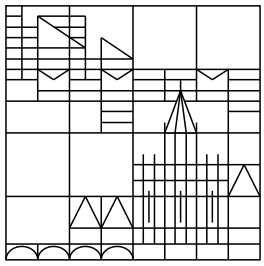

Universität Konstanz

Projected Runge-Kutta methods for differential algebraic equations of index 3

Johannes Schropp

Konstanzer Schriften in Mathematik und Informatik
Nr. 191, Juni 2003

ISSN 1430–3558

Universität Konstanz
Fach D 188, 78457 Konstanz, Germany
Email: preprints@informatik.uni–konstanz.de
WWW: http://www.informatik.uni–konstanz.de/Schriften/

Projected Runge-Kutta methods for differential algebraic equations of index 3

Johannes Schropp

FB Mathematik, Universität zu Köln,

D-50931 Köln, Germany

E-mail: jschropp@math.uni-koeln.de

Abstract

In the present paper we introduce a new class of methods, *Projected Runge-Kutta methods*, for the solution of index 3 differential algebraic equations (DAEs) in Hessenberg form. The methods admit the integration of index 3 DAEs without any drift effects. This makes them particularly well suited for long term integration. Finally, implemented on the basis of the Radau5 code, the projected Runge-Kutta method admits larger step sizes for a prescribed tolerance than the corresponding classical scheme without projection.

Keywords. Differential algebraic systems, projected Runge-Kutta methods
AMS classification. 34C05, 34C40, 65L05

1 Introduction

Many problems of scientific interest occur naturally as an index 3 DAE, i.e., mechanical systems with constraints on the position level. Obviously much attention has recently been devoted to the development of appropriate numerical methods. Problems in the numerical treatment of index k DAEs ($k \geq 2$) arise from the fact that a solution satisfies, apart from the explicitely statet constraint, $(k-1)$-additional conditions the so-called hidden constraints.

A challenging task in Numerical Analysis is to design numerical schemes satisfying as many of these constraints. Well known are numerical schemes, i.e., stiffly accurate or projected Runge-Kutta methods solving index 2 problems satisfactory. Other problems of higher index can be brought in this form by differentiating the constraints. The original constraints remain in the system and are enforced by means of additional Lagrange multipliers (see, e.g., Gear, Gupta and Leimkuhler [6]). This approach is only convenient if

the constraint derivative is directly available. But even in that case the efficiency of the applied code decreases due to the enlargement of the system.

To overcome these difficulties we will present a different method. We combine classical Runge-Kutta schemes with projection techniques. The Runge-Kutta iteration works as a predictor and the projection step back to the constraint manifold as a corrector. This approach was introduced first by Ascher and Petzold [1] for index 2 DAEs in Hessenberg form. Here we generalize their methods to index 3 DAEs. An advantage of this strategy is that the number of variables of the original problems remains untouched.

The rest of the paper is organized as follows. In section 2 we introduce the projected implicit Runge-Kutta methods for index 3 DAEs and present the corresponding convergence theorem. In section 3 the reader will find the proof of the convergence statements. It will be shown that the projected Runge-Kutta method inherits the order of convergence from the underlying classical Runge-Kutta scheme applied to the index 3 DAE, i.e., the projection step does not change the order of convergence. The last section is devoted to numerical computations. Here we have implemented a variable step size projected Runge-Kutta method for index 3 DAEs on the basis of the excellent and widely spread code Radau5 designed by Hairer and Wanner (see, [9] p.566-574). Finally we choose two examples from real applications and compare the resulting code with the classical Radau5 software.

A forthcoming task is to analyze the geometric properties of the projected index 3 Runge-Kutta methods. Of particular interest is the relation between the discrete state and control variables as well as the discrete phase portrait near equilibria, periodic orbits and attracting sets as it has been worked out for the projected Runge-Kutta methods applied to index 2 DAEs (see, e.g., Schropp [12], [13], [14]) or classical Runge-Kutta schemes applied to ordinary differential equations (compare Garay [5], Beyn [2], [3] and Kloeden and Lorenz [11]).

2 Projected Runge-Kutta methods

We consider the DAE

$$\begin{aligned} \dot{u} &= f(u,v), \quad u(0) = u_0, \\ \dot{v} &= k(u,v,\lambda), \quad v(0) = v_0, \\ 0 &= g(u), \quad \lambda(0) = \lambda_0, \end{aligned} \tag{2.1}$$

$u \in \mathbb{R}^N$, $v \in \mathbb{R}^M$ and $\lambda \in \mathbb{R}^l$ in Hessenberg form. Let C_b^ν denote the space of functions of class C^ν with bounded derivatives up to order ν. We make the following assumptions.

(A1) $f \in C_b^{\nu+1}(\mathbb{R}^{N+M}, \mathbb{R}^N)$, $k \in C_b^\nu(\mathbb{R}^{N+M+l}, \mathbb{R}^M)$, $g \in C_b^{\nu+2}(\mathbb{R}^N, \mathbb{R}^l)$ for ν sufficiently big.

(A2) There is a C_b^ν-function ψ_0 satisfying $Dg^2(u)(f(u,v),f(u,v))+Dg(u)\frac{\partial f}{\partial u}(u,v)f(u,v)+Dg(u)\frac{\partial f}{\partial v}(u,v)k(u,v,\psi_0(u,v))=0$ for $(u,v)\in D_\tau := \{(u,v)\in\mathbb{R}^{N+M} \mid \max(\|\,g(u)\,\|, \|\,Dg(u)f(u,v)\,\|)<\tau\}$, $\tau>0$.

(A3) $Dg(u)\frac{\partial f}{\partial v}(u,v)\frac{\partial k}{\partial \lambda}(u,v,\psi_0(u,v))$ is invertible for $(u,v)\in D_\tau$ and the inverse has bounded norm.

In particular, problem (2.1) is of index 3 and consistent initial values (u_0,v_0,λ_0) for (2.1) must satisfy $g(u_0)=0$, $Dg(u_0)f(u_0,v_0)=0$ and $Dg^2(u_0)(f(u_0,v_0),f(u_0,v_0))+Dg(u_0)\frac{\partial f}{\partial u}(u_0,v_0)f(u_0,v_0)+Dg(u_0)\frac{\partial f}{\partial v}(u_0,v_0)k(u_0,v_0,\lambda_0)=0$. Additionally, (A3) says that $Dg(u)$, $Dg(u)\frac{\partial f}{\partial v}(u,v)$ are of full rank. Thus the solutions of the equations $g(u)=0$, $Dg(u)f(u,v)=0$ define the $(N+M-2l)$-dimensional submanifold

$$S_0 \;:=\; \{(u,v)\in\mathbb{R}^{N+M} \mid\; g(u)=0,\; Dg(u)f(u,v)=0\} \tag{2.2}$$

of $\mathbb{R}^{N+M}$ and the underlying index 0 ODE reads

$$\begin{aligned} \dot u &= f(u,v),\;\; u(0)=u_0 \\ \dot v &= k(u,v,\psi_0(u,v)),\;\; v(0)=v_0,\quad (u_0,v_0)\in S_0 \end{aligned} \tag{2.3}$$

(for an illustration of S_0 and the dynamics on it, see Hairer, Wanner [9], p.458). We denote the solution flow of (2.3) with $(\bar u(t,u_0,v_0),\bar v(t,u_0,v_0))$, $(u_0,v_0)\in S_0$. Then, (A2) implies the flow $(\bar u(t,u_0,v_0),\bar v(t,u_0,v_0),\bar\lambda(t,u_0,v_0))$, $\bar\lambda(t,u_0,v_0)=\psi_0(\bar u(t,u_0,v_0),\bar v(t,u_0,v_0))$ for equation (2.1).

We are interested in the qualitative, geometric features of s-stage Runge-Kutta type methods with Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array},\;\; A=(a_{ij})_{1\le i,j\le s}\in\mathbb{R}^{s,s},\; b,c\in\mathbb{R}^s \tag{2.4}$$

and constant step size Δt when applied to (2.1). The Runge-Kutta method possesses stage order q, if

$$\sum_{j=1}^{s} a_{ij}c_j^{k-1} \;=\; \frac{c_i^k}{k},\; k=1,\dots,q,\; i=1,\dots,s.$$

To eliminate drift problems in the discrete long time run we are interested in Runge-Kutta type methods whose iterates retain the first and second order constraints $g(u)=0$, $Dg(u)f(u,v)=0$ exactly. To that purpose we combine classical Runge-Kutta methods with projection techniques. This approach was first proposed by Ascher and Petzold [1] for index 2 DAEs in Hessenberg form.
For the Butcher tableau of the projected Runge-Kutta method we impose the conditions:

(B1) The Runge-Kutta matrix A is invertible.

(B2) $R(\infty) = 1 - b^T A^{-1}\mathbb{I}$, $\mathbb{I} = (1, \ldots, 1)$ satisfies $|R(\infty)| < 1$.

(B3) The method is of classical order p and possesses stage order q with $p \geq q + 1$ and $q \geq 2$.

Applied to equation (2.1) the proposed projected Runge-Kutta method with step size Δt is a combination of a Runge-Kutta step with a projection. We denote the projected Runge-Kutta iterates at time $t_n = n\Delta t$ shortly by (u_n, v_n, λ_n) or more precisely

$$(\hat{u}(n\Delta t, u_0, v_0, \lambda_0), \hat{v}(n\Delta t, u_0, v_0, \lambda_0), \hat{\lambda}(n\Delta t, u_0, v_0, \lambda_0)),$$

if the dependence of the initial value (u_0, v_0, λ_0) is emphasized. The Runge-Kutta step has the form

$$\begin{aligned} \tilde{u}_{n+1} &= u_n + \Delta t(b^T \otimes I)\bar{f}(U^n, V^n), \\ \tilde{v}_{n+1} &= v_n + \Delta t(b^T \otimes I)\bar{k}(U^n, V^n, \Lambda^n), \\ \lambda_{n+1} &= (1 - b^T A^{-1}\mathbb{I})\lambda_n + (b^T A^{-1} \otimes I)\Lambda^n \end{aligned} \tag{2.5}$$

where $U^n = (U_1^n, \ldots, U_s^n) \in \mathbb{R}^{Ns}$, $V^n = (V_1^n, \ldots, V_s^n) \in \mathbb{R}^{Ms}$, $\Lambda^n = (\Lambda_1^n, \ldots, \Lambda_s^n) \in \mathbb{R}^{ls}$ denote the solution of the algebraic system

$$\begin{aligned} U - (\mathbb{I} \otimes u_n) &= \Delta t(A \otimes I)\bar{f}(U, V), \\ V - (\mathbb{I} \otimes v_n) &= \Delta t(A \otimes I)\bar{k}(U, V, \Lambda), \\ 0 &= \bar{g}(U) \end{aligned} \tag{2.6}$$

and the functions $\bar{f}, \bar{k}, \bar{g}$ stand for $\bar{f}(U^n, \Lambda^n) = (f(U_1^n, \Lambda_1^n), \ldots, f(U_s^n, \Lambda_s^n))$, $\bar{k}(U^n, V^n, \Lambda^n) = (k(U_1^n, V_1^n, \Lambda_1^n), \ldots, k(U_s^n, V_s^n, \Lambda_s^n))$, $\bar{g}(U^n) = (g(U_1^n), \ldots, g(U_s^n))$.
Finally, the projection step

$$\begin{aligned} u_{n+1} &= \tilde{u}_{n+1} + \frac{\partial f}{\partial v}(u_{n+1}, v_{n+1})\frac{\partial k}{\partial \lambda}(u_{n+1}, v_{n+1}, \lambda_{n+1})\mu_1, \\ v_{n+1} &= \tilde{v}_{n+1} + \frac{\partial k}{\partial \lambda}(u_{n+1}, v_{n+1}, \lambda_{n+1})\frac{\mu_2}{\Delta t}, \\ 0 &= g(u_{n+1}) \\ 0 &= \Delta t\, Dg(u_{n+1})f(u_{n+1}, v_{n+1}) \end{aligned} \tag{2.7}$$

determines u_{n+1} and v_{n+1}. In (2.7) the variables μ_1, μ_2 are needed for the projection only.

A Runge-Kutta method satisfying $a_{sj} = b_j$, $j = 1, \ldots, s$ is called stiffly accurate. Stiffly accurate Runge-Kutta solutions satisfy the first order constraint $g(u) = 0$. In this case we obtain $u_{n+1} = \tilde{u}_{n+1}$, $\mu_1 = 0$ in (2.7) and the projection step can be reduced to

$$\begin{aligned} v_{n+1} &= \tilde{v}_{n+1} + \frac{\partial k}{\partial \lambda}(u_{n+1}, v_{n+1}, \lambda_{n+1})\frac{\mu_2}{\Delta t}, \\ 0 &= \Delta t\, Dg(u_{n+1})f(u_{n+1}, v_{n+1}). \end{aligned} \tag{2.8}$$

The convergence properties of the proposed scheme are characterized in

Theorem 2.1 *Consider the DAE (2.1) and assume (A1)-(A3). Let (u_n, v_n, λ_n) denote the sequences generated with a projected Runge-Kutta method satisfying (B1)-(B3), when applied to (2.1) with consistent initial values (u_0, v_0, λ_0), $(u_0, v_0) \in S_0$, $\lambda_0 = \psi_0(u_0, v_0)$. Then with $\gamma(u_0, v_0) := (u_0, v_0, \psi_0(u_0, v_0))$ we have the following estimates for the global error. There exists a positive constant Δt_0 such that for $0 < \Delta t \leq \Delta t_0$*

$$\begin{aligned} \hat{u}(n\Delta t, \gamma(u_0, v_0)) - \bar{u}(n\Delta t, u_0, v_0) &= O(\Delta t^q), \\ \hat{v}(n\Delta t, \gamma(u_0, v_0)) - \bar{v}(n\Delta t, u_0, v_0) &= O(\Delta t^q), \\ \hat{\lambda}(n\Delta t, \gamma(u_0, v_0)) - \bar{\lambda}(n\Delta t, u_0, v_0) &= O(\Delta t^{q-1}) \ \text{ for } n \in \mathbb{N} \ \text{ with } \ 0 \leq n\Delta t \leq t_{end}. \end{aligned} \tag{2.9}$$

Remark: Hairer, Lubich and Roche (see Theorem 6.4 in [7]) have shown the corresponding statement to (2.9) including convergence order for the classical Runge-Kutta scheme. Thus, the accuracy of the projected Runge-Kutta iterates (2.5)-(2.7) is untouched by the projection step. In addition, for $p \geq q + 2$, $q \geq 2$ in (B3) the consistency and the convergence order in the u-component can be improved by one to $q + 1$ in (2.9).

3 Proofs: differential algebraic equations

In this section we will give a proof of Theorem 2.1. In the process of proving Theorem 2.1 refined stability inequalities distinguishing the variables u, v and λ are needed. To that purpose we work with the concept of vectornorms.
A functional $|\cdot|: W \to \mathbb{R}^k$ on a vector space W is called a generalized norm, if

$$\begin{aligned} &|v| \geq 0, \ \ |v| = 0 \iff v = 0, \\ &|v_1 + v_2| \leq |v_1| + |v_2|, \\ &|\alpha v| = |\alpha|_{\mathbb{R}} \, |v| \end{aligned} \tag{3.1}$$

holds with the natural ordering "$\leq$" on $\mathbb{R}^k$. Here $|\cdot|_{\mathbb{R}}$ denotes the absolute value in $\mathbb{R}$. Every norm $\|\cdot\|_*$ in $\mathbb{R}^k$ defines a norm $\|\cdot\|$ in W via $\|v\| = \||v|\|_*$. The main tool to solve equations in spaces equipped with vectornorms is the following version of Banachs fixed point theorem in a ball.

Lemma 3.1 *Let $(W, |\cdot|)$ be a Banach space with generalized norm $|\cdot|$ and let $B_r(v_0) := \{v \in W \mid |v - v_0| \leq r\}$ for $r > 0$. Let the map $F : B_r(v_0) \mapsto W$ be continuously differentiable with invertible $DF(v_0)$. Moreover, for some nonnegative matrices $P, K \in \mathbb{R}^{k,k}$ we assume*

$$\begin{aligned} |DF(v_0)^{-1} z| &\leq P\,|z|, \ z \in W, \\ |(DF(v_0) - DF(v))z| &\leq K\,|z|, \ z \in W, \ v \in B_r(v_0), \\ P\,|F(v_0)| &< (I - PK)r. \end{aligned}$$

Then, the equation $F(v) = 0$ has a unique solution in $B_r(v_0)$. In addition, the matrix $I - PK$ is nonsingular and we have the stability inequality

$$|\, v - w \,| \;\leq\; (I - PK)^{-1}\, |\, DF(v_0)^{-1}(F(v) - F(w))\,| \quad \forall v, w \in B_r(v_0).$$

A proof of Lemma 3.1 can be found in Beyn, Schropp [4].

Now we present existence results for the projected Runge-Kutta schemes.

Lemma 3.2 *Let the assumptions of Theorem 2.1 hold and let $(u_0, v_0) \in S_0, \lambda_0 = \psi_0(u_0, v_0)$ be a consistent initial value for the DAE (2.1).*
Then for $0 < \Delta t \leq \Delta t_0$, $\Delta t_0 > 0$ sufficiently small the projected Runge-Kutta iterates (u_n, v_n, λ_n) exist for $n \in \mathbb{N}$. For the stages (U, V, Λ) of the projected Runge-Kutta dynamics we have with $w_0(u, v) := (\mathbb{I} \otimes u, \mathbb{I} \otimes v, \mathbb{I} \otimes \psi_0(u, v))$, $(u, v) \in D_\tau$ the inequality

$$|\, (U, V, \Lambda) - w_0(u, v)\,| \;\leq\; O(\Delta t)\,(1, 1, 1) \tag{3.2}$$

Moreover, the functions $\mu_i = \mu_i(\Delta t, u, v, \lambda)$ from the projection step (2.7) satisfy

$$\|\, \mu_i \,\| \;=\; O(\Delta t^{q+1}) \text{ for } (u, v) \in S_0,\ \|\, \lambda - \psi_0(u, v)\,\| \text{ sufficiently small.} \tag{3.3}$$

Proof: The first step of a projected Runge-Kutta method is a classical Runge-Kutta step. Following Hairer, Lubich and Roche [7], Ch. 6, we use the intermediate value theorem two times and obtain

$$\begin{aligned}
\bar g(U) &= \bar g(\mathbb{I} \otimes u) + \int_0^1 D\bar g(\mathbb{I} \otimes u + s(U - \mathbb{I} \otimes u))\, ds\ \Delta t (A \otimes I)\bar f(U, V) \\
&= \mathbb{I} \otimes g(u) + \Delta t\, A\mathbb{I} \otimes Dg(u) f(u, v) + \\
&\quad \Delta t^2 \Big(\int_0^1 \int_0^1 s D^2 \bar g(\mathbb{I} \otimes u + \sigma(U - \mathbb{I} \otimes u))\, ds d\sigma\ ((A \otimes I)\bar f(U, V), (A \otimes I)\bar f(U, V)) \\
&\quad + \int_0^1 D\bar g(\mathbb{I} \otimes u + s(U - \mathbb{I} \otimes u))\, ds\ (A \otimes I) \frac{\partial \bar f}{\partial U}(U, V)(A \otimes I)\bar f(U, V) \\
&\quad + \int_0^1 D\bar g(\mathbb{I} \otimes u + s(U - \mathbb{I} \otimes u))\, ds\ (A \otimes I) \frac{\partial \bar f}{\partial V}(U, V)(A \otimes I)\bar k(U, V, \Lambda) \Big) \\
&=: \mathbb{I} \otimes g(u) + \Delta t A\mathbb{I} \otimes Dg(u) f(u, v) + \Delta t^2 \bar h(U, V, \Lambda).
\end{aligned} \tag{3.4}$$

Using (3.4) we replace the equation $0 = \bar g(U)$ in (2.6) for $\Delta t > 0$ by

$$0 \;=\; \frac{1}{\Delta t^2} \mathbb{I} \otimes g(u) + \frac{1}{\Delta t} A\mathbb{I} \otimes Dg(u) f(u, v) + \bar h(U, V, \Lambda). \tag{3.5}$$

We prove the first statements of Lemma 3.2 by applying Lemma 3.1 to the equation

$$\begin{aligned}
T_1(\Delta t, u, v, U, V, \Lambda) &:= \begin{pmatrix} U - (\mathbb{I} \otimes u) - \Delta t (A \otimes I)\bar f(U, V) \\ V - (\mathbb{I} \otimes v) - \Delta t (A \otimes I)\bar k(U, V, \Lambda) \\ \frac{1}{\Delta t^2} \mathbb{I} \otimes g(u) + \frac{1}{\Delta t} A\mathbb{I} \otimes Dg(u) f(u, v) + \bar h(U, V, \Lambda) \end{pmatrix} \\
&= 0, \quad (u, v) \in D_\tau.
\end{aligned} \tag{3.6}$$

Let

$$M_{c_1,c_2,\Delta t} \;:=\; \{(u,v)\in\mathbb{R}^{N+M} \mid \|\, g(u)\,\| < c_1\Delta t^2,\ \|\, Dg(u)f(u,v)\,\| < c_2\Delta t\}$$

and let $(u,v)\in M_{c_1,c_2,\Delta t}\cap D_\tau$. We define $w_0(u,v) := (\mathbb{I}\otimes u, \mathbb{I}\otimes v, \mathbb{I}\otimes\psi_0(u,v))$ and calculate with $q\geq 2$ and (A2)

$$\begin{aligned}\bar{h}(w_0(u,v)) \;&=\; \frac{1}{2}(I\otimes D^2g(u))(A\mathbb{I}\otimes f(u,v), A\mathbb{I}\otimes f(u,v))\\ &\quad + A^2\mathbb{I}\otimes Dg(u)\frac{\partial f}{\partial u}(u,v)f(u,v)\\ &\quad + A^2\mathbb{I}\otimes Dg(u)\frac{\partial f}{\partial v}(u,v)k(u,v,\psi_0(u,v)) \;=\; 0.\end{aligned} \tag{3.7}$$

Introducing the generalized norm $|\,(U,V,\Lambda)\,| = (\|\,U\,\|, \|\,V\,\|, \|\,\Lambda\,\|)\in\mathbb{R}^3$ and using (3.7) we can compute

$$\begin{aligned}T_1(\Delta t,u,v,w_0(u,v)) \;&=\; \Big(O(\Delta t), O(\Delta t), \frac{1}{\Delta t^2}\mathbb{I}\otimes g(u) + \frac{1}{\Delta t}A\mathbb{I}\otimes Dg(u)f(u,v)\Big)\\ &=\; (O(\Delta t), O(\Delta t), O(c_1+c_2)) \text{ for } (u,v)\in D_\tau.\end{aligned} \tag{3.8}$$

Moreover, we obtain

$$\begin{aligned}\frac{\partial\bar{h}}{\partial\Lambda}(w_0(u,v)) \;&=\; I\otimes Dg(u)(A\otimes I)I\otimes\frac{\partial f}{\partial v}(u,v)(A\otimes I)I\otimes\frac{\partial k}{\partial\lambda}(u,v,\psi_0(u,v))\\ &=\; A^2\otimes Dg(u)\frac{\partial f}{\partial v}(u,v)\frac{\partial k}{\partial\lambda}(u,v,\psi_0(u,v))\end{aligned} \tag{3.9}$$

which is invertible bei (A3) and (B1).
For the derivative of T_1 with respect to (U,V,Λ) we find with

$$\Gamma(u,v) \;:=\; Dg(u)\frac{\partial f}{\partial v}(u,v)\frac{\partial k}{\partial\lambda}(u,v,\psi_0(u,v)) \tag{3.10}$$

the representation

$$\frac{\partial}{\partial(U,V,\Lambda)}T_1(\Delta t,u,v,w_0(u,v)) \;=\; \begin{pmatrix} I+O(\Delta t) & O(\Delta t) & 0\\ O(\Delta t) & I+O(\Delta t) & O(\Delta t)\\ O(1) & O(1) & A^2\otimes\Gamma(u,v)\end{pmatrix}.$$

By (3.9) the matrix $\frac{\partial}{\partial(U,V,\Lambda)}T_1(\Delta t,u,v,w_0(u,v))$ is invertible for $0<\Delta t\leq\Delta t_0$, $\Delta t_0>0$ sufficiently small and the inverse is of the form

$$\frac{\partial}{\partial(U,V,\Lambda)}T_1(\Delta t,u,v,w_0(u,v))^{-1} \;=\; \begin{pmatrix} I & 0 & 0\\ 0 & I & 0\\ O(1) & O(1) & (A^2\otimes\Gamma(u,v))^{-1}\end{pmatrix} + O(\Delta t).$$

In terms of vector norms this leads to $|\ \frac{\partial}{\partial(U,V,\Lambda)} T_1(\Delta t, u, v, w_0(u,v))^{-1}\ | \leq P_{\Delta t}$ with

$$P_{\Delta t} := \begin{pmatrix} 1+O(\Delta t) & O(\Delta t) & O(\Delta t) \\ O(\Delta t) & 1+O(\Delta t) & O(\Delta t) \\ O(1) & O(1) & O(1) \end{pmatrix} \in \mathbb{R}^{3,3}.$$

Then, following the lines of the proof of Lemma 4.1 in Beyn, Schropp [4] we obtain the unique solvability of (3.6) in $B_r(w_0(u,v)) := \{(U,V,\Lambda) \in \mathbb{R}^{(N+M+l)s} \mid |\ (U,V,\Lambda) - (\mathbb{I}\otimes u, \mathbb{I}\otimes v, \mathbb{I}\otimes \psi_0(u,v))\ | \leq r\}$, $r=(r_1,r_2,r_3)>0$ for $(u,v)\in M_{c_1,c_2,\Delta t}$, $0<\Delta t\leq \Delta t_0$ provided that $c_1, c_2, \Delta t_0 > 0$ are sufficiently small. We remark that an application of the implicit function theorem ensures the smooth dependency of the solution (U,V,Λ) from $(\Delta t, u, v)$. In addition, the claimed stability inequality (3.2) holds.

The second step is the projection of the classical Runge-Kutta iterates onto the constrained manifold S_0. We define the discrete time Δt step forward functions of the Runge-Kutta map

$$\begin{aligned} \tilde{u} &= \tilde{u}(\Delta t,u,v) = u + \Delta t (b^T\otimes I)\bar{f}(U(\Delta t,u,v),V(\Delta t,u,v)), \\ \tilde{v} &= \tilde{v}(\Delta t,u,v) = v + \Delta t (b^T\otimes I)\bar{k}(U(\Delta t,u,v),V(\Delta t,u,v),\Lambda(\Delta t,u,v)), \\ \tilde{\lambda} &= \tilde{\lambda}(\Delta t,u,v,\lambda) = R(\infty)\lambda + (b^T A^{-1}\otimes I)\Lambda(\Delta t,u,v) \end{aligned} \tag{3.11}$$

and consider the equation

$$\begin{aligned} T_2(\Delta t,u,v,\lambda,\hat{u},\hat{v},\mu_1,\mu_2) &= \begin{pmatrix} \hat{u} - \tilde{u}(\Delta t,u,v) - \frac{\partial f}{\partial v}(\hat{u},\hat{v})\frac{\partial k}{\partial \lambda}(\hat{u},\hat{v},\tilde{\lambda}(\Delta t,u,v,\lambda))\mu_1 \\ \hat{v} - \tilde{v}(\Delta t,u,v) - \frac{\partial k}{\partial \lambda}(\hat{u},\hat{v},\tilde{\lambda}(\Delta t,u,v,\lambda))\frac{\mu_2}{\Delta t} \\ g(\hat{u}) \\ \Delta t\, Dg(\hat{u}) f(\hat{u},\hat{v}) \end{pmatrix} \\ &= 0 \ \text{ for } 0<\Delta t<\Delta t_0,\ (u,v)\in M_{c_1,c_2,\Delta t}\cap D_\tau. \end{aligned} \tag{3.12}$$

With the central point $\tilde{z}(\Delta t,u,v) = (\tilde{u}(\Delta t,u,v),\tilde{v}(\Delta t,u,v),0,0)$ we can compute

$$T_2(\Delta t,u,v,\lambda,\tilde{z}(\Delta t,u,v)) = (0,0,g(\tilde{u}(\Delta t,u,v)),\Delta t\, Dg(\tilde{u})f(\tilde{u},\tilde{v})(\Delta t,u,v)). \tag{3.13}$$

Obviously, $g(\tilde{u}) = O(\Delta t)$ and $\Delta t Dg(\tilde{u})f(\tilde{u},\tilde{v}) = O(\Delta t)$ hold for $(u,v)\in M_{c_1,c_2,\Delta t}\cap D_\tau$. Moreover, for $(u,v)\in S_0$ we obtain

$$\begin{aligned} g(\tilde{u})(\Delta t,u,v) &= O(\Delta t^{q+1}), \\ Dg(\tilde{u})f(\tilde{u},\tilde{v})(\Delta t,u,v) &= O(\Delta t^q) \end{aligned} \tag{3.14}$$

from the local error analysis of the underlying classical Runge-Kutta map (see, e.g., Ch. 6 in Hairer, Lubich and Roche [7]).
Next, we can calculate

$$\frac{\partial}{\partial(\hat{u},\hat{v},\mu_1,\mu_2)} T_2(\Delta t,u,v,\lambda,\tilde{z}(\Delta t,u,v)) = \begin{pmatrix} I & \tilde{B}(\Delta t,u,v,\lambda) \\ \tilde{C}^T(\Delta t,u,v) & 0 \end{pmatrix} \tag{3.15}$$

with

$$\begin{aligned}\tilde{B}(\Delta t,u,v,\lambda) &= \begin{pmatrix} -\frac{\partial f}{\partial v}(\tilde{u},\tilde{v})\frac{\partial k}{\partial \lambda}(\tilde{u},\tilde{v},\tilde{\lambda})(\Delta t,u,v,\lambda) & 0 \\ 0 & -\frac{1}{\Delta t}\frac{\partial k}{\partial \lambda}(\tilde{u},\tilde{v},\tilde{\lambda})(\Delta t,u,v,\lambda) \end{pmatrix},\\ \tilde{C}^T(\Delta t,u,v) &= \begin{pmatrix} Dg(\tilde{u})(\Delta t,u,v) & 0 \\ O(\Delta t) & \Delta t\, Dg(\tilde{u})\frac{\partial f}{\partial v}(\tilde{u},\tilde{v})(\Delta t,u,v) \end{pmatrix}. \end{aligned} \tag{3.16}$$

The matrix $\frac{\partial}{\partial(\hat{u},\hat{v},\mu_1,\mu_2)}T_2(\Delta t,u,v,\lambda,\tilde{z}(\Delta t,u,v))$ in (3.15) is regular, if and only if its Schur-complement

$$\begin{aligned}-\tilde{C}^T(\Delta t,u,v)\tilde{B}(\Delta t,u,v,\lambda) &= \begin{pmatrix} a(\Delta t,u,v,\lambda) & 0 \\ O(\Delta t) & a(\Delta t,u,v,\lambda) \end{pmatrix}\\ a(\Delta t,u,v,\lambda) &= Dg(\tilde{u})\frac{\partial f}{\partial v}(\tilde{u},\tilde{v})\frac{\partial k}{\partial \lambda}(\tilde{u},\tilde{v},\tilde{\lambda})(\Delta t,u,v,\lambda) \end{aligned} \tag{3.17}$$

is invertible. Hence, by (A3) and with $(\tilde{\lambda}-\psi_0(\tilde{u},\tilde{v}))(\Delta t,u,v,\lambda)=\lambda-\psi_0(u,v)+O(\Delta t)$ the operator $\frac{\partial}{\partial(\hat{u},\hat{v},\mu_1,\mu_2)}T_2(\Delta t,u,v,\lambda,\tilde{z}(\Delta t,u,v))$ is invertible for Δt and $\lambda-\psi_0(u,v)$ sufficiently small.
Moreover, the equation $T_2(\Delta t,u,v,\lambda,.,.,.,.)=0$ possesses a unique solution in $B_r(z_0)$ for $r=(r_1,\ldots,r_4)>0$ appropriate and $(u,v)\in M_{c_1,c_2,\Delta t}$, $c_1,c_2,\Delta t$, $\lambda-\psi_0(u,v)$ sufficiently small. By the implicit function theorem this solution is smooth in $(\Delta t,u,v,\lambda)$. Finally, combining (3.13), (3.14) and using the stability inequality of Lemma 3.1 this yields $\mu_i(\Delta t,\gamma(u,v))=O(\Delta t^{q+1})$ for $i=1,2$, $(u,v)\in S_0$.

Now, let $(u_0,v_0)\in S_0$, $\lambda_0=\psi_0(u_0,v_0)$ be consistent initial values. We show the existence of the iterates (u_n,v_n,λ_n), $(u_n,v_n)\in S_0$ iteratively for $n\in\mathbb{N}$. Obviously we obtain (u_1,v_1,λ_1), $(u_1,v_1)\in S_0$ by solving (3.6), (3.12) with $u=u_0$, $v=v_0$, $\lambda=\lambda_0$ since $\lambda_0=\psi_0(u_0,v_0)$. To make the induction work for $n\in\mathbb{N}$ it remains to show

$$\|\,\lambda_n-\psi_0(u_n,v_n)\,\| \;\le\; \epsilon,\;\; n\in\mathbb{N},\;\; \epsilon>0 \text{ arbitrary.} \tag{3.18}$$

Let $\eta_n:=\lambda_n-\psi_0(u_n,v_n)$. The iteration scheme of the η-sequence reads

$$\begin{aligned}\eta_{n+1} &= R(\infty)\eta_n+\psi_0(u_n,v_n)-\psi_0(u_{n+1},v_{n+1})\\ &\quad +(b^TA^{-1}\otimes I)(\Lambda(\Delta t,u_n,v_n)-\mathbb{I}\otimes\psi_0(u_n,v_n))\\ &=: R(\infty)\eta_n+\beta_n,\;\; \eta_0=0 \end{aligned} \tag{3.19}$$

with $\beta_n:=\beta(\Delta t,u_n,v_n,\eta_n)=O(\Delta t)$. Here, due to the construction of the method u_{n+1} and v_{n+1} are functions of $(\Delta t,u_n,v_n,\eta_n)$. Using $|\,R(\infty)\,|<1$, the theory of difference equations yields

$$\|\,\eta_n\,\| \;\le\; \|\,\eta_0\,\|+\frac{1}{1-R(\infty)}\,\sup\{\|\,\beta_n\,\|\;|\;\; n\in\mathbb{N}\} \;=\; O(\Delta t)\;\;\forall n\in\mathbb{N} \tag{3.20}$$

and (3.18) is verified possibly after diminishing Δt. This finishes the proof of Lemma 3.2.

In order to show the convergence statements of Theorem 2.1 we have to analyze the local error and the stability properties of the projected Runge-Kutta method.

Lemma 3.3 *Suppose that the underlying Runge-Kutta method is of classical order p and stage order q with $p \geq q+1$ and $q \geq 2$. Then for $(u_0, v_0) \in S_0$, $\gamma(u_0, v_0) = (u_0, v_0, \psi_0(u_0, v_0))$ the local error $l_{err} = (l_{err,u}, l_{err,v}, l_{err,\lambda})$ is of magnitude*

$$\begin{aligned} l_{err,u} &= \bar{u}(\Delta t, u_0, v_0) - \hat{u}(\Delta t, \gamma(u_0, v_0)) = O(\Delta t^{min(q+2,p)}), \\ l_{err,v} &= \bar{v}(\Delta t, u_0, v_0) - \hat{v}(\Delta t, \gamma(u_0, v_0)) = O(\Delta t^{q+1}), \\ l_{err,\lambda} &= \bar{\lambda}(\Delta t, u_0, v_0) - \hat{\lambda}(\Delta t, \gamma(u_0, v_0)) = O(\Delta t^{q-1}), \end{aligned} \tag{3.21}$$

Proof: The local error of the projected Runge-Kutta method is strongly connected with the local error of the classical Runge-Kutta method which is, e.g., analyzed in Hairer, Lubich and Roche [7], Lemma 6.3. For the λ-component which is untouched by the projection (2.7) we obtain directly

$$\| \bar{\lambda}(\Delta t, u_0, v_0) - \hat{\lambda}(\Delta t, \gamma(u_0, v_0)) \| = O(\Delta t^{q-1})$$

from Lemma 6.3 in Hairer, Lubich and Roche [7].

To estimate the local error in the u- and v-component, we consider the equation (3.12). Evaluating (3.12) at the central point $\bar{z}(\Delta t, u_0, v_0) = (\bar{u}(\Delta t, u_0, v_0), \bar{v}(\Delta t, u_0, v_0), 0, 0)$ we can compute

$$T_2(\Delta t, \gamma(u_0, v_0), \bar{z}(\Delta t, u_0, v_0)) = ((\bar{u} - \tilde{u})(\Delta t, u_0, v_0), (\bar{v} - \tilde{v})(\Delta t, u_0, v_0), 0, 0). \tag{3.22}$$

Following the formulae (3.15)-(3.17) we can calculate

$$\frac{\partial}{\partial(\hat{u}, \hat{v}, \mu_1, \mu_2)} T_2(\Delta t, \gamma(u_0, v_0), \bar{z}(\Delta t, u_0, v_0)) = \begin{pmatrix} I & \bar{B}(\Delta t, u_0, v_0) \\ \bar{C}^T(\Delta t, u_0, v_0) & 0 \end{pmatrix} \tag{3.23}$$

with

$$\begin{aligned} \bar{B}(\Delta t, u_0, v_0) &= \begin{pmatrix} -\frac{\partial f}{\partial v}(\bar{u}, \bar{v}) \frac{\partial k}{\partial \lambda}(\bar{u}, \bar{v}, \bar{\lambda})(\Delta t, u_0, v_0) & 0 \\ 0 & -\frac{1}{\Delta t} \frac{\partial k}{\partial \lambda}(\bar{u}, \bar{v}, \bar{\lambda})(\Delta t, u_0, v_0) \end{pmatrix}, \\ \bar{C}^T(\Delta t, u_0, v_0) &= \begin{pmatrix} Dg(\bar{u})(\Delta t, u_0, v_0) & 0 \\ O(\Delta t) & \Delta t Dg(\bar{u}) \frac{\partial f}{\partial v}(\bar{u}, \bar{v})(\Delta t, u_0, v_0) \end{pmatrix}. \end{aligned} \tag{3.24}$$

Here we analyze the structure of $(\frac{\partial}{\partial(\hat{u}, \hat{v}, \mu_1, \mu_2)} T_2(\Delta t, \gamma(u_0, v_0), \bar{z}(\Delta t, u_0, v_0)))^{-1}$ in more detail. A straightforward calculation shows that with $\bar{S}(\Delta t, u, v) = -(\bar{C}^T \bar{B})(\Delta t, u, v)$ this inverse has the form

$$\left(\frac{\partial}{\partial(\hat{u}, \hat{v}, \mu_1, \mu_2)} T_2(\Delta t, \gamma(u_0, v_0), \bar{z}(\Delta t, u_0, v_0)) \right)^{-1} = \begin{pmatrix} I + \bar{B}\bar{S}^{-1}\bar{C}^T & -\bar{B}\bar{S}^{-1} \\ -\bar{S}^{-1}\bar{C}^T & \bar{S}^{-1} \end{pmatrix} (\Delta t, u_0, v_0) \tag{3.25}$$

with

$$
\begin{aligned}
(I + \bar{B}\bar{S}^{-1}\bar{C}^T)(\Delta t, u, v) &= \begin{pmatrix} a_{11}(\Delta t, u, v) & 0 \\ O(1) & a_{22}(\Delta t, u, v) \end{pmatrix}, \\
a_{11}(\Delta t, u, v) &= \left(I - \frac{\partial f}{\partial v}\frac{\partial k}{\partial \lambda}\left(Dg\frac{\partial f}{\partial v}\frac{\partial k}{\partial \lambda}\right)^{-1} Dg \right)(\Delta t, \gamma(u, v)), \quad (3.26) \\
a_{22}(\Delta t, u, v) &= \left(I - \frac{\partial k}{\partial \lambda}\left(Dg\frac{\partial f}{\partial v}\frac{\partial k}{\partial \lambda}\right)^{-1} Dg\frac{\partial f}{\partial v} \right)(\Delta t, \gamma(u, v)).
\end{aligned}
$$

In order to compute the local error of the projected Runge-Kutta method we apply the stability inequality of Lemma 3.1. Using (3.22), (3.25) and (3.26) this yields

$$
\begin{aligned}
|\, z(\Delta t, u_0, v_0) - (\hat{u}, \hat{v}, \mu_1, \mu_2)(\Delta t, \gamma(u_0, v_0))\,| \;&\le\; C\,|((DT_2)^{-1}T_2)(\Delta t, \gamma(u_0, v_0), \bar{z}(\Delta t, u_0, v_0)\,| \\
&= \left| \begin{pmatrix} (a_{11})(\bar{u} - \tilde{u})(\Delta t, u_0, v_0) \\ (O(1)(\bar{u} - \tilde{u}) + (a_{22})(\bar{v} - \tilde{v}))(\Delta t, u_0, v_0) \\ * \\ * \end{pmatrix} \right|. \qquad (3.27)
\end{aligned}
$$

We insert the relations

$$
\begin{aligned}
\|\,(\bar{u} - \tilde{u})(\Delta t, u, v)\,\| &= O(\Delta t^{q+1}), \\
\|\,(\bar{v} - \tilde{v})(\Delta t, u, v)\,\| &= O(\Delta t^{q}), \\
\|\,a_{11}(\Delta t, u, v)(\bar{u} - \tilde{u})(\Delta t, u, v)\,\| &= O(\Delta t^{min(q+2,p)}), \\
\|\,a_{22}(\Delta t, u, v)(\bar{v} - \tilde{v})(\Delta t, u, v)\,\| &= O(\Delta t^{q+1})
\end{aligned}
$$

(see Lemma 6.3 in Hairer, Lubich, Roche [7]) into (3.27) and obtain

$$
\begin{aligned}
\|\,\bar{u}(\Delta t, u_0, v_0) - \hat{u}(\Delta t, \gamma(u_0, v_0))\,\| &= O(\Delta t^{min(q+2,p)}), \\
\|\,\bar{v}(\Delta t, u_0, v_0) - \hat{v}(\Delta t, \gamma(u_0, v_0))\,\| &= O(\Delta t^{q+1}).
\end{aligned}
$$

This finishes the proof.

Next we rewrite the discrete scheme as dynamical system. With $P_1 = \hat{u}(\Delta t, \cdot, \cdot, \cdot)$, $P_2 = \hat{v}(\Delta t, \cdot, \cdot, \cdot)$ and $P_3(\Delta t, u, v, \lambda) = R(\infty)\lambda + (b^T A^{-1} \otimes I)\Lambda(\Delta t, u, v)$, $P = (P_1, P_2, P_3)$ we obtain

$$
\begin{aligned}
u_{n+1} &= P_1(\Delta t, u_n, v_n, \lambda_n), \\
v_{n+1} &= P_2(\Delta t, u_n, v_n, \lambda_n), \qquad (3.28) \\
\lambda_{n+1} &= P_3(\Delta t, u_n, v_n, \lambda_n).
\end{aligned}
$$

The dynamics of the iteration (3.28) is guided essentially by the dynamics of $(u_{n+1}, v_{n+1}) = \tilde{P}_{\Delta t}(u_n, v_n)$ with the map $\tilde{P}_{\Delta t} : M_{c_1,c_2,\Delta t} \to \mathbb{R}^{N+M}$ defined by

$$\tilde{P}_{\Delta t}(u,v) \;:=\; \begin{pmatrix} P_1(\Delta t, \gamma(u,v)) \\ P_2(\Delta t, \gamma(u,v)) \end{pmatrix} \;=\; \begin{pmatrix} \tilde{P}_{1,\Delta t}(\Delta t, u, v) \\ \tilde{P}_{2,\Delta t}(\Delta t, u, v) \end{pmatrix}. \tag{3.29}$$

Now we show the stability of the projected Runge-Kutta method.

Lemma 3.4 *Let the assumptions of Theorem 2.1 hold. Then the projected Runge-Kutta method is stable and we have for $(u_0, v_0) \in S_0$ the following estimates for the global error.*

$$\begin{aligned} \bar{u}(n\Delta t, u_0, v_0) - \hat{u}(n\Delta t, \gamma(u_0, v_0)) &= O(\Delta t^q), \\ \bar{v}(n\Delta t, u_0, v_0) - \hat{v}(n\Delta t, \gamma(u_0, v_0)) &= O(\Delta t^q), \\ \bar{\lambda}(n\Delta t, u_0, v_0) - \hat{\lambda}(n\Delta t, \gamma(u_0, v_0)) &= O(\Delta t^{q-1}) \;\; \textit{for} \;\; 0 \le n\Delta t \le t_{end} \end{aligned} \tag{3.30}$$

and $0 < \Delta t \le \Delta t_0$, $\Delta t_0 > 0$ sufficiently small.

Proof: Let $\Omega_{\Delta t} = \{j\Delta t \mid \; j = 0, \ldots, \sigma\}$, $\sigma = \sigma(\Delta t)$ with $\sigma(\Delta t)\Delta t \in [t_{end}, t_{end} + \Delta t[$ be an equidistant grid on $[0, t_{end}]$ and let $(\vec{w}, \vec{z}) = (w_0, z_0, w_1, z_1, \ldots, w_\sigma, z_\sigma) \in (M_{c_1,c_2,\Delta t})^{\Omega_{\Delta t}}$. For a grid function $(\vec{w}, \vec{z})$ we use the norm

$$\| (\vec{w}, \vec{z}) \|_\infty \;=\; \sup\{ \| (w_i, z_i) \| \mid \; i = 0, \ldots, \sigma \}.$$

We consider the operator

$$S_{\Delta t}(\vec{w}, \vec{z}) \;=\; \left(\begin{pmatrix} w_0 - u_0 \\ z_0 - v_0 \end{pmatrix}, \; \frac{1}{\Delta t} \begin{pmatrix} w_{n+1} - \tilde{P}_{1,\Delta t}(w_n, z_n) \\ z_{n+1} - \tilde{P}_{2,\Delta t}(w_n, z_n) \end{pmatrix}, \; n = 0, \ldots, \sigma - 1 \right) \tag{3.31}$$

Obviously, the equation $S_{\Delta t}(\vec{w}, \vec{z}) = 0$ is uniquely solvable for $(u_0, v_0) \in S_0$. To show the convergence of the projected Runge-Kutta scheme we will derive a stability inequality for the operator $S_{\Delta t}$. This will be done by appyling Lemma 3.1 to the equation $S_{\Delta t}(\vec{w}, \vec{z}) = 0$. The central point is

$$\begin{pmatrix} \bar{u}_{\Delta t} \\ \bar{v}_{\Delta t} \end{pmatrix} \;=\; \left(\begin{pmatrix} \bar{u}(i\Delta t, u_0, v_0) \\ \bar{v}(i\Delta t, u_0, v_0) \end{pmatrix}, \; i = 0, \ldots, \sigma \right). \tag{3.32}$$

With the abbreviation $\bar{u}_i = \bar{u}(i\Delta t, u_0, v_0)$, $\bar{v}_i = \bar{v}(i\Delta t, u_0, v_0)$, $i = 0, \ldots, \sigma$ we can compute

$$\begin{aligned} S_{\Delta t}(\bar{u}_{\Delta t}, \bar{v}_{\Delta t}) &= \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; \frac{1}{\Delta t} \begin{pmatrix} \bar{u}_{i+1} - \tilde{P}_{1,\Delta t}(\bar{u}_i, \bar{v}_i) \\ \bar{v}_{i+1} - \tilde{P}_{2,\Delta t}(\bar{u}_i, \bar{v}_i) \end{pmatrix}, \; i = 0, \ldots, \sigma - 1 \right) \\ &= \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; \frac{1}{\Delta t} \begin{pmatrix} \bar{u}_{i+1} - P_1(\Delta t, \bar{u}_i, \bar{v}_i, \bar{\lambda}_i) \\ \bar{v}_{i+1} - P_2(\Delta t, \bar{u}_i, \bar{v}_i, \bar{\lambda}_i) \end{pmatrix}, \; i = 0, \ldots, \sigma - 1 \right) \\ &= \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; \begin{pmatrix} O(\Delta t^{min(q+1, p-1)}) \\ O(\Delta t^q) \end{pmatrix}, \; i = 0, \ldots, \sigma - 1 \right) \;=\; O(\Delta t^q). \end{aligned} \tag{3.33}$$

Moreover, the linearization at the central point $(\bar{u}_{\Delta t}, \bar{v}_{\Delta t})$ reads

$$DS_{\Delta t}(\bar{u}_{\Delta t}, \bar{v}_{\Delta t}) = \begin{pmatrix} E & & & & \\ -B_{\Delta t,0} & A_{\Delta t} & & & \\ & -B_{\Delta t,1} & A_{\Delta t} & & \\ & & \ddots & \ddots & \\ & & & -B_{\Delta t,\sigma-1} & A_{\Delta t} \end{pmatrix}$$

with the matrices

$$\begin{aligned} E &= \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \in \mathbb{R}^{N+M,N+M}, \\ A_{\Delta t} &= \frac{1}{\Delta t} \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \in \mathbb{R}^{N+M,N+M}, \\ B_{\Delta t,i} &= \frac{1}{\Delta t} \begin{pmatrix} \frac{\partial \tilde{P}_{\Delta t,1}}{\partial u}(\bar{u}_i, \bar{v}_i) & \frac{\partial \tilde{P}_{\Delta t,1}}{\partial v}(\bar{u}_i, \bar{v}_i) \\ \frac{\partial \tilde{P}_{\Delta t,2}}{\partial u}(\bar{u}_i, \bar{v}_i) & \frac{\partial \tilde{P}_{\Delta t,2}}{\partial v}(\bar{u}_i, \bar{v}_i) \end{pmatrix} \in \mathbb{R}^{N+M,N+M}, \quad i = 0, \ldots, \sigma - 1. \end{aligned}$$

Using

$$\frac{\partial \tilde{P}_{\Delta t}}{\partial (u,v)}(\bar{u}_i, \bar{v}_i) = I + O(\Delta t)$$

we can deduce

$$\| A_{\Delta t}^{-1} B_{\Delta t,i} \| \leq 1 + C\Delta t \text{ for } i \in \{0, 1, \ldots, \sigma\}. \tag{3.34}$$

Now we show that $\| DS_{\Delta t}(\bar{u}_{\Delta t}, \bar{v}_{\Delta t})^{-1} \|_\infty$ remains bounded as $\Delta t \to 0$. To that purpose we consider the linear system of equations

$$DS_{\Delta t}(\bar{u}_{\Delta t}, \bar{v}_{\Delta t}) y = r, \quad y = (y_0, \ldots, y_\sigma), \; r = (r_0, \ldots, r_\sigma). \tag{3.35}$$

(3.35) is equivalent to

$$y_i = \prod_{j=1}^{i} (A_{\Delta t}^{-1} B_{\Delta t,i-j}) r_0 + \sum_{k=1}^{i} \left(\prod_{j=1}^{i-k} (A_{\Delta t}^{-1} B_{\Delta t,i-j}) \right) A_{\Delta t}^{-1} r_k, \; i = 0, \ldots, \sigma.$$

Hence we obtain

$$\begin{aligned} \| y_i \| \leq \; & \| \prod_{j=1}^{i} (A_{\Delta t}^{-1} B_{\Delta t,i-j}) \| \cdot \| r_0 \| + \sum_{k=1}^{i} \| \prod_{j=1}^{i-k} (A_{\Delta t}^{-1} B_{\Delta t,i-j}) \| \cdot \| A_{\Delta t}^{-1} r_k \|, \\ & i = 0, \ldots, \sigma. \end{aligned}$$

Using (3.34) and $\| A_{\Delta t}^{-1} \| = \Delta t$ we can compute

$$\begin{aligned} \| y_i \| &\leq (1 + C\Delta t)^i \ \| r_0 \| + \sum_{k=1}^{i} (1 + C\Delta t)^k \, \Delta t \ \| r_k \| \\ &\leq \exp(Ci\Delta t) \ \| r_0 \| + \frac{(1 + C\Delta t)^{i+1} - 1}{C} \ \| r \|_\infty \\ &\leq \exp(C(t_{end} + \Delta t_0))(1 + \frac{1}{C}) \cdot \ \| r \|_\infty, \ i = 0, \ldots, \sigma \end{aligned} \tag{3.36}$$

with $\| r \|_\infty = \ \sup\{ \| r_i \| \mid \ i = 0, \ldots, \sigma \}$. Formula (3.36) shows

$$\| DS_{\Delta t}(\bar{u}_{\Delta t}, \bar{v}_{\Delta t})^{-1} \|_\infty \ \leq \ \exp(C(t_{end} + \Delta t_0))(1 + \frac{1}{C}), \ \ 0 < \Delta t \leq \Delta t_0.$$

Moreover, the stability inequality

$$\| (\vec{u}, \vec{v}) - (\vec{w}, \vec{z}) \|_\infty \ \leq \ \tilde{C} \, \| S_{\Delta t}(\vec{u}, \vec{v}) - S_{\Delta t}(\vec{w}, \vec{z}) \|_\infty \tag{3.37}$$

holds. We apply the stability inequality (3.37) for

$$\begin{aligned} (\vec{u}, \vec{v}) &= (\bar{u}_0, \bar{v}_0, \bar{u}_1, \bar{v}_1, \ldots, \bar{u}_\sigma, \bar{v}_\sigma), \\ (\vec{w}, \vec{z}) &= (u_0, v_0, u_1, v_1, \ldots, u_\sigma, v_\sigma). \end{aligned}$$

Using $u_i = (P^i(\Delta t, \gamma(u_0, v_0)))_1$, $v_i = (P^i(\Delta t, \gamma(u_0, v_0)))_2$, $i = 0, \ldots, \sigma$ we can compute

$$\begin{aligned} S_{\Delta t}(\vec{w}, \vec{z}) &= \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{\Delta t} \begin{pmatrix} P(\Delta t, u_i, v_i, \lambda_i)_1 - \tilde{P}_{1,\Delta t}(u_i, v_i) \\ P(\Delta t, u_i, v_i, \lambda_i)_2 - \tilde{P}_{2,\Delta t}(u_i, v_i) \end{pmatrix}, i = 0, \ldots, \sigma - 1 \right) \\ &= \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{\Delta t} \begin{pmatrix} \frac{\partial P_1}{\partial \lambda}(u_i, v_i, \alpha_i)(\psi_0(u_i, v_i) - \lambda_i) \\ \frac{\partial P_2}{\partial \lambda}(u_i, v_i, \alpha_i)(\psi_0(u_i, v_i) - \lambda_i) \end{pmatrix}, i = 0, \ldots, \sigma - 1 \right). \end{aligned}$$

Here the values $\alpha_i = \lambda_i + s_i \psi_0(u_i, v_i)$, $s_i \in [0, 1]$, $i = 0, \ldots, \sigma - 1$ arise from the application of the intermediate value theorem. Moreover, we know

$$P(\Delta t, u, v, \lambda)_i \ = \ \tilde{P}_{i,\Delta t}(u, v) + \Delta t^q f_i(\Delta t, u, v, \lambda), \ i = 1, 2 \tag{3.38}$$

with functions f_1, f_2 bounded for $\Delta t \to 0$ (compare (3.3)). Differentiation of (3.38) yields

$$\frac{\partial P_i}{\partial \lambda}(u_i, v_i, \alpha_i) \ = \ O(\Delta t^q), \ i = 0, \ldots, \sigma - 1, \ i = 1, 2$$

and with (3.20) we can derive $S_{\Delta t}(\vec{w}, \vec{z}) = O(\Delta t^q)$. Combining this with (3.33) the convergence results

$$\begin{aligned} \bar{u}(i\Delta t, u_0, v_0) - (P^i(\Delta t, \gamma(u_0, v_0)))_1 &= O(\Delta t^q), \\ \bar{v}(i\Delta t, u_0, v_0) - (P^i(\Delta t, \gamma(u_0, v_0)))_2, &= O(\Delta t^q), \ \ i = 0, \ldots, \sigma \end{aligned}$$

follow.

It remains to prove the convergence of the λ-component. Due to the definition of the projected Runge-Kutta scheme the λ-iteration reads

$$\lambda_{n+1} \;=\; R(\infty)\lambda_n + (b^T A^{-1} \otimes I)\Lambda(\Delta t, u_n, v_n) \;=\; P_3(\Delta t, u_n, v_n, \lambda_n).$$

Now let

$$W_\epsilon^{\Omega_{\Delta t}} \;=\; \{\rho = (\rho_0, \ldots, \rho_\sigma) \mid \| \rho_i - \psi_0(\bar{u}_i, \bar{v}_i) \| < \epsilon,\ i = 0, \ldots, \sigma\}.$$

For $(u_0, v_0) \in S_0$ fix we set up the operator $\Gamma_{\Delta t} : W_\epsilon^{\Omega_{\Delta t}} \to (\mathbb{R}^l)^{\Omega_{\Delta t}}$ defined via

$$\Gamma_{\Delta t}(\rho) \;:=\; (\rho_0 - \psi_0(u_0, v_0), (\rho_{n+1} - R(\infty)\rho_n - r_n), n = 0, \ldots, \sigma - 1)$$

with $r_n = (b^T A^{-1} \otimes I)\Lambda(\Delta t, (P^n(\Delta t, \gamma(u_0, v_0)))_1, (P^n(\Delta t, \gamma(u_0, v_0)))_2)$.
We have $(\lambda_0, \lambda_1, \ldots, \lambda_\sigma) \in W_\epsilon^{\Omega_{\Delta t}}$ by (3.20) and apply Lemma 3.1 to $\Gamma_{\Delta t}(\rho) = 0$. The central point is $z_{0,\Delta t} = (\psi_0(\bar{u}_0, \bar{v}_0), \ldots, \psi_0(\bar{u}_\sigma, \bar{v}_\sigma))$. Using (3.21), (3.28) we can compute

$$\begin{aligned}
\Gamma_{\Delta t}(z_{0,\Delta t}) \;&=\; (0, (\psi_0(\bar{u}_{i+1}, \bar{v}_{i+1}) - R(\infty)\psi_0(\bar{u}_i, \bar{v}_i) - (b^T A^{-1} \otimes I)\Lambda(\Delta t, \bar{u}_i, \bar{v}_i) \\
&\quad + (b^T A^{-1} \otimes I)(\Lambda(\Delta t, \bar{u}_i, \bar{v}_i) - \Lambda(\Delta t, u_i, v_i))),\ i = 0, \ldots, \sigma - 1) \\
&=\; O(\Delta t^{q-1}).
\end{aligned}$$

We compute the linearization at the central point $z_{0,\Delta t}$ and obtain

$$D\Gamma_{\Delta t}(z_{0,\Delta t}) \;=\; \begin{pmatrix}
I & & & & & \\
-R(\infty)I & I & & & & \\
 & -R(\infty)I & I & & & \\
 & & \ddots & \ddots & & \\
 & & & -R(\infty)I & I
\end{pmatrix}.$$

Finally, we estimate the norm of $\| D\Gamma_{\Delta t}(z_{0,\Delta t})^{-1} \|_\infty$. We consider

$$D\Gamma_{\Delta t}(z_{0,\Delta t})y \;=\; r,\ \ y = (y_0, \ldots, y_\sigma),\ r = (r_0, \ldots, r_\sigma). \tag{3.39}$$

(3.39) is equivalent to

$$y_i \;=\; \sum_{k=0}^{i} R(\infty)^{i-k} r_k,\ \ i = 0, \ldots, \sigma.$$

We can calculate

$$\| y_i \| \;\le\; \sum_{k=0}^{i} | R(\infty) |^{i-k} \cdot \| r \|_\infty \le\; \frac{1}{1 - | R(\infty) |} \| r \|_\infty,\ \ i = 0, \ldots \sigma.$$

This shows $\| D\Gamma_{\Delta t}(z_{0,\Delta t})^{-1} \|_\infty \le \frac{1}{1-|R(\infty)|}$. The stability inequality then serves

$$\| (\lambda_0, \ldots, \lambda_\sigma) - (\psi_0(\bar{u}_0, \bar{v}_0), \ldots, \psi_0(\bar{u}_\sigma, \bar{v}_\sigma)) \|_\infty \;\le\; C \| \Gamma_{\Delta t}(z_{0,\Delta t}) \|_\infty = \; O(\Delta t^{q-1})$$

and completes the convergence proof of the projected Runge-Kutta methods.

4 Numerical Applications

In this section we present the results of our numerical computations. The projected Radau IIa methods have been implemented on the basis of the well known excellent Radau5-code from Hairer and Wanner [9], p.566-574. Moreover, we assume the internal parameter *iwork(5), iwork(6), iwork(7)* of Radau5 standing for the dimensions of the index 1,2 and 3 variable in the underlying DAE are set correctly.

We apply the Radau5-code with and without projection step to our index 3 DAE test examples.

Example 1: The famous normalized pendulum

$$\begin{aligned} \dot{u}_1 &= v_1, \quad u_1(0) = u_{10}, \\ \dot{u}_2 &= v_2, \quad u_2(0) = u_{20}, \\ \dot{v}_1 &= -2u_1\lambda, \quad v_1(0) = v_{10}, \\ \dot{v}_2 &= -1 - 2u_2\lambda, \quad v_2(0) = v_{20}, \\ 0 &= u_1^2 + u_2^2 - 1, \quad \lambda(0) = \lambda_0. \end{aligned} \tag{4.1}$$

Here (u_1, u_2) are Cartesian coordinates and describe the motion of an infinitesimal ball of mass 1 with gravitational force 1 and pendulum length 1. Consistent initial values read

$$(u_{10}, u_{20}, v_{10}, v_{20}, \lambda_0) = (1, 0, 0, 0, 0). \tag{4.2}$$

The solution of (4.1), (4.2) is plotted in Fig. 1.

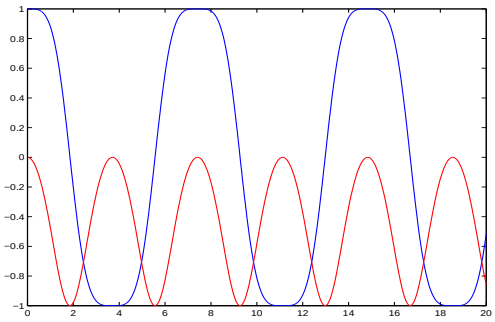

Fig. 1: $(u_1(t), u_2(t))$ for $t \in [0, 20]$

As usual we monitor the index-i deviation d_i, $i = 1, 2, 3$,

$$\begin{aligned} d_1(t) &= u_1(t)^2 + u_2(t)^2 - 1, \\ d_2(t) &= 2(u_1(t)v_1(t) + u_2(t)v_2(t)), \\ d_3(t) &= -2\lambda(t)(u_1(t)^2 + u_2(t)^2) - 2u_2(t) + 2v_1(t)^2 + 2v_2(t)^2 \end{aligned}$$

along the solution. The functions $d_1(t)$ (blue) and $d_2(t)$ (red) computed by Radau5 ($rtol = atol = 10^{-8}$) without and with projection are shown in Fig. 2.

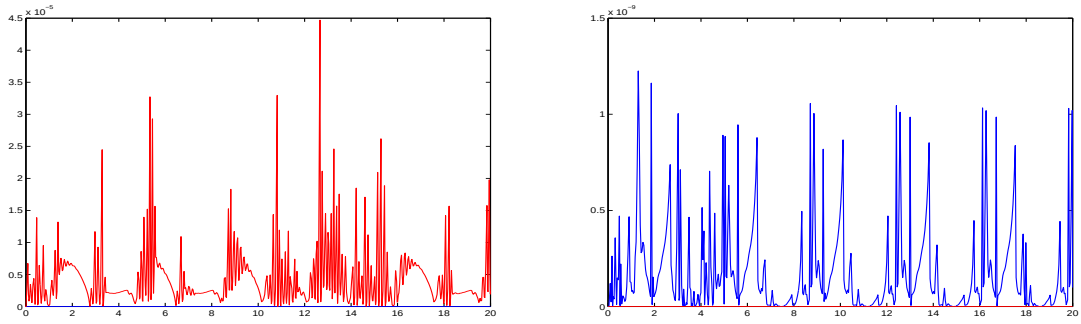

Fig. 2: $(d_1(t), d_2(t))$ without proj. (left) and with proj. (right)

Due to the structure of the problem (4.1) we able to monitor the index 3 defect d_3 too. This is shown in Fig. 3.

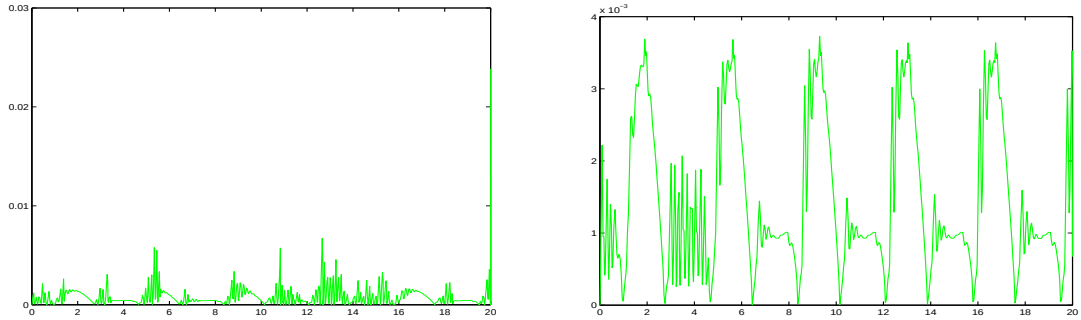

Fig. 3: $d_3(t)$ without proj. (left) and with proj. (right)

As computer independent measure for the computational effort we have chosen the number of function evaluations (fev) and evaluations of the jacobian (jacev) for various tolerances $tol = rtol = atol$.

tol	10^{-6}	10^{-8}	10^{-10}	10^{-12}
fev	2966/2580	6217/4996	12979/9963	24531/20576
jacev	276/238	570/481	1124/956	2225/1912

Tab. 1: comp. effort example 1 without/with proj.

Example 2: Andrews squeezing mechanism.

The squeezing mechanism consists of 7 rigid bodies connected by joints without friction in plane motion. A detailed description including all parameters and consistent initial values can be found in the book of Hairer and Wanner, [9], Ch. VII.7. The equations of motion have the form

$$\begin{aligned} \dot{u} &= v, \quad u(0) = u_0, \\ M(u)\dot{v} &= f(u,v) - Dg(u)^T\lambda, \quad v(0) = v_0, \\ 0 &= g(u), \quad \lambda(0) = \lambda_0 \end{aligned} \tag{4.3}$$

with a symmetric, positive definite matrix $M(u)$. The reader may notice that (4.3) is of index 3 if $Dg(u)M(u)^{-1}Dg(u)^T$ is invertible. Here $u \in \mathbb{R}^7$ is defined in angle-coordinates by $u = (\beta, \theta, \gamma, \Phi, \delta, \Omega, \epsilon)$. The position of these angles in the mechanical system can be seen the picture of the physical system shown in Hairer and Wanner, [9], p.531. Moreover, we take the consistent initial values (u_0, v_0, λ_0) from [9], p.536-537. The corresponding solution in the state variables is plotted in Fig. 4.

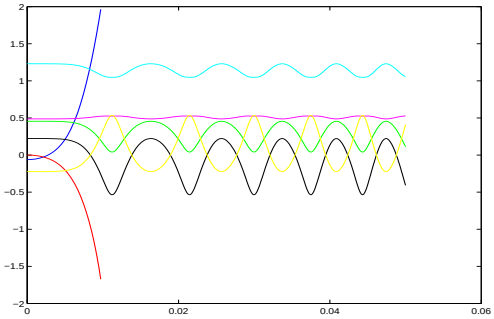

Fig. 4: $u(t) = (\beta, \theta, \gamma, \Phi, \delta, \Omega, \epsilon)(t))$ for $t \in [0, 0.05]$

Again we monitor the deviation of the index 1 constraint $d_1(t) = g(q(t))$ in blue and the index 2 constraint $d_2(t) = Dg(q(t))v(t)$ in red (see Fig. 5).

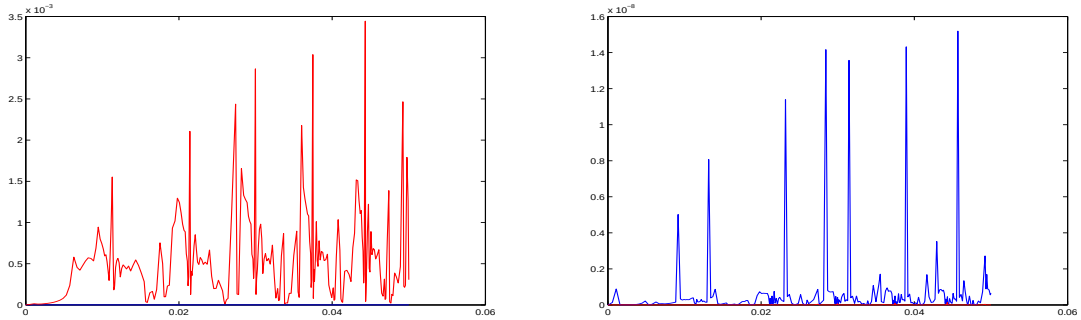

Fig. 5: $(d_1(t), d_2(t))$ without proj. (left) and with proj. (right)

Finally, the computational effort for various tolerances $tol = rtol = atol$ is summarized in Table 2.

tol	10^{-6}	10^{-8}	10^{-10}	10^{-12}
fev	2147/2073	3429/3251	6098/5760	12078/11190
jacev	138/131	241/227	472/447	984/926

Tab. 2: comp. effort example 2 without/with proj.

Our computations show that the projected index 3 Runge-Kutta method satisfies the index 1 and index 2 constraints correctly. Moreover, in combination with the excellent Radau5 code the projected 3-stage Radau IIa method of order 5 is able to gain efficiency (see Tab. 1 and 2). The reason is that the step size Δt chosen by the internal local error estimator for a prescribed tolerance $tol = rtol = atol$ increases when combining Radau5 with the propagated projection.

References

[1] Ascher, U., Petzold, L.R., *Projected Runge-Kutta methods for Differential-Algebraic Equations*, SIAM J. of Numer. Anal., **28** (1991), 1097–1120.

[2] Beyn, W.-J., *On invariant closed curves for one-step methods*, Numer. Math., **51** (1987), 103–122.

[3] Beyn, W.-J., *On the numerical approximation of phase portraits near stationary points*, SIAM J. Numer. Anal., **24** (1987), 1095–1113.

[4] Beyn, W.-J., Schropp, J., *Runge-Kutta discretizations of singularly perturbed gradient equations*, BIT Numerical Mathematics, **40** (2000), pp. 415–433.

[5] Garay, B., *Discretization and some qualitative properties of ordinary differential equations about equilibria*, Acta Math. Com. Univ., **62** (1993), 249–275.

[6] Gear, C.W., Gupta, G.K., Leimkuhler, B., *Automatic Integration of Euler-Lagrange Equations with Constraints*, J. Comp. Math., **12 & 13** (1985), pp. 77–90.

[7] Hairer, E., Lubich, Ch., Roche, M., *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, Lecture Notes in Mathematics, 1409 (1989), Springer.

[8] Hairer, E., Wanner, G., *Solving Ordinary Differential Equations I*, second edition, Springer, Heidelberg, (1993).

[9] Hairer, E., Wanner, G., *Solving Ordinary Differential Equations II*, second edition, Springer, Heidelberg, (1996).

[10] Humphries, A.R., Stuart, A.M., *Runge-Kutta methods for dissipative and gradient dynamical systems*, SIAM J. Numer. Anal., **31** (1994), 1452–1485.

[11] Kloeden, P., Lorenz, J., *Stable attracting sets in dynamical systems and in their one-step discretizations*, SIAM J. Numer. Anal., **23** (1986), 986–993.

[12] Schropp, J., *Behavior of Runge-Kutta Discretizations near Equilibria of Index 2 Differential Algebraic Systems*, Applied Numerical Mathematics, **42** (2002), p.425-435.

[13] Schropp, J., *Geometric properties of Runge-Kutta discretizations for index 2 Differential Algebraic Systems*, SIAM J. of Numer. Anal., **40** (2002), p.872-890.

[14] Schropp, J., *Attracting sets in index 2 Differential Algebraic Systems and in their Runge-Kutta discretizations*, Nonlin. Anal.: Theory, Methods & Appl., **52** (2003), p.1185-1197.