(t, q)V(t, q, p). \tag{3.6d}$$

Thus, the system (3.3) can be equivalently expressed in the form (3.1). For the analysis presented in this chapter, we consider systems with $q(t, y) = y$ and $p(t, y, z) = z$, but the results are also valid in the more general case of (3.3).

## 3.2 SPARK Methods

We introduce here SPARK methods applied to problems with index 3 constraints.

**Definition 3.2.1.** *One step of an $(s, \widetilde{s})$-stage specialized partitioned additive Runge-Kutta (SPARK) method applied to the system (3.1) with stepsize $h$ starting at $(y_0, z_0)$ at time $t_0$ is given by the solution of the nonlinear system of equations*

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} v(t_0 + c_j h, Y_j, Z_j), \quad i = 1, \ldots, s \tag{3.7a}$$

$$\widetilde{Y}_i = y_0 + h \sum_{j=1}^{s} \bar{a}_{ij} v(t_0 + c_j h, Y_j, Z_j), \quad i = 0, \ldots, \widetilde{s} \tag{3.7b}$$

$$Z_i = z_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} f(t + c_j h, Y_j, Z_j) + h \sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \tag{3.7c}$$

$$y_1 = y_0 + h \sum_{j=1}^{s} b_j v(t_0 + c_j h, Y_j, Z_j) \tag{3.7d}$$

$$z_1 = z_0 + h \sum_{j=1}^{s} \widehat{b}_j f(t_0 + c_j h, Y_j, Z_j) + h \sum_{j=0}^{s} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) \tag{3.7e}$$

$$0 = g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \ldots, \widetilde{s} \tag{3.7f}$$

$$0 = g(t_1, y_1) \tag{3.7g}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1), \tag{3.7h}$$

*where* $t_1 := t_0 + h$. *The coefficients* $c_i$ *and* $\widetilde{c}_i$ *are given by the relations*

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad i = 1, \ldots, s, \qquad \widetilde{c}_i = \sum_{j=1}^{s} \bar{a}_{ij}, \quad i = 0, \ldots, \widetilde{s}. \tag{3.8}$$

We also define the coefficients

$$\alpha := \begin{bmatrix} A \\ b^T \end{bmatrix}, \quad \widehat{\alpha} := \begin{bmatrix} \widehat{A} \\ \widehat{b}^T \end{bmatrix}, \quad \widetilde{\alpha} := \begin{bmatrix} \widetilde{A} \\ \widetilde{b}^T \end{bmatrix}.$$

The RK coefficients are assumed to satisfy the properties

$$\bar{a}_{0j} = 0, \quad j = 1, \ldots, s, \tag{3.9a}$$

$$\bar{a}_{\widetilde{s}j} = b_j, \quad j = 1, \ldots, s, \tag{3.9b}$$

$$\sum_{j=1}^{s} \bar{a}_{ij} c_j = \sum_{j=1}^{s} \bar{a}_{ij} \sum_{k=1}^{s} \widehat{a}_{jk} = \sum_{j=1}^{s} \bar{a}_{ij} \sum_{k=0}^{\widetilde{s}} \widetilde{a}_{jk} = \frac{\widetilde{c}_i^2}{2}, \quad i = 0, \ldots, \widetilde{s}, \tag{3.9c}$$

$$\bar{A}\widetilde{A} = \begin{bmatrix} 0 & \ldots & 0 \\ & N & \end{bmatrix}, \quad \begin{bmatrix} N \\ \widetilde{b}^T \end{bmatrix} \text{ is invertible.} \tag{3.9d}$$

We also assume that the coefficients $b_i$, $\widehat{b}_i$, and $\widetilde{b}_i$ satisfy

$$\sum_{j=1}^{s} b_j = 1, \qquad \sum_{j=1}^{s} \widehat{b}_j = 1, \qquad \sum_{j=0}^{\widetilde{s}} \widetilde{b}_j = 1. \tag{3.10}$$

From (3.8) and assumptions (3.9a), we have $\widetilde{c}_0 = 0$. Similarly, from (3.9b) and (3.10), we have $\widetilde{c}_{\widetilde{s}} = 1$. Thus, condition (3.9b) gives that $\widetilde{Y}_{\widetilde{s}} = y_1$, and (3.7g) is a consequence of (3.7f) for $i = \widetilde{s}$. The existence and uniqueness of solutions to these methods is considered in [16].

### 3.2.1 Gauss-Lobatto SPARK Methods

An example of a class of SPARK methods is given by the $(s, s)$–Gauss-Lobatto SPARK methods. These are presented in [16]. In this section, we focus primarily

on results concerning the coefficients of these methods. The coefficients have the properties $\widetilde{s} = s$, $\widehat{a}_{ij} = a_{ij}$, $\widehat{b}_i = b_i$, as well as $\widetilde{c}_0 = 0$, $\widetilde{c}_s = 1$, and satisfy

$$B(2s): \quad \sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s, \tag{3.11}$$

$$C(s): \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \qquad i = 1, \ldots, s, \ k = 1, \ldots, s, \tag{3.12}$$

$$\widetilde{B}(2s): \quad \sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s. \tag{3.13}$$

The condition (3.11) is from the Gaussian quadrature with $s$ nodes, (3.12) is from the Gauss RK coefficients, and (3.13) is from the Lobatto quadrature with $s+1$ nodes. It is well known that the Gauss and Lobatto quadrature formulas satisfy $b_i \neq 0$, $c_i \neq c_j$ for $i \neq j$ and $\widetilde{b}_i \neq 0$, $\widetilde{c}_i \neq \widetilde{c}_j$ for $i \neq j$. The coefficients $\bar{a}_{ij}$ and $\widetilde{a}_{ij}$ are chosen as in [16] to satisfy

$$\overline{C}(s): \quad \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad i = 0, \ldots, s, \ k = 1, \ldots, s, \tag{3.14}$$

$$\widetilde{a}_{ij} = \widetilde{b}_j \left(1 - \frac{\bar{a}_{ji}}{b_i}\right), \qquad i = 1, \ldots, s, \ j = 0, \ldots, s. \tag{3.15}$$

The next task is to show that the methods satisfying these conditions also satisfy the hypotheses (3.9). The first two of these are shown in the following lemma.

**Lemma 3.2.2.** *The coefficients of the $(s,s)$–Gauss-Lobatto SPARK methods satisfy the assumptions (3.9a,b), i.e. $\bar{a}_{0j} = 0$ and $\bar{a}_{sj} = b_j$, as well as the conditions*

$$\widetilde{a}_{i0} = \widetilde{b}_0, \qquad \widetilde{a}_{is} = 0, \qquad i = 1, \ldots, s. \tag{3.16}$$

*Proof.* Condition (3.9a) follows from (3.14). Taking $i = 0$ in (3.14), and using

$\widetilde{c}_0 = 0$, this becomes

$$\begin{bmatrix} 1 & \dots & 1 \\ c_1 & \dots & c_s \\ \vdots & & \vdots \\ c_1^{s-1} & \dots & c_s^{s-1} \end{bmatrix} \begin{bmatrix} \bar{a}_{01} \\ \bar{a}_{02} \\ \vdots \\ \bar{a}_{0s} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Because the coefficient matrix is a Vandermonde matrix, it is invertible, providing the desired result (3.9a).

Condition (3.9b) is proved in a similar fashion. Taking $i = s$ in (3.14) and using $\widetilde{c}_s = 1$ leads to

$$\begin{bmatrix} 1 & \dots & 1 \\ c_1 & \dots & c_s \\ \vdots & & \vdots \\ c_1^{s-1} & \dots & c_s^{s-1} \end{bmatrix} \begin{bmatrix} \bar{a}_{s1} \\ \bar{a}_{s2} \\ \vdots \\ \bar{a}_{ss} \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/s \end{bmatrix}.$$

Because the coefficient matrix is invertible, there is a unique solution of $\bar{a}_{si}$ for $i = 1, \dots, s$. This system is satisfied by $\bar{a}_{sj} = b_j$, by (3.11).

Finally, we show (3.16). Taking $j = 0$ in (3.15) gives

$$\widetilde{a}_{i0} = \widetilde{b}_0 \left(1 - \frac{\bar{a}_{0i}}{b_i}\right) = \widetilde{b}_0,$$

since $\bar{a}_{0i} = 0$. Similarly, taking $j = s$ in (3.15) gives

$$\widetilde{a}_{is} = \widetilde{b}_s \left(1 - \frac{\bar{a}_{si}}{b_i}\right) = 0,$$

as $\bar{a}_{si} = b_i$. □

Before showing that the Gauss-Lobatto SPARK methods satisfy assumption (3.9c,d), we show that the coefficients satisfy other useful properties.

**Lemma 3.2.3.** *For the $(s, s)$–Gauss-Lobatto methods, the choice of the definition*

*(3.15) is equivalent to the choice of the definition*

$$\widetilde{C}(s): \quad \sum_{j=0}^{s} \widetilde{a}_{ij}\widetilde{c}_j^{k-1} = \frac{c_i^k}{k}, \quad i = 1, \ldots, s, \ k = 1, \ldots, s \tag{3.17}$$

*and*

$$\widetilde{a}_{i0} = \widetilde{b}_0, \quad i = 1, \ldots, s \tag{3.18}$$

*for the coefficients $\widetilde{a}_{ij}$, for $i = 1, \ldots, s$ and $j = 0, \ldots, s$.*

*Proof.* We first show that (3.15) implies (3.17) and (3.18). The property $\widetilde{a}_{i0} = \widetilde{b}_0$ for $i = 1, \ldots, s$ follows from (3.16). Consider the matrices

$$B := \operatorname{diag}(b_1, \ldots, b_s),$$
$$V := \begin{bmatrix} 1 & \ldots & 1 \\ c_1 & \ldots & c_s \\ \vdots & & \vdots \\ c_1^{s-1} & \ldots & c_s^{s-1} \end{bmatrix}, \quad R := \left[ \sum_{j=0}^{s} \widetilde{a}_{ij}\widetilde{c}_j^{k-1} - \frac{c_i^k}{k} \right]_{\substack{i=1,\ldots,s \\ k=1,\ldots,s}}.$$

Since $b_i \neq 0$ for the Gauss coefficients, $B$ is invertible. Because $V$ is a Vandermonde matrix, the product $VB$ is thus invertible. We will now show that $R = 0$ by showing that $VBR = 0$. Using properties of the RK coefficients, the $(m, k)$-th entry of $VBR$, for $m = 1, \ldots, s$ and $k = 1, \ldots, s$, is

$$\begin{aligned}
\sum_{i=1}^{s} b_i c_i^{m-1} \left( \sum_{j=0}^{s} \widetilde{a}_{ij}\widetilde{c}_j^{k-1} - \frac{c_i^k}{k} \right) &= \sum_{i=1}^{s}\sum_{j=0}^{s} b_i c_i^{m-1} \widetilde{a}_{ij}\widetilde{c}_j^{k-1} - \frac{1}{k}\sum_{i=1}^{s} b_i c_i^{k+m-1} \\
&= \sum_{i=1}^{s}\sum_{j=0}^{s} b_i c_i^{m-1} \widetilde{b}_j \left(1 - \frac{\bar{a}_{ji}}{b_i}\right) \widetilde{c}_j^{k-1} - \frac{1}{k} \cdot \frac{1}{(k+m)} \\
&= \sum_{i=1}^{s} b_i c_i^{m-1} \sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k-1} - \sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k-1} \sum_{i=1}^{s} \bar{a}_{ji} c_i^{m-1} - \frac{1}{k(k+m)} \\
&= \frac{1}{mk} - \frac{1}{m}\sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k+m-1} - \frac{1}{k(k+m)} \\
&= \frac{1}{mk} - \frac{1}{m(k+m)} - \frac{1}{k(k+m)} \\
&= 0.
\end{aligned}$$

It thus follows that $R = 0$, and the desired result is obtained.

Next, we will show that (3.17) and (3.18) imply (3.15). Let $M$ be the matrix with $m_{ij} := \widetilde{b}_i \bar{a}_{ij} + b_j \widetilde{a}_{ji} - \widetilde{b}_i b_j$, and denote

$$V := \begin{bmatrix} 1 & \dots & 1 \\ c_1 & \dots & c_s \\ \vdots & & \vdots \\ c_1^{s-1} & \dots & c_s^{s-1} \end{bmatrix}, \qquad \widetilde{V} := \begin{bmatrix} 1 & \dots & 1 \\ \widetilde{c}_1 & \dots & \widetilde{c}_s \\ \vdots & & \vdots \\ \widetilde{c}_1^{s-1} & \dots & \widetilde{c}_s^{s-1} \end{bmatrix}. \tag{3.19}$$

Note that $m_{0j} = \widetilde{b}_0 \bar{a}_{0j} + b_j \widetilde{a}_{j0} - \widetilde{b}_0 b_j = b_j \widetilde{b}_0 - \widetilde{b}_0 b_j = 0$, for $j = 1, \dots, s$, by (3.9a) and (3.18). It remains to show that $m_{ij} = 0$ for $i = 1, \dots, s$, and $j = 1, \dots, s$. But because $V$ and $\widetilde{V}$ are Vandermonde matrices and invertible, this is equivalent to $\widetilde{V} M V^T = 0$, or $\sum_{i=1}^{s} \sum_{j=1}^{s} \widetilde{c}_i^{k-1} m_{ij} c_j^{l-1} = 0$ for $k, l = 1, \dots, s$. Since $m_{0j} = 0$, this is equal to

$$\begin{aligned} \sum_{i=0}^{s} \sum_{j=1}^{s} \left[ \widetilde{c}_i^{k-1} \widetilde{b}_i \bar{a}_{ij} c_j^{l-1} + \widetilde{c}_i^{k-1} b_j \widetilde{a}_{ji} c_j^{l-1} - \widetilde{c}_i^{k-1} \widetilde{b}_i b_j c_j^{l-1} \right] & \\ = \frac{1}{l} \sum_{i=0}^{s} \widetilde{c}_i^{k-1} \widetilde{b}_i \widetilde{c}_i^{l} + \frac{1}{k} \sum_{j=1}^{s} b_j c_j^{l-1} c_j^{k} - \frac{1}{kl} & \\ = \frac{1}{l} \sum_{i=0}^{s} \widetilde{b}_i \widetilde{c}_i^{k+l-1} + \frac{1}{k} \sum_{j=1}^{s} b_j c_j^{k+l-1} - \frac{1}{kl} & \\ = \frac{1}{l(k+l)} + \frac{1}{k(k+l)} - \frac{1}{kl} & \\ = 0. & \end{aligned}$$

□

In the previous result, we could have assumed that $\bar{a}_{is} = 0$ for $i = 1, \dots, s$ instead of (3.18). The equivalence follows in a very similar manner to the proof presented above.

Although, by definition, the $(s, s)$–Gauss-Lobatto SPARK methods satisfy (3.14) for $k = 1, \dots, s$, it will now be shown that in fact, the condition holds for $k = s + 1$. This property will be useful for showing (3.9d) shortly.

**Lemma 3.2.4.** *The $(s,s)$–Gauss-Lobatto SPARK methods satisfy*

$$\overline{C}(s+1): \quad \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \quad i = 0, \ldots, s, \quad k = 1, \ldots, s+1. \tag{3.20}$$

*Proof.* By the definition (3.14) of the $\bar{a}_{ij}$ coefficients, only the case of $k = s+1$ need be considered. Define $\delta_i := \sum_{j=1}^{s} \bar{a}_{ij} c_j^s - \frac{\widetilde{c}_i^{s+1}}{s+1}$, for $i = 0, \ldots, s$. For $i = 0$ and $i = s$, this becomes

$$\begin{aligned}
\delta_0 &= \sum_{j=1}^{s} \bar{a}_{0j} c_j^s - \frac{\widetilde{c}_0^{s+1}}{s+1} = 0, \\
\delta_s &= \sum_{j=1}^{s} \bar{a}_{sj} c_j^s - \frac{\widetilde{c}_s^{s+1}}{s+1} = \sum_{j=1}^{s} b_j c_j^s - \frac{1}{s+1} = 0,
\end{aligned}$$

since $\bar{a}_{0j} = 0$, $\widetilde{c}_0 = 0$, $\bar{a}_{sj} = b_j$, and $\widetilde{c}_s = 1$. We denote by $\delta \in \mathbb{R}^{s-1}$ the vector $[\delta_1, \ldots, \delta_{s-1}]^T$, and show that $\delta = 0$. Define also the matrices

$$\widetilde{B} := \operatorname{diag}(\widetilde{b}_1, \widetilde{b}_2, \ldots, \widetilde{b}_{s-1}), \qquad \widetilde{V} := \begin{bmatrix} 1 & \ldots & 1 \\ \widetilde{c}_1 & \ldots & \widetilde{c}_{s-1} \\ \vdots & & \vdots \\ \widetilde{c}_1^{s-2} & \ldots & \widetilde{c}_{s-1}^{s-2} \end{bmatrix}.$$

We now consider the vector $\widetilde{V}\widetilde{B}\delta \in \mathbb{R}^{s-1}$. Since $\delta_0 = \delta_s = 0$, the $k$-th entry of this vector, for $k = 1, \ldots, s-1$, is

$$\begin{aligned}
\sum_{i=1}^{s-1} \widetilde{b}_i \widetilde{c}_i^{k-1} \delta_i &= \sum_{i=0}^{s} \widetilde{b}_i \widetilde{c}_i^{k-1} \delta_i \\
&= \sum_{i=0}^{s} \sum_{j=1}^{s} \widetilde{b}_i \widetilde{c}_i^{k-1} \bar{a}_{ij} c_j^s - \frac{1}{s+1} \sum_{i=0}^{s} \widetilde{b}_i \widetilde{c}_i^{k+s} \\
&= \sum_{i=0}^{s} \sum_{j=1}^{s} \left[ \widetilde{b}_i \widetilde{c}_i^{k-1} b_j c_j^s - \widetilde{a}_{ji} \widetilde{c}_i^{k-1} b_j c_j^s \right] - \frac{1}{s+1} \cdot \frac{1}{s+k+1} \\
&= \sum_{i=0}^{s} \widetilde{b}_i \widetilde{c}_i^{k-1} \sum_{j=1}^{s} b_j c_j^s - \frac{1}{k} \sum_{j=1}^{s} b_j c_j^{s+k} - \frac{1}{(s+1)(s+k+1)} \\
&= \frac{1}{k(s+1)} - \frac{1}{k(s+k+1)} - \frac{1}{(s+1)(s+k+1)}
\end{aligned}$$

$$= 0.$$

Here we have made use of (3.15). Therefore $\widetilde{V}\widetilde{B}\delta = 0$. But since $\widetilde{B}$ is invertible, and $\widetilde{V}$ is a Vandermonde matrix, $\delta = 0$. □

Finally, using properties (3.17) and (3.20), the Gauss-Lobatto SPARK methods can be shown to satisfy the assumptions (3.9c,d).

**Lemma 3.2.5.** *The $(s, s)$–Gauss-Lobatto SPARK methods satisfy the assumptions (3.9c,d).*

*Proof.* We will show how each of the first three terms in (3.9c) is equal to $\frac{\widetilde{c}_i^2}{2}$. The first equality follows immediately from (3.14) with $k = 2$. Using the definition $c_j = \sum_{k=1}^{s} a_{jk} (= \sum_{k=1}^{s} \widehat{a}_{jk})$, and again (3.14), we get that second equality $\sum_{j=1}^{s}\sum_{k=1}^{s} \bar{a}_{ij}a_{jk} = \frac{\widetilde{c}_i^2}{2}$. Lastly, using (3.17) with $k = 1$, we find that $\sum_{k=0}^{s} \widetilde{a}_{ij} = c_i$. Therefore, using (3.14) once more, we derive the relation $\sum_{j=1}^{s}\sum_{k=0}^{s} \bar{a}_{ij}\widetilde{a}_{jk} = \frac{\widetilde{c}_i^s}{2}$. This completes the proof of (3.9c).

The first equality of (3.9d) follows trivially from $\bar{a}_{0j} = 0$. We focus now upon the invertibility of $[N^T\ \widetilde{b}]^T \in \mathbb{R}^{(s+1)\times(s+1)}$. Because $\widetilde{a}_{is} = 0$ for $i = 1, \ldots, s$ (see (3.16)), the last column of $N$ must be 0. Thus, to show the invertibility of the matrix $[N^T\ \widetilde{b}]^T$, since $\widetilde{b}_s \neq 0$, it suffices to show the invertibility of the matrix $N^*$, with $N^* \in \mathbb{R}^{s\times s}$ equal to the matrix $N$ with its $(s+1)$-st column removed.

Define the matrix $\widetilde{V} \in \mathbb{R}^{s\times s}$ by

$$\widetilde{V} := \begin{bmatrix} 1 & \widetilde{c}_0 & \ldots & \widetilde{c}_0^{s-1} \\ 1 & \widetilde{c}_1 & \ldots & \widetilde{c}_1^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & \widetilde{c}_{s-1} & \ldots & \widetilde{c}_{s-1}^{s-1} \end{bmatrix}.$$

Because $\widetilde{V}$ is a Vandermonde matrix with $\widetilde{c}_i \neq \widetilde{c}_j$ for $i \neq j$, it is invertible. Thus, to prove that $N^*$ is invertible, we show that the product $N^*\widetilde{V}$ is invertible. Since

$\widetilde{a}_{ls} = 0$ for $l = 1, \ldots, s$, the $(i,k)$-th entry of $N^* \widetilde{V}$, for $i = 1, \ldots, s$ and $k = 1, \ldots, s$, is given by

$$\sum_{j=0}^{s-1} \sum_{l=1}^{s} \bar{a}_{il} \widetilde{a}_{lj} \widetilde{c}_j^{k-1} = \sum_{l=1}^{s} \bar{a}_{il} \sum_{j=0}^{s} \widetilde{a}_{lj} \widetilde{c}_j^{k-1} = \frac{1}{k} \sum_{l=1}^{s} \bar{a}_{il} c_l^k = \frac{\widetilde{c}_i^{k+1}}{k(k+1)}.$$

This follows successively from (3.17) and (3.20). This product $N^* \widetilde{V}$ can be expressed as

$$N^* \widetilde{V} = \begin{bmatrix} \widetilde{c}_1^2 & \widetilde{c}_1^3 & \ldots & \widetilde{c}_1^{s+1} \\ \widetilde{c}_2^2 & \widetilde{c}_2^3 & \ldots & \widetilde{c}_2^{s+1} \\ \vdots & \vdots & & \vdots \\ \widetilde{c}_s^2 & \widetilde{c}_s^3 & \ldots & \widetilde{c}_s^{s+1} \end{bmatrix} \begin{bmatrix} \frac{1}{1\cdot 2} & & & O \\ & \frac{1}{2\cdot 3} & & \\ & & \ddots & \\ O & & & \frac{1}{s(s+1)} \end{bmatrix},$$

or

$$\begin{bmatrix} \widetilde{c}_1^2 & & & O \\ & \widetilde{c}_2^2 & & \\ & & \ddots & \\ O & & & \widetilde{c}_s^2 \end{bmatrix} \begin{bmatrix} 1 & \widetilde{c}_1 & \ldots & \widetilde{c}_1^{s-1} \\ 1 & \widetilde{c}_2 & \ldots & \widetilde{c}_2^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & \widetilde{c}_s & \ldots & \widetilde{c}_s^{s-1} \end{bmatrix} \begin{bmatrix} \frac{1}{1\cdot 2} & & & O \\ & \frac{1}{2\cdot 3} & & \\ & & \ddots & \\ O & & & \frac{1}{s(s+1)} \end{bmatrix}.$$

Because each of these three matrices is invertible, their product is invertible. This shows that $N^*$ must be invertible, and thus $[N^T \ \widetilde{b}]^T$ is invertible. □

The Gauss-Lobatto SPARK methods applied to problems with originally index 3 constraints are symmetric methods. Because of this, their local order must be even. This will play a role later in the derivation of their order.

**Theorem 3.2.6.** *Assume the initial conditions $(y_0, z_0)$ at time $t_0$ are consistent, i.e.,*

$$g(t_0, y_0) = 0$$

$$g_t(t_0, y_0) + g_y(t_0, y_0) v(t_0, y_0, z_0) = 0.$$

*Then the Gauss-Lobatto SPARK methods applied to originally index 3 DAEs are*

*symmetric.*

*Proof.* First, we rewrite (3.7g,h). Because the initial conditions are assumed consistent, these can be written as

$$0 = g(t_1, y_1) + g(t_0, y_0) \tag{3.21a}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) + g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) \tag{3.21b}$$

Using the method (3.7a-f) and (3.21) with Gauss-Lobatto coefficients, we apply the method with stepsize $-h$ starting at time $t_1$ to obtain the system

$$Y_i = y_1 - h\sum_{j=1}^{s} a_{ij}v(t_1 - c_j h, Y_j, Z_j), \quad i = 1, \ldots, s \tag{3.22a}$$

$$\widetilde{Y}_i = y_1 - h\sum_{j=1}^{s} \bar{a}_{ij}v(t_1 - c_j h, Y_j, Z_j), \quad i = 0, \ldots, s \tag{3.22b}$$

$$\begin{aligned} Z_i = z_1 &- h\sum_{j=1}^{s} a_{ij}f(t_1 - c_j h, Y_j, Z_j) \\ &- h\sum_{j=0}^{s} \widetilde{a}_{ij}r(t_1 - \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \end{aligned} \tag{3.22c}$$

$$y_0 = y_1 - h\sum_{j=1}^{s} b_j v(t_1 - c_j h, Y_j, Z_j) \tag{3.22d}$$

$$z_0 = z_1 - h\sum_{j=1}^{s} b_j f(t_1 - c_j h, Y_j, Z_j) - h\sum_{j=0}^{s} \widetilde{b}_j r(t_1 - \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) \tag{3.22e}$$

$$0 = g(t_1 - \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{3.22f}$$

$$0 = g(t_0, y_0) + g(t_1, y_1) \tag{3.22g}$$

$$0 = g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) + g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{3.22h}$$

Using the definition $t_1 = t_0 + h$, (3.22d,e) become

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j v(t_0 + (1 - c_j)h, Y_j, Z_j)$$

$$z_1 = z_0 + h\sum_{j=1}^{s} b_j f(t_0 + (1 - c_j)h, Y_j, Z_j) + h\sum_{j=0}^{s} \widetilde{b}_j r(t_0 + (1 - \widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j).$$

Substituting these back into (3.22) and applying the consistency of the initial conditions $(y_0, z_0)$ at time $t_0$ gives

$$Y_i = y_0 + h\sum_{j=1}^{s}(b_j - a_{ij})v(t_0 + (1-c_j)h, Y_j, Z_j), \quad i = 1, \ldots, s \tag{3.23a}$$

$$\widetilde{Y}_i = y_0 + h\sum_{j=1}^{s}(b_i - \bar{a}_{ij})v(t_0 + (1-c_j)h, Y_j, Z_j), \quad i = 0, \ldots, s \tag{3.23b}$$

$$\begin{aligned} Z_i = z_0 &+ h\sum_{j=1}^{s}(b_j - a_{ij})f(t_0 + (1-c_j)h, Y_j, Z_j) \\ &+ h\sum_{j=0}^{s}(\widetilde{b}_j - \widetilde{a}_{ij})r(t_0 + (1-\widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \end{aligned} \tag{3.23c}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j v(t_0 + (1-c_j)h, Y_j, Z_j) \tag{3.23d}$$

$$\begin{aligned} z_1 = z_0 &+ h\sum_{j=1}^{s} b_j f(t_0 + (1-c_j)h, Y_j, Z_j) \\ &+ h\sum_{j=0}^{s} \widetilde{b}_j r(t_0 + (1-\widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j) \end{aligned} \tag{3.23e}$$

$$0 = g(t_0 + (1-\widetilde{c}_i)h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{3.23f}$$

$$0 = g(t_1, y_1) \tag{3.23g}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{3.23h}$$

Lastly, we would like to reformulate (3.23) in the format of (3.7). To do so, we must reindex each $Y_i$, $\widetilde{Y}_i$, $Z_i$, and $\Lambda_i$ so as to preserve the usual ordering of the $c_i$, $\widetilde{c}_i$ coefficients. This results in the system

$$Y_i^* = y_0 + h\sum_{j=1}^{s} a_{ij}^* v(t_0 + c_j^* h, Y_j^*, Z_j^*), \quad i = 1, \ldots, s \tag{3.24a}$$

$$\widetilde{Y}_i^{\,*} = y_0 + h\sum_{j=1}^{s} \bar{a}_{ij}^* v(t_0 + c_j^* h, Y_j^*, Z_j^*), \quad i = 0, \ldots, s \tag{3.24b}$$

$$\begin{aligned} Z_i^* &= z_0 + h\sum_{j=1}^{s} a_{ij}^* f(t_0 + c_j^* h, Y_j^*, Z_j^*) \\ &\quad + h\sum_{j=0}^{s} \widetilde{a}_{ij}^* r(t_0 + \widetilde{c}_j^* h, \widetilde{Y}_j^{\,*}, \Lambda_j^*), \quad i = 1, \ldots, s \end{aligned} \tag{3.24c}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j^* v(t_0 + c_j^* h, Y_j^*, Z_j^*) \tag{3.24d}$$

$$z_1 = z_0 + h\sum_{j=1}^{s} b_j^* f(t_0 + c_j^* h, Y_j^*, Z_j^*) + h\sum_{j=0}^{s} \widetilde{b}_j^* r(t_0 + \widetilde{c}_j^* h, \widetilde{Y}_j^{\,*}, \Lambda_j^*) \tag{3.24e}$$

$$0 = g(t_0 + \widetilde{c}_i^* h, \widetilde{Y}_i^{\,*}), \quad i = 0, \ldots, s \tag{3.24f}$$

$$0 = g(t_1, y_1) \tag{3.24g}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1) v(t_1, y_1, z_1) \tag{3.24h}$$

where the internal stages are defined by

$$Y_i^* := Y_{s+1-i}, \quad \widetilde{Y}_i^{\,*} := \widetilde{Y}_{s-i}, \quad Z_i^* := Z_{s+1-i}, \quad \Lambda_i^* := \Lambda_{s-i}$$

and where the RK coefficients are defined by

$$c_i^* = 1 - c_{s+1-i}, \quad \widetilde{c}_i^* = 1 - \widetilde{c}_{s-i} \tag{3.25a}$$

$$b_i^* = b_{s+1-i}, \quad \widetilde{b}_i^* = \widetilde{b}_{s-i} \tag{3.25b}$$

$$\begin{aligned} &a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j}, \quad \widetilde{a}_{ij}^* = \widetilde{b}_{s-j} - \widetilde{a}_{s+1-i,s-j}, \\ &\bar{a}_{ij}^* = b_{s+1-j} - \bar{a}_{s-i,s+1-j}. \end{aligned} \tag{3.25c}$$

The method given by (3.24) is the *adjoint* of (3.7) with Gauss-Lobatto coefficients. To show the symmetry of the method, we must show that the method given by (3.24) is the same as the method (3.7) with Gauss-Lobatto coefficients, i.e. we must show

$$\begin{aligned} &c_i^* = c_i, \quad \widetilde{c}_i^* = \widetilde{c}_i, \quad b_i^* = b_i, \quad \widetilde{b}_i^* = \widetilde{b}_i, \\ &a_{ij}^* = a_{ij}, \quad \widetilde{a}_{ij}^* = \widetilde{a}_{ij}, \quad \bar{a}_{ij}^* = \bar{a}_{ij}. \end{aligned}$$

The first line of equalities come immediately from the symmetry of the Gauss and

the Lobatto coefficients, as does the equality $a^*_{ij} = a_{ij}$. Because the coefficients $\bar{a}_{ij}$ are determined by the system (3.14), we can show that $\bar{a}^*_{ij} = \bar{a}_{ij}$ by showing that

$$\sum_{j=1}^{s} \bar{a}^*_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^{\,k}}{k}, \qquad i = 0, \ldots, s,\ k = 1, \ldots, s.$$

But by applying $B(s)$ and $\overline{C}(s)$, this is true because

$$\begin{aligned}
\sum_{j=1}^{s} \bar{a}^*_{ij} c_j^{k-1} &= \sum_{j=1}^{s} (b_j - \bar{a}_{s-i,s+1-j}) c_j^{k-1} \\
&= \frac{1}{k} - \sum_{j=1}^{s} \bar{a}_{s-i,s+1-j} (1 - c_{s+1-j})^{k-1} \\
&= \frac{1}{k} - \sum_{j=1}^{s} \sum_{l=0}^{k-1} \bar{a}_{s-i,s+1-j} \binom{k-1}{l} (-1)^l \ c_{s+1-j}^l \\
&= \frac{1}{k} - \sum_{l=0}^{k-1} \binom{k-1}{l} (-1)^l \frac{\widetilde{c}_{s-i}^{\,l+1}}{l+1} \\
&= \frac{1}{k} \left( 1 + \sum_{l=1}^{k} \binom{k}{l} (-1)^l \widetilde{c}_{s-i}^{\,l} \right) \\
&= \frac{1}{k} (1 - \widetilde{c}_{s-i})^k \\
&= \frac{\widetilde{c}_i^{\,k}}{k}.
\end{aligned}$$

To show that $\widetilde{a}^*_{ij} = \widetilde{a}_{ij}$, we use (3.15) to get

$$\begin{aligned}
\widetilde{a}_{ij} - \widetilde{a}^*_{ij} &= \widetilde{a}_{ij} + \widetilde{a}_{s-i+1,s-j} - \widetilde{b}_{s-j} \\
&= \widetilde{b}_j \left( 1 - \frac{\bar{a}_{ji}}{b_i} \right) + \widetilde{b}_{s-j} \left( 1 - \frac{\bar{a}_{s-j,s-i+1}}{b_{s-i+1}} \right) - \widetilde{b}_{s-j} \\
&= \widetilde{b}_j \left( 1 - \frac{\bar{a}_{ji}}{b_i} \right) + \widetilde{b}_j \left( 1 - \frac{b_i - \bar{a}_{ji}}{b_i} \right) - \widetilde{b}_j \\
&= 0.
\end{aligned}$$

Therefore, the Gauss-Lobatto SPARK methods applied to originally index 3 DAEs are symmetric. □

### 3.3 Analysis of Existing Literature

The Gauss-Lobatto SPARK methods for overdetermined systems of originally index 3 DAEs have been analyzed in detail by Jay [16]. However, the proof presented there for the local error of the methods is incorrect, due to an overlooked technical detail. More specifically, the defect $\mu(t)$ defined in (5.6c) of [16] as

$$\dot{Z}(t) = f(Y(t), Z(t)) + r(\widetilde{Y}(t), \Lambda(t)) + \mu(t)$$

does not satisfy the condition

$$\mu(t_0 + c_i h) = 0, \quad i = 1, \ldots, s,$$

as claimed. The Gauss-Lobatto SPARK methods are therefore not equivalent to a class of collocation methods as stated, but rather to a class of discontinuous collocation type methods. Due to this oversight, the proof of the results of [16] is not straightforward and will be done in the remainder of this chapter.

To correct the proof, we introduce two new internal stages for the SPARK methods (3.7) to handle the Gauss and the Lobatto discretizations separately. We express (3.7c) equivalently as

$$Z_i^f = z_0 + h \sum_{j=1}^{s} a_{ij} f(t_0 + c_i h, Y_j, Z_j), \quad i = 1, \ldots, s, \tag{3.26a}$$

$$Z_i^r = h \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_i h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s. \tag{3.26b}$$

Notice that $Z_i = Z_i^f + Z_i^r$, $i = 1, \ldots, s$. We will also consider two additional differential equations

$$\dot{z}^f = f(t, y, z), \quad z^f(t_0) = z_0 \tag{3.27a}$$

$$\dot{z}^r = r(t, y, \lambda), \quad z^r(t_0) = 0. \tag{3.27b}$$

As a result of these definitions, we will have $z(t) = z^f(t) + z^r(t)$.

In this chapter, a corrected argument showing the local error of the Gauss-Lobatto SPARK methods is presented based upon discontinuous collocation techniques which can be found in [8]. We show the equivalence of the Gauss-Lobatto methods to a class of discontinuous collocation methods, present a perturbation analysis, and finally analyze the local error of the methods.

## 3.4 Influence of Perturbations

We first consider the influence of perturbations on the solution of the method (3.7). Consider the perturbed system

$$\widehat{Y}_i = \widehat{y}_0 + h\sum_{j=1}^{s} a_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\delta_i^y, \quad i = 1, \ldots, s \tag{3.28a}$$

$$\widehat{\widetilde{Y}}_i = \widehat{y}_0 + h\sum_{j=1}^{s} \bar{a}_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\widetilde{\delta}_i^y, \quad i = 0, \ldots, \widetilde{s} \tag{3.28b}$$

$$\begin{aligned}\widehat{Z}_i = \widehat{z}_0 + h\sum_{j=1}^{s} \widehat{a}_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) \\ + h\delta_i^z, \quad i = 1, \ldots, s\end{aligned} \tag{3.28c}$$

$$\widehat{y}_1 = \widehat{y}_0 + h\sum_{j=1}^{s} b_j v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\delta_{s+1}^y \tag{3.28d}$$

$$\widehat{z}_1 = \widehat{z}_0 + h\sum_{j=1}^{s} \widehat{b}_j f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_{s+1}^z \tag{3.28e}$$

$$0 = g(t_0 + \widetilde{c}_i h, \widehat{\widetilde{Y}}_i) + h\delta_i^\lambda, \quad i = 0, \ldots, \widetilde{s} \tag{3.28f}$$

$$0 = g_t(t_1, \widehat{y}_1) + g_y(t_1, \widehat{y}_1) v(t_1, \widehat{y}_1, \widehat{z}_1) + \delta_{\widetilde{s}+1}^\lambda. \tag{3.28g}$$

Consider also the perturbed form of (3.26),

$$\widehat{Z}_i^f = \widehat{z}_0 + h\sum_{j=1}^{s} \widehat{a}_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\delta_i^f, \quad i = 1, \ldots, s \tag{3.29a}$$

$$\widehat{Z}_i^r = h\sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_i^r, \quad i = 1, \ldots, s. \tag{3.29b}$$

We examine the influence of the perturbations

$$\begin{aligned}
\delta^y &:= [\delta_1^{yT}, \ldots, \delta_{s+1}^{y\ T}]^T, \quad \delta^z := [\delta_1^{zT}, \ldots, \delta_{s+1}^{z\ T}]^T, \\
\widetilde{\delta}^y &:= [\widetilde{\delta}_0^{yT}, \ldots, \widetilde{\delta}_{\tilde{s}}^{y\ T}]^T, \quad \delta^\lambda := [\delta_0^{\lambda\ T}, \ldots, \delta_{\tilde{s}+1}^{\lambda\ T}]^T, \\
\delta^f &:= [\delta_1^{fT}, \ldots, \delta_s^{f\ T}]^T, \quad \delta^r := [\delta_1^{rT}, \ldots, \delta_s^{rT}]^T.
\end{aligned}$$

For simplicity, the following notations are introduced

$$\begin{aligned}
Y &:= [Y_1^T, Y_2^T, \ldots, Y_s^T]^T, \quad \widehat{Y} := [\widehat{Y}_1^T, \widehat{Y}_2^T, \ldots, \widehat{Y}_s^T]^T \\
Z &:= [Z_1^T, Z_2^T, \ldots, Z_s^T]^T, \quad \widehat{Z} := [\widehat{Z}_1^T, \widehat{Z}_2^T, \ldots, \widehat{Z}_s^T]^T \\
\widetilde{Y} &:= [\widetilde{Y}_0^T, \widetilde{Y}_1^T, \ldots, \widetilde{Y}_{\tilde{s}}^T]^T, \quad \widehat{\widetilde{Y}} := [\widehat{\widetilde{Y}}_0^T, \widehat{\widetilde{Y}}_1^T, \ldots, \widehat{\widetilde{Y}}_{\tilde{s}}^T]^T \\
\Lambda &:= [\Lambda_0^T, \Lambda_1^T, \Lambda_2^T, \ldots, \Lambda_{\tilde{s}}^T]^T, \quad \widehat{\Lambda} := [\widehat{\Lambda}_0^T, \widehat{\Lambda}_1^T, \widehat{\Lambda}_2^T, \ldots, \widehat{\Lambda}_{\tilde{s}}^T]^T \\
\Delta Y_i &:= \widehat{Y}_i - Y_i, \quad \Delta Z_i := \widehat{Z}_i - Z_i, \quad \Delta \widetilde{Y}_i := \widehat{\widetilde{Y}}_i - \widetilde{Y}_i, \quad \Delta \Lambda_i := \widehat{\Lambda}_i - \Lambda_i \\
\Delta y_1 &:= \widehat{y}_1 - y_1, \quad \Delta z_1 := \widehat{z}_1 - z_1, \quad \Delta y_0 := \widehat{y}_0 - y_0, \quad \Delta z_0 := \widehat{z}_0 - z_0 \\
\Delta Y &:= \widehat{Y} - Y, \quad \Delta Z := \widehat{Z} - Z, \quad \Delta \widetilde{Y} := \widehat{\widetilde{Y}} - \widetilde{Y}, \quad \Delta \Lambda := \widehat{\Lambda} - \Lambda \\
\Delta \overline{Y} &:= [\widehat{Y}^T - Y^T, \widehat{y}_1^T - y_1^T]^T, \quad \Delta \overline{Z} := [\widehat{Z}^T - Z^T, \widehat{z}_1^T - z_1^T]^T.
\end{aligned}$$

We also define $||Y|| := \max_i\{||Y_i||\}$, $||\Lambda|| := \max_i\{||\Lambda_i||\}$, etc. We will make use of the coefficient matrices

$$\alpha := \begin{bmatrix} A \\ b^T \end{bmatrix}, \quad \widehat{\alpha} := \begin{bmatrix} \widehat{A} \\ \widehat{b}^T \end{bmatrix}, \quad \widetilde{\alpha} := \begin{bmatrix} \widetilde{A} \\ \widetilde{b}^T \end{bmatrix}, \quad \bar{\alpha} := \begin{bmatrix} \bar{A}^* & 0_s \\ 0_s^T & 1 \end{bmatrix},$$

where $\bar{A}^* \in \mathbb{R}^{\tilde{s} \times s}$ equals $\bar{A}$ with the first row removed.

**Theorem 3.4.1.** *Suppose the initial conditions satisfy*

$$\begin{aligned}
g(t_0, y_0) &= o(h^2) \\
g_t(t_0, y_0) + g_y(t_0, y_0) v(t_0, y_0, z_0) &= o(h).
\end{aligned}$$

*Further, assume that the matrix* $g_y(t,y)v_z(t,y,z)r_\lambda(t,y,\lambda)$ *is invertible in a neighborhood of* $(y_0,z_0,\lambda_0)$*. Lastly, assume*

$$\begin{aligned}
&\Delta y_0 = \mathcal{O}(h^3), \quad \Delta z_0 = \mathcal{O}(h^2), \quad \widehat{\Lambda}_k - \lambda_0 = \mathcal{O}(h)\\
&\delta_i^y = \mathcal{O}(h), \quad \widetilde{\delta}_k^y = \mathcal{O}(h^2), \quad \delta_i^z = \mathcal{O}(h),\\
&\delta_l^\lambda = \mathcal{O}(h^3), \quad \delta_j^f = \mathcal{O}(1), \quad \delta_j^r = \mathcal{O}(1),
\end{aligned}$$

*for* $i = 1,\dots,s+1$, $j = 1,\dots,s$, $k = 0,\dots,\widetilde{s}$, *and* $l = 0,\dots,\widetilde{s}+1$. *Then for* $|h| \le h_0$*, we have the bounds*

$$\begin{aligned}
\Delta Y_i = \Delta y_0 + \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y||\\
&+ h||\delta^\lambda|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\eta_0||)
\end{aligned} \tag{3.30a}$$

$$\begin{aligned}
\Delta \widetilde{Y}_i = \Delta y_0 + \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y||\\
&+ h||\delta^\lambda|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\eta_0||)
\end{aligned} \tag{3.30b}$$

$$\begin{aligned}
\Delta Z_i = \Delta z_0 + \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y||\\
&+ ||\delta^\lambda|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\eta_0||)
\end{aligned} \tag{3.30c}$$

$$\begin{aligned}
\Delta y_1 = \Delta y_0 + \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y||\\
&+ h||\delta^\lambda|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\eta_0||)
\end{aligned} \tag{3.30d}$$

$$\begin{aligned}
\Delta z_1 = \Delta z_0 + \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y||\\
&+ ||\delta^\lambda|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\eta_0||)
\end{aligned} \tag{3.30e}$$

$$\begin{aligned}
h\Delta\Lambda_i = \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y||\\
&+ ||\delta^\lambda|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\eta_0||).
\end{aligned} \tag{3.30f}$$

*In addition, we have the bounds*

$$\begin{aligned}
\Delta Z_i^f = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y||\\
&+ h||\delta^\lambda|| + h||\delta^f|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\eta_0||\big)
\end{aligned} \tag{3.31a}$$

$$\begin{aligned}
\Delta Z_i^r = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y||\\
&+ ||\delta^\lambda|| + h||\delta^r|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\eta_0||\big),
\end{aligned} \tag{3.31b}$$

*where we have used the notation*

$$\begin{aligned}\eta_0 := g_y(t_0,y_0)v_y(t_0,y_0,z_0)\Delta y_0 + g_y(t_0,y_0)v_z(t_0,y_0,z_0)\Delta z_0 \\ + g_{ty}(t_0,y_0)\Delta y_0 + g_{yy}(t_0,y_0)(\Delta y_0, v(t_0,y_0,z_0)).\end{aligned} \tag{3.32}$$

*Proof.* The proof presented here uses ideas presented in [12] and [22]. We begin by showing that (3.30b) holds for $\Delta\widetilde{Y}_0$. This follows immediately from (3.28b) and (3.7b), as

$$\Delta\widetilde{Y}_0 = \Delta y_0 + h\widetilde{\delta}_0^y = \mathcal{O}(||\Delta y_0|| + h||\widetilde{\delta}^y||).$$

Subtracting (3.7) from (3.28), and expanding around $(Y_j, Z_j, \Lambda_j)$ gives

$$\begin{aligned}\Delta Y_i = \Delta y_0 + h\sum_{j=1}^{s} a_{ij}(v_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j \\ + v_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) + h\delta_i^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)\end{aligned} \tag{3.33a}$$

$$\begin{aligned}\Delta\widetilde{Y}_i = \Delta y_0 + h\sum_{j=1}^{s} \bar{a}_{ij}(v_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j \\ + v_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) + h\widetilde{\delta}_i^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)\end{aligned} \tag{3.33b}$$

$$\begin{aligned}\Delta Z_i = \Delta z_0 + h\sum_{j=1}^{s} \widehat{a}_{ij}(f_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j + f_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) \\ + h\sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij}(r_y(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j) \\ + h\delta_i^z + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2\right)\end{aligned} \tag{3.33c}$$

$$\begin{aligned}\Delta y_1 = \Delta y_0 + h\sum_{j=1}^{s} b_j(v_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j \\ + v_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) + h\delta_{s+1}^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)\end{aligned} \tag{3.33d}$$

$$\begin{aligned}\Delta z_1 = \Delta z_0 + h\sum_{j=1}^{s} \widehat{b}_j(f_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j + f_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) \\ + h\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j(r_y(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j) \\ + h\delta_{s+1}^z + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2\right)\end{aligned} \tag{3.33e}$$

$$
\begin{aligned}
0 = &\frac{1}{h} g_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_i)\Delta y_0 \\
&+ g_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_i) \sum_{j=1}^{s} \bar{a}_{ij} v_y(t_0 + c_j h, Y_j, Z_j)\Delta Y_j \\
&+ g_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_i) \sum_{j=1}^{s} \bar{a}_{ij} v_z(t_0 + c_j h, Y_j, Z_j)\Delta Z_j \\
&+ \mathcal{O}(h||\Delta y_0|| + ||\Delta Y||^2 + ||\Delta Z||^2 + ||\widetilde{\delta}^y|| + ||\delta^\lambda||)
\end{aligned} \tag{3.33f}
$$

$$
\begin{aligned}
0 = &\, g_{ty}(t_1, y_1)\Delta y_1 + h g_y(t_1, y_1) v_y(t_1, y_1, z_1)\Delta y_1 \\
&+ g_y(t_1, y_1) v_z(t_1, y_1, z_1)\Delta z_1 + g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) \\
&+ \delta^\lambda_{\widetilde{s}+1} + \mathcal{O}\left(||\Delta y_1||^2 + ||\Delta z_1||^2\right)
\end{aligned} \tag{3.33g}
$$

Note that in (3.33f) we have divided both sides by $h$. We can express this more compactly using tensor notation. We rewrite (3.33a-f) as

$$
\begin{aligned}
\Delta\overline{Y} = &\, \mathbb{1}_{s+1} \otimes \Delta y_0 + h(\alpha \otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z) \\
&+ \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\delta^y||\right)
\end{aligned} \tag{3.34a}
$$

$$
\begin{aligned}
\Delta\widetilde{Y} = &\, \mathbb{1}_{\widetilde{s}+1} \otimes \Delta y_0 + h(\bar{A}^* \otimes I_{n_z})(\{v_y\}\Delta Y + \{v_z\}\Delta Z) \\
&+ \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\widetilde{\delta}^y||\right)
\end{aligned} \tag{3.34b}
$$

$$
\begin{aligned}
\Delta\overline{Z} = &\, \mathbb{1}_{s+1} \otimes \Delta z_0 + h(\widehat{\alpha} \otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z) \\
&+ h(\widetilde{\alpha} \otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda) \\
&+ \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2 + h||\delta^z||\right)
\end{aligned} \tag{3.34c}
$$

$$
\begin{aligned}
0 = &\frac{1}{h}\{g_y\}\left(\mathbb{1}_{\widetilde{s}} \otimes \Delta y_0 + h(\bar{A}^* \otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right) \\
&+ \mathcal{O}\left(h||\Delta y_0|| + ||\Delta Y||^2 + ||\Delta Z||^2 + ||\widetilde{\delta}^y|| + ||\delta^\lambda||\right).
\end{aligned} \tag{3.34d}
$$

Here and in the remainder of this proof, we use the notation

$$
\begin{aligned}
\{v_y\} &:= \text{blockdiag}(v_y(t_0 + c_1 h, Y_1, Z_1), \ldots, v_y(t_0 + c_s h, Y_s, Z_s)) \\
\{g_y\} &:= \text{blockdiag}(g_y(t_0 + \widetilde{c}_1 h, \widetilde{Y}_1), \ldots, g_y(t_0 + \widetilde{c}_{\widetilde{s}} h, \widetilde{Y}_{\widetilde{s}})) \\
\langle v_y\rangle &:= \text{blockdiag}(v_y(t_0 + c_1 h, Y_1, Z_1), \ldots, v_y(t_0 + c_s h, Y_s, Z_s), v_y(t_1, y_1, z_1)) \\
[r_y] &:= \text{blockdiag}(r_y(t_0 + \widetilde{c}_0 h, \widetilde{Y}_0, \Lambda_0), \ldots, r_y(t_0 + \widetilde{c}_{\widetilde{s}} h, \widetilde{Y}_{\widetilde{s}}, \Lambda_{\widetilde{s}})) \\
\widetilde{[g_y]} &:= \text{blockdiag}(g_y(t_0 + \widetilde{c}_1 h, \widetilde{Y}_1), \ldots, g_y(t_0 + \widetilde{c}_{\widetilde{s}} h, \widetilde{Y}_{\widetilde{s}}), g_y(t_1, y_1))
\end{aligned}
$$

and so on. We will now solve for the term $h\Delta\Lambda$. Combining (3.34d) with (3.33g) gives

$$\begin{aligned} 0 = {} & \widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y}) \left( \langle v_y \rangle \Delta \overline{Y} + \langle v_z \rangle \Delta \overline{Z} \right) \\ & + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_{\widetilde{s}} \otimes \Delta y_0) \\ g_{ty}(t_1, y_1)\Delta y_1 + g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) \end{bmatrix} \\ & + \mathcal{O}\left( h||\Delta y_0|| + ||\Delta \overline{Y}||^2 + ||\Delta \overline{Z}||^2 + ||\widetilde{\delta}^y|| + ||\delta^\lambda|| \right). \end{aligned} \tag{3.35}$$

Substituting (3.34a,c) into (3.35), we arrive at

$$\begin{aligned} 0 = {} & \widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y}) \Bigg( \langle v_y \rangle \left( \mathbb{1}_{s+1} \otimes \Delta y_0 + h(\alpha \otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z) \right) \\ & + \langle v_z \rangle \big( \mathbb{1}_{s+1} \otimes \Delta z_0 + h(\widehat{\alpha} \otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z) \\ & + h(\widetilde{\alpha} \otimes I_{n_z})([r_y]\Delta \widetilde{Y} + [r_\lambda]\Delta\Lambda) \big) \Bigg) \\ & + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_{\widetilde{s}} \otimes \Delta y_0) \\ g_{ty}(t_1, y_1)\Delta y_1 + g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) \end{bmatrix} \\ & + \mathcal{O}(h||\Delta y_0|| + ||\Delta \overline{Y}||^2 + ||\Delta \overline{Z}||^2 + h||\Delta\Lambda||^2 \\ & \qquad + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda||). \end{aligned}$$

This expression can be solved for $h\Delta\Lambda$. The expression can be rewritten as

$$\begin{aligned} & -\widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y})\langle v_z \rangle(\widetilde{\alpha} \otimes I_{n_z})[r_\lambda](h\Delta\Lambda) = \\ & \widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y}) \Bigg( \langle v_y \rangle(\mathbb{1}_{s+1} \otimes \Delta y_0) + h\langle v_y \rangle(\alpha \otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z) \\ & \quad + \langle v_z \rangle(\mathbb{1}_{s+1} \otimes \Delta z_0) + h\langle v_z \rangle(\widehat{\alpha} \otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z) \\ & \quad + h\langle v_z \rangle(\widetilde{\alpha} \otimes I_{n_z})[r_y]\Delta \widetilde{Y} \Bigg) \\ & + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_{\widetilde{s}} \otimes \Delta y_0) \\ g_{ty}(t_1, y_1)\Delta y_1 + g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) \end{bmatrix} \\ & + \mathcal{O}(h||\Delta y_0|| + ||\Delta \overline{Y}||^2 + ||\Delta \overline{Z}||^2 + h||\Delta\Lambda||^2 \\ & \qquad + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda||). \end{aligned}$$

For $i = 1, \ldots, s$ and $j = 0, \ldots, s$, the coefficient matrix multiplying $h\Delta\Lambda$ satisfies

$$\begin{aligned}
&\widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha} \otimes I_{n_z})[r_\lambda] \\
&= \begin{bmatrix} \left[\sum_{k=1}^{s} \bar{a}_{ik}\widetilde{a}_{kj} g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) v_z(t_0 + c_k h, Y_k, Z_k) r_\lambda(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\right] \\ \left[\widetilde{b}_j g_y(t_1, y_1) v_z(t_1, y_1, z_1) r_\lambda(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\right] \end{bmatrix} \\
&= \begin{bmatrix} \bar{A}^* \widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y(t_0, y_0) v_z(t_0, y_0, z_0) r_\lambda(t_0, y_0, \lambda_0) + \mathcal{O}(h)
\end{aligned}$$

From the properties of the Gauss-Lobatto SPARK methods, and for $h$ sufficiently small, this matrix is invertible. Taking

$$D := -\left(\widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha} \otimes I_{n_z})[r_\lambda]\right)^{-1},$$

and after a little shuffling, we arrive at

$$\begin{aligned}
h\Delta\Lambda = D\widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y})\Big(&\langle v_y\rangle(\mathbb{1}_{s+1} \otimes \Delta y_0) \\
&+ h\langle v_y\rangle(\alpha \otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z) \\
&+ \langle v_z\rangle(\mathbb{1}_{s+1} \otimes \Delta z_0) + h\langle v_z\rangle(\widehat{\alpha} \otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z) \\
&+ h\langle v_z\rangle(\widetilde{\alpha} \otimes I_{n_z})[r_y]\Delta\widetilde{Y}\Big) \\
&+ D\begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_{\widetilde{s}} \otimes \Delta y_0) \\ g_{ty}(t_1, y_1)\Delta y_1 + g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) \end{bmatrix} \\
&+ \mathcal{O}(h||\Delta y_0|| + ||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + h||\Delta\Lambda||^2 \\
&\qquad + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda||).
\end{aligned} \tag{3.36}$$

Several terms can be reexpressed as

$$\begin{aligned}
\frac{1}{h} g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\Delta y_0 &= \frac{1}{h} g_y(t_0 + \widetilde{c}_i h, y_0 + (\widetilde{Y}_i - y_0))\Delta y_0 \\
&= \frac{1}{h} g_y(t_0, y_0)\Delta y_0 + \widetilde{c}_i g_{ty}(t_0, y_0)\Delta y_0 \\
&\qquad + \frac{1}{h} g_{yy}(t_0, y_0)(\Delta y_0, \widetilde{Y}_i - y_0) + \mathcal{O}(h||\Delta y_0||) \\
&= \frac{1}{h} g_y(t_0, y_0)\Delta y_0 + \widetilde{c}_i g_{ty}(t_0, y_0)\Delta y_0
\end{aligned}$$

$$+ \widetilde{c}_i g_{yy}(t_0, y_0)(\Delta y_0, v(t_0, y_0, z_0)) + \mathcal{O}(h||\Delta y_0||)$$

$$g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) \sum_{j=1}^{s} \bar{a}_{ij} v_y(t_0 + c_j h, Y_j, Z_j)\Delta y_0 = \widetilde{c}_i g_y(t_0, y_0) v_y(t_0, y_0, z_0)\Delta y_0$$
$$+ \mathcal{O}(h||\Delta y_0||)$$

$$g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) \sum_{j=1}^{s} \bar{a}_{ij} v_z(t_0 + c_j h, Y_j, Z_j)\Delta z_0 = \widetilde{c}_i g_y(t_0, y_0) v_z(t_0, y_0, z_0)\Delta z_0$$
$$+ \mathcal{O}(h||\Delta z_0||)$$

$$\Delta y_1 = \Delta y_0 + h \sum_{j=1}^{s} b_j (v(t_0 + c_j h, \widetilde{Y}_j, \widetilde{Z}_j) - v(t_0 + c_j h, Y_j, Z_j)) + h\widetilde{\delta}_s^y$$
$$= \Delta y_0 + \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\widetilde{\delta}^y||)$$

$$g_{ty}(t_1, y_1)\Delta y_1 = g_{ty}(t_0, y_0)\Delta y_0 + \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||)$$

$$g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) = g_{yy}(t_0, y_0)(\Delta y_0, v(t_0, y_0, z_0))$$
$$+ \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||).$$

Using these with (3.36), we obtain the bound

$$\begin{aligned} h\Delta\Lambda = \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\Delta \overline{Y}|| + h||\Delta Z|| \\ &+ h||\Delta\Lambda||^2 + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda|| \\ &+ \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\eta_0||). \end{aligned} \tag{3.37}$$

Applying (3.36) to (3.33a-c) results in

$$\Delta \overline{Y} = \Delta y_0 + \mathcal{O}\left(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||\right)$$
$$\Delta \widetilde{Y} = \Delta y_0 + \mathcal{O}\left(h||\Delta Y|| + h||\Delta Z|| + h||\widetilde{\delta}^y||\right)$$
$$\Delta \overline{Z} = \Delta z_0 + \mathcal{O}\left(h||\Delta \overline{Y}|| + h||\Delta Z|| + h||\Delta\Lambda|| + h||\delta^z||\right).$$

Reinserting these equations for $\Delta Y$, $\Delta\widetilde{Y}$, and $\Delta Z$ into each other and using (3.37) gives

$$\begin{aligned} \Delta Y = \Delta y_0 + \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ &+ h||\delta^\lambda|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\eta_0||) \end{aligned} \tag{3.38a}$$

$$\begin{aligned}\Delta\widetilde{Y} = \Delta y_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ + h||\delta^\lambda|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\eta_0||)\end{aligned} \tag{3.38b}$$

$$\begin{aligned}\Delta Z = \Delta z_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ + ||\delta^\lambda|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\eta_0||).\end{aligned} \tag{3.38c}$$

In addition, using (3.37) and (3.38) gives

$$\begin{aligned}h\Delta\Lambda = \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ + ||\delta^\lambda|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\eta_0||).\end{aligned} \tag{3.39}$$

The equations (3.38) and (3.39) show the first result (3.30). Subtracting (3.26) from (3.29) and linearizing yields

$$\begin{aligned}\Delta Z_i^f = \Delta z_0 + h\sum_{j=1}^{s}\widehat{a}_{ij}(f_y(t_0 + c_j h, Y_j, Z_j)\Delta Y_j + f_z(t_0 + c_j h, Y_j, Z_j)\Delta Z_j) \\ + h\delta_i^f + \mathcal{O}(h||\Delta Y||^2 + h||\Delta Z||^2)\end{aligned}$$

$$\begin{aligned}\Delta Z_i^r = h\sum_{j=0}^{s}\widetilde{a}_{ij}(r_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j) \\ + h\delta_i^r + \mathcal{O}(h||\Delta\widetilde{Y}||^2 + h||\Delta\Lambda||^2).\end{aligned}$$

Substituting in the expressions for $\Delta Y$, $\Delta\widetilde{Y}$, $\Delta Z$, and $\Delta\Lambda$, we arrive at

$$\begin{aligned}\Delta Z^f = \Delta z_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ + h||\delta^\lambda|| + h||\delta^f|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\eta_0||)\end{aligned}$$

$$\begin{aligned}\Delta Z^r = \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ + ||\delta^\lambda|| + h||\delta^r|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\eta_0||).\end{aligned}$$

This completes the proof. □

Each of the constants in the result of Theorem 3.4.1 depend only upon the derivatives of the functions $v$, $f$, $r$, and $g$, not upon any of the constants from the hypotheses. With some additional assumptions, the bounds of this perturbation theorem can be improved.

**Corollary 3.4.2.** *If, in addition to the conditions of Theorem 3.4.1, we assume that*

$$g(t_0, y_0) = 0 = g(t_0, \widehat{y}_0)$$

$$g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) = 0 = g_t(t_0, \widehat{y}_0) + g_y(t_0, \widehat{y}_0)v(t_0, \widehat{y}_0, \widehat{z}_0),$$

*then we have the bounds*

$$\begin{aligned}\Delta Y_i = \Delta y_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| \\ + h||\widetilde{\delta}^y|| + h||\delta^\lambda||)\end{aligned} \tag{3.40a}$$

$$\begin{aligned}\Delta \widetilde{Y}_i = \Delta y_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| \\ + h||\widetilde{\delta}^y|| + h||\delta^\lambda||)\end{aligned} \tag{3.40b}$$

$$\Delta Z_i = \Delta z_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda||) \tag{3.40c}$$

$$\begin{aligned}\Delta y_1 = \Delta y_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| \\ + h||\widetilde{\delta}^y|| + h||\delta^\lambda||)\end{aligned} \tag{3.40d}$$

$$\Delta z_1 = \Delta z_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda||) \tag{3.40e}$$

$$h\Delta\Lambda_i = \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda||) \tag{3.40f}$$

$$\begin{aligned}\Delta Z_i^f = \Delta z_0 + \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ + h||\delta^\lambda|| + h||\delta^f||)\end{aligned} \tag{3.40g}$$

$$\Delta Z_i^r = \mathcal{O}(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + ||\delta^\lambda|| + h||\delta^r||). \tag{3.40h}$$

*Proof.* With these stronger assumptions on the constraints, we subtract and linearize, giving

$$0 = g(t_0, \widehat{y}_0) - g(t_0, y_0) = g_y(t_0, y_0)\Delta y_0 + \mathcal{O}(||\Delta y_0||^2)$$

$$\begin{aligned}0 &= g_t(t_0, \widehat{y}_0) + g_y(t_0, \widehat{y}_0)v(t_0, \widehat{y}_0, \widehat{z}_0) - g_t(t_0, y_0) - g_y(t_0, y_0)v(t_0, y_0, z_0) \\ &= \eta_0 + \mathcal{O}(||\Delta y_0||^2 + ||\Delta z_0||^2),\end{aligned}$$

with $\eta_0$ as defined in (3.32). But this implies that

$$g_y(t_0, y_0)\Delta y_0 = \mathcal{O}(h^2||\Delta y_0||)$$

$$\eta_0 = \mathcal{O}(h^3||\Delta y_0|| + h^2||\Delta z_0||),$$

as $\Delta y_0 = \mathcal{O}(h^3)$ and $\Delta z_0 = \mathcal{O}(h^2)$. The conclusion (3.40) now follows by applying these bounds to the results of Theorem 3.4.1. □

### 3.5 Discontinuous Collocation Type Methods

We present here discontinuous collocation type methods for solving problems with originally index 3 constraints, (3.1) and (3.27). Similar results can be found in [8] a different class of methods applied to index 3 problems.

**Definition 3.5.1.** *Let $c_1, \ldots, c_s$ be distinct real numbers, and $\widetilde{c}_0, \ldots, \widetilde{c}_s$ also be distinct real numbers, with $\widetilde{c}_0 = 0$, $\widetilde{c}_s = 1$. We then define the $s$-degree polynomials $Y(t)$, $Z^f(t)$, and $\Lambda(t)$, and the $(s+1)$-degree polynomials $Z(t)$, $Z^r(t)$, as the polynomials satisfying the initial conditions*

$$\begin{gathered} Y(t_0) = y_0, \\ Z^f(t_0) = z_0, \quad Z^r(t_0) = -h\widetilde{b}_0\widetilde{\mu}(t_0), \quad Z(t_0) = Z^f(t_0) + Z^r(t_0), \end{gathered} \tag{3.41}$$

*where*

$$\widetilde{\mu}(t) := \dot{Z}^r(t) - r(t, Y(t), \Lambda(t)),$$

*as well as the conditions*

$$\dot{Y}(t_0 + c_i h) = v(t_0 + c_i h, Y(t_0 + c_i h), Z(t_0 + c_i h)), \quad i = 1, \ldots, s \tag{3.42a}$$

$$\dot{Z}^f(t_0 + c_i h) = f(t_0 + c_i h, Y(t_0 + c_i h), Z(t_0 + c_i h)), \quad i = 1, \ldots, s \tag{3.42b}$$

$$\dot{Z}^r(t_0 + \widetilde{c}_i h) = r(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h), \Lambda(t_0 + \widetilde{c}_i h)), \quad i = 1, \ldots, s-1 \tag{3.42c}$$

$$Z(t) = Z^f(t) + Z^r(t) \tag{3.42d}$$

$$0 = g(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h)), \quad i = 0, \ldots, s \tag{3.42e}$$

$$0 = g_t(t_1, Y(t_1)) + g_y(t_1, Y(t_1)) v(t_1, Y(t_1), Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)). \tag{3.42f}$$

*The polynomials $Y(t)$, $Z(t)$, $Z^f(t)$, $Z^r(t)$, and $\Lambda(t)$ are referred to as* discontinuous collocation type polynomials. *The expressions $Y(t_1)$, $Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)$, and $\Lambda(t_1)$, are used as approximations to the exact solutions $y(t)$, $z(t)$, and $\lambda(t)$, respectively,*

*of (3.1) at time* $t_1 := t_0 + h$.

**Theorem 3.5.2.** *The discontinuous collocation type polynomials defined by (3.42) are equivalent to an* $(s, s)$*-stage SPARK method (3.7). Given* $\widetilde{b}_0$ *and* $\widetilde{b}_s$*, the remaining coefficients are determined by*

$$a_{ij} = \widehat{a}_{ij} = \int_0^{c_i} \ell_j(\tau)d\tau, \quad b_j = \widehat{b}_j = \int_0^1 \ell_j(\tau)d\tau, \quad i, j = 1, \ldots, s, \tag{3.43a}$$

$$\bar{a}_{ij} = \int_0^{\widetilde{c}_i} \ell_j(\tau)d\tau, \quad i = 0, \ldots, s, \ j = 1, \ldots, s, \tag{3.43b}$$

$$\widetilde{a}_{ij} = \int_0^{c_i} \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0), \quad i = 1, \ldots, s, \ j = 0, \ldots, s,$$

$$\widetilde{a}_{i0} = \widetilde{b}_0, \quad \widetilde{a}_{is} = 0, \quad i = 1, \ldots, s, \tag{3.43c}$$

$$\widetilde{b}_j = \int_0^1 \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0) - \widetilde{b}_s\widehat{\ell}_j(\widetilde{c}_s), \quad j = 1, \ldots, s-1,$$

*where the functions* $\ell_j(\tau)$ *and* $\widehat{\ell}_j(\tau)$ *are Lagrange polynomials given by*

$$\ell_j(\tau) = \prod_{\substack{k=1 \\ k\neq j}}^{s} \left(\frac{\tau - c_k}{c_j - c_k}\right), \qquad \widehat{\ell}_j(\tau) = \prod_{\substack{k=1 \\ k\neq j}}^{s-1} \left(\frac{\tau - \widetilde{c}_k}{\widetilde{c}_j - \widetilde{c}_k}\right).$$

*Proof.* Using Lagrangian interpolation, we write

$$\dot{Y}(t_0 + \tau h) = \sum_{j=1}^{s} \ell_j(\tau)\dot{Y}(t_0 + c_j h). \tag{3.44}$$

Next, using the Fundamental Theorem of Calculus gives

$$Y(t_0 + c_i h) = y_0 + h\int_0^{c_i} \dot{Y}(t_0 + \tau h)d\tau, \quad i = 1, \ldots, s. \tag{3.45}$$

Inserting (3.44), we get

$$Y(t_0 + c_i h) = y_0 + h\sum_{j=1}^{s}\int_0^{c_i} \ell_j(\tau)d\tau \ \dot{Y}(t_0 + c_j h), \quad i = 1, \ldots, s. \tag{3.46}$$

Take $a_{ij} := \int_0^{c_i} \ell_j(\tau)d\tau$. If we define $Y_i := Y(t_0 + c_i h)$ and $Z_i := Z(t_0 + c_i h)$, then (3.46) becomes

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} v(t_0 + c_j h, Y_j, Z_j), \quad i = 1, \ldots, s, \tag{3.47}$$

which matches (3.7a). Additionally, we can write, for $i = 0, \ldots, s$,

$$Y(t_0 + \widetilde{c}_i h) = y_0 + h \int_0^{\widetilde{c}_i} \dot{Y}(t_0 + \tau h) d\tau \tag{3.48}$$

$$= y_0 + h \sum_{j=1}^{s} \int_0^{\widetilde{c}_i} \ell_j(\tau) d\tau \; v(t_0 + c_j h, Y_j, Z_j), \tag{3.49}$$

and defining $\bar{a}_{ij} := \int_0^{\widetilde{c}_i} \ell_j(\tau) d\tau$, and $\widetilde{Y}_i := Y(t_0 + \widetilde{c}_i h)$, we find that (3.49) can be seen as being equivalent to (3.7b). The discontinuous collocation type polynomial $Z^f(t)$ can be treated similarly. We have

$$\dot{Z}^f(t_0 + \tau h) = \sum_{j=1}^{s} \ell_j(\tau) \dot{Z}^f(t_0 + c_j h).$$

The polynomial $Z^f(t)$ can thus be expressed as

$$\begin{aligned} Z^f(t_0 + c_i h) &= z_0 + h \int_0^{c_i} \dot{Z}^f(t_0 + \tau h) d\tau \\ &= z_0 + h \sum_{j=1}^{s} \int_0^{c_i} \ell_j(\tau) d\tau \; f(t_0 + c_j h, Y_j, Z_j), \end{aligned}$$

for $i = 1, \ldots, s$. Taking $Z_i^f := Z^f(t_0 + c_i h)$ and $\widehat{a}_{ij} := \int_0^{c_i} \ell_j(\tau) d\tau$, we arrive at (3.26a). For the polynomial $Z^r(t)$ we have

$$\dot{Z}^r(t_0 + \tau h) = \sum_{j=1}^{s-1} \widehat{\ell}_j(\tau) \dot{Z}^r(t_0 + \widetilde{c}_j h) = \sum_{j=1}^{s-1} \widehat{\ell}_j(\tau) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j),$$

where $\Lambda_j := \Lambda(t_0 + \widetilde{c}_j h)$. Thus, for $i = 1, \ldots, s$,

$$\begin{aligned} Z^r(t_0 + c_i h) &= -h\widetilde{b}_0 (\dot{Z}^r(t_0) - r(t_0, Y(t_0), \Lambda(t_0))) + h \int_0^{c_i} \dot{Z}^r(t_0 + \tau h) d\tau \\ &= -h\widetilde{b}_0 \left( \sum_{j=1}^{s-1} \widehat{\ell}_j(\widetilde{c}_0) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) - r(t_0, Y(t_0), \Lambda(t_0)) \right) \\ &\qquad + h \int_0^{c_i} \sum_{j=1}^{s-1} \widehat{\ell}_j(\tau) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) d\tau \\ &= h\widetilde{b}_0 r(t_0, Y(t_0), \Lambda(t_0)) \\ &\quad + h \sum_{j=1}^{s-1} \left( \int_0^{c_i} \widehat{\ell}_j(\tau) d\tau - \widetilde{b}_0 \widehat{\ell}_j(\widetilde{c}_0) \right) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j). \end{aligned} \tag{3.50}$$

Define $\widetilde{a}_{ij} := \int_0^{c_i} \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0)$ for $i = 1, \ldots, s$ and $j = 1, \ldots, s-1$, and $\widetilde{a}_{i0} := \widetilde{b}_0$, $\widetilde{a}_{is} := 0$. If we take $Z_i^r := Z^r(t_0 + c_i h)$, then (3.50) becomes (3.26b).

For the polynomial $Z(t)$, using $Z(t) = Z^f(t) + Z^r(t)$ and the results above immediately give an internal stage of a SPARK method, by taking $Z_i := Z(t_0 + c_i h)$.

Similar to the work above, we can write

$$Y(t_1) = y_0 + h\int_0^1 \dot{Y}(t_0 + \tau h)d\tau. \tag{3.51}$$

Again inserting (3.44), we get

$$\begin{aligned} Y(t_1) &= y_0 + h\sum_{j=1}^{s}\int_0^1 \ell_j(\tau)d\tau \cdot \dot{Y}(t_0 + c_j h) \\ &= y_0 + h\sum_{j=1}^{s}\int_0^1 \ell_j(\tau)d\tau \cdot v(t_0 + c_j h, Y_j, Z_j). \end{aligned} \tag{3.52}$$

Taking $b_j := \int_0^1 \ell_j(\tau)d\tau$ and $y_1 := Y(t_0 + h)$, (3.52) becomes the numerical solution $y_1$ given by a SPARK method in (3.7d). Working with $Z^f(t_0 + h)$ and $Z^r(t_0 + h)$ gives the similar results

$$Z^f(t_1) = z_0 + h\sum_{j=1}^{s}\int_0^1 \ell_j(\tau)d\tau \cdot f(t_0 + c_j h, Y_j, Z_j) \tag{3.53}$$

$$\begin{aligned} Z^r(t_1) = {} & h\widetilde{b}_0 r(t_0, Y(t_0), \Lambda(t_0)) \\ & + h\sum_{j=1}^{s-1}\left(\int_0^1 \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widehat{c}_0)\right) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j). \end{aligned} \tag{3.54}$$

Applying these formulas, and using $z_1 := Z^f(t_0 + h) + Z^r(t_0 + h) - h\widetilde{b}_s\widetilde{\mu}(t_0 + h)$, the numerical approximation of $z(t_0 + h)$ becomes

$$\begin{aligned} z_1 = {} & z_0 + h\sum_{j=1}^{s}\int_0^1 \ell_j(\tau)d\tau \cdot f(t_0 + c_j h, Y_j, Z_j) \\ & + h\sum_{j=1}^{s-1}\left(\int_0^1 \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0) - \widetilde{b}_s\widehat{\ell}_j(\widetilde{c}_s)\right) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) \\ & + h\widetilde{b}_0 r(t_0, Y(t_0), \Lambda(t_0)) + h\widetilde{b}_s r(t_1, Y(t_1), \Lambda(t_1)). \end{aligned}$$

If we define $\widehat{b}_j := \int_0^1 \ell_j(\tau)d\tau$ for $j = 1, \ldots, s$ and $\widetilde{b}_j := \int_0^1 \widehat{\ell}_j(t)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0) - \widetilde{b}_s\widehat{\ell}_j(\widetilde{c}_s)$ for $j = 1, \ldots, s-1$, then this approximation agrees with that of (3.7e).

Lastly, we must check that the discontinuous collocation type method satisfies the constraints (3.7f-h) of a SPARK method. However, both (3.7f,h) follow immediately from the definitions of $Y_i$,$Z_i$, and $\widetilde{Y}_i$ above, and from (3.42). The condition (3.7g) comes from the fact that

$$g(t_1, y_1) = g(t_1, Y(t_1)) = g(t_0 + \widetilde{c}_s h, Y(t_0 + \widetilde{c}_s h)) = 0$$

since $\widetilde{c}_s = 1$. □

The Gauss-Lobatto methods can thus be expressed as discontinuous collocation type methods. This fact will be useful for determining the order of SPARK methods. We take advantage of this later for computing the local error of the Gauss-Lobatto methods. For now, we give the equivalence of SPARK methods and discontinuous collocation methods in the following theorem.

**Theorem 3.5.3.** *A SPARK method with distinct values $c_1, \ldots, c_s$, distinct values $\widetilde{c}_0, \ldots, \widetilde{c}_s$, and coefficients $\widehat{a}_{ij} = a_{ij}$, $\widehat{b}_j = b_j$, is a discontinuous collocation type method (3.42) if and only if the coefficients satisfy*

$$\sum_{j=1}^{s} a_{ij}c_j^{k-1} = \frac{c_i^k}{k}, \qquad \sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, s \tag{3.55a}$$

$$\sum_{j=0}^{s} \widetilde{a}_{ij}\widetilde{c}_j^{k-1} = \frac{c_i^k}{k}, \qquad \sum_{j=0}^{s} \widetilde{b}_j\widetilde{c}_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, s-1 \tag{3.55b}$$

$$\widetilde{a}_{i0} = \widetilde{b}_0, \qquad \widetilde{a}_{is} = 0 \tag{3.55c}$$

$$\sum_{j=1}^{s} \bar{a}_{ij}c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad k = 1, \ldots, s-1. \tag{3.55d}$$

*Proof.* The proof of this theorem is similar to proofs found in [8] and [9]. The coefficients $a_{ij}$ and $b_j$ are uniquely determined by the conditions (3.55a). These

conditions are equivalent to

$$\begin{aligned} \sum_{j=1}^{s} a_{ij} p(c_j) &= \int_0^{c_i} p(\tau) d\tau \\ \sum_{j=1}^{s} b_j p(c_j) &= \int_0^{1} p(\tau) d\tau \end{aligned} \tag{3.56}$$

for $p$ any polynomial of degree less than or equal to $s-1$. Thus, if a SPARK method satisfies (3.55a), then the $a_{ij}$ and $b_j$ coefficients are equal to those defined in Theorem 3.5.2. This is because the coefficients in (3.43) satisfy the above conditions (3.56), as this is just the Lagrange interpolation formula. The equality of the coefficients $\bar{a}_{ij}$ follow similarly for (3.55d).

For the $\widetilde{a}_{ij}$ and $\widetilde{b}_j$ coefficients, the conditions (3.55b) are equivalent to

$$\begin{aligned} \sum_{j=0}^{s} \widetilde{a}_{ij} p(\widetilde{c}_j) &= \int_0^{c_i} p(\tau) d\tau \\ \sum_{j=0}^{s} \widetilde{b}_j p(\widetilde{c}_j) &= \int_0^{1} p(\tau) d\tau \end{aligned} \tag{3.57}$$

for $p$ any polynomial with degree less than or equal to $s-2$. But using the coefficients defined in Theorem 3.5.2, the first of these conditions is equivalent to

$$\widetilde{b}_0 p(\widetilde{c}_0) + \sum_{j=1}^{s-1} \widetilde{a}_{ij} p(\widetilde{c}_j) = \int_0^{c_i} p(\tau) d\tau \tag{3.58}$$

$$\widetilde{b}_0 \sum_{j=1}^{s-1} \widehat{\ell}_j(\widetilde{c}_0) p(\widetilde{c}_j) + \sum_{j=1}^{s-1} \left( \int_0^{c_i} \widehat{\ell}_j(\tau) d\tau - \widetilde{b}_0 \widehat{\ell}_j(\widetilde{c}_0) \right) p(\widetilde{c}_j) = \int_0^{c_i} p(\tau) d\tau \tag{3.59}$$

$$\sum_{j=1}^{s-1} \int_0^{c_i} \widehat{\ell}_j(\tau) p(\widetilde{c}_j) d\tau = \int_0^{c_i} p(\tau) d\tau. \tag{3.60}$$

Through similar computations, the second condition in (3.55b) is equivalent to

$$\sum_{j=1}^{s-1} \int_0^{1} \widehat{\ell}_j(\tau) p(\widetilde{c}_j) d\tau = \int_0^{1} p(\tau) d\tau. \tag{3.61}$$

Again, if a SPARK method satisfies (3.55b,c), then the coefficients $\widetilde{a}_{ij}$ and $\widetilde{b}_j$ are equal to those defined in Theorem 3.5.2. This is because the coefficients in (3.43)

satisfy the above conditions above, as this is just the Lagrange interpolation formula again.

Finally, the expressions for the constraints in a SPARK method (3.7) are equivalent to those of a discontinuous collocation method type (3.42). This completes the proof. □

We present here a lemma regarding the error of the internal stages of a collocation or SPARK method for problems with originally index 3 constraints, considering Gauss-Lobatto coefficients. This lemma will be useful for showing the effectiveness of the derivatives of discontinuous collocation type methods, as well as for a proof of local error.

**Lemma 3.5.4.** *Suppose the internal stages $Y_i$, $\widetilde{Y}_j$, $Z_i$, and $\Lambda_j$ are as defined in (3.7) with Gauss-Lobatto coefficients, for $i = 1, \ldots, s$, and $j = 0, \ldots, s$. Let $y(t), z(t), \lambda(t)$ be the exact solutions to (3.1), and let $z^f(t), z^r(t)$ be the exact solutions to (3.27). Then we have the bounds*

$$\begin{aligned} Y_i - y(t_0 + c_i h) &= \mathcal{O}(h^{s+1}), \qquad \widetilde{Y}_i - y(t_0 + \widetilde{c}_i h) = \mathcal{O}(h^{s+2}), \\ Z_i - z(t_0 + c_i h) &= \mathcal{O}(h^{s+1}), \qquad \Lambda_i - \lambda(t_0 + \widetilde{c}_i h) = \mathcal{O}(h^{s}), \\ Z_i^f - z^f(t_0 + c_i h) &= \mathcal{O}(h^{s+1}), \quad Z_i^r - z^r(t_0 + \widetilde{c}_i h) = \mathcal{O}(h^{s+1}). \end{aligned} \tag{3.62}$$

*Proof.* This lemma can be proved by applying Corollary 3.4.2 with the exact solution for the perturbed values. More concretely, we take

$$\begin{aligned} \widehat{Y}_i &= y(t_0 + c_i h), \quad \widehat{y}_1 = y(t_0 + h), \\ \widehat{Z}_i &= z(t_0 + c_i h), \quad \widehat{z}_1 = z(t_0 + h), \\ \widehat{\Lambda}_i &= \lambda(t_0 + \widetilde{c}_i h), \quad \widehat{\widetilde{Y}}_i = Y(t_0 + \widetilde{c}_i h), \\ \widehat{y}_0 &= y(t_0), \quad \widehat{z}_0 = z(t_0), \\ \widehat{Z}_i^f &= z^f(t_0 + c_i h), \quad \widehat{Z}_i^r = z^r(t_0 + \widetilde{c}_i h). \end{aligned}$$

Because the exact solution satisfies

$$g(t_0 + \widetilde{c}_i h, y(t_0 + \widetilde{c}_i h)) = g(t_0 + h, y(t_0 + h)) = 0, \quad i = 0, \ldots, s,$$
$$g_t(t_0 + h, y(t_0 + h)) + g_y v(t_0 + h, y(t_0 + h), z(t_0 + h)) = 0,$$

the constraints (3.28f,g) give that $\delta_i^\lambda = 0$ for all $i = 1, \ldots, s+1$. Using a Taylor series expansion in $h$ around $h = 0$, the values $\widehat{Y}_i$, $\widehat{\widetilde{Y}}_i$, and $\widehat{Z}_i$ can be expressed as

$$\widehat{Y}_i = y(t_0 + c_i h) = \sum_{k=0}^{s} \frac{h^k}{k!} c_i^k y^{(k)}(t_0) + \mathcal{O}(h^{s+1}) \tag{3.63}$$
$$\widehat{\widetilde{Y}}_i = y(t_0 + \widetilde{c}_i h) = \sum_{k=0}^{s} \frac{h^k}{k!} \widetilde{c}_i^k y^{(k)}(t_0) + \mathcal{O}(h^{s+1}) \tag{3.64}$$
$$\widehat{Z}_i = z(t_0 + c_i h) = \sum_{k=0}^{s} \frac{h^k}{k!} c_i^k z^{(k)}(t_0) + \mathcal{O}(h^{s+1}). \tag{3.65}$$

Thus (3.28a) gives us, for $i = 1, \ldots, s$,

$$\begin{aligned}
\delta_i^y &= \frac{1}{h}(\widehat{Y}_i - \widehat{y}_0) - \sum_{j=1}^{s} a_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{k!} c_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} \sum_{k=0}^{s} a_{ij} \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \left[ \frac{h^{k-1}}{k!} c_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} a_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{c_i^k}{k} - \sum_{j=1}^{s} a_{ij} c_j^{k-1} \right] + \mathcal{O}(h^{s+1}) \\
&= \frac{h^s}{s!} y^{(s+1)}(t_0) \left[ \frac{c_i^{s+1}}{s+1} - \sum_{j=1}^{s} a_{ij} c_j^s \right] + \mathcal{O}(h^{s+1}) \\
&= \mathcal{O}(h^s).
\end{aligned}$$

We have made use of the important fact that the Gauss coefficients $a_{ij}$ satisfy $C(s)$ defined in (3.12). For $\delta_{s+1}^y$, a similar derivation can be made, using the fact that the Gauss coefficients $b_i$ also satisfy $B(2s)$ defined in (3.11). This, along with (3.28d),

gives

$$\begin{aligned}
\delta_{s+1}^y &= \frac{1}{h}(\widehat{y}_1 - \widehat{y}_0) - \sum_{j=1}^{s} b_j v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) \\
&= \sum_{k=1}^{2s+1} \frac{h^{k-1}}{k!} y^{(k)}(t_0) - \sum_{j=1}^{2s} \sum_{k=0}^{s} b_j \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{2s+1}) \\
&= \sum_{k=1}^{2s+1} \left[ \frac{h^{k-1}}{k!} y^{(k)}(t_0) - \sum_{j=1}^{s} b_j \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{2s+1}) \\
&= \sum_{k=1}^{2s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{1}{k} - \sum_{j=1}^{s} b_j c_j^{k-1} \right] + \mathcal{O}(h^{2s+1}) \\
&= \frac{h^s}{s!} y^{(s+1)}(t_0) \left[ \frac{1}{s+1} - \sum_{j=1}^{s} b_j c_j^s \right] + \mathcal{O}(h^{2s+1}) \\
&= \mathcal{O}(h^{2s}).
\end{aligned}$$

This calculation can be repeated for $\widetilde{\delta}_i^y$. Using (3.28b), we get, for $i = 0, \ldots, s$,

$$\begin{aligned}
\widetilde{\delta}_i^y &= \frac{1}{h}(\widehat{\widetilde{Y}}_i - \widehat{y}_0) - \sum_{j=1}^{s} \bar{a}_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{k!} \widetilde{c}_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} \sum_{k=0}^{s} \bar{a}_{ij} \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \left[ \frac{h^{k-1}}{k!} \widetilde{c}_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} \bar{a}_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{\widetilde{c}_i^k}{k} - \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} \right] + \mathcal{O}(h^{s+1}) \\
&= \mathcal{O}(h^{s+1}).
\end{aligned}$$

We have made use of the important fact that the coefficients $\bar{a}_{ij}$ satisfy $\overline{C}(s+1)$ from (3.20) in Lemma 3.2.4. A similar process can be repeated for the $\delta_i^z$. Using (3.28c), we get

$$\delta_i^z = \frac{1}{h}(\widehat{Z}_i - \widehat{z}_0) - \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) - \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, , \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j)$$

$$
\begin{aligned}
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k z^{(k)}(t_0) \right] - \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, y(t_0 + c_j h), z(t_0 + c_j h)) \\
&\qquad - \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, y(t_0 + \widetilde{c}_j h), \lambda(t_0 + \widetilde{c}_j h)) + \mathcal{O}(h^s) \\
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k \left( z^{f(k)}(t_0) + z^{r(k)}(t_0) \right) - \sum_{j=1}^{s} a_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} z^{f(k)}(t_0) \right. \\
&\qquad \left. - \sum_{j=0}^{s} \widetilde{a}_{ij} \frac{h^{k-1}}{(k-1)!} \widetilde{c}_j^{k-1} z^{r(k)}(t_0) \right] + \mathcal{O}(h^s) \\
&= \sum_{k=1}^{s} \frac{h^{k-1}}{(k-1)!} \left[ \left( \frac{c_i^k}{k} - \sum_{j=1}^{s} a_{ij} c_j^{k-1} \right) z^{f(k)}(t_0) \right. \\
&\qquad \left. + \left( \frac{c_i^k}{k} - \sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} \right) z^{r(k)}(t_0) \right] + \mathcal{O}(h^s) \\
&= \mathcal{O}(h^s).
\end{aligned}
$$

To get to the last step, we use the properties $C(s)$ (3.12) and $\widetilde{C}(s)$ (3.17). For $\delta_{s+1}^z$, we get the similar result

$$
\begin{aligned}
\delta_{s+1}^z &= \frac{1}{h}(\widehat{z}_1 - \widehat{z}_0) - \sum_{j=1}^{s} b_j f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) - \sum_{j=0}^{s} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) \\
&= \sum_{k=1}^{2s} \left[ \frac{h^{k-1}}{k!} z^{(k)}(t_0) \right] - \sum_{j=1}^{s} b_j f(t_0 + c_j h, y(t_0 + c_j h), z(t_0 + c_j h)) \\
&\qquad - \sum_{j=0}^{s} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, y(t_0 + \widetilde{c}_j h), \lambda(t_0 + \widetilde{c}_j h)) + \mathcal{O}(h^{2s}) \\
&= \sum_{k=1}^{2s} \left[ \frac{h^{k-1}}{k!} \left( z^{f(k)}(t_0) + z^{r(k)}(t_0) \right) - \sum_{j=1}^{s} b_j \frac{h^{k-1}}{(k-1)!} c_j^{k-1} z^{f(k)}(t_0) \right. \\
&\qquad \left. - \sum_{j=0}^{s} \widetilde{b}_j \frac{h^{k-1}}{(k-1)!} \widetilde{c}_j^{k-1} z^{r(k)}(t_0) \right] + \mathcal{O}(h^{2s}) \\
&= \sum_{k=1}^{2s} \frac{h^{k-1}}{(k-1)!} \left[ \left( \frac{1}{k} - \sum_{j=1}^{s} b_j c_j^{k-1} \right) z^{f(k)}(t_0) \right. \\
&\qquad \left. + \left( \frac{1}{k} - \sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k-1} \right) z^{r(k)}(t_0) \right] + \mathcal{O}(h^{2s}) \\
&= \mathcal{O}(h^{2s}).
\end{aligned}
$$

The final step is derived using $B(2s)$ in (3.11), and $\widetilde{B}(2s)$ in (3.13). The additional perturbations $\delta_i^f$ and $\delta_i^r$ can be expressed in a manner similar to that of $\delta_i^z$. The relation (3.29a) can be expressed as

$$\begin{aligned}
\delta_i^f &= \frac{1}{h}(\widehat{Z}_i^f - \widehat{z}_0) - \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) \\
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k z^{f^{(k)}}(t_0) - \sum_{j=1}^{s} a_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} z^{f^{(k)}}(t_0) \right] + \mathcal{O}(h^s) \\
&= \sum_{k=1}^{s} \frac{h^{k-1}}{(k-1)!} \left( \frac{c_i^k}{k} - \sum_{j=1}^{s} a_{ij} c_j^{k-1} \right) z^{f^{(k)}}(t_0) + \mathcal{O}(h^s) \\
&= \mathcal{O}(h^s).
\end{aligned}$$

The relation (3.29b) can also be expressed as

$$\begin{aligned}
\delta_i^r &= \frac{1}{h}\widehat{Z}_i^r - \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) \\
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k z^{r(k)}(t_0) - \sum_{j=0}^{s} \widetilde{a}_{ij} \frac{h^{k-1}}{(k-1)!} \widetilde{c}_j^{k-1} z^{r(k)}(t_0) \right] + \mathcal{O}(h^s) \\
&= \sum_{k=1}^{s} \frac{h^{k-1}}{(k-1)!} \left[ \left( \frac{c_i^k}{k} - \sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} \right) z^{r(k)}(t_0) \right] + \mathcal{O}(h^s) \\
&= \mathcal{O}(h^s).
\end{aligned}$$

Finally, by applying Corollary 3.4.2, we get

$$\begin{aligned}
&Y_i - y(t_0 + c_i h) = \mathcal{O}\left( h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \right) = \mathcal{O}\left( h^{s+1} \right), \\
&\widetilde{Y}_i - y(t_0 + \widetilde{c}_i h) = \mathcal{O}\left( h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \right) = \mathcal{O}\left( h^{s+2} \right), \\
&Z_i - z(t_0 + c_i h) = \mathcal{O}\left( h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \right) = \mathcal{O}\left( h^{s+1} \right), \\
&\Lambda_i - \lambda(t_0 + \widetilde{c}_i h) = \frac{1}{h}\mathcal{O}\left( h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \right) = \mathcal{O}\left( h^s \right), \\
&y_1 - y(t_0 + h) = \mathcal{O}\left( h^{s+1} \right), \\
&z_1 - z(t_0 + h) = \mathcal{O}\left( h^{s+1} \right), \\
&Z_i^f - z^f(t_0 + c_i h) = \mathcal{O}\left( h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| + h||\delta^f|| \right) = \mathcal{O}\left( h^{s+1} \right),
\end{aligned}$$

$$Z_i^r - z^r(t_0 + \widetilde{c}_i h) = \mathcal{O}\left(h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\delta^r||\right) = \mathcal{O}\left(h^{s+1}\right). \qquad \square$$

The next theorem gives the quality of the derivatives of the approximations by discontinuous collocation type methods. This result is similar to that of [9, Theorem II.7.10] and [10, Theorem VII.4.8].

**Theorem 3.5.5.** *Let $y(t), z(t), \lambda(t)$ be the exact solutions to the problem (3.1). The discontinuous collocation type polynomials $Y(t)$, $Z(t)$, and $\Lambda(t)$ defined by (3.42) with Gauss coefficients $c_i$ and Lobatto coefficients $\widetilde{c}_i$ satisfy for $k = 0, \ldots, s$ and $t \in [t_0, t_1]$*

$$||Y^{(k)}(t) - y^{(k)}(t)|| \leq Ch^{s+1-k}, \tag{3.66a}$$

$$||Z^{f(k)}(t) - z^{f(k)}(t)|| \leq Ch^{s+1-k}, \tag{3.66b}$$

$$||Z^{r(k)}(t) - z^{r(k)}(t)|| \leq Ch^{s-1-k}, \tag{3.66c}$$

$$||Z^{(k)}(t) - z^{(k)}(t)|| \leq Ch^{s-1-k}, \tag{3.66d}$$

$$||\Lambda^{(k)}(t) - \lambda^{(k)}(t)|| \leq Ch^{s+1-k}. \tag{3.66e}$$

*Proof.* As in the proof of Theorem 3.5.2, take $Y_i := Y(t_0 + c_i h)$, $Z_i := Z(t_0 + c_i h)$, and $\Lambda_i := \Lambda(t_0 + \widetilde{c}_i h)$. Using the convention $c_0 := 0$, we can write the collocation polynomials as

$$Y(t_0 + \tau h) = y_0 \ell_0(\tau) + \sum_{i=1}^{s} Y_i \ell_i(\tau) \tag{3.67a}$$

$$Z^f(t_0 + \tau h) = z_0 \ell_0(\tau) + \sum_{i=1}^{s} Z_i^f \ell_i(\tau) \tag{3.67b}$$

$$Z^r(t_0 + \tau h) = -h\widetilde{b}_0 \widetilde{\mu}(t_0)\widehat{\ell}_0(\tau) + \sum_{i=1}^{s-1} Z_i^r \widehat{\ell}_i(\tau) \tag{3.67c}$$

$$Z(t_0 + \tau h) = Z^f(t_0 + \tau h) + Z^r(t_0 + \tau h) \tag{3.67d}$$

$$\Lambda(t_0 + \tau h) = \sum_{i=0}^{s} \Lambda_i \widetilde{\ell}_i(\tau) \tag{3.67e}$$

with $\tau \in \mathbb{R}$ and Lagrange polynomials $\ell_i$, $\widehat{\ell}_i$, and $\widetilde{\ell}_i$ defined as

$$\ell_i(\tau) := \prod_{\substack{j=0 \\ j\neq i}}^{s} \left( \frac{\tau - c_j}{c_i - c_j} \right), \quad \widehat{\ell}_i(\tau) := \prod_{\substack{j=0 \\ j\neq i}}^{s-1} \left( \frac{\tau - \widetilde{c}_j}{\widetilde{c}_i - \widetilde{c}_j} \right), \quad \widetilde{\ell}_i(\tau) := \prod_{\substack{j=0 \\ j\neq i}}^{s} \left( \frac{\tau - \widetilde{c}_j}{\widetilde{c}_i - \widetilde{c}_j} \right).$$

Applying the Lagrange interpolation formula to the exact solutions gives

$$y(t_0 + \tau h) = y_0 \ell_0(\tau) + \sum_{i=1}^{s} y(t_0 + c_i h)\ell_i(\tau) + \mathcal{O}(h^{s+1}) \tag{3.68a}$$

$$z^f(t_0 + \tau h) = z_0 \ell_0(\tau) + \sum_{i=1}^{s} z^f(y_0 + c_i h)\ell_i(\tau) + \mathcal{O}(h^{s+1}) \tag{3.68b}$$

$$z^r(t_0 + \tau h) = \sum_{i=1}^{s-1} z^r(t_0 + \widetilde{c}_i h)\widehat{\ell}_i(\tau) + \mathcal{O}(h^s) \tag{3.68c}$$

$$z(t_0 + \tau h) = z^f(t_0 + \tau h) + z^r(t_0 + \tau h) \tag{3.68d}$$

$$\lambda(t_0 + \tau h) = \sum_{i=0}^{s} \lambda(t_0 + \widetilde{c}_i h)\widetilde{\ell}_i(\tau) + \mathcal{O}(h^{s+1}). \tag{3.68e}$$

Define the functions

$$\overline{Y}(\tau) := y(t_0 + \tau h) - \left( y_0 \ell_0(\tau) + \sum_{i=1}^{s} y(t_0 + c_i h)\ell_i(\tau) \right)$$

$$\overline{Z}^f(\tau) := z^f(t_0 + \tau h) - \left( z_0 \ell_0(\tau) + \sum_{i=1}^{s} z^f(y_0 + c_i h)\ell_i(\tau) \right)$$

$$\overline{Z}^r(\tau) := z^r(t_0 + \tau h) - \sum_{i=1}^{s-1} z^r(t_0 + \widetilde{c}_i h)\widehat{\ell}_i(\tau)$$

$$\overline{Z}(\tau) := z(t_0 + \tau h) - \left( z_0 \ell_0(\tau) + \sum_{i=1}^{s} z^f(y_0 + c_i h)\ell_i(\tau) + \sum_{i=1}^{s-1} z^r(t_0 + \widetilde{c}_i h)\widehat{\ell}_i(\tau) \right)$$

$$\overline{\Lambda}(\tau) := \lambda(t_0 + \tau h) - \sum_{i=0}^{s} \lambda(t_0 + \widetilde{c}_i h)\widetilde{\ell}_i(\tau).$$

The functions $\overline{Y}(\tau)$ and $\overline{Z}^f(\tau)$ have $s+1$ zeros at each $c_i$, for $i = 0, \ldots, s$. The function $\overline{Z}^r(\tau)$ has $s-1$ zeros at each $\widetilde{c}_1, \ldots, \widetilde{c}_{s-1}$, while $\overline{\Lambda}$ has $s+1$ zeros at each $\widetilde{c}_i$, for $i = 0, \ldots, s$. We have

$$\overline{Y}^{(k)}(\tau) = h^k y^{(k)}(t_0 + \tau h) - \left( y_0 \ell_0^{(k)}(\tau) + \sum_{i=1}^{s} y(t_0 + c_i h)\ell_i^{(k)}(\tau) \right) \tag{3.69a}$$

$$\overline{Z}^{f^{(k)}}(\tau) = h^k z^{f^{(k)}}(t_0+\tau h) - \left( z_0 \ell_0^{(k)}(\tau) - \sum_{i=1}^{s} z^f(y_0 + c_i h)\ell_i^{(k)}(\tau) \right) \tag{3.69b}$$

$$\overline{Z}^{r^{(k)}}(\tau) = h^k z^{r^{(k)}}(t_0+\tau h) - \left( \sum_{i=1}^{s-1} z^r(t_0 + \widetilde{c}_i h)\widehat{\ell}_i^{(k)}(\tau) \right) \tag{3.69c}$$

$$\overline{\Lambda}^{(k)}(\tau) = h^k \lambda^{(k)}(t_0+\tau h) - \left( \sum_{i=0}^{s} \lambda(t_0 + \widetilde{c}_i h)\widetilde{\ell}_i^{(k)}(\tau) \right). \tag{3.69d}$$

Applying Rolle's Theorem to each open interval $]c_i, c_{i+1}[$ or $]\widetilde{c}_i, \widetilde{c}_{i+1}[$, we see that $\overline{Y}^{(k)}$, $\overline{Z}^{f^{(k)}}$, and $\overline{\Lambda}^{(k)}$ have $s+1-k$ zeros, and $\overline{Z}^{r^{(k)}}$ has $s-1-k$ zeros. Thus, the terms in brackets in (3.69) can be viewed as interpolation polynomials, with $\overline{Y}^{(k)}$, $\overline{Z}^{f^{(k)}}$, and $\overline{\Lambda}^{(k)}$ of degree $s-k$ and $\overline{Z}^{r^{(k)}}$ of degree $s-k-2$, for $h^k y^{(k)}(t_0+\tau h)$, $h^k z^{f^{(k)}}(t_0+\tau h)$, $h^k \lambda^{(k)}(t_0+\tau h)$, and $h^k z^{r^{(k)}}(t_0+\tau h)$, respectively. The interpolation error for each is $\mathcal{O}(h^{s+1})$, $\mathcal{O}(h^{s+1})$, $\mathcal{O}(h^{s-1})$, and $\mathcal{O}(h^{s+1})$, respectively. In other words,

$$h^k y^{(k)}(t_0+\tau h) = y_0 \ell_0^{(k)}(\tau) + \sum_{i=1}^{s} y(t_0 + c_i h)\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}) \tag{3.70a}$$

$$h^k z^{f^{(k)}}(t_0+\tau h) = z_0 \ell_0^{(k)}(\tau) - \sum_{i=1}^{s} z^f(y_0 + c_i h)\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}) \tag{3.70b}$$

$$h^k z^{r^{(k)}}(t_0+\tau h) = \sum_{i=1}^{s-1} z^r(t_0 + \widetilde{c}_i h)\widehat{\ell}_i^{(k)}(\tau) + \mathcal{O}(h^{s-1}) \tag{3.70c}$$

$$h^k \lambda^{(k)}(t_0+\tau h) = \sum_{i=0}^{s} \lambda(t_0 + \widetilde{c}_i h)\widetilde{\ell}_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}). \tag{3.70d}$$

Therefore, taking $k$ derivatives of (3.67) with respect to $s$ and subtracting (3.70), we arrive at the relations

$$h^k \left( Y^{(k)}(t_0+\tau h) - y^{(k)}(t_0+\tau h) \right) = \sum_{i=1}^{s} (Y_i - y(t_0 + c_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1})$$

$$h^k \left( Z^{f^{(k)}}(t_0+\tau h) - z^{f^{(k)}}(t_0+\tau h) \right) = \sum_{i=1}^{s} (Z_i^f - z^f(t_0 + c_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1})$$

$$h^k \left( Z^{r^{(k)}}(t_0+\tau h) - z^{r^{(k)}}(t_0+\tau h) \right) = -h\widetilde{b}_0 \widetilde{\mu}(t_0)\widehat{\ell}_0(\tau) + \sum_{i=1}^{s} (Z_i^r - z^r(t_0 + \widetilde{c}_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s-1})$$

$$h^k \left( Z^{(k)}(t_0+\tau h) - z^{(k)}(t_0+\tau h) \right) = h^k \left( Z^f(t_0+\tau h) - z^f(t_0+\tau h) \right)$$

$$+ h^k \left(Z^r(t_0+\tau h) - z^r(t_0+\tau h)\right)$$

$$h^k \left(\Lambda^{(k)}(t_0+\tau h) - \lambda^{(k)}(t_0+\tau h)\right) = \sum_{i=0}^{s} (\Lambda_i - \lambda(t_0+\widetilde{c}_i h))\widetilde{\ell}_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}).$$

Invoking Lemma 3.5.4, and dividing by $h^k$, we arrive at

$$Y^{(k)}(t_0+\tau h) - y^{(k)}(t_0+\tau h) = \mathcal{O}(h^{s+1-k}) \tag{3.71a}$$

$$Z^{f(k)}(t_0+\tau h) - z^{f(k)}(t_0+\tau h) = \mathcal{O}(h^{s+1-k}) \tag{3.71b}$$

$$Z^{r(k)}(t_0+\tau h) - z^{r(k)}(t_0+\tau h) = -h^{1-k}\widetilde{b}_0\widetilde{\mu}(t_0)\widehat{\ell}_0(\tau) + \mathcal{O}(h^{s-1-k}) \tag{3.71c}$$

$$Z^{(k)}(t_0+\tau h) - z^{(k)}(t_0+\tau h) = -h^{1-k}\widetilde{b}_0\widetilde{\mu}(t_0)\widehat{\ell}_0(\tau) + \mathcal{O}(h^{s-1-k}) \tag{3.71d}$$

$$\Lambda^{(k)}(t_0+\tau h) - \lambda^{(k)}(t_0+\tau h) = \mathcal{O}(h^{s+1-k}). \tag{3.71e}$$

However, $\dot{Z}^r(\tau) = \sum_{j=1}^{s-1} \widehat{\ell}_j(\tau) r(t_0+\widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)$ and $r(\tau, Y(\tau), \Lambda(\tau))$ agree for $\tau = \widetilde{c}_j$, with $j = 1, \ldots, s-1$. Therefore,

$$\widetilde{\mu}(t_0) = \dot{Z}^r(t_0) - r(t_0, Y(t_0), \Lambda(t_0)) = \mathcal{O}(h^{s-1}), \tag{3.72}$$

as $\dot{Z}^r(\tau)$ can be viewed as an interpolation polynomial for $r(\tau, Y(\tau), \Lambda(\tau))$. This gives

$$Z^{r(k)}(t_0+\tau h) - z^{r(k)}(t_0+\tau h) = \mathcal{O}(h^{s-1-k}),$$
$$Z^{(k)}(t_0+\tau h) - z^{(k)}(t_0+\tau h) = \mathcal{O}(h^{s-1-k}),$$

proving the result. □

## 3.6 Local Error Analysis

Using the fact that the Gauss-Lobatto SPARK methods for index 3 problems are equivalent to a class of discontinuous collocation type method, we can determine the local error for these methods.

**Theorem 3.6.1.** *For the $(s,s)$-Gauss-Lobatto SPARK methods (3.7) with consistent initial values $(y_0, z_0, \lambda_0)$ at time $t_0$, assume $g_y(t,y)v_z(t,y,z)r_\lambda(t,y,\lambda)$ is invertible. Then for $|h| \leq h_0$, the local error is of order $2s$, i.e.,*

$$y_1 - y(t_1) = \mathcal{O}(h^{2s+1}), \qquad z_1 - z(t_1) = \mathcal{O}(h^{2s+1}). \tag{3.73}$$

*Proof.* For the proof, we utilize the discontinuous collocation type polynomials $Y(t)$, $Z^f(t)$, $Z^r(t)$, $Z(t)$, and $\Lambda(t)$ defined by (3.42). We define the defects $\delta(t)$, $\mu(t)$, $\widetilde{\mu}(t)$, and $\widetilde{\theta}(t)$ by

$$\dot{Y}(t) = v(t, Y(t), Z(t)) + \delta(t) \tag{3.74a}$$

$$\dot{Z}^f(t) = f(t, Y(t), Z(t)) + \mu(t) \tag{3.74b}$$

$$\dot{Z}^r(t) = r(t, Y(t), \Lambda(t)) + \widetilde{\mu}(t) \tag{3.74c}$$

$$0 = g(t, Y(t)) + \widetilde{\theta}(t) \tag{3.74d}$$

$$0 = g_t(t, Y(t)) + g_y(t, Y(t))\left(v(t, Y(t), Z(t)) + \delta(t)\right) + \dot{\widetilde{\theta}}(t). \tag{3.74e}$$

From the definition of the discontinuous collocation type polynomials, the defects satisfy

$$\delta(t_0 + c_i h) = 0, \quad i = 1, \ldots, s \tag{3.75a}$$

$$\mu(t_0 + c_i h) = 0, \quad i = 1, \ldots, s \tag{3.75b}$$

$$\widetilde{\mu}(t_0 + \widetilde{c}_i h) = 0, \quad i = 1, \ldots, s-1 \tag{3.75c}$$

$$\widetilde{\theta}(t_0 + \widetilde{c}_i h) = 0, \quad i = 0, \ldots, s. \tag{3.75d}$$

Writing $t_1 := t_0 + h$, the derivative of $\widetilde{\theta}$ also satisfies

$$\begin{aligned}\dot{\widetilde{\theta}}(t_0) &= -g_t(t_0,y_0) - g_y(t_0,y_0)(v(t_0,y_0,z_0 - h\widetilde{b}_0\widetilde{\mu}(t_0)) + \delta(t_0)) \\ &= -g_t(t_0,y_0) - g_y(t_0,y_0)\delta(t_0) - g_y(t_0,y_0)v(t_0,y_0,z_0) \\ &\qquad + h\widetilde{b}_0 g_y(t_0,y_0)v_z(t_0,y_0,z_0)\widetilde{\mu}(t_0) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_0)||^2\right) \\ &= -g_y(t_0,Y(t_0))\delta(t_0) \\ &\quad + h\widetilde{b}_0 g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_0)||^2\right),\end{aligned} \tag{3.76}$$

$$\begin{aligned}\dot{\widetilde{\theta}}(t_1) &= -g_t(t_1,y_1) - g_y(t_1,y_1)(v(t_1,y_1,z_1 + h\widetilde{b}_s\widetilde{\mu}(t_1)) + \delta(t_1)) \\ &= -g_t(t_1,y_1) - g_y(t_1,y_1)\delta(t_1) - g_y(t_1,y_1)v(t_1,y_1,z_1) \\ &\qquad - h\widetilde{b}_s g_y(t_1,y_1)v_z(t_1,y_1,z_1)\widetilde{\mu}(t_1) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_1)||^2\right) \\ &= -g_y(t_1,Y(t_1))\delta(t_1) \\ &\quad - h\widetilde{b}_s g_y(t_1,Y(t_1))v_z(t_1,Y(t_1),Z(t_1))\widetilde{\mu}(t_1) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_1)||^2\right).\end{aligned} \tag{3.77}$$

From the proof of Theorem 3.5.5 (see (3.72)), we get that

$$\mathcal{O}(h^2||\widetilde{\mu}(t_0)||^2) = \mathcal{O}(h^2||\widetilde{\mu}(t_1)||^2) = \mathcal{O}(h^{2s}).$$

Note that the exact solution satisfies these same relations, but with $\delta \equiv 0$, $\widetilde{\theta} \equiv 0$, and $\mu + \widetilde{\mu} \equiv 0$. Using the relations for the exact solution and (3.74), we obtain

$$\begin{aligned}\dot{Y}(t) - \dot{y}(t) &= v(t,Y(t),Z(t)) - v(t,y(t),z(t)) + \delta(t) \\ \dot{Z}(t) - \dot{z}(t) &= f(t,Y(t),Z(t)) - f(t,y(t),z(t)) \\ &\qquad + r(t,Y(t),\Lambda(t)) - r(t,y(t),\lambda(t)) + \mu(t) + \widetilde{\mu}(t).\end{aligned}$$

After linearizing and using the notation $\Delta Y(t) := Y(t) - y(t)$, $\Delta Z(t) := Z(t) - z(t)$, and $\Delta\Lambda(t) := \Lambda(t) - \lambda(t)$, these reduce to

$$\begin{aligned}\dot{Y}(t) - \dot{y}(t) = v_y(t,y(t),z(t))\Delta Y(t) + v_z(t,y(t),z(t))\Delta Z(t) + \delta(t) \\ + \mathcal{O}\left(||\Delta Y(t)||^2 + ||\Delta Z(t)||^2\right)\end{aligned} \tag{3.78}$$

$$\begin{aligned}\dot{Z}(t) - \dot{z}(t) = {} & f_y(t, y(t), z(t))\Delta Y(t) + f_z(t, y(t), z(t))\Delta Z(t) \\ & + r_y(t, y(t), \lambda(t))\Delta Y(t) + r_\lambda(t, y(t), \lambda(t))\Delta\Lambda(t) + \mu(t) \\ & + \widetilde{\mu}(t) + \mathcal{O}\left(||\Delta Y(t)||^2 + ||\Delta Z(t)||^2 + ||\Delta\Lambda(t)||^2\right).\end{aligned} \tag{3.79}$$

The next goal is to solve for $\Delta\Lambda$. Taking the derivative of (3.74e) gives

$$\begin{aligned}0 = {} & g_{tt}(t, Y(t)) + g_{ty}(t, Y(t))(v(t, Y(t), Z(t)) + \delta(t)) \\ & + [g_{yt}(t, Y(t)) + g_{yy}(t, Y(t))(v(t, Y(t), Z(t)) + \delta(t))] \cdot \\ & \qquad [v(t, Y(t), Z(t)) + \delta(t)] \\ & + g_y(t, Y(t))\big[v_t(t, Y(t), Z(t)) + v_y(t, Y(t), Z(t))(v(t, Y(t), Z(t)) + \delta(t)) \\ & \qquad + v_z(t, Y(t), Z(t))(f(t, Y(t), Z(t)) + r(t, Y(t), \Lambda(t)) \\ & \qquad + \mu(t) + \widetilde{\mu}(t)) + \dot{\delta}(t)\big] \\ & + \ddot{\widetilde{\theta}}(t).\end{aligned}$$

The exact solution satisfies the similar condition

$$\begin{aligned}0 = {} & g_{tt}(t, y(t)) + g_{ty}(t, y(t))v(t, y(t), z(t)) \\ & + [g_{yt}(t, y(t)) + g_{yy}(t, y(t))v(t, y(t), z(t))]\, v(t, y(t), z(t)) \\ & + g_y(t, y(t))\big[v_t(t, y(t), z(t)) + v_y(t, y(t), z(t))v(t, y(t), z(t)) \\ & \qquad + v_z(t, y(t), z(t))(f(t, y(t), z(t)) + r(t, y(t), \lambda(t)))\big].\end{aligned}$$

Suppressing the argument $t$ on the exact solution, we subtract these two equations and linearize, giving a relation of the form

$$\begin{aligned}0 = {} & F_1(t)\Delta Y(t) + F_2(t)\Delta Z(t) + g_y(t, y)v_z(t, y, z)r_\lambda(t, y, \lambda)\Delta\Lambda(t) \\ & + G(t)\delta(t) + g_y(t, Y)v_z(t, Y(t), Z(t))(\mu(t) + \widetilde{\mu}(t)) \\ & + g_y(t, Y(t))\dot{\delta}(t) + \ddot{\widetilde{\theta}}(t).\end{aligned} \tag{3.80}$$

The functions $F_1(t)$ and $F_2(t)$ depend only upon the exact solution, not on the discontinuous collocation type polynomials, while the function $G(t)$ depends only upon the discontinuous collocation type polynomials. Note that there should be a term $\mathcal{O}(||\Delta Y||^2 + ||\Delta Z||^2 + ||\delta|| \cdot ||\Delta Y|| + \cdots)$. This term is omitted for simplicity,

as it does not change the result. Because $g_y(t,y)v_z(t,y,z)r_\lambda(t,y,\lambda)$ is invertible, we can solve for $\Delta\Lambda$ from this equation. Defining $\Upsilon := r_\lambda(-g_y v_z r_\lambda)^{-1}$, and suppressing dependence upon time for the exact solution $y = y(t)$, $z = z(t)$, and $\lambda = \lambda(t)$,

$$\begin{aligned}
\Delta\dot{Y}(t) &= v_y(t,y,z)\Delta Y(t) + v_z(t,y,z)\Delta Z(t) + \delta(t) \\
\Delta\dot{Z}(t) &= \big[f_y(t,y,z) + r_\lambda(t,y,\lambda) + \Upsilon(t)F_1(t)\big]\Delta Y(t) \\
&\quad + \big[f_z(t,y,z) + \Upsilon(t)F_2(t)\big]\Delta Z(t) \\
&\quad + \Upsilon(t)G(t)\delta(t) + \big[\Upsilon(t)g_y(t,Y(t))v_z(t,Y(t),Z(t)) + I_{n_z}\big](\mu(t) + \widetilde{\mu}(t)) \\
&\quad + \Upsilon(t)g_y(t,Y(t))\dot{\delta}(t) + \Upsilon(t)\ddot{\widetilde{\theta}}(t).
\end{aligned}$$

Taking as the resolvent

$$R(t,s) = \begin{bmatrix} R_{11}(t,s) & R_{12}(t,s) \\ R_{21}(t,s) & R_{22}(t,s) \end{bmatrix}, \quad R(t,t) = I_{n_y+n_z},$$

the variation of constants formula ([9, Theorem I.11.2]) gives

$$\begin{aligned}
\Delta Y(t_1) &= R_{12}(t_1,t_0)\Delta Z(t_0) + \int_{t_0}^{t_1} \Big[R_{11}(t_1,s)\delta(s) + R_{12}(t_1,s)G(s)\delta(s) \\
&\quad + R_{12}(t_1,s)\left[\Upsilon(s)g_y(s,Y(s))v_z(s,Y(s),Z(s)) + I_{n_z}\right](\mu(s) + \widetilde{\mu}(s)) \\
&\quad + R_{12}(t_1,s)\Upsilon(s)g_y(s,Y(s))\dot{\delta}(s) + R_{12}(t_1,s)\Upsilon(s)\ddot{\widetilde{\theta}}(s)\Big]ds \qquad (3.81) \\
\Delta Z(t_1) &= R_{22}(t_1,t_0)\Delta Z(t_0) + \int_{t_0}^{t_1} \Big[R_{21}(t_1,s)\delta(s) + R_{22}(t_1,s)G(s)\delta(s) \\
&\quad + R_{22}(t_1,s)\left[\Upsilon(s)g_y(s,Y(s))v_z(s,Y(s),Z(s)) + I_{n_z}\right](\mu(s) + \widetilde{\mu}(s)) \\
&\quad + R_{22}(t_1,s)\Upsilon(s)g_y(s,Y(s))\dot{\delta}(s) + R_{22}(t_1,s)\Upsilon(s)\ddot{\widetilde{\theta}}(s)\Big]ds. \qquad (3.82)
\end{aligned}$$

Beginning with $\Delta Y(t_1) = y_1 - y(t_1)$, we apply an integration by parts formula to two of the integrands. This gives

$$\begin{aligned}
\int_{t_0}^{t_1} R_{12}(t_1,s)\Upsilon(s)g_y(s,Y(s))\dot{\delta}(s)ds &= R_{12}(t_1,s)\Upsilon(s)g_y(s,Y(s))\delta(s)\Big|_{s=t_0}^{t_1} \\
&\quad - \int_{t_0}^{t_1} \frac{d}{ds}\left(R_{12}(t_1,s)\Upsilon(s)g_y(s,Y(s))\right)\delta(s)ds
\end{aligned}$$

$$= -R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))\delta(t_0) + \mathcal{O}(h^{2s+1}),$$

and

$$\begin{aligned}
\int_{t_0}^{t_1} R_{12}(t_1,s)\Upsilon(s)\ddot{\widetilde{\theta}}(s)ds &= R_{12}(t_1,s)\Upsilon(s)\dot{\widetilde{\theta}}(s)\Big|_{s=t_0}^{t_1} \\
&\qquad - \int_{t_0}^{t_1} \frac{d}{ds}(R_{12}(t_1,s)\Upsilon(s))\dot{\widetilde{\theta}}(s)ds \\
&= R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))\delta(t_0) \\
&\quad - h\widetilde{b}_0 R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\
&\quad - \frac{d}{ds}(R_{12}(t_1,s)\Upsilon(s))\widetilde{\theta}(s)\Big|_{s=t_0}^{t_1} + \int_{t_0}^{t_1} \frac{d^2}{ds^2}(R_{12}(t_1,s)\Upsilon(s))\widetilde{\theta}(s)ds \\
&\quad + \mathcal{O}\left(h^2||\widetilde{\mu}(t_0)||^2\right) \\
&= R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))\delta(t_0) \\
&\qquad - h\widetilde{b}_0 R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\
&\qquad + \mathcal{O}\left(h^2||\widetilde{\mu}(t_0)||^2 + h^{2s}\right).
\end{aligned}$$

Terms of the form $\mathcal{O}(h^{2s+1})$ are introduced in the two expressions above by applying Gauss and Lobatto quadratures, respectively, to the integrals in the last steps. Substituting these into (3.81), applying Gaussian quadrature on each term with a $\delta(s)$ or $\mu(s)$, and Lobatto quadrature on the term with a $\widetilde{\mu}(s)$ gives

$$\begin{aligned}
y_1 - y(t_1) = \Delta Y(t_1) &= R_{12}(t_1,t_0)\Delta Z(t_0) \\
&\quad + h\widehat{b}_0 R_{12}(t_1,t_0)(\Upsilon(t_0)g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0)) + I_{n_z})\widetilde{\mu}(t_0) \\
&\quad + R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))\delta(t_0) \\
&\quad - R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))\delta(t_0) \\
&\quad - h\widetilde{b}_0 R_{12}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\
&\quad + \mathcal{O}(h^{2s}) \\
&= R_{12}(t_1,t_0)(\Delta Z(t_0) + h\widetilde{b}_0\widetilde{\mu}(t_0)) + \mathcal{O}(h^{2s}) \\
&= \mathcal{O}(h^{2s}).
\end{aligned}$$

This means that the numerical approximation $y_1$ is of local order at least $2s-1$. However, because the Gauss-Lobatto methods are symmetric, the local order must therefore be even and equal to $2s$. The $\Delta Z(t_1)$ expression can be handled similarly. We integrate by parts the terms

$$\int_{t_0}^{t_1} R_{22}(t_1,s)\Upsilon(s)g_y(t_1,Y(s))\dot{\delta}(s)ds = \Upsilon(t_1)g_y(t_1,Y(t_1))\delta(t_1) - R_{22}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))\delta(t_0) + \mathcal{O}(h^{2s+1})$$

and

$$\begin{aligned}\int_{t_0}^{t_1} R_{22}(t_1,s)\Upsilon(s)\ddot{\widetilde{\theta}}(s)ds = &-h\widetilde{b}_s\Upsilon(t_1)g_y(t_1,Y(t_1))v_z(t_1,Y(t_1),Z(t_1))\\ &- h\widetilde{b}_0R_{22}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0)\\ &+ R_{22}(t_1,t_0)\Upsilon(t_0)g_y(t_0,Y(t_0))\delta(t_0)\\ &- \Upsilon(t_1)g_y(t_1,Y(t_1))\delta(t_1) + \mathcal{O}(h^{2s}).\end{aligned}$$

Once again, applying these relations, Gaussian quadrature, Lobatto quadrature, and (3.75) to (3.82) gives

$$\begin{aligned}z_1 - z(t_1) &= \Delta Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)\\ &= R_{22}(t_1,t_0)\left(\Delta Z(t_0) + h\widetilde{b}_0\widetilde{\mu}(t_0)\right) + \mathcal{O}(h^{2s})\\ &= \mathcal{O}(h^{2s}).\end{aligned}$$

Again, because the Gauss-Lobatto methods are symmetric, this result shows that $z_1$ is of local order $2s$. □

# CHAPTER 4
# BACKWARD ERROR ANALYSIS FOR SPARK METHODS FOR HAMILTONIAN SYSTEMS WITH HOLONOMIC CONSTRAINTS

## 4.1 Introduction

In this chapter, we show that symplectic SPARK methods applied to autonomous Hamiltonian systems preserve the total energy of the system. An autonomous Hamiltonian system with holonomic constraints is of the form

$$\dot{q} = \nabla_p H(q,p) \tag{4.1a}$$

$$\dot{p} = -\nabla_q H(q,p) - G(q)^T \lambda \tag{4.1b}$$

$$0 = g(q) \tag{4.1c}$$

$$0 = g_q(q)\nabla_p H(q,p) \tag{4.1d}$$

where $q(t) \in \mathbb{R}^{n_q}$, $p(t) \in \mathbb{R}^{n_q}$, $\lambda(t) \in \mathbb{R}^{n_g}$, and the functions

$$H : \mathbb{R}^{n_q} \times \mathbb{R}^{n_q} \to \mathbb{R}$$

$$G : \mathbb{R}^{n_q} \to \mathbb{R}^{n_g \times n_q}$$

$$g : \mathbb{R}^{n_q} \to \mathbb{R}^{n_g}.$$

We shall assume time independent (scleronomic) constraints for simplicity. We also assume ideal constraints, i.e., we have that $G(q) = g_q(q)$. The system (4.1) is assumed to have consistent initial values $(q_0, p_0)$ at time $t_0$.

We also make the assumption that the matrix

$$G(q)\nabla^2_{pp} H(q,p) G(q)^T \text{ is invertible.} \tag{4.2}$$

We define the manifold

$$\mathcal{M} := \{(q,p) \in \mathbb{R}^n \times \mathbb{R}^n \mid 0 = g(q), 0 = g_q(q)\nabla_p H(q,p)\}. \tag{4.3}$$

If we have consistent initial values (that is, $(q_0, p_0) \in \mathcal{M}$), then a unique solution exists, as shown in [10]. We assume for this chapter that the functions $H(q,p)$ and $g(q)$ are sufficiently smooth in an open set $D$ around $\mathcal{M}$.

For this chapter, we assume the Hamiltonian function $H$ is independent of time. For systems (4.1) with time dependent Hamiltonians, the results presented in this chapter can be extended. The flow of time independent Hamiltonian systems is symplectic and preserves the Hamiltonian, i.e.,

$$\frac{d}{dt}H(q(t),p(t)) = 0$$

for $q(t)$ and $p(t)$ the solution to (4.1). For mechanical systems, the Hamiltonian corresponds to the total energy of the system, and thus the preservation of $H$ reflects the conservation of the total energy of the system.

## 4.2 SPARK Methods

We present SPARK methods for solving systems of the form (4.1). These methods were considered in greater detail in Chapter 3. We restate the methods here for convenience.

**Definition 4.2.1.** *One step of an $(s,\widetilde{s})$-stage specialized partitioned additive Runge-Kutta (SPARK) method applied to the system (4.1) with stepsize $h$ starting at $(q_0,p_0)$ at time $t_0$ is given by the solution of the nonlinear system of equations*

$$Q_i = q_0 + h\sum_{j=1}^{s} a_{ij}\nabla_p H(Q_j,P_j), \quad i = 1,\dots,s \tag{4.4a}$$

$$\widetilde{Q}_i = q_0 + h\sum_{j=1}^{s} \bar{a}_{ij}\nabla_p H(Q_j,P_j), \quad i = 0,\dots,\widetilde{s} \tag{4.4b}$$

$$P_i = p_0 - h\sum_{j=1}^{s} \widehat{a}_{ij}\nabla_q H(Q_j,P_j) - h\sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij} g_q(\widetilde{Q}_j)^T\Lambda_j, \quad i = 1,\dots,s \tag{4.4c}$$

$$q_1 = q_0 + h\sum_{j=1}^{s} b_j\nabla_p H(Q_j,P_j) \tag{4.4d}$$

$$p_1 = p_0 - h\sum_{j=1}^{s} \widehat{b}_j \nabla_q H(Q_j, P_j) - h\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j g_q(\widetilde{Q}_j)^T \Lambda_j \tag{4.4e}$$

$$0 = g(\widetilde{Q}_i), \quad i = 0, \ldots, \widetilde{s} \tag{4.4f}$$

$$0 = g(q_1) \tag{4.4g}$$

$$0 = g_q(q_1)\nabla_p H(q_1, p_1). \tag{4.4h}$$

*The coefficients $c_i$ and $\widetilde{c}_i$ are determined by*

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad \widetilde{c}_i = \sum_{j=1}^{s} \bar{a}_{ij}. \tag{4.5}$$

It is assumed that the RK coefficients satisfy

$$\bar{a}_{0j} = 0, \quad j = 1, \ldots, s, \tag{4.6a}$$

$$\bar{a}_{\widetilde{s}j} = b_j, \quad j = 1, \ldots, s, \tag{4.6b}$$

$$\sum_{j=1}^{s} \bar{a}_{ij} c_j = \sum_{j=1}^{s} \bar{a}_{ij} \sum_{k=1}^{s} \widehat{a}_{jk} = \sum_{j=1}^{s} \bar{a}_{ij} \sum_{k=0}^{\widetilde{s}} \widetilde{a}_{jk} = \frac{\widetilde{c}_i^2}{2}, \quad i = 0, \ldots, \widetilde{s}, \tag{4.6c}$$

$$\bar{A}\widetilde{A} = \begin{bmatrix} 0 & \ldots & 0 \\ & \bar{A}^*\widetilde{A} & \end{bmatrix}, \quad \begin{bmatrix} \bar{A}^*\widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \text{ is invertible,} \tag{4.6d}$$

where $\bar{A}^* \in \mathbb{R}^{s\times s}$ equals $\bar{A}$ with the first row removed. These assumptions are made in [16] and Chapter 3 for the existence and uniqueness of a numerical solution for the system (4.1). It is shown in [16] that, with the following assumptions on the coefficients, the SPARK methods are symplectic mappings.

**Theorem 4.2.2.** *If the SPARK method (4.4) applied to (4.1) satisfies*

$$\widehat{b}_i = b_i, \quad i = 1, \ldots, s \tag{4.7a}$$

$$\widehat{b}_i a_{ij} + b_j \widehat{a}_{ji} = \widehat{b}_i b_j, \quad i, j = 1, \ldots, s \tag{4.7b}$$

$$\widetilde{b}_i \bar{a}_{ij} + b_j \widetilde{a}_{ji} = \widetilde{b}_i b_j, \quad i = 0, 1, \ldots, \widetilde{s},\ j = 1, \ldots, s, \tag{4.7c}$$

*then the numerical flow $(q_0, p_0) \to (q_1, p_1)$ preserves on $\mathcal{M}$ the 2-form $\sum_{i=1}^{n} dp^i \wedge dq^i$.*

For example, the Gauss-Lobatto SPARK methods presented in Chapter 3

satisfy these conditions.

Our goal is to show that the method (4.4) preserves approximately the Hamiltonian of the original problem (4.1). We do this by using a backward error analysis approach that utilizes the symplectic structure of the SPARK methods. In other words, we want to show that the modified equation for the method (4.1) is a Hamiltonian system. For standard RK-methods for problems with holonomic constraints, this type of analysis was performed in [6] and [8].

Our first goal is to extend the numerical method to a neighborhood of the manifold $\mathcal{M}$.

### 4.3 Generating Function

The numerical solution (4.4) is well-defined only if the initial conditions $(q_0, p_0)$ lie in the manifold $\mathcal{M}$. We can change this by replacing the conditions (4.4f,g,h) by

$$0 = g(\widetilde{Q}_i) - g(q_0) - \widetilde{c}_i h g_q(q_0)\nabla_p H(q_0, p_0), \quad i = 1, \ldots, s \tag{4.8a}$$

$$0 = g(q_1) - g(q_0) - h g_q(q_0)\nabla_p H(q_0, p_0) \tag{4.8b}$$

$$0 = g_q(q_1)\nabla_p H(q_1, p_1) - g_q(q_0)\nabla_p H(q_0, p_0). \tag{4.8c}$$

Using these, we can now consider $(q_0, p_0) \in D$, an open neighborhood around $\mathcal{M}$. Notice that for $(q_0, p_0) \in \mathcal{M}$, the constraints (4.8) reduce to (4.4f,g,h). The extended method (4.4a-e), (4.8), however, may not be symplectic.

Because of (4.4d,e), we can view $q_1$, $p_0$, $Q_i$, $\widetilde{Q}_i$, $P_i$, $\Lambda_i$, as functions of $q_0$, $p_1$, and $h$. For notational simplicity, we define $H[i] := H(Q_i, P_i)$. We now define the function $S$ as

$$\begin{aligned} S(q_0, p_1, h) := {} & h\sum_{i=1}^{s} b_i H[i] + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g(\widetilde{Q}_i)^T \Lambda_i \\ & - h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} \widehat{b}_i a_{ij} \nabla_q H[i]^T \nabla_p H[j] - h^2 \sum_{j=1}^{s}\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \bar{a}_{ij} (g_q(\widetilde{Q}_i)^T \Lambda_i)^T \nabla_p H[j]. \end{aligned} \tag{4.9}$$

In light of the following lemma, the function $S$ is referred to as a *generating function (of type I)* for the method (4.4a-e), (4.8).

**Lemma 4.3.1.** *Let the coefficients $b_i$, $\widehat{b}_i$, $\widetilde{b}_i$, $a_{ij}$, $\bar{a}_{ij}$, $\widehat{a}_{ij}$, $\widetilde{a}_{ij}$ satisfy the symplecticity conditions (4.7). Then, for the function $S$ in (4.9) the numerical method given by*

$$q_1 = q_0 + \nabla_{p_1} S(q_0, p_1, h) \qquad p_0 = p_1 + \nabla_{q_0} S(q_0, p_1, h) \tag{4.10}$$

*defines a symplectic extension of the SPARK method (4.4) to an open neighborhood of $\mathcal{M}$.*

*Proof.* The symplecticity of the mapping $(q_0, p_0) \to (q_1, p_1)$, which is implicitly defined in (4.10), follows from the theory of generating functions mentioned in Section 1.2.2 and presented in [8].

We must also prove that (4.10) is an extension to the SPARK method (4.4). To do this, we calculate explicitly the gradients of the function $S$. For $\nabla_{q_0} S(q_0, p_1, h)$, we get

$$\begin{aligned}
\nabla_{q_0} S(q_0, p_1, h) = {} & h \sum_{i=1}^{s} b_i \left( (\partial_{q_0} Q_i)^T \nabla_q H[i] + (\partial_{q_0} P_i)^T \nabla_p H[i] \right) \\
& + h \sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \left( (\partial_{q_0} \widetilde{Q}_i)^T g_q(\widetilde{Q}_i)^T \Lambda_i + (\partial_{q_0} \Lambda_i)^T g(\widetilde{Q}_i) \right) \\
& - h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} \widehat{b}_i a_{ij} \left( (\partial_{q_0} Q_i)^T \nabla^2_{qq} H[i] \nabla_p H[j] + (\partial_{q_0} P_i)^T \nabla^2_{pq} H[i] \nabla_p H[j] \right. \\
& \quad \left. + (\partial_{q_0} Q_j)^T \nabla^2_{qp} H[j] \nabla_q H[i] + (\partial_{q_0} P_j)^T \nabla^2_{pp} H[j] \nabla_q H[i] \right) \\
& - h^2 \sum_{j=1}^{s} \sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \bar{a}_{ij} \left( (\partial_{q_0} \widetilde{Q}_i)^T (\partial_{qq} g(\widetilde{Q}_i))^T (\nabla_p H[j], \Lambda_i) \right. \\
& \quad + (\nabla_{q_0} \Lambda_i)^T g_q(\widetilde{Q}_i) \nabla_p H[j] + (\partial_{q_0} P_j)^T \nabla^2_{pp} H[j] g_q(\widetilde{Q}_i)^T \Lambda_i \\
& \quad \left. + (\partial_{q_0} Q_j)^T \nabla^2_{qp} H[j] g_q(\widetilde{Q}_i)^T \Lambda_i \right).
\end{aligned}$$

From here, we can substitute the values for $\partial_{q_0} Q_i$, $\partial_{q_0} \widetilde{Q}_i$, and $\partial_{q_0} P_i$ from (4.4).

Applying the symplecticity conditions (4.7), the above calculation reduces to

$$\begin{aligned}
\nabla_{q_0} S(q_0, p_1, h) &= h\sum_{i=1}^{s} b_i \nabla_q H[i] + h\sum_{i=1}^{s} b_i (\partial_{q_0} p_0)^T \nabla_p H[i] \\
&+ h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g_q(\widetilde{Q}_i)^T \Lambda_i + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i (\partial_{q_0}\Lambda_i)^T g(\widetilde{Q}_i) \\
&- h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i \widehat{b}_j (\partial_{q_0}\{\nabla_q H[j]\})^T \nabla_p H[i] \\
&- h^2 \sum_{i=1}^{s}\sum_{j=0}^{\widetilde{s}} b_i \widetilde{b}_j (\partial_{q_0}\{g_q(\widetilde{Q}_j)^T \Lambda_j\})^T \nabla_p H[i] \\
&= h\sum_{i=1}^{s} b_i \nabla_q H[i] + h\sum_{i=1}^{s} b_i (\partial_{q_0} p_0)^T \nabla_p H[i] \\
&+ h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g_q(\widetilde{Q}_i)^T \Lambda_i + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i (\partial_{q_0}\Lambda_i)^T g(\widetilde{Q}_i) \\
&- h^2 \sum_{i=1}^{s} b_i \left( \sum_{j=1}^{s} \widehat{b}_j (\partial_{q_0}\{\nabla_q H[j]\})^T + \sum_{j=0}^{\widetilde{s}} \widetilde{b}_j (\partial_{q_0}\{g_q(\widetilde{Q}_j)^T \Lambda_j\})^T \right) \nabla_p H[i].
\end{aligned} \tag{4.11}$$

(4.12)

However, taking the partial derivatives with respect to $q_0$ of (4.4e), we find that

$$(\partial_{q_0} p_0)^T = h\sum_{j=1}^{s} \widehat{b}_j (\partial_{q_0}\{\nabla_q H[j]\})^T + h\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j (\partial_{q_0}\{g_q(\widetilde{Q}_j)^T \Lambda_j\})^T.$$

This, along with the symplecticity condition $b_i = \widehat{b}_i$ for $i = 1, ..., s$, show that (4.12) is

$$\nabla_{q_0} S(q_0, p_1, h) = h\sum_{i=1}^{s} \widehat{b}_i \nabla_q H[i] + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g_q(\widetilde{Q}_i)^T \Lambda_i + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i (\partial_{q_0}\Lambda_i)^T g(\widetilde{Q}_i).$$

We can also calculate $\nabla_{p_1} S$ in a similar manner. We have

$$\begin{aligned}
\nabla_{p_1} S(q_0, p_1, h) &= h\sum_{i=1}^{s} b_i \left( (\partial_{p_1} Q_i)^T \nabla_q H[i] + (\partial_{p_1} P_i)^T \nabla_p H[i] \right) \\
&+ h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \left( (\partial_{p_1} \widetilde{Q}_i)^T g_q(\widetilde{Q}_i)^T \Lambda_i + (\partial_{p_1}\Lambda_i)^T g(\widetilde{Q}_i) \right) \\
&- h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} \widehat{b}_j a_{ij} \left( (\partial_{p_1} Q_i)^T \nabla^2_{qq} H[i] \nabla_p H[j] + (\partial_{p_1} P_i)^T \nabla^2_{pq} H[i] \nabla_p H[j] \right.
\end{aligned}$$

$$
\begin{aligned}
&+(\partial_{p_1} Q_j)^T \nabla^2_{qp} H[j] \nabla_q H[i] + (\partial_{p_1} P_j)^T \nabla^2_{pp} H[j] \nabla_q H[j]\big) \\
&- h^2 \sum_{j=1}^{s} \sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \bar{a}_{ij} \Big( (\partial_{p_1} \widetilde{Q}_i)^T (\partial_{qq} g(\widetilde{Q}_i))^T (\nabla_p H[j], \Lambda_i) \\
&\quad + (\partial_{p_1} \Lambda_i)^T g_q(\widetilde{Q}_i) \nabla_p H[j] \\
&\quad + (\partial_{p_1} P_j)^T \nabla^2_{pp} H[j] g_q(\widetilde{Q}_i)^T \Lambda_i + (\partial_{p_1} Q_j)^T \nabla^2_{qp} H[j] g_q(\widetilde{Q}_i)^T \Lambda_i \Big).
\end{aligned}
$$

Again, we can substitute the values for $\partial_{p_1} Q_i$, $\partial_{p_1} \widetilde{Q}_i$, and $\partial_{p_1} P_i$ from (4.4). Applying the symplecticity conditions (4.7), this calculation reduces to

$$
\begin{aligned}
\nabla_{p_1} S(q_0, p_1, h) &= h \sum_{i=1}^{s} b_i (\partial_{p_1} p_0)^T \nabla_p H[i] + h \sum_{i=0}^{\widetilde{s}} \widetilde{b}_i (\partial_{p_1} \Lambda_i g(\widetilde{Q}_i))^T \\
&- h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{b}_j (\partial_{p_1} \{\nabla_q H[j]\})^T \nabla_p H[i] \\
&- h^2 \sum_{i=1}^{s} \sum_{j=0}^{\widetilde{s}} b_i \widetilde{b}_j (\partial_{p_1} \{g_q(\widetilde{Q}_j)^T \Lambda_j\})^T \nabla_p H[i] \\
&= h \sum_{i=1}^{s} b_i (\partial_{p_1} p_0)^T \nabla_p H[i] + h \sum_{i=0}^{\widetilde{s}} \widetilde{b}_i (\partial_{p_1} \Lambda_i)^T g(\widetilde{Q}_i) \\
&- h^2 \sum_{i=1}^{s} b_i \left( \sum_{j=1}^{s} \widehat{b}_j (\partial_{p_1} \{\nabla_q H[j]\})^T + \sum_{j=0}^{\widetilde{s}} \widetilde{b}_j (\partial_{p_1} \{g_q(\widetilde{Q}_j)^T \Lambda_j\})^T \right) \nabla_p H[i].
\end{aligned}
$$

Taking the partial derivative with respect to $p_1$ of the equation (4.4e), we can find $\nabla_{p_1} p_0$ as we did with $\nabla_{q_0} p_0$. This gives

$$
I_{n_q} = \nabla_{p_1} p_0 - h \sum_{j=1}^{s} \widehat{b}_j \nabla_{p_1} \{\nabla_q H[j]\} - h \sum_{j=0}^{\widetilde{s}} \widetilde{b}_j (\partial_{p_1} \{g_q(\widetilde{Q}_j)^T \Lambda_j\})^T.
$$

Using this and simplifying, we finally get

$$
\nabla_{p_1} S(q_0, p_1, h) = h \sum_{i=0}^{\widetilde{s}} \widetilde{b}_i (\partial_{p_1} \Lambda_i)^T g(\widetilde{Q}_i) + h \sum_{i=1}^{s} b_i \nabla_p H[i].
$$

At last, we substitute the expressions found for $\nabla_{q_0} S$ and $\nabla_{p_1} S$ into (4.10). Recalling that $g(\widetilde{Q}_i) = 0$ for $(q_0, p_0) \in \mathcal{M}$, the equations in (4.10) can be written as

$$
q_1 = q_0 + h \sum_{i=1}^{s} b_i \nabla_p H[i]
$$

$$p_1 = p_0 - h\sum_{i=1}^{s} \widehat{b}_i \nabla_q H[i] - h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g_q(\widetilde{Q}_i)^T \Lambda_i$$

if $(q_0, p_0) \in \mathcal{M}$. We note that these expressions agree with those of $q_1$ and $p_1$ from (4.4d,e). This shows that (4.10) is an extension of the SPARK methods (4.4). □

We note that the generating function $S$ can be expanded in $h$ as

$$S(q, p, h) = hS_1(q, p) + h^2 S_2(q, p) + h^3 S_3(q, p) + ... \tag{4.13}$$

where each $S_i$ is smooth and well-defined on the domain $D$, and where the assumption (4.2) holds. If fact, each $S_i$ can be expressed in terms of the functions $H$, $g$, and of each $\Lambda_j$. However, because the $\Lambda_j$ terms depend upon the stepsize $h$ as well, they must also be expanded. We thus write

$$\Lambda_i(q, p, h) = \lambda_i^0(q, p) + h\lambda_i^1(q, p) + h^2\lambda_i^2(q, p) + \dots \tag{4.14}$$

We illustrate this with an example.

*Example.* We consider the SPARK midpoint-trapezoidal rule

$$Q_1 = q_0 + \frac{h}{2}\nabla_q H(Q_1, P_1) \tag{4.15a}$$

$$\widetilde{Q}_1 = q_0 + h\nabla_q H(Q_1, P_1) \tag{4.15b}$$

$$P_1 = p_0 - \frac{h}{2}\nabla_q H(Q_1, P_1) - \frac{h}{2} g_q(q_0)^T \Lambda_0 \tag{4.15c}$$

$$q_1 = q_0 + h\nabla_q H(Q_1, P_1) \tag{4.15d}$$

$$p_1 = p_0 - h\nabla_q H(Q_1, P_1) - \frac{h}{2} g_q(q_0)^T \Lambda_0 - \frac{h}{2} g_q(\widetilde{Q}_1)^T \Lambda_1 \tag{4.15e}$$

$$0 = g(\widetilde{Q}_1) \tag{4.15f}$$

$$0 = g_q(q_1)\nabla_p H(q_1, p_1). \tag{4.15g}$$

Notice that using (4.15e), $P_1$ can be expressed as

$$P_1 = p_1 + \frac{h}{2}\nabla_q H(Q_1, P_1) + \frac{h}{2} g_q(q_0)^T \Lambda_0 + \frac{h}{2} g_q(q_1)^T \Lambda_1. \tag{4.16}$$

By (4.9), we can write the generating function for this method as

$$\begin{aligned} S(q_0,p_1,h) = \frac{h}{2}\Big(2H(Q_1,P_1) + g(q_0)^T\Lambda_0 + g(q_1)^T\Lambda_1\Big) \\ - \frac{h^2}{2}\Big(\nabla_q H(Q_1,P_1)^T + (\nabla_q g(q_1)\Lambda_1)^T\Big)\nabla_p H(Q_1,P_1), \end{aligned} \tag{4.17}$$

with $Q_1$, $P_1$, $q_1$, $\Lambda_0$, and $\Lambda_1$ being interpreted as functions of $(q_0, p_1)$. To expand $S$ as a series in $h$, we must expand the functions $H(Q_1, P_1)$, $g(q_1)$, $\Lambda_0$, and $\Lambda_1$ in terms of $h$. Doing so gives

$$\begin{aligned} S(q_0,p_1,h) = h\left[H + \frac{1}{2}g^T(\lambda_0^0 + \lambda_1^0)\right] \\ + h^2\left[\frac{1}{2}g^T(\lambda_0^1 + \lambda_1^1) - \frac{1}{2}\nabla_q H^T\nabla_p H\right] \\ + h^3\ldots \end{aligned}$$

with each function evaluated at $(q_0, p_1)$. The functions $\lambda_i^j$ can be solved for in terms of derivatives of $H$ and $g$ by expanding in $h$ the constraints for the SPARK midpoint-trapezoidal rule

$$0 = g(q_1) \tag{4.18a}$$

$$0 = g_q(q_1)^T\nabla_p H(q_1,p_1). \tag{4.18b}$$

For instance, expanding the first constraint gives

$$\begin{aligned} 0 = g + h\,(g_q\nabla_p H) + \frac{h^2}{2}\Big(g_{qq}(\nabla_p H,\nabla_p H) + g_q\nabla^2_{pq}H\nabla_p H \\ + g_q\nabla^2_{pp}H\nabla_q H + g_q\nabla^2_{pp}H\nabla_q g\lambda_0^0 + g_q\nabla^2_{pp}H\nabla_q g\lambda_1^0\Big) \\ + \mathcal{O}(h^3) \end{aligned} \tag{4.19}$$

with all function evaluated at $(q_0, p_1)$. This can be used to solve for $\lambda_0^0 + \lambda_1^0$ since the matrix $g_q\nabla^2_{pp}Hg_q^T$ is assumed invertible in (4.2). Expanding out more terms and expanding (4.18b) allows for more $\lambda_i^j$ terms to be found.

## 4.4 Modified Hamiltonian

The modified Hamiltonian for numerical methods of the form (4.10) with a generating function can be obtained from the Hamilton-Jacobi equations. This construction process is given in [6] and [8]. We describe it here.

*First Step.* Let $\widetilde{H}(q,p,h)$ be the modified Hamiltonian for the system given by (4.10). We can expand this in $h$ as

$$\widetilde{H}(q,p,h) = H_1(q,p) + hH_2(q,p) + h^2H_3(q,p) + ... \tag{4.20}$$

and use the fact that the exact solution $(Q,P) = \widetilde{\varphi}_t(q_0,p_0)$ for the system (4.10) is given by

$$Q = q_0 + \nabla_p \widetilde{S}(q_0,P,t,h) \qquad p_0 = P + \nabla_q \widetilde{S}(q_0,P,t,h). \tag{4.21}$$

The perturbed generating function $\widetilde{S}(q,p,t)$ is the solution to the Hamilton-Jacobi equation

$$\begin{aligned} \frac{\partial \widetilde{S}}{\partial t}(q,p,t) &= \widetilde{H}(q + \nabla_p \widetilde{S}(q,p,t),p,h) \\ \widetilde{S}(q,p,0) &= 0. \end{aligned} \tag{4.22}$$

Note that because $\widetilde{H}$ depends upon the stepsize $h$, the perturbed generating function must also depend upon $h$. Hence we write $\widetilde{S}(q,p,t) = \widetilde{S}(q,p,t,h)$. We express the function $\widetilde{S}$ as a series in $t$

$$\widetilde{S}(q,p,t,h) = t\widetilde{S}_1(q,p,h) + t^2\widetilde{S}_2(q,p,h) + t^3\widetilde{S}_3(q,p,h) + \ldots, \tag{4.23}$$

and insert this into (4.22), and expand $\widetilde{H}(q + \nabla_p \widetilde{S}(q,p,t,h),p)$ into a series in $t$. Matching powers of $t$, we obtain terms for each $\widetilde{S}_j(q,p,h)$. For example, we obtain $\widetilde{S}_1(q,p,h) = \widetilde{H}(q,p)$, and $2\widetilde{S}_2(q,p,h) = (\nabla_q \widetilde{H}^T \nabla_p \widetilde{S}_1)(q,p,h)$.

*Second step.* We now expand each $\widetilde{S}_j(q,p,h)$ as a series in $h$. We write the series as

$$\widetilde{S}_j = \widetilde{S}_{j1}(q,p) + h\widetilde{S}_{j2}(q,p) + h^2\widetilde{S}_{j3}(q,p) + \ldots.$$

Inserting this and (4.20) into the results of the first step, we can again match like powers of $h$. This results in the function $\widetilde{S}_{jk}$ expressed in terms of derivatives of the terms in the expansion of the modified Hamiltonian. For instance, we calculate $\widetilde{S}_{1k}(q,p) = H_k(q,p)$, $2\widetilde{S}_{2k} = \sum_{l=1}^{k} \nabla_q H_l^T \nabla_p H_{k-1}$. In the general case, we will have that each $\widetilde{S}_{jk}(q,p)$ can be expressed in terms of derivatives of $H_l$, for $l < k$.

*Third step.* Finally, we require the generating function (4.13) for the numerical method to be equal to the perturbed generating function $\widetilde{S}(q,p,h,h)$, i.e.

$$\widetilde{S}(q,p,h,h) = S(q,p,h).$$

Once more, by matching like powers of $h$, we obtain expressions such as $S_1(q,p) = \widetilde{S}_{11}(q,p)$, $S_2(q,p) = \widetilde{S}_{12}(q,p) + \widetilde{S}_{21}(q,p)$, etc. Using these with results of the second step, we find that each $S_j(q,p)$ can be expressed as $H_j(q,p)$ plus additional terms with the derivatives of $H_k(q,p)$ for $k < j$. For example,

$$\begin{aligned}
H_1 &= S_1 \\
H_2 &= S_2 - \frac{1}{2}\nabla_q H_1^T \nabla_p H_1 \\
H_3 &= S_3 - \frac{1}{2}\nabla_q H_1^T \nabla_p H_2 - \frac{1}{2}\nabla_q H_2^T \nabla_p H_1 \\
&\quad + \frac{1}{6}\nabla_q H_1^T \nabla_p \left(\nabla_q H_1^T \nabla_p H_1\right) + \frac{1}{3}\nabla_p H_1^T \nabla_{qq}^2 H_1 \nabla_p H_1.
\end{aligned}$$

Note also that from these formulas, each $H_i$ will have the same domain as $S_j$. This completes the construction of the modified Hamiltonian.

## 4.5 Main Result

With the construction above, we can present the main result of this chapter.

**Theorem 4.5.1.** *Let $H(q,p)$ and $g(q)$ be defined and smooth on a neighborhood $D$ of $\mathcal{M}$ where (4.2) is fulfilled. Let $\Phi_h(q,p) : \mathcal{M} \to \mathcal{M}$ be the discrete flow of a SPARK method that satisfies (4.7) when applied to a problem of the form (4.1).*

*Then there exist functions $H_k(q,p)$ defined and smooth on $D$ satisfying*

$$g_q(q)\nabla_p H_k(q,p) = 0 \quad \textit{for} \quad (q,p) \in \mathcal{M} \tag{4.24}$$

*so that for any $N \geq 1$ with*

$$\widetilde{H}_N^* := H(q,p) + hH_2^*(q,p) + ... + h^{N-1}H_N^*(q,p), \tag{4.25}$$

*we have*

$$\Phi_h(y) - \widetilde{\varphi}_h(y) = \mathcal{O}(h^{N+1}), \tag{4.26}$$

*where $\widetilde{\varphi}_t : \mathcal{M} \to \mathcal{M}$ denotes the exact flow of the system*

$$\dot{q} = \nabla_p \widetilde{H}_N^*(q,p) \tag{4.27a}$$

$$\dot{p} = -\nabla_q \widetilde{H}_N^*(q,p) - g_q(q)^T\lambda \tag{4.27b}$$

$$0 = g(q). \tag{4.27c}$$

*Proof.* To find $\widetilde{H}_N^*$, we use the function $\widetilde{H}$ constructed in the previous section. There are several important properties of this function which give us the result.

1. For $\widetilde{H}_N$ a truncation of the first $N$ terms of the perturbed Hamiltonian $\widetilde{H}$, the system
$$\dot{q} = \nabla_p \widetilde{H}_N(q,p), \quad \dot{p} = -\nabla_q \widetilde{H}_N(q,p) \tag{4.28}$$
is a differential equation on the manifold $\mathcal{M}$. This result is given in [8, Theorem IX.5.1]. This is the underlying perturbed ODE for the modified equation.

2. From part 1, we therefore have, for any $H_k(q,p)$,
$$g_q(q)\nabla_p H_k(q,p) = 0 \quad \text{for } (q,p) \in \mathcal{M}. \tag{4.29}$$

3. In (4.28), any term of the form $g(q)^T\mu(q,p)$, for some vector-valued function $\mu$, can be removed from the Hamiltonian $\widetilde{H}_N$, since these terms will be 0 when restricted to the manifold $\mathcal{M}$. However, each of these terms generates

additional terms of the form

$$g_q(q)^T\mu(q,p) + \mu_q(q,p)g(q) \quad \text{and} \quad \mu_p(q,p)g(q)$$

upon taking the gradients for the ODE (4.28). On $\mathcal{M}$, there will be non-zero terms $g_q(q)^T\mu(q,p)$ from the gradient with respect to $q$. We can separate these terms from the gradient of the perturbed Hamiltonian, giving a differential equation of the form

$$\dot{q} = \nabla_p \widetilde{H}_N^*(q,p) \tag{4.30}$$

$$\dot{p} = -\nabla_q \widetilde{H}_N^*(q,p) - g_q(q)^T\lambda \tag{4.31}$$

$$0 = g(q) \tag{4.32}$$

where $\lambda$ is the sum of all such $\mu(q,p)$ terms, and $\widetilde{H}_N^*$ is the same as $\widetilde{H}_N$ but with all terms of the form $g(q)^T\mu(q,p)$ removed.

4. By the construction of $\widetilde{H}$, we have that

$$H_1(q,p) = H(q,p) + \frac{1}{2}g(q)^T(\lambda_0^0 + \lambda_1^0).$$

This term will also be first term of $\widetilde{H}_N^*$, but will be reduced to simply $H(q,p)$. This shows that the first term of $\widetilde{H}_N^*$ will be $H(q,p)$.

All of these properties together prove the theorem. □

# CHAPTER 5
# SPARK METHODS FOR MIXED INDEX 2 AND INDEX 3 DAES

## 5.1 Introduction

In this chapter, we examine SPARK methods applied to problems with mixed index 2 and index 3 constraints. These methods are introduced in [22]. We consider the overdetermined system of mixed index 2 and 3 DAEs

$$\dot{y} = v(t, y, z) \tag{5.1a}$$

$$\dot{z} = f(t, y, z, \psi) + r(t, y, \lambda) \tag{5.1b}$$

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

$$0 = g_t(t, y) + g_y(t, y)v(t, y, z) \tag{5.1d}$$

$$0 = k(t, y, z) \tag{5.1e}$$

where $y(t) \in \mathbb{R}^{n_y}$, $z(t) \in \mathbb{R}^{n_z}$, $\lambda(t) \in \mathbb{R}^{n_g}$, and $\psi(t) \in \mathbb{R}^{n_k}$, and the functions

$$v : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_y}$$

$$f : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_k} \to \mathbb{R}^{n_z}$$

$$r : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_g} \to \mathbb{R}^{n_z}$$

$$g : \mathbb{R} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_g}$$

$$k : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_k}.$$

We also make the assumption that the matrices

$$g_y(t, y)v_z(t, y, z)r_\lambda(t, y, \lambda) \tag{5.2a}$$

$$\begin{bmatrix} g_y(t, y)v_z(t, y, z)r_\lambda(t, y, \lambda) & g_y(t, y)v_z(t, y, z)f_\psi(t, y, z, \psi) \\ k_z(t, y, z)r_\lambda(t, y, \lambda) & k_z(t, y, z)f_\psi(t, y, z, \psi) \end{bmatrix} \tag{5.2b}$$

are invertible. The invertibility of the matrix (5.2b) allows the system (5.1) to be expressed as a system of ODEs. This gives the existence and uniqueness of a solution to (5.1). The system

$$\frac{d}{dt}q(t,y) = v(t,y,z) \tag{5.3a}$$
$$\frac{d}{dt}p(t,y,z) = f(t,y,z,\psi) + r(t,y,\lambda) \tag{5.3b}$$
$$0 = g(t,y) \tag{5.3c}$$
$$0 = g_t(t,y) + g_y(t,y)q_y(t,y)^{-1}(v(t,y,z) - q_t(t,y)) \tag{5.3d}$$
$$0 = k(t,y,z) \tag{5.3e}$$

is a generalization of the system (5.1). Both constrained Hamiltonian and Lagrangian systems can be expressed in this form, with $q(t,y) = y$, $p(t,y,z) = z$ for Hamiltonian systems and $q(t,y) = y$, $p(t,y,z) = \nabla_z L(t,y,z)$ for Lagrangian systems. To insure existence and uniqueness of a solution, the following matrices are assumed invertible:

$$q_y(t,y) \tag{5.4a}$$
$$p_z(t,y,z) \tag{5.4b}$$
$$\begin{bmatrix} g_y(t,y)q_y(t,y)^{-1}v_z(t,y,z)p_z(t,y,z)^{-1} \\ k_z(t,y,z)p_z(t,y,z)^{-1} \end{bmatrix} \begin{bmatrix} r_\lambda(t,y,\lambda)^T \\ f_\psi(t,y,z,\psi)^T \end{bmatrix}^T. \tag{5.4c}$$

The matrix (5.4c) is the generalization of the matrix (5.2b). Under these assumptions, differentiating the left sides of (5.3a) and (5.3b) gives

$$\dot{y} = q_y(t,y)^{-1}(v(t,y,z) - q_t(t,y)) \tag{5.5a}$$
$$\begin{aligned}\dot{z} = p_z(t,y,z)^{-1}\big[&f(t,y,z,\psi) + r(t,y,\lambda) - p_t(t,y,z) \\ &- p_y(t,y,z)q_y(t,y)^{-1}(v(t,y,z) - q_t(t,y))\big].\end{aligned} \tag{5.5b}$$

Taking the derivative of (5.3d) and (5.3e), and substituting in (5.5), we arrive at a system of the form

$$g_y(t,y)q_y(t,y)^{-1}v_z(t,y,z)p_z(t,y,z)^{-1}(f(t,y,z,\psi)+r(t,y,\lambda))+A(t,y,z)=0$$
$$k_z(t,y,z)p_z(t,y,z)^{-1}(f(t,y,z,\psi)+r(t,y,\lambda))+B(t,y,z)=0,$$

with $A(t,y,z)$ and $B(t,y,z)$ contains sums and products of derivatives of $v$, $f$, $r$, $g$, $k$, $q_t$, $p_t$, $q_y^{-1}$, and $p_z^{-1}$. This determines uniquely the terms $\lambda$ and $\psi$ by (5.4c) and the implicit function theorem.

Following [16], we define the new variables $q$ and $p$ satisfying the relations

$$q=q(t,y),\quad p=p(t,y,z).$$

By (5.4a) and (5.4b) we can express $y$ and $z$ as functions of $t$, $q$, and $p$. Defining

$$V(t,q,p):=v(t,y(t,q,p),z(t,q,p)),$$
$$F(t,q,p,\psi):=f(t,y(t,q,p),z(t,q,p),\psi),\quad R(t,q,\lambda):=r(t,y(t,q),\lambda),$$
$$G(t,q):=g(t,y(t,q)),\quad K(t,q,p):=k(t,y(t,q),z(t,q,p)),$$

the system (5.3) can be expressed as

$$\dot{q}=V(t,q,p) \tag{5.6a}$$
$$\dot{p}=F(t,q,p,\psi)+R(t,q,\lambda) \tag{5.6b}$$
$$0=G(t,q) \tag{5.6c}$$
$$0=G_t(t,q)+G_q(t,q)V(t,q,p) \tag{5.6d}$$
$$0=K(t,q,p). \tag{5.6e}$$

Thus, the system (5.3) can be equivalently expressed in the form (5.1). For the analysis presented in this chapter, we consider systems with $p(t,y,z)=z$, but the results are also valid in the general case.

## 5.2 SPARK Methods

We introduce the SPARK methods applied to problems with mixed index 2 and 3 constraints.

**Definition 5.2.1.** *One step of an $(s, s)$-stage specialized partitioned additive Runge-Kutta (SPARK) method applied to the system (5.1) with stepsize $h$ starting at $(y_0, z_0)$ at time $t_0$ is given by the solution of the nonlinear system of equations*

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij}V_j, \quad i = 1, \ldots, s \tag{5.7a}$$

$$\widetilde{Y}_i = y_0 + h\sum_{j=1}^{s} \bar{a}_{ij}V_j, \quad i = 0, \ldots, s \tag{5.7b}$$

$$Z_i = z_0 + h\sum_{j=1}^{s} \widehat{a}_{ij}F_j + h\sum_{j=0}^{s} \widetilde{a}_{ij}R_j, \quad i = 1, \ldots, s \tag{5.7c}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_jV_j \tag{5.7d}$$

$$z_1 = z_0 + h\sum_{j=1}^{s} \widehat{b}_jF_j + h\sum_{j=0}^{s} \widetilde{b}_jR_j \tag{5.7e}$$

$$0 = g(t_0 + \widetilde{c}_ih, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{5.7f}$$

$$0 = g(t_1, y_1) \tag{5.7g}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{5.7h}$$

$$0 = \sum_{j=1}^{s} \omega_{ij}k(t_0 + c_jh, Y_j, Z_j) + \omega_{i,s+1}k(t_1, y_1, z_1), \quad i = 1, \ldots, s, \tag{5.7i}$$

*where we have used the definitions $t_1 := t_0 + h$, $V_j := v(t_0 + c_jh, Y_j, Z_j)$, $F_j := f(t_0 + c_jh, Y_j, Z_j, \Psi_j)$, and $R_j := r(t_0 + \widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)$. The coefficients $c_i$ and $\widetilde{c}_i$ are determined by*

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad \widetilde{c}_i = \sum_{j=1}^{s} \bar{a}_{ij}. \tag{5.8}$$

*The coefficients $\omega_{ij}$ are from the matrix $\widetilde{\Omega}_0$ defined by*

$$\widetilde{\Omega}_0 := \begin{bmatrix} 0_s^T & 1 \\ b^T & 0 \\ b^T C & 0 \\ \vdots & \vdots \\ b^T C^{s-2} & 0 \end{bmatrix} \in \mathbb{R}^{s\times(s+1)}, \qquad C^k := \operatorname{diag}(c_1^k, \ldots, c_s^k). \tag{5.9}$$

We also define

$$\alpha := \begin{bmatrix} A \\ b^T \end{bmatrix}, \quad \widehat{\alpha} := \begin{bmatrix} \widehat{A} \\ \widehat{b}^T \end{bmatrix}, \quad \widetilde{\alpha} := \begin{bmatrix} \widetilde{A} \\ \widetilde{b}^T \end{bmatrix}.$$

It is assumed that the RK coefficients satisfy

$$\bar{a}_{0j} = 0, \quad j = 1, \ldots, s, \tag{5.10a}$$

$$\bar{a}_{sj} = b_j, \quad j = 1, \ldots, s, \tag{5.10b}$$

$$\sum_{j=1}^{s} \bar{a}_{ij} c_j = \sum_{j=1}^{s} \bar{a}_{ij} \sum_{k=1}^{s} \widehat{a}_{jk} = \sum_{j=1}^{s} \bar{a}_{ij} \sum_{k=0}^{s} \widetilde{a}_{jk} = \frac{\widetilde{c}_i^2}{2}, \quad i = 0, \ldots, s, \tag{5.10c}$$

$$\bar{A}\widetilde{A} = \begin{bmatrix} 0 & \ldots & 0 \\ & \bar{A}^*\widetilde{A} & \end{bmatrix}, \quad \begin{bmatrix} \bar{A}^*\widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \text{ is invertible,} \tag{5.10d}$$

where $\bar{A}^* \in \mathbb{R}^{s\times s}$ equals $\bar{A}$ with the first row removed. These assumptions are made in [16] and Chapter 3 for the existence and uniqueness of a numerical solution for the system (5.1) with only index 3 constraints. The coefficients will also be assumed to satisfy

$$\sum_{i=1}^{s} b_i = \sum_{i=1}^{s} \widehat{b}_i = \sum_{i=0}^{s} \widetilde{b}_i = 1 \tag{5.11a}$$

$$\bar{A}^* \in \mathbb{R}^{s\times s} \text{ and } \widetilde{\alpha} \in \mathbb{R}^{(s+1)\times(s+1)} \text{ are invertible,} \tag{5.11b}$$

$$M := \begin{bmatrix} b^T \\ b^T - b^T A \\ b^T - 2b^T CA \\ \vdots \\ b^T - (s-1)b^T C^{s-2} A \end{bmatrix} \in \mathbb{R}^{s \times s} \text{ is invertible,} \tag{5.11c}$$

$$c_i = \sum_{j=1}^{s} \widehat{a}_{ij} = \sum_{j=0}^{s} \widetilde{a}_{ij}, \quad i = 1, \ldots, s. \tag{5.11d}$$

To help with calculations, we introduce the internal stages

$$Z_i^f = z_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} f(t_0 + c_j h, Y_j, Z_j, \Psi_j), \quad i = 1, \ldots, s \tag{5.12a}$$

$$Z_i^r = h \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s. \tag{5.12b}$$

Notice that $Z_i = Z_i^f + Z_i^r$. We also introduce the differential equations as part of (5.1)

$$\dot{z}^f = f(t, y, z), \quad z^f(t_0) = z_0 \tag{5.13a}$$

$$\dot{z}^r = r(t, y, \lambda), \quad z^r(t_0) = 0. \tag{5.13b}$$

Again, notice that $z(t) = z^f(t) + z^r(t)$.

### 5.2.1 Gauss-Lobatto SPARK Methods

An example of a class of SPARK methods is the $(s, s)$-Gauss-Lobatto SPARK methods. These are presented in [16]. In this section, we focus on results concerning the coefficients of these methods. The coefficients have the properties $\widehat{a}_{ij} = a_{ij}$, $\widehat{b}_i = b_i$, as well as $\widetilde{c}_0 = 0$, $\widetilde{c}_s = 1$, and satisfy

$$B(2s): \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s, \tag{5.14}$$

$$C(s): \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \qquad i = 1, \ldots, s, \ k = 1, \ldots, s, \tag{5.15}$$

$$\widetilde{B}(2s): \quad \sum_{i=0}^{s} \widetilde{b}_i \widetilde{c}_i^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s, \tag{5.16}$$

$$D(s): \quad \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{b_j}{k}(1 - c_j^k), \qquad j = 1, \ldots, s, \ k = 1, \ldots, s. \tag{5.17}$$

The condition (5.14) is from the Gaussian quadrature formula with $s$ nodes, (5.15) is from the Gauss RK coefficients, and (5.16) is from the Lobatto quadrature formula with $s+1$ nodes. It is well known that the Gauss and Lobatto quadrature formulas satisfy $b_i \neq 0$, $c_i \neq 0$ for $i = 1, \ldots, s$ and $\widetilde{b}_j \neq 0$, $\widetilde{c}_i \neq 0$ for $j = 0, \ldots, s$, $i = 1, \ldots, s$. The coefficients also satisfy

$$\sum_{i=1}^{s} b_i = \sum_{i=0}^{s} \widetilde{b}_i = 1.$$

The coefficients $\bar{a}_{ij}$ and $\widetilde{a}_{ij}$ are chosen to satisfy, respectively, the conditions

$$\overline{C}(s): \quad \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad i = 0, \ldots, s, \ k = 1, \ldots, s, \tag{5.18}$$

$$\widetilde{b}_j \left(1 - \frac{\bar{a}_{ji}}{b_i}\right) = \widetilde{a}_{ij}, \qquad i = 1, \ldots, s, \ j = 0, \ldots, s. \tag{5.19}$$

In Chapter 3, we presented a proof showing that the condition (5.19) is equivalent to the conditions

$$\widetilde{a}_{i0} = \widetilde{b}_0, \quad i = 1, \ldots, s,$$

$$\widetilde{C}(s): \quad \sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} = \frac{c_i^k}{k}, \qquad i = 1, \ldots, s, \ k = 1, \ldots, s. \tag{5.20}$$

We also proved Lemma 3.2.4 showing that the Gauss-Lobatto coefficients satisfy the extended condition

$$\overline{C}(s+1): \quad \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad i = 0, \ldots, s, \ k = 1, \ldots, s+1, \tag{5.21}$$

as well as the properties (5.10), and the properties

$$\widetilde{a}_{i0} = \widetilde{b}_0, \qquad \widetilde{a}_{is} = 0, \qquad i = 1, \ldots, s. \tag{5.22}$$

For the Gauss coefficients, the invertibility of the matrix $A$ is known. By definition, $c_i = \sum_{j=1}^{s} \widehat{a}_{ij}$. We have also shown in Theorem 2.2.3 that the matrix

$$M = \begin{bmatrix} b^T \\ b^T - b^T A \\ b^T - 2b^T CA \\ \vdots \\ b^T - (s-1)b^T C^{s-2} A \end{bmatrix}$$

is invertible. We now present a proof of the remaining properties in (5.11).

**Theorem 5.2.2.** *For the Gauss-Lobatto methods, the matrix $\bar{A}^*$ from (5.10d) is invertible, with entries determined by (5.18) for $i = 1, \ldots, s$. In addition, the matrix $\widetilde{\alpha}$ is invertible, with entries determined by (5.19).*

*Proof.* Let $V, \widetilde{V} \in \mathbb{R}^{s \times s}$ be the matrices defined by

$$V := \begin{bmatrix} 1 & c_1 & \ldots & c_1^{s-1} \\ 1 & c_2 & \ldots & c_2^{s-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & c_s & \ldots & c_s^{s-1} \end{bmatrix}, \quad \widetilde{V} := \begin{bmatrix} 1 & \widetilde{c}_0 & \ldots & \widetilde{c}_0^{s-1} \\ 1 & \widetilde{c}_1 & \ldots & \widetilde{c}_1^{s-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \widetilde{c}_{s-1} & \ldots & \widetilde{c}_{s-1}^{s-1} \end{bmatrix}.$$

These are Vandermonde matrices, and will thus be invertible since $c_i \neq c_j$ and $\widetilde{c}_i \neq \widetilde{c}_j$, for $i \neq j$. To show that $\bar{A}^*$ is invertible, it suffices to show that the product $\bar{A}^* V$ is invertible. However, the $(i,k)$ entry of this product $(\bar{A}^* V)_{ik}$ is given by $\sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1}$, for $i = 1, \ldots, s$, $k = 1, \ldots, s$. By (5.18), the $(i,k)$ entry can be expressed as $\widetilde{c}_i^k / k$. Thus

$$\bar{A}^* V = \begin{bmatrix} \widetilde{c}_1 & & & O \\ & \widetilde{c}_2 & & \\ & & \ddots & \\ O & & & \widetilde{c}_s \end{bmatrix} \begin{bmatrix} 1 & \widetilde{c}_1 & \ldots & \widetilde{c}_1^{s-1} \\ 1 & \widetilde{c}_2 & \ldots & \widetilde{c}_2^{s-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \widetilde{c}_s & \ldots & \widetilde{c}_s^{s-1} \end{bmatrix} \begin{bmatrix} 1 & & & O \\ & \frac{1}{2} & & \\ & & \ddots & \\ O & & & \frac{1}{s} \end{bmatrix}.$$

The three matrices in this product are each invertible. Therefore, the invertibility of $\bar{A}^*$ follows. The invertibility of $\widetilde{\alpha}$ follows in a similar fashion. Because $\widetilde{a}_{is} = 0$ for $i = 1, \ldots, s$ by (5.22), and $\widetilde{b}_s \neq 0$ for the Lobatto coefficients, the invertibility of $\widetilde{\alpha}$ is equivalent to the invertibility of the matrix

$$\begin{bmatrix} \widetilde{a}_{10} & \widetilde{a}_{11} & \ldots & \widetilde{a}_{1,s-1} \\ \widetilde{a}_{20} & \widetilde{a}_{21} & \ldots & \widetilde{a}_{2,s-1} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{a}_{s0} & \widetilde{a}_{s1} & \ldots & \widetilde{a}_{s,s-1} \end{bmatrix}.$$

The invertibility of this matrix is seen by multiplying by $\widetilde{V}$:

$$\begin{bmatrix} \widetilde{a}_{10} & \widetilde{a}_{11} & \ldots & \widetilde{a}_{1,s-1} \\ \widetilde{a}_{20} & \widetilde{a}_{21} & \ldots & \widetilde{a}_{2,s-1} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{a}_{s0} & \widetilde{a}_{s1} & \ldots & \widetilde{a}_{s,s-1} \end{bmatrix} \begin{bmatrix} 1 & \widetilde{c}_0 & \ldots & \widetilde{c}_0^{s-1} \\ 1 & \widetilde{c}_1 & \ldots & \widetilde{c}_1^{s-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \widetilde{c}_{s-1} & \ldots & \widetilde{c}_{s-1}^{s-1} \end{bmatrix}.$$

The $(i,k)$ entry of this product is $\sum_{j=0}^{s-1} \widetilde{a}_{ij}\widetilde{c}_j^{k-1} = \sum_{j=0}^{s} \widetilde{a}_{ij}\widetilde{c}_j^{k-1}$, for $i = 1, \ldots, s$, $k = 1, \ldots, s$. But by (5.20), this is the same as $c_i^k/k$. This can be written as

$$\begin{bmatrix} c_1 & & & O \\ & c_2 & & \\ & & \ddots & \\ O & & & c_s \end{bmatrix} \begin{bmatrix} 1 & c_1 & \ldots & c_1^{s-1} \\ 1 & c_2 & \ldots & c_2^{s-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & c_s & \ldots & c_s^{s-1} \end{bmatrix} \begin{bmatrix} 1 & & & O \\ & \frac{1}{2} & & \\ & & \ddots & \\ O & & & \frac{1}{s} \end{bmatrix}.$$

Each of these matrices is invertible, and thus the invertibility of $\widetilde{\alpha}$ follows. □

The Gauss-Lobatto SPARK methods applied to mixed index 2 and 3 DAEs are symmetric methods. Because of this, their local order must be even. This will play a role later in the derivation of their order.

**Theorem 5.2.3.** *Assume the initial conditions $(y_0, z_0)$ at time $t_0$ are consistent,*

*i.e.,*

$$\begin{aligned} g(t_0, y_0) &= 0 \\ g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) &= 0 \\ k(t_0, y_0, z_0) &= 0. \end{aligned}$$

*Then the Gauss-Lobatto SPARK methods applied to mixed index 2 and 3 DAEs are symmetric.*

*Proof.* First, we rewrite (5.7g,h,i). Because the initial conditions are assumed consistent, these can be written as

$$0 = g(t_1, y_1) + g(t_0, y_0) \tag{5.23a}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) + g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) \tag{5.23b}$$

$$\begin{aligned} 0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j, Z_j) + \omega_{i,s+1} k(t_1, y_1, z_1) \\ + \omega_{i,s+1} k(t_0, y_0, z_0), \quad i = 1, \ldots, s. \end{aligned} \tag{5.23c}$$

Using the method (5.7a-f) and (5.23) with Gauss-Lobatto coefficients, we apply the method with stepsize $-h$ starting at time $t_1$. Exchanging $y_0 \longleftrightarrow y_1$, $z_0 \longleftrightarrow z_1$, we obtain the system

$$Y_i = y_1 - h \sum_{j=1}^{s} a_{ij} v(t_1 - c_j h, Y_j, Z_j), \quad i = 1, \ldots, s \tag{5.24a}$$

$$\widetilde{Y}_i = y_1 - h \sum_{j=1}^{s} \bar{a}_{ij} v(t_1 - c_j h, Y_j, Z_j), \quad i = 0, \ldots, s \tag{5.24b}$$

$$\begin{aligned} Z_i = z_1 - h \sum_{j=1}^{s} a_{ij} f(t_1 - c_j h, Y_j, Z_j, \Psi_j) \\ - h \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_1 - \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \end{aligned} \tag{5.24c}$$

$$y_0 = y_1 - h \sum_{j=1}^{s} b_j v(t_1 - c_j h, Y_j, Z_j) \tag{5.24d}$$

$$z_0 = z_1 - h\sum_{j=1}^{s} b_j f(t_1 - c_j h, Y_j, Z_j, \Psi_j) - h\sum_{j=0}^{s} \widetilde{b}_j r(t_1 - \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) \tag{5.24e}$$

$$0 = g(t_1 - \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{5.24f}$$

$$0 = g(t_0, y_0) + g(t_1, y_1) \tag{5.24g}$$

$$0 = g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) + g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{5.24h}$$

$$\begin{aligned} 0 = \sum_{j=1}^{s} \omega_{ij} k(t_1 - c_j h, Y_j, Z_j) + \omega_{i,s+1} k(t_0, y_0, z_0) \\ + \omega_{i,s+1} k(t_1, y_1, z_1), \quad i = 1, \ldots, s. \end{aligned} \tag{5.24i}$$

Using the definition $t_1 = t_0 + h$, (5.24d,e) become

$$\begin{aligned} y_1 &= y_0 + h\sum_{j=1}^{s} b_j v(t_0 + (1 - c_j)h, Y_j, Z_j) \\ z_1 &= z_0 + h\sum_{j=1}^{s} b_j f(t_0 + (1 - c_j)h, Y_j, Z_j, \Psi_j) \\ &\quad + h\sum_{j=0}^{s} \widetilde{b}_j r(t_0 + (1 - \widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j). \end{aligned}$$

Substituting these back into (5.24) and applying the consistency of the initial conditions $(y_0, z_0)$ at time $t_0$ gives

$$Y_i = y_0 + h\sum_{j=1}^{s} (b_j - a_{ij}) v(t_0 + (1 - c_j)h, Y_j, Z_j), \quad i = 1, \ldots, s \tag{5.25a}$$

$$\widetilde{Y}_i = y_0 + h\sum_{j=1}^{s} (b_i - \bar{a}_{ij}) v(t_0 + (1 - c_j)h, Y_j, Z_j), \quad i = 0, \ldots, s \tag{5.25b}$$

$$\begin{aligned} Z_i = z_0 + h\sum_{j=1}^{s} (b_j - a_{ij}) f(t_0 + (1 - c_j)h, Y_j, Z_j, \Psi_j) \\ + h\sum_{j=0}^{s} (\widetilde{b}_j - \widetilde{a}_{ij}) r(t_0 + (1 - \widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \end{aligned} \tag{5.25c}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j v(t_0 + (1 - c_j)h, Y_j, Z_j) \tag{5.25d}$$

$$z_1 = z_0 + h\sum_{j=1}^{s} b_j f(t_0 + (1-c_j)h, Y_j, Z_j, \Psi_j) + h\sum_{j=0}^{s} \widetilde{b}_j r(t_0 + (1-\widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j) \tag{5.25e}$$

$$0 = g(t_0 + (1-\widetilde{c}_i)h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{5.25f}$$

$$0 = g(t_1, y_1) \tag{5.25g}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{5.25h}$$

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + (1-c_j)h, Y_j, Z_j) + \omega_{i,s+1} k(t_1, y_1, z_1), \quad i = 1, \ldots, s. \tag{5.25i}$$

Lastly, we would like to write (5.25) in the format of (5.7). To do so, we must reindex each $Y_i$, $\widetilde{Y}_i$, $Z_i$, $\Lambda_i$, and $\Psi_i$ so as to preserve the usual ordering of the $c_i$, $\widetilde{c}_i$ coefficients. This results in the system

$$Y_i^* = y_0 + h\sum_{j=1}^{s} a_{ij}^* v(t_0 + c_j^* h, Y_j^*, Z_j^*), \quad i = 1, \ldots, s \tag{5.26a}$$

$$\widetilde{Y}_i^* = y_0 + h\sum_{j=1}^{s} \bar{a}_{ij}^* v(t_0 + c_j^* h, Y_j^*, Z_j^*), \quad i = 0, \ldots, s \tag{5.26b}$$

$$Z_i^* = z_0 + h\sum_{j=1}^{s} a_{ij}^* f(t_0 + c_j^* h, Y_j^*, Z_j^*, \Psi_j^*) + h\sum_{j=0}^{s} \widetilde{a}_{ij}^* r(t_0 + \widetilde{c}_j^* h, \widetilde{Y}_j^*, \Lambda_j^*), \quad i = 1, \ldots, s \tag{5.26c}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j^* v(t_0 + c_j^* h, Y_j^*, Z_j^*) \tag{5.26d}$$

$$z_1 = z_0 + h\sum_{j=1}^{s} b_j^* f(t_0 + c_j^* h, Y_j^*, Z_j^*, \Psi_j^*) + h\sum_{j=0}^{s} \widetilde{b}_j^* r(t_0 + \widetilde{c}_j^* h, \widetilde{Y}_j^*, \Lambda_j^*) \tag{5.26e}$$

$$0 = g(t_0 + \widetilde{c}_i^* h, \widetilde{Y}_i^*), \quad i = 0, \ldots, s \tag{5.26f}$$

$$0 = g(t_1, y_1) \tag{5.26g}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{5.26h}$$

$$0 = \sum_{j=1}^{s} \omega_{ij}^* k(t_0 + c_j^* h, Y_j^*, Z_j^*) + \omega_{i,s+1} k(t_1, y_1, z_1), \quad i = 1, \ldots, s, \tag{5.26i}$$

where the stages are defined by

$$Y_i^* := Y_{s+1-i}, \quad \widetilde{Y}_i^* := \widetilde{Y}_{s-i}, \quad Z_i^* := Z_{s+1-i}, \quad \Lambda_i^* := \Lambda_{s-i}, \quad \Psi_i^* := \Psi_{s+1-i}$$

and where the RK coefficients are defined by

$$c_i^* = 1 - c_{s+1-i}, \quad \widetilde{c}_i^* = 1 - \widetilde{c}_{s-i} \tag{5.27a}$$

$$b_i^* = b_{s+1-i}, \quad \widetilde{b}_i^* = \widetilde{b}_{s-i} \tag{5.27b}$$

$$\begin{aligned} a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j}, \quad \widetilde{a}_{ij}^* = \widetilde{b}_{s-j} - \widetilde{a}_{s+1-i,s-j}, \\ \bar{a}_{ij}^* = b_{s+1-j} - \bar{a}_{s-i,s+1-j} \end{aligned} \tag{5.27c}$$

$$\omega_{ij}^* = \omega_{i,s+1-j}. \tag{5.27d}$$

The method given by (5.27) is the *adjoint* of (5.7) with Gauss-Lobatto coefficients. To show the symmetry of the method, we must show that the method given by (5.26) is the same as the method (5.7) with Gauss-Lobatto coefficients, i.e. we must show

$$\begin{aligned} c_i^* = c_i, \quad \widetilde{c}_i^* = \widetilde{c}_i, \quad b_i^* = b_i, \quad \widetilde{b}_i^* = \widetilde{b}_i, \\ a_{ij}^* = a_{ij}, \quad \widetilde{a}_{ij}^* = \widetilde{a}_{ij}, \quad \bar{a}_{ij}^* = \bar{a}_{ij}, \end{aligned}$$

and that (5.26i) implies

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j^*, Z_j^*) + \omega_{i,s+1} k(t_1, y_1, z_1), \quad i = 1, \ldots, s. \tag{5.28}$$

The first line of equalities comes immediately from the symmetry of the Gauss and the Lobatto coefficients, as does the equality $a_{ij}^* = a_{ij}$. The proof of Theorem 3.2.6 shows that $\bar{a}_{ij}^* = \bar{a}_{ij}$ and $\widetilde{a}_{ij}^* = \widetilde{a}_{ij}$. To show that (5.26i) implies (5.28), it suffices to show that

$$0 = \sum_{j=1}^{s} \omega_{ij}^* k(t_0 + c_j^* h, Y_j^*, Z_j^*) \quad \Rightarrow \quad 0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j^*, Z_j^*),$$

for $i = 2, \ldots, s$. From the definition of $\widetilde{\Omega}_0$ (5.9), the condition (5.26i) for $i = 2, \ldots, s$

is the same as

$$\begin{aligned} 0 &= \sum_{j=1}^{s} \omega_{ij}^{*} k(t_0 + c_j^{*} h, Y_j^{*}, Z_j^{*}) \\ &= \sum_{j=1}^{s} b_{s+1-j}\ c_{s+1-j}^{i-1}\ k(t_0 + c_j h, Y_j^{*}, Z_j^{*}) \\ &= \sum_{j=1}^{s} b_j (1 - c_j)^{i-1} k(t_0 + c_j h, Y_j^{*}, Z_j^{*}). \end{aligned}$$

We will show by induction that $0 = \sum_{j=1}^{s} b_j (1 - c_j)^{i-1} k(t_0 + c_j h, Y_j^{*}, Z_j^{*})$ implies $0 = \sum_{j=1}^{s} b_j c_j^{i-1} k(t_0 + c_j h, Y_j^{*}, Z_j^{*})$ for $i = 1, \ldots, s$. For $i = 1$, the result is trivial. For any $i = 2, \ldots, s$, we have

$$\begin{aligned} 0 &= \sum_{j=1}^{s} b_j (1 - c_j)^{i-1} k(t_0 + c_j h, Y_j^{*}, Z_j^{*}) \\ &= \sum_{j=1}^{s} \sum_{l=0}^{i-1} \binom{i-1}{l} b_j (-1)^l c_j^l k(t_0 + c_j h, Y_j^{*}, Z_j^{*}) \\ &= \sum_{l=0}^{i-1} \binom{i-1}{l} (-1)^l \sum_{j=1}^{s} b_j c_j^l k(t_0 + c_j h, Y_j^{*}, Z_j^{*}). \end{aligned}$$

Hence,

$$0 = \sum_{j=1}^{s} b_j c_j^{i-1} k(t_0 + c_j h, Y_j^{*}, Z_j^{*}),$$

where we have applied the induction hypothesis. This gives the desired result, as for $i = 1, \ldots, s$, $b_j c_j^{i-1} = \omega_{ij}$.

Therefore, the Gauss-Lobatto SPARK methods applied to mixed index 2 and 3 problems are symmetric. □

### 5.3 Existence, Uniqueness, and Influence of Perturbations

In this section, we give results concerning the existence and uniqueness of solutions to the system (5.7). We also discuss some results regarding the influence of perturbations. First, however, we present a lemma useful for this section.

**Lemma 5.3.1.** *Assume the invertibility of the matrices (5.2), and the conditions*

*(5.10) and (5.11). Then the block matrix*

$$\begin{bmatrix} \begin{bmatrix} \bar{A}^* \widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y v_z r_\lambda(t,y,z,\lambda) & \begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} \otimes g_y v_z f_\psi(t,y,z,\psi) \\ \widetilde{\Omega}_0 \widetilde{\alpha} \otimes k_z r_\lambda(t,y,z,\lambda) & \widetilde{\Omega}_0 \alpha \otimes k_z f_\psi(t,y,z,\psi) \end{bmatrix} \tag{5.29}$$

*is invertible.*

*Proof.* This proof will be broken into two parts. First, we construct an invertible matrix $R$. Secondly, we compute the product of the matrices $R$ and of (5.29), and show that this product is invertible. This will prove the invertibility of the matrix of (5.29).

*First Step.* Let $D \in \mathbb{R}^{s\times s}$ be the matrix

$$D := M^{-1} \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & -(s-1) \end{bmatrix}.$$

The matrix $D$ is invertible, as it is the product of an invertible matrix and a triangular matrix with no zeros on the diagonal. Denote by $\Omega$ the matrix $\Omega := D\widetilde{\Omega}_0 \in \mathbb{R}^{s\times(s+1)}$. We have shown in the proof of Lemma 2.2.2 that $\Omega\alpha = I_s$. Let $\gamma = [\widetilde{\gamma}^T \ \gamma_{s+1}]^T \in \mathbb{R}^{s+1}$, with $\widetilde{\gamma} \in \mathbb{R}^s$ and $\gamma_{s+1} \neq 0$, be a fixed vector defined by the orthogonality condition

$$\gamma^T \alpha = 0. \tag{5.30}$$

Because $\alpha \in \mathbb{R}^{(s+1)\times s}$, this is an underdetermined system of linear equations for $\gamma$. In fact, an infinite number of values exist for $\gamma$, because $\alpha = [A^T \ b]^T$, and $A$ is invertible. Now, let $Q \in \mathbb{R}^{(s+1)\times(s+1)}$ be given by

$$Q := \begin{bmatrix} A\Omega \\ \gamma^T \end{bmatrix}.$$

The matrix $Q$ is invertible. To see this, we multiply Q by an invertible matrix:

$$Q \begin{bmatrix} A & 0_s \\ b^T & 1 \end{bmatrix} = \begin{bmatrix} A\Omega\alpha & A\Omega e_{s+1} \\ \gamma^T\alpha & \gamma_{s+1} \end{bmatrix} = \begin{bmatrix} A & A\Omega e_{s+1} \\ 0_s^T & \gamma_{s+1} \end{bmatrix}.$$

Here, $e_i$ is the $i$-th column of $I_{s+1}$. Because $A$ is invertible and $\gamma_{s+1} \neq 0$, we see that $Q$ must therefore be invertible. Lastly, define $\check{Q} \in \mathbb{R}^{(s+1)\times(s+1)}$ by

$$\check{Q} := Q \begin{bmatrix} \bar{A^*}^{-1} & 0_s \\ 0_s^T & 1 \end{bmatrix},$$

which is invertible, as $\check{Q}$ is the product of two invertible matrices.

Combining all of these together, we define the matrix $R$ by

$$R := \begin{bmatrix} \check{Q} \otimes I_{n_g} & O \\ O & AD \otimes I_{n_k} \end{bmatrix} \in \mathbb{R}^{((s+1)n_g+sn_k)\times((s+1)n_g+sn_k)}. \tag{5.31}$$

Because this is block diagonal with each block invertible, the matrix $R$ itself must be invertible.

*Second Step.* We now compute the product of $R$ and (5.29). This can be expressed as

$$\begin{bmatrix} \check{Q} \begin{bmatrix} \bar{A^*}\widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y v_z r_\lambda & \check{Q} \begin{bmatrix} \bar{A^*}A \\ b^T \end{bmatrix} \otimes g_y v_z f_\psi \\ A\Omega\widetilde{\alpha} \otimes k_z r_\lambda & A\Omega\alpha \otimes k_z f_\psi \end{bmatrix}. \tag{5.32}$$

Each function above is evaluated at $(t, y, z, \lambda, \psi)$. The function arguments will be suppressed for the remainder of this proof. However, the blocks of this matrix can be expressed as

$$\check{Q} \begin{bmatrix} \bar{A^*}\widetilde{A} \\ \widetilde{b}^T \end{bmatrix} = Q \begin{bmatrix} \bar{A^*}^{-1} & 0_s \\ 0_s^T & 1 \end{bmatrix} \begin{bmatrix} \bar{A^*}\widetilde{A} \\ \widetilde{b}^T \end{bmatrix}$$

$$
\begin{aligned}
&= Q \begin{bmatrix} \bar{A}^{*-1} & 0_s \\ 0_s^T & 1 \end{bmatrix} \begin{bmatrix} \bar{A}^* & 0_s \\ 0_s^T & 1 \end{bmatrix} \widetilde{\alpha} \\
&= Q\widetilde{\alpha} \\
&= \begin{bmatrix} A\Omega \\ \gamma^T \end{bmatrix} \widetilde{\alpha},
\end{aligned}
$$

$$
\begin{aligned}
\check{Q} \begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} &= Q \begin{bmatrix} \bar{A}^{*-1} & 0_s \\ 0_s^T & 1 \end{bmatrix} \begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} \\
&= Q \begin{bmatrix} \bar{A}^{*-1} & 0_s \\ 0_s^T & 1 \end{bmatrix} \begin{bmatrix} \bar{A}^* & 0_s \\ 0_s^T & 1 \end{bmatrix} \alpha \\
&= Q\alpha \\
&= \begin{bmatrix} A \\ 0_s^T \end{bmatrix},
\end{aligned}
$$

$$A\Omega\alpha = A.$$

Applying these to (5.32) results in the matrix

$$
\begin{bmatrix} \begin{bmatrix} A\Omega \\ \gamma^T \end{bmatrix} \widetilde{\alpha} \otimes g_y v_z r_\lambda & \begin{bmatrix} A \\ 0_s^T \end{bmatrix} \otimes g_y v_z f_\psi \\ A\Omega\widetilde{\alpha} \otimes k_z r_\lambda & A \otimes k_z f_\psi \end{bmatrix} =
$$

$$
\begin{bmatrix} I_{s+1} \otimes g_y v_z r_\lambda & \begin{bmatrix} I_s \\ 0_s^T \end{bmatrix} \otimes g_y v_z f_\psi \\ [I_s \ 0_s] \otimes k_z r_\lambda & I_s \otimes k_z f_\psi \end{bmatrix} \begin{bmatrix} Q\widetilde{\alpha} \otimes I_{n_g} & O \\ O & A \otimes I_{n_k} \end{bmatrix}. \tag{5.33}
$$

In the final product above, the second matrix is invertible, since $Q$, $\widetilde{\alpha}$, and $A$ are all invertible. Because $g_y v_z r_\lambda$ is assumed invertible by (5.2a), the invertibility of the

first matrix in the product of (5.33) is equivalent to the invertibility of the matrix

$$\begin{bmatrix} I_s \otimes g_y v_z r_\lambda & I_s \otimes g_y v_z f_\psi \\ I_s \otimes k_z r_\lambda & I_s \otimes k_z f_\psi \end{bmatrix}.$$

Because the matrix (5.2b) is invertible, the above matrix is also invertible. This can be seen by applying Gaussian elimination by block.

This completes the proof. □

### 5.3.1 Existence and Uniqueness

We give here a proof regarding the existence and uniqueness of a solution to (5.7). This proof is a combination of the existence proofs presented in [15] and [16]. For this section, we consider $y_0$, $y_1$, $z_0$, $\psi_0$ as functions of $h$.

**Theorem 5.3.2.** *Suppose that $y_0 = y_0(h)$, $z_0 = z_0(h)$, $\lambda_0 = \lambda_0(h)$, $\psi_0 = \psi_0(h)$ satisfy*

$$o(h^2) = g(t_0, y_0) \tag{5.34a}$$

$$o(h) = g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) \tag{5.34b}$$

$$\begin{aligned} o(1) = {} & g_{tt}(t_0, y_0) + 2g_{ty}(t_0, y_0)v(t_0, y_0, z_0) + g_y(t_0, y_0)v_t(t_0, y_0, z_0) \\ & + g_{yy}(t_0, y_0)\left(v(t_0, y_0, z_0), v(t_0, y_0, z_0)\right) \\ & + g_y(t_0, y_0)v_y(t_0, y_0, z_0)v(t_0, y_0, z_0) \\ & + g_y(t_0, y_0)v_z(t_0, y_0, z_0)\left[f(t_0, y_0, z_0, \psi_0) + r(t_0, y_0, \lambda_0)\right] \end{aligned} \tag{5.34c}$$

$$o(h) = k(t_0, y_0, z_0) \tag{5.34d}$$

$$\begin{aligned} o(1) = {} & k_t(t_0, y_0, z_0) + k_y(t_0, y_0, z_0)v(t_0, y_0, z_0) \\ & + k_z(t_0, y_0, z_0)\left[f(t_0, y_0, z_0, \psi_0) + r(t_0, y_0, \lambda_0)\right], \end{aligned} \tag{5.34e}$$

*where the matrices given in (5.2) are invertible. Then for $|h| \leq h_0$, there exists a locally unique solution to (5.7) that satisfies*

$$Y_i - y_0 = \mathcal{O}(h), \quad i = 1, \ldots, s$$

$$
\begin{aligned}
Z_i - z_0 &= \mathcal{O}(h), \quad i = 1, \ldots, s \\
\widetilde{Y}_i - y_0 &= \mathcal{O}(h), \quad i = 0, \ldots, s \\
\Lambda_i - \lambda_0 &= \mathcal{O}(h), \quad i = 0, \ldots, s \\
\Psi_i - \psi_0 &= \mathcal{O}(h), \quad i = 1, \ldots, s \\
y_1 - y_0 &= \mathcal{O}(h), \\
z_1 - z_0 &= \mathcal{O}(h).
\end{aligned}
$$

*Proof.* We begin by reformulating the system (5.7) as

$$0 = Y_i - y_0 - h \sum_{j=1}^{s} a_{ij} V_j, \quad i = 1, \ldots, s \tag{5.35a}$$

$$0 = \widetilde{Y}_i - y_0 - h \sum_{j=1}^{s} \bar{a}_{ij} V_j, \quad i = 0, \ldots, s \tag{5.35b}$$

$$0 = Z_i - z_0 - h \sum_{j=1}^{s} \widehat{a}_{ij} F_j - h \sum_{j=0}^{s} \widetilde{a}_{ij} R_j, \quad i = 1, \ldots, s \tag{5.35c}$$

$$0 = y_1 - y_0 - h \sum_{j=1}^{s} b_j V_j \tag{5.35d}$$

$$0 = z_1 - z_0 - h \sum_{j=1}^{s} \widehat{b}_j F_j - h \sum_{j=0}^{s} \widetilde{b}_j R_j \tag{5.35e}$$

$$0 = \frac{1}{h^2} g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{5.35f}$$

$$0 = \frac{1}{h} g(t_1, y_1) \tag{5.35g}$$

$$0 = \frac{1}{h} \left( g_t(t_1, y_1) + g_y(t_1, y_1) v(t_1, y_1, z_1) \right) \tag{5.35h}$$

$$0 = \frac{1}{h} \left( \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j, Z_j) + \omega_{i,s+1} k(t_1, y_1, z_1) \right), \quad i = 1, \ldots, s. \tag{5.35i}$$

The proof of this theorem can be done by application of the implicit function theorem. We first examine the constraints $0 = g(t, y)$, and then examine the constraints $0 = k(t, y, z)$.

We have $\widetilde{Y}_0 = y_0$, hence $g(t_0, \widetilde{Y}_0) = 0$ is automatically satisfied by assumption. We have $\widetilde{Y}_s = y_1$ since $\widetilde{c}_s = 1$, hence the equation (5.35g) can be removed since

it is equivalent to (5.35f) for $i = s$. To keep the following calculations clean, we introduce the notation

$$\begin{aligned}
&Y_i(\tau) := y_0 + \tau(Y_i - y_0), \quad && Z_i(\tau) := z_0 + \tau(Z_i - z_0),\\
&\widetilde{Y}_i(\tau) := y_0 + \tau(\widetilde{Y}_i - y_0), && \\
&y_1(\tau) := y_0 + \tau(y_1 - y_0), \quad && z_1(\tau) := z_0 + \tau(z_1 - z_0),\\
&T_i(\tau) := t_0 + \tau c_i h, && \widetilde{T}_i(\tau) := t_0 + \tau \widetilde{c}_i h,\\
&t_1(\tau) := t_0 + \tau h. &&
\end{aligned}$$

We now expand $g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)$ for $i = 1, \ldots, s$ and $v(t_0 + c_i h, Y_i, Z_i)$ for $i = 1, \ldots, s$ into a Taylor series around $(t_0, y_0, z_0)$, resulting in

$$\begin{aligned}
g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) = g(t_0, y_0) &+ g_t(t_0, y_0)\widetilde{c}_i h + g_y(t_0, y_0)(\widetilde{Y}_i - y_0)\\
&+ \int_0^1 (1-\tau) g_{tt}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau)) d\tau \cdot \widetilde{c}_i^2 h^2\\
&+ 2\int_0^1 (1-\tau) g_{ty}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau)) d\tau \cdot \widetilde{c}_i h (\widetilde{Y}_i - y_0)\\
&+ \int_0^1 (1-\tau) g_{yy}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau)) d\tau (\widetilde{Y}_i - y_0, \widetilde{Y}_i - y_0),
\end{aligned}$$

$$\begin{aligned}
V_i = v(t_0 + c_i h, Y_i, Z_i) &= v(t_0, y_0, z_0) + \int_0^1 v_t(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot c_i h\\
&+ \int_0^1 v_y(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot (Y_i - y_0)\\
&+ \int_0^1 v_z(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot (Z_i - z_0)\\
&= v(t_0, y_0, z_0) + c_i h \int_0^1 v_t(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau\\
&+ h \int_0^1 v_y(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot \sum_{j=1}^{s} a_{ij} V_j\\
&+ h \int_0^1 v_z(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot \Big( \sum_{j=1}^{s} \widehat{a}_{ij} F_j + \sum_{j=0}^{s} \widetilde{a}_{ij} R_j \Big).
\end{aligned}$$

Dividing $g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)$ by $h^2$, replacing the terms $Y_i - y_0$, $\widetilde{Y}_i - y_0$, and $Z_i - z_0$ by using (5.35a,b,c), and utilizing the relation above for $V_i = v(t_0 + c_i h, Y_i, Z_i)$, we

obtain

$$\begin{aligned}
\frac{1}{h^2}g(t_0+\widetilde{c}_ih,\widetilde{Y}_i) &= \frac{1}{h^2}g(t_0,y_0)+\frac{1}{h}g_t(t_0,y_0)\widetilde{c}_i+\frac{1}{h}\sum_{j=1}^{s}\bar{a}_{ij}g_y(t_0,y_0)V_j\\
&\quad+\widetilde{c}_i^2\int_0^1(1-\tau)g_{tt}(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau))d\tau\\
&\quad+2\widetilde{c}_i\sum_{j=1}^{s}\bar{a}_{ij}\int_0^1(1-\tau)g_{ty}(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau))d\tau\cdot V_j\\
&\quad+\sum_{j=1}^{s}\sum_{k=1}^{s}\bar{a}_{ij}\bar{a}_{ik}\int_0^1(1-\tau)g_{yy}(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau))d\tau(V_j,V_k)\\
&= \frac{1}{h^2}g(t_0,y_0)+\frac{1}{h}g_t(t_0,y_0)\widetilde{c}_i+\frac{1}{h}\sum_{j=1}^{s}\bar{a}_{ij}g_y(t_0,y_0)v(t_0,y_0,z_0)\\
&\quad+\widetilde{c}_i^2\int_0^1(1-\tau)g_{tt}(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau))d\tau\\
&\quad+2\widetilde{c}_i\sum_{j=1}^{s}\bar{a}_{ij}\int_0^1(1-\tau)g_{ty}(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau))d\tau\cdot V_j\\
&\quad+\sum_{j=1}^{s}\bar{a}_{ij}c_jg_y(t_0,y_0)\int_0^1v_t(T_j(\tau),Y_j(\tau),Z_j(\tau))d\tau\\
&\quad+\sum_{j=1}^{s}\sum_{k=1}^{s}\bar{a}_{ij}a_{jk}g_y(t_0,y_0)\int_0^1v_y(T_j(\tau),Y_j(\tau),Z_j(\tau))d\tau\cdot V_k\\
&\quad+\sum_{j=1}^{s}\sum_{k=1}^{s}\bar{a}_{ij}\widehat{a}_{jk}g_y(t_0,y_0)\int_0^1v_z(T_j(\tau),Y_j(\tau),Z_j(\tau))d\tau\cdot F_k\\
&\quad+\sum_{j=1}^{s}\sum_{k=0}^{s}\bar{a}_{ij}\widetilde{a}_{jk}g_y(t_0,y_0)\int_0^1v_z(T_j(\tau),Y_j(\tau),Z_j(\tau))d\tau\cdot R_k\\
&\quad+\sum_{j=1}^{s}\sum_{k=1}^{s}\bar{a}_{ij}\bar{a}_{ik}\int_0^1(1-\tau)g_{yy}(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau))d\tau(V_j,V_k).
\end{aligned}\tag{5.36}$$

By (5.10c), for the values $Y_i := y_0$, $\widetilde{Y}_i := y_0$, $Z_i := z_0$, $\Lambda_i := \lambda_0$, and $\Psi_i := \psi_0$ we obtain, as $h \to 0$,

$$\begin{aligned}
&\frac{1}{h^2}g(t_0+\widetilde{c}_ih,\widetilde{Y}_i) = \frac{1}{h^2}g(t_0,y_0)+\frac{\widetilde{c}_i}{h}\big(g_t(t_0,y_0)+g_y(t_0,y_0)v(t_0,y_0,z_0)\big)\\
&+\frac{\widetilde{c}_i^2}{2}\big(g_{tt}(t_0,y_0)+2g_{ty}(t_0,y_0)v(t_0,y_0,z_0)+g_y(t_0,y_0)v_t(t_0,y_0,z_0)\\
&\quad+g_y(t_0,y_0)v_y(t_0,y_0,z_0)v(t_0,y_0,z_0)+g_y(t_0,y_0)v_z(t_0,y_0,z_0)f(t_0,y_0,z_0,\psi_0)
\end{aligned}$$

$$+ g_y(t_0,y_0)v_z(t_0,y_0,z_0)r(t_0,y_0,\lambda_0) + g_{yy}(t_0,y_0)(v(t_0,y_0,z_0),v(t_0,y_0,z_0))\big) + \mathcal{O}(h)$$
$$= o(1),$$

since (5.34a-c) holds. Hence the values $Y_i(0) := y_0(0)$, $\widetilde{Y}_i(0) := y_0(0)$, $Z_i(0) := z_0(0)$, $\Lambda_i(0) := \lambda_0(0)$, and $\Psi_i(0) := \psi_0(0)$ satisfy (5.35a,b,c) and the constraints

$$0 = \frac{1}{h^2}g(t_0+\widetilde{c}_i h, \widetilde{Y}_i) = \frac{1}{h^2}g(t_0,y_0) + \frac{1}{h}g_t(t_0,y_0)\widetilde{c}_i + \dots \ . \tag{5.37}$$

The additional terms are those from (5.36). Similarly, we have

$$\begin{aligned}
g_y(t_1,y_1) &= g_y(t_0,y_0) + h\int_0^1 g_{ty}(t_1(\tau),y_1(\tau))d\tau \\
&\qquad + \int_0^1 g_{yy}(t_1(\tau),y_1(\tau))d\tau(y_1-y_0,\cdot) \\
&= g_y(t_0,y_0) + h\int_0^1 g_{ty}(t_1(\tau),y_1(\tau))d\tau \\
&\qquad + h\sum_{j=1}^{s} b_j \int_0^1 g_{yy}(t_1(\tau),y_1(\tau))d\tau(V_j,\cdot), \\
v(t_1,y_1,z_1) &= v(t_0,y_0,z_0) + h\int_0^1 v_t(t_1(\tau),y_1(\tau),z_1(\tau))d\tau \\
&\qquad + \int_0^1 v_y(t_1(\tau),y_1(\tau),z_1(\tau))d\tau\cdot(y_1-y_0) \\
&\qquad + \int_0^1 v_z(t_1,y_1(\tau),z_1(\tau))d\tau\cdot(z_1-z_0) \\
&= v(t_0,y_0,z_0) + h\int_0^1 v_t(t_1(\tau),y_1(\tau),z_1(\tau))d\tau \\
&\qquad + h\int_0^1 v_y(t_1(\tau),y_1(\tau),z_1(\tau))d\tau\cdot\sum_{j=1}^{s} b_j V_j \\
&\qquad + h\int_0^1 v_z(t_1(\tau),y_1(\tau),z_1(\tau))d\tau\cdot\left(\sum_{j=1}^{s}\widehat{b}_j F_j + \sum_{j=0}^{s}\widetilde{b}_j R_j\right), \\
g_t(t_1,y_1) &= g_t(t_0,y_0) + h\int_0^1 g_{tt}(t_1(\tau),y_1(\tau))d\tau \\
&\qquad + \int_0^1 g_{ty}(t_1(\tau),y_1(\tau))d\tau\cdot(y_1-y_0) \\
&= g_t(t_0,y_0) + h\int_0^1 g_{tt}(t_1(\tau),y_1(\tau))d\tau
\end{aligned}$$

$$+ h \int_0^1 g_{ty}(t_1(\tau), y_1(\tau))d\tau \cdot \sum_{j=1}^{s} b_j V_j.$$

Hence, dividing $g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1)$ by $h$, we obtain

$$\begin{aligned}
&\frac{1}{h}g_t(t_1, y_1) + \frac{1}{h}g_y(t_1, y_1)v(t_1, y_1, z_1) = \qquad (5.38)\\
&\quad \frac{1}{h}g_t(t_0, y_0) + \frac{1}{h}g_y(t_0, y_0)v(t_0, y_0, z_0) + \int_0^1 g_{tt}(t_1(\tau), y_1(\tau))d\tau\\
&+ \int_0^1 g_{ty}(t_1(\tau), y_1(\tau))d\tau \cdot v(t_1, y_1, z_1) + \sum_{j=1}^{s} b_j \int_0^1 g_{ty}(t_1(\tau), y_1(\tau))d\tau \cdot V_j\\
&+ g_y(t_0, y_0) \int_0^1 v_t(t_1(\tau), y_1(\tau), z_1(\tau))d\tau\\
&+ \sum_{j=1}^{s} b_j g_y(t_0, y_0) \int_0^1 v_y(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \cdot V_j\\
&+ \sum_{j=1}^{s} \widehat{b}_j g_y(t_0, y_0) \int_0^1 v_z(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \cdot F_j\\
&+ \sum_{j=0}^{s} \widetilde{b}_j g_y(t_0, y_0) \int_0^1 v_z(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \cdot R_j\\
&+ \sum_{j=1}^{s} b_j \int_0^1 g_{yy}(t_1(\tau), y_1(\tau))d\tau(V_j, v(t_1, y_1, z_1)).
\end{aligned}$$

Because of assumption (5.11a), for the values $Y_i := y_0$, $\widetilde{Y}_i := y_0$, $y_1 := y_0$, $Z_i := z_0$, $z_1 := z_0$, $\Lambda_i := \lambda_0$, and $\Psi_i := \psi_0$, we use (5.34) to obtain, for $h \to 0$,

$$\begin{aligned}
&\frac{1}{h}g_t(t_1, y_1) + \frac{1}{h}g_y(t_1, y_1)v(t_1, y_1, z_1) = \frac{1}{h}\left(g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0)\right)\\
&\quad + g_{tt}(t_0, y_0) + g_y(t_0, y_0)v_t(t_0, y_0, z_0) + 2g_{ty}(t_0, y_0)v(t_0, y_0, z_0)\\
&\quad + g_y(t_0, y_0)v_y(t_0, y_0, z_0)v(t_0, y_0, z_0) + g_y(t_0, y_0)v_z(t_0, y_0, z_0)f(t_0, y_0, z_0, \psi_0)\\
&\quad + g_y(t_0, y_0)v_z(t_0, y_0, z_0)r(t_0, y_0, \lambda_0) + g_{yy}(t_0, y_0)(v(t_0, y_0, z_0), v(t_0, y_0, z_0)) + \mathcal{O}(h)\\
&\quad = o(1).
\end{aligned}$$

Hence the values $Y_i(0) := y_0(0)$, $\widetilde{Y}_i(0) := y_0(0)$, $y_1(0) := y_0(0)$, $Z_i(0) := z_0(0)$,

$z_1(0) := z_0(0)$, $\Lambda_i(0) := \lambda_0(0)$, and $\Psi_i(0) := \psi_0(0)$ satisfy

$$\begin{aligned} 0 &= \frac{1}{h} g_t(t_1, y_1) + \frac{1}{h} g_y(t_1, y_1) v(t_1, y_1, z_1) \\ &= \frac{1}{h} g_t(t_0, y_0) + \frac{1}{h} g_y(t_0, y_0) v(t_0, y_0, z_0) + \dots, \end{aligned} \tag{5.39}$$

with the additional terms coming from (5.38).

Next, we examine the constraints $0 = k(t, y, z)$. To simplify, we will use the notation $Y_{s+1} := y_1$, $Z_{s+1} := z_1$, and $c_{s+1} := 1$. Dividing the right side of (5.35i) by $h$ gives

$$0 = \frac{1}{h} \sum_{j=1}^{s+1} \omega_{ij} k(t_0 + c_j h, Y_j, Z_j), \qquad i = 1, \dots, s. \tag{5.40}$$

Writing

$$\begin{aligned} k(t_0 + c_j h, Y_j, Z_j) = k(t_0, y_0, z_0) &+ c_j h \int_0^1 k_t(T_j(\tau), Y_j(\tau), Z_j(\tau)) d\tau \\ &+ \int_0^1 k_y(T_j(\tau), Y_j(\tau), Z_j(\tau)) d\tau \cdot (Y_j - y_0) \\ &+ \int_0^1 k_z(T_j(\tau), Y_j(\tau), Z_j(\tau)) d\tau \cdot (Z_j - z_0), \end{aligned}$$

we substitute into this (5.35a,c,d,e), and get

$$\begin{aligned} \frac{1}{h} \sum_{j=1}^{s+1} \omega_{ij} k(t_0 + c_j h, Y_j, Z_j) &= \frac{1}{h} \sum_{j=1}^{s+1} \omega_{ij} k(t_0, y_0, z_0) \\ &+ \sum_{j=1}^{s+1} \omega_{ij} c_j \int_0^1 k_t(T_j(\tau), Y_j(\tau), Z_j(\tau)) d\tau \\ &+ \sum_{j=1}^{s+1} \sum_{l=1}^{s} \omega_{ij} \alpha_{jl} \int_0^1 k_y(T_j(\tau), Y_j(\tau), Z_j(\tau)) d\tau \cdot v(t_0 + c_l h, Y_l, Z_l) \\ &+ \sum_{j=1}^{s+1} \sum_{l=1}^{s} \omega_{ij} \widehat{\alpha}_{jl} \int_0^1 k_z(T_j(\tau), Y_j(\tau), Z_j(\tau)) d\tau \cdot f(t_0 + c_l h, Y_l, Z_l, \Psi_l) \\ &+ \sum_{j=1}^{s+1} \sum_{l=0}^{s} \omega_{ij} \widetilde{\alpha}_{jl} \int_0^1 k_z(T_j(\tau), Y_j(\tau), Z_j(\tau)) d\tau \cdot r(t_0 + \widetilde{c}_l h, \widetilde{Y}_l, \Lambda_l) \end{aligned} \tag{5.41}$$

for $i = 1, \dots, s$. By (5.34d,e), for the values $Y_i(0) = y_0$, $Z_i(0) = z_0$, $\Psi_i(0) = \psi_0$, for $i = 1, \dots, s$, and $\widetilde{Y}_i(0) = y_0$, $\Lambda_i(0) = \lambda_0$ for $i = 0, \dots, s$, and $y_1(0) = y_0$, $z_1(0) = z_0$,

we obtain, for $h \to 0$,

$$\begin{aligned}\frac{1}{h}\sum_{j=1}^{s+1}\omega_{ij}k(t_0+c_jh,Y_j,Z_j) &= \frac{1}{h}\sum_{j=1}^{s+1}\omega_{ij}k(t_0,y_0,z_0)\\ &+\sum_{j=1}^{s+1}\omega_{ij}c_j\big[k_t(t_0,y_0,z_0)+k_y(t_0,y_0,z_0)v(t_0,y_0,z_0)\\ &\qquad +k_z(t_0,y_0,z_0)f(t_0,y_0,z_0,\psi_0)+k_z(t_0,y_0,z_0)r(t_0,y_0,\lambda_0)\big]+\mathcal{O}(h)\\ &= o(1).\end{aligned}$$

Hence, the values $Y_i(0)=y_0$, $Z_i(0)=z_0$, $\Psi_i(0)=\psi_0$, for $i=1,\ldots,s$, and $\widetilde{Y}_i(0)=y_0$, $\Lambda_i(0)=\lambda_0$ for $i=0,\ldots,s$, and $y_1(0)=y_0$, $z_1(0)=z_0$ satisfy (5.35a,b,c) and the constraints

$$0=\frac{1}{h}\sum_{j=1}^{s+1}\omega_{ij}k(t_0,y_0,z_0)+\ldots, \tag{5.42}$$

with the additional terms coming from (5.41).

Putting everything together, we replace $y_1$ and $z_1$ in (5.37), (5.39) by using (5.35d,e). Using tensor matrix product notation, the Jacobian of equations (5.35a,d), (5.35b), (5.35c,e), (5.37), (5.39), and (5.42) with respect to $Y_i$ ($i=1,\ldots,s$), $\widetilde{Y}_i$ ($i=1,\ldots,s$), $Z_i$ ($i=1,\ldots,s$), $\Lambda_i$ ($i=0,1,\ldots,s$), and $\Psi_i$ ($i=1,\ldots,s$) with $h=0$, is of the form

$$\begin{bmatrix} I_{sn_y} & 0 & 0 & 0 & 0\\ 0 & I_{sn_y} & 0 & 0 & 0\\ 0 & 0 & I_{sn_z} & 0 & 0\\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) & \chi(t_0,y_0,z_0,\lambda_0) & \gamma(t_0,y_0,z_0,\psi_0)\\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) & \xi(t_0,y_0,z_0,\lambda_0) & \rho(t_0,y_0,z_0,\psi_0)\end{bmatrix}$$

with the matrix

$$\begin{bmatrix} \chi(t_0,y_0,z_0,\lambda_0) & \gamma(t_0,y_0,z_0,\psi_0)\\ \xi(t_0,y_0,z_0,\lambda_0) & \rho(t_0,y_0,z_0,\psi_0)\end{bmatrix}$$

given by (5.29). This Jacobian matrix is invertible, as a result of Lemma 5.3.1.

Therefore, if $|h| \leq h_0$, the implicit function theorem yields the existence of a locally unique solution to (5.35abc)-(5.37)-(5.39)-(5.42), hence to the corresponding SPARK method (5.35). □

### 5.3.2 Influence of Perturbations

We now consider the influence of perturbations on the solution of the method (5.7). We consider the perturbed system

$$\widehat{Y}_i = \widehat{y}_0 + h\sum_{j=1}^{s} a_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\delta_i^y, \quad i = 1, \ldots, s \tag{5.43a}$$

$$\widehat{\widetilde{Y}}_i = \widehat{y}_0 + h\sum_{j=1}^{s} \bar{a}_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\widetilde{\delta}_i^y, \quad i = 0, \ldots, s \tag{5.43b}$$

$$\begin{aligned}\widehat{Z}_i = \widehat{z}_0 &+ h\sum_{j=1}^{s} \widehat{a}_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) \\ &+ h\sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_i^z, \quad i = 1, \ldots, s\end{aligned} \tag{5.43c}$$

$$\widehat{y}_1 = \widehat{y}_0 + h\sum_{j=1}^{s} b_j v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\delta_{s+1}^y \tag{5.43d}$$

$$\begin{aligned}\widehat{z}_1 = \widehat{z}_0 &+ h\sum_{j=1}^{s} \widehat{b}_j f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) \\ &+ h\sum_{j=0}^{s} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_{s+1}^z\end{aligned} \tag{5.43e}$$

$$0 = g(t_0 + \widetilde{c}_i h, \widehat{\widetilde{Y}}_i) + h\delta_i^\lambda, \quad i = 0, \ldots, s \tag{5.43f}$$

$$0 = g_t(t_1, \widehat{y}_1) + g_y(t_1, \widehat{y}_1) v(t_1, \widehat{y}_1, \widehat{z}_1) + \delta_{s+1}^\lambda \tag{5.43g}$$

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + \omega_{i,s+1} k(t_1, \widehat{y}_1, \widehat{z}_1) + \delta_i^\psi, \quad i = 1, \ldots, s. \tag{5.43h}$$

Consider also the perturbed form of (5.12)

$$\widehat{Z}_i^f = \widehat{z}_0 + h\sum_{j=1}^{s} \widehat{a}_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) + h\delta_i^f, \quad i = 1, \ldots, s \tag{5.44a}$$

$$\widehat{Z}_i^r = h\sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_i^r, \quad i = 1, \ldots, s. \tag{5.44b}$$

We examine the influence of the perturbations

$$\begin{aligned}
&\delta^y := [\delta_1^{yT}, \ldots, \delta_{s+1}^{y\ T}]^T, \quad \delta^z := [\delta_1^{zT}, \ldots, \delta_{s+1}^{z\ T}]^T, \quad \widetilde{\delta}^y := [\widetilde{\delta}_0^{y^T}, \ldots, \widetilde{\delta}_s^{y^T}]^T,\\
&\delta^\lambda := [\delta_0^{\lambda^T}, \ldots, \delta_{s+1}^{\lambda\ T}]^T, \quad \delta^\psi := [\delta_1^{\psi^T}, \ldots, \delta_s^{\psi^T}]^T,\\
&\delta^f := [\delta_1^{f^T}, \ldots, \delta_s^{f^T}]^T, \quad \delta^r := [\delta_1^{rT}, \ldots, \delta_s^{rT}]^T.
\end{aligned}$$

For simplicity, we introduce the notations

$$\begin{aligned}
Y &:= [Y_1^T, Y_2^T, \ldots, Y_s^T]^T, \quad \widehat{Y} := [\widehat{Y}_1^T, \widehat{Y}_2^T, \ldots, \widehat{Y}_s^T]^T,\\
Z &:= [Z_1^T, Z_2^T, \ldots, Z_s^T]^T, \quad \widehat{Z} := [\widehat{Z}_1^T, \widehat{Z}_2^T, \ldots, \widehat{Z}_s^T]^T,\\
\widetilde{Y} &:= [\widetilde{Y}_0^T, \widetilde{Y}_1^T, \ldots, \widetilde{Y}_s^T]^T, \quad \widehat{\widetilde{Y}} := [\widehat{\widetilde{Y}}_0^T, \widehat{\widetilde{Y}}_1^T, \ldots, \widehat{\widetilde{Y}}_s^T]^T,\\
\Lambda &:= [\Lambda_0^T, \Lambda_1^T, \ldots, \Lambda_s^T]^T, \quad \widehat{\Lambda} := [\widehat{\Lambda}_0^T, \widehat{\Lambda}_1^T, \ldots, \widehat{\Lambda}_s^T]^T,\\
\Psi &:= [\Psi_1^T, \Psi_2^T, \ldots, \Psi_s^T]^T, \quad \widehat{\Psi} := [\widehat{\Psi}_1^T, \widehat{\Psi}_2^T, \ldots, \widehat{\Psi}_s^T]^T,\\
\Delta Y_i &:= \widehat{Y}_i - Y_i, \quad \Delta Z_i := \widehat{Z}_i - Z_i, \quad \Delta \widetilde{Y}_i := \widehat{\widetilde{Y}}_i - \widetilde{Y}_i,\\
\Delta \Lambda_i &:= \widehat{\Lambda}_i - \Lambda_i, \quad \Delta \Psi_i := \widehat{\Psi}_i - \Psi_i,\\
\Delta y_1 &:= \widehat{y}_1 - y_1, \quad \Delta z_1 := \widehat{z}_1 - z_1, \quad \Delta y_0 := \widehat{y}_0 - y_0, \quad \Delta z_0 := \widehat{z}_0 - z_0,\\
\Delta Y &:= \widehat{Y} - Y, \quad \Delta Z := \widehat{Z} - Z, \quad \Delta \widetilde{Y} := \widehat{\widetilde{Y}} - \widetilde{Y},\\
\Delta \Lambda &:= \widehat{\Lambda} - \Lambda, \quad \Delta \Psi := \widehat{\Psi} - \Psi,\\
\Delta \overline{Y} &:= [\widehat{Y}^T - Y^T, \widehat{y}_1^T - y_1^T]^T, \quad \Delta \overline{Z} := [\widehat{Z}^T - Z^T, \widehat{z}_1^T - z_1^T]^T.
\end{aligned}$$

We also define $||Y|| := \max_i\{||Y_i||\}$, $||\Lambda|| := \max_i\{||\Lambda_i||\}$, etc. We will make use of the coefficient matrices

$$\alpha := \begin{bmatrix} A \\ b^T \end{bmatrix}, \quad \widehat{\alpha} := \begin{bmatrix} \widehat{A} \\ \widehat{b}^T \end{bmatrix}, \quad \widetilde{\alpha} := \begin{bmatrix} \widetilde{A} \\ \widetilde{b}^T \end{bmatrix}, \quad \bar{\alpha} := \begin{bmatrix} \bar{A}^* & 0_s \\ 0_s^T & 1 \end{bmatrix},$$

where $\bar{A}^* \in \mathbb{R}^{s\times s}$ equals $\bar{A}$ with the first row removed.

**Theorem 5.3.3.** *Suppose the initial conditions satisfy (5.34). Further, assume that the matrices in (5.2) are invertible. Lastly, we assume*

$$\begin{aligned}
&\Delta y_0 = \mathcal{O}(h^3), \quad \Delta z_0 = \mathcal{O}(h^2),\\
&\Lambda_k - \lambda_0 = \mathcal{O}(h), \quad \Psi_j - \psi_0 = \mathcal{O}(h),\\
&\delta_i^y = \mathcal{O}(h), \quad \widetilde{\delta}_k^y = \mathcal{O}(h^2), \quad \delta_i^z = \mathcal{O}(h), \quad \delta_l^\lambda = \mathcal{O}(h^2),\\
&\delta_j^\psi = \mathcal{O}(h^2), \quad \delta_j^f = \mathcal{O}(1), \quad \delta_j^r = \mathcal{O}(1),
\end{aligned} \tag{5.45}$$

*for* $i = 1, \dots, s+1$, $j = 1, \dots, s$, $k = 0, \dots, s$, *and* $l = 0, \dots, s+1$. *Then for* $|h| \le h_0$, *we have the bounds*

$$\begin{aligned}
\Delta Y_i = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\\
&+ h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)
\end{aligned} \tag{5.46a}$$

$$\begin{aligned}
\Delta \widetilde{Y}_i = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\\
&+ h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)
\end{aligned} \tag{5.46b}$$

$$\begin{aligned}
\Delta Z_i = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\\
&+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)
\end{aligned} \tag{5.46c}$$

$$\begin{aligned}
\Delta y_1 = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\\
&+ h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)
\end{aligned} \tag{5.46d}$$

$$\begin{aligned}
\Delta z_1 = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\\
&+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)
\end{aligned} \tag{5.46e}$$

$$\begin{aligned}
h\Delta\Lambda_i = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\\
&+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)
\end{aligned} \tag{5.46f}$$

$$\begin{aligned}
h\Delta\Psi_i = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\\
&+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).
\end{aligned} \tag{5.46g}$$

*Further, we have the bounds*

$$\begin{aligned}
\Delta Z_i^f = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\\
&+ ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^f|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)
\end{aligned} \tag{5.47a}$$

$$\begin{aligned}\Delta Z_i^r = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^r|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{5.47b}$$

*where we have used the notations*

$$\begin{aligned}\kappa_0 &:= k_y(t_0,y_0,z_0)\Delta y_0 + k_z(t_0,y_0,z_0)\Delta z_0 \\ \eta_0 &:= g_y(t_0,y_0)v_y(t_0,y_0,z_0)\Delta y_0 + g_y(t_0,y_0)v_z(t_0,y_0,z_0)\Delta z_0 \\ &\qquad + g_{ty}(t_0,y_0)\Delta y_0 + g_{yy}(t_0,y_0)(\Delta y_0, v(t_0,y_0,z_0)).\end{aligned}$$

*Proof.* The proof given here uses ideas presented in [12] and [22]. We begin by showing that (5.46b) holds for $\Delta\widetilde{Y}_0$. This comes immediately from (5.43b) and (5.7b), as

$$\Delta\widetilde{Y}_0 = \Delta y_0 + h\widetilde{\delta}_0^y = \mathcal{O}(||\Delta y_0|| + h||\widetilde{\delta}^y||).$$

Subtracting (5.7) from (5.43), and expanding around $Y_j$, $\widetilde{Y}_j$, $Z_j$, $\Lambda_j$, and $\Psi_j$ gives

$$\begin{aligned}\Delta Y_i = \Delta y_0 &+ h\sum_{j=1}^{s} a_{ij}(v_y(t_0+c_jh,Y_j,Z_j)\Delta Y_j + v_z(t_0+c_jh,Y_j,Z_j)\Delta Z_j) \\ &+ h\delta_i^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)\end{aligned} \tag{5.48a}$$

$$\begin{aligned}\Delta\widetilde{Y}_i = \Delta y_0 &+ h\sum_{j=1}^{s} \bar{a}_{ij}(v_y(t_0+c_jh,Y_j,Z_j)\Delta Y_j + v_z(t_0+c_jh,Y_j,Z_j)\Delta Z_j) \\ &+ h\widetilde{\delta}_i^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)\end{aligned} \tag{5.48b}$$

$$\begin{aligned}\Delta Z_i = \Delta z_0 &+ h\sum_{j=1}^{s} \widehat{a}_{ij}(f_y(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta Y_j \\ &\quad + f_z(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta Z_j + f_\psi(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta\Psi_j) \\ &+ h\sum_{j=0}^{s} \widetilde{a}_{ij}(r_y(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\Delta\Lambda_j) + h\delta_i^z \\ &+ \mathcal{O}(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2)\end{aligned} \tag{5.48c}$$

$$\begin{aligned}\Delta y_1 = \Delta y_0 &+ h\sum_{j=1}^{s} b_j(v_y(t_0+c_jh,Y_j,Z_j)\Delta Y_j + v_z(t_0+c_jh,Y_j,Z_j)\Delta Z_j) \\ &+ h\delta_{s+1}^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)\end{aligned} \tag{5.48d}$$

$$
\begin{aligned}
\Delta z_1 = \Delta z_0 + h\sum_{j=1}^{s} \widehat{b}_j (f_y(t_0 + c_j h, Y_j, Z_j)\Delta Y_j + f_z(t_0 + c_j h, Y_j, Z_j)\Delta Z_j \\
+ f_\psi(t_0 + c_j h, Y_j, Z_j, \Psi_j)\Delta\Psi_j) + h\sum_{j=0}^{s} \widetilde{b}_j (r_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j \\
+ r_\lambda(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j) + h\delta^z_{s+1} \\
+ \mathcal{O}(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2)
\end{aligned} \tag{5.48e}
$$

$$
\begin{aligned}
0 = \frac{1}{h} g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\Delta y_0 + g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\sum_{j=0}^{s} \bar{a}_{ij} v_y(t_0 + c_j h, Y_j, Z_j)\Delta Y_j \\
+ g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\sum_{j=0}^{s} \bar{a}_{ij} v_z(t_0 + c_j h, Y_j, Z_j)\Delta Z_j \\
+ \mathcal{O}\left(h||\Delta y_0|| + ||\Delta Y||^2 + ||\Delta Z||^2 + ||\widetilde{\delta}^y|| + ||\delta^\lambda||\right)
\end{aligned} \tag{5.48f}
$$

$$
\begin{aligned}
0 = g_{ty}(t_1, y_1)\Delta y_1 + g_y(t_1, y_1)v_y(t_1, y_1, z_1)\Delta y_1 \\
+ g_y(t_1, y_1)v_z(t_1, y_1, z_1)\Delta z_1 + g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) \\
+ \delta^\lambda_{s+1} + \mathcal{O}\left(||\Delta y_1||^2 + ||\Delta z_1||^2\right)
\end{aligned} \tag{5.48g}
$$

$$
\begin{aligned}
0 = \sum_{j=1}^{s} \omega_{ij} \left(k_y(t_0 + c_j h, Y_j, Z_j)\Delta Y_j + k_z(t_0 + c_j h, Y_j, Z_j)\Delta Z_j\right) \\
+ \omega_{i,s+1}(k_y(t_1, y_1, z_1)\Delta y_1 + k_z(t_1, y_1, z_1)\Delta z_1) \\
+ \delta^\psi_i + \mathcal{O}\left(||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2\right).
\end{aligned} \tag{5.48h}
$$

Note that in (5.48f) we have divided both sides by $h$. This system can be written more compactly using tensor notation. We will use the notation

$$\{v_y\} := \text{blockdiag}(v_y(t_0 + c_1 h, Y_1, Z_1), \ldots, v_y(t_0 + c_s h, Y_s, Z_s))$$

$$\{g_y\} := \text{blockdiag}(g_y(t_0 + \widetilde{c}_1 h, \widetilde{Y}_1), \ldots, g_y(t_0 + \widetilde{c}_s h, \widetilde{Y}_s))$$

$$[r_y] := \text{blockdiag}(r_y(t_0 + \widetilde{c}_0 h, \widetilde{Y}_0, \Lambda_0), \ldots, r_y(t_0 + \widetilde{c}_s h, \widetilde{Y}_s, \Lambda_s))$$

$$\langle k_y\rangle := \text{blockdiag}(k_y(t_0 + c_1 h, Y_1, Z_1), \ldots, k_y(t_0 + c_s h, Y_s, Z_s), k_y(t_1, y_1, z_1))$$

$$\widetilde{[g_y]} := \text{blockdiag}(g_y(t_0 + \widetilde{c}_1 h, \widetilde{Y}_1), \ldots, g_y(t_0 + \widetilde{c}_s h, \widetilde{Y}_s), g_y(t_1, y_1))$$

and so on. We rewrite (5.48a-e,h) as

$$\begin{aligned}\Delta\overline{Y} &= \mathbb{1}_{s+1}\otimes\Delta y_0 + h(\alpha\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\\ &\quad + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\delta^y||\right)\end{aligned} \tag{5.49a}$$

$$\begin{aligned}\Delta\widetilde{Y} &= \mathbb{1}_{s+1}\otimes\Delta y_0 + h(\bar{A}^*\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\\ &\quad + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\widetilde{\delta}^y||\right)\end{aligned} \tag{5.49b}$$

$$\begin{aligned}\Delta\overline{Z} &= \mathbb{1}_{s+1}\otimes\Delta z_0 + h(\widehat{\alpha}\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z + \{f_\psi\}\Delta\Psi)\\ &\quad + h(\widetilde{\alpha}\otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda)\\ &\quad + \mathcal{O}\big(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2 + h||\delta^z||\big)\end{aligned} \tag{5.49c}$$

$$0 = (\widetilde{\Omega}_0\otimes I_{n_k})(\langle k_y\rangle\Delta\widetilde{Y} + \langle k_z\rangle\Delta\widetilde{Z}) + \mathcal{O}\left(||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + ||\delta^\psi||\right). \tag{5.49d}$$

We will now solve for the terms $\Delta\Lambda$ and $\Delta\Psi$. First, we address the constraints $0 = k(t,y,z)$. Substituting (5.49a,c) into (5.49d) gives

$$\begin{aligned}0 &= (\widetilde{\Omega}_0\otimes I_{n_k})\langle k_y\rangle\left[\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\alpha\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right]\\ &\quad + (\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle\big[\mathbb{1}_{s+1}\otimes\Delta z_0 + h(\widehat{\alpha}\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z + \{f_\psi\}\Delta\Psi)\\ &\qquad + h(\widetilde{\alpha}\otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda)\big]\\ &\quad + \mathcal{O}\left(||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2 + h||\delta^y|| + h||\delta^z|| + ||\delta^\psi||\right).\end{aligned}$$

Rewriting with $\Delta\Lambda$ and $\Delta\Psi$ more isolated, we get

$$\begin{aligned}&-h(\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda]\Delta\Lambda - h(\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle(\widehat{\alpha}\otimes I_{n_z})\{f_\psi\}\Delta\Psi =\\ &\qquad (\widetilde{\Omega}_0\otimes I_{n_k})\langle k_y\rangle\left(\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\alpha\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right)\\ &\qquad + (\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle\left(\mathbb{1}_{s+1}\otimes\Delta z_0 + h(\widehat{\alpha}\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z)\right)\\ &\qquad + h(\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_y]\Delta\widetilde{Y}\\ &\qquad + \mathcal{O}\big(||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2\\ &\qquad\qquad + h||\delta^y|| + h||\delta^z|| + ||\delta^\psi||\big).\end{aligned} \tag{5.50}$$

Next, we address the constraints (5.48f,g). Combining (5.48f) with (5.48g), and

rewriting in tensor notation gives

$$\begin{aligned}0 = {} & \widetilde{[g_y]}(\bar{\alpha}\otimes I_n)\left(\langle v_y\rangle\Delta\widetilde{Y} + \langle v_z\rangle\Delta\widetilde{Z}\right)\\ & + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_s\otimes\Delta y_0) \\ g_{ty}(t_1,y_1)\Delta y_1 + g_{yy}(t_1,y_1)(\Delta y_1, v(t_1,y_1,z_1)) \end{bmatrix}\\ & + \mathcal{O}\left(h||\Delta y_0|| + ||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + ||\widetilde{\delta^y}|| + ||\delta^\lambda||\right).\end{aligned} \tag{5.51}$$

Substituting (5.49a,c) into this, we arrive at

$$\begin{aligned}0 = {} & \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_y\rangle\left(\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\alpha\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right)\\ & + \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_z})\langle v_z\rangle\left(\mathbb{1}_{s+1}\otimes\Delta z_0\right.\\ & \qquad \left. + h(\widehat{\alpha}\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z + \{f_\psi\}\Delta\Psi)\right)\\ & + \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_z})\langle v_z\rangle\left(h(\widetilde{\alpha}\otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda)\right)\\ & + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_s\otimes\Delta y_0) \\ g_{ty}(t_1,y_1)\Delta y_1 + g_{yy}(t_1,y_1)(\Delta y_1, v(t_1,y_1,z_1)) \end{bmatrix}\\ & + \mathcal{O}\big(h||\Delta y_0|| + ||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2\\ & \qquad + h||\delta^y|| + ||\widetilde{\delta^y}|| + h||\delta^z|| + ||\delta^\lambda||\big).\end{aligned}$$

Solving this for $\Delta\Lambda$ and $\Delta\Psi$ gives

$$\begin{aligned} & -\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda](h\Delta\Lambda)\\ & \quad - \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widehat{\alpha}\otimes I_{n_z})\{f_\psi\}(h\Delta\Psi) =\\ & \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_y\rangle\left(\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\alpha\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right)\\ & + \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle\left(\mathbb{1}_{s+1}\otimes\Delta z_0 + h(\widehat{\alpha}\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z)\right)\\ & + h\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_y]\Delta\widetilde{Y}\\ & + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_s\otimes\Delta y_0) \\ g_{ty}(t_1,y_1)\Delta y_1 + g_{yy}(t_1,y_1)(\Delta y_1, v(t_1,y_1,z_1)) \end{bmatrix}\\ & + \mathcal{O}\big(h||\Delta y_0|| + ||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2\\ & \qquad + h||\delta^y|| + ||\widetilde{\delta^y}|| + h||\delta^z|| + ||\delta^\lambda||\big).\end{aligned} \tag{5.52}$$

Bringing together (5.52) and (5.50), we get

$$
\begin{aligned}
-\begin{bmatrix} \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda] & \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widehat{\alpha}\otimes I_{n_z})\{f_\psi\} \\ (\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda] & (\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle(\widehat{\alpha}\otimes I_{n_z})\{f_\psi\} \end{bmatrix}. \\
\begin{bmatrix} h\Delta\Lambda \\ h\Delta\Psi \end{bmatrix} = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix},
\end{aligned}
\tag{5.53}
$$

where $M_1$ is the right-hand side of (5.52) and $M_2$ is the right-hand side of (5.50). With $i = 1, \ldots, s$, and $j = 0, \ldots, s$, the first block of the coefficient matrix is

$$
\begin{aligned}
&\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda] \\
&= \begin{bmatrix} \left[\sum_{k=1}^{s} \bar{a}_{ik}\widetilde{a}_{kj} g_y(t_0+\widetilde{c}_i h, \widetilde{Y}_i) v_z(t_0+c_k h, Y_k, Z_k) r_\lambda(t_0+\widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\right] \\ \widetilde{b}_j g_y(t_1, y_1) v_z(t_1, y_1, z_1) r_\lambda(t_0+\widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) \end{bmatrix} \\
&= \begin{bmatrix} \bar{A}^* \widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y(t_0, y_0) v_z(t_0, y_0, z_0) r_\lambda(t_0, y_0, \lambda_0) + \mathcal{O}(h).
\end{aligned}
$$

With $i = 1, \ldots, s$ and $j = 0, \ldots, s$, the lower left block becomes

$$
\begin{aligned}
&(\widetilde{\Omega}_0\otimes I_{n_k})\langle k_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda] \\
&\qquad = \left[\sum_{k=1}^{s+1} \omega_{ik}\widetilde{\alpha}_{kj} k_z(t_0+c_k h, Y_k, Z_k) r_\lambda(t_0+\widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\right] \\
&\qquad = \left[\sum_{k=1}^{s+1} \omega_{ik}\widetilde{\alpha}_{kj} k_z(t_0, y_0, z_0) r_\lambda(t_0, y_0, \lambda_0)\right] + \mathcal{O}(h) \\
&\qquad = \widetilde{\Omega}_0\widetilde{\alpha} \otimes k_z(t_0, y_0, z_0) r_\lambda(t_0, y_0, \lambda_0) + \mathcal{O}(h).
\end{aligned}
$$

With $i = 1, \ldots, s$ and $j = 1, \ldots, s$, the upper right block becomes

$$
\begin{aligned}
&\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widehat{\alpha}\otimes I_{n_z})\{f_\psi\} = \\
&\begin{bmatrix} \sum_{k=1}^{s} \bar{a}_{ik}\widehat{a}_{kj} g_y(t_0+\widetilde{c}_i h, \widetilde{Y}_i) v_z(t_0+c_k h, Y_k, Z_k) f_\psi(t_0+c_j h, Y_j, Z_j, \Psi_j) \\ \widehat{b}_j g_y(t_1, y_1) v_z(t_1, y_1, z_1) f_\psi(t_0+c_j h, Y_j, Z_j, \Psi_j) \end{bmatrix}
\end{aligned}
$$

$$= \begin{bmatrix} \bar{A}^* \widehat{A} \\ \widehat{b}^T \end{bmatrix} \otimes g_y(t_0, y_0) v_z(t_0, y_0, z_0) f_\psi(t_0, y_0, z_0, \psi_0) + \mathcal{O}(h).$$

With $i = 1, \ldots, s$ and $j = 1, \ldots, s$, the last block in the lower right becomes

$$\begin{aligned}
&(\widetilde{\Omega}_0 \otimes I_{n_k}) \langle k_z \rangle (\widehat{\alpha} \otimes I_{n_z}) \{f_\psi\} \\
&\qquad = \left[ \sum_{k=1}^{s+1} \omega_{ik} \widehat{\alpha}_{kj} k_z(t_0 + c_k h, Y_k, Z_k) f_\psi(t_0 + c_j h, Y_j, Z_j, \Psi_j) \right] \\
&\qquad = \left[ \sum_{k=1}^{s+1} \omega_{ik} \widehat{\alpha}_{kj} k_z(t_0, y_0, z_0) f_\psi(t_0, y_0, z_0, \psi_0) \right] + \mathcal{O}(h) \\
&\qquad = \widetilde{\Omega}_0 \widehat{\alpha} \otimes k_z(t_0, y_0, z_0) f_\psi(t_0, y_0, z_0, \psi_0) + \mathcal{O}(h).
\end{aligned}$$

Therefore, as $h \to 0$, the coefficient matrix in (5.53) becomes

$$\begin{bmatrix} \begin{bmatrix} \bar{A}^* \widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y v_z r_\lambda(t_0, y_0, z_0, \lambda_0) & \begin{bmatrix} \bar{A}^* \widehat{A} \\ \widehat{b}^T \end{bmatrix} \otimes g_y v_z f_\psi(t_0, y_0, z_0, \psi_0) \\ \widetilde{\Omega}_0 \widetilde{\alpha} \otimes k_z r_\lambda(t_0, y_0, z_0, \lambda_0) & \widetilde{\Omega}_0 \widehat{\alpha} \otimes k_z f_\psi(t_0, y_0, z_0, \psi_0) \end{bmatrix}. \tag{5.54}$$

Thus, for $h$ sufficiently small, the coefficient matrix in (5.53) is invertible. In (5.52), we express several terms as

$$\begin{aligned}
\frac{1}{h} g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) \Delta y_0 &= \frac{1}{h} g_y(t_0 + (\widetilde{c}_i h), y_0 + (\widetilde{Y}_i - y_0)) \Delta y_0 \\
&= \frac{1}{h} g_y(t_0, y_0) \Delta y_0 + \widetilde{c}_i g_{ty}(t_0, y_0) \Delta y_0 \\
&\qquad + \frac{1}{h} g_{yy}(t_0, y_0)(\Delta y_0, \widetilde{Y}_i - y_0) + \mathcal{O}(h||\Delta y_0||) \\
&= \frac{1}{h} g_y(t_0, y_0) \Delta y_0 + \widetilde{c}_i g_{ty}(t_0, y_0) \Delta y_0 \\
&\qquad + \widetilde{c}_i g_{yy}(t_0, y_0)(\Delta y_0, v(t_0, y_0, z_0)) + \mathcal{O}(h||\Delta y_0||) \\
g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) \sum_{j=1}^{s} \bar{a}_{ij} v_y(t_0 + c_j h, Y_j, Z_j) \Delta y_0 &= \widetilde{c}_i g_y(t_0, y_0) v_y(t_0, y_0, z_0) \Delta y_0 \\
&\qquad + \mathcal{O}(h||\Delta y_0||), \\
g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) \sum_{j=1}^{s} \bar{a}_{ij} v_z(t_0 + c_j h, Y_j, Z_j) \Delta z_0 &= \widetilde{c}_i g_y(t_0, y_0) v_z(t_0, y_0, z_0) \Delta z_0 \\
&\qquad + \mathcal{O}(h||\Delta z_0||),
\end{aligned}$$

$$\Delta y_1 = \Delta y_0 + h\sum_{j=1}^{s} b_j(v(t_0+c_jh,\hat{Y}_j,\hat{Z}_j) - v(t_0+c_jh,Y_j,Z_j)) + h\delta^y_{s+1}$$
$$= \Delta y_0 + \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||),$$
$$g_{ty}(t_1,y_1)\Delta y_1 = g_{ty}(t_0,y_0)\Delta y_0 + \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||),$$
$$g_{yy}(t_1,y_1)(\Delta y_1, v(t_1,y_1,z_1)) = g_{yy}(t_0,y_0)(\Delta y_0, v(t_0,y_0,z_0))$$
$$+ \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||).$$

In addition, (5.50) contains terms that may be written as

$$k_y(t_0+c_ih,Y_i,Z_i)\Delta y_0 = k_y(t_0,y_0,z_0)\Delta y_0 + \mathcal{O}(h||\Delta y_0||)$$
$$k_z(t_0+c_ih,Y_i,Z_i)\Delta z_0 = k_z(t_0,y_0,z_0)\Delta z_0 + \mathcal{O}(h||\Delta z_0||).$$

We therefore get that both $h\Delta\Lambda$ and $h\Delta\Psi$ can be expressed by

$$\begin{aligned}\mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\Delta Y|| + h||\Delta\widetilde{Y}|| + h||\Delta Z|| + h||\Delta\Lambda||^2\\ &+ h||\Delta\Psi||^2 + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi||\\ &+ \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||).\end{aligned} \tag{5.55}$$

The equations (5.48a-e) result in

$$\Delta\overline{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\left(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||\right)$$
$$\Delta\widetilde{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\left(h||\Delta Y|| + h||\Delta Z|| + h||\widetilde{\delta}^y||\right)$$
$$\Delta\overline{Z} = \mathbb{1}_{s+1}\otimes\Delta z_0 + \mathcal{O}\big(h||\Delta Y|| + h||\Delta\widetilde{Y}|| + h||\Delta Z||$$
$$+ h||\Delta\Lambda|| + h||\Delta\Psi|| + h||\delta^z||\big).$$

Reinserting these equations for $\Delta Y$ and $\Delta Z$ into each other and using (5.55) gives

$$\begin{aligned}\Delta\overline{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\\ &+ h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)\end{aligned} \tag{5.56a}$$
$$\begin{aligned}\Delta\widetilde{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\\ &+ h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)\end{aligned} \tag{5.56b}$$

$$\begin{aligned}\Delta\overline{Z} = \mathbb{1}_{s+1}\otimes\Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).\end{aligned} \tag{5.56c}$$

In addition, inserting (5.56) into (5.55) gives

$$\begin{aligned}h||\Delta\Lambda|| = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| + ||\delta^\lambda|| \\ &+ ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{5.57a}$$

$$\begin{aligned}h||\Delta\Psi|| = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| + ||\delta^\lambda|| \\ &+ ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).\end{aligned} \tag{5.57b}$$

The results (5.56) and (5.57) show (5.46).

Subtracting (5.12) from (5.44) and linearizing gives

$$\begin{aligned}\Delta Z_i^f = \Delta z_0 + h\sum_{j=1}^{s}\widehat{a}_{ij}\Big(&f_y(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta Y_j \\ &+ f_z(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta Z_j + f_\psi(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta\Psi_j\Big) \\ &+ h\delta_i^f + \mathcal{O}(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\Delta\Psi||^2) \\ \Delta Z_i^r = h\sum_{j=0}^{s}\widetilde{a}_{ij}&(r_y(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\Delta\Lambda_j) \\ &+ h\delta_i^r + \mathcal{O}(h||\Delta\widetilde{Y}||^2 + h||\Delta\Lambda||^2).\end{aligned}$$

Substituting in the expressions for $\Delta Y$, $\Delta\widetilde{Y}$, $\Delta Z$, $\Delta\Lambda$, and $\Delta\Psi$, we arrive at

$$\begin{aligned}\Delta Z^f = \mathbb{1}_{s+1}\otimes\Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| + ||\delta^\lambda|| \\ &+ ||\delta^\psi|| + h||\delta^f|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big) \\ \Delta Z^r = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi|| \\ &+ h||\delta^r|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).\end{aligned}$$

This completes the proof. □

Each of the constants in the result of Theorem 5.3.3 depend only upon the derivatives of the functions $v$, $f$, $r$, $g$, and $k$, not upon any of the constants from

the hypothesis. With some additional assumptions, the bounds of this perturbation theorem can be simplified.

**Corollary 5.3.4.** *If, in addition to the conditions of Theorem 5.3.3, we assume that*

$$g(t_0, y_0) = 0 = g(t_0, \widehat{y}_0)$$

$$g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) = 0 = g_t(t_0, \widehat{y}_0) + g_y(t_0, \widehat{y}_0)v(t_0, \widehat{y}_0, \widehat{z}_0)$$

$$k(t_0, y_0, z_0) = 0 = k(t_0, \widehat{y}_0, \widehat{z}_0),$$

*then we have the bounds*

$$\begin{aligned}\Delta Y_i = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| \\ &+ h||\widetilde{\delta^y}|| + h^2||\delta^z|| + h||\delta^\lambda|| + h||\delta^\psi||\big)\end{aligned} \tag{5.58a}$$

$$\begin{aligned}\Delta \widetilde{Y}_i = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| \\ &+ h||\widetilde{\delta^y}|| + h^2||\delta^z|| + h||\delta^\lambda|| + h||\delta^\psi||\big)\end{aligned} \tag{5.58b}$$

$$\begin{aligned}\Delta Z_i = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| \\ &+ ||\widetilde{\delta^y}|| + h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big)\end{aligned} \tag{5.58c}$$

$$\begin{aligned}\Delta y_1 = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| \\ &+ h||\widetilde{\delta^y}|| + h^2||\delta^z|| + h||\delta^\lambda|| + h||\delta^\psi||\big)\end{aligned} \tag{5.58d}$$

$$\begin{aligned}\Delta z_1 = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| \\ &+ ||\widetilde{\delta^y}|| + h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big)\end{aligned} \tag{5.58e}$$

$$\begin{aligned}h\Delta \Lambda_i = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| \\ &+ ||\widetilde{\delta^y}|| + h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big)\end{aligned} \tag{5.58f}$$

$$\begin{aligned}h\Delta \Psi_i = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| \\ &+ ||\widetilde{\delta^y}|| + h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big)\end{aligned} \tag{5.58g}$$

$$\begin{aligned}\Delta Z_i^f = \Delta z_0 + \mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + ||\widetilde{\delta^y}|| \\ &+ h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^f||\big)\end{aligned} \tag{5.58h}$$

$$\begin{aligned}\Delta Z_i^r = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| \\ &+ ||\widetilde{\delta^y}|| + h||\delta^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^r||\big).\end{aligned} \tag{5.58i}$$

*Proof.* With these stronger assumptions on the constraints, we subtract and linearize, giving

$$\begin{aligned}
0 &= g(t_0, \widehat{y}_0) - g(t_0, y_0) = g_y(t_0, y_0)\Delta y_0 + \mathcal{O}(||\Delta y_0||^2) \\
0 &= k(t_0, \widehat{y}_0, \widehat{z}_0) - k(t_0, y_0, z_0) = \kappa_0 + \mathcal{O}(||\Delta y_0||^2 + ||\Delta z_0||^2) \\
0 &= g_t(t_0, \widehat{y}_0) + g_y(t_0, \widehat{y}_0)v(t_0, \widehat{y}_0, \widehat{z}_0) - g_t(t_0, y_0) - g_y(t_0, y_0)v(t_0, y_0, z_0) \\
&= \eta_0 + \mathcal{O}(||\Delta y_0||^2 + ||\Delta z_0||^2),
\end{aligned}$$

with $\kappa_0$ and $\eta_0$ as defined in the statement of Theorem 5.3.3. But this means

$$\begin{aligned}
g_y(t_0, y_0)\Delta y_0 &= \mathcal{O}(h^3||\Delta y_0||) \\
\kappa_0 &= \mathcal{O}(h^3||\Delta y_0|| + h^2||\Delta z_0||) \\
\eta_0 &= \mathcal{O}(h^3||\Delta y_0|| + h^2||\Delta z_0||),
\end{aligned}$$

as $\Delta y_0 = \mathcal{O}(h^3)$, $\Delta z_0 = \mathcal{O}(h^2)$. The conclusion (5.58) now follows by applying these bounds to the results of Theorem 5.3.3. □

## 5.4 Discontinuous Collocation Type Methods

We present here discontinuous collocation type methods for solving problem (5.1) with mixed index 2 and 3 constraints. Similar results can be found in [8], [9], and [10] for a different class of methods applied to index 3 and unconstrained problems.

**Definition 5.4.1.** *Let $c_1, \ldots, c_s$ be distinct real numbers, and $\widetilde{c}_0, \ldots, \widetilde{c}_s$ also be distinct real numbers, with $\widetilde{c}_0 = 0$ and $\widetilde{c}_s = 1$. Assume also that $\widetilde{b}_0$ and $\widetilde{b}_s$ are positive real numbers. We then define the s-degree polynomials $Y(t)$, $\Lambda(t)$, $\Psi(t)$, and $Z^f(t)$ and the $(s+1)$-degree polynomials $Z(t)$ and $Z^r(t)$ as the polynomials*

*satisfying the initial conditions*

$$\begin{aligned} Y(t_0) &= y_0, \\ Z^f(t_0) &= z_0, \quad Z^r(t_0) = -h\widetilde{b}_0\widetilde{\mu}(t_0), \\ Z(t_0) &= Z^f(t_0) + Z^r(t_0) = z_0 - h\widetilde{b}_0\widetilde{\mu}(t_0), \end{aligned} \tag{5.59}$$

*where*

$$\widetilde{\mu}(t) := \dot{Z}^r(t) - r(t, Y(t), \Lambda(t)),$$

*as well as the conditions*

$$\dot{Y}(t_0 + c_i h) = v(t_0 + c_i h, Y(t_0 + c_i h), Z(t_0 + c_i h)), \quad i = 1, \ldots, s \tag{5.60a}$$

$$\dot{Z}^f(t_0 + c_i h) = f(t_0 + c_i h, Y(t_0 + c_i h), Z(t_0 + c_i h), \Psi(t_0 + c_i h)), \quad i = 1, \ldots, s \tag{5.60b}$$

$$\dot{Z}^r(t_0 + \widetilde{c}_i h) = r(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h), \Lambda(t_0 + \widetilde{c}_i h)), \quad i = 1, \ldots, s-1 \tag{5.60c}$$

$$Z(t) = Z^f(t) + Z^r(t) \tag{5.60d}$$

$$0 = g(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h)), \quad i = 0, \ldots, s \tag{5.60e}$$

$$0 = g_t(t_1, Y(t_1)) + g_y(t_1, Y(t_1))v(t_1, Y(t_1), Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)) \tag{5.60f}$$

$$\begin{aligned} 0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y(t_0 + c_j h), Z(t_0 + c_j h)) \\ + \omega_{i,s+1} k(t_1, Y(t_1), Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)), \quad i = 1, \ldots, s. \end{aligned} \tag{5.60g}$$

*The coefficients $\omega_{ij}$ are from a matrix $\widetilde{\Omega}_0$ of the form*

$$\widetilde{\Omega}_0 := \begin{bmatrix} 0_s^T & 1 \\ b^T & 0 \\ b^T C & 0 \\ \vdots & \vdots \\ b^T C^{s-2} & 0 \end{bmatrix} \in \mathbb{R}^{s \times (s+1)},$$

*with $b \in \mathbb{R}^s$ and $C^k \in \mathbb{R}^{s \times s}$ defined by*

$$b_j := \int_0^1 \ell_j(\tau) d\tau, \qquad \ell_j(\tau) := \prod_{\substack{k=1 \\ k \neq j}}^{s} \left( \frac{\tau - c_k}{c_j - c_k} \right), \qquad C^k := \operatorname{diag}(c_1^k, \ldots, c_s^k).$$

*The polynomials $Y(t)$, $Z(t)$, $\Lambda(t)$, $\Psi(t)$, $Z^f(t)$, and $Z^r(t)$ are referred to as* discontinuous collocation type polynomials. *The values of $Y(t_1)$ and $Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)$ are used as approximations to the exact solutions $y(t)$ and $z(t)$, respectively, of (5.1) at time $t_1 := t_0 + h$.*

In general, the discontinuous collocation type polynomials at the times $t_0$ and $t_1$ do not satisfy the constraints $0 = k(t, y, z)$. However, as a consequence of $\widetilde{c}_0 = 0$ and $\widetilde{c}_s = 1$, condition (5.60e) does give that the constraints $0 = g(t, y)$ are satisfied by the discontinuous collocation polynomials at both $t_0$ and $t_1$.

**Theorem 5.4.2.** *The discontinuous collocation type polynomials $Y(t)$, $Z(t)$, $\Lambda(t)$, and $\Psi(t)$ defined by (5.60) are equivalent to an $(s, s)$-stage SPARK method for mixed index 2 and 3 problems. Given $\widetilde{b}_0$ and $\widetilde{b}_s$, the remaining coefficients are determined by*

$$a_{ij} = \widehat{a}_{ij} = \int_0^{c_i} \ell_j(\tau)d\tau, \quad b_j = \widehat{b}_j = \int_0^1 \ell_j(\tau)d\tau, \quad i, j = 1, \ldots, s, \tag{5.61a}$$

$$\bar{a}_{ij} = \int_0^{\widetilde{c}_i} \ell_j(\tau)d\tau, \quad i = 0, \ldots, s, \ j = 1, \ldots, s, \tag{5.61b}$$

$$\left.\begin{array}{ll} \widetilde{a}_{ij} = \int_0^{c_i} \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0), & i = 1, \ldots, s, \ j = 0, \ldots, s, \\ \widetilde{a}_{i0} = \widetilde{b}_0, \quad \widetilde{a}_{is} = 0, & i = 1, \ldots, s, \\ \widetilde{b}_j = \int_0^1 \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0) - \widetilde{b}_s\widehat{\ell}_j(\widetilde{c}_s), & j = 1, \ldots, s-1, \end{array}\right\} \tag{5.61c}$$

*where the functions $\ell_j(\tau)$ and $\widehat{\ell}_j(\tau)$ are Lagrange polynomials given by*

$$\ell_j(\tau) = \prod_{\substack{k=1 \\ k\neq j}}^{s} \left(\frac{\tau - c_k}{c_j - c_k}\right) \qquad \widehat{\ell}_j(\tau) = \prod_{\substack{k=1 \\ k\neq j}}^{s-1} \left(\frac{\tau - \widetilde{c}_k}{\widetilde{c}_j - \widetilde{c}_k}\right).$$

*Proof.* Most of this proof is the same as the proof of Theorem 3.5.2. We give here only the pieces which are different. For the discontinuous collocation type polynomial $Z^f(t)$, we have

$$\dot{Z}^f(t_0 + \tau h) = \sum_{j=1}^{s} \ell_j(\tau)\dot{Z}^f(t_0 + c_j h).$$

The polynomial $Z^f(t)$ can thus be expressed as

$$\begin{aligned} Z^f(t_0 + c_i h) &= z_0 + h \int_0^{c_i} \dot{Z}^f(t_0 + \tau h) d\tau \\ &= z_0 + h \sum_{j=1}^{s} \int_0^{c_i} \ell_j(\tau) d\tau \; f(t_0 + c_j h, Y_j, Z_j, \Psi_j), \end{aligned}$$

for $i = 1, \ldots, s$. Taking $Z_i^f := Z^f(t_0 + c_i h)$, $\Psi_j := \Psi(t_0 + c_j h)$, and $\widehat{a}_{ij} := \int_0^{c_i} \ell_j(\tau) d\tau$, we arrive at (5.12a). Working with $Z^f(t_1)$ and $Z^r(t_1)$ gives

$$Z^f(t_1) = z_0 + h \sum_{j=1}^{s} \int_0^1 \ell_j(\tau) d\tau \; f(t_0 + c_j h, Y_j, Z_j, \Psi_j) \tag{5.62}$$

$$\begin{aligned} Z^r(t_1) = {} & h \widetilde{b}_0 r(t_0, Y(t_0), \Lambda(t_0)) \\ & + h \sum_{j=1}^{s-1} \left( \int_0^1 \widehat{\ell}_j(\tau) d\tau - \widetilde{b}_0 \widehat{\ell}_j(\widetilde{c}_0) \right) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j). \end{aligned} \tag{5.63}$$

Applying these formulas, and using $z_1 := Z^f(t_1) + Z^r(t_1) - h \widetilde{b}_s \widetilde{\mu}(t_1)$, the numerical approximation of $z(t_1)$ becomes

$$\begin{aligned} z_1 = {} & z_0 + h \sum_{j=1}^{s} \int_0^1 \ell_j(\tau) d\tau \; f(t_0 + c_j h, Y_j, Z_j, \Psi_j) \\ & + h \sum_{j=1}^{s-1} \left( \int_0^1 \widehat{\ell}_j(\tau) d\tau - \widetilde{b}_0 \widehat{\ell}_j(\widetilde{c}_0) - \widetilde{b}_s \widehat{\ell}_j(\widetilde{c}_s) \right) r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) \\ & + h \widetilde{b}_0 r(t_0, Y(t_0), \Lambda(t_0)) + h \widetilde{b}_s r(t_1, Y(t_1), \Lambda(t_1)). \end{aligned}$$

If we define $\widehat{b}_j := \int_0^1 \ell_j(\tau) d\tau$ for $j = 1, \ldots, s$ and $\widetilde{b}_j := \int_0^1 \widehat{\ell}_j(t) d\tau - \widetilde{b}_0 \widehat{\ell}_j(\widetilde{c}_0) - \widetilde{b}_s \widehat{\ell}_j(\widetilde{c}_s)$ for $j = 1, \ldots, s-1$, then this approximation agrees with that of (5.7e).

We must check that the discontinuous collocation type method satisfies the constraint $0 = k(t_1, y_1, z_1)$. However, this follows immediately from the definitions of $y_1 := Y(t_1)$, $z_1 := Z^f(t_1) + Z^r(t_1) - h \widetilde{b}_s \widetilde{\mu}(t_1)$, and from (5.60). □

Certain SPARK methods applied to mixed index 2 and 3 problems, including the Gauss-Lobatto methods, can be expressed as discontinuous collocation methods. This fact is useful for determining the order of SPARK methods. We take advantage

of this later for computing the local error of the Gauss-Lobatto methods. For now, we give the equivalence of SPARK methods and discontinuous collocation type methods in the following theorem.

**Theorem 5.4.3.** *A SPARK method with distinct values $c_1, \ldots, c_s$, distinct values $\widetilde{c}_0, \ldots, \widetilde{c}_s$, and coefficients $\widehat{a}_{ij} = a_{ij}$, $\widehat{b}_j = b_j$, is a discontinuous collocation type method (5.60) if and only if the coefficients satisfy*

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \qquad \sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, s \tag{5.64a}$$

$$\sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} = \frac{c_i^k}{k}, \qquad \sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, s-1 \tag{5.64b}$$

$$\widetilde{a}_{i0} = \widetilde{b}_0, \qquad \widetilde{a}_{is} = 0 \tag{5.64c}$$

$$\sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad k = 1, \ldots, s-1. \tag{5.64d}$$

*Proof.* The proof of this theorem is identical to that of Theorem 3.5.3 from Chapter 3. We therefore omit it. □

We present here a lemma regarding the error of the internal stages of a discontinuous collocation type or SPARK method for mixed index 2 and 3 DAEs, assuming Gauss-Lobatto coefficients. This lemma will be useful for showing the effectiveness of the derivatives of discontinuous collocation type methods, as well as for a proof determining the local error.

**Lemma 5.4.4.** *Suppose the internal stages $Y_i$, $\widetilde{Y}_j$, $Z_i$, $\Lambda_j$, and $\Psi_i$ are as defined in (5.7) with Gauss-Lobatto coefficients, for $i = 1, \ldots, s$, and $j = 0, \ldots, s$. Let $y(t), z(t), \lambda(t), \psi(t)$ be the exact solutions to (5.1), and let $z^f(t)$ and $z^r(t)$ be the*

*exact solutions to (5.13). Then we have the bounds*

$$\begin{aligned}
Y_i - y(t_0 + c_i h) &= \mathcal{O}(h^{s+1}), \qquad Z_i - z(t_0 + c_i h) = \mathcal{O}(h^{s+1}), \\
\widetilde{Y}_i - y(t_0 + \widetilde{c}_i h) &= \mathcal{O}(h^{s+2}), \\
\Lambda_i - \lambda(t_0 + \widetilde{c}_i h) &= \mathcal{O}(h^s), \qquad \Psi_i - \psi(t_0 + c_i h) = \mathcal{O}(h^s), \\
Z_i^f - z^f(t_0 + c_i h) &= \mathcal{O}(h^{s+1}), \quad Z_i^r - z^r(t_0 + \widetilde{c}_i h) = \mathcal{O}(h^{s+1}), \\
y_1 - y(t_1) &= \mathcal{O}(h^{s+1}), \qquad z_1 - z(t_1) = \mathcal{O}(h^{s+1}).
\end{aligned} \tag{5.65}$$

*Proof.* We apply Corollary 5.3.4 using the exact solution for the perturbed values. So we take

$$\begin{aligned}
&\widehat{Y}_i = y(t_0 + c_i h), \quad \widehat{Z}_i = z(t_0 + c_i h), \quad \widehat{\widetilde{Y}}_i = Y(t_0 + \widetilde{c}_i h), \\
&\widehat{\Lambda}_i = \lambda(t_0 + \widetilde{c}_i h), \quad \widehat{\Psi}_i = \psi(t_0 + c_i h), \\
&\widehat{y}_1 = y(t_1), \quad \widehat{z}_1 = z(t_1), \\
&\widehat{y}_0 = y(t_0), \quad \widehat{z}_0 = z(t_0), \\
&\widehat{Z}_i^f = z^f(t_0 + c_i h), \quad \widehat{Z}_i^r = z^r(t_0 + \widetilde{c}_i h).
\end{aligned}$$

Because the exact solution satisfies

$$\begin{aligned}
&g(t_0 + \widetilde{c}_i h, y(t_0 + \widetilde{c}_i h)) = g(t_1, y(t_1)) = 0 \\
&g_t(t_1, y_1) + g_y(t_1, y(t_1)) v(t_1, y(t_1), z(t_1)) = 0 \\
&k(t_0 + c_i h, y(t_0 + c_i h), z(t_0 + c_i h)) = k(t_1, y(t_1), z(t_1)) = 0,
\end{aligned}$$

the constraints (5.43f,g,h) give that $\delta_j^\lambda = \delta_i^\psi = 0$ for all $i = 1, \ldots, s$ and $j = 0, \ldots, s+1$. Using a Taylor series expansion around $h = 0$, the values $\widehat{Y}_i = y(t_0 + c_i h)$ and $\widehat{Z}_i = z(t_0 + c_i h)$ can be expressed as

$$\widehat{Y}_i = y(t_0 + c_i h) = \sum_{k=0}^{s} \frac{h^k}{k!} c_i^k y^{(k)}(t_0) + \mathcal{O}(h^{s+1}) \tag{5.66}$$

$$\widehat{Z}_i = z(t_0 + c_i h) = \sum_{k=0}^{s} \frac{h^k}{k!} c_i^k z^{(k)}(t_0) + \mathcal{O}(h^{s+1}). \tag{5.67}$$

Thus (5.43a) gives us, for $i = 1, \ldots, s$,

$$\begin{aligned}
\delta_i^y &= \frac{1}{h}(\widehat{Y}_i - \widehat{y}_0) - \sum_{j=1}^{s} a_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{k!} c_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} \sum_{k=0}^{s} a_{ij} \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \left[ \frac{h^{k-1}}{k!} c_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} a_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{c_i^k}{k} - \sum_{j=1}^{s} a_{ij} c_j^{k-1} \right] + \mathcal{O}(h^{s+1}) \\
&= \frac{h^s}{s!} y^{(s+1)}(t_0) \left[ \frac{c_i^{s+1}}{s+1} - \sum_{j=1}^{s} a_{ij} c_j^s \right] + \mathcal{O}(h^{s+1}) \\
&= \mathcal{O}(h^s).
\end{aligned}$$

We have made use of the important fact that the Gauss coefficients $a_{ij}$ satisfy $C(s)$ in (5.15). For $\delta_{s+1}^y$, a similar derivation can be made, using the fact that the Gauss coefficients $b_i$ also satisfy $B(2s)$, (5.14). This, along with (5.43d), gives

$$\begin{aligned}
\delta_{s+1}^y &= \frac{1}{h}(\widehat{y}_1 - \widehat{y}_0) - \sum_{j=1}^{s} b_j v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) \\
&= \sum_{k=1}^{2s+1} \frac{h^{k-1}}{k!} y^{(k)}(t_0) - \sum_{j=1}^{2s} \sum_{k=0}^{s} b_j \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{2s+1}) \\
&= \sum_{k=1}^{2s+1} \left[ \frac{h^{k-1}}{k!} y^{(k)}(t_0) - \sum_{j=1}^{s} b_j \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{2s+1}) \\
&= \sum_{k=1}^{2s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{1}{k} - \sum_{j=1}^{s} b_j c_j^{k-1} \right] + \mathcal{O}(h^{2s+1}) \\
&= \frac{h^{2s}}{2s!} y^{(2s+1)}(t_0) \left[ \frac{1}{2s+1} - \sum_{j=1}^{s} b_j c_j^{2s} \right] + \mathcal{O}(h^{2s+1}) \\
&= \mathcal{O}(h^{2s}).
\end{aligned}$$

This calculation may be repeated for $\widetilde{\delta}_i^y$. Using (5.43b), for $i = 0, \ldots, s$, gives

$$\widetilde{\delta}_i^y = \frac{1}{h}(\widehat{\widetilde{Y}}_i - \widehat{y}_0) - \sum_{j=1}^{s} \bar{a}_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j)$$

$$
\begin{aligned}
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{k!} \widetilde{c}_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} \sum_{k=0}^{s} \bar{a}_{ij} \frac{h^k}{k!} c_j^k y^{(k+1)}(t_0) + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \left[ \frac{h^{k-1}}{k!} \widetilde{c}_i^k y^{(k)}(t_0) - \sum_{j=1}^{s} \bar{a}_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} y^{(k)}(t_0) \right] + \mathcal{O}(h^{s+1}) \\
&= \sum_{k=1}^{s+1} \frac{h^{k-1}}{(k-1)!} y^{(k)}(t_0) \left[ \frac{\widetilde{c}_i^k}{k} - \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} \right] + \mathcal{O}(h^{s+1}) \\
&= \mathcal{O}(h^{s+1}).
\end{aligned}
$$

We have made use of the important fact that the coefficients $\bar{a}_{ij}$ satisfy $\overline{C}(s+1)$ in (5.21). We can repeat a similar process for the $\delta_i^z$. Using (5.43c), we get

$$
\begin{aligned}
\delta_i^z &= \frac{1}{h}(\widehat{Z}_i - \widehat{z}_0) - \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) - \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) \\
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k z^{(k)}(t_0) \right] - \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, y(t_0 + c_j h), z(t_0 + c_j h), \psi(t_0 + c_j h)) \\
&\qquad - \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, y(t_0 + \widetilde{c}_j h), \lambda(t_0 + \widetilde{c}_j h)) + \mathcal{O}(h^s) \\
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k \left( z^{f^{(k)}}(t_0) + z^{r^{(k)}}(t_0) \right) - \sum_{j=1}^{s} a_{ij} \frac{h^{k-1}}{(k-1)!} c_j^{k-1} z^{f^{(k)}}(t_0) \right. \\
&\qquad \left. - \sum_{j=0}^{s} \widetilde{a}_{ij} \frac{h^{k-1}}{(k-1)!} \widetilde{c}_j^{k-1} z^{r^{(k)}}(t_0) \right] + \mathcal{O}(h^s) \\
&= \sum_{k=1}^{s} \frac{h^{k-1}}{(k-1)!} \left[ \left( \frac{c_i^k}{k} - \sum_{j=1}^{s} a_{ij} c_j^{k-1} \right) z^{f^{(k)}}(t_0) \right. \\
&\qquad \left. + \left( \frac{c_i^k}{k} - \sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} \right) z^{r^{(k)}}(t_0) \right] + \mathcal{O}(h^s) \\
&= \mathcal{O}(h^s).
\end{aligned}
$$

To get to the last step, we use the properties $C(s)$ (5.15) and $\widetilde{C}(s)$ (5.20). For $\delta_{s+1}^z$, we get the similar result

$$
\delta_{s+1}^z = \frac{1}{h}(\widehat{z}_1 - \widehat{z}_0) - \sum_{j=1}^{s} b_j f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) - \sum_{j=0}^{s} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j)
$$

$$
\begin{aligned}
&= \sum_{k=1}^{2s} \left[ \frac{h^{k-1}}{k!} z^{(k)}(t_0) \right] - \sum_{j=1}^{s} b_j f(t_0 + c_j h, y(t_0 + c_j h), z(t_0 + c_j h), \psi(t_0 + c_j h)) \\
&\qquad\qquad - \sum_{j=0}^{s} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, y(t_0 + \widetilde{c}_j h), \lambda(t_0 + \widetilde{c}_j h)) + \mathcal{O}(h^{2s}) \\
&= \sum_{k=1}^{2s} \left[ \frac{h^{k-1}}{k!} \left( z^{f^{(k)}}(t_0) + z^{r(k)}(t_0) \right) - \sum_{j=1}^{s} b_j \frac{h^{k-1}}{(k-1)!} c_j^{k-1} z^{f^{(k)}}(t_0) \right. \\
&\qquad\qquad \left. - \sum_{j=0}^{s} \widetilde{b}_j \frac{h^{k-1}}{(k-1)!} \widetilde{c}_j^{k-1} z^{r(k)}(t_0) \right] + \mathcal{O}(h^{2s}) \\
&= \sum_{k=1}^{2s} \frac{h^{k-1}}{(k-1)!} \left[ \left( \frac{1}{k} - \sum_{j=1}^{s} b_j c_j^{k-1} \right) z^{f^{(k)}}(t_0) \right. \\
&\qquad\qquad \left. + \left( \frac{1}{k} - \sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k-1} \right) z^{r(k)}(t_0) \right] + \mathcal{O}(h^{2s}) \\
&= \mathcal{O}(h^{2s}).
\end{aligned}
$$

The final step is derived using $B(2s)$ in (5.14), and $\widetilde{B}(2s)$ in (5.16). The additional perturbations $\delta_i^f$ and $\delta_i^r$ can be expressed in a manner similar to that of $\delta_i^z$. The relation (5.44a) can be expressed as

$$
\begin{aligned}
\delta_i^f &= \frac{1}{h}(\widehat{Z}_i^f - \widehat{z}_0) - \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) \\
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k z^{f^{(k)}}(t_0) - \sum_{j=1}^{s} a_{ij} \frac{h^{k-1}}{(k-1_!} c_j^{k-1} z^{f^{(k)}}(t_0) \right] + \mathcal{O}(h^s) \\
&= \sum_{k=1}^{s} \frac{h^{k-1}}{(k-1)!} \left( \frac{c_i^k}{k} - \sum_{j=1}^{s} a_{ij} c_j^{k-1} \right) z^{f^{(k)}}(t_0) + \mathcal{O}(h^s) \\
&= \mathcal{O}(h^s).
\end{aligned}
$$

The relation (5.44b) can also be expressed as

$$
\begin{aligned}
\delta_i^r &= \frac{1}{h} \widehat{Z}_i^r - \sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widehat{Y}}_j, \widehat{\Lambda}_j) \\
&= \sum_{k=1}^{s} \left[ \frac{h^{k-1}}{k!} c_i^k z^{r(k)}(t_0) - \sum_{j=0}^{s} \widetilde{a}_{ij} \frac{h^{k-1}}{(k-1)!} \widetilde{c}_j^{k-1} z^{r(k)}(t_0) \right] + \mathcal{O}(h^s)
\end{aligned}
$$

$$= \sum_{k=1}^{s} \frac{h^{k-1}}{(k-1)!} \left[ \left( \frac{c_i^k}{k} - \sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} \right) z^{r(k)}(t_0) \right] + \mathcal{O}(h^s)$$
$$= \mathcal{O}(h^s).$$

Finally, by applying Corollary 5.3.4, we get

$$Y_i - y(t_0 + c_i h) = \mathcal{O}\left(h||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\right) = \mathcal{O}\left(h^{s+1}\right)$$
$$\widetilde{Y}_i - y(t_0 + \widetilde{c}_i h) = \mathcal{O}\left(h^2||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\right) = \mathcal{O}(h^{s+2})$$
$$Z_i - z(t_0 + c_i h) = \mathcal{O}\left(h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\right) = \mathcal{O}\left(h^{s+1}\right)$$
$$\Lambda_i - \lambda(t_0 + \widetilde{c}_i h) = \frac{1}{h}\mathcal{O}\left(h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\right) = \mathcal{O}\left(h^s\right)$$
$$\Psi_i - \psi(t_0 + c_i h) = \frac{1}{h}\mathcal{O}\left(h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\right) = \mathcal{O}\left(h^s\right)$$
$$y_1 - y(t_1) = \mathcal{O}\left(h||\delta^y|| + h||\widetilde{\delta}^y|| + h^2||\delta^z||\right) = \mathcal{O}(h^{s+1})$$
$$z_1 - z(t_1) = \mathcal{O}\left(h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z||\right) = \mathcal{O}(h^{s+1})$$
$$Z_i^f - z^f(t_0 + c_i h) = \mathcal{O}\left(h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| + h||\delta^f||\right) = \mathcal{O}(h^{s+1})$$
$$Z_i^r - z^r(t_0 + \widetilde{c}_i h) = \mathcal{O}\left(h||\delta^y|| + ||\widetilde{\delta}^y|| + h||\delta^z|| + h||\delta^r||\right) = \mathcal{O}(h^{s+1}). \qquad \square$$

The next theorem gives the quality of the derivatives of the approximations by discontinuous collocation type methods. This result is similar to that of [9, Theorem II.7.10] and [10, Theorem VII.4.8].

**Theorem 5.4.5.** *Let $y(t), z(t), \lambda(t), \psi(t)$ be the exact solutions to the problem (5.1). The discontinuous collocation type polynomials $Y(t)$, $Z(t)$, $\Lambda(t)$, and $\Psi(t)$ defined by (5.60) with Gauss coefficients $c_i$ and Lobatto coefficients $\widetilde{c}_i$ satisfy for $k = 0, \ldots, s$ and $t \in [t_0, t_1]$*

$$||Y^{(k)}(t) - y^{(k)}(t)|| \leq C h^{s+1-k}, \tag{5.68a}$$
$$||Z^{f(k)}(t) - z^{f(k)}(t)|| \leq C h^{s+1-k}, \tag{5.68b}$$
$$||Z^{r(k)}(t) - z^{r(k)}(t)|| \leq C h^{s-k}, \tag{5.68c}$$
$$||Z^{(k)}(t) - z^{(k)}(t)|| \leq C h^{s+1-k}, \tag{5.68d}$$

$$||\Lambda^{(k)}(t) - \lambda^{(k)}(t)|| \leq Ch^{s-k}, \tag{5.68e}$$

$$||\Psi^{(k)}(t) - \psi^{(k)}(t)|| \leq Ch^{s-1-k}. \tag{5.68f}$$

*Proof.* As in the proof of Theorem 5.4.2, let $Y_i := Y(t_0 + c_i h)$, $Z_i := Z(t_0 + c_i h)$, $\Lambda_i := \Lambda(t_0 + \widetilde{c}_i h)$, and $\Psi_i := \Psi(t_0 + c_i h)$. Using the convention $c_0 := 0$, we can write the collocation polynomials as

$$Y(t_0 + \tau h) = y_0 \ell_0(\tau) + \sum_{i=1}^{s} Y_i \ell_i(\tau) \tag{5.69a}$$

$$Z^f(t_0 + \tau h) = z_0 \ell_0(\tau) + \sum_{i=1}^{s} Z_i^f \ell_i(\tau) \tag{5.69b}$$

$$Z^r(t_0 + \tau h) = -h\widetilde{b}_0 \widetilde{\mu}(t_0)\widehat{\ell}_0(\tau) + \sum_{i=1}^{s-1} Z_i^r \widehat{\ell}_i(\tau) \tag{5.69c}$$

$$Z(t_0 + \tau h) = Z^f(t_0 + \tau h) + Z^r(t_0 + \tau h) \tag{5.69d}$$

$$\Lambda(t_0 + \tau h) = \sum_{i=0}^{s} \Lambda_i \widetilde{\ell}_i(\tau) \tag{5.69e}$$

$$\Psi(t_0 + \tau h) = \sum_{i=1}^{s} \Psi_i \overline{\ell}_i(\tau) \tag{5.69f}$$

with $\tau \in \mathbb{R}$ Lagrange polynomials

$$\ell_i(\tau) := \prod_{\substack{j=0 \\ j \neq i}}^{s} \left( \frac{\tau - c_j}{c_i - c_j} \right), \qquad \overline{\ell}_i(\tau) := \prod_{\substack{j=1 \\ j \neq i}}^{s} \left( \frac{\tau - c_j}{c_i - c_j} \right),$$

$$\widehat{\ell}_i(\tau) := \prod_{\substack{j=0 \\ j \neq i}}^{s-1} \left( \frac{\tau - \widetilde{c}_j}{\widetilde{c}_i - \widetilde{c}_j} \right), \qquad \widetilde{\ell}_i(\tau) := \prod_{\substack{j=0 \\ j \neq i}}^{s} \left( \frac{\tau - \widetilde{c}_j}{\widetilde{c}_i - \widetilde{c}_j} \right).$$

Applying the Lagrange interpolation formula to the exact solutions gives

$$y(t_0 + \tau h) = y_0 \ell_0(\tau) + \sum_{i=1}^{s} y(t_0 + c_i h)\ell_i(\tau) + \mathcal{O}(h^{s+1}) \tag{5.70a}$$

$$z^f(t_0 + \tau h) = z_0 \ell_0(\tau) + \sum_{i=1}^{s} z^f(t_0 + c_i h)\ell_i(\tau) + \mathcal{O}(h^{s+1}) \tag{5.70b}$$

$$z^r(t_0 + \tau h) = \sum_{i=1}^{s-1} z^r(t_0 + \widetilde{c}_i h)\widehat{\ell}_i(\tau) + \mathcal{O}(h^s) \tag{5.70c}$$

$$z(t_0 + \tau h) = z^f(t_0 + \tau h) + z^r(t_0 + \tau h) \tag{5.70d}$$

$$\lambda(t_0+\tau h) = \sum_{i=0}^{s} \lambda(t_0+\widetilde{c}_i h)\widetilde{\ell}_i(\tau) + \mathcal{O}(h^{s+1}) \tag{5.70e}$$

$$\psi(t_0+\tau h) = \sum_{i=1}^{s} \psi(t_0+c_i h)\overline{\ell}_i(\tau) + \mathcal{O}(h^{s}). \tag{5.70f}$$

We define the functions

$$\overline{Y}(\tau) := y(t_0+\tau h) - \left( y_0\ell_0(\tau) + \sum_{i=1}^{s} y(t_0+c_i h)\ell_i(\tau) \right)$$

$$\overline{Z}^f(\tau) := z^f(t_0+\tau h) - \left( z_0\ell_0(\tau) + \sum_{i=1}^{s} z^f(t_0+c_i h)\ell_i(\tau) \right)$$

$$\overline{Z}^r(\tau) := z^r(t_0+\tau h) - \sum_{i=1}^{s-1} z^r(t_0+\widetilde{c}_i h)\widehat{\ell}_i(\tau)$$

$$\overline{Z}(\tau) := z(t_0+\tau h) - \left( z_0\ell_0(\tau) + \sum_{i=1}^{s} z^f(t_0+c_i h)\ell_i(\tau) + \sum_{i=1}^{s-1} z^r(t_0+\widetilde{c}_i h)\widehat{\ell}_i(\tau) \right)$$

$$\overline{\Lambda}(\tau) := \lambda(t_0+\tau h) - \sum_{i=0}^{s} \lambda(t_0+\widetilde{c}_i h)\widetilde{\ell}_i(\tau)$$

$$\overline{\Psi}(\tau) := \psi(t_0+\tau h) - \sum_{i=1}^{s} \psi(t_0+c_i h)\overline{\ell}_i(\tau).$$

The functions $\overline{Y}(\tau)$ and $\overline{Z}^f(\tau)$ have at least $s+1$ zeros at each $c_i$, for $i = 0, \ldots, s$. $\overline{Z}^r(\tau)$ has $s-1$ zeros at each $\widetilde{c}_i$, for $i = 1, \ldots, s-1$. The function $\overline{\Lambda}$ has at least $s+1$ zeros at each $\widetilde{c}_i$, for $i = 0, \ldots, s$, and $\overline{\Psi}$ has at least $s$ zeros at $c_i$, for $i = 1, \ldots, s$. We have

$$\overline{Y}^{(k)}(\tau) = h^k y^{(k)}(t_0+\tau h) - \left( y_0\ell_0^{(k)}(\tau) + \sum_{i=1}^{s} y(t_0+c_i h)\ell_i^{(k)}(\tau) \right) \tag{5.71a}$$

$$\overline{Z}^{f(k)}(\tau) = h^k z^{f(k)}(t_0+\tau h) - \left( z_0\ell_0^{(k)}(\tau) + \sum_{i=1}^{s} z^f(t_0+c_i h)\ell_i^{(k)}(\tau) \right) \tag{5.71b}$$

$$\overline{Z}^{r(k)}(\tau) = h^k z^{r(k)}(t_0+\tau h) - \left( \sum_{i=1}^{s-1} z^r(t_0+\widetilde{c}_i h)\widehat{\ell}_i^{(k)}(\tau) \right) \tag{5.71c}$$

$$\overline{\Lambda}^{(k)}(\tau) = h^k \lambda^{(k)}(t_0+\tau h) - \left( \sum_{i=0}^{s} \lambda(t_0+\widetilde{c}_i h)\widetilde{\ell}_i^{(k)}(\tau) \right) \tag{5.71d}$$

$$\overline{\Psi}^{(k)}(\tau) = h^k \psi^{(k)}(t_0+\tau h) - \left( \sum_{i=1}^{s} \psi(t_0+c_i h)\overline{\ell}_i^{(k)}(\tau) \right). \tag{5.71e}$$

Applying Rolle's Theorem to each open interval $]c_i, c_{i+1}[$ or $]\widetilde{c}_i, \widetilde{c}_{i+1}[$, we see that $\overline{Y}^{(k)}$, $\overline{Z}^{f(k)}$, and $\overline{\Lambda}^{(k)}$ have $s+1-k$ zeros, $\overline{Z}^{r(k)}$ has $s-1-k$ zeros, and $\overline{\Psi}^{(k)}$ has $s-k$ zeros. Thus, the terms in brackets in (5.71) can be viewed as interpolation polynomials, with $\overline{Y}^{(k)}$, $\overline{Z}^{f(k)}$, and $\overline{\Lambda}^{(k)}$ of degree $s-k$, $\overline{Z}^{r(k)}$ of degree $s-2-k$, and $\overline{\Psi}^{(k)}$ of degree $s-1-k$, for $h^k y^{(k)}(t_0+\tau h)$, or $h^k \psi^{(k)}(t_0+\tau h)$, etc. In other words,

$$h^k Y^{(k)}(t_0+\tau h) = y_0 \ell_0^{(k)}(\tau) + \sum_{i=1}^{s} y(t_0+c_i h)\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}) \tag{5.72a}$$

$$h^k z^{f(k)}(t_0+\tau h) = z_0 \ell_0^{(k)}(\tau) - \sum_{i=1}^{s} z^f(y_0+c_i h)\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}) \tag{5.72b}$$

$$h^k z^{r(k)}(t_0+\tau h) = \sum_{i=1}^{s-1} z^r(t_0+\widetilde{c}_i h)\widehat{\ell}_i^{(k)}(\tau) + \mathcal{O}(h^{s-1}) \tag{5.72c}$$

$$h^k \lambda^{(k)}(t_0+\tau h) = \sum_{i=0}^{s} \lambda(t_0+\widetilde{c}_i h)\widetilde{\ell}_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}) \tag{5.72d}$$

$$h^k \psi^{(k)}(t_0+\tau h) = \sum_{i=1}^{s} \psi(t_0+c_i h)\overline{\ell}_i^{(k)}(\tau) + \mathcal{O}(h^{s}). \tag{5.72e}$$

Therefore, taking $k$ derivatives of (5.69) with respect to $\tau$ and subtracting (5.72), we arrive at the relations

$$\begin{aligned}
h^k \left(Y^{(k)}(t_0+\tau h) - y^{(k)}(t_0+\tau h)\right) &= \sum_{i=1}^{s}(Y_i - y(t_0+c_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}) \\
h^k \left(Z^{f(k)}(t_0+\tau h) - z^{f(k)}(t_0+\tau h)\right) &= \sum_{i=1}^{s}(Z_i^f - z^f(t_0+c_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s+1}) \\
h^k \left(Z^{r(k)}(t_0+\tau h) - z^{r(k)}(t_0+\tau h)\right) &= -h\widetilde{b}_0\widetilde{\mu}(t_0)\widehat{\ell}_0(\tau) \\
&\quad + \sum_{i=1}^{s}(Z_i^r - z^r(t_0+\widetilde{c}_i h))\ell_i^{(k)}(\tau) + \mathcal{O}(h^{s-1}) \\
h^k \left(Z^{(k)}(t_0+\tau h) - z^{(k)}(t_0+\tau h)\right) &= h^k \left(Z^f(t_0+\tau h) - z^f(t_0+\tau h)\right) \\
&\quad + h^k \left(Z^r(t_0+\tau h) - z^r(t_0+\tau h)\right) \\
h^k \left(\Lambda^{(k)}(t_0+\tau h) - \lambda^{(k)}(t_0+\tau h)\right) &= \sum_{i=0}^{s}(\Lambda_i - \lambda(t_0+\widetilde{c}_i h))\widetilde{\ell}_i^{(k)}(\tau) + \mathcal{O}(h^{s+1})
\end{aligned}$$

$$h^k\left(\Psi^{(k)}(t_0+\tau h)-\psi^{(k)}(t_0+\tau h)\right)=\sum_{i=1}^{s}(\Psi_i-\psi(t_0+c_ih))\overline{\ell}_i^{(k)}(\tau)+\mathcal{O}(h^s).$$

Invoking Lemma 5.4.4, and dividing by $h^k$, we arrive at

$$Y^{(k)}(t_0+\tau h)-y^{(k)}(t_0+\tau h)=\mathcal{O}(h^{s+1-k}) \tag{5.73a}$$

$$Z^{f(k)}(t_0+\tau h)-z^{f(k)}(t_0+\tau h)=\mathcal{O}(h^{s+1-k}) \tag{5.73b}$$

$$Z^{r(k)}(t_0+\tau h)-z^{r(k)}(t_0+\tau h)=-h^{1-k}\widetilde{b}_0\widetilde{\mu}(t_0)\widehat{\ell}_0(\tau)+\mathcal{O}(h^{s-1-k}) \tag{5.73c}$$

$$Z^{(k)}(t_0+\tau h)-z^{(k)}(t_0+\tau h)=-h^{1-k}\widetilde{b}_0\widetilde{\mu}(t_0)\widehat{\ell}_0(\tau)+\mathcal{O}(h^{s-1-k}) \tag{5.73d}$$

$$\Lambda^{(k)}(t_0+\tau h)-\lambda^{(k)}(t_0+\tau h)=\mathcal{O}(h^{s+1-k}) \tag{5.73e}$$

$$\Psi^{(k)}(t_0+\tau h)-\psi^{(k)}(t_0+\tau h)=\mathcal{O}(h^{s-k}). \tag{5.73f}$$

However, $\dot{Z}^r(\tau)=\sum_{j=1}^{s-1}\widehat{\ell}_j(\tau)r(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)$ and $r(\tau,Y(\tau),\Lambda(\tau))$ agree for $\tau=\widetilde{c}_j$, with $j=1,\ldots,s-1$. Therefore,

$$\widetilde{\mu}(t_0)=\dot{Z}^r(t_0)-r(t_0,Y(t_0),\Lambda(t_0))=\mathcal{O}(h^{s-1}), \tag{5.74}$$

as $\dot{Z}^r(\tau)$ can be viewed as an interpolation polynomial for $r(\tau,Y(\tau),\Lambda(\tau))$. This gives

$$Z^{r(k)}(t_0+\tau h)-z^{r(k)}(t_0+\tau h)=\mathcal{O}(h^{s-1-k})$$
$$Z^{(k)}(t_0+\tau h)-z^{(k)}(t_0+\tau h)=\mathcal{O}(h^{s-1-k}),$$

proving the result. □

## 5.5 Local Error Analysis

Using the fact that the Gauss-Lobatto SPARK methods are equivalent to a class discontinuous collocation methods, we can determine the local error for these methods.

**Theorem 5.5.1.** *For the $(s,s)$-Gauss-Lobatto SPARK methods (5.7) with consistent initial values $(y_0,z_0,\lambda_0,\psi_0)$ at time $t_0$, assume the matrices given by (5.2) are*

*invertible. Then for $|h| \leq h_0$, the local error is of order $2s$, i.e.,*

$$y_1 - y(t_1) = \mathcal{O}(h^{2s+1}), \qquad z_1 - z(t_1) = \mathcal{O}(h^{2s+1}). \tag{5.75}$$

*Proof.* For the proof, we utilize the discontinuous collocation polynomials $Y(t)$, $Z^f(t)$, $Z^r(t)$, $Z(t)$, $\Lambda(t)$, and $\Psi(t)$ defined by (5.60). We define the defects $\delta(t)$, $\mu(t)$, $\widetilde{\mu}(t)$, $\widetilde{\theta}(t)$, and $\theta(t)$ by

$$\dot{Y}(t) = v(t, Y(t), Z(t)) + \delta(t) \tag{5.76a}$$

$$\dot{Z}^f(t) = f(t, Y(t), Z(t), \Psi(t)) + \mu(t) \tag{5.76b}$$

$$\dot{Z}^r(t) = r(t, Y(t), \Lambda(t)) + \widetilde{\mu}(t) \tag{5.76c}$$

$$0 = g(t, Y(t)) + \widetilde{\theta}(t) \tag{5.76d}$$

$$0 = g_t(t, Y(t)) + g_y(t, Y(t))\left(v(t, Y(t), Z(t)) + \delta(t)\right) + \dot{\widetilde{\theta}}(t) \tag{5.76e}$$

$$0 = k(t, Y(t), Z(t)) + \theta(t). \tag{5.76f}$$

From the definition of the discontinuous collocation type polynomials, the defects satisfy

$$\delta(t_0 + c_i h) = 0, \quad i = 1, \ldots, s \tag{5.77a}$$

$$\mu(t_0 + c_i h) = 0, \quad i = 1, \ldots, s \tag{5.77b}$$

$$\widetilde{\mu}(t_0 + \widetilde{c}_i h) = 0, \quad i = 1, \ldots, s-1 \tag{5.77c}$$

$$\widetilde{\theta}(t_0 + \widetilde{c}_i h) = 0, \quad i = 0, \ldots, s. \tag{5.77d}$$

Writing $t_1 := t_0 + h$, the derivative of $\widetilde{\theta}$ also satisfies

$$\begin{aligned}
\dot{\widetilde{\theta}}(t_0) &= -g_t(t_0, y_0) - g_y(t_0, y_0)(v(t_0, y_0, z_0 - h\widetilde{b}_0\widetilde{\mu}(t_0)) + \delta(t_0)) \\
&= -g_t(t_0, y_0) - g_y(t_0, y_0)\delta(t_0) - g_y(t_0, y_0)v(t_0, y_0, z_0) \\
&\qquad + h\widetilde{b}_0 g_y(t_0, y_0)v_z(t_0, y_0, z_0)\widetilde{\mu}(t_0) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_0)||^2\right) \qquad (5.78) \\
&= -g_y(t_0, Y(t_0))\delta(t_0) \\
&\quad + h\widetilde{b}_0 g_y(t_0, Y(t_0))v_z(t_0, Y(t_0), Z(t_0))\widetilde{\mu}(t_0) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_0)||^2\right),
\end{aligned}$$

$$\begin{aligned}
\dot{\widetilde{\theta}}(t_1) &= -g_t(t_1, y_1) - g_y(t_1, y_1)(v(t_1, y_1, z_1 + h\widetilde{b}_s\widetilde{\mu}(t_1)) + \delta(t_1)) \\
&= -g_t(t_1, y_1) - g_y(t_1, y_1)\delta(t_1) - g_y(t_1, y_1)v(t_1, y_1, z_1) \\
&\qquad - h\widetilde{b}_s g_y(t_1, y_1)v_z(t_1, y_1, z_1)\widetilde{\mu}(t_1) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_1)||^2\right) \qquad (5.79) \\
&= -g_y(t_1, Y(t_1))\delta(t_1) \\
&\quad - h\widetilde{b}_s g_y(t_1, Y(t_1))v_z(t_1, Y(t_1), Z(t_1))\widetilde{\mu}(t_1) + \mathcal{O}\left(h^2||\widetilde{\mu}(t_1)||^2\right).
\end{aligned}$$

From the proof of Theorem 5.4.5 (see (5.74)), we get that

$$\mathcal{O}(h^2||\widetilde{\mu}(t_0)||^2) = \mathcal{O}(h^2||\widetilde{\mu}(t_1)||^2) = \mathcal{O}(h^{2s}).$$

Note that the exact solution satisfies these same relations, but with $\delta \equiv 0$, $\theta \equiv 0$, $\widetilde{\theta} \equiv 0$, $\mu \equiv 0$, and $\widetilde{\mu} \equiv 0$. Using the relations for the exact solution and (5.76), we obtain

$$\begin{aligned}
\dot{Y}(t) - \dot{y}(t) &= v(t, Y(t), Z(t)) - v(t, y(t), z(t)) + \delta(t) \\
\dot{Z}(t) - \dot{z}(t) &= f(t, Y(t), Z(t), \Psi(t)) - f(t, y(t), z(t), \psi(t)) \\
&\qquad + r(t, Y(t), \Lambda(t)) - r(t, y(t), \lambda(t)) + \mu(t) + \widetilde{\mu}(t).
\end{aligned}$$

After linearizing and using the notation $\Delta Y(t) := Y(t) - y(t)$, $\Delta Z(t) := Z(t) - z(t)$,

$\Delta\Lambda(t) := \Lambda(t) - \lambda(t)$, and $\Delta\Psi(t) := \Psi(t) - \psi(t)$, these reduce to

$$\begin{aligned}\dot{Y}(t) - \dot{y}(t) &= v_y(t, y(t), z(t))\Delta Y(t) + v_z(t, y(t), z(t))\Delta Z(t) + \delta(t) \\ &\qquad + \mathcal{O}\left(||\Delta Y(t)||^2 + ||\Delta Z(t)||^2\right)\end{aligned} \tag{5.80}$$

$$\begin{aligned}\dot{Z}(t) - \dot{z}(t) &= f_y(t, y(t), z(t), \psi(t))\Delta Y(t) + f_z(t, y(t), z(t), \psi(t))\Delta Z(t) \\ &\quad + f_\psi(t, y(t), z(t), \psi(t))\Delta\Psi(t) + r_y(t, y(t), \lambda(t))\Delta Y(t) \\ &\quad + r_\lambda(t, y(t), \lambda(t))\Delta\Lambda(t) + \mu(t) + \widetilde{\mu}(t) \\ &\quad + \mathcal{O}\left(||\Delta Y(t)||^2 + ||\Delta Z(t)||^2 + ||\Delta\Lambda(t)||^2 + ||\Delta\Psi(t)||^2\right).\end{aligned} \tag{5.81}$$

The next goal is to solve for the terms $\Delta\Lambda$ and $\Delta\Psi$. Taking the derivative of (5.76e) gives

$$\begin{aligned}0 &= g_{tt}(t, Y(t)) + 2g_{ty}(t, Y(t))(v(t, Y(t), Z(t)) + \delta(t)) \\ &\quad + g_{yy}(t, Y(t))(v(t, Y(t), Z(t)) + \delta(t), v(t, Y(t), Z(t)) + \delta(t)) \\ &\quad + g_y(t, Y(t))\big[v_t(t, Y(t), Z(t)) + v_y(t, Y(t), Z(t))(v(t, Y(t), Z(t)) + \delta(t)) \\ &\qquad + v_z(t, Y(t), Z(t))\big(f(t, Y(t), Z(t), \Psi(t)) \\ &\qquad\qquad + r(t, Y(t), \Lambda(t)) + \mu(t) + \widetilde{\mu}(t)\big) + \dot{\delta}(t)\big] \\ &\quad + \dddot{\widetilde{\theta}}(t).\end{aligned}$$

The exact solution satisfies the similar condition

$$\begin{aligned}0 &= g_{tt}(t, y(t)) + 2g_{ty}(t, y(t))v(t, y(t), z(t)) \\ &\quad + g_{yy}(t, y(t))(v(t, y(t), z(t)), v(t, y(t), z(t))) \\ &\quad + g_y(t, y(t))\big[v_t(t, y(t), z(t)) + v_y(t, y(t), z(t))v(t, y(t), z(t)) \\ &\qquad + v_z(t, y(t), z(t))(f(t, y(t), z(t), \psi(t)) + r(t, y(t), \lambda(t)))\big].\end{aligned}$$

Suppressing the argument $t$ on the exact solution, we subtract these two equations and linearize, giving a relation of the form

$$\begin{aligned}0 &= F_1(t)\Delta Y(t) + F_2(t)\Delta Z(t) + g_y(t, y)v_z(t, y, z)r_\lambda(t, y, \lambda)\Delta\Lambda(t) \\ &\quad + g_y(t, y)v_z(t, y, z)f_\psi(t, y, z, \psi)\Delta\Psi(t) + G_1(t)\delta(t) \\ &\quad + g_y(t, Y(t))v_z(t, Y(t), Z(t))(\mu(t) + \widetilde{\mu}(t)) + g_y(t, Y(t))\dot{\delta}(t) + \dddot{\widetilde{\theta}}(t).\end{aligned} \tag{5.82}$$

The functions $F_1(t)$ and $F_2(t)$ depend only upon the exact solution, not the discontinuous collocation type polynomials, while the function $G_1(t)$ depends only upon the discontinuous collocation type polynomials. Note that there should be a term $\mathcal{O}(||\Delta Y||^2 + ||\Delta Z||^2 + ||\delta|| \cdot ||\Delta Y|| + \cdots)$. This term is omitted for simplicity, as it does not change the results. Taking the derivative of (5.76f) gives

$$\begin{aligned}
0 &= k_t(t, Y(t), Z(t)) + k_y(t, Y(t), Z(t))\dot{Y}(t) + k_z(t, Y(t), Z(t))\dot{Z}(t) + \dot{\theta}(t) \\
&= k_t(t, Y(t), Z(t)) + k_y(t, Y(t), Z(t))\left[v(t, Y(t), Z(t)) + \delta(t)\right] \\
&\quad + k_z(t, Y(t), Z(t))\big[f(t, Y(t), Z(t), \Psi(t)) + r(t, Y(t), \Lambda(t)) + \mu(t) + \widetilde{\mu}(t)\big] + \dot{\theta}(t).
\end{aligned}$$

The exact solution satisfies the similar relation

$$\begin{aligned}
0 = k_t(t, y(t), z(t)) &+ k_y(t, y(t), z(t))v(t, y(t), z(t)) \\
&+ k_z(t, y(t), z(t))\big[f(t, y(t), z(t), \psi(t)) + r(t, y(t), \lambda(t))\big].
\end{aligned}$$

Subtracting these and linearizing gives

$$\begin{aligned}
0 = F_3(t)\Delta Y(t) &+ F_4(t)\Delta Z(t) + G_2(t)\delta(t) \\
&+ k_z(t, y, z)r_\lambda(t, y, \lambda)\Delta\Lambda + k_z(t, y, z)f_\psi(t, y, z, \psi)\Delta\Psi(t) \\
&+ k_z(t, Y(t), Z(t))\left[\mu(t) + \widetilde{\mu}(t)\right] + \dot{\theta}(t),
\end{aligned} \tag{5.83}$$

where $F_3$ and $F_4$ depend only upon the exact solution, and $G_2$ depends only upon the collocation polynomials. Again, we omit higher order terms. Combining (5.82) and (5.83), we get an equation of the form

$$\begin{aligned}
-\begin{bmatrix} g_y v_z r_\lambda(t, y, z, \lambda) & g_y v_z f_\psi(t, y, z, \psi) \\ k_z r_\lambda(t, y, z, \lambda) & k_z f_\psi(t, y, z, \psi) \end{bmatrix} \begin{bmatrix} \Delta\Lambda \\ \Delta\Psi \end{bmatrix} = \\
\begin{bmatrix} F_1(t) \\ F_3(t) \end{bmatrix} \Delta Y(t) + \begin{bmatrix} F_2(t) \\ F_4(t) \end{bmatrix} \Delta Z(t) + \begin{bmatrix} G_1(t) \\ G_2(t) \end{bmatrix} \delta(t) \\
+ \begin{bmatrix} g_y v_z(t, Y(t), Z(t)) \\ k_z(t, Y(t), Z(t)) \end{bmatrix} (\mu(t) + \widetilde{\mu}(t))
\end{aligned}$$

$$+\begin{bmatrix} g_y(t,Y(t)) \\ 0 \end{bmatrix}\dot{\delta}(t)+\begin{bmatrix} 0 \\ I_{n_k} \end{bmatrix}\dot{\theta}(t)+\begin{bmatrix} I_{n_g} \\ 0 \end{bmatrix}\ddot{\widetilde{\theta}}(t).$$

Because the matrix (5.2b) is assumed invertible, we may solve for $\Delta\Lambda$ and $\Delta\Psi$. The equations (5.80) and (5.81) can then be written in the form

$$\begin{aligned}
\Delta\dot{Y}(t) &= W_1(t)\Delta Y(t)+W_2(t)\Delta Z(t)+\delta(t)\\
\Delta\dot{Z}(t) &= W_3(t)\Delta Y(t)+W_4(t)\Delta Z(t)+W_5(t)\delta(t)\\
&\quad+\big[\Upsilon_{11}(t)g_y v_z(t,Y(t),Z(t))+\Upsilon_{12}(t)k_z(t,Y(t),Z(t))\\
&\quad+\Upsilon_{21}(t)g_y v_z(t,Y(t),Z(t))+\Upsilon_{22}(t)k_z(t,Y(t),Z(t))+I_{n_z}\big](\mu(t)+\widetilde{\mu}(t))\\
&\quad+\left[\Upsilon_{11}(t)g_y(t,Y(t))+\Upsilon_{21}(t)g_y(t,Y(t))\right]\dot{\delta}(t)\\
&\quad+\left[\Upsilon_{12}(t)+\Upsilon_{22}(t)\right]\dot{\theta}(t)+\left[\Upsilon_{11}(t)+\Upsilon_{21}(t)\right]\ddot{\widetilde{\theta}}(t),
\end{aligned}$$

where we use the notation

$$\begin{bmatrix} \Upsilon_{11}(t) & \Upsilon_{12}(t) \\ \Upsilon_{21}(t) & \Upsilon_{22}(t) \end{bmatrix}=-\begin{bmatrix} r_\lambda & 0 \\ 0 & f_\psi \end{bmatrix}\begin{bmatrix} g_y v_z r_\lambda & g_y v_z f_\psi \\ k_z r_\lambda & k_z f_\psi \end{bmatrix}^{-1}.$$

The functions $W_i(t)$ for $i=1,\ldots,5$ are in terms of the discontinuous collocation type polynomials, the functions from (5.1), and the defects defined in (5.76). Taking as the resolvent

$$R(t,s)=\begin{bmatrix} R_{11}(t,s) & R_{12}(t,s) \\ R_{21}(t,s) & R_{22}(t,s) \end{bmatrix},\quad R(t,t)=I_{n_y+n_z}$$

the variation of constants formula (see [9, Theorem I.11.2]) gives

$$\begin{aligned}
\Delta Y(t_1) &= R_{12}(t_1,t_0)\Delta Z(t_0) + \int_{t_0}^{t_1} \Big[ R_{11}(t_1,s)\delta(s) + R_{12}(t_1,s)W_5(s)\delta(s) \\
&\quad + R_{12}(t_1,s)\big[\Upsilon_{11}(s)g_y v_z(s,Y(s),Z(s)) + \Upsilon_{12}(s)k_z(s,Y(s),Z(s)) \\
&\qquad + \Upsilon_{21}(s)g_y v_z(s,Y(s),Z(s)) + \Upsilon_{22}(s)k_z(s,Y(s),Z(s)) + I_{n_z}\big]\cdot \\
&\qquad\qquad (\mu(s)+\widetilde{\mu}(s)) \\
&\quad + R_{12}(t_1,s)\left[\Upsilon_{11}(s)g_y(s,Y(s)) + \Upsilon_{21}(s)g_y(s,Y(s))\right]\dot{\delta}(s) \\
&\quad + R_{12}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\dot{\theta}(s) + R_{12}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right]\ddot{\widetilde{\theta}}(s)\Big] ds
\end{aligned} \tag{5.84}$$

$$\begin{aligned}
\Delta Z(t_1) &= R_{22}(t_1,t_0)\Delta Z(t_0) + \int_{t_0}^{t_1} \Big[ R_{21}(t_1,s)\delta(s) + R_{22}(t_1,s)W_5(s)\delta(s) \\
&\quad + R_{22}(t_1,s)\big[\Upsilon_{11}(s)g_y v_z(s,Y(s),Z(s)) + \Upsilon_{12}(s)k_z(s,Y(s),Z(s)) \\
&\qquad + \Upsilon_{21}(s)g_y v_z(s,Y(s),Z(s)) + \Upsilon_{22}(s)k_z(s,Y(s),Z(s)) + I_{n_z}\big]\cdot \\
&\qquad\qquad (\mu(s)+\widetilde{\mu}(s)) \\
&\quad + R_{22}(t_1,s)\left[\Upsilon_{11}(s)g_y(s,Y(s)) + \Upsilon_{21}(s)g_y(s,Y(s))\right]\dot{\delta}(s) \\
&\quad + R_{22}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\dot{\theta}(s) + R_{22}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right]\ddot{\widetilde{\theta}}(s)\Big] ds.
\end{aligned} \tag{5.85}$$

We have made use of the fact that $\Delta Y(t_0) = Y(t_0) - y_0 = 0$. Looking first at $\Delta Y(t_1) = y_1 - y(t_1)$, we apply an integration by parts formula to three of the integrands. The first is

$$\begin{aligned}
&\int_{t_0}^{t_1} R_{12}(t_1,s)\big[\Upsilon_{11}(s)g_y(s,Y(s)) + \Upsilon_{21}(s)g_y(s,Y(s))\big]\dot{\delta}(s)ds \\
&= R_{12}(t_1,s)\big[\Upsilon_{11}(s)g_y(s,Y(s)) + \Upsilon_{21}(s)g_y(s,Y(s))\big]\delta(s)\Big|_{s=t_0}^{t_1} \\
&\qquad - \int_{t_0}^{t_1} \frac{d}{ds}\big(\Upsilon_{11}(s)g_y(s,Y(s)) + \Upsilon_{21}(s)g_y(s,Y(s))\big)\delta(s)ds \\
&= -R_{12}(t_1,t_0)\big[\Upsilon_{11}(t_0)g_y(t_0,Y(t_0)) + \Upsilon_{21}(t_0)g_y(t_0,Y(t_0))\big]\delta(t_0) + \mathcal{O}(h^{2s+1}).
\end{aligned}$$

Next, applying the integration by parts formula gives

$$\begin{aligned}\int_{t_0}^{t_1} R_{12}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\dot{\theta}(s)ds = \\ R_{12}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\theta(s)\Big|_{s=t_0}^{t_1} \\ -\int_{t_0}^{t_1}\frac{d}{ds}\left(R_{12}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\right)\theta(s)ds \\ = -R_{12}(t_1,t_0)\left[\Upsilon_{12}(t_0)+\Upsilon_{22}(t_0)\right]\theta(t_0) \\ -\int_{t_0}^{t_1}\frac{d}{ds}\left(R_{12}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\right)\theta(s)ds. \end{aligned} \tag{5.86}$$

However, we can replace $\theta(t_0)$ by

$$\begin{aligned}\theta(t_0) &= -k(t_0,Y(t_0),Z(t_0)) \\ &= -k(t_0,y_0,z_0-h\widetilde{b}_0\widetilde{\mu}(t_0)) \\ &= -k(t_0,y_0,z_0)+h\widetilde{b}_0k_z(t_0,y_0,z_0)\widetilde{\mu}(t_0)+\mathcal{O}(h^2||\widetilde{\mu}(t_0)||^2) \\ &= h\widetilde{b}_0k_z(t_0,y_0,z_0)\widetilde{\mu}(t_0)+\mathcal{O}(h^2||\widetilde{\mu}(t_0)||^2) \\ &= h\widetilde{b}_0k_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0)+\mathcal{O}(h^{2s}). \end{aligned} \tag{5.87}$$

Again, we have from (5.74) that $\mathcal{O}(||\widetilde{\mu}(t_0)||^2)=\mathcal{O}(h^{2s})$. Define

$$\sigma_{12}(s) := \frac{d}{ds}\left(R_{12}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\right).$$

Therefore (5.86) becomes

$$\begin{aligned}\int_{t_0}^{t_1} R_{12}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\dot{\theta}(s)ds = \\ -h\widetilde{b}_0R_{12}(t_1,t_0)\left[\Upsilon_{12}(t_0)+\Upsilon_{22}(t_0)\right]k_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\ -\int_{t_0}^{t_1}\sigma_{12}(s)\theta(s)ds+\mathcal{O}(h^{2s}).\end{aligned}$$

Finally, the third integration by parts gives

$$\int_{t_0}^{t_1} R_{12}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right]\ddot{\widetilde{\theta}}(s)ds = R_{12}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right]\dot{\widetilde{\theta}}(s)\Big|_{s=t_0}^{t_1}$$

$$
\begin{aligned}
&\qquad\qquad - \int_{t_0}^{t_1} \frac{d}{ds}(R_{12}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right])\dot{\widetilde{\theta}}(s)ds \\
&= R_{12}(t_1,t_0)\left[\Upsilon_{11}(t_0)+\Upsilon_{21}(t_0)\right] g_y(t_0,Y(t_0))\delta(t_0) \\
&\quad - h\widetilde{b}_0 R_{12}(t_1,t_0)\left[\Upsilon_{11}(t_0)+\Upsilon_{21}(t_0)\right] g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\
&\quad - \frac{d}{ds}(R_{12}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right])\widetilde{\theta}(s)\Big|_{s=t_0}^{t_1} \\
&\quad + \int_{t_0}^{t_1} \frac{d^2}{ds^2}(R_{12}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right])\widetilde{\theta}(s)ds + \mathcal{O}\left(h^{2s}\right) \\
&= R_{12}(t_1,t_0)\left[\Upsilon_{11}(t_0)+\Upsilon_{21}(t_0)\right] g_y(t_0,Y(t_0))\delta(t_0) \\
&\quad - h\widetilde{b}_0 R_{12}(t_1,t_0)\left[\Upsilon_{11}(t_0)+\Upsilon_{21}(t_0)\right] g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\
&\quad + \mathcal{O}(h^{2s}).
\end{aligned}
$$

Terms of the form $\mathcal{O}(h^{2s+1})$ are introduced in the two expressions above by applying Gauss and Lobatto quadratures, respectively, to the integrals in the last steps. Substituting these into (5.84), applying Gaussian quadrature on each term with a $\delta(s)$, $\mu^f(s)$, or $\mu^r(s)$, and Lobatto quadrature on the term with a $\widetilde{\mu}(s)$ gives

$$
\begin{aligned}
&y_1 - y(t_1) = \Delta Y(t_1) = R_{12}(t_1,t_0)\Delta Z(t_0) \\
&\quad + h\widetilde{b}_0 R_{12}(t_1,t_0)\big[\Upsilon_{11}(t_0)g_y v_z(t_0,Y(t_0),Z(t_0)) + \Upsilon_{12}(t_0)k_z(t_0,Y(t_0),Z(t_0)) \\
&\quad + \Upsilon_{21}(t_0)g_y v_z(t_0,Y(t_0),Z(t_0)) + \Upsilon_{22}(t_0)k_z(t_0,Y(t_0),Z(t_0)) + I_{n_z}\big]\widetilde{\mu}(t_0) \\
&\quad - R_{12}(t_1,t_0)\big[\Upsilon_{11}(t_0)g_y(t_0,Y(t_0)) + \Upsilon_{21}(t_0)g_y(t_0,Y(t_0))\big]\delta(t_0) \\
&\quad + R_{12}(t_1,t_0)\left[\Upsilon_{11}(t_0)+\Upsilon_{21}(t_0)\right] g_y(t_0,Y(t_0))\delta(t_0) \\
&\quad - h\widetilde{b}_0 R_{12}(t_1,t_0)\left[\Upsilon_{11}(t_0)+\Upsilon_{21}(t_0)\right] g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\
&\quad - h\widetilde{b}_0 R_{12}(t_1,t_0)\left[\Upsilon_{12}(t_0)+\Upsilon_{22}(t_0)\right] k_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\
&\quad - \int_{t_0}^{t_1} \sigma_{12}(s)\theta(s)ds + \mathcal{O}(h^{2s}) \\
&= R_{12}(t_1,t_0)\Delta Z(t_0) + h\widetilde{b}_0 R_{12}(t_1,t_0)\widetilde{\mu}(t_0) + \int_{t_0}^{t_1} \sigma_{12}(s)\theta(s)ds + \mathcal{O}(h^{2s}) \\
&= \int_{t_0}^{t_1} \sigma_{12}(s)\theta(s)ds + \mathcal{O}(h^{2s}).
\end{aligned}
$$

Note that $\theta(t_0 + c_j h)$ satisfies

$$\begin{aligned}\theta(t_0 + c_j h) &= -k(t_0 + c_j h, Y_j, Z_j) \\ &= -k(t_0 + c_j h, y(t_0 + c_j h), z(t_0 + c_j h)) \\ &\qquad + \mathcal{O}(||Y_j - y(t_0 + c_j h)|| + ||Z_j - z(t_0 + c_j h)||) \\ &= \mathcal{O}(h^{s+1}),\end{aligned}$$

because of Lemma 5.4.4. However, using Gaussian quadrature and expanding $\sigma_{12}(t_0 + c_j h)$ in a Taylor series around $h = 0$, the integral therefore becomes

$$\begin{aligned}\int_{t_0}^{t_1} \sigma_{12}(s)\theta(s)ds &= h\sum_{j=1}^{s} b_j \sigma_{12}(t_0 + c_j h)\theta(t_0 + c_j h) + \mathcal{O}(h^{2s+1}) \\ &= \sum_{i=1}^{s}\sum_{j=1}^{s} b_j c_j^{i-1} \frac{h^i}{(i-1)!}\sigma_{12}^{(i-1)}(t_0)\theta(t_0 + c_j h) + \mathcal{O}(h^{s+1})\theta(t_0 + c_j h) + \mathcal{O}(h^{2s+1}) \\ &= \sum_{i=1}^{s}\sum_{j=1}^{s} \omega_{ij} \frac{h^i}{(i-1)!}\sigma_{12}^{(i-1)}(t_0)\theta(t_0 + c_j h) + \mathcal{O}(h^{2s+1}) \\ &= -\sum_{i=1}^{s}\left(\sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Y_j, Z_j)\right) \frac{h^i}{(i-1)!}\sigma_{12}^{(i-1)}(t_0) + \mathcal{O}(h^{2s+1}) \\ &= \mathcal{O}(h^{2s+1}).\end{aligned}$$

Therefore, we obtain that $y_1 - y(t_1) = \mathcal{O}(h^{2s})$. This means that the numerical approximation $y_1$ is of local order at least $2s - 1$. However, because the Gauss-Lobatto methods are symmetric according to Theorem 5.2.3, the local order must be an even $2s$. The $\Delta Z(t_1)$ expression can be handled similarly. We integrate by parts on three of the terms in the integral for $\Delta Z(t_1)$. First, we get

$$\begin{aligned}&\int_{t_0}^{t_1} R_{22}(t_1, s)\left[\Upsilon_{11}(s)g_y(s, Y(s)) + \Upsilon_{21}(s)g_y(s, Y(s))\right]\dot{\delta}(s)ds = \\ &\quad \left[\Upsilon_{11}(t_1)g_y(t_1, Y(t_1)) + \Upsilon_{21}(t_1)g_y(t_1, Y(t_1))\right]\delta(t_1) \\ &\quad - R_{22}(t_1, t_0)\left[\Upsilon_{11}(t_0)g_y(t_0, Y(t_0)) + \Upsilon_{21}(t_0)g_y(t_0, Y(t_0))\right]\delta(t_0) + \mathcal{O}(h^{2s+1}).\end{aligned}$$

Second, we integrate by parts to get

$$\begin{aligned}\int_{t_0}^{t_1} R_{22}(t_1,s)\left[\Upsilon_{11}(s)+\Upsilon_{21}(s)\right]\ddot{\widetilde{\theta}}(s)ds = \\ &-\left[\Upsilon_{11}(t_1)+\Upsilon_{21}(t_1)\right]\cdot \\ &\qquad (g_y(t_1,Y(t_1))\delta(t_1)+h\widetilde{b}_s g_y(t_1,Y(t_1))v_z(t_1,Y(t_1),Z(t_1))\widetilde{\mu}(t_1)) \\ &+R_{22}(t_1,t_0)\left[\Upsilon_{11}(t_0)+\Upsilon_{21}(t_0)\right]\cdot \\ &\qquad (g_y(t_0,Y(t_0))\delta(t_0)+h\widetilde{b}_0 g_y(t_0,Y(t_0))v_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0)) \\ &+\mathcal{O}\left(h^{2s}\right).\end{aligned}$$

Finally, defining $\sigma_{22}(s) := \frac{d}{ds}(R_{22}(t_1,s)[\Upsilon_{12}(s)+\Upsilon_{22}(s)])$, the last integration by parts gives

$$\begin{aligned}&\int_{t_0}^{t_1} R_{22}(t_1,s)\left[\Upsilon_{12}(s)+\Upsilon_{22}(s)\right]\dot{\theta}(s)ds = \\ &\quad\left[\Upsilon_{12}(t_1)+\Upsilon_{22}(t_1)\right]\theta(t_1)-R_{22}(t_1,t_0)\left[\Upsilon_{12}(t_0)+\Upsilon_{22}(t_0)\right]\theta(t_0)-\int_{t_0}^{t_1}\sigma_{22}(s)\theta(s)ds \\ &=\left[\Upsilon_{12}(t_1)+\Upsilon_{22}(t_1)\right]\theta(t_1)-R_{22}(t_1,t_0)\left[\Upsilon_{12}(t_0)+\Upsilon_{22}(t_0)\right]\theta(t_0)+\mathcal{O}(h^{2s+1}),\end{aligned}$$

with the last step coming from results similar to those presented for the integral for $\Delta Y(t_1)$. Once again, applying these relations, Gaussian quadrature, Lobatto quadrature, and (5.77) to (5.85) gives

$$\begin{aligned}z_1-z(t_1) &= \Delta Z(t_1)-h\widetilde{b}_s\widetilde{\mu}(t_1) \\ &= h\widetilde{b}_0 R_{22}(t_1,t_0)\left[\Upsilon_{12}(t_0)+\Upsilon_{22}(t_0)\right]k_z(t_0,Y(t_0),Z(t_0))\widetilde{\mu}(t_0) \\ &\quad +h\widetilde{b}_s\left[\Upsilon_{12}(t_1)+\Upsilon_{22}(t_1)\right]k_z(t_1,Y(t_1),Z(t_1))\widetilde{\mu}(t_1) \\ &\quad +\left[\Upsilon_{12}(t_1)+\Upsilon_{22}(t_1)\right]\theta(t_1)-R_{22}\left[\Upsilon_{12}(t_0)+\Upsilon_{22}(t_0)\right]\theta(t_0) \\ &\quad +\mathcal{O}(h^{2s}).\end{aligned}$$

By (5.87) and the similar result

$$\theta(t_1) = -k(t_1,Y(t_1),Z(t_1))$$

$$
\begin{aligned}
&= -k(t_1, y_1, z_1 + h\widetilde{b}_s\widetilde{\mu}(t_1)) \\
&= -k(t_1, y_1, z_1) - h\widetilde{b}_s k_z(t_1, y_1, z_1)\widetilde{\mu}(t_1) + \mathcal{O}(h^2||\widetilde{\mu}(t_1)||^2) \\
&= -h\widetilde{b}_s k_z(t_1, y_1, z_1)\widetilde{\mu}(t_1) + \mathcal{O}(h^2||\widetilde{\mu}(t_1)||^2) \\
&= -h\widetilde{b}_s k_z(t_1, Y(t_1), Z(t_1))\widetilde{\mu}(t_1) + \mathcal{O}(h^{2s}),
\end{aligned} \tag{5.88}
$$

we find that the expression above for $z_1 - z(t_1)$ simplifies down to

$$z_1 - z(t_1) = \mathcal{O}(h^{2s}).$$

Again, because the Gauss-Lobatto methods are symmetric, this result shows that the local error of $z_1$ is of order $2s$. □

## 5.6 Convergence

In this section, we present a proof for the global convergence of the Gauss-Lobatto SPARK methods applied to problems of the form (5.1).

**Theorem 5.6.1.** *Consider the $(s, s)$-Gauss-Lobatto SPARK methods applied to problem (5.1) with consistent initial conditions $(y_0, z_0)$ at time $t_0$. Suppose that the matrices (5.2) are invertible. Then the $(s, s)$-Gauss-Lobatto SPARK methods are convergent of order $2s$, i.e.*

$$y_N - y(t_N) = \mathcal{O}(h^{2s}), \quad z_N - z(t_N) = \mathcal{O}(h^{2s}), \tag{5.89}$$

*where $y_N$ and $z_N$ are the numerical solution at time $t_N := t_0 + Nh$, for $Nh \leq Const.$*

*Proof.* For the proof of this theorem, we apply techniques similar to those found in [7] and [10] for a different class of methods applied to index 3 problems. This proof uses the Lady Windermere's Fan technique.

Denote by $y_n^0$ and $z_n^0$ the numerical solution for a Gauss-Lobatto SPARK method at time $t_n$ with initial conditions $(y_0, z_0)$ at time $t_0$. Then, let $y_n^l$ and $z_n^l$ be the numerical solution at $t_n$ for the same Gauss-Lobatto SPARK method with initial conditions $(y(t_l), z(t_l))$ at time $t_l$. We will estimate $\Delta y_n^l := y_n^{l+1} - y_n^l$ and

$\Delta z_n^l := z_n^{l+1} - z_n^l$ with $0 \le l \le n$ using the local error of the method.

Using the notation above with $0 \le l \le n-1$, we will show that

$$||y_n^{l+1} - y_n^l|| \le C_0 h^{2s+1}, \qquad ||z_n^{l+1} - z_n^l|| \le C_1 h^{2s+1} \tag{5.90a}$$

$$||y(t_n) - y_n^0|| \le C_2 h^{2s}, \qquad ||z(t_n) - z_n^0|| \le C_3 h^{2s}, \tag{5.90b}$$

for $nh \le Const.$ Showing the this completes the proof, as (5.90b) is the desired result. We proceed by induction on $n$.

For $n = 1$, every condition in (5.90) comes immediately from the fact that the $(s, s)$-Gauss-Lobatto methods are of local order $2s$. Suppose now that the above conditions hold for some value of $n$. From Corollary 5.3.4, we see that

$$\begin{aligned} ||\Delta y_{n+1}^l|| + ||\Delta z_{n+1}^l|| &\le (1 + Ch) \left( ||\Delta y_n^l|| + ||\Delta z_n^l|| \right) \\ &\le (1 + Ch)^{n-l} \left( ||\Delta y_{l+1}^l|| + ||\Delta z_{l+1}^l|| \right) \\ &\le e^C \left( ||\Delta y_{l+1}^l|| + ||\Delta z_{l+1}^l|| \right) \end{aligned}$$

since $(1 + Ch)^{n-l} \le e^C$ if $(n - l)h \le Const.$ Therefore,

$$||\Delta y_{n+1}^l|| \le K \left( ||\Delta y_{l+1}^l|| + ||\Delta z_{l+1}^l|| \right) \tag{5.91a}$$

$$||\Delta z_{n+1}^l|| \le K \left( ||\Delta y_{l+1}^l|| + ||\Delta z_{l+1}^l|| \right). \tag{5.91b}$$

We therefore get the bound

$$\begin{aligned} ||y_{n+1}^{l+1} - y_{n+1}^l|| &= ||\Delta y_{n+1}^l|| \\ &\le K \left( ||\Delta y_{l+1}^l|| + ||\Delta z_{l+1}^l|| \right) \\ &\le K \left( ||y_{l+1}^{l+1} - y_{l+1}^l|| + ||z_{l+1}^{l+1} - z_{l+1}^l|| \right) \\ &= K \left( ||y(t_{l+1}) - y_{l+1}^l|| + ||z(t_{l+1}) - z_{l+1}^l|| \right) \\ &\le C_0 h^{2s+1}, \end{aligned}$$

and similarly,

$$||z_{n+1}^{l+1} - z_{n+1}^{l}|| \leq C_1 h^{2s+1},$$

from the local order of the method. This shows (5.90a) for $n+1$. Finally, summing up all the differences in approximations gives

$$||y(t_{n+1}) - y_{n+1}^{0}|| \leq \sum_{i=0}^{n} ||y_{n+1}^{i+1} - y_{n+1}^{i}|| \leq C_2 h^{2s}$$
$$||z(t_{n+1}) - z_{n+1}^{0}|| \leq \sum_{i=0}^{n} ||z_{n+1}^{i+1} - z_{n+1}^{i}|| \leq C_3 h^{2s}.$$

This shows (5.90b) for $n+1$ and completes the proof. □

# CHAPTER 6
# AN EXTENSION OF MPRK METHODS TO MIXED INDEX 2 AND INDEX 3 DAES

## 6.1 Introduction

This chapter presents an extension to MPRK method (see [21]) for index 2 DAEs to DAEs with mixed index 2 and 3 constraints. We refer to these extended methods as *extended Murua's partitioned Runge-Kutta (EMPRK) methods*. We again consider the overdetermined system of mixed index 2 and 3 DAEs

$$\dot{y} = v(t, y, z) \tag{6.1a}$$

$$\dot{z} = f(t, y, z, \psi) + r(t, y, \lambda) \tag{6.1b}$$

$$0 = g(t, y) \tag{6.1c}$$

$$0 = g_t(t, y) + g_y(t, y)v(t, y, z) \tag{6.1d}$$

$$0 = k(t, y, z) \tag{6.1e}$$

where $y(t) \in \mathbb{R}^{n_y}$, $z(t) \in \mathbb{R}^{n_z}$, $\lambda(t) \in \mathbb{R}^{n_g}$, and $\psi(t) \in \mathbb{R}^{n_k}$, and the functions

$$v : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_y}$$

$$f : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \times \mathbb{R}^{n_k} \to \mathbb{R}^{n_z}$$

$$r : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_g} \to \mathbb{R}^{n_z}$$

$$g : \mathbb{R} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_g}$$

$$k : \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \to \mathbb{R}^{n_k}.$$

We also make the assumption that the matrices

$$g_y(t, y)v_z(t, y, z)r_\lambda(t, y, \lambda) \tag{6.2a}$$

$$\begin{bmatrix} g_y(t, y)v_z(t, y, z)r_\lambda(t, y, \lambda) & g_y(t, y)v_z(t, y, z)f_\psi(t, y, z, \psi) \\ k_z(t, y, z)r_\lambda(t, y, \lambda) & k_z(t, y, z)f_\psi(t, y, z, \psi) \end{bmatrix} \tag{6.2b}$$

are invertible. The invertibility of the matrix (6.2b) allows the system (6.1) to be expressed as a system of ODEs. This gives the existence and uniqueness of a solution to (6.1). As explained in Chapter 5, the class of system (6.1) encompasses constrained Lagrangian and Hamiltonian systems, with reasonable assumptions.

## 6.2 EMPRK Methods

In [21], a class of MPRK methods is proposed for solving DAEs of index 2. This class of methods is given as

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j, \Lambda_j), \quad i = 1, \ldots, s, \tag{6.3a}$$

$$\widetilde{Y}_i = y_0 + h\sum_{j=1}^{s} \bar{a}_{ij} f(t_0 + c_j h, Y_j, \Lambda_j), \quad i = 0, \ldots, s, \tag{6.3b}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j f(t_0 + c_j h, Y_j, \Lambda_j), \tag{6.3c}$$

$$0 = k(t_0 + \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \ldots, s. \tag{6.3d}$$

These methods can be expressed with additional internal stages to handle the constraints $0 = k(t, y, z)$ instead of considering a linear combination. This is similar to the approach taken by the SPARK methods applied to index 3 problems. We propose here an extension of this method to include mixed index 2 and 3 DAEs.

**Definition 6.2.1.** *One step of an $(s, s)$-stage EMPRK method applied to the system (6.1) with stepsize $h$ starting at $(y_0, z_0)$ at time $t_0$ is given by the solution of the nonlinear system of equations*

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} V_j, \quad i = 1, \ldots, s \tag{6.4a}$$

$$Z_i = z_0 + h\sum_{j=1}^{s} a_{ij} F_j + h\sum_{j=0}^{s} \widetilde{a}_{ij} R_j, \quad i = 1, \ldots, s \tag{6.4b}$$

$$\widetilde{Y}_i = y_0 + h\sum_{j=1}^{s} \bar{a}_{ij} V_j, \quad i = 0, \ldots, s \tag{6.4c}$$

$$\widetilde{Z}_i = z_0 + h\sum_{j=1}^{s} \bar{a}_{ij} F_j + h\sum_{j=0}^{s} \check{a}_{ij} R_j, \quad i = 0, \ldots, s \tag{6.4d}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j V_j \tag{6.4e}$$

$$z_1 = z_0 + h\sum_{j=1}^{s} b_j F_j + h\sum_{j=0}^{s} \widetilde{b}_j R_j \tag{6.4f}$$

$$0 = g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{6.4g}$$

$$0 = g(t_1, y_1) \tag{6.4h}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1) v(t_1, y_1, z_1) \tag{6.4i}$$

$$0 = k(t_0 + \widetilde{c}_i h, \widetilde{Y}_i, \widetilde{Z}_i), \quad i = 0, \ldots, s \tag{6.4j}$$

$$0 = k(t_1, y_1, z_1), \tag{6.4k}$$

*where we have used the definitions* $t_1 := t_0 + h$, $V_j := v(t_0 + c_j h, Y_j, Z_j)$, $F_j := f(t_0 + c_j h, Y_j, Z_j, \Psi_j)$, *and* $R_j := r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)$. *The coefficients* $c_i$ *and* $\widetilde{c}_i$ *are determined by*

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad \widetilde{c}_i = \sum_{j=1}^{s} \bar{a}_{ij}. \tag{6.5}$$

We also define the coefficients

$$\alpha := \begin{bmatrix} A \\ b^T \end{bmatrix}, \quad \widetilde{\alpha} := \begin{bmatrix} \widetilde{A} \\ \widetilde{b}^T \end{bmatrix}.$$

It is assumed that the RK coefficients satisfy

$$\bar{a}_{0j} = 0, \quad j = 1, \ldots, s, \tag{6.6a}$$

$$\bar{a}_{sj} = b_j, \quad j = 1, \ldots, s, \tag{6.6b}$$

$$\check{a}_{0j} = 0, \quad j = 0, \ldots, s, \tag{6.6c}$$

$$\check{a}_{sj} = \widetilde{b}_j, \quad j = 0, \ldots, s, \tag{6.6d}$$

$$\sum_{j=1}^{s} \bar{a}_{ij} c_j = \sum_{j=1}^{s} \bar{a}_{ij} \sum_{k=0}^{s} \widetilde{a}_{jk} = \frac{\widetilde{c}_i^2}{2}, \quad i = 0, \ldots, s, \tag{6.6e}$$

$$\bar{A}\widetilde{A} = \begin{bmatrix} 0 & \dots & 0 \\ & \bar{A}^*\widetilde{A} & \end{bmatrix}, \quad \begin{bmatrix} \bar{A}^*\widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \text{ is invertible.} \tag{6.6f}$$

These assumptions are made to ensure the existence and uniqueness of a solution. The coefficients will also be assumed to satisfy

$$\sum_{i=1}^{s} b_i = \sum_{i=0}^{s} \widetilde{b}_i = 1 \tag{6.7a}$$

$$\bar{A}^* \in \mathbb{R}^{s\times s} \text{ and } \widetilde{\alpha} \in \mathbb{R}^{(s+1)\times(s+1)} \text{ are invertible,} \tag{6.7b}$$

$$M := \begin{bmatrix} b^T \\ b^T - b^T A \\ b^T - 2b^T CA \\ \vdots \\ b^T - (s-1)b^T C^{s-2}A \end{bmatrix} \in \mathbb{R}^{s\times s} \text{ is invertible,} \tag{6.7c}$$

$$c_i = \sum_{j=0}^{s} \widetilde{a}_{ij}, \quad i = 1, \dots, s, \tag{6.7d}$$

$$\widetilde{c}_i = \sum_{j=0}^{s} \check{a}_{ij}, \quad i = 0, \dots, s, \tag{6.7e}$$

where $\bar{A}^* \in \mathbb{R}^{s\times s}$ equals $\bar{A}$ with the first row removed. Because of (6.6b,d), the methods satisfy $\widetilde{Y}_s = y_1$ and $\widetilde{Z}_s = z_1$. This means that condition (6.4h) is satisfied by (6.4g) for $i = s$, and that (6.4k) is satisfied by (6.4j). Further, because of conditions (6.6a,c), EMPRK methods satisfy $\widetilde{Y}_0 = y_0$ and $\widetilde{Z}_0 = z_0$, and thus (6.4g,j) are satisfied automatically for $i = 0$. We also note that for problems without constraints (6.1e), these extended methods reduce to the SPARK methods for originally index 3 problems.

As in previous chapters, we introduce extra internal stages into (6.4):

$$Z_i^f = z_0 + h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, Y_j, Z_j, \Psi_j), \quad i = 1, \dots, s \tag{6.8a}$$

$$Z_i^r = h\sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \dots, s \tag{6.8b}$$

$$\widetilde{Z}_i^f = z_0 + h\sum_{j=1}^{s} \bar{a}_{ij} f(t_0 + c_j h, Y_j, Z_j, \Psi_j), \quad i = 0, \ldots, s \tag{6.8c}$$

$$\widetilde{Z}_i^r = h\sum_{j=0}^{s} \check{a}_{ij} r(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 0, \ldots, s. \tag{6.8d}$$

We clearly have $Z_i = Z_i^f + Z_i^r$ and $\widetilde{Z}_i = \widetilde{Z}_i^f + \widetilde{Z}_i^r$. We also introduce the differential equations as part of (6.1)

$$\dot{z}^f = f(t, y, z), \quad z^f(t_0) = z_0 \tag{6.9a}$$

$$\dot{z}^r = r(t, y, \lambda), \quad z^r(t_0) = 0. \tag{6.9b}$$

Again, notice that $z(t) = z^f(t) + z^r(t)$.

### 6.2.1 Gauss-Lobatto EMPRK Methods

An example of a class of EMPRK methods is the $(s, s)$-Gauss-Lobatto EMPRK methods. Many of the coefficients are chosen to be the same as for the Gauss-Lobatto SPARK methods presented in Chapter 5. For the choice of $\check{A}$, we use the Lobatto IIIA coefficients. These methods satisfy (6.6c,d) and (6.7e). The coefficients for the Gauss-Lobatto methods satisfy

$$B(2s): \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s, \tag{6.10}$$

$$C(s): \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \qquad i = 1, \ldots, s,\ k = 1, \ldots, s, \tag{6.11}$$

$$\widetilde{B}(2s): \quad \sum_{i=0}^{s} \widetilde{b}_i \widetilde{c}_i^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s, \tag{6.12}$$

$$D(s): \quad \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{b_j}{k}(1 - c_j^k), \qquad j = 1, \ldots, s,\ k = 1, \ldots, s. \tag{6.13}$$

The condition (6.10) is from the $s$-stage Gaussian quadrature, (6.11) is from the Gauss RK coefficients, and (6.12) is from the $s + 1$ stage Lobatto quadrature. The

coefficients $\bar{a}_{ij}$ and $\widetilde{a}_{ij}$ are chosen to satisfy

$$\overline{C}(s): \quad \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad i = 0, \ldots, s, \ k = 1, \ldots, s, \tag{6.14}$$

$$\widetilde{b}_j \left(1 - \frac{\bar{a}_{ji}}{b_i}\right) = \widetilde{a}_{ij}, \qquad i = 1, \ldots, s, \ j = 0, \ldots, s. \tag{6.15}$$

In Lemma 3.17, we showed that the condition (6.15) is equivalent to the conditions

$$\begin{aligned} &\widetilde{a}_{i0} = \widetilde{b}_0, \quad i = 1, \ldots, s, \\ &\widetilde{C}(s): \quad \sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} = \frac{c_i^k}{k}, \qquad i = 1, \ldots, s, \ k = 1, \ldots, s. \end{aligned} \tag{6.16}$$

We also presented Lemma 3.20 showing that the Gauss-Lobatto coefficients satisfy the extended condition

$$\overline{C}(s+1): \quad \sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad i = 0, \ldots, s, \ k = 1, \ldots, s+1. \tag{6.17}$$

The Lobatto IIIA coefficients $\check{A}$ satisfy

$$\check{C}(s+1): \quad \sum_{j=0}^{s} \check{a}_{ij} \widetilde{c}_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad i = 0, \ldots, s, \ k = 1, \ldots, s+1. \tag{6.18}$$

In Chapters 3 and 5, we showed that the coefficients $A$, $\widetilde{A}$, $\bar{A}$, $b$, $\widetilde{b}$, $c$, and $\widetilde{c}$ satisfy the conditions (6.6) and (6.7).

The Gauss-Lobatto EMPRK methods applied to mixed index 2 and 3 DAEs are symmetric methods. We present this result next.

**Theorem 6.2.2.** *Assume the initial conditions $(y_0, z_0)$ at time $t_0$ are consistent, i.e.,*

$$\begin{aligned} g(t_0, y_0) &= 0 \\ g_t(t_0, y_0) + g_y(t_0, y_0) v(t_0, y_0, z_0) &= 0 \\ k(t_0, y_0, z_0) &= 0. \end{aligned}$$

*Then the Gauss-Lobatto EMPRK methods applied to mixed index 2 and 3 DAEs are*

*symmetric.*

*Proof.* First, we rewrite (6.4h,i,k). Because the initial conditions are assumed consistent, these can be written as

$$0 = g(t_1, y_1) + g(t_0, y_0) \tag{6.19a}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) + g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) \tag{6.19b}$$

$$0 = k(t_1, y_1, z_1) + k(t_0, y_0, z_0). \tag{6.19c}$$

Using the method (6.4a-g,j) and (6.19), we apply the method with stepsize $-h$ starting at time $t_1$ to obtain the system

$$Y_i = y_1 - h\sum_{j=1}^{s} a_{ij}v(t_1 - c_j h, Y_j, Z_j), \quad i = 1, \ldots, s \tag{6.20a}$$

$$Z_i = z_1 - h\sum_{j=1}^{s} a_{ij}f(t_1 - c_j h, Y_j, Z_j, \Psi_j) - h\sum_{j=0}^{s} \widetilde{a}_{ij}r(t_1 - \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \tag{6.20b}$$

$$\widetilde{Y}_i = y_1 - h\sum_{j=1}^{s} \bar{a}_{ij}v(t_1 - c_j h, Y_j, Z_j), \quad i = 0, \ldots, s \tag{6.20c}$$

$$\widetilde{Z}_i = z_1 - h\sum_{j=1}^{s} \bar{a}_{ij}f(t_1 - c_j h, Y_j, Z_j, \Psi_j) - h\sum_{j=0}^{s} \check{a}_{ij}r(t_1 - \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j), \quad i = 0, \ldots, s \tag{6.20d}$$

$$y_0 = y_1 - h\sum_{j=1}^{s} b_j v(t_1 - c_j h, Y_j, Z_j) \tag{6.20e}$$

$$z_0 = z_1 - h\sum_{j=1}^{s} b_j f(t_1 - c_j h, Y_j, Z_j, \Psi_j) - h\sum_{j=0}^{s} \widetilde{b}_j r(t_1 - \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j) \tag{6.20f}$$

$$0 = g(t_1 - \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{6.20g}$$

$$0 = g(t_0, y_0) + g(t_1, y_1) \tag{6.20h}$$

$$0 = g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) + g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{6.20i}$$

$$0 = k(t_1 - \widetilde{c}_i h, \widetilde{Y}_i, \widetilde{Z}_i), \quad i = 0, \ldots, s \tag{6.20j}$$

$$0 = k(t_0, y_0, z_0) + k(t_1, y_1, z_1). \tag{6.20k}$$

Using the definition $t_1 = t_0 + h$, (6.20e,f) become

$$y_1 = y_0 + h \sum_{j=1}^{s} b_j v(t_0 + (1 - c_j)h, Y_j, Z_j)$$

$$z_1 = z_0 + h \sum_{j=1}^{s} b_j f(t_0 + (1 - c_j)h, Y_j, Z_j, \Psi_j) + h \sum_{j=0}^{s} \widetilde{b}_j r(t_0 + (1 - \widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j).$$

Substituting these back into (6.20) and applying the consistency of the initial conditions $(y_0, z_0)$ at time $t_0$ gives

$$Y_i = y_0 + h \sum_{j=1}^{s} (b_j - a_{ij}) v(t_0 + (1 - c_j)h, Y_j, Z_j), \quad i = 1, \ldots, s \tag{6.21a}$$

$$Z_i = z_0 + h \sum_{j=1}^{s} (b_j - a_{ij}) f(t_0 + (1 - c_j)h, Y_j, Z_j, \Psi_j) + h \sum_{j=0}^{s} (\widetilde{b}_j - \widetilde{a}_{ij}) r(t_0 + (1 - \widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \tag{6.21b}$$

$$\widetilde{Y}_i = y_0 + h \sum_{j=1}^{s} (b_i - \bar{a}_{ij}) v(t_0 + (1 - c_j)h, Y_j, Z_j), \quad i = 0, \ldots, s \tag{6.21c}$$

$$\widetilde{Z}_i = z_0 + h \sum_{j=1}^{s} (b_j - \bar{a}_{ij}) f(t_0 + (1 - c_j)h, Y_j, Z_j, \Psi_j) + h \sum_{j=0}^{s} (\widetilde{b}_j - \check{a}_{ij}) r(t_0 + (1 - \widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j), \quad i = 1, \ldots, s \tag{6.21d}$$

$$y_1 = y_0 + h \sum_{j=1}^{s} b_j v(t_0 + (1 - c_j)h, Y_j, Z_j) \tag{6.21e}$$

$$z_1 = z_0 + h\sum_{j=1}^{s} b_j f(t_0 + (1-c_j)h, Y_j, Z_j, \Psi_j) + h\sum_{j=0}^{s} \widetilde{b}_j r(t_0 + (1-\widetilde{c}_j)h, \widetilde{Y}_j, \Lambda_j) \tag{6.21f}$$

$$0 = g(t_0 + (1-\widetilde{c}_i)h, \widetilde{Y}_i), \quad i = 0, \ldots, s \tag{6.21g}$$

$$0 = g(t_1, y_1) \tag{6.21h}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1) \tag{6.21i}$$

$$0 = k(t_0 + (1-\widetilde{c}_i)h, \widetilde{Y}_i, \widetilde{Z}_i), \quad i = 0, \ldots, s. \tag{6.21j}$$

$$0 = k(t_1, y_1, z_1). \tag{6.21k}$$

Lastly, we would like to write (6.21) in the format of (6.4). To do so, we must reindex each $Y_i$, $Z_i$, $\widetilde{Y}_i$, $\widetilde{Z}_i$, $\Lambda_i$, and $\Psi_i$ so as to preserve the usual ordering of the $c_i$, $\widetilde{c}_i$ coefficients. This results in the system

$$Y_i^* = y_0 + h\sum_{j=1}^{s} a_{ij}^* v(t_0 + c_j^* h, Y_j^*, Z_j^*), \quad i = 1, \ldots, s \tag{6.22a}$$

$$Z_i^* = z_0 + h\sum_{j=1}^{s} a_{ij}^* f(t_0 + c_j^* h, Y_j^*, Z_j^*, \Psi_j^*) + h\sum_{j=0}^{s} \widetilde{a}_{ij}^* r(t_0 + \widetilde{c}_j^* h, \widetilde{Y}_j^*, \Lambda_j^*), \quad i = 1, \ldots, s \tag{6.22b}$$

$$\widetilde{Y}_i^* = y_0 + h\sum_{j=1}^{s} \bar{a}_{ij}^* v(t_0 + c_j^* h, Y_j^*, Z_j^*), \quad i = 0, \ldots, s \tag{6.22c}$$

$$\widetilde{Z}_i^* = z_0 + h\sum_{j=1}^{s} \bar{a}_{ij}^* f(t_0 + c_j^* h, Y_j^*, Z_j^*, \Psi_j^*) + h\sum_{j=0}^{s} \check{a}_{ij}^* r(t_0 + \widetilde{c}_j^* h, \widetilde{Y}_j^*, \Lambda_j^*), \quad i = 0, \ldots, s \tag{6.22d}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j^* v(t_0 + c_j^* h, Y_j^*, Z_j^*) \tag{6.22e}$$

$$z_1 = z_0 + h\sum_{j=1}^{s} b_j^* f(t_0 + c_j^* h, Y_j^*, Z_j^*, \Psi_j^*) + h\sum_{j=0}^{s} \widetilde{b}_j^* r(t_0 + \widetilde{c}_j^* h, \widetilde{Y}_j^*, \Lambda_j^*) \tag{6.22f}$$

$$0 = g(t_0 + \widetilde{c}_i^* h, \widetilde{Y}_i^*), \quad i = 0, \ldots, s \tag{6.22g}$$

$$0 = g(t_1, y_1) \tag{6.22h}$$

$$0 = g_t(t_1, y_1) + g_y(t_1, y_1) v(t_1, y_1, z_1) \tag{6.22i}$$

$$0 = k(t_0 + \widetilde{c}_i^* h, \widetilde{Y}_i^*, \widetilde{Z}_i^*), \quad i = 0, \ldots, s \tag{6.22j}$$

$$0 = k(t_1, y_1, z_1), \tag{6.22k}$$

where the stages are defined by

$$Y_i^* := Y_{s+1-i}, \quad Z_i^* := Z_{s+1-i}, \quad \widetilde{Y}_i^* := \widetilde{Y}_{s-i}, \quad \widetilde{Z}_i^* := \widetilde{Z}_{s-i},$$
$$\Lambda_i^* := \Lambda_{s-i}, \quad \Psi_i^* := \Psi_{s+1-i},$$

and where the RK coefficients are defined by

$$c_i^* = 1 - c_{s+1-i}, \quad \widetilde{c}_i^* = 1 - \widetilde{c}_{s-i} \tag{6.23a}$$

$$b_i^* = b_{s+1-i}, \quad \widetilde{b}_i^* = \widetilde{b}_{s-i} \tag{6.23b}$$

$$a_{ij}^* = b_{s+1-j} - a_{s+1-i,s+1-j}, \quad \widetilde{a}_{ij}^* = \widetilde{b}_{s-j} - \widetilde{a}_{s+1-i,s-j},$$
$$\bar{a}_{ij}^* = b_{s+1-j} - \bar{a}_{s-i,s+1-j}, \quad \check{a}_{ij}^* = \widetilde{b}_{s-j} - \check{a}_{s-i,s-j}. \tag{6.23c}$$

The method given by (6.22) and (6.23) is the *adjoint* of (6.4) with Gauss-Lobatto coefficients. To show the symmetry of the method, we need the method given by (6.22) to be the same as the method (6.4) with Gauss-Lobatto coefficients, i.e. we must show

$$c_i^* = c_i, \quad \widetilde{c}_i^* = \widetilde{c}_i, \quad b_i^* = b_i, \quad \widetilde{b}_i^* = \widetilde{b}_i,$$
$$a_{ij}^* = a_{ij}, \quad \widetilde{a}_{ij}^* = \widetilde{a}_{ij}, \quad \bar{a}_{ij}^* = \bar{a}_{ij}, \quad \check{a}_{ij}^* = \check{a}_{ij}.$$

However, the Lobatto IIIA RK method is symmetric, so $\check{a}_{ij}^* = \check{a}_{ij}$. The remaining

coefficients have been shown to satisfy these conditions in Chapter 3. Therefore, the Gauss-Lobatto EMPRK methods for mixed index 2 and 3 problems are symmetric.

□

### 6.3 Existence, Uniqueness, and Influence of Perturbations

We give here a proof regarding the existence and uniqueness of a solution to (6.4). This proof is a combination of the existence proofs presented in [15] and [16]. The approach is similar to that of Chapter 5. For this section, consider $y_0$, $y_1$, $z_0$, $\lambda_0$, $\psi_0$ as functions of $h$. First, we give a lemma.

**Lemma 6.3.1.** *Assume the invertibility of the matrices (6.2), as well as the conditions (6.6) and (6.7). Then the matrix*

$$\begin{bmatrix} \begin{bmatrix} \bar{A}^* \widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y v_z r_\lambda(t,y,z,\lambda) & \begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} \otimes g_y v_z f_\psi(t,y,z,\psi) \\ \check{A}^* \otimes k_z r_\lambda(t,y,z,\lambda) & \bar{A}^* \otimes k_z f_\psi(t,y,z,\psi) \end{bmatrix} \tag{6.24}$$

*is invertible.*

*Proof.* This proof will be broken into two parts. First, we construct an invertible matrix $R$. Secondly, we compute the product of the matrices $R$ and (6.24), and show that this product is invertible. This will prove the invertibility of (6.24).

*First Step.* Let $N \in \mathbb{R}^{s\times s}$ be the matrix

$$N := \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & -(s-1) \end{bmatrix}.$$

The matrix $N$ is invertible, as it is a triangular matrix with no zeros on the diagonal.

Let $\widetilde{\Omega}_0$ be as in Chapter 2

$$\widetilde{\Omega}_0 := \begin{bmatrix} 0_s^T & 1 \\ b^T & 0 \\ b^T C & 0 \\ \vdots & \vdots \\ b^T C^{s-2} & 0 \end{bmatrix} \in \mathbb{R}^{s\times(s+1)},$$

and denote by $\Omega$ the matrix $\Omega := M^{-1}N\widetilde{\Omega}_0 \in \mathbb{R}^{s\times(s+1)}$. In [15] and in the proof of Lemma 2.2.2, the matrix $\Omega$ is shown to satisfy the property $\Omega\alpha = I_s$. Let $\gamma = [\widetilde{\gamma}^T \ \gamma_{s+1}]^T \in \mathbb{R}^{s+1}$, with $\widetilde{\gamma} \in \mathbb{R}^s$ and $\gamma_{s+1} \neq 0$, be a fixed vector defined by the orthogonality condition

$$\gamma^T \alpha = 0. \tag{6.25}$$

Because $\alpha \in \mathbb{R}^{(s+1)\times s}$, this is an underdetermined system of linear equations for $\gamma$. In fact, an infinite number of values exist for $\gamma$, because $\alpha = [A^T \ b]^T$, and $A$ is invertible. Now, let $Q \in \mathbb{R}^{(s+1)\times(s+1)}$ be given by

$$Q := \begin{bmatrix} A\Omega \\ \gamma^T \end{bmatrix}.$$

The matrix $Q$ is invertible, as shown in Chapter 5. Lastly, define $\check{Q} \in \mathbb{R}^{(s+1)\times(s+1)}$ by

$$\check{Q} := Q \begin{bmatrix} \bar{A^*}^{-1} & 0_s \\ 0_s^T & 1 \end{bmatrix},$$

which is invertible, as $\check{Q}$ is the product of two invertible matrices.

Combining all of these together, we can define the matrix $R$ by

$$R := \begin{bmatrix} \check{Q} \otimes I_{n_g} & O \\ O & A\bar{A^*}^{-1} \otimes I_{n_k} \end{bmatrix} \in \mathbb{R}^{((s+1)n_g+sn_k)\times((s+1)n_g+sn_k)}. \tag{6.26}$$

Because this is block diagonal with each block invertible, the matrix $R$ itself must be invertible.

*Second Step.* We now compute the product of $R$ and (6.24). This can be expressed as

$$\begin{bmatrix} \check{Q}\begin{bmatrix} \bar{A}^*\widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y v_z r_\lambda & \check{Q}\begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} \otimes g_y v_z f_\psi \\ A\bar{A}^{*^{-1}}\check{A}^* \otimes k_z r_\lambda & A \otimes k_z f_\psi \end{bmatrix}. \tag{6.27}$$

Each function above is evaluated at $(t, y, z, \lambda, \psi)$. The function arguments will be suppressed for the remainder of this proof. However, the matrices in the blocks can be expressed as

$$\check{Q}\begin{bmatrix} \bar{A}^*\widetilde{A} \\ \widetilde{b}^T \end{bmatrix} = \begin{bmatrix} A\Omega \\ \gamma^T \end{bmatrix}\widetilde{\alpha}, \tag{6.28a}$$

$$\check{Q}\begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} = \begin{bmatrix} A \\ 0_s^T \end{bmatrix}, \tag{6.28b}$$

$$A\bar{A}^{*^{-1}}\check{A}^* = A\Omega\widetilde{\alpha}. \tag{6.28c}$$

The simplifications (6.28a,b) are shown in Chapter 5. For (6.28c), it is enough to show that

$$\check{A}^*\ V = \bar{A}^*\ M^{-1}N\widetilde{\Omega}_0\ \widetilde{\alpha}\ V, \tag{6.29}$$

where $\widetilde{V}$ is the invertible Vandermonde matrix

$$\widetilde{V} := \begin{bmatrix} 1 & \widetilde{c}_0 & \dots & \widetilde{c}_0^s \\ 1 & \widetilde{c}_1 & \dots & \widetilde{c}_1^s \\ \vdots & \vdots & & \vdots \\ 1 & \widetilde{c}_s & \dots & \widetilde{c}_s^s \end{bmatrix} \mathbb{R}^{(s+1)\times(s+1)}.$$

The product $\check{A}^*\ \widetilde{V}$ is easily seen to be

$$(\check{A}^*\ \widetilde{V})_{ik} = \sum_{j=0}^{s} \check{a}_{ij}\widetilde{c}_j^{k-1} = \frac{\widetilde{c}_i^k}{k},$$

by using (6.18). Now, we focus on the right side of (6.29). The product $\widetilde{\alpha}\ \widetilde{V}$ is given

by

$$\widetilde{\alpha}\ \widetilde{V} = \left[\begin{array}{c|c} \frac{c_i^k}{k} & \sum_{j=0}^{s} \widetilde{a}_{ij}\widetilde{c}_j^s \\ \frac{1}{k} & \frac{1}{s+1} \end{array}\right]_{\substack{i=1,\ldots,s \\ k=1,\ldots,s}}$$

by using (6.16) and (6.12). Applying to this $\widetilde{\Omega}_0$ and using (6.11) gives

$$\widetilde{\Omega}_0\widetilde{\alpha}\ \widetilde{V} = \left[\begin{array}{cc|c} 1 & \frac{1}{2} & \cdots & \frac{1}{s+1} \\ \frac{1}{k}\cdot\frac{1}{(i+k)} & & & \sum_{j=1}^{s}\sum_{l=0}^{s} b_j\widetilde{a}_{jl}\widetilde{c}_l^s c_j^{i-1} \end{array}\right]_{\substack{i=1,\ldots,s-1 \\ k=1,\ldots,s}}.$$

Applying the matrix $N$ gives

$$N\ \widetilde{\Omega}_0\widetilde{\alpha}\ \widetilde{V} = \left[\begin{array}{cc|c} 1 & \frac{1}{2} & \cdots & \frac{1}{s+1} \\ & \frac{1}{(i+k)} & & \frac{1}{s+1} - i\sum_{j=1}^{s}\sum_{l=0}^{s} b_j\widetilde{a}_{jl}\widetilde{c}_l^s c_j^{i-1} \end{array}\right]_{\substack{i=1,\ldots,s-1 \\ k=1,\ldots,s}}.$$

But using (6.10), (6.12), (6.14), and (6.15), this can be reduced to

$$N\ \widetilde{\Omega}_0\widetilde{\alpha}\ \widetilde{V} = \left[\frac{1}{i+k}\right]_{\substack{i=0,\ldots,s-1 \\ k=1,\ldots,s+1}} = M\ V,$$

for the matrix $V$ equal to

$$V := \begin{bmatrix} 1 & c_1 & \ldots & c_1^s \\ 1 & c_2 & \ldots & c_2^s \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \ldots & c_s^s \end{bmatrix}.$$

Finally, applying $\bar{A}^* M^{-1}$ gives

$$\bar{A}^* M^{-1} N\ \widetilde{\Omega}_0\widetilde{\alpha}\ \widetilde{V} = \left[\frac{\widetilde{c}_i^k}{k}\right]_{\substack{i=1,\ldots,s \\ k=1,\ldots,s+1}}.$$

This shows the product (6.28c).

Applying the products (6.28) to (6.27) results in the matrix

$$\begin{bmatrix} \begin{bmatrix} A\Omega \\ \gamma^T \end{bmatrix} \widetilde{\alpha} \otimes g_y v_z r_\lambda & \begin{bmatrix} A \\ 0_s^T \end{bmatrix} \otimes g_y v_z f_\psi \\ A\Omega\widetilde{\alpha} \otimes k_z r_\lambda & A \otimes k_z f_\psi \end{bmatrix} = \\ \begin{bmatrix} I_{s+1} \otimes g_y v_z r_\lambda & \begin{bmatrix} I_s \\ 0_s^T \end{bmatrix} \otimes g_y v_z f_\psi \\ [I_s \ 0_s] \otimes k_z r_\lambda & I_s \otimes k_z f_\psi \end{bmatrix} \begin{bmatrix} Q\widetilde{\alpha} \otimes I_{n_g} & O \\ O & A \otimes I_{n_k} \end{bmatrix}. \tag{6.30}$$

But this matrix is the same as in the proof of Lemma 5.3.1. As shown previously, (6.30) is invertible. □

### 6.3.1 Existence and Uniqueness

Using Lemma 6.3.1, we now show that extended EMPRK methods applied to problems with mixed index 2 and index 3 constraints has a unique solution. This result follows from an application of the implicit function theorem.

**Theorem 6.3.2.** *Suppose that* $y_0 = y_0(h)$, $z_0 = z_0(h)$, $\lambda_0 = \lambda_0(h)$, $\psi_0 = \psi_0(h)$ *satisfy*

$$o(h^2) = g(t_0, y_0) \tag{6.31a}$$

$$o(h) = g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) \tag{6.31b}$$

$$\begin{aligned} o(1) = {} & g_{tt}(t_0, y_0) + 2g_{ty}(t_0, y_0)v(t_0, y_0, z_0) + g_y(t_0, y_0)v_t(t_0, y_0, z_0) \\ & + g_{yy}(t_0, y_0)\left(v(t_0, y_0, z_0), v(t_0, y_0, z_0)\right) \\ & + g_y(t_0, y_0)v_y(t_0, y_0, z_0)v(t_0, y_0, z_0) \\ & + g_y(t_0, y_0)v_z(t_0, y_0, z_0)\left[f(t_0, y_0, z_0, \psi_0) + r(t_0, y_0, \lambda_0)\right] \end{aligned} \tag{6.31c}$$

$$o(h) = k(t_0, y_0, z_0) \tag{6.31d}$$

$$\begin{aligned} o(1) = {} & k_t(t_0, y_0, z_0) + k_y(t_0, y_0, z_0)v(t_0, y_0, z_0) \\ & + k_z(t_0, y_0, z_0)\left[f(t_0, y_0, z_0, \psi_0) + r(t_0, y_0, \lambda_0)\right], \end{aligned} \tag{6.31e}$$

*where the matrices given in (6.2) are invertible. Then for* $|h| \leq h_0$, *there exists a*

*locally unique solution to (6.4) that satisfies*

$$
\begin{aligned}
Y_i - y_0 &= \mathcal{O}(h), \quad i = 1, \dots, s \\
Z_i - z_0 &= \mathcal{O}(h), \quad i = 1, \dots, s \\
\widetilde{Y}_i - y_0 &= \mathcal{O}(h), \quad i = 0, \dots, s \\
\widetilde{Z}_i - z_0 &= \mathcal{O}(h), \quad i = 0, \dots, s \\
\Lambda_i - \lambda_0 &= \mathcal{O}(h), \quad i = 0, \dots, s \\
\Psi_i - \psi_0 &= \mathcal{O}(h), \quad i = 1, \dots, s \\
y_1 - y_0 &= \mathcal{O}(h), \\
z_1 - z_0 &= \mathcal{O}(h).
\end{aligned}
$$

*Proof.* We begin by reformulating the system (6.4) as

$$0 = Y_i - y_0 - h \sum_{j=1}^{s} a_{ij} V_j, \quad i = 1, \dots, s \tag{6.32a}$$

$$0 = Z_i - z_0 - h \sum_{j=1}^{s} a_{ij} F_j - h \sum_{j=0}^{s} \widetilde{a}_{ij} R_j, \quad i = 1, \dots, s \tag{6.32b}$$

$$0 = \widetilde{Y}_i - y_0 - h \sum_{j=1}^{s} \bar{a}_{ij} V_j, \quad i = 0, \dots, s \tag{6.32c}$$

$$0 = \widetilde{Z}_i - z_0 - h \sum_{j=1}^{s} \bar{a}_{ij} F_j - h \sum_{j=0}^{s} \check{a}_{ij} R_j, \quad i = 1, \dots, s \tag{6.32d}$$

$$0 = y_1 - y_0 - h \sum_{j=1}^{s} b_j V_j \tag{6.32e}$$

$$0 = z_1 - z_0 - h \sum_{j=1}^{s} b_j F_j - h \sum_{j=0}^{s} \widetilde{b}_j R_j \tag{6.32f}$$

$$0 = \frac{1}{h^2} g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i), \quad i = 0, \dots, s \tag{6.32g}$$

$$0 = \frac{1}{h^2} g(t_1, y_1) \tag{6.32h}$$

$$0 = \frac{1}{h} g_t(t_1, y_1) + \frac{1}{h} g_y(t_1, y_1) v(t_1, y_1, z_1) \tag{6.32i}$$

$$0 = \frac{1}{h} k(t_0 + \widetilde{c}_i h, \widetilde{Y}_i, \widetilde{Z}_i), \quad i = 1, \dots, s \tag{6.32j}$$

$$0 = \frac{1}{h} k(t_1, y_1). \tag{6.32k}$$

The proof of this theorem can be done by application of the implicit function theorem. We first expand the constraints $0 = g(t, y)$, and then expand the constraints $0 = k(t, y, z)$.

We have $\widetilde{Y}_0 = y_0$, hence $g(t_0, \widetilde{Y}_0) = 0$ is automatically satisfied by assumption. We have $\widetilde{Y}_s = y_1$ since $\widetilde{c}_s = 1$, hence the equation (6.32h) can be removed since it is equivalent to (6.32g) for $i = s$. Similarly, (6.32k) and (6.32j) for $i = 0$ can be removed. To keep the following calculations clean, we introduce the notation

$$\begin{aligned}
&Y_i(\tau) := y_0 + \tau(Y_i - y_0), && Z_i(\tau) := z_0 + \tau(Z_i - z_0),\\
&\widetilde{Y}_i(\tau) := y_0 + \tau(\widetilde{Y}_i - y_0), && \widetilde{Z}_i(\tau) := z_0 + \tau(\widetilde{Z}_i - z_0),\\
&y_1(\tau) := y_0 + \tau(y_1 - y_0), && z_1(\tau) := z_0 + \tau(z_1 - z_0),\\
&T_i(\tau) := t_0 + \tau c_i h, && \widetilde{T}_i(\tau) := t_0 + \tau \widetilde{c}_i h,\\
&t_1(\tau) := t_0 + \tau h.
\end{aligned}$$

We now expand $g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)$ for $i = 1, \ldots, s$ and $v(t_0 + c_i h, Y_i, Z_i)$ for $i = 1, \ldots, s$ into a Taylor series around $(t_0, y_0, z_0)$, resulting in

$$\begin{aligned}
g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) = g(t_0, y_0) &+ g_t(t_0, y_0)\widetilde{c}_i h + g_y(t_0, y_0)(\widetilde{Y}_i - y_0)\\
&+ \int_0^1 (1-\tau) g_{tt}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau)) d\tau \cdot \widetilde{c}_i^2 h^2\\
&+ 2\int_0^1 (1-\tau) g_{ty}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau)) d\tau \cdot \widetilde{c}_i h(\widetilde{Y}_i - y_0)\\
&+ \int_0^1 (1-\tau) g_{yy}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau)) d\tau (\widetilde{Y}_i - y_0, \widetilde{Y}_i - y_0),
\end{aligned}$$

$$\begin{aligned}
V_i = v(t_0 + c_i h, Y_i, Z_i) = v(t_0, y_0, z_0) &+ \int_0^1 v_t(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot c_i h\\
&+ \int_0^1 v_y(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot (Y_i - y_0)\\
&+ \int_0^1 v_z(T_i(\tau), Y_i(\tau), Z_i(\tau)) d\tau \cdot (Z_i - z_0)
\end{aligned}$$

$$
\begin{aligned}
&= v(t_0, y_0, z_0) + c_i h \int_0^1 v_t(T_i(\tau), Y_i(\tau), Z_i(\tau))d\tau \\
&\quad + h \int_0^1 v_y(T_i(\tau), Y_i(\tau), Z_i(\tau))d\tau \cdot \sum_{j=1}^{s} a_{ij} V_j \\
&\quad + h \int_0^1 v_z(T_i(\tau), Y_i(\tau), Z_i(\tau))d\tau \cdot \Big( \sum_{j=1}^{s} a_{ij} F_j + \sum_{j=0}^{s} \widetilde{a}_{ij} R_j \Big).
\end{aligned}
$$

Dividing $g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)$ by $h^2$, replacing the terms $Y_i - y_0$, $Z_i - z_0$, and $\widetilde{Y}_i - y_0$ by using (6.32a,b,c), and utilizing the relation above for $V_i = v(t_0 + c_i h, Y_i, Z_i)$, we obtain

$$
\begin{aligned}
&\frac{1}{h^2} g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) = \frac{1}{h^2} g(t_0, y_0) + \frac{1}{h} g_t(t_0, y_0)\widetilde{c}_i + \frac{1}{h} \sum_{j=1}^{s} \bar{a}_{ij} g_y(t_0, y_0) V_j \\
&\quad + \widetilde{c}_i^2 \int_0^1 (1-\tau) g_{tt}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau))d\tau + 2\widetilde{c}_i \sum_{j=1}^{s} \bar{a}_{ij} \int_0^1 (1-\tau) g_{ty}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau))d\tau \cdot V_j \\
&\quad + \sum_{j=1}^{s} \sum_{k=1}^{s} \bar{a}_{ij} \bar{a}_{ik} \int_0^1 (1-\tau) g_{yy}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau))d\tau (V_j, V_k) \\
&= \frac{1}{h^2} g(t_0, y_0) + \frac{1}{h} g_t(t_0, y_0)\widetilde{c}_i + \frac{1}{h} \sum_{j=1}^{s} \bar{a}_{ij} g_y(t_0, y_0) v(t_0, y_0, z_0) \\
&\quad + \widetilde{c}_i^2 \int_0^1 (1-\tau) g_{tt}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau))d\tau \\
&\quad + 2\widetilde{c}_i \sum_{j=1}^{s} \bar{a}_{ij} \int_0^1 (1-\tau) g_{ty}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau))d\tau \cdot V_j \\
&\quad + \sum_{j=1}^{s} \bar{a}_{ij} c_j g_y(t_0, y_0) \int_0^1 v_t(T_j(\tau), Y_j(\tau), Z_j(\tau))d\tau \\
&\quad + \sum_{j=1}^{s} \sum_{k=1}^{s} \bar{a}_{ij} a_{jk} g_y(t_0, y_0) \int_0^1 v_y(T_j(\tau), Y_j(\tau), Z_j(\tau))d\tau \cdot V_k \\
&\quad + \sum_{j=1}^{s} \sum_{k=1}^{s} \bar{a}_{ij} a_{jk} g_y(t_0, y_0) \int_0^1 v_z(T_j(\tau), Y_j(\tau), Z_j(\tau))d\tau \cdot F_k \\
&\quad + \sum_{j=1}^{s} \sum_{k=0}^{s} \bar{a}_{ij} \widetilde{a}_{jk} g_y(t_0, y_0) \int_0^1 v_z(T_j(\tau), Y_j(\tau), Z_j(\tau))d\tau \cdot R_k \\
&\quad + \sum_{j=1}^{s} \sum_{k=1}^{s} \bar{a}_{ij} \bar{a}_{ik} \int_0^1 (1-\tau) g_{yy}(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau))d\tau (V_j, V_k).
\end{aligned}
\tag{6.33}
$$

By (6.6e), for the values $Y_i := y_0$, $Z_i := z_0$, $\widetilde{Y}_i := y_0$ $\Lambda_i := \lambda_0$, and $\Psi_i := \psi_0$ we obtain, for $h$ sufficiently small,

$$\begin{aligned}
&\frac{1}{h^2} g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) = \frac{1}{h^2} g(t_0, y_0) + \frac{\widetilde{c}_i}{h} \big( g_t(t_0, y_0) + g_y(t_0, y_0) v(t_0, y_0, z_0) \big) \\
&+ \frac{\widetilde{c}_i^2}{2} \big( g_{tt}(t_0, y_0) + 2 g_{ty}(t_0, y_0) v(t_0, y_0, z_0) + g_y(t_0, y_0) v_t(t_0, y_0, z_0) \\
&\quad + g_y(t_0, y_0) v_y(t_0, y_0, z_0) v(t_0, y_0, z_0) + g_y(t_0, y_0) v_z(t_0, y_0, z_0) f(t_0, y_0, z_0, \psi_0) \\
&\quad + g_y(t_0, y_0) v_z(t_0, y_0, z_0) r(t_0, y_0, \lambda_0) + g_{yy}(t_0, y_0)(v(t_0, y_0, z_0), v(t_0, y_0, z_0)) \big) + \mathcal{O}(h) \\
&= o(1),
\end{aligned}$$

since (6.31) holds. Hence the values $Y_i(0) := y_0(0)$, $Z_i(0) := z_0(0)$, $\widetilde{Y}_i(0) := y_0(0)$, $\Lambda_i(0) := \lambda_0(0)$, and $\Psi_i(0) := \psi_0(0)$ satisfy (6.32a,b,c) and the constraints

$$0 = \frac{1}{h^2} g(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) = \frac{1}{h^2} g(t_0, y_0) + \frac{1}{h} g_t(t_0, y_0) \widetilde{c}_i + \dots \ . \tag{6.34}$$

The additional terms are those from (6.33). Similarly, we have

$$\begin{aligned}
g_y(t_1, y_1) &= g_y(t_0, y_0) + h \int_0^1 g_{ty}(t_1(\tau), y_1(\tau)) d\tau \\
&\qquad + \int_0^1 g_{yy}(t_1(\tau), y_1(\tau)) d\tau (y_1 - y_0, \cdot) \\
&= g_y(t_0, y_0) + h \int_0^1 g_{ty}(t_1(\tau), y_1(\tau)) d\tau \\
&\qquad + h \sum_{j=1}^{s} b_j \int_0^1 g_{yy}(t_1(\tau), y_1(\tau)) d\tau (V_j, \cdot), \\
v(t_1, y_1, z_1) &= v(t_0, y_0, z_0) + h \int_0^1 v_t(t_1(\tau), y_1(\tau), z_1(\tau)) d\tau \\
&\qquad + \int_0^1 v_y(t_1(\tau), y_1(\tau), z_1(\tau)) d\tau \cdot (y_1 - y_0) \\
&\qquad + \int_0^1 v_z(t_1, y_1(\tau), z_1(\tau)) d\tau \cdot (z_1 - z_0) \\
&= v(t_0, y_0, z_0) + h \int_0^1 v_t(t_1(\tau), y_1(\tau), z_1(\tau)) d\tau \\
&\qquad + h \int_0^1 v_y(t_1(\tau), y_1(\tau), z_1(\tau)) d\tau \cdot \sum_{j=1}^{s} b_j V_j
\end{aligned}$$

$$+ h \int_0^1 v_z(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \cdot \left( \sum_{j=1}^{s} b_j F_j + \sum_{j=0}^{s} \widetilde{b}_j R_j \right),$$

$$\begin{aligned} g_t(t_1, y_1) &= g_t(t_0, y_0) + h \int_0^1 g_{tt}(t_1(\tau), y_1(\tau))d\tau \\ &\qquad + \int_0^1 g_{ty}(t_1(\tau), y_1(\tau))d\tau \cdot (y_1 - y_0) \\ &= g_t(t_0, y_0) + h \int_0^1 g_{tt}(t_1(\tau), y_1(\tau))d\tau \\ &\qquad + h \int_0^1 g_{ty}(t_1(\tau), y_1(\tau))d\tau \cdot \sum_{j=1}^{s} b_j V_j. \end{aligned}$$

Hence, dividing $g_t(t_1, y_1) + g_y(t_1, y_1)v(t_1, y_1, z_1)$ by $h$, we obtain

$$\begin{aligned} \frac{1}{h} g_t(t_1, y_1) + \frac{1}{h} g_y(t_1, y_1)v(t_1, y_1, z_1) = & \qquad (6.35) \\ & \frac{1}{h} g_t(t_0, y_0) + \frac{1}{h} g_y(t_0, y_0)v(t_0, y_0, z_0) \\ & + \int_0^1 g_{tt}(t_1(\tau), y_1(\tau))d\tau + \int_0^1 g_{ty}(t_1(\tau), y_1(\tau))d\tau \cdot v(t_1, y_1, z_1) \\ & + \sum_{j=1}^{s} b_j \int_0^1 g_{ty}(t_1(\tau), y_1(\tau))d\tau \cdot V_j \\ & + g_y(t_0, y_0) \int_0^1 v_t(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \\ & + \sum_{j=1}^{s} b_j g_y(t_0, y_0) \int_0^1 v_y(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \cdot V_j \\ & + \sum_{j=1}^{s} b_j g_y(t_0, y_0) \int_0^1 v_z(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \cdot F_j \\ & + \sum_{j=0}^{s} \widetilde{b}_j g_y(t_0, y_0) \int_0^1 v_z(t_1(\tau), y_1(\tau), z_1(\tau))d\tau \cdot R_j \\ & + \sum_{j=1}^{s} b_j \int_0^1 g_{yy}(t_1(\tau), y_1(\tau))d\tau(V_j, v(t_1, y_1, z_1)). \end{aligned}$$

Because of assumption (6.7a), for the values $Y_i := y_0$, $y_1 := y_0$, $Z_i := z_0$, $z_1 := z_0$, $\widetilde{Y}_i := y_0$, $\Lambda_i := \lambda_0$, and $\Psi_i := \psi_0$, we use (6.31) to obtain

$$\begin{aligned} \frac{1}{h} g_t(t_1, y_1) + \frac{1}{h} g_y(t_1, y_1)v(t_1, y_1, z_1) &= \frac{1}{h} \left( g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) \right) \\ + g_{tt}(t_0, y_0) + g_y(t_0, y_0)v_t(t_0, y_0, z_0) &+ 2g_{ty}(t_0, y_0)v(t_0, y_0, z_0) \end{aligned}$$

$$+ g_y(t_0,y_0)v_y(t_0,y_0,z_0)v(t_0,y_0,z_0) + g_y(t_0,y_0)v_z(t_0,y_0,z_0)f(t_0,y_0,z_0,\psi_0)$$
$$+ g_y(t_0,y_0)v_z(t_0,y_0,z_0)r(t_0,y_0,\lambda_0) + g_{yy}(t_0,y_0)(v(t_0,y_0,z_0),v(t_0,y_0,z_0)) + \mathcal{O}(h)$$
$$= o(1).$$

Hence the values $Y_i(0) := y_0(0)$, $y_1(0) := y_0(0)$, $Z_i(0) := z_0(0)$, $z_1(0) := z_0(0)$, $\widetilde{Y}_i(0) := y_0(0)$, $\Lambda_i(0) := \lambda_0(0)$, and $\Psi_i(0) := \psi_0(0)$ satisfy

$$\begin{aligned} 0 &= \frac{1}{h}g_t(t_1,y_1) + \frac{1}{h}g_y(t_1,y_1)v(t_1,y_1,z_1) \\ &= \frac{1}{h}g_t(t_0,y_0) + \frac{1}{h}g_y(t_0,y_0)v(t_0,y_0,z_0) + \dots, \end{aligned} \tag{6.36}$$

with the additional terms coming from (6.35).

Next, we expand the constraints $0 = k(t,y,z)$. Writing

$$\begin{aligned} k(t_0+\widetilde{c}_ih,\widetilde{Y}_i,\widetilde{Z}_i) = k(t_0,y_0,z_0) &+ \widetilde{c}_ih\int_0^1 k_t(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau),\widetilde{Z}_i(\tau))d\tau \\ &+ \int_0^1 k_y(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau),\widetilde{Z}_i(\tau))d\tau\cdot(\widetilde{Y}_i - y_0) \\ &+ \int_0^1 k_z(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau),\widetilde{Z}_i(\tau))d\tau\cdot(\widetilde{Z}_i - z_0), \end{aligned}$$

we substitute into this (6.32c,d), and get

$$\begin{aligned} \frac{1}{h}k(t_0+\widetilde{c}_ih,\widetilde{Y}_i,\widetilde{Z}_i) &= \frac{1}{h}k(t_0,y_0,z_0) + \widetilde{c}_i\int_0^1 k_t(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau),\widetilde{Z}_i(\tau))d\tau \\ &+ \sum_{j=1}^s \bar{a}_{ij}\int_0^1 k_y(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau),\widetilde{Z}_i(\tau))d\tau\cdot v(T_j,Y_j,Z_j) \\ &+ \sum_{j=1}^s \bar{a}_{ij}\int_0^1 k_z(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau),\widetilde{Z}_i(\tau))d\tau\cdot f(T_j,Y_j,Z_j,\Psi_j) \\ &+ \sum_{j=0}^s \check{a}_{ij}\int_0^1 k_z(\widetilde{T}_i(\tau),\widetilde{Y}_i(\tau),\widetilde{Z}_i(\tau))d\tau\cdot r(\widetilde{T}_j,\widetilde{Y}_j,\Lambda_j) \end{aligned} \tag{6.37}$$

for $i = 1,\dots,s$. By (6.31), for the values $Y_i(0) = y_0$, $Z_i(0) = z_0$, $\widetilde{Y}_i(0) = y_0$, $\widetilde{Z}_i(0) = z_0$, $\Psi_i(0) = \psi_0$, for $i = 1,\dots,s$, and $\Lambda_i(0) = \lambda_0$ for $i = 0,\dots,s$, and

$y_1(0) = y_0$, $z_1(0) = z_0$, we obtain, for $h \to 0$,

$$\begin{aligned}\frac{1}{h}k(t_0 + \widetilde{c}_i h, \widetilde{Y}_i, \widetilde{Z}_i) &= \frac{1}{h}k(t_0, y_0, z_0) + c_j\big[k_t(t_0, y_0, z_0) + k_y(t_0, y_0, z_0)v(t_0, y_0, z_0) \\ &\quad + k_z(t_0, y_0, z_0)f(t_0, y_0, z_0, \psi_0) + k_z(t_0, y_0, z_0)r(t_0, y_0, \lambda_0)\big] + \mathcal{O}(h) \\ &= o(1).\end{aligned}$$

Hence, the values $Y_i(0) = y_0$, $Z_i(0) = z_0$, $\widetilde{Y}_i(0) = y_0$, $\widetilde{Z}_i(0) = z_0$, $\Psi_i(0) = \psi_0$, for $i = 1, \ldots, s$, and $\Lambda_i(0) = \lambda_0$ for $i = 0, \ldots, s$, and $y_1(0) = y_0$, $z_1(0) = z_0$ satisfy (6.32a,b,c,d) and the constraints

$$0 = \frac{1}{h}k(t_0, y_0, z_0) + \widetilde{c}_i \int_0^1 k_t(\widetilde{T}_i(\tau), \widetilde{Y}_i(\tau), \widetilde{Z}_i(\tau))d\tau + \ldots, \tag{6.38}$$

with the additional terms coming from (6.37).

Putting everything together, and using tensor matrix product notation, the Jacobian of equations (6.32a,e), (6.32b,f), (6.32c,d), (6.34), (6.36), and (6.38) with respect to $Y_i$ $(i = 1, \ldots, s)$, $y_1$, $Z_i$ $(i = 1, \ldots, s)$, $z_1$, $\widetilde{Y}_i$ $(i = 1, \ldots, s)$, $\widetilde{Z}_i$ $(i = 1, \ldots, s)$, $\Lambda_i$ $(i = 0, 1, \ldots, s)$, and $\Psi_i$ $(i = 1, \ldots, s)$ with $h = 0$, is of the form

$$\begin{bmatrix} I_{(s+1)n_y} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{(s+1)n_z} & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{sn_y} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{sn_z} & 0 & 0 \\ \mathcal{O}(1) & \mathcal{O}(1) & \mathcal{O}(1) & 0 & \chi(t_0, y_0, z_0, \lambda_0) & \gamma(t_0, y_0, z_0, \psi_0) \\ 0 & 0 & \mathcal{O}(1) & \mathcal{O}(1) & \xi(t_0, y_0, z_0, \lambda_0) & \rho(t_0, y_0, z_0, \psi_0) \end{bmatrix}$$

with the matrix

$$\begin{bmatrix} \chi(t_0, y_0, z_0, \lambda_0) & \gamma(t_0, y_0, z_0, \psi_0) \\ \xi(t_0, y_0, z_0, \lambda_0) & \rho(t_0, y_0, z_0, \psi_0) \end{bmatrix}$$

given by (6.24). This Jacobian matrix is invertible, as a result of Lemma 6.3.1. Therefore, if $|h| \leq h_0$, the implicit function theorem yields the existence of a locally

unique solution to (6.32abcd)-(6.34)-(6.36)-(6.38), and hence to the corresponding SPARK method (6.32). □

### 6.3.2 Influence of Perturbations

We now consider the influence of perturbations on the solution of the method (6.4). We consider the perturbed system

$$\widehat{Y}_i = \widehat{y}_0 + h\sum_{j=1}^{s} a_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\delta_i^y, \quad i = 1, \ldots, s \tag{6.39a}$$

$$\begin{aligned}\widehat{Z}_i = \widehat{z}_0 &+ h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) \\ &+ h\sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_i^z, \quad i = 1, \ldots, s\end{aligned} \tag{6.39b}$$

$$\widehat{\widetilde{Y}}_i = \widehat{y}_0 + h\sum_{j=1}^{s} \bar{a}_{ij} v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\widetilde{\delta}_i^y, \quad i = 0, \ldots, s \tag{6.39c}$$

$$\begin{aligned}\widehat{\widetilde{Z}}_i = \widehat{z}_0 &+ h\sum_{j=1}^{s} \bar{a}_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) \\ &+ h\sum_{j=0}^{s} \check{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\widetilde{\delta}_i^z, \quad i = 0, \ldots, s\end{aligned} \tag{6.39d}$$

$$\widehat{y}_1 = \widehat{y}_0 + h\sum_{j=1}^{s} b_j v(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j) + h\delta_{s+1}^y \tag{6.39e}$$

$$\begin{aligned}\widehat{z}_1 = \widehat{z}_0 &+ h\sum_{j=1}^{s} b_j f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) \\ &+ h\sum_{j=0}^{s} \widetilde{b}_j r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_{s+1}^z\end{aligned} \tag{6.39f}$$

$$0 = g(t_0 + \widetilde{c}_i h, \widehat{\widetilde{Y}}_i) + h\delta_i^\lambda, \quad i = 0, \ldots, s \tag{6.39g}$$

$$0 = g_t(t_1, \widehat{y}_1) + g_y(t_1, \widehat{y}_1) v(t_1, \widehat{y}_1, \widehat{z}_1) + \delta_{s+1}^\lambda \tag{6.39h}$$

$$0 = k(t_0 + c_i h, \widehat{\widetilde{Y}}_i, \widehat{\widetilde{Z}}_i) + \delta_i^\psi, \quad i = 0, \ldots, s. \tag{6.39i}$$

Consider also the perturbed form of (6.8)

$$\widehat{Z}_i^f = \widehat{z}_0 + h\sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) + h\delta_i^f, \quad i = 1, \ldots, s \tag{6.40a}$$

$$\widehat{Z}_i^r = h\sum_{j=0}^{s} \widetilde{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\delta_i^r, \quad i = 1, \ldots, s \tag{6.40b}$$

$$\widehat{\widetilde{Z}}_i^f = \widehat{z}_0 + h\sum_{j=1}^{s} \bar{a}_{ij} f(t_0 + c_j h, \widehat{Y}_j, \widehat{Z}_j, \widehat{\Psi}_j) + h\widetilde{\delta}_i^f, \quad i = 0, \ldots, s \tag{6.40c}$$

$$\widehat{\widetilde{Z}}_i^r = h\sum_{j=0}^{s} \check{a}_{ij} r(t_0 + \widetilde{c}_j h, \widehat{\widetilde{Y}}_j, \widehat{\Lambda}_j) + h\widetilde{\delta}_i^r, \quad i = 0, \ldots, s. \tag{6.40d}$$

We examine the influence of the perturbations

$$\begin{aligned}
&\delta^y := [\delta_1^{yT}, \ldots, \delta_{s+1}^{y\ T}]^T, \quad \delta^z := [\delta_1^{zT}, \ldots, \delta_{s+1}^{z\ T}]^T,\\
&\widetilde{\delta}^y := [\widetilde{\delta}_0^{yT}, \ldots, \widetilde{\delta}_s^{y\ T}]^T, \quad \widetilde{\delta}^z := [\widetilde{\delta}_0^{zT}, \ldots, \widetilde{\delta}_s^{z\ T}]^T,\\
&\delta^\lambda := [\delta_0^{\lambda T}, \ldots, \delta_{s+1}^{\lambda\ T}]^T, \quad \delta^\psi := [\delta_1^{\psi T}, \ldots, \delta_s^{\psi T}]^T,\\
&\delta^f := [\delta_1^{f\ T}, \ldots, \delta_s^{f\ T}]^T, \quad \delta^r := [\delta_1^{rT}, \ldots, \delta_s^{rT}]^T,\\
&\widetilde{\delta}^f := [\widetilde{\delta}_0^{f\ T}, \ldots, \widetilde{\delta}_s^{f\ T}]^T, \quad \widetilde{\delta}^r := [\widetilde{\delta}_0^{r\ T}, \ldots, \widetilde{\delta}_s^{r\ T}]^T.
\end{aligned}$$

For simplicity, we introduce the notations

$$\begin{aligned}
Y &:= [Y_1^T, Y_2^T, \ldots, Y_s^T]^T, \quad \widehat{Y} := [\widehat{Y}_1^T, \widehat{Y}_2^T, \ldots, \widehat{Y}_s^T]^T\\
Z &:= [Z_1^T, Z_2^T, \ldots, Z_s^T]^T, \quad \widehat{Z} := [\widehat{Z}_1^T, \widehat{Z}_2^T, \ldots, \widehat{Z}_s^T]^T\\
\widetilde{Y} &:= [\widetilde{Y}_0^T, \widetilde{Y}_1^T, \ldots, \widetilde{Y}_s^T]^T, \quad \widehat{\widetilde{Y}} := [\widehat{\widetilde{Y}}_0^T, \widehat{\widetilde{Y}}_1^T, \ldots, \widehat{\widetilde{Y}}_s^T]^T\\
\widetilde{Z} &:= [\widetilde{Z}_0^T, \widetilde{Z}_1^T, \ldots, \widetilde{Z}_s^T]^T, \quad \widehat{\widetilde{Z}} := [\widehat{\widetilde{Z}}_0^T, \widehat{\widetilde{Z}}_1^T, \ldots, \widehat{\widetilde{Z}}_s^T]^T\\
\Lambda &:= [\Lambda_0^T, \Lambda_1^T, \ldots, \Lambda_s^T]^T, \quad \widehat{\Lambda} := [\widehat{\Lambda}_0^T, \widehat{\Lambda}_1^T, \ldots, \widehat{\Lambda}_s^T]^T\\
\Psi &:= [\Psi_1^T, \Psi_2^T, \ldots, \Psi_s^T]^T, \quad \widehat{\Psi} := [\widehat{\Psi}_1^T, \widehat{\Psi}_2^T, \ldots, \widehat{\Psi}_s^T]^T\\
\Delta Y_i &:= \widehat{Y}_i - Y_i, \quad \Delta Z_i := \widehat{Z}_i - Z_i, \quad \Delta\widetilde{Y}_i := \widehat{\widetilde{Y}}_i - \widetilde{Y}_i\\
\Delta\widetilde{Z}_i &:= \widehat{\widetilde{Z}}_i - \widetilde{Z}_i, \quad \Delta\Lambda_i := \widehat{\Lambda}_i - \Lambda_i, \quad \Delta\Psi_i := \widehat{\Psi}_i - \Psi_i\\
\Delta y_1 &:= \widehat{y}_1 - y_1, \quad \Delta z_1 := \widehat{z}_1 - z_1, \quad \Delta y_0 := \widehat{y}_0 - y_0, \quad \Delta z_0 := \widehat{z}_0 - z_0
\end{aligned}$$

$$\Delta Y := \widehat{Y} - Y, \quad \Delta Z := \widehat{Z} - Z, \quad \Delta \widetilde{Y} := \widehat{\widetilde{Y}} - \widetilde{Y}$$
$$\Delta \widetilde{Z} := \widehat{\widetilde{Z}} - \widetilde{Z}, \quad \Delta \Lambda := \widehat{\Lambda} - \Lambda, \quad \Delta \Psi := \widehat{\Psi} - \Psi$$
$$\Delta \overline{Y} := [\widehat{Y}^T - Y^T, \widehat{y}_1^T - y_1^T]^T, \quad \Delta \overline{Z} := [\widehat{Z}^T - Z^T, \widehat{z}_1^T - z_1^T]^T$$
$$\Delta \underline{\widetilde{Y}} := [\widehat{\widetilde{Y}}_1^{\,T} - \widetilde{Y}_1^T, \ldots, \widehat{\widetilde{Y}}_s^{\,T} - \widetilde{Y}_s^T]^T, \quad \Delta \underline{\widetilde{Z}} := [\widehat{\widetilde{Z}}_1^T - \widetilde{Z}_1^T, \ldots, \widehat{\widetilde{Z}}_s^T - \widetilde{Z}_s^T]^T,$$

We also define $||Y|| := \max_i\{||Y_i||\}$, $||\Lambda|| := \max_i\{||\Lambda_i||\}$, etc. We will make use of the coefficient matrices

$$\alpha := \begin{bmatrix} A \\ b^T \end{bmatrix}, \quad \widetilde{\alpha} := \begin{bmatrix} \widetilde{A} \\ \widetilde{b}^T \end{bmatrix}, \quad \bar{\alpha} := \begin{bmatrix} \bar{A}^* & 0_s \\ 0_s^T & 1 \end{bmatrix},$$

where $\bar{A}^*$, $\check{A}^* \in \mathbb{R}^{s\times s}$ equal $\bar{A}$ and $\check{A}$, respectively, with the first row removed.

**Theorem 6.3.3.** *Suppose the initial conditions satisfy (6.31). Further, assume that the matrices in (6.2) are invertible. Lastly, we assume*

$$\begin{aligned} &\Delta y_0 = \mathcal{O}(h^3), \quad \Delta z_0 = \mathcal{O}(h^2), \\ &\Lambda_k - \lambda_0 = \mathcal{O}(h), \quad \Psi_j - \psi_0 = \mathcal{O}(h), \\ &\delta_i^y = \mathcal{O}(h), \quad \widetilde{\delta}_k^y = \mathcal{O}(h^2), \quad \widetilde{\delta}_k^z = \mathcal{O}(h^2), \quad \delta_i^z = \mathcal{O}(h), \\ &\delta_l^\lambda = \mathcal{O}(h^2), \quad \delta_j^\psi = \mathcal{O}(h^2), \quad \delta_j^f = \mathcal{O}(1), \quad \delta_j^r = \mathcal{O}(1), \end{aligned} \tag{6.41}$$

*for $i = 1, \ldots, s+1$, $j = 1, \ldots, s$, $k = 0, \ldots, s$, and $l = 0, \ldots, s+1$. Then for $|h| \le h_0$, we have the bounds*

$$\begin{aligned} \Delta Y_i = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ &+ h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big) \end{aligned} \tag{6.42a}$$

$$\begin{aligned} \Delta Z_i = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ &+ h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big) \end{aligned} \tag{6.42b}$$

$$\begin{aligned} \Delta \widetilde{Y}_i = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ &+ h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0, y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big) \end{aligned} \tag{6.42c}$$

$$\begin{aligned} \Delta \widetilde{Z}_i = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ &+ h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big) \end{aligned} \tag{6.42d}$$

$$\begin{aligned}\Delta y_1 = \Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ &+ h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)\end{aligned} \tag{6.42e}$$

$$\begin{aligned}\Delta z_1 = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ &+ h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.42f}$$

$$\begin{aligned}h\Delta\Lambda_i = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.42g}$$

$$\begin{aligned}h\Delta\Psi_i = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).\end{aligned} \tag{6.42h}$$

*Further, we have the bounds*

$$\begin{aligned}\Delta Z_i^f = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^f|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.43a}$$

$$\begin{aligned}\Delta Z_i^r = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^r|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.43b}$$

$$\begin{aligned}\Delta \widetilde{Z}_i^f = \Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + h||\widetilde{\delta}^f|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.43c}$$

$$\begin{aligned}\Delta \widetilde{Z}_i^r = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + h||\widetilde{\delta}^r|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.43d}$$

*where we have used the notations*

$$\begin{aligned}\kappa_0 &:= k_y(t_0,y_0,z_0)\Delta y_0 + k_z(t_0,y_0,z_0)\Delta z_0 \\ \eta_0 &:= g_y(t_0,y_0)v_y(t_0,y_0,z_0)\Delta y_0 + g_y(t_0,y_0)v_z(t_0,y_0,z_0)\Delta z_0 \\ &\qquad + g_{ty}(t_0,y_0)\Delta y_0 + g_{yy}(t_0,y_0)(\Delta y_0, v(t_0,y_0,z_0)).\end{aligned}$$

*Proof.* The proof given here uses ideas presented in [12] and [22], and is similar to that for SPARK methods applied to mixed index 2 and 3 problems. We begin by showing that (6.42c,d) hold for $\Delta\widetilde{Y}_0$ and $\Delta\widetilde{Z}_0$. These come immediately from

(6.39c,d) and (6.4c,d), as

$$\Delta\widetilde{Y}_0 = \Delta y_0 + h\widetilde{\delta}_0^y = \mathcal{O}(||\Delta y_0|| + h||\widetilde{\delta}^y||)$$
$$\Delta\widetilde{Z}_0 = \Delta z_0 + h\widetilde{\delta}_0^z = \mathcal{O}(||\Delta z_0|| + h||\widetilde{\delta}^z||).$$

Subtracting (6.4) from (6.39), and expanding around $Y_j$, $Z_j$, $\widetilde{Y}_j$, $\widetilde{Z}_j$, $\Lambda_j$, and $\Psi_j$ gives

$$\begin{aligned}
\Delta Y_i &= \Delta y_0 + h\sum_{j=1}^{s} a_{ij}(v_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j \\
&\quad + v_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) + h\delta_i^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)
\end{aligned} \tag{6.44a}$$

$$\begin{aligned}
\Delta Z_i &= \Delta z_0 + h\sum_{j=1}^{s} a_{ij}(f_y(t_0+c_jh, Y_j, Z_j, \Psi_j)\Delta Y_j \\
&\quad + f_z(t_0+c_jh, Y_j, Z_j, \Psi_j)\Delta Z_j + f_\psi(t_0+c_jh, Y_j, Z_j, \Psi_j)\Delta\Psi_j) \\
&\quad + h\sum_{j=0}^{s} \widetilde{a}_{ij}(r_y(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j) + h\delta_i^z \\
&\quad + \mathcal{O}(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2)
\end{aligned} \tag{6.44b}$$

$$\begin{aligned}
\Delta\widetilde{Y}_i &= \Delta y_0 + h\sum_{j=1}^{s} \bar{a}_{ij}(v_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j \\
&\quad + v_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) + h\widetilde{\delta}_i^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)
\end{aligned} \tag{6.44c}$$

$$\begin{aligned}
\Delta\widetilde{Z}_i &= \Delta z_0 + h\sum_{j=1}^{s} \bar{a}_{ij}(f_y(t_0+c_jh, Y_j, Z_j, \Psi_j)\Delta Y_j \\
&\quad + f_z(t_0+c_jh, Y_j, Z_j, \Psi_j)\Delta Z_j + f_\psi(t_0+c_jh, Y_j, Z_j, \Psi_j)\Delta\Psi_j) \\
&\quad + h\sum_{j=0}^{s} \check{a}_{ij}(r_y(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0+\widetilde{c}_jh, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j) + h\widetilde{\delta}_i^z \\
&\quad + \mathcal{O}(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2)
\end{aligned} \tag{6.44d}$$

$$\begin{aligned}
\Delta y_1 &= \Delta y_0 + h\sum_{j=1}^{s} b_j(v_y(t_0+c_jh, Y_j, Z_j)\Delta Y_j \\
&\quad + v_z(t_0+c_jh, Y_j, Z_j)\Delta Z_j) + h\delta_{s+1}^y + \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2\right)
\end{aligned} \tag{6.44e}$$

$$
\begin{aligned}
\Delta z_1 = \Delta z_0 + h\sum_{j=1}^{s} b_j (f_y(t_0 + c_j h, Y_j, Z_j)\Delta Y_j \\
+ f_z(t_0 + c_j h, Y_j, Z_j)\Delta Z_j + f_\psi(t_0 + c_j h, Y_j, Z_j, \Psi_j)\Delta\Psi_j) \\
+ h\sum_{j=0}^{s} \widetilde{b}_j (r_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j) + h\delta^z_{s+1} \\
+ \mathcal{O}(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2)
\end{aligned} \tag{6.44f}
$$

$$
\begin{aligned}
0 = \frac{1}{h} g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\Delta y_0 \\
+ g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\sum_{j=0}^{s} \bar{a}_{ij} v_y(t_0 + c_j h, Y_j, Z_j)\Delta Y_j \\
+ g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\sum_{j=0}^{s} \bar{a}_{ij} v_z(t_0 + c_j h, Y_j, Z_j)\Delta Z_j \\
+ \mathcal{O}\left(h||\Delta y_0|| + ||\Delta Y||^2 + ||\Delta Z||^2 + ||\widetilde{\delta}^y|| + ||\delta^\lambda||\right)
\end{aligned} \tag{6.44g}
$$

$$
\begin{aligned}
0 = g_{ty}(t_1, y_1)\Delta y_1 + g_y(t_1, y_1)v_y(t_1, y_1, z_1)\Delta y_1 \\
+ g_y(t_1, y_1)v_z(t_1, y_1, z_1)\Delta z_1 + g_{yy}(t_1, y_1)(\Delta y_1, v(t_1, y_1, z_1)) \\
+ \delta^\lambda_{s+1} + \mathcal{O}\left(||\Delta y_1||^2 + ||\Delta z_1||^2\right)
\end{aligned} \tag{6.44h}
$$

$$
\begin{aligned}
0 = k_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \widetilde{Z}_j)\Delta\widetilde{Y}_j + k_z(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \widetilde{Z}_j)\Delta\widetilde{Z}_j + \delta^\psi_i \\
+ \mathcal{O}\left(||\Delta\widetilde{Y}||^2 + ||\Delta\widetilde{Z}||^2\right)
\end{aligned} \tag{6.44i}
$$

Note that in (6.44g) we have divided both sides by $h$. This system can be written more compactly using tensor notation. We rewrite (6.44a-f,i) as

$$
\begin{aligned}
\Delta\overline{Y} = \mathbb{1}_{s+1} \otimes \Delta y_0 + h(\alpha \otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z) \\
+ \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\delta^y||\right)
\end{aligned} \tag{6.45a}
$$

$$
\begin{aligned}
\Delta\overline{Z} = \mathbb{1}_{s+1} \otimes \Delta z_0 + h(\widehat{\alpha} \otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z + \{f_\psi\}\Delta\Psi) \\
+ h(\widetilde{\alpha} \otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda) \\
+ \mathcal{O}\big(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2 \\
+ h||\Delta\Lambda||^2 + h||\Delta\Psi||^2 + h||\delta^z||\big)
\end{aligned} \tag{6.45b}
$$

$$
\begin{aligned}
\Delta\widetilde{Y} = \mathbb{1}_{s+1} \otimes \Delta y_0 + h(\bar{A}^* \otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z) \\
+ \mathcal{O}\left(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\widetilde{\delta}^y||\right)
\end{aligned} \tag{6.45c}
$$

$$\begin{aligned}\Delta\widetilde{Z} &= \mathbb{1}_{s+1}\otimes\Delta z_0 + h(\bar{A}^*\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z + \{f_\psi\}\Delta\Psi)\\ &\quad + h(\check{A}^*\otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda)\\ &\quad + \mathcal{O}\big(h||\Delta Y||^2 + h||\Delta\widetilde{Y}||^2 + h||\Delta Z||^2\\ &\qquad\qquad + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2 + h||\widetilde{\delta^z}||\big)\end{aligned} \tag{6.45d}$$

$$0 = \langle k_y\rangle\Delta\underline{\widetilde{Y}} + \langle k_z\rangle\Delta\underline{\widetilde{Z}} + \mathcal{O}\left(||\Delta\underline{\widetilde{Y}}||^2 + ||\Delta\underline{\widetilde{Z}}||^2 + ||\delta^\psi||\right). \tag{6.45e}$$

We will use the notation

$$\begin{aligned}\{v_y\} &:= \text{blockdiag}(v_y(t_0+c_1h, Y_1, Z_1), \ldots, v_y(t_0+c_sh, Y_s, Z_s))\\ \{g_y\} &:= \text{blockdiag}(g_y(t_0+\widetilde{c}_1h, \widetilde{Y}_1), \ldots, g_y(t_0+\widetilde{c}_sh, \widetilde{Y}_s))\\ [r_y] &:= \text{blockdiag}(r_y(t_0+\widetilde{c}_0h, \widetilde{Y}_0, \Lambda_0), \ldots, r_y(t_0+\widetilde{c}_sh, \widetilde{Y}_s, \Lambda_s))\\ \langle k_y\rangle &:= \text{blockdiag}(k_y(t_0+\widetilde{c}_1h, \widetilde{Y}_1, \widetilde{Z}_1), \ldots, k_y(t_0+\widetilde{c}_sh, \widetilde{Y}_s, \widetilde{Z}_s))\\ \widetilde{[g_y]} &:= \text{blockdiag}(g_y(t_0+\widetilde{c}_1h, \widetilde{Y}_1), \ldots, g_y(t_0+\widetilde{c}_sh, \widetilde{Y}_s), g_y(t_1, y_1))\end{aligned}$$

and so on. We will now solve for the terms $\Delta\Lambda$ and $\Delta\Psi$. First, we address the constraints $0 = k(t, y, z)$. Substituting (6.45c,d) into (6.45e) gives

$$\begin{aligned}0 &= \langle k_y\rangle\left[\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\bar{A}^*\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right]\\ &\quad + \langle k_z\rangle\big[\mathbb{1}_{s+1}\otimes\Delta z_0 + h(\bar{A}^*\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z + \{f_\psi\}\Delta\Psi)\\ &\qquad + h(\check{A}^*\otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda)\big]\\ &\quad + \mathcal{O}\Big(||\Delta Y||^2 + ||\Delta Z||^2 + ||\Delta\widetilde{Y}||^2 + ||\Delta\widetilde{Z}||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2\\ &\qquad + h||\widetilde{\delta^y}|| + h||\widetilde{\delta^z}|| + ||\delta^\psi||\Big).\end{aligned}$$

Rewriting with $\Delta\Lambda$ and $\Delta\Psi$ more isolated, we get

$$\begin{aligned}
&-h\langle k_z\rangle(\check{A}^*\otimes I_{n_z})[r_\lambda]\Delta\Lambda - h\langle k_z\rangle(\bar{A}^*\otimes I_{n_z})\{f_\psi\}\Delta\Psi = \\
&\qquad \langle k_y\rangle\left(\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\bar{A}^*\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right) \\
&\qquad + \langle k_z\rangle\left(\mathbb{1}_{s+1}\otimes\Delta z_0 + h(\bar{A}^*\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z)\right) \\
&\qquad + h\langle k_z\rangle(\check{A}^*\otimes I_{n_z})[r_y]\Delta\widetilde{Y} \\
&\qquad + \mathcal{O}\big(||\Delta Y||^2 + ||\Delta Z||^2 + ||\Delta\widetilde{Y}||^2 + ||\Delta\widetilde{Z}||^2 \\
&\qquad\qquad + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2 + h||\widetilde{\delta^y}|| + h||\widetilde{\delta^z}|| + ||\delta^\psi||\big).
\end{aligned} \tag{6.46}$$

Next, we address the constraints $0 = g(t, y)$. Combining (6.44g) with (6.44h), and rewriting in tensor notation gives

$$\begin{aligned}
0 = {} & \widetilde{[g_y]}(\bar{\alpha}\otimes I_n)\left(\langle v_y\rangle\Delta\overline{Y} + \langle v_z\rangle\Delta\overline{Z}\right) \\
& + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_s\otimes\Delta y_0) \\ g_{ty}(t_1,y_1)\Delta y_1 + g_{yy}(t_1,y_1)(\Delta y_1, v(t_1,y_1,z_1)) \end{bmatrix} \\
& + \mathcal{O}\left(h||\Delta y_0|| + ||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + ||\widetilde{\delta^y}|| + ||\delta^\lambda||\right).
\end{aligned} \tag{6.47}$$

Substituting (6.45a,b) into this, we arrive at

$$\begin{aligned}
0 = {} & \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_y\rangle\left(\mathbb{1}_{s+1}\otimes\Delta y_0 + h(\alpha\otimes I_{n_y})(\{v_y\}\Delta Y + \{v_z\}\Delta Z)\right) \\
& + \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_z})\langle v_z\rangle\left(\mathbb{1}_{s+1}\otimes\Delta z_0 + h(\alpha\otimes I_{n_z})(\{f_y\}\Delta Y + \{f_z\}\Delta Z + \{f_\psi\}\Delta\Psi)\right) \\
& + \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_z})\langle v_z\rangle\left(h(\widetilde{\alpha}\otimes I_{n_z})([r_y]\Delta\widetilde{Y} + [r_\lambda]\Delta\Lambda)\right) \\
& + \begin{bmatrix} \frac{1}{h}\{g_y\}(\mathbb{1}_s\otimes\Delta y_0) \\ g_{ty}(t_1,y_1)\Delta y_1 + g_{yy}(t_1,y_1)(\Delta y_1, v(t_1,y_1,z_1)) \end{bmatrix} \\
& + \mathcal{O}\big(h||\Delta y_0|| + ||\Delta\overline{Y}||^2 + ||\Delta\overline{Z}||^2 + h||\Delta\Lambda||^2 + h||\Delta\Psi||^2 \\
& \qquad + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta^y}|| + ||\delta^\lambda||\big).
\end{aligned}$$

Solving this for $\Delta\Lambda$ and $\Delta\Psi$ gives

$$
\begin{aligned}
&-\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda](h\Delta\Lambda)\\
&\quad-\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\alpha\otimes I_{n_z})\{f_\psi\}(h\Delta\Psi)=\\
&\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_y\rangle\left(\mathbb{1}_{s+1}\otimes\Delta y_0+h(\alpha\otimes I_{n_y})(\{v_y\}\Delta Y+\{v_z\}\Delta Z)\right)\\
&+\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle\left(\mathbb{1}_{s+1}\otimes\Delta z_0+h(\alpha\otimes I_{n_z})(\{f_y\}\Delta Y+\{f_z\}\Delta Z)\right)\\
&+h\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_y]\Delta\widetilde{Y}\\
&+\begin{bmatrix}\frac{1}{h}\{g_y\}(\mathbb{1}_s\otimes\Delta y_0)\\ g_{ty}(t_1,y_1)\Delta y_1+g_{yy}(t_1,y_1)(\Delta y_1,v(t_1,y_1,z_1))\end{bmatrix}\\
&+\mathcal{O}\big(h||\Delta y_0||+||\Delta\overline{Y}||^2+||\Delta\overline{Z}||^2+h||\Delta\Lambda||^2+h||\Delta\Psi||^2\\
&\qquad\qquad+h||\delta^y||+h||\delta^z||+||\widetilde{\delta}^y||+||\delta^\lambda||\big).
\end{aligned}
\tag{6.48}
$$

Bringing together (6.48) and (6.46), we get

$$
\begin{aligned}
&-\begin{bmatrix}\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda] & \widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\alpha\otimes I_{n_z})\{f_\psi\}\\ \langle k_z\rangle(\check{A}^*\otimes I_{n_z})[r_\lambda] & \langle k_z\rangle(\bar{A}^*\otimes I_{n_z})\{f_\psi\}\end{bmatrix}\cdot\\
&\qquad\qquad\begin{bmatrix}h\Delta\Lambda\\ h\Delta\Psi\end{bmatrix}=\begin{bmatrix}M_1\\ M_2\end{bmatrix},
\end{aligned}
\tag{6.49}
$$

where $M_1$ is the right-hand side of (6.48) and $M_2$ is the right-hand side of (6.46). With $i=1,\dots,s$, and $j=0,\dots,s$, the first block of the coefficient matrix is

$$
\begin{aligned}
&\widetilde{[g_y]}(\bar{\alpha}\otimes I_{n_y})\langle v_z\rangle(\widetilde{\alpha}\otimes I_{n_z})[r_\lambda]\\
&=\begin{bmatrix}\left[\sum_{k=1}^s\bar{a}_{ik}\widetilde{a}_{kj}g_y(t_0+\widetilde{c}_ih,\widetilde{Y}_i)v_z(t_0+c_kh,Y_k,Z_k)r_\lambda(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\right]\\ \widetilde{b}_jg_y(t_1,y_1)v_z(t_1,y_1,z_1)r_\lambda(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\end{bmatrix}\\
&=\begin{bmatrix}\bar{A}^*\widetilde{A}\\ \widetilde{b}^T\end{bmatrix}\otimes g_y(t_0,y_0)v_z(t_0,y_0,z_0)r_\lambda(t_0,y_0,\lambda_0)+\mathcal{O}(h).
\end{aligned}
$$

With $i=1,\dots,s$ and $j=0,\dots,s$, the lower left block becomes

$$
\langle k_z\rangle(\check{A}^*\otimes I_{n_z})[r_\lambda]=\left[\check{a}_{ij}k_z(t_0+c_ih,Y_i,Z_i)r_\lambda(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\right]
$$

$$= [\check{a}_{ij} k_z(t_0, y_0, z_0) r_\lambda(t_0, y_0, \lambda_0)] + \mathcal{O}(h)$$
$$= \check{A}^* \otimes k_z(t_0, y_0, z_0) r_\lambda(t_0, y_0, \lambda_0) + \mathcal{O}(h).$$

With $i = 1, \ldots, s$ and $j = 1, \ldots, s$, the upper right block becomes

$$\widetilde{[g_y]}(\bar{\alpha} \otimes I_{n_y})\langle v_z \rangle(\alpha \otimes I_{n_z})\{f_\psi\} =$$
$$\begin{bmatrix} \sum_{k=1}^{s} \bar{a}_{ik} a_{kj} g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i) v_z(t_0 + c_k h, Y_k, Z_k) f_\psi(t_0 + c_j h, Y_j, Z_j, \Psi_j) \\ b_j g_y(t_1, y_1) v_z(t_1, y_1, z_1) f_\psi(t_0 + c_j h, Y_j, Z_j, \Psi_j) \end{bmatrix}$$
$$= \begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} \otimes g_y(t_0, y_0) v_z(t_0, y_0, z_0) f_\psi(t_0, y_0, z_0, \psi_0) + \mathcal{O}(h).$$

With $i = 1, \ldots, s$ and $j = 1, \ldots, s$, the last block in the lower right becomes

$$\langle k_z \rangle(\bar{A}^* \otimes I_{n_z})\{f_\psi\} = [\bar{a}_{ij} k_z(t_0 + c_i h, Y_i, Z_i) f_\psi(t_0 + c_j h, Y_j, Z_j, \Psi_j)]$$
$$= \left[\bar{a}_{ij} k_z(t_0, y_0, z_0) f_\psi(t_0, y_0, z_0, \psi_0)\right] + \mathcal{O}(h)$$
$$= \bar{A}^* \otimes k_z(t_0, y_0, z_0) f_\psi(t_0, y_0, z_0, \psi_0) + \mathcal{O}(h).$$

Therefore, as $h \to 0$, the coefficient matrix in (6.49) becomes

$$\begin{bmatrix} \begin{bmatrix} \bar{A}^* \widetilde{A} \\ \widetilde{b}^T \end{bmatrix} \otimes g_y v_z r_\lambda(t_0, y_0, z_0, \lambda_0) & \begin{bmatrix} \bar{A}^* A \\ b^T \end{bmatrix} \otimes g_y v_z f_\psi(t_0, y_0, z_0, \psi_0) \\ \check{A}^* \otimes k_z r_\lambda(t_0, y_0, z_0, \lambda_0) & \bar{A}^* \otimes k_z f_\psi(t_0, y_0, z_0, \psi_0) \end{bmatrix}. \quad (6.50)$$

Thus, for $h$ sufficiently small, the coefficient matrix in (6.49) is invertible. In (6.48), we express several terms as

$$\frac{1}{h} g_y(t_0 + \widetilde{c}_i h, \widetilde{Y}_i)\Delta y_0 = \frac{1}{h} g_y(t_0 + (\widetilde{c}_i h), y_0 + (\widetilde{Y}_i - y_0))\Delta y_0$$
$$= \frac{1}{h} g_y(t_0, y_0)\Delta y_0 + \widetilde{c}_i g_{ty}(t_0, y_0)\Delta y_0$$
$$+ \frac{1}{h} g_{yy}(t_0, y_0)(\Delta y_0, \widetilde{Y}_i - y_0) + \mathcal{O}(h||\Delta y_0||)$$
$$= \frac{1}{h} g_y(t_0, y_0)\Delta y_0 + \widetilde{c}_i g_{ty}(t_0, y_0)\Delta y_0$$
$$+ \widetilde{c}_i g_{yy}(t_0, y_0)(\Delta y_0, v(t_0, y_0, z_0)) + \mathcal{O}(h||\Delta y_0||)$$

$$g_y(t_0+\widetilde{c}_i h,\widetilde{Y}_i)\sum_{j=1}^{s}\bar{a}_{ij}v_y(t_0+c_j h,Y_j,Z_j)\Delta y_0 = \widetilde{c}_i g_y(t_0,y_0)v_y(t_0,y_0,z_0)\Delta y_0 + \mathcal{O}(h||\Delta y_0||),$$

$$g_y(t_0+\widetilde{c}_i h,\widetilde{Y}_i)\sum_{j=1}^{s}\bar{a}_{ij}v_z(t_0+c_j h,Y_j,Z_j)\Delta z_0 = \widetilde{c}_i g_y(t_0,y_0)v_z(t_0,y_0,z_0)\Delta z_0 + \mathcal{O}(h||\Delta z_0||),$$

$$\Delta y_1 = \Delta y_0 + h\sum_{j=1}^{s} b_j(v(t_0+c_j h,\widehat{Y}_j,\widehat{Z}_j) - v(t_0+c_j h,Y_j,Z_j)) + h\delta^y_{s+1} = \Delta y_0 + \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||),$$

$$g_{ty}(t_1,y_1)\Delta y_1 = g_{ty}(t_0,y_0)\Delta y_0 + \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||),$$

$$g_{yy}(t_1,y_1)(\Delta y_1, v(t_1,y_1,z_1)) = g_{yy}(t_0,y_0)(\Delta y_0, v(t_0,y_0,z_0)) + \mathcal{O}(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||).$$

In addition, (6.46) contains terms that may be written as

$$k_y(t_0+\widetilde{c}_i h,\widetilde{Y}_i,\widetilde{Z}_i)\Delta y_0 = k_y(t_0,y_0,z_0)\Delta y_0 + \mathcal{O}(h||\Delta y_0||)$$

$$k_z(t_0+\widetilde{c}_i h,\widetilde{Y}_i,\widetilde{Z}_i)\Delta z_0 = k_z(t_0,y_0,z_0)\Delta z_0 + \mathcal{O}(h||\Delta z_0||).$$

We therefore get that both $h\Delta\Lambda$ and $h\Delta\Psi$ can be expressed by

$$\begin{aligned}\mathcal{O}(&h||\Delta y_0|| + h||\Delta z_0|| + h||\Delta Y|| + h||\Delta Z|| + h||\Delta\widetilde{Y}|| + h||\Delta\widetilde{Z}|| \\ &+ h||\Delta\Lambda||^2 + h||\Delta\Psi||^2 + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||).\end{aligned} \tag{6.51}$$

The equations (6.44a-f) result in

$$\Delta\overline{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\left(h||\Delta Y|| + h||\Delta Z|| + h||\delta^y||\right)$$

$$\Delta\overline{Z} = \mathbb{1}_{s+1}\otimes\Delta z_0 + \mathcal{O}\big(h||\Delta Y|| + h||\Delta Z|| + h||\Delta\widetilde{Y}|| + h||\Delta\Lambda|| + h||\Delta\Psi|| + h||\delta^z||\big)$$

$$\Delta\widetilde{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\left(h||\Delta Y|| + h||\Delta Z|| + h||\widetilde{\delta}^y||\right)$$

$$\Delta\widetilde{Z} = \mathbb{1}_{s+1}\otimes\Delta z_0 + \mathcal{O}\big(h||\Delta Y|| + h||\Delta Z|| + h||\Delta\widetilde{Y}||$$

$$+ h||\Delta\Lambda|| + h||\Delta\Psi|| + h||\widetilde{\delta}^z||\big).$$

Reinserting these equations for $\Delta Y$ and $\Delta Z$ into each other and using (6.51) gives

$$\begin{aligned}\Delta\overline{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ &+ h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)\end{aligned} \tag{6.52a}$$

$$\begin{aligned}\Delta\overline{Z} = \mathbb{1}_{s+1}\otimes\Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ &+ h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.52b}$$

$$\begin{aligned}\Delta\widetilde{Y} = \mathbb{1}_{s+1}\otimes\Delta y_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| \\ &+ h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi|| + ||g_y(t_0,y_0)\Delta y_0|| + h||\kappa_0|| + h||\eta_0||\big)\end{aligned} \tag{6.52c}$$

$$\begin{aligned}\Delta\widetilde{Z} = \mathbb{1}_{s+1}\otimes\Delta z_0 + \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| \\ &+ h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).\end{aligned} \tag{6.52d}$$

In addition, inserting (6.52) into (6.51) gives

$$\begin{aligned}h||\Delta\Lambda|| = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)\end{aligned} \tag{6.53a}$$

$$\begin{aligned}h||\Delta\Psi|| = \mathcal{O}\big(&h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| \\ &+ ||\delta^\lambda|| + ||\delta^\psi|| + \frac{1}{h}||g_y(t_0,y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).\end{aligned} \tag{6.53b}$$

The results (6.52) and (6.53) show (6.42).

Subtracting (6.8) from (6.40) and linearizing gives

$$\begin{aligned}\Delta Z_i^f = \Delta z_0 + h\sum_{j=1}^{s} a_{ij}\Big(&f_y(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta Y_j \\ &+ f_z(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta Z_j + f_\psi(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta\Psi_j\Big) \\ &+ h\delta_i^f + \mathcal{O}(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\Delta\Psi||^2)\end{aligned}$$

$$\begin{aligned}\Delta Z_i^r = h\sum_{j=0}^{s} \widetilde{a}_{ij}(&r_y(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0+\widetilde{c}_jh,\widetilde{Y}_j,\Lambda_j)\Delta\Lambda_j) \\ &+ h\delta_i^r + \mathcal{O}(h||\Delta\widetilde{Y}||^2 + h||\Delta\Lambda||^2)\end{aligned}$$

$$\begin{aligned}\Delta\widetilde{Z}_i^f = \Delta z_0 + h\sum_{j=1}^{s} \bar{a}_{ij}\Big(&f_y(t_0+c_jh,Y_j,Z_j,\Psi_j)\Delta Y_j\end{aligned}$$

$$+ f_z(t_0 + c_j h, Y_j, Z_j, \Psi_j)\Delta Z_j + f_\psi(t_0 + c_j h, Y_j, Z_j, \Psi_j)\Delta\Psi_j\Big)$$
$$+ h\widetilde{\delta}_i^f + \mathcal{O}(h||\Delta Y||^2 + h||\Delta Z||^2 + h||\Delta\Psi||^2)$$
$$\Delta\widetilde{Z}_i^r = h\sum_{j=0}^{s} \check{a}_{ij}(r_y(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\widetilde{Y}_j + r_\lambda(t_0 + \widetilde{c}_j h, \widetilde{Y}_j, \Lambda_j)\Delta\Lambda_j)$$
$$+ h\widetilde{\delta}_i^r + \mathcal{O}(h||\Delta\widetilde{Y}||^2 + h||\Delta\Lambda||^2).$$

Substituting in the expressions for $\Delta Y$, $\Delta Z$, $\Delta\widetilde{Y}$, $\Delta\widetilde{Z}$, $\Delta\Lambda$, and $\Delta\Psi$, we arrive at

$$\Delta Z^f = \mathbb{1}_{s+1} \otimes \Delta z_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y||$$
$$+ h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^f|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)$$
$$\Delta Z^r = \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda||$$
$$+ ||\delta^\psi|| + h||\delta^r|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)$$
$$\Delta\widetilde{Z}^f = \mathbb{1}_{s+1} \otimes \Delta z_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y||$$
$$+ h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\widetilde{\delta}^f|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big)$$
$$\Delta\widetilde{Z}^r = \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda||$$
$$+ ||\delta^\psi|| + h||\widetilde{\delta}^r|| + \frac{1}{h}||g_y(t_0, y_0)\Delta y_0|| + ||\kappa_0|| + ||\eta_0||\big).$$

This completes the proof. □

Each of the constants in the result of Theorem 6.3.3 depend only upon the derivatives of the functions $v$, $f$, $r$, $g$, and $k$, not upon any of the constants from the hypothesis. With some additional assumptions, the bounds of this perturbation theorem can be simplified.

**Corollary 6.3.4.** *If, in addition to the conditions of Theorem 6.3.3, we assume that*

$$g(t_0, y_0) = 0 = g(t_0, \widehat{y}_0)$$
$$g_t(t_0, y_0) + g_y(t_0, y_0)v(t_0, y_0, z_0) = 0 = g_t(t_0, \widehat{y}_0) + g_y(t_0, \widehat{y}_0)v(t_0, \widehat{y}_0, \widehat{z}_0)$$
$$k(t_0, y_0, z_0) = 0 = k(t_0, \widehat{y}_0, \widehat{z}_0),$$

*then we have the bounds*

$$\Delta Y_i = \Delta y_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| + h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi||\big) \tag{6.54a}$$

$$\Delta Z_i = \Delta z_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big) \tag{6.54b}$$

$$\Delta \widetilde{Y}_i = \Delta y_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h^2||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| + h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi||\big) \tag{6.54c}$$

$$\Delta \widetilde{Z}_i = \Delta z_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big) \tag{6.54d}$$

$$\Delta y_1 = \Delta y_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h^2||\delta^z|| + h||\widetilde{\delta}^y|| + h^2||\widetilde{\delta}^z|| + h||\delta^\lambda|| + h||\delta^\psi||\big) \tag{6.54e}$$

$$\Delta z_1 = \Delta z_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big) \tag{6.54f}$$

$$h\Delta\Lambda_i = \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big) \tag{6.54g}$$

$$h\Delta\Psi_i = \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi||\big). \tag{6.54h}$$

$$\Delta Z_i^f = \Delta z_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^f||\big) \tag{6.54i}$$

$$\Delta Z_i^r = \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\delta^r||\big) \tag{6.54j}$$

$$\Delta \widetilde{Z}_i^f = \Delta z_0 + \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\widetilde{\delta}^f||\big) \tag{6.54k}$$

$$\Delta \widetilde{Z}_i^r = \mathcal{O}\big(h||\Delta y_0|| + h||\Delta z_0|| + h||\delta^y|| + h||\delta^z|| + ||\widetilde{\delta}^y|| + h||\widetilde{\delta}^z|| + ||\delta^\lambda|| + ||\delta^\psi|| + h||\widetilde{\delta}^r||\big) \tag{6.54l}$$

*Proof.* With these stronger assumptions on the constraints, we subtract and linearize, giving

$$\begin{aligned}
0 &= g(t_0, \widehat{y}_0) - g(t_0, y_0) = g_y(t_0, y_0)\Delta y_0 + \mathcal{O}(||\Delta y_0||^2) \\
0 &= k(t_0, \widehat{y}_0, \widehat{z}_0) - k(t_0, y_0, z_0) = \kappa_0 + \mathcal{O}(||\Delta y_0||^2 + ||\Delta z_0||^2) \\
0 &= g_t(t_0, \widehat{y}_0) + g_y(t_0, \widehat{y}_0)v(t_0, \widehat{y}_0, \widehat{z}_0) - g_t(t_0, y_0) - g_y(t_0, y_0)v(t_0, y_0, z_0) \\
&= \eta_0 + \mathcal{O}(||\Delta y_0||^2 + ||\Delta z_0||^2),
\end{aligned}$$

with $\kappa_0$ and $\eta_0$ as defined in the statement of Theorem 6.3.3. But this means

$$\begin{aligned}
g_y(t_0, y_0)\Delta y_0 &= \mathcal{O}(h^3||\Delta y_0||) \\
\kappa_0 &= \mathcal{O}(h^3||\Delta y_0|| + h^2||\Delta z_0||) \\
\eta_0 &= \mathcal{O}(h^3||\Delta y_0|| + h^2||\Delta z_0||),
\end{aligned}$$

as $\Delta y_0 = \mathcal{O}(h^3)$, $\Delta z_0 = \mathcal{O}(h^2)$. The conclusion (6.54) now follows by applying these bounds to the results of Theorem 6.3.3. □

## 6.4 Discontinuous Collocation Type Methods

We present here discontinuous collocation type methods for solving (6.1) with mixed index 2 and 3 constraints. Similar results can be found in Chapter 5.

**Definition 6.4.1.** *Let $c_1, \ldots, c_s$ be distinct real numbers, and $\widetilde{c}_0, \ldots, \widetilde{c}_s$ also be distinct real numbers, with $\widetilde{c}_0 = 0$ and $\widetilde{c}_s = 1$. Assume also that $\widetilde{b}_0$ and $\widetilde{b}_s$ are positive real numbers. We then define the s-degree polynomials $Y(t)$, $\Lambda(t)$, $\Psi(t)$, and $Z^f(t)$ and the $(s+1)$-degree polynomials $Z(t)$ and $Z^r(t)$ and the $(s+2)$-degree*

*polynomials* $\widetilde{Z}(t)$ *and* $\widetilde{Z}^r(t)$ *as the polynomials satisfying the initial conditions*

$$\begin{aligned} &Y(t_0) = y_0, \\ &Z^f(t_0) = z_0, \quad Z^r(t_0) = -h\widetilde{b}_0\widetilde{\mu}(t_0), \\ &\widetilde{Z}^r(t_0) = 0, \\ &Z(t_0) = Z^f(t_0) + Z^r(t_0) = z_0 - h\widetilde{b}_0\widetilde{\mu}(t_0), \\ &\widetilde{Z}(t_0) = Z^f(t_0) + \widetilde{Z}^r(t_0) = z_0, \end{aligned} \tag{6.55}$$

*where*

$$\widetilde{\mu}(t) := \dot{Z}^r(t) - r(t, Y(t), \Lambda(t)),$$

*as well as the conditions*

$$\dot{Y}(t_0 + c_i h) = v(t_0 + c_i h, Y(t_0 + c_i h), Z(t_0 + c_i h)), \quad i = 1, \ldots, s \tag{6.56a}$$

$$\dot{Z}^f(t_0 + c_i h) = f(t_0 + c_i h, Y(t_0 + c_i h), Z(t_0 + c_i h), \Psi(t_0 + c_i h)), \quad i = 1, \ldots, s \tag{6.56b}$$

$$\dot{Z}^r(t_0 + \widetilde{c}_i h) = r(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h), \Lambda(t_0 + \widetilde{c}_i h)), \quad i = 1, \ldots, s-1 \tag{6.56c}$$

$$Z(t) = Z^f(t) + Z^r(t) \tag{6.56d}$$

$$\dot{\widetilde{Z}}^r(t_0 + \widetilde{c}_i h) = r(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h), \Lambda(t_0 + \widetilde{c}_i h)), \quad i = 0, \ldots, s \tag{6.56e}$$

$$\widetilde{Z}(t) = Z^f(t) + \widetilde{Z}^r(t) \tag{6.56f}$$

$$0 = g(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h)), \quad i = 0, \ldots, s \tag{6.56g}$$

$$0 = g_t(t_1, Y(t_1)) + g_y(t_1, Y(t_1))v(t_1, Y(t_1), Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)) \tag{6.56h}$$

$$0 = k(t_0 + \widetilde{c}_i h, Y(t_0 + \widetilde{c}_i h), \widetilde{Z}(t_0 + \widetilde{c}_i h)), \quad i = 0, \ldots, s. \tag{6.56i}$$

*The polynomials* $Y(t)$, $Z(t)$, $\widetilde{Z}(t)$, $\Lambda(t)$, $\Psi(t)$, $Z^f(t)$, $Z^r(t)$, $\widetilde{Z}^r(t)$ *are referred to as* discontinuous collocation type polynomials. *The values of* $Y(t_1)$ *and* $Z(t_1) - h\widetilde{b}_s\widetilde{\mu}(t_1)$ *are used as approximations to the exact solutions* $y(t)$ *and* $z(t)$*, respectively, of (6.1) at time* $t_1 := t_0 + h$.

The discontinuous collocation type methods presented here are quite similar to the those presented for the SPARK methods in Chapter 5. We shall therefore

omit much of the detail, and refer the reader to the previous proofs. The interesting point to note is the addition of the polynomial $\widetilde{Z}(t)$. Unlike $Z(t)$, this polynomial can be treated as a (continuous) collocation polynomial.

**Theorem 6.4.2.** *The discontinuous collocation type polynomials $Y(t)$, $Z(t)$, $\Lambda(t)$, and $\Psi(t)$ defined by (6.56) are equivalent to an $(s,s)$-stage EMPRK method for mixed index 2 and 3 problems. Given $\widetilde{b}_0$ and $\widetilde{b}_s$, the remaining coefficients are determined by*

$$a_{ij} = \int_0^{c_i} \ell_j(\tau)d\tau, \quad b_j = \int_0^1 \ell_j(\tau)d\tau, \quad i,j = 1,\ldots,s, \tag{6.57a}$$

$$\bar{a}_{ij} = \int_0^{\widetilde{c}_i} \ell_j(\tau)d\tau, \quad i = 0,\ldots,s,\ j = 1,\ldots,s, \tag{6.57b}$$

$$\check{a}_{ij} = \int_0^{\widetilde{c}_i} \widetilde{\ell}_j(\tau)d\tau, \quad i = 0,\ldots,s,\ j = 0,\ldots,s, \tag{6.57c}$$

$$\left.\begin{aligned} &\widetilde{a}_{ij} = \textstyle\int_0^{c_i} \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0), \quad i = 1,\ldots,s,\ j = 0,\ldots,s,\\ &\widetilde{a}_{i0} = \widetilde{b}_0, \quad \widetilde{a}_{is} = 0, \qquad\qquad i = 1,\ldots,s,\\ &\widetilde{b}_j = \textstyle\int_0^1 \widehat{\ell}_j(\tau)d\tau - \widetilde{b}_0\widehat{\ell}_j(\widetilde{c}_0) - \widetilde{b}_s\widehat{\ell}_j(\widetilde{c}_s), \quad j = 1,\ldots,s-1, \end{aligned}\right\} \tag{6.57d}$$

*where the functions $\ell_j(\tau)$ and $\widehat{\ell}_j(\tau)$ are Lagrange polynomials given by*

$$\ell_j(\tau) = \prod_{\substack{k=1\\k\neq j}}^{s} \left(\frac{\tau - c_k}{c_j - c_k}\right) \qquad \widehat{\ell}_j(\tau) = \prod_{\substack{k=1\\k\neq j}}^{s-1} \left(\frac{\tau - \widetilde{c}_k}{\widetilde{c}_j - \widetilde{c}_k}\right) \qquad \widetilde{\ell}_j(\tau) = \prod_{\substack{k=0\\k\neq j}}^{s} \left(\frac{\tau - \widetilde{c}_k}{\widetilde{c}_j - \widetilde{c}_k}\right).$$

*Proof.* Most of this proof is the same as the proof of Theorem 5.4.2. The only difference is the use of the internal stages $\widetilde{Z}_i$ and the $\check{a}_{ij}$ coefficients. The polynomial $\widetilde{Z}_i^r$ can be expressed as

$$\widetilde{Z}^r(t_0 + \widetilde{c}_i h) = h\sum_{j=0}^{s} \int_0^{\widetilde{c}_i} \widetilde{\ell}_j(\tau)d\tau\ r(t_0 + \widetilde{c}_j h, Y(t_0 + \widetilde{c}_j h), \Lambda(t_0 + \widetilde{c}_j h)).$$

Taking $\widetilde{Z}_i^r := \widetilde{Z}^r(t_0 + \widetilde{c}_i h)$, and $\check{a}_{ij} := \int_0^{\widetilde{c}_i} \widetilde{\ell}_j(\tau)d\tau$, we arrive at (6.8d). Using the definition of $\widetilde{Z}(t)$ allows us to take $\widetilde{Z}_i := \widetilde{Z}(t_0 + \widetilde{c}_i h)$. □

By examining the value $\widetilde{Z}(t_1)$, we find that the numerical solution $z_1$ can be

expressed in terms of either continuous collocation type polynomials, or discontinuous type polynomials. The proof for the equivalence of EMPRK methods is similar to the equivalence for SPARK methods. We therefore omit it here.

**Theorem 6.4.3.** *An EMPRK method with distinct values $c_1, \ldots, c_s$, distinct values $\widetilde{c}_0, \ldots, \widetilde{c}_s$ is a discontinuous collocation type method (6.56) if and only if the coefficients satisfy*

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \qquad \sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, s \tag{6.58a}$$

$$\sum_{j=0}^{s} \widetilde{a}_{ij} \widetilde{c}_j^{k-1} = \frac{c_i^k}{k}, \qquad \sum_{j=0}^{s} \widetilde{b}_j \widetilde{c}_j^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, s-1 \tag{6.58b}$$

$$\widetilde{a}_{i0} = \widetilde{b}_0, \qquad \widetilde{a}_{is} = 0 \tag{6.58c}$$

$$\sum_{j=1}^{s} \bar{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad k = 1, \ldots, s-1 \tag{6.58d}$$

$$\sum_{j=0}^{s} \check{a}_{ij} \widetilde{c}_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad k = 1, \ldots, s+1. \tag{6.58e}$$

We present here a lemma regarding the error of the internal stages of a discontinuous collocation type or EMPRK method for mixed index 2 and 3 DAEs, assuming Gauss-Lobatto coefficients. This lemma will be useful for showing the effectiveness of the derivatives of discontinuous collocation type methods, as well as for a proof determining the local error. See the proof of Theorem 3.5.3 for details.

**Lemma 6.4.4.** *Suppose the internal stages $Y_i$, $Z_i$, $\widetilde{Y}_j$, $\widetilde{Z}_j$, $\Lambda_j$, and $\Psi_i$ are as defined in (6.4) with Gauss-Lobatto coefficients, for $i = 1, \ldots, s$, and $j = 0, \ldots, s$. Let $y(t), z(t), \lambda(t), \psi(t)$ be the exact solutions to (6.1), and let $z^f(t)$, $z^r(t)$, $\widetilde{z}^f(t)$, $\widetilde{z}^r(t)$*

*be the exact solutions to (6.9). Then we have the bounds*

$$\begin{aligned}
Y_i - y(t_0 + c_i h) &= \mathcal{O}(h^{s+1}), & Z_i - z(t_0 + c_i h) &= \mathcal{O}(h^{s+1}),\\
\widetilde{Y}_i - y(t_0 + \widetilde{c}_i h) &= \mathcal{O}(h^{s+2}), & \widetilde{Z}_i - z(t_0 + \widetilde{c}_i h) &= \mathcal{O}(h^{s+2}),\\
\Lambda_i - \lambda(t_0 + \widetilde{c}_i h) &= \mathcal{O}(h^{s}), & \Psi_i - \psi(t_0 + c_i h) &= \mathcal{O}(h^{s}),\\
Z_i^f - z^f(t_0 + c_i h) &= \mathcal{O}(h^{s+1}), & Z_i^r - z^r(t_0 + \widetilde{c}_i h) &= \mathcal{O}(h^{s+1}),\\
\widetilde{Z}_i^f - \widetilde{z}^f(t_0 + c_i h) &= \mathcal{O}(h^{s+2}), & \widetilde{Z}_i^r - \widetilde{z}^r(t_0 + \widetilde{c}_i h) &= \mathcal{O}(h^{s+2}),\\
y_1 - y(t_1) &= \mathcal{O}(h^{s+1}), & z_1 - z(t_1) &= \mathcal{O}(h^{s+1}).
\end{aligned} \tag{6.59}$$

## 6.5 Local Error Analysis and Convergence

Using the fact that the Gauss-Lobatto EMPRK methods are equivalent to a class discontinuous collocation methods, we can determine the local error for these methods.

**Theorem 6.5.1.** *For the $(s, s)$-Gauss-Lobatto EMPRK methods (6.4) with consistent initial values $(y_0, z_0, \lambda_0, \psi_0)$ at time $t_0$, assume the matrices given by (6.2) are invertible. Then for $|h| \leq h_0$, the local error is of order $2s$, i.e.,*

$$y_1 - y(t_1) = \mathcal{O}(h^{2s+1}), \qquad z_1 - z(t_1) = \mathcal{O}(h^{2s+1}). \tag{6.60}$$

*Proof.* This proof is similar to the proof for SPARK methods given in Theorem 5.5.1. We therefore omit it. □

We now state a result concerning the global convergence of the Gauss-Lobatto EMPRK methods applied to problems of the form (6.1). The proof is identical to that of Theorem 5.6.1 for the SPARK methods.

**Theorem 6.5.2.** *Consider the $(s, s)$-Gauss-Lobatto EMPRK methods applied to problem (6.1) with consistent initial conditions $(y_0, z_0)$ at time $t_0$. Suppose that the matrices (6.2) are invertible. Then the $(s, s)$-Gauss-Lobatto EMPRK methods are convergent of order $2s$, i.e.*

$$y_N - y(t_N) = \mathcal{O}(h^{2s}), \quad z_N - z(t_N) = \mathcal{O}(h^{2s}), \tag{6.61}$$

*where $y_N$ and $z_N$ are the numerical solution at time $t_N := t_0 + Nh$, for $Nh \leq Const$.*

# CHAPTER 7
# LAGRANGE-D'ALEMBERT INTEGRATORS APPLIED TO LAGRANGIAN SYSTEMS WITH CONSTRAINTS

## 7.1 Introduction

In this chapter, we introduce the Lagrange-d'Alembert principle for systems with mixed index 2 and 3 constraints. The goal is to show that the Gauss-Lobatto SPARK and the EMPRK methods for DAEs of mixed index 2 and 3 are Lagrange-d'Alembert integrators. We consider the constrained Lagrangian system with mixed index 2 and 3 given by

$$\frac{d}{dt}q = v \tag{7.1a}$$

$$\frac{d}{dt}\nabla_v L(t,q,v) = \nabla_q L(t,q,v) - G(t,q)^T\lambda - K(t,q,v)^T\psi \tag{7.1b}$$

$$0 = g(t,q) \tag{7.1c}$$

$$0 = g_t(t,q) + g_q(t,q)v \tag{7.1d}$$

$$0 = k(t,q,v), \tag{7.1e}$$

with $G(t,q) := g_q(t,q)$ and $K(t,q,v) := k_v(t,q,v)$. The matrices

$$L_{vv}(t,q,v) \tag{7.2a}$$

$$G(t,q)L_{vv}(t,q,v)^{-1}G(t,q)^T \tag{7.2b}$$

$$\begin{bmatrix} G(t,q)L_{vv}(t,q,v)^{-1}G(t,q)^T & G(t,q)L_{vv}(t,q,v)^{-1}K(t,q,v)^T \\ K(t,q,v)L_{vv}(t,q,v)^{-1}G(t,q)^T & K(t,q,v)L_{vv}(t,q,v)^{-1}K(t,q,v)^T \end{bmatrix} \tag{7.2c}$$

are assumed invertible. The functions in this system satisfy

$$L : \mathbb{R} \times \mathbb{R}^{n_q} \times \mathbb{R}^{n_q} \to \mathbb{R}$$

$$g : \mathbb{R} \times \mathbb{R}^{n_q} \to \mathbb{R}^{n_g}$$

$$k : \mathbb{R} \times \mathbb{R}^{n_q} \times \mathbb{R}^{n_q} \to \mathbb{R}^{n_k},$$

where we assume $n_g < n_q$ and $n_k < n_q$. These allow for the system (7.1) to be written as a system of DAEs. One can therefore consider $\lambda = \lambda(t, q, v)$ and $\psi = \psi(t, q, v)$. This system is also an example of a forced Lagrangian system, with forcing terms defined by

$$f_L(t, q, v) := -G(t, q)^T \lambda(t, q, v) - K(t, q, v)^T \psi(t, q, v). \tag{7.3}$$

## 7.2 The Lagrange-d'Alembert Principle

The forced Lagrange-d'Alembert principle is discussed in, for example, [17] and [18]. For the remainder of this chapter, we consider the system (7.1) over a time range denoted by $[t_0, t_N]$, which is partitioned by $\{t_i\}_{i=0}^N$. Following [11], for the forced Lagrangian system (7.1), we define the *action integral*

$$A(q) := \int_{t_0}^{t_N} L(t, q(t), \dot{q}(t))dt.$$

The corresponding Lagrange-d'Alembert principle is then

$$0 = \delta A(q)(\delta q) + \int_{t_0}^{t_N} f_L(t, q(t), \dot{q}(t))^T \delta q(t) dt \tag{7.4a}$$

$$0 = g(t, q(t)) \tag{7.4b}$$

$$0 = g_t(t, q(t)) + g_q(t, q(t))\dot{q}(t) \tag{7.4c}$$

$$0 = k(t, q(t), \dot{q}(t)) \tag{7.4d}$$

for all sufficiently smooth virtual displacements $\delta q(t)$ with $\delta q(t_0) = \delta q(t_N) = 0$. A corresponding forced discrete Lagrange-d'Alembert principle is presented in [18]. It is given as

$$\begin{aligned} &\delta \sum_{k=0}^{N-1} L_d(t_k, q_k, t_{k+1}, q_{k+1}) \\ &+ \sum_{k=0}^{N-1} (f_d^-(t_k, q_k, t_{k+1}, q_{k+1})^T \delta q_k + f_d^+(t_k, q_k, t_{k+1}, q_{k+1})^T \delta q_{k+1}) = 0 \end{aligned} \tag{7.5a}$$

$$0 = b(t_{k+1}, q_{k+1}) \tag{7.5b}$$

$$0 = c(t_k, q_k, t_{k+1}, q_{k+1}) \tag{7.5c}$$

$$0 = d(t_k, q_k, t_{k+1}, q_{k+1}) \tag{7.5d}$$

for all variations $\{\delta q_k\}_{k=0}^N$ with $\delta q_k \in \mathbb{R}^n$ satisfying $\delta q_0 = \delta q_N = 0$, and

$$\begin{aligned}
b(t_{k+1}, q_{k+1}) &:= g(t_{k+1}, q_{k+1}) \\
c(t_k, q_k, t_{k+1}, q_{k+1}) &:= g_t(t_{k+1}, q_{k+1}) \\
&\qquad + g_q(t_{k+1}, q_{k+1}) v(t_{k+1}, q_{k+1}, u(t_{k+1}, t_k, q_k, t_{k+1}, q_{k+1})) \\
d(t_k, q_k, t_{k+1}, q_{k+1}) &:= k(t_{k+1}, q_{k+1}, u(t_{k+1}, t_k, q_k, t_{k+1}, q_{k+1})) \\
u(t, t_k, q_k, t_{k+1}, q_{k+1}) &\approx \frac{d}{dt} q(t, t_k, q_k, t_{k+1}, q_{k+1}).
\end{aligned}$$

The discrete Lagrangian function $L_d$ is an approximation to the exact discrete Lagrangian

$$L_d(t_k, q_k, t_{k+1}, q_{k+1}) \approx L_d^E(t_k, q_k, t_{k+1}, q_{k+1}) := \int_{t_k}^{t_{k+1}} L(t, q(t), v(t)) dt,$$

where $q(t) := q(t, t_k, q_k, t_{k+1}, q_{k+1})$. The discrete forces $f_d^-$ and $f_d^+$ are approximations to the exact discrete forces

$$\begin{aligned}
f_d^-(t_k, q_k, t_{k+1}, q_{k+1})^T &\approx f_d^{E-}(t_k, q_k, t_{k+1}, q_{k+1})^T \\
&:= \int_{t_k}^{t_{k+1}} f_L(t, q(t), v(t))^T \delta_{q_k} q(t) dt, \\
f_d^+(t_k, q_k, t_{k+1}, q_{k+1})^T &\approx f_d^{E+}(t_k, q_k, t_{k+1}, q_{k+1})^T \\
&:= \int_{t_k}^{t_{k+1}} f_L(t, q(t), v(t))^T \delta_{q_{k+1}} q(t) dt.
\end{aligned}$$

The first term in (7.5a) can be expressed as

$$\begin{aligned}
&\delta \sum_{k=0}^{N-1} L_d(t_k, q_k, t_{k+1}, q_{k+1}) \\
&\qquad = \sum_{k=0}^{N-1} \left( \nabla_2 L_d(t_k, q_k, t_{k+1}, q_{k+1}) \delta q_k + \nabla_4 L_d(t_k, q_k, t_{k+1}, q_{k+1}) \delta q_{k+1} \right).
\end{aligned}$$

Because of this, the left side of (7.5a) can be written as

$$\begin{aligned}
&\sum_{k=0}^{N-1} \left(\nabla_2 L_d(t_k, q_k, t_{k+1}, q_{k+1})\delta q_k + \nabla_4 L_d(t_k, q_k, t_{k+1}, q_{k+1})\delta q_{k+1}\right) \\
&\qquad + \sum_{k=0}^{N-1} \left(f_d^-(t_k, q_k, t_{k+1}, q_{k+1})^T \delta q_k + f_d^+(t_k, q_k, t_{k+1}, q_{k+1})^T \delta q_{k+1}\right) \\
&= \nabla_2 L_d(t_0, q_0, t_1, q_1)\delta q_0 + \nabla_4 L_d(t_{N-1}, q_{N-1}, t_N, q_N)\delta q_N \\
&\qquad + \sum_{k=1}^{N-1} \left(\nabla_2 L_d(t_k, q_k, t_{k+1}, q_{k+1}) + \nabla_4 L_d(t_{k-1}, q_{k-1}, t_k, q_k)\right) \delta q_k \\
&\quad + f_d^-(t_0, q_0, t_1, q_1)^T \delta q_0 + f_d^+(t_{N-1}, q_{N-1}, t_N, q_N)\delta q_N \\
&\qquad + \sum_{k=1}^{N-1} \left(f_d^-(t_k, q_k, t_{k+1}, q_{k+1})^T + f_d^+(t_{k-1}, q_{k-1}, t_k, q_k)^T\right) \delta q_k \\
&= \sum_{k=1}^{N-1} \big(\nabla_2 L_d(t_k, q_k, t_{k+1}, q_{k+1}) + \nabla_4 L_d(t_{k-1}, q_{k-1}, t_k, q_k) \\
&\qquad + f_d^-(t_k, q_k, t_{k+1}, q_{k+1})^T + f_d^+(t_{k-1}, q_{k-1}, t_k, q_k)^T\big)\delta q_k.
\end{aligned}$$

The calculation above makes use of the fact that $\delta q_0 = \delta q_N = 0$. Because the choice of the variations $\{\delta q_k\}_{k=1}^{N-1}$ is arbitrary, the discrete principle (7.5) is equivalent to the *discrete Lagrange-d'Alembert equations*

$$\begin{aligned}
&\nabla_2 L_d(t_k, q_k, t_{k+1}, q_{k+1}) + \nabla_4 L_d(t_{k-1}, q_{k-1}, t_k, q_k) \\
&\qquad + f_d^-(t_k, q_k, t_{k+1}, q_{k+1}) + f_d^+(t_{k-1}, q_{k-1}, t_k, q_k) = 0
\end{aligned} \tag{7.6a}$$

$$0 = b(t_k, q_k) \tag{7.6b}$$

$$0 = c(t_{k-1}, q_{k-1}, t_k, q_k) \tag{7.6c}$$

$$0 = d(t_{k-1}, q_{k-1}, t_k, q_k) \tag{7.6d}$$

for $k = 1, \ldots, N-1$.

### 7.3 Exact Discrete Forcing Terms

In this section, we derive expressions for the discrete forcing terms $f_d^+$ and $f_d^-$. Let $q(t) = q(t, t_0, q_0, t_1, q_1)$ and $v(t) = v(t, t_0, q_0, t_1, q_1)$ be solutions of (7.1) passing through $q_0$ at $t_0$ and $q_1$ at $t_1$. We follow the work presented in [11]. The exact

discrete Lagrangian is given by

$$L_d^E(t_0, q_0, t_1, q_1) := \int_{t_0}^{t_1} L(t, q(t), v(t))dt. \tag{7.7}$$

Denote by $p_0$ and $p_1$ the quantities

$$p_0 := \nabla_v L(t_0, q_0, v_0), \qquad p_1 := \nabla_v L(t_1, q_1, v_1).$$

Using the exact discrete Lagrangian and (7.6), we derive expressions for the exact forcing terms. The partial derivatives of $L_d^E$ are

$$\begin{aligned}
\partial_{q_0} L_d^E(t_0, q_0, t_1, q_1) &= \int_{t_0}^{t_1} \left(L_q(t, q(t), v(t))\partial_{q_0} q(t) + L_v(t, q(t), v(t))\partial_{q_0} v(t)\right) dt \\
&= \int_{t_0}^{t_1} (L_q(t, q(t), v(t))\partial_{q_0} q(t) - \frac{d}{dt} L_v(t, q(t), v(t))\partial_{q_0} q(t))dt \\
&\qquad + L_v(t, q(t), v(t))\partial_{q_0} q(t)\Big|_{t_0}^{t_1} \\
&= \int_{t_0}^{t_1} \left(L_q(t, q(t), v(t)) - \frac{d}{dt} L_v(t, q(t), v(t))\right) \partial_{q_0} q(t)dt \\
&\qquad + L_v(t_1, q_1, v_1)\partial_{q_0} q_1 - L_v(t_0, q_0, v_0)\partial_{q_0} q_0 \\
&= \int_{t_0}^{t_1} \big(\Lambda(t, q(t), v(t))^T G(t, q(t)) \\
&\qquad + \Psi(t, q(t), v(t))^T K(t, q(t), v(t))\big)\partial_{q_0} q(t)dt - p_0^T, \\
\partial_{q_1} L_d^E(t_0, q_0, t_1, q_1) &= \int_{t_0}^{t_1} \left(L_q(t, q(t), v(t))\partial_{q_1} q(t) + L_v(t, q(t), v(t))\partial_{q_1} v(t)\right) dt \\
&= \int_{t_0}^{t_1} (L_q(t, q(t), v(t))\partial_{q_1} q(t) - \frac{d}{dt} L_v(t, q(t), v(t))\partial_{q_1} q(t))dt \\
&\qquad + L_v(t, q(t), v(t))\partial_{q_1} q(t)\Big|_{t_0}^{t_1} \\
&= \int_{t_0}^{t_1} \left(L_q(t, q(t), v(t)) - \frac{d}{dt} L_v(t, q(t), v(t))\right) \partial_{q_1} q(t)dt \\
&\qquad + L_v(t_1, q_1, v_1)\partial_{q_1} q_1 - L_v(t_0, q_0, v_0)\partial_{q_1} q_0 \\
&= \int_{t_0}^{t_1} \big(\Lambda(t, q(t), v(t))^T G(t, q(t)) \\
&\qquad + \Psi(t, q(t), v(t))^T K(t, q(t), v(t))\big)\partial_{q_1} q(t)dt + p_1^T.
\end{aligned}$$

We use $\Lambda(t, q(t), v(t))$ and $\Psi(t, q(t), v(t))$ to represent $\lambda$ and $\psi$, respectively, as functions of $t$, $q$, and $v$. To help keep the presentation clean, we introduce the

function

$$\begin{aligned}\Upsilon(t) &= \Upsilon(t, t_0, q_0, t_1, q_1) \\ &:= \Lambda(t, q(t), v(t))^T G(t, q(t)) + \Psi(t, q(t), v(t))^T K(t, q(t), v(t)).\end{aligned}$$

A natural definition for the exact forcing terms is then

$$f_d^{E-}(t_0, q_0, t_1, q_1) := -p_0 - \nabla_{q_0} L_d^E(t_0, q_0, t_1, q_1) \tag{7.8}$$

$$f_d^{E+}(t_0, q_0, t_1, q_1) := p_1 - \nabla_{q_1} L_d^E(t_0, q_0, t_1, q_1), \tag{7.9}$$

as these easily satisfy (7.6). Therefore,

$$f_d^{E-}(t_0, q_0, t_1, q_1)^T := -\int_{t_0}^{t_1} \Upsilon(t) \partial_{q_0} q(t) dt \tag{7.10a}$$

$$f_d^{E+}(t_0, q_0, t_1, q_1)^T := -\int_{t_0}^{t_1} \Upsilon(t) \partial_{q_1} q(t) dt. \tag{7.10b}$$

The expressions for the exact forcing terms in (7.10) can be reexpressed using the Fundamental Theorem of Calculus. This gives

$$\begin{aligned}q(t) &= q_0 + \int_{t_0}^{t} \frac{d}{ds} q(s) ds = q_0 + \int_{t_0}^{t} v(s) ds \\ q(t) &= q_1 + \int_{t_1}^{t} \frac{d}{ds} q(s) ds = q_1 - \int_{t}^{t_1} v(s) ds\end{aligned}$$

and

$$\begin{aligned}\partial_{q_0} q(t) &= -\int_{t}^{t_1} \partial_{q_0} v(s) ds \\ \partial_{q_1} q(t) &= \int_{t_0}^{t} \partial_{q_1} v(s) ds.\end{aligned}$$

Thus, the exact forcing terms can be expressed as

$$\begin{aligned}f_d^{E-}(t_0, q_0, t_1, q_1)^T &= \int_{t_0}^{t_1} \Upsilon(t) \left( \int_{t}^{t_1} \partial_{q_0} v(s) ds \right) dt \\ &= \int_{t_0}^{t_1} \int_{t_0}^{s} \Upsilon(t) \partial_{q_0} v(s) dt ds,\end{aligned}$$

$$f_d^{E+}(t_0,q_0,t_1,q_1)^T = -\int_{t_0}^{t_1} \Upsilon(t)\left(\int_{t_0}^{t}\partial_{q_1}v(s)ds\right)dt$$
$$= -\int_{t_0}^{t_1}\int_{s}^{t_1}\Upsilon(t)\partial_{q_1}v(s)dtds.$$

For a general $t_j := t_0 + jh$ with $j = 0,\ldots,N$, we can then see that

$$\nabla_4 L_d^E(t_{k-1},q_{k-1},t_k,q_k) + \nabla_2 L_d^E(t_k,q_k,t_{k+1},q_{k+1})$$
$$+ f_d^{E+}(t_{k-1},q_{k-1},t_k,q_k) + f_d^{E-}(t_k,q_k,t_{k+1},q_{k+1}) = 0$$

for $k = 1,\ldots,N-1$, with

$$f_d^{E-}(t_k,q_k,t_{k+1},q_{k+1})^T = \int_{t_k}^{t_{k+1}}\int_{t_k}^{s}\big(\Lambda(t,q(t),v(t))^T G(t,q(t)) + \Psi(t,q(t),v(t))^T K(t,q(t),v(t))\big)\partial_{q_k}v(s)dtds \tag{7.11a}$$

$$f_d^{E+}(t_{k-1},q_{k-1},t_k,q_k)^T = -\int_{t_{k-1}}^{t_k}\int_{s}^{t_k}\big(\Lambda(t,q(t),v(t))^T G(t,q(t)) + \Psi(t,q(t),v(t))^T K(t,q(t),v(t))\big)\partial_{q_k}v(s)dtds. \tag{7.11b}$$

## 7.4 SPARK Methods as Lagrange-d'Alembert Integrators for Mixed Index 2 and Index 3 Lagrangian Systems

Combining the methods presented in [15] and [16], the application of an $(s,s)$-stage SPARK method to the Lagrangian system (7.1) with stepsize $h$ and consistent initial values $q_0$, $v_0$ at time $t_0$, can be expressed as

$$Q_i = q_0 + h\sum_{j=1}^{s} a_{ij}V_j, \quad i = 1,\ldots,s \tag{7.12a}$$

$$\widetilde{Q}_i = q_0 + h\sum_{j=1}^{s} \bar{a}_{ij}V_j, \quad i = 0,\ldots,s \tag{7.12b}$$

$$P_i = p_0 + h\sum_{j=1}^{s}\widehat{a}_{ij}F_j + h\sum_{j=0}^{s}\widetilde{a}_{ij}R_j + h\sum_{j=1}^{s}\widehat{a}_{ij}S_j, \quad i = 1,\ldots,s \tag{7.12c}$$

$$q_1 = q_0 + h\sum_{j=1}^{s} b_jV_j \tag{7.12d}$$

$$p_1 = p_0 + h\sum_{j=1}^{s}\widehat{b}_jF_j + h\sum_{j=0}^{s}\widetilde{b}_jR_j + h\sum_{j=1}^{s}\widehat{b}_jS_j \tag{7.12e}$$

$$0 = g(t_0 + \widetilde{c}_i h, \widetilde{Q}_i), \quad i = 0,\ldots,s \tag{7.12f}$$

$$0 = g(t_1, q_1) \tag{7.12g}$$

$$0 = g_t(t_1, q_1) + g_q(t_1, q_1)v_1 \tag{7.12h}$$

$$0 = \sum_{j=1}^{s} \omega_{ij} k(t_0 + c_j h, Q_j, V_j), \quad i = 1, \ldots, s \tag{7.12i}$$

$$0 = k(t_1, q_1, v_1), \tag{7.12j}$$

where we have used the notation

$$t_1 := t_0 + h, \quad P_i := \nabla_v L(t_0 + c_i h, Q_i, V_i), \quad F_i := \nabla_q L(t_0 + c_i h, Q_i, V_i),$$
$$R_i := -G(t_0 + \widetilde{c}_i h, \widetilde{Q}_i)^T \Lambda_i, \quad S_i := -K(t_0 + c_i h, Q_i, V_i)^T \Psi_i,$$
$$p_0 := \nabla_v L(t_0, q_0, v_0), \quad p_1 := \nabla_v L(t_1, q_1, v_1).$$

These methods, particularly those with Gauss-Lobatto coefficients, are analyzed in Chapter 5. The assumptions for existence and uniqueness of a solution from that chapter are assumed here. Note that in (7.12c,e) we have separated out the terms with the algebraic variables. We consider as functions of $(t_0, q_0, t_1, q_1)$ the quantities $v_0$, $v_1$, $p_0$, $p_1$, $Q_i$, $\widetilde{Q}_j$, $V_i$, $\Lambda_j$, $\Psi_i$, $P_i$, $F_i$, $R_j$, $S_i$, for $i = 1, \ldots, s$ and $j = 0, \ldots, s$. The coefficients $\omega_{ij}$ can be taken as $\omega_{ij} := b_j c_j^{i-1}$.

**Theorem 7.4.1.** *For the Lagrangian system (7.1) and the $(s, s)$-stage SPARK method (7.12), suppose $t_0$, $q_0$ and $t_1$, $q_1$ are given. If the SPARK coefficients satisfy*

$$\widehat{b}_i = b_i, \quad i = 1, \ldots, s, \tag{7.13a}$$

$$\widehat{b}_i a_{ij} + b_j \widehat{a}_{ji} - \widehat{b}_i b_j = 0, \quad i, j = 1, \ldots, s, \tag{7.13b}$$

*then we have a discrete Lagrange-d'Alembert integrator in the sense of (7.5) with*

$$L_d(t_0, q_0, t_1, q_1) := h \sum_{i=1}^{s} b_i L(t_0 + c_i h, Q_i, V_i) \tag{7.14a}$$

$$
\begin{aligned}
f_d^-(t_0,q_0,t_1,q_1) &:= h^2\sum_{i=1}^{s}\sum_{j=0}^{s} b_i\widetilde{a}_{ij}\Lambda_j^T G(t_0+\widetilde{c}_j h,\widetilde{Q}_j)\partial_{q_0}V_i \\
&\quad + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i\widehat{a}_{ij}\Psi_j^T K(t_0+c_j h,Q_j,V_j)\partial_{q_0}V_i
\end{aligned}
\tag{7.14b}
$$

$$
\begin{aligned}
f_d^+(t_0,q_0,t_1,q_1) &:= h^2\sum_{i=1}^{s}\sum_{j=0}^{s} b_i(\widetilde{a}_{ij}-\widetilde{b}_j)\Lambda_j^T G(t_0+\widetilde{c}_j h,\widetilde{Q}_j)\partial_{q_1}V_i \\
&\quad + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i(\widehat{a}_{ij}-\widehat{b}_j)\Psi_j^T K(t_0+c_j h,Q_j,V_j)\partial_{q_1}V_i.
\end{aligned}
\tag{7.14c}
$$

*Proof.* We prove the SPARK method (7.12) is a discrete Lagrange-d'Alembert integrator by repeating the derivation for the exact forcing terms. The constraint (7.6b) is satisfied by (7.12g). Also, (7.6c) is satisfied by (7.12h) and (7.6d) by (7.12j), using $u(t,t_0,q_0,t_1,q_1) = v_1(t,t_0,q_0,t_1,q_1)$. We have

$$
\begin{aligned}
\partial_{q_0}L_d(t_0,q_0,t_1,q_1) &= h\sum_{i=1}^{s} b_i L_q(t_0+c_i h,Q_i,V_i)\partial_{q_0}Q_i \\
&\qquad + h\sum_{i=1}^{s} b_i L_v(t_0+c_i h,Q_i,V_i)\partial_{q_0}V_i \\
&= h\sum_{i=1}^{s} b_i F_i^T\left(I_n + h\sum_{j=1}^{s} a_{ij}\partial_{q_0}V_j\right) + h\sum_{i=1}^{s} b_i P_i^T\partial_{q_0}V_i
\end{aligned}
$$

since

$$
\partial_{q_0}Q_i = I_n + h\sum_{j=1}^{s} a_{ij}\partial_{q_0}V_j
$$

by (7.12a). But then by (7.12c),

$$
\begin{aligned}
\partial_{q_0}L_d(t_0,q_0,t_1,q_1) &= h\sum_{j=1}^{s} b_i F_i^T I_n + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij}F_i^T\partial_{q_0}V_j + h\sum_{i=1}^{s} b_i p_0^T\partial_{q_0}V_i \\
&\quad + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i\widehat{a}_{ij}F_j^T\partial_{q_0}V_i + h^2\sum_{i=1}^{s}\sum_{j=0}^{s} b_i\widetilde{a}_{ij}R_j^T\partial_{q_0}V_i \\
&\quad + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i\widehat{a}_{ij}S_j^T\partial_{q_0}V_i \\
&= h\sum_{i=1}^{s} b_i F_i^T I_n + h^2\sum_{i=1}^{s}\sum_{j=1}^{s}(b_j a_{ji} + b_i\widehat{a}_{ij})F_j^T\partial_{q_0}V_i
\end{aligned}
$$

$$+ p_0^T h \sum_{i=1}^{s} b_i \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_0} V_i$$
$$+ h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} S_j^T \partial_{q_0} V_i.$$

Taking the derivative with respect to $q_1$ of equation (7.12d) gives the relations

$$0 = I_n + h \sum_{i=1}^{s} b_i \partial_{q_0} V_i \quad \Rightarrow \quad I_n = -h \sum_{i=1}^{s} b_i \partial_{q_0} V_i.$$

Using this in the relation above gives

$$\begin{aligned}
\partial_{q_0} L_d(t_0, q_0, t_1, q_1) &= h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_j a_{ji} + b_i \widehat{a}_{ij} - b_j b_i) F_j^T \partial_{q_0} V_i - p_0^T \\
&\qquad + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} S_j^T \partial_{q_0} V_i \\
&= -p_0^T + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} S_j^T \partial_{q_0} V_i \\
&= -p_0^T - h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} \Lambda_j^T G(t_0 + \widetilde{c}_j h, \widetilde{Q}_j) \partial_{q_0} V_i \\
&\qquad - h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} \Psi_j^T K(t_0 + c_j h, Q_j, V_j) \partial_{q_0} V_i.
\end{aligned}$$

Here we have made use of the assumption (7.13b). But this is a discrete analog to (7.8) and (7.11a). So we have that

$$\begin{aligned}
f_d^-(t_0, q_0, t_1, q_1) &= h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} \Lambda_j^T G(t_0 + \widetilde{c}_j h, \widetilde{Q}_j) \partial_{q_0} V_i \\
&\qquad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} \Psi_j^T K(t_0 + c_j h, Q_j, V_j) \partial_{q_0} V_i.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\partial_{q_1} L_d(t_0, q_0, t_1, q_1) &= h \sum_{i=1}^{s} b_i L_q(t_0 + c_j h, Q_i, V_i) \partial_{q_1} Q_i \\
&\qquad + h \sum_{i=1}^{s} b_i L_v(t_0 + c_j h, Q_i, V_i) \partial_{q_1} V_i
\end{aligned}$$

$$= h \sum_{i=1}^{s} b_i F_i^T \left( h \sum_{j=1}^{s} a_{ij} \partial_{q_1} V_j \right) + h \sum_{i=1}^{s} b_i P_i^T \partial_{q_1} V_i$$

since

$$\partial_{q_1} Q_i = h \sum_{j=1}^{s} a_{ij} \partial_{q_1} V_j$$

by (7.12a). Using (7.12c) gives

$$\begin{aligned}
\partial_{q_1} L_d(t_0, q_0, t_1, q_1) &= h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} F_i^T \partial_{q_1} V_j + h \sum_{i=1}^{s} b_i p_0^T \partial_{q_1} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} F_j^T \partial_{q_1} V_i + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_1} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} S_j^T \partial_{q_1} V_i \\
&= h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_j a_{ji} + b_i \widehat{a}_{ij}) F_j^T \partial_{q_1} V_i + p_0^T h \sum_{i=1}^{s} b_i \partial_{q_1} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_1} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i \widehat{a}_{ij} S_j^T \partial_{q_1} V_i.
\end{aligned}$$

Writing (7.12e) as

$$p_0^T = p_1^T - h \sum_{j=1}^{s} \widehat{b}_j F_j^T - h \sum_{j=0}^{s} \widetilde{b}_j R_j^T - h \sum_{j=1}^{s} \widehat{b}_j S_j^T,$$

the calculation above becomes

$$\begin{aligned}
\partial_{q_1} L_d(t_0, q_0, t_1, q_1) &= h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_j a_{ji} + b_i \widehat{a}_{ij} - \widehat{b}_j b_i) F_j^T \partial_{q_1} V_i + p_1^T h \sum_{i=1}^{s} b_i \partial_{q_1} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i (\widetilde{a}_{ij} - \widetilde{b}_j) R_j^T \partial_{q_1} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i (\widehat{a}_{ij} - \widehat{b}_j) S_j^T \partial_{q_1} V_i.
\end{aligned}$$

Taking the derivative of (7.12d) with respect to $q_1$ gives

$$I_n = h \sum_{j=1}^{s} b_j \partial_{q_1} V_j.$$

Using this along with (7.13b) gives

$$\begin{aligned}\partial_{q_1} L_d(t_0,q_0,t_1,q_1) &= p_1^T + h^2 \sum_{i=1}^{s}\sum_{j=0}^{s} b_i(\widetilde{a}_{ij}-\widetilde{b}_j)R_j^T \partial_{q_1} V_i \\ &\quad + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i(\widehat{a}_{ij}-\widehat{b}_j)S_j^T \partial_{q_1} V_i \\ &= p_1^T - h^2 \sum_{i=1}^{s}\sum_{j=0}^{s} b_i(\widetilde{a}_{ij}-\widetilde{b}_j)\Lambda_j^T G(t_0+\widetilde{c}_j h, \widetilde{Q}_j)\partial_{q_1} V_i \\ &\quad - h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i(\widehat{a}_{ij}-\widehat{b}_j)\Psi_j^T K(t_0+c_j h, Q_j, V_j)\partial_{q_1} V_i.\end{aligned}$$

But this is a discrete analog of (7.9) and (7.11b). So we have

$$\begin{aligned}f_d^+(t_0,q_0,t_1,q_1) &= h^2 \sum_{i=1}^{s}\sum_{j=0}^{s} b_i(\widetilde{a}_{ij}-\widetilde{b}_j)\Lambda_j^T G(t_0+\widetilde{c}_j h, \widetilde{Q}_j)\partial_{q_1} V_i \\ &\quad + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i(\widehat{a}_{ij}-\widehat{b}_j)\Psi_j^T K(t_0+c_j h, Q_j, V_j)\partial_{q_1} V_i. \qquad \square\end{aligned}$$

## 7.5 MPRK Methods as Lagrange-d'Alembert Integrators for Index 2 Lagrangian Systems

We give here MPRK method as presented in [21] applied to Lagrangian system (7.1) with only nonholonomic constraints. One step of this method with stepsize $h$ and consistent initial values $q_0$, $v_0$ is the solution to the system

$$Q_i = q_0 + h\sum_{j=1}^{s} a_{ij} V_j, \quad i = 1,\dots,s \tag{7.15a}$$

$$P_i = p_0 + h\sum_{j=1}^{s} a_{ij} F_j + h\sum_{j=1}^{s} a_{ij} S_j, \quad i = 1,\dots,s \tag{7.15b}$$

$$\widetilde{Q}_i = q_0 + h\sum_{j=1}^{s} \bar{a}_{ij} V_j, \quad i = 0,\dots,s \tag{7.15c}$$

$$\widetilde{P}_i = p_0 + h\sum_{j=1}^{s} \bar{a}_{ij} F_j + h\sum_{j=1}^{s} \bar{a}_{ij} S_j, \quad i = 0,\dots,s \tag{7.15d}$$

$$q_1 = q_0 + h\sum_{j=1}^{s} b_j V_j \tag{7.15e}$$

$$p_1 = p_0 + h\sum_{j=1}^{s} b_j F_j + h\sum_{j=1}^{s} b_j S_j \tag{7.15f}$$

$$0 = k(t_0 + \widetilde{c}_i h, \widetilde{Q}_i, \widetilde{V}_i), \quad i = 0, \ldots, s, \tag{7.15g}$$

where

$$P_i := \nabla_v L(t_0 + c_i h, Q_i, V_i), \quad \widetilde{P}_i := \nabla_v L(t_0 + \widetilde{c}_i h, \widetilde{Q}_i, \widetilde{V}_i),$$

$$F_i := \nabla_q L(t_0 + c_i h, Q_i, V_i), \quad S_i := -K(t_0 + \widetilde{c}_i h, \widetilde{Q}_i, \widetilde{V}_i)^T \Psi_i.$$

We assume that

$$\bar{a}_{sj} = b_j, \quad j = 1, \ldots, s, \tag{7.16}$$

$$\bar{a}_{0j} = 0, \quad j = 1, \ldots, s, \tag{7.17}$$

$$\sum_{j=1}^{s} \bar{a}_{ij} = \widetilde{c}_i, \quad i = 0, \ldots, s. \tag{7.18}$$

This gives that

$$\widetilde{Q}_0 = q_0, \quad \widetilde{P}_0 = p_0$$

$$\widetilde{Q}_s = q_1, \quad \widetilde{P}_s = p_1.$$

Thus, (7.15g) is satisfied for $i = 0$ automatically, and (7.15g) for $i = s$ gives that $0 = k(t_1, q_1, v_1)$. We consider as functions of $t_0$, $q_0$, $t_1$, and $q_1$ the quantities $v_0$, $v_1$, $p_0$, $p_1$, $Q_i$, $P_i$, $\widetilde{Q}_j$, $\widetilde{P}_j$, $\Psi_i$, $F_i$, and $S_i$ for $i = 1, \ldots, s$ and $j = 0, \ldots, s$.

**Theorem 7.5.1.** *For the Lagrangian system (7.1) with only nonholonomic constraints and the s-stage MPRK method (7.15), suppose $t_0$, $q_0$ and $t_1$, $q_1$ are given. If the coefficients satisfy*

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \quad i, j = 1, \ldots, s, \tag{7.19}$$

*then we have a discrete Lagrange-d'Alembert integrator in the sense of (7.5) with*

$$L_d(t_0, q_0, t_1, q_1) := h \sum_{i=1}^{s} b_i L(t_0 + c_i h, Q_i, V_i) \tag{7.20a}$$

$$f_d^-(t_0, q_0, t_1, q_1) := h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_0} V_i \tag{7.20b}$$

$$f_d^+(t_0, q_0, t_1, q_1) := h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i (a_{ij} - b_j) \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_1} V_i. \tag{7.20c}$$

*Proof.* We will prove that the method (7.15) is a discrete Lagrange-dAlembert integrator by repeating the derivation for the exact forcing terms. The term (7.6d) is satisfied by (7.15g) for $i = s$ and $u(t, t_0, q_0, t_1, q_1) = v_1(t, t_0, q_0, t_1, q_1)$. We have

$$\begin{aligned}
\partial_{q_0} L_d(t_0, q_0, t_1, q_1) &= h \sum_{i=1}^{s} b_i L_q(t_0 + c_i h, Q_i, V_i) \partial_{q_0} Q_i \\
&\qquad + h \sum_{i=1}^{s} b_i L_v(t_0 + c_i h, Q_i, V_i) \partial_{q_0} V_i \\
&= h \sum_{i=1}^{s} b_i F_i^T \left( I_n + h \sum_{j=1}^{s} a_{ij} \partial_{q_0} V_j \right) + h \sum_{i=1}^{s} b_i P_i^T \partial_{q_0} V_i
\end{aligned}$$

since

$$\partial_{q_0} Q_i = I_n + h \sum_{j=1}^{s} a_{ij} \partial_{q_0} V_j$$

from (7.15a). So (7.15b) gives

$$\begin{aligned}
\partial_{q_0} L_d(t_0, q_0, t_1, q_1) &= h \sum_{i=1}^{s} b_i F_i^T + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} F_i^T \partial_{q_0} V_j + h \sum_{i=1}^{s} b_i p_0^T \partial_{q_0} V_i \\
&\qquad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} F_j^T \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i \\
&= h \sum_{i=1}^{s} b_i F_i^T + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_j a_{ji} + b_i a_{ij}) F_j^T \partial_{q_0} V_i \\
&\qquad + p_0^T h \sum_{i=1}^{s} b_i I_n \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i.
\end{aligned}$$

Taking the derivative of (7.15e) with respect to $q_0$ gives

$$0 = I_n + h\sum_{i=1}^{s} b_i \partial_{q_0} V_i.$$

Substituting this into the calculation above gives

$$\begin{aligned}\partial_{q_0} L_d(t_0, q_0, t_1, q_1) &= h^2 \sum_{i=1}^{s}\sum_{j=1}^{s}(b_j a_{ji} + b_i a_{ij} - b_i b_j) F_j^T \partial_{q_0} V_i \\ &\qquad - p_0^T + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i \\ &= -p_0^T + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i \\ &= -p_0^T - h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_0} V_i\end{aligned}$$

which follows from the assumption (7.19). But this is the discrete analog of (7.8) and (7.11a). So we have that

$$f_d^-(t_0, q_0, t_1, q_1) = h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_0} V_i.$$

Similarly, we calculate

$$\begin{aligned}\partial_{q_1} L_d(t_0, q_0, t_1, q_1) &= h \sum_{i=1}^{s} b_i L_q(t_0 + c_i h, Q_i, V_i) \partial_{q_1} Q_i \\ &\qquad + h \sum_{i=1}^{s} b_i L_v(t_0 + c_i h, Q_i, V_i) \partial_{q_1} V_i \\ &= h \sum_{i=1}^{s} b_i F_i^T \left( h \sum_{j=1}^{s} a_{ij} \partial_{q_1} V_j \right) + h \sum_{i=1}^{s} b_i P_i^T \partial_{q_1} V_i\end{aligned}$$

since the derivative of (7.15a) gives that

$$\partial_{q_1} Q_i = h \sum_{j=1}^{s} a_{ij} \partial_{q_1} V_j.$$

Applying the equation (7.15b) gives

$$\begin{aligned}\partial_{q_1} L_d(t_0,q_0,t_1,q_1) &= h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} F_i^T \partial_{q_1} V_j + h\sum_{i=1}^{s} b_i p_0^T \partial_{q_1} V_i \\ &\quad + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} F_j^T \partial_{q_1} V_i + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_1} V_i \\ &= h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} (b_j a_{ji} + b_i a_{ij}) F_j^T \partial_{q_1} V_i + h\sum_{i=1}^{s} b_i p_0^T \partial_{q_1} V_i \\ &\quad + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_1} V_i.\end{aligned}$$

Solving (7.15f) for $p_0$ and substituting into the above calculation gives

$$\begin{aligned}\partial_{q_1} L_d(t_0,q_0,t_1,q_1) &= h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} (b_j a_{ji} + b_i a_{ij} - b_i b_j) F_j^T \partial_{q_1} V_i + h\sum_{i=1}^{s} b_i p_1^T \partial_{q_1} V_i \\ &\quad + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i (a_{ij} - b_j) S_j^T \partial_{q_1} V_i \\ &= h\sum_{i=1}^{s} b_i p_1^T \partial_{q_1} V_i + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i (a_{ij} - b_j) S_j^T \partial_{q_1} V_i.\end{aligned}$$

But from (7.15e),

$$I_n = h\sum_{i=1}^{s} b_i \partial_{q_1} V_i.$$

Thus,

$$\begin{aligned}\partial_{q_1} L_d(t_0,q_0,t_1,q_1) &= p_1^T + h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i (a_{ij} - b_j) S_j^T \partial_{q_1} V_i \\ &= p_1^T - h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i (a_{ij} - b_j) \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_1} V_i.\end{aligned}$$

This is a discretization of (7.9) and (7.11b). Thus,

$$f_d^+(t_0,q_1,t_1,q_1) = h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i (a_{ij} - b_j) \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_1} V_i.$$

□

## 7.6 EMPRK Methods as Lagrange-d'Alembert Integrators for Mixed Index 2 and 3 Lagrangian Systems

We give here an extension of EMPRK methods applied to the Lagrangian system (7.1) with both holonomic and nonholonomic constraints. This method is first presented in Chapter 6. One step of this method with stepsize $h$ and consistent initial values $q_0$, $v_0$ is the solution to the system

$$Q_i = q_0 + h\sum_{j=1}^{s} a_{ij}V_j, \quad i = 1,\dots,s \tag{7.21a}$$

$$P_i = p_0 + h\sum_{j=1}^{s} a_{ij}F_j + h\sum_{j=0}^{s} \widetilde{a}_{ij}R_j + h\sum_{j=1}^{s} a_{ij}S_j, \quad i = 1,\dots,s \tag{7.21b}$$

$$\widetilde{Q}_i = q_0 + h\sum_{j=1}^{s} \bar{a}_{ij}V_j, \quad i = 0,\dots,s \tag{7.21c}$$

$$\widetilde{P}_i = p_0 + h\sum_{j=1}^{s} \bar{a}_{ij}F_j + h\sum_{j=0}^{s} \check{a}_{ij}R_j + h\sum_{j=1}^{s} \bar{a}_{ij}S_j, \quad i = 0,\dots,s \tag{7.21d}$$

$$q_1 = q_0 + h\sum_{j=1}^{s} b_jV_j \tag{7.21e}$$

$$p_1 = p_0 + h\sum_{j=1}^{s} b_jF_j + h\sum_{j=0}^{s} \widetilde{b}_jR_j + h\sum_{j=1}^{s} b_jS_j \tag{7.21f}$$

$$0 = g(t_0 + \widetilde{c}_ih, \widetilde{Q}_i), \quad i = 0,\dots,s \tag{7.21g}$$

$$0 = g(t_1, q_1) \tag{7.21h}$$

$$0 = g_t(t_1, q_1) + g_q(t_1, q_1)v_1 \tag{7.21i}$$

$$0 = k(t_0 + \widetilde{c}_ih, \widetilde{Q}_i, \widetilde{V}_i), \quad i = 0,\dots,s \tag{7.21j}$$

$$0 = k(t_1, q_1, v_1), \tag{7.21k}$$

where

$$P_i := \nabla_v L(t_0 + c_ih, Q_i, V_i), \quad \widetilde{P}_i := \nabla_v L(t_0 + \widetilde{c}_ih, \widetilde{Q}_i, \widetilde{V}_i),$$

$$F_i := \nabla_q L(t_0 + c_ih, Q_i, V_i), \quad R_i := -G(t_0 + \widetilde{c}_ih, \widetilde{Q}_i)^T\Lambda_i,$$

$$S_i := -K(t_0 + \widetilde{c}_ih, \widetilde{Q}_i, \widetilde{P}_i)^T\Psi_i.$$

The assumptions made in Chapter 6 for existence and uniqueness of a solution are also assumed here. We consider as functions of $t_0$, $q_0$, $t_1$, and $q_1$ the quantities $v_0$, $v_1$, $p_0$, $p_1$, $Q_i$, $P_i$, $\widetilde{Q}_i$, $\widetilde{P}_i$, $\Lambda_j$, $\Psi_i$, $F_i$, $R_j$ and $S_i$ for $i = 1, \ldots, s$ and $j = 0, \ldots, s$.

**Theorem 7.6.1.** *For the Lagrangian system (7.1) and the s-stage EMPRK method (7.21), suppose $t_0$, $q_0$ and $t_1$, $q_1$ are given. If the coefficients satisfy*

$$b_i a_{ij} + b_j a_{ji} - b_i b_j = 0, \quad i, j = 1, \ldots, s, \tag{7.22}$$

*then we have a discrete Lagrange-d'Alembert integrator in the sense of (7.5) with*

$$L_d(t_0, q_0, t_1, q_1) := h \sum_{i=1}^{s} b_i L(t_0 + c_i h, Q_i, V_i) \tag{7.23a}$$

$$\begin{aligned} f_d^-(t_0, q_0, t_1, q_1) := {} & h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} \Lambda_j^T G(t_0 + \widetilde{c}_j h, \widetilde{Q}_j) \partial_{q_0} V_i \\ & + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_0} V_i \end{aligned} \tag{7.23b}$$

$$\begin{aligned} f_d^+(t_0, q_0, t_1, q_1) := {} & h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i (\widetilde{a}_{ij} - \widetilde{b}_j) \Lambda_j^T G(t_0 + \widetilde{c}_j h, \widetilde{Q}_j) \partial_{q_1} V_i \\ & + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i (a_{ij} - b_j) \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_1} V_i. \end{aligned} \tag{7.23c}$$

*Proof.* We will prove that the method (7.21) is a discrete Lagrange-d'Alembert integrator by repeating the derivation for the exact forcing terms. The constraint (7.6b) is satisfied by (7.21h). The condition (7.6c) is satisfied by (7.21i) and (7.6d) by (7.21k) for $u(t, t_0, q_0, t_1, q_1) = v_1(t, t_0, q_0, t_1, q_1)$. We have

$$\begin{aligned} \partial_{q_0} L_d(t_0, q_0, t_1, q_1) = {} & h \sum_{i=1}^{s} b_i L_q(t_0 + c_i h, Q_i, V_i) \partial_{q_0} Q_i \\ & + h \sum_{i=1}^{s} b_i L_v(t_0 + c_i h, Q_i, V_i) \partial_{q_0} V_i \\ = {} & h \sum_{i=1}^{s} b_i F_i^T \left( I_n + h \sum_{j=1}^{s} a_{ij} \partial_{q_0} V_j \right) + h \sum_{i=1}^{s} b_i P_i^T \partial_{q_0} V_i \end{aligned}$$

since

$$\partial_{q_0} Q_i = I_n + h \sum_{j=1}^{s} a_{ij} \partial_{q_0} V_j$$

from (7.21a). So

$$\begin{aligned}
\partial_{q_0} L_d(t_0, q_0, t_1, q_1) &= h \sum_{i=1}^{s} b_i F_i^T + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} F_i^T \partial_{q_0} V_j + h \sum_{i=1}^{s} b_i p_0^T \partial_{q_0} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} F_j^T \partial_{q_0} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i \\
&= h \sum_{i=1}^{s} b_i F_i^T + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_j a_{ji} + b_i a_{ij}) F_j^T \partial_{q_0} V_i \\
&\quad + p_0^T h \sum_{i=1}^{s} b_i \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_0} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i.
\end{aligned}$$

Taking the derivative of (7.21e) with respect to $q_0$ gives

$$0 = I_n + h \sum_{i=1}^{s} b_i \partial_{q_0} V_i.$$

Substituting this into the calculation above gives

$$\begin{aligned}
\partial_{q_0} L_d(t_0, q_0, t_1, q_1) &= h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_j a_{ji} + b_i a_{ij} - b_i b_j) F_j^T \partial_{q_0} V_i \\
&\quad - p_0^T + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i \\
&= -p_0^T + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} R_j^T \partial_{q_0} V_i + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} S_j^T \partial_{q_0} V_i \\
&= -p_0^T - h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} b_i \widetilde{a}_{ij} \Lambda_j^T G(t_0 + \widetilde{c}_j h, \widetilde{Q}_j) \partial_{q_0} V_i \\
&\quad - h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} b_i a_{ij} \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_0} V_i,
\end{aligned}$$

which follows from the assumption (7.22). But this is the discrete analog of (7.8)

and (7.11a). So we have that

$$f_d^-(t_0,q_0,t_1,q_1) = h^2\sum_{i=1}^{s}\sum_{j=0}^{s} b_i\widetilde{a}_{ij}\Lambda_j^T G(t_0+\widetilde{c}_j h,\widetilde{Q}_j)\partial_{q_0}V_i \\ + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij}\Psi_j^T K(t_0+\widetilde{c}_j h,\widetilde{Q}_j,\widetilde{V}_j)\partial_{q_0}V_i.$$

Similarly, we calculate

$$\begin{aligned}\partial_{q_1}L_d(t_0,q_0,t_1,q_1) &= h\sum_{i=1}^{s} b_i L_q(t_0+c_i h,Q_i,V_i)\partial_{q_1}Q_i \\ &\quad + h\sum_{i=1}^{s} b_i L_v(t_0+c_i h,Q_i,V_i)\partial_{q_1}V_i \\ &= h\sum_{i=1}^{s} b_i F_i^T\left(h\sum_{j=1}^{s} a_{ij}\partial_{q_1}V_j\right) + h\sum_{i=1}^{s} b_i P_i^T\partial_{q_1}V_i\end{aligned}$$

since the derivative of (7.21a) gives that

$$\partial_{q_1}Q_i = h\sum_{j=1}^{s} a_{ij}\partial_{q_1}V_j.$$

Applying the equation (7.21b) gives

$$\begin{aligned}\partial_{q_1}L_d(t_0,q_0,t_1,q_1) &= h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij}F_i^T\partial_{q_1}V_j + h\sum_{i=1}^{s} b_i p_0^T\partial_{q_1}V_i \\ &\quad + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij}F_j^T\partial_{q_1}V_i \\ &\quad + h^2\sum_{i=1}^{s}\sum_{j=0}^{s} b_i\widetilde{a}_{ij}R_j^T\partial_{q_1}V_i + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij}S_j^T\partial_{q_1}V_i \\ &= h^2\sum_{i=1}^{s}\sum_{j=1}^{s}(b_j a_{ji}+b_i a_{ij})F_j^T\partial_{q_1}V_i + h\sum_{i=1}^{s} b_i p_0^T\partial_{q_1}V_i \\ &\quad + h^2\sum_{i=1}^{s}\sum_{j=0}^{s} b_i\widetilde{a}_{ij}R_j^T\partial_{q_1}V_i + h^2\sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij}S_j^T\partial_{q_1}V_i.\end{aligned}$$

Solving (7.21f) for $p_0$ and substituting into the above calculation gives

$$\partial_{q_1}L_d(t_0,q_0,t_1,q_1) = h^2\sum_{i=1}^{s}\sum_{j=1}^{s}(b_j a_{ji}+b_i a_{ij}-b_i b_j)F_j^T\partial_{q_1}V_i + h\sum_{i=1}^{s} b_i p_1^T\partial_{q_1}V_i$$

$$
\begin{aligned}
&+ h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} (b_i \widetilde{a}_{ij} - b_i \widetilde{b}_j) R_j^T \partial_{q_1} V_i \\
&+ h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_i a_{ij} - b_i b_j) S_j^T \partial_{q_1} V_i \\
&= p_1^T h \sum_{i=1}^{s} b_i \partial_{q_1} V_i + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} (b_i \widetilde{a}_{ij} - b_i \widetilde{b}_j) R_j^T \partial_{q_1} V_i \\
&+ h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_i a_{ij} - b_i b_j) S_j^T \partial_{q_1} V_i.
\end{aligned}
$$

But from (7.21e),

$$
I_n = h \sum_{i=1}^{s} b_i \partial_{q_1} V_i.
$$

Thus,

$$
\begin{aligned}
\partial_{q_1} L_d(t_0, q_0, t_1, q_1) &= p_1^T + h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} (b_i \widetilde{a}_{ij} - b_i \widetilde{b}_j) R_j^T \partial_{q_1} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_i a_{ij} - b_i b_j) S_j^T \partial_{q_1} V_i \\
&= p_1^T - h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} (b_i \widetilde{a}_{ij} - b_i \widetilde{b}_j) \Lambda_j^T G(t_0 + \widetilde{c}_j h, \widetilde{Q}_j) \partial_{q_1} V_i \\
&\quad - h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_i a_{ij} - b_i b_j) \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_1} V_i.
\end{aligned}
$$

This is a discretization of (7.9) and (7.11b). Thus,

$$
\begin{aligned}
f_d^+(t_0, q_1, t_1, q_1) &= h^2 \sum_{i=1}^{s} \sum_{j=0}^{s} (b_i \widetilde{a}_{ij} - b_i \widetilde{b}_j) \Lambda_j^T G(t_0 + \widetilde{c}_j h, \widetilde{Q}_j) \partial_{q_1} V_i \\
&\quad + h^2 \sum_{i=1}^{s} \sum_{j=1}^{s} (b_i a_{ij} - b_i b_j) \Psi_j^T K(t_0 + \widetilde{c}_j h, \widetilde{Q}_j, \widetilde{V}_j) \partial_{q_1} V_i.
\end{aligned}
$$

□

# CHAPTER 8
# NUMERICAL EXPERIMENTS

## 8.1 Introduction

In this chapter, we apply SPARK methods and EMPRK methods to problems arising in mechanics. The problems we consider here can be expressed as an overdetermined Lagrangian system with mixed holonomic (index 3) and nonholonomic (index 2) constraints:

$$\frac{d}{dt}q = v \tag{8.1a}$$

$$\frac{d}{dt}\nabla_v L(t,q,v) = \nabla_q L(t,q,v) - G(t,q)^T\lambda - K(t,q,v)^T\psi \tag{8.1b}$$

$$0 = g(t,q) \tag{8.1c}$$

$$0 = g_t(t,q) + g_q(t,q)v \tag{8.1d}$$

$$0 = k(t,q,v), \tag{8.1e}$$

with the functions $G$ and $K$ given by

$$G(t,q) = g_q(t,q), \qquad K(t,q,v) = k_v(t,q,v).$$

The computations for each experiment were done using *MATLAB*. An implementation for the SPARK methods was graciously provided by Laurent Jay.

## 8.2 Example Methods

For the experiments performed in this chapter, we focus on the $(1,1)$- and $(2,2)$-Gauss-Lobatto SPARK and EMPRK methods. We provide in Tables 8.2 and 8.2 the corresponding coefficients for these methods as presented in Chapters 5 and 6. Note that the two EMPRK methods use the same coefficients as the Gauss-Lobatto SPARK methods, and need the additional coefficients $\check{A}$. Each of the

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 0 | | | | | 0 | 0 |
| 1/2 | 1/2 | | 1/2 | 1/2 | | 1/2 | 0 | | 1 | 1 |
| $A$ | 1 | $\check{A}$ | | | $\widetilde{A}$ | 1/2 | 1/2 | $\bar{A}$ | | |

Table 8.1: $(1,1)$-Gauss-Lobatto coefficients

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | 0 | 0 | 0 |
| $1/2-\sqrt{3}/6$ | | $1/4$ | | $1/4-\sqrt{3}/6$ | | | 5/24 | 1/3 | $-1/24$ |
| $1/2+\sqrt{3}/6$ | | $1/4+\sqrt{3}/6$ | | $1/4$ | | | 1/6 | 2/3 | 1/6 |
| $A$ | | $1/2$ | | $1/2$ | | $\check{A}$ | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 0 | 0 | | 0 |
| | 1/6 | $1/3-\sqrt{3}/6$ | | 0 | 1/2 | $1/4+\sqrt{3}/8$ | | $1/4-\sqrt{3}/8$ |
| | 1/6 | $1/3+\sqrt{3}/6$ | | 0 | 1 | 1/2 | | 1/2 |
| $\widetilde{A}$ | 1/6 | 2/3 | | 1/6 | $\bar{A}$ | | | |

Table 8.2: $(2,2)$-Gauss-Lobatto coefficients

methods is expressed in Butcher tableaux in the format

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \qquad \begin{array}{c|} \check{A} \\ \hline \end{array} \qquad \begin{array}{c|} \widetilde{A} \\ \hline \widetilde{b}^T \end{array} \qquad \begin{array}{c|c} \widetilde{c} & \bar{A} \\ \hline & \end{array} \tag{8.2}$$

## 8.3 Numerical Experiments

### 8.3.1 The Simple Pendulum

We consider here the simple pendulum which consists of a mass $m$ suspended from a string of length $\ell$. This system is generally expressed using polar coordinates, resulting in an unconstrained Lagrangian (or Hamiltonian) system of the form (8.1). If Cartesian coordinates are instead considered, then we can express the system as a Lagrangian system with a holonomic constraint for the length of the string. The position of the mass is given by the two Cartesian coordinates $q_1$ and $q_2$, while the corresponding velocities are given by $v_1$ and $v_2$. The Lagrangian is given by

$$L(t,q,v) := T - U, \qquad T := \frac{1}{2}m(v_1^2 + v_2^2), \qquad U := -m\gamma q_2,$$

where $T$ is the kinetic energy and $U$ is the potential energy, and the constant $\gamma$ is the acceleration due to gravity. The holonomic constraint is given by

$$0 = g(t,q) = \frac{1}{2}(q_1^2 + q_2^2 - \ell^2).$$

We consider the following values for the constants and initial conditions:

$$m = 1, \qquad \ell = 1, \qquad \gamma = 1,$$
$$q_0 = (1\ 0)^T, \qquad v_0 = (0\ 0)^T.$$

We apply the $(1,1)$ and $(2,2)$–Gauss-Lobatto SPARK methods. Because this problem has only holonomic constraints, the SPARK and EMPRK methods are the same. The results for the global convergence are given in Figures 8.1. Note that the scales of the axes are logarithmic. The global order of convergence can be seen to be $2s$ for the pendulum. The exact trajectory preserves the total energy of the system $H = T + U$. The energy errors for $h = .01$ are given in Figures 8.2 and 8.3. The energy error in either figure is bounded and does not wander far from 0. This

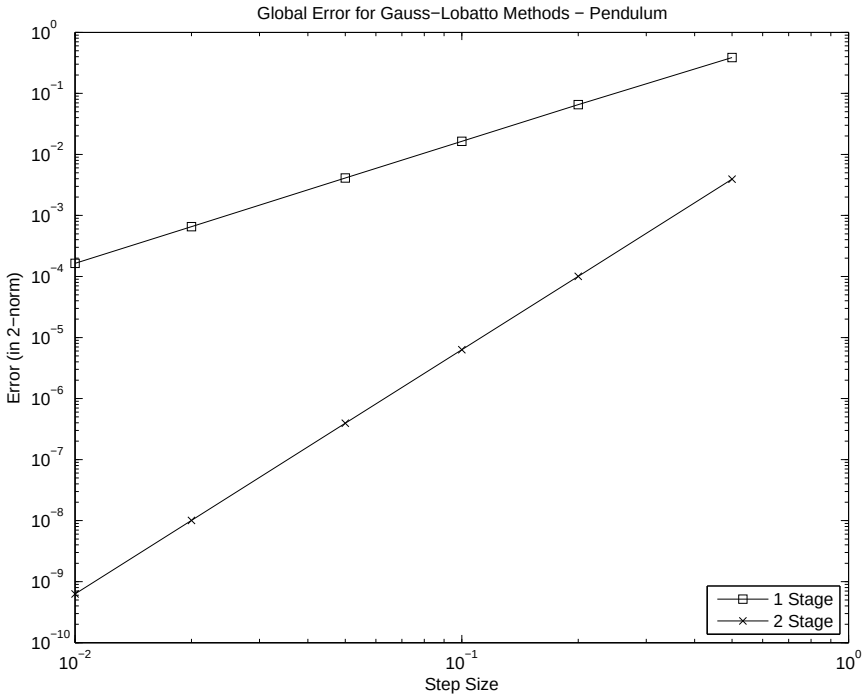


Figure 8.1: Global Error for the $q$ component for the $(1,1)$ and $(2,2)$–Gauss-Lobatto SPARK methods applied to the pendulum problem

shows that the numerical methods preserve well the energy of the system.

### 8.3.2 Skate on an Inclined Plane

We consider a thin rigid rod of fixed length $\ell$ and mass $m$ moving on a plane inclined by an angle $\beta$. The rod is constrained to move tangent to the direction in which it points. This can be viewed as a model for a skate on a tilted floor. A similar system can be found in [24]. However, we formulate the system here as a Lagrangian system (8.1) in Cartesian coordinates. This results in a system with a holonomic and a nonholonomic constraint. The coordinates $q_1$, $q_2$ represent the Cartesian coordinates of one end of the rod and $q_3$, $q_4$ the Cartesian coordinates of the other end. The vector $v = (v_1\ v_2\ v_3\ v_4)^T$ is the corresponding velocities. The

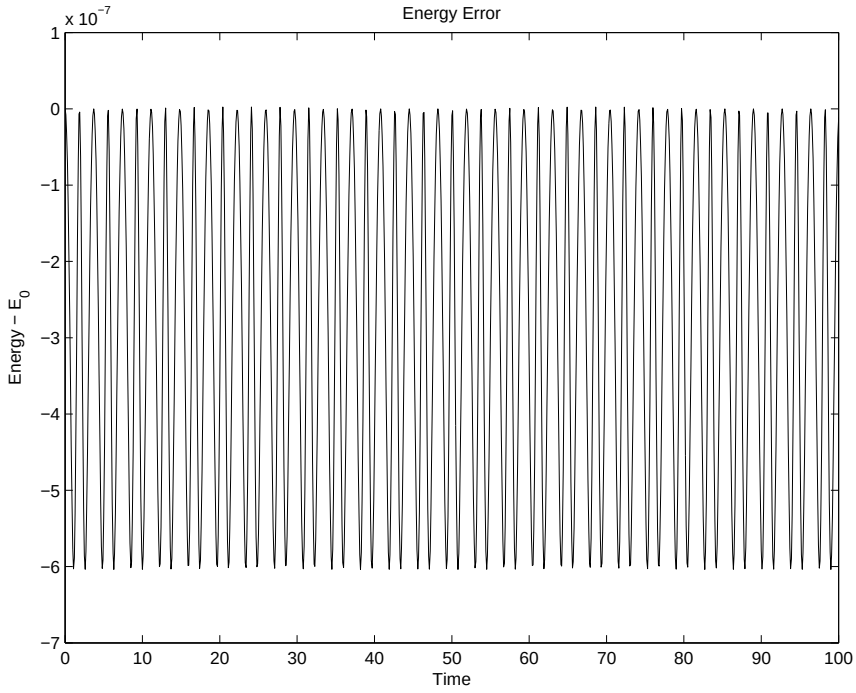


Figure 8.2: Energy error for the $(2,2)$–Gauss-Lobatto SPARK method applied to the pendulum problem with $h = .1$

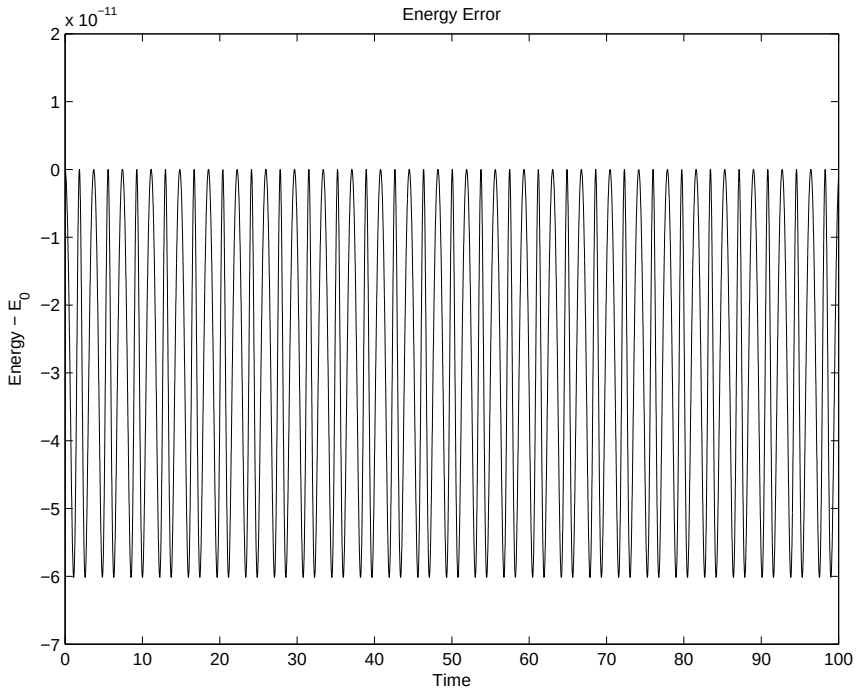


Figure 8.3: Energy error for the $(2,2)$–Gauss-Lobatto SPARK method applied to the pendulum problem with $h = .01$

Lagrangian is given by

$$L(t,q,v) := T - U,$$
$$T := \frac{1}{4}m\left(v_1^2 + v_2^2 + v_3^2 + v_4^2\right), \qquad U := -\frac{1}{2}m\gamma\sin(\beta)(q_1 + q_3).$$

The constant $\gamma$ here represents the acceleration due to gravity. The two constraints are given by

$$0 = g(t,q) = \frac{1}{2}\left((q_3 - q_1)^2 + (q_4 - q_2)^2 - \ell^2\right)$$
$$0 = k(t,q,v) = -(q_4 - q_2)(v_1 + v_3) + (q_3 - q_1)(v_2 + v_4).$$

We apply the $(1,1)$ and $(2,2)$–Gauss-Lobatto SPARK and EMPRK methods. The constants and initial conditions used are

$$m = 1, \quad \ell = 2, \quad \gamma\sin(\beta) = 1,$$
$$q_0 = \begin{pmatrix} -\frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}^T$$
$$v_0 = \begin{pmatrix} 0 & -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}^T.$$

The global errors for these methods are given in Figures 8.4 and 8.5. Again, these methods can be seen to be of order $2s$. The exact solution to this system also preserves the total energy $H := T + U$. Energy errors for $h = .1$ for the $(2,2)$–Gauss-Lobatto SPARK and EMPRK methods are given in Figures 8.6 and 8.7.

### 8.3.3 Ball on a Rotating Table

We consider a homogeneous ball rolling on a table, which is itself rotating counterclockwise at a constant angular speed $\Omega$. A formulation of this problem using Euler angles is given in [20], while an approach using essentially quaternions was done by [24]. Here we construct the system using a rotational matrix approach. This results in a Lagrangian system with both holonomic and nonholonomic constraints.

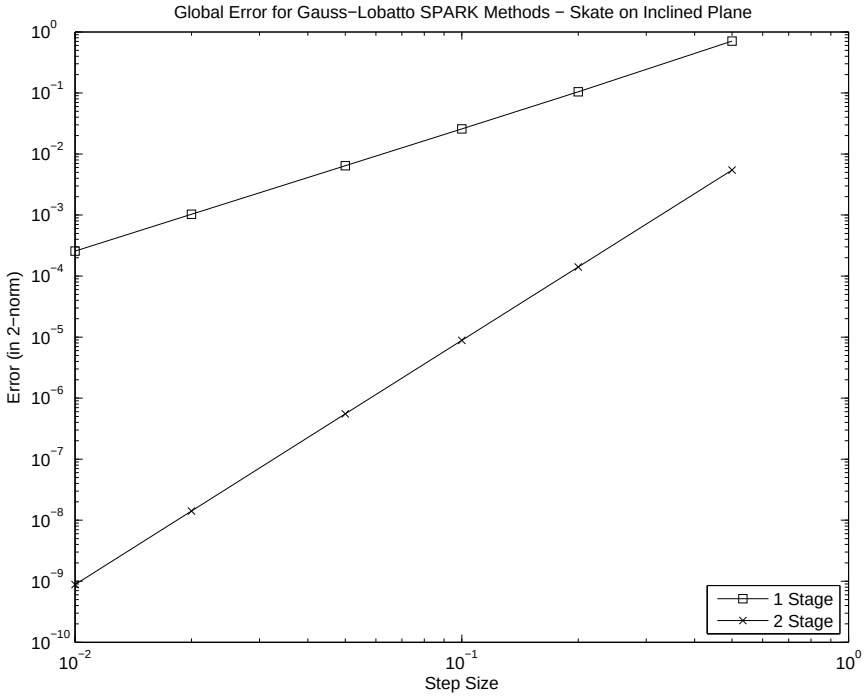


Figure 8.4: Global Error for the $q$ component for the $(1,1)$ and $(2,2)$–Gauss-Lobatto SPARK methods applied to the skate problem

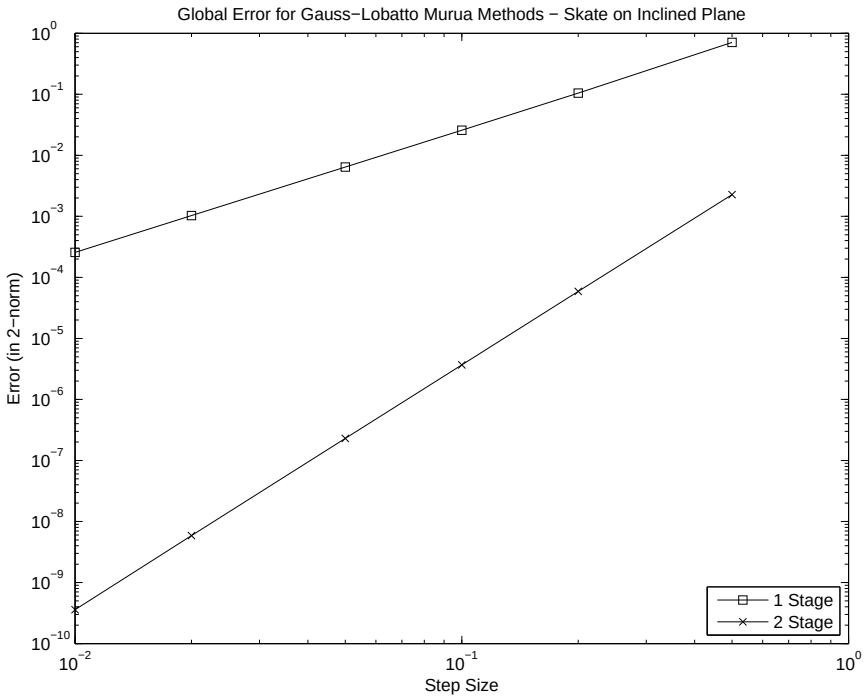


Figure 8.5: Global Error for the $q$ component for the $(1,1)$ and $(2,2)$–Gauss-Lobatto EMPRK methods applied to the skate problem

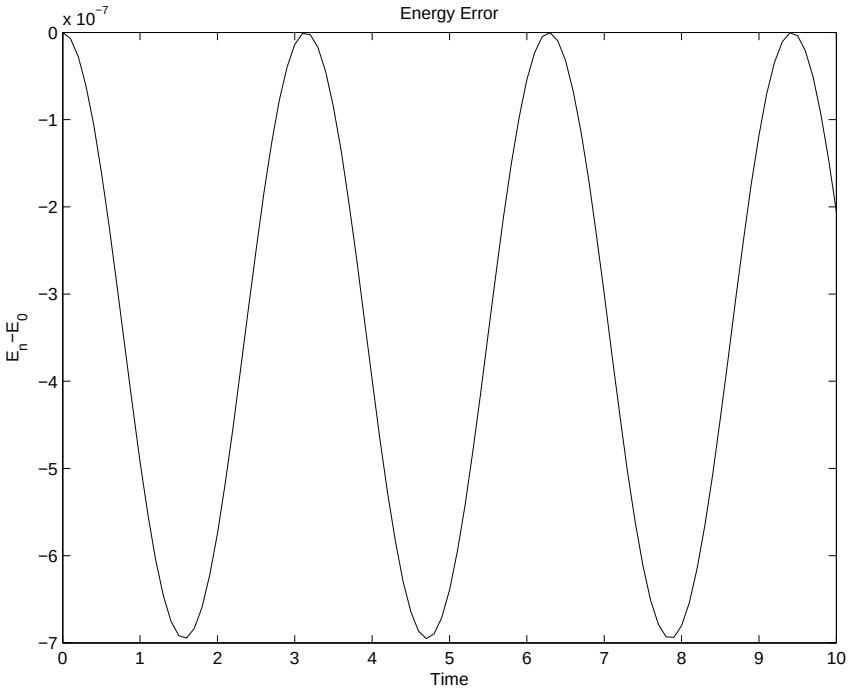


Figure 8.6: Energy error for the $(2,2)$–Gauss-Lobatto SPARK method applied to the skate on an inclined plane problem with $h = .1$

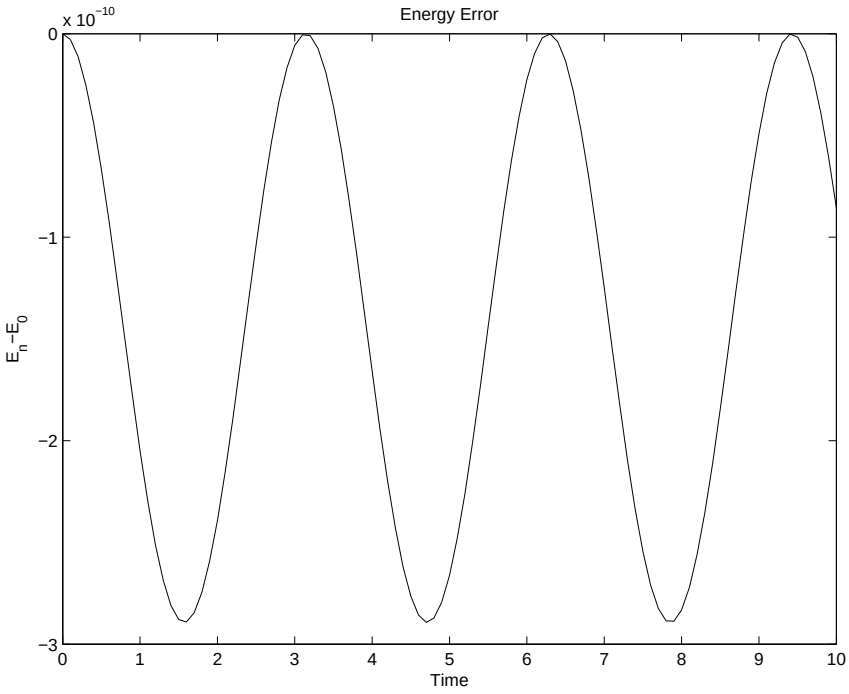


Figure 8.7: Energy error for the $(2,2)$–Gauss-Lobatto EMPRK method applied to the skate on an inclined plane problem with $h = .1$

The Lagrangian is given by

$$L(t,q,v) = T - U, \qquad T = \frac{1}{2}m(v_1^2 + v_2^2) + \frac{1}{2}I^2\sum_{i=3}^{11} v_i^2, \qquad U = 0$$

for $I$ the moment of inertia of the ball. The variables $q_1$ and $q_2$ refer to the coordinates of the center of mass of the ball. The remaining coordinates of $q$ describe the orientation of the ball. The orientation can be represented by the orthonormal matrix

$$Q = \begin{bmatrix} q_3 & q_4 & q_5 \\ q_6 & q_7 & q_8 \\ q_9 & q_{10} & q_{11} \end{bmatrix}.$$

The nonholonomic constraints for the rolling of the ball can be expressed as

$$0 = k_1(t,q,v) = v_1 - R(v_3q_9 + v_4q_{10} + v_5q_{11}) + \Omega q_2$$
$$0 = k_2(t,q,v) = v_2 - R(v_9q_6 + v_{10}q_7 + v_{11}q_8) - \Omega q_1,$$

where $R$ is the radius of the ball. Holonomic constraints are introduced from the orthonormality, giving

$$0 = g_1(t,q) = q_3^2 + q_6^2 + q_9^2 - 1$$
$$0 = g_2(t,q) = q_3q_4 + q_6q_7 + q_9q_{10}$$
$$0 = g_3(t,q) = q_3q_5 + q_6q_8 + q_9q_{11}$$
$$0 = g_4(t,q) = q_4^2 + q_7^2 + q_{10}^2 - 1$$
$$0 = g_5(t,q) = q_4q_5 + q_7q_8 + q_{10}q_{11}$$
$$0 = g_6(t,q) = q_5^2 + q_8^2 + q_{11}^2 - 1.$$

For initial conditions, we use

$$q_0 = (3\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1)^T$$

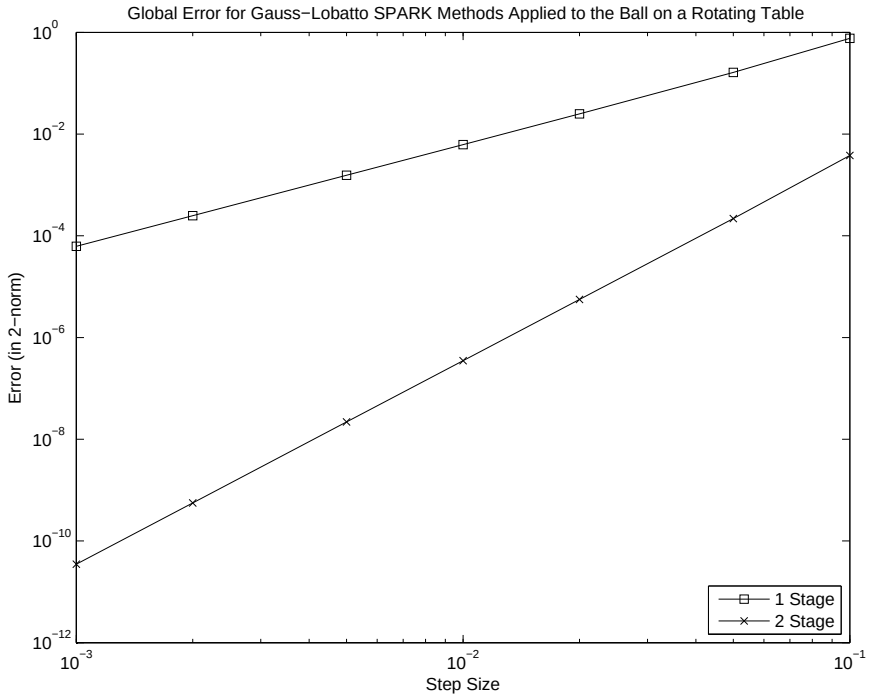


Figure 8.8: Global Error for the $q$ component for the $(1,1)$ and $(2,2)$–Gauss-Lobatto SPARK methods applied to the ball on a rotating table problem

$$v_0 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 6\ 0\ \ -6\ 0)^T,$$

and for the constants in the system, we use

$$I = 1 \qquad R = 1 \qquad \Omega = 1 \qquad m = 1.$$

The global error for the $(1,1)$ and $(2,2)$–Gauss-Lobatto SPARK and EMPRK methods are given in Figures 8.8 and 8.9, respectively. Once again, the convergence results for this problem for the SPARK methods agree with theory.

### 8.3.4 The Seven Body Mechanism

We consider here a mechanical system which consists of seven rigid bodies connected without friction. This has become a common test problem for constrained differential equations in the literature. Although we skip some of the details of this

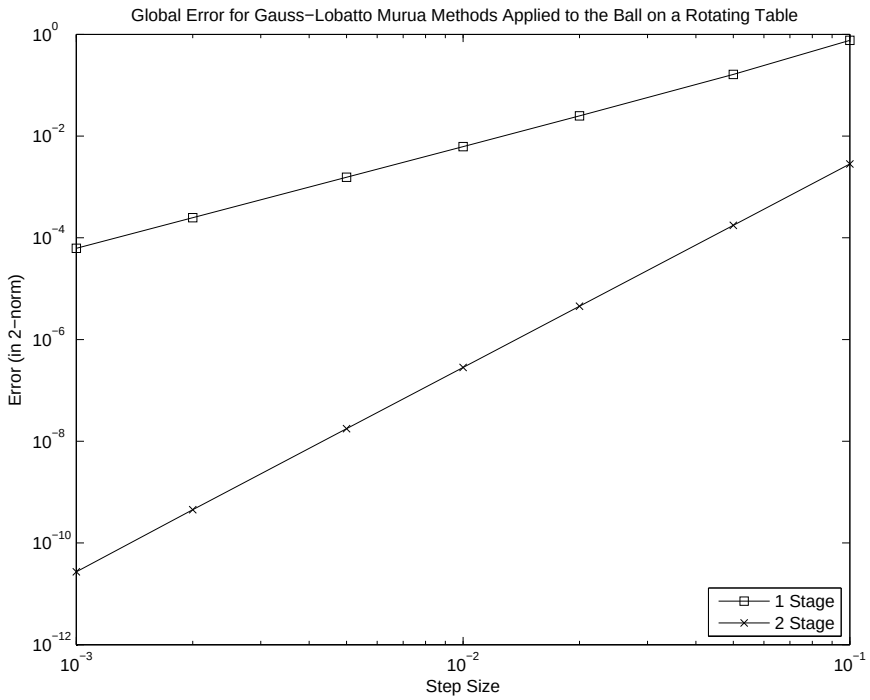


Figure 8.9: Global Error for the $q$ component for the $(1,1)$ and $(2,2)$–Gauss-Lobatto EMPRK methods applied to the ball on a rotating table problem

system, it is considered in greater detail in, for example, [10, Section VII.7]. The system has seven equations with six holonomic constraints. We use for as the position coordinates for the system the variables

$$q_1 = \beta, \quad q_2 = \Theta, \quad q_3 = \gamma, \quad q_4 = \Psi, \quad q_5 = \delta, \quad q_6 = \Omega, \quad q_7 = \epsilon.$$

Following [14], the gradient of the Lagrangian with respect to $q$ for the system is given by

$$\begin{bmatrix} MOM \\ m_2 \cdot da \cdot rr \cdot v_1(v_1 + v_2) \cdot \sin(q_2) \\ FX \cdot (sc \cdot \cos(q_3) - sd \cdot \sin(q_3)) + FY \cdot (sd \cdot \cos(q_3) + sc \cdot \sin(q_3)) \\ m_4 \cdot zt \cdot (e - ea) \cdot v_5 \cdot (v_5 + v_4) \cdot \cos(q_4) \\ 0.0 \\ -m_6 \cdot u \cdot (zf - fa) \cdot v_7 \cdot (v_7 + v_6) \cdot \cos(q_6) \\ 0.0 \end{bmatrix},$$

with the additional definitions

$$\begin{aligned} xd &= sd \cdot \cos(q_3) + sc \cdot \sin(q_3) + xb \\ yd &= sd \cdot \sin(q_3) - sc \cdot \cos(q_3) + yb \\ LANG &= \sqrt{(xd - xc)^2 + (yd - yc)^2} \\ FORCE &= -c_0 \cdot (LANG - l_0)/LANG \\ FX &= FORCE \cdot (xd - xc) \\ FY &= FORCE \cdot (yd - yc). \end{aligned}$$

The gradient of the Lagrangian with respect to $v$ is given by the product $M(q) \cdot v$, with the nonzero entries of $M(q)$

$$\begin{aligned} M_{11}(q) &= m_1 \cdot ra^2 + m_2 \cdot \left(rr^2 - 2 \cdot da \cdot rr \cdot \cos(q_2) + da^2\right) + I_1 + I_2 \\ M_{21}(q) &= m_2 \cdot (da^2 - da \cdot rr \cdot \cos(q_2)) + I_2 \\ M_{22}(q) &= m_2 \cdot da^2 + I_2 \\ M_{33}(q) &= m_3 \cdot (sa^2 + sb^2) + I_3 \\ M_{44}(q) &= m_4 \cdot (e - ea)^2 + I_4 \\ M_{54}(q) &= m_4 \cdot ((e - ea)^2 + zt \cdot (e - ea) \cdot \sin(q_4)) + I_4 \end{aligned}$$

| | | | |
|---|---|---|---|
| $d = .028$ | $da = .0115$ | $e = .02$ | $xa = -.06934$ |
| $ea = .01421$ | $zf = .02$ | $fa = .01421$ | $ya = -.00227$ |
| $rr = .007$ | $ra = .00092$ | $ss = .035$ | $xb = -.03635$ |
| $sa = .01874$ | $sb = .01043$ | $sc = .018$ | $yb = .03273$ |
| $sd = .02$ | $zt = .04$ | $ta = .02308$ | $xc = .014$ |
| $tb = .00916$ | $u = .04$ | $ua = .01228$ | $yc = .072$ |
| $ub = .00449$ | $c_0 = 4530$ | $l_0 = .07785$ | |
| $m_1 = .04325$ | $m_2 = .00365$ | $m_3 = .02373$ | $m_4 = .00706$ |
| $m_5 = .07050$ | $m_6 = .00706$ | $m_7 = .05498$ | $I_1 = 2.194 \cdot 10^{-6}$ |
| $I_2 = 4.410 \cdot 10^{-7}$ | $I_3 = 5.255 \cdot 10^{-6}$ | $I_4 = 5.667 \cdot 10^{-7}$ | $I_5 = 1.169 \cdot 10^{-5}$ |
| $I_6 = 5.667 \cdot 10^{-7}$ | $I_7 = 1.912 \cdot 10^{-5}$ | | |

Table 8.3: Coefficients for the seven body mechanism

$$
\begin{aligned}
M_{55}(q) &= m_4 \cdot (zt^2 + 2 \cdot zt \cdot (e - ea) \cdot \sin(q_4) + (e - ea)^2) \\
&\qquad + m_5 \cdot (ta^2 + tb^2) + I_4 + I_5 \\
M_{66}(q) &= m_6 \cdot (zf - fa)^2 + I_6 \\
M_{76}(q) &= m_6 \cdot ((zf - fa)^2 - u \cdot (zf - fa) \cdot \sin(q_6)) + I_6 \\
M_{77}(q) &= m_6 \cdot ((zf - fa)^2 - 2 \cdot u \cdot (zf - fa) \cdot \sin(q_6) + u^2) \\
&\qquad + m_7 \cdot (ua^2 + ub^2) + I_6 + I_7.
\end{aligned}
$$

We use the same initial conditions and constant values as in [10, Section VII.7] and [14], with a constant drive torque $MOM = .033$. These are given in Table 8.3. The global convergence for the SPARK methods are given in Figure 8.10. We also give the energy error for two different stepsizes in Figures 8.11 and 8.12. The exact solution to the seven body problem preserves the total energy of the system. In general, these methods applied to problems with holonomic constraints are symplectic, and thus preserve the total energy of the systems. However, the

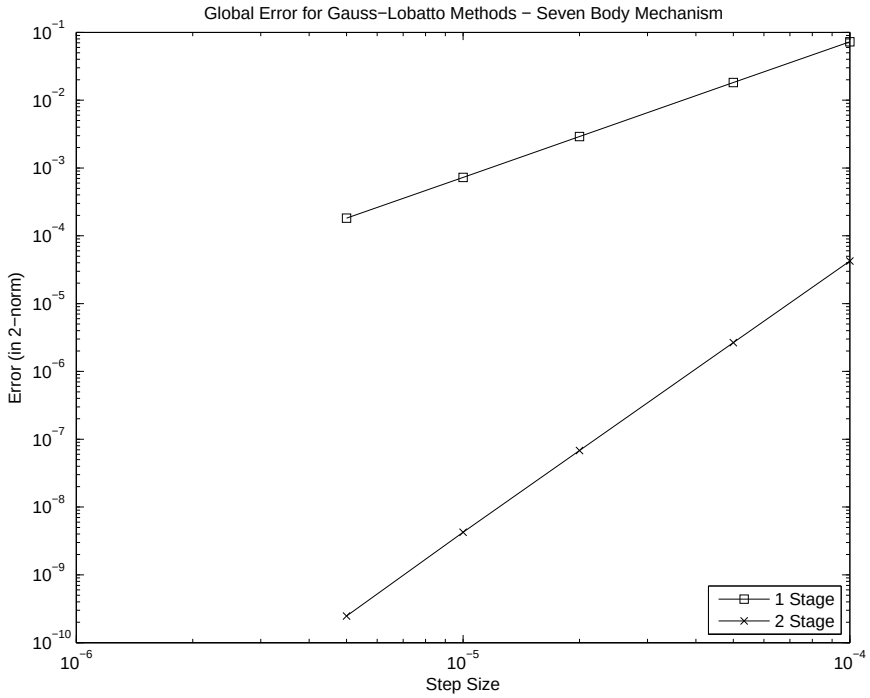


Figure 8.10: Global Error for the $q$ component for the $(1,1)$ and $(2,2)$–Gauss-Lobatto SPARK methods applied to the seven body mechanism

kinetic and potential energies here are unbounded, and we see that the SPARK method does not preserve the total energy of the system. This matches the findings of [14].

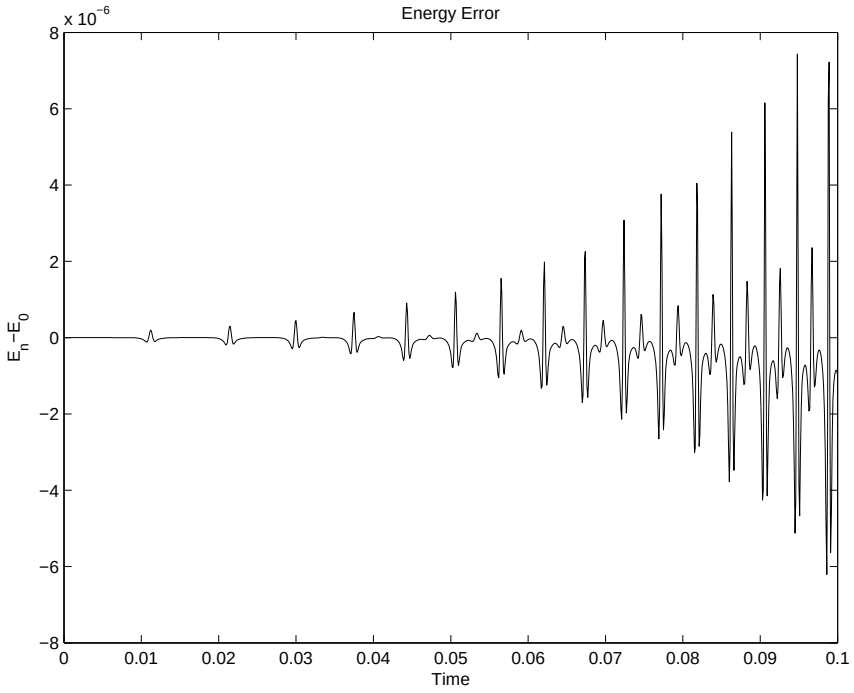


Figure 8.11: Energy error for the $(2,2)$–Gauss-Lobatto SPARK method applied to the seven body mechanism with $h = .0001$

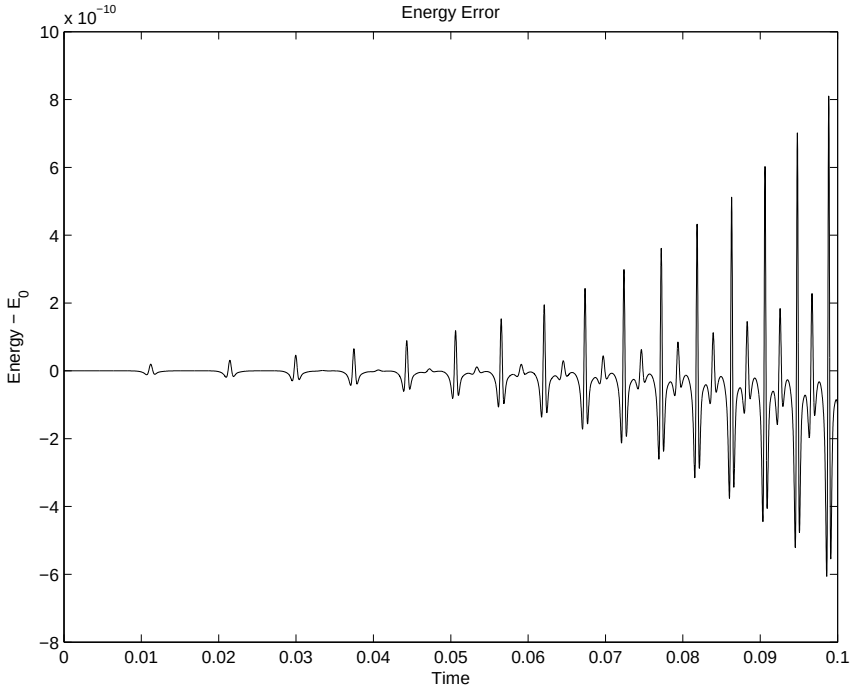


Figure 8.12: Energy error for the $(2,2)$–Gauss-Lobatto SPARK method applied to the seven body mechanism with $h = 10^{-5}$

# CHAPTER 9
# CONCLUSION

## 9.1 Introduction

The goal of this thesis was to present an analysis for some numerical solvers for certain classes of DAEs having a particular structure. We summarize the achievements of each chapter, and give considerations for future work.

## 9.2 Summary of the Results

For SPARK methods applied to index 2 DAEs, the main result is the equivalence of the Gauss methods to a class of discontinuous collocation methods. Combining this with an examination of the influence of perturbations, the Gauss SPARK methods are shown to have order of convergence $2s$. This is an alternative proof to that shown by Jay in [15].

Gauss-Lobatto SPARK methods applied to originally index 3 DAEs are shown to be equivalent to a class of discontinuous collocation methods. This, along with an analysis of perturbations, gives a proof that the local error of the methods is of order $2s$. The results presented in this thesis correct the proof presented in [16], which erroneously states that these methods are equivalent to a class of (continuous) collocation methods. Symplectic SPARK methods applied to mechanical systems with time independent total energy and holonomic constraints are shown to preserve the total energy of the system well.

Two types of numerical solutions for solving systems with mixed index 2 and 3 constraints are presented. With the help of the results for index 2 and index 3 constraints, the order of the Gauss-Lobatto SPARK methods is shown to be $2s$. We also gave an extension to the method Murua presented in [21] for problems with index 2 constraints to mixed index 2 and 3 DAEs. The existence and uniqueness of these methods is shown. Each of these two types of methods (SPARK and

EMPRK) for mixed index problems are a type of Lagrange-d'Alembert integrator in the sense of [18]. This approach views a constrained Lagrangian as a forced Lagrangian system with constraints seen as invariants.

We have also presented a few numerical examples of the 1-stage and 2-stage Gauss-Lobatto SPARK and EMPRK methods. In every example, we observed the predicted order of $2s$. Further, in problems with a time independent Hamiltonian, we saw that the energy of these methods was preserved well over long time intervals. This can be explained by the backward error analysis presented in Chapter 4 for originally index 3 DAEs.

## 9.3 Future Work

From a practical standpoint, much work remains for the numerical methods explored in this thesis.

- More experimentation with these methods should be considered in the future. The need for efficient solvers in complicated systems is a major motivation of numerical analysis and scientific computing. This thesis considered only relatively simple problems for experiments. Further real world examples could be considered.

- An efficient implementation of these methods could be developed. All code written for this thesis used *MATLAB*. A faster programming language (such as C or FORTRAN) could be used for the implementation.

- Practical error estimation can be considered for the methods presented. This would allow for automatic stepsize control.

There is also theoretical work directly related to the results presented in this thesis that remains.

- The discontinuous collocation method proposed in Chapter 5 has an obvious extension. For example, in (5.60a), we have used a "continuous" collocation

for the $Y$ polynomial. A discontinuous collocation method could also be used here. This extended discontinuous collocation method could possibly allow for other coefficients to be used, such as the Lobatto or Radau coefficients. To fit with this extension, the Gauss coefficients would need to be extended by taking $b_0 = 0$, $b_{s+1} = 0$, with the first and last rows and columns of $A \in \mathbb{R}^{(s+2)\times(s+2)}$ taken as 0. This extension could work with the SPARK methods, and presumably with the EMPRK methods as well.

- The Gauss-Lobatto SPARK and EMPRK methods applied to systems with nonholonomic constraints appear to conserve the total energy of the system well (see [11]). Showing this theoretically is a topic for future research. Unlike for holonomic mechanical systems, the flow of a nonholonomic system is not a symplectic mapping. Thus the theory of generating functions used in this thesis and used elsewhere for other classes of methods (see for example [6]) does not apply. It is possible that the modified equation of a nonholonomically constrained Hamiltonian system is not necessarily a Hamiltonian system, and that rather the energy preservation comes from the existence of an invariant to the modified equation.

## REFERENCES


[1] A.M. Bloch. *Nonholonomic Mechanics and Control*. Springer Science and Business Media, LLC, 2003.

[2] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. SIAM, Second edition, 1996.

[3] C. Cortés and S. Martínez. Non-holonomic integrators. *Nonlinearity*, 14:1365–1392, 2001.

[4] I.M. Gelfand and S.V. Fomin. *Calculus of Variations*. Dover Publications, Inc., 1991.

[5] E. Hairer. Backward analysis of numerical integrators and symplectic methods. In *Stiff and Differential-Algebraic Problems*. Springer-Verlag, 1994.

[6] E. Hairer. Global modified Hamiltonian for constrained symplectic integrators. *Numer. Math.*, 95:325–336, 2003.

[7] E. Hairer, C. Lubich, and M Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Springer-Verlag Berlin Heidelberg, 1989.

[8] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*. Springer-Verlag Berlin Heidelberg, 2006.

[9] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I, Nonstiff Problems*. Springer New York, 2000.

[10] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II, Stiff and Differential - Algebraic Problems*. Springer-Verlag New York, 2006.

[11] L.O. Jay. Lagrange-d'Alembert integrators for nonholonomic systems. Submitted.

[12] L.O. Jay. Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2. *BIT*, 33:137–150, 1993.

[13] L.O. Jay. Symplectic partitioned Runge-Kutta methods for constrained Hamiltonian systems. *SIAM Journal on Numerical Analysis*, 33(1), 1996.

[14] L.O. Jay. Structure preservation for constrained dynamics with super partitioned additive Runge Kutta methods. *SIAM J. Sci. Comput.*, 20:416–446, 1998.

[15] L.O. Jay. Specialized Runge-Kutta methods for index 2 differential-algebraic equations. *Mathematics of Computation*, 75(254):641–654, 2006.

[16] L.O. Jay. Specialized partitioned additive Runge-Kutta methods for systems of overdetermined DAEs with holonomic constraints. *SIAM Journal on Numerical Analysis*, 45(5):1814–1842, 2007.

[17] C. Kane, J.E. Marsden, M. Ortiz, and M. West. Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems. *Int. J. Numer. Meth. Engng.*, 49:1295–1325, 2000.

[18] J.E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica*, pages 357–514, 2001.

[19] D. Meiss. *Differential Dynamical Systems*. SIAM, 2007.

[20] J.C. Monforte. *Geometric, Control and Numerical Aspects of Nonholonomic Systems*. Springer-Verlag Berlin Heidelberg, 2002.

[21] A. Murua. Partitioned Runge-Kutta methods for semi-explicit differential-algebraic systems of index 2. Unpublished, 1996.

[22] H. Oh. *SPARK Methods for Mixed DAEs of Index 2 and 3 and Their Application in Mechanics*. PhD thesis, University of Iowa, 2005.

[23] L. Petzold. Differential/algebraic equations are not ODE's. *SIAM J. Sci. Stat. Computing*, 3, 1982.

[24] P.J. Rabier and W.C. Rheinboldt. *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*. SIAM, 1987.

[25] S. Reich. Backward error analysis for numerical integrators. *SIAM J. Numer. Anal.*, 36(2), 1999.

[26] S.T. Thornton and J.B. Marion. *Classical Dynamics of Particles and Systems*. Brooks Cole, Fifth edition, 2003.