Computational Optimization and Applications (2023) 86:1299–1325
https://doi.org/10.1007/s10589-023-00484-1

Error estimates for Runge–Kutta schemes of optimal control problems with index 1 DAEs

Björn Martens[1]

Received: 17 June 2022 / Accepted: 27 March 2023 / Published online: 19 April 2023

Abstract
In this paper we derive error estimates for Runge–Kutta schemes of optimal control problems subject to index one differential–algebraic equations (DAEs). Usually, Runge–Kutta methods applied to DAEs approximate the differential and algebraic state in an analogous manner. These schemes can be considered as discretizations of the index reduced system where the algebraic equation is solved for the algebraic variable to get an explicit ordinary differential equation. However, in optimal control this approach yields discrete necessary conditions that are not consistent with the continuous necessary conditions which are essential for deriving error estimates. Therefore, we suggest to treat the algebraic variable like a control, obtaining a new type of Runge–Kutta scheme. For this method we derive consistent necessary conditions and compare the discrete and continuous systems to get error estimates up to order three for the states and control as well as the multipliers.

Keywords Optimal control · Differential–algebraic equation · Discrete approximations · Convergence analysis · Runge–Kutta schemes

Mathematics Subject Classification 49J15 · 49K15 · 49M25 · 34A09 · 65L06

1 Introduction

Direct discretization methods are often utilized to numerically solve optimal control problems because they are robust and able to solve difficult problems with state and control constraints (cf. Betts [7], Kraft [23], and von Stryk [37]). In order to justify the application of approximation schemes, an investigation of error estimates between the solutions of the continuous and discrete problem is crucial. In addition, the analysis

✉ Björn Martens
bjoern.martens@unibw.de

[1] Institute of Applied Mathematics and Scientific Computing, Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg/Munich, Germany

reveals conditions under which discretization schemes yield a solution and at which rate it converges. In process engineering, mechanical engineering, and path planning, optimal control problems subject to differential–algebraic equations (DAEs) might occur (cf. Kunkel and Mehrmann [24]). These can be solved efficiently with Runge–Kutta schemes as proposed in Gerdts [16]. However, a theoretical analysis of these discretizations applied to problems with DAEs is missing in the literature. Runge–Kutta schemes are especially important for DAEs of index 3 and higher since the Euler method does not converge in that case (cf. Brennan et al. [9]). As a first step, to reduce the gap between applications and theory, we analyze Runge–Kutta schemes applied to optimal control problems with index 1 DAEs. The knowledge gained from these investigations will then be utilized for problems with higher index DAEs.

The study of discretized optimal control problems is still a current field of research particularly in the context of DAEs. The Euler-scheme applied to nonlinear problems with ordinary differential equations (ODEs) and smooth controls has been analyzed in [8, 12, 14, 25]. Herein, Malanowski et al. [25] consider problems with mixed control-state constraints, Dontchev et al. [12] also include pure state constraints, whereas Bonnans and Festa [8] and Dontchev and Hager [14] consider problems solely with pure state constraints. Gerdts and Kunkel [17] analyze nonlinear problems with control-state constraints and controls of bounded variation. They derive error estimates of order $\frac{1}{p}$ with respect to the L_p-norm. Runge–Kutta schemes for problems with convex control constraints are examined by Dontchev et al. [13] for order 2 methods and by Hager [18] up to order 4 methods.

First order discretization methods applied to problems with bang-bang optimal control have been discussed in [1–6, 32, 35, 36]. Alt et al. [1–3, 5, 36] examine linear and linear-quadratic problems. They assume that the switching function does not have singular subarcs. Linear-quadratic problems with additional L_1-sparsity terms in the cost functional are analyzed by Alt and Schneider [4] and Schneider and Wachsmuth [35]. Alt et al. [6] and Osmolovskii and Veliov [32] study affine problems. In terms of higher order discretization schemes, these types of problems have been examined by Veliov [38] for Runge–Kutta methods, by Haunschmied et al. [21] using the stability concept of strong bi-metric regularity, and by Pietrus et al. [33] based on second order Volterra-Fliess approximations (compare also Scarinci and Veliov [34]).

Recently, using the implicit Euler-scheme, Martens and Gerdts [26–30] and Martens and Schneider [31] derived error estimates for different types of optimal control problems with DAEs. The nonlinear index 1 case was discussed in [26]. Convergence for the index 2 case was analyzed for problems with mixed control-state constraints in [28] and with pure state constraints in [30]. Linear quadratic problems and affine problems with bang-bang controls have been discussed in [29, 31], respectively.

In this paper we consider the following type of problem:

$$\text{Minimize} \quad \varphi(x(1)) \tag{OCP}$$

$$\begin{aligned} \text{subject to} \quad \dot{x}(t) &= f\,(x(t), y(t), u(t)) && \text{a.e. in } [0,1]\,, && (1)\\ 0 &= g\,(x(t), y(t), u(t)) && \text{a.e. in } [0,1]\,, && (2)\\ x(0) &= x^0 && && (3) \end{aligned}$$

with the functional $\varphi : \mathbb{R}^{n_x} \to \mathbb{R}$ and the functions $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$ and $g : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_y}$. Herein, (1), (2) is a DAE in semi-explicit form with Lipschitz continuous differential state $x \in W_{1,\infty}^{n_x}$ and essentially bounded algebraic variable and control $y \in L_\infty^{n_y}$, $u \in L_\infty^{n_u}$. The algebraic state y is implicitly determined by the algebraic constraint (2). DAEs are characterized by a quantity called index, which has various concepts, e.g., differentiation and perturbation index (cf. [24]). By differentiating the algebraic equation (2) with respect to t we get

$$\begin{aligned} 0 &= \frac{d}{dt} g(x(t), y(t), u(t)) \\ &= g'_x(x(t), y(t), u(t))\dot{x}(t) + g'_y(x(t), y(t), u(t))\dot{y}(t) + g'_u(x(t), y(t), u(t))\dot{u}(t). \end{aligned} \tag{4}$$

If we assume that the matrix g'_y is non-singular along a trajectory (compare Sect. 3), then we are able to solve this equation for $\dot{y}$ to obtain an explicit ODE for the algebraic state. Therefore, the DAE (1), (2) has differentiation index 1 since differentiating once with respect to t was sufficient to derive an explicit ODE.

Remark 1 The non singularity of the matrix g'_y along a trajectory implies that the Eq. (2) is (implicitly) solvable for y with respect to x and u. In theory, it would then be possible to reduce (1), (2) to an ODE and to apply a numerical scheme afterwards. However, in the context of DAEs, this has the drawback that the algebraic constraints are no longer enforced in the discrete system. Thus, depending on the dynamic, the discrete solution may suffer from the drift-off effect (cf. [10, 20]). Therefore, we suggest to discretize the system (OCP) directly and then to solve the discrete optimization problem.

Runge–Kutta schemes are often implemented to get accurate numerical solutions of DAEs. Hairer et al. [19] proved convergence of Runge–Kutta methods for Hessenberg DAEs up to index 3. Usually, in order to approximate a DAE with these schemes, we proceed as follows: for $N \in \mathbb{N}$, $N \geq 2$ we define the mesh size $h := \frac{1}{N}$ and choose coefficients b_j, a_{jk} for $j, k = 1, \ldots, s$. Then, we approximate the differential and algebraic state by

$$x_{i+1} = x_i + h \sum_{j=1}^{s} b_j p_i^j, \quad y_{i+1} = y_i + h \sum_{j=1}^{s} b_j q_i^j, \quad i = 0, \ldots, N-1$$

with stage derivatives p_i^j, q_i^j and stage approximations z_i^j, w_i^j determined by

$$p_i^j = f(z_i^j, w_i^j, u_i^j), \quad z_i^j = x_i + h \sum_{k=1}^{s} a_{jk} p_i^k, \quad i = 0, \ldots, N-1 \quad j = 1, \ldots, s,$$

$$0 = g(z_i^j, w_i^j, u_i^j), \quad w_i^j = y_i + h \sum_{k=1}^{s} a_{jk} q_i^k, \quad i = 0, \ldots, N-1 \quad j = 1, \ldots, s.$$

Herein, z_i^j and w_i^j approximate the differential and algebraic state at $t = t_i + c_j h$ with $t_i := ih$ and $c_j := \sum_{k=1}^{s} a_{jk}$. Moreover, q_i^j is an approximation of $\dot{y}$ at $t = t_i + c_j h$. Thus, we have a discretization of the index reduced system, which we get by solving (4) for $\dot{y}$, i.e.,

$$\begin{aligned}\dot{x}(t) &= f(x(t), y(t), u(t)),\\ \dot{y}(t) &= -g_y'(x(t), y(t), u(t))^{-1}\left[g_x'(x(t), y(t), u(t))f(x(t), y(t), u(t))\right.\\ &\quad \left.+g_u'(x(t), y(t), u(t))\dot{u}(t))\right].\end{aligned}$$

Note that the second equation also depends on $\dot{u}$. Furthermore, if we derive discrete necessary conditions with the standard Runge–Kutta scheme, we do not obtain an approximation of the continuous necessary conditions associated with (OCP) (compare (9)–(12)). Instead, we get a discretization for the necessary conditions of the index reduced ODE system. To generate an approximation for the necessary conditions of (OCP) and avoid the dependency on $\dot{u}$, we suggest to treat the algebraic state analogous to the control (cf. [13, 18]). This yields the following approximation:

$$\begin{aligned}x_{i+1} &= x_i + h\sum_{j=1}^{s} b_j f(z_i^j, y_i^j, u_i^j), && i = 0, \ldots, N-1\\ z_i^j &= x_i + h\sum_{k=1}^{s} a_{jk} f(z_i^k, y_i^k, u_i^k), && i = 0, \ldots, N-1, \quad j = 1, \ldots, s\\ 0 &= g(z_i^j, y_i^j, u_i^j), && i = 0, \ldots, N-1, \quad j = 1, \ldots, s\\ x_0 &= x^0\end{aligned}$$

with stage approximations $z_i^j \approx x(t_i + c_j h)$ as well as intermediate algebraic variable $y_i^j \approx y(t_i + c_j h)$ and control $u_i^j \approx u(t_i + c_j h)$. Then, with the abbreviation $x_i' := \frac{x_{i+1} - x_i}{h}$ for $i = 0, \ldots, N-1$, we have the discrete optimization problem

$$\text{Minimize} \quad \varphi(x_N) \tag{DOCP}$$

$$\text{subject to} \quad x_i' = \sum_{j=1}^{s} b_j f(z_i^j, y_i^j, u_i^j), \qquad i = 0, \ldots, N-1 \tag{5}$$

$$z_i^j = x_i + h\sum_{k=1}^{s} a_{jk} f(z_i^k, y_i^k, u_i^k), \quad i = 0, \ldots, N-1, \quad j = 1, \ldots, s \tag{6}$$

$$0 = g(z_i^j, y_i^j, u_i^j), \qquad i = 0, \ldots, N-1, \quad j = 1, \ldots, s \tag{7}$$

$$x_0 = x^0. \tag{8}$$

The objective of this paper is to prove that this system has a local solution, which satisfies certain error estimates with respect to the continuous solution. To that end, we proceed as follows:

In Sect. 2 we derive discrete necessary conditions for (DOCP), which are consistent with the continuous necessary conditions of (OCP). In Sect. 3 we introduce the main result (Theorem 1) of this paper and the assumptions required to prove it. We transform the discrete problem into an abstract setting for which we can apply a convergence theorem in Sect. 4. In Sect. 5 we estimate the consistency error and show that the discretization scheme is stable with respect to small perturbations. This allows us to apply Proposition 2 and prove the main result Theorem 1. Numerical experiments confirming the theoretical deductions are provided in Sect. 6 for different schemes. In Sect. 7 we summarize the results of this paper and give an outlook for the index 2 case. We moved some technical details that would disturb the reading flow to Appendix 1.

Notation We denote by $\mathbb{R}^n$ the $n-$dimensional Euclidean space with the norm $|\cdot|$. The space of $n \times m$ matrices A is endowed with the spectral norm $\|A\|$ and the $n-$dimensional unit matrix is denoted by I_n. Let $\mathfrak{B}_r(w)$ be the open ball with center w and radius $r > 0$. For generic, non-negative constants we use $\Gamma, \Gamma_1, \Gamma_2, \ldots$. Furthermore, for vector-valued functions $w : [0, 1] \to \mathbb{R}^n$, $p \in [1, \infty]$, and $k \in \mathbb{N}$ we introduce the Banach spaces

L_p^n space of equivalence classes which consist of measurable functions that are bounded in the norm $\|\cdot\|_p$,
$W_{k,p}^n$ Sobolev space of absolutely continuous functions that are bounded in the norm $\|\cdot\|_{k,p}$,

equipped, respectively, with the norms

$$\|w\|_p := \left(\int_0^1 |w(t)|^p \, \mathrm{d}t \right)^{\frac{1}{p}}, \quad p \in [1, \infty), \qquad \|w\|_\infty := \operatorname*{ess\,sup}_{t \in [0,1]} |w(t)|,$$

$$\|w\|_{k,p} := \left(\sum_{j=0}^{k} \left\| \frac{\mathrm{d}^j w}{\mathrm{d}t^j} \right\|_p^p \right)^{\frac{1}{p}}, \quad p \in [1, \infty), \qquad \|w\|_{k,\infty} := \sum_{j=0}^{k} \left\| \frac{\mathrm{d}^j w}{\mathrm{d}t^j} \right\|_\infty.$$

Moreover, we associate discrete sequences $(w_i)_{i=0,\ldots,N} \subset \mathbb{R}^n$ with the spaces

$L_{p,h}^n \subset L_p^n$ space of functions that are piecewise constant on $\big[t_i, t_{i+1}\big)$,
$W_{1,p,h}^n \subset W_{1,p}^n$ space of continuous functions that are piecewise linear on $\big[t_i, t_{i+1}\big)$

for $i = 0, \ldots, N-1$, $p \in [1, \infty]$ and define the discrete norms

$$\|w\|_{\infty,h} := \max_{0 \le i \le N} |w_i|, \quad \|w\|_{1,\infty,h} := \|w\|_{\infty,h} + \max_{1 \le i \le N} \left| \frac{w_i - w_{i-1}}{h} \right|.$$

Throughout the paper we use i as the index for the discrete time t_i and j, k for the index of coefficients. Finally, to simplify notation, we often use the abbreviation $F[t]$ for functions of type $F(\hat{w}(t))$ where $\hat{w}$ is a local minimizer or Karush–Kuhn–Tucker (KKT) point.

2 Necessary conditions

The general procedure to derive error estimates for solutions of optimal control problems is to compare the associated necessary conditions and the dynamic with its discrete counterparts. Therefore, in this section we derive necessary conditions associated with (DOCP), which are consistent with the continuous necessary conditions. These hold if (OCP) has a (local) solution. A tuple $(\hat{x}, \hat{y}, \hat{u}) \in W_{1,\infty}^{n_x} \times L_\infty^{n_y} \times L_\infty^{n_u}$ satisfying (1)–(3) is a local minimizer of (OCP) if there exists $\epsilon > 0$ such that

$$\varphi(\hat{x}(1)) \leq \varphi(x(1))$$

for all feasible $(x, y, u) \in \mathfrak{B}_\epsilon(\hat{x}, \hat{y}, \hat{u}) \subset W_{1,\infty}^{n_x} \times L_\infty^{n_y} \times L_\infty^{n_u}$ satisfying (1)–(3). Consequently, if (OCP) has a solution $(\hat{x}, \hat{y}, \hat{u})$ and the index 1 condition in Sect. 3 holds, then there exist Lagrange multipliers $\lambda \in W_{1,\infty}^{n_x}$ and $\mu \in L_\infty^{n_y}$ such that the normalized necessary conditions for (OCP) (cf. [16, Theorem 3.4.3])

$$\begin{aligned}
\dot{\lambda}(t) &= -\nabla_x \mathcal{H}(\hat{x}(t), \hat{y}(t), \hat{u}(t), \lambda(t), \mu(t)), && \text{a.e. in } [0,1], && (9)\\
0 &= \nabla_y \mathcal{H}(\hat{x}(t), \hat{y}(t), \hat{u}(t), \lambda(t), \mu(t)), && \text{a.e. in } [0,1], && (10)\\
\lambda(1) &= \nabla \varphi(\hat{x}(1)), && && (11)\\
0 &= \nabla_u \mathcal{H}(\hat{x}(t), \hat{y}(t), \hat{u}(t), \lambda(t), \mu(t)), && \text{a.e. in } [0,1] && (12)
\end{aligned}$$

are satisfied with the Hamilton function

$$\mathcal{H}(x, y, u, \lambda, \mu) := \lambda^\top f(x, y, u) + \mu^\top g(x, y, u).$$

Herein, we have the adjoint DAE (9), (10) with adjoint differential state λ and adjoint algebraic variable μ as well as the endpoint condition (11). Next, deriving necessary conditions associated with the discrete system (DOCP) yields adjoint equations for multipliers $\lambda_i, \eta_i^j, \mu_i^j, i = 0, \ldots, N-1, j = 1, \ldots, s$:

$$\begin{aligned}
\lambda_{i+1} &= \lambda_i - \sum_{j=1}^{s} \eta_i^j,\\
\eta_i^j &= h b_j f_x'\left(z_i^j, y_i^j, u_i^j\right)^\top \lambda_{i+1}\\
&\quad + h \sum_{k=1}^{s} a_{kj} f_x'\left(z_i^j, y_i^j, u_i^j\right)^\top \eta_i^k + h b_j g_x'\left(z_i^j, y_i^j, u_i^j\right)^\top \mu_i^j,\\
0 &= h b_j f_y'\left(z_i^j, y_i^j, u_i^j\right)^\top \lambda_{i+1}
\end{aligned}$$

$$+ h \sum_{k=1}^{s} a_{kj} f_y' \left(z_i^j, y_i^j, u_i^j \right)^\top \eta_i^k + h b_j g_y' \left(z_i^j, y_i^j, u_i^j \right)^\top \mu_i^j,$$

$$\lambda_N = \nabla \varphi \left(x_N \right),$$

$$0 = h b_j f_u' \left(z_i^j, y_i^j, u_i^j \right)^\top \lambda_{i+1} + h \sum_{k=1}^{s} a_{kj} f_u' \left(z_i^j, y_i^j, u_i^j \right)^\top \eta_i^k + h b_j g_u' \left(z_i^j, y_i^j, u_i^j \right)^\top \mu_i^j.$$

Assuming $b_j > 0$ for all $j = 1, \ldots, s$ and introducing the new multiplier

$$\nu_i^j := \lambda_{i+1} + \sum_{k=1}^{s} \frac{a_{kj}}{b_j} \eta_i^k, \quad i = 0, \ldots, N-1, \quad j = 1, \ldots, s$$

gives us

$$\begin{aligned} \eta_i^j &= h b_j \left(f_x' \left(z_i^j, y_i^j, u_i^j \right)^\top \nu_i^j + g_x' \left(z_i^j, y_i^j, u_i^j \right)^\top \mu_i^j \right) \\ &= h b_j \nabla_x \mathcal{H}(z_i^j, y_i^j, u_i^j, \nu_i^j, \mu_i^j), \end{aligned} \tag{13}$$

and therefore

$$\begin{aligned} \lambda_{i+1} &= \lambda_i - h \sum_{j=1}^{s} b_j \nabla_x \mathcal{H}(z_i^j, y_i^j, u_i^j, \nu_i^j, \mu_i^j), \\ \nu_i^j &= \lambda_{i+1} + h \sum_{k=1}^{s} \frac{b_k a_{kj}}{b_j} \nabla_x \mathcal{H}(z_i^k, y_i^k, u_i^k, \nu_i^k, \mu_i^k). \end{aligned}$$

Inserting λ_{i+1} into the second equation yields the new necessary conditions

$$\lambda_i' = - \sum_{j=1}^{s} b_j \nabla_x \mathcal{H} \left(z_i^j, y_i^j, u_i^j, \nu_i^j, \mu_i^j \right), \qquad i = 0, \ldots, N-1 \tag{14}$$

$$\nu_i^j = \lambda_i - h \sum_{k=1}^{s} \tilde{a}_{jk} \nabla_x \mathcal{H} \left(z_i^k, y_i^k, u_i^k, \nu_i^k, \mu_i^k \right), \quad i = 0, \ldots, N-1, \quad j = 1, \ldots, s \tag{15}$$

$$0 = \nabla_y \mathcal{H} \left(z_i^j, y_i^j, u_i^j, \nu_i^j, \mu_i^j \right), \qquad i = 0, \ldots, N-1, \quad j = 1, \ldots, s \tag{16}$$

$$\lambda_N = \nabla \varphi \left(x_N \right), \tag{17}$$

$$0 = \nabla_u \mathcal{H} \left(z_i^j, y_i^j, u_i^j, \nu_i^j, \mu_i^j \right), \qquad i = 0, \ldots, N-1, \quad j = 1, \ldots, s \tag{18}$$

with the coefficients

$$\tilde{a}_{jk} = \frac{b_k \left(b_j - a_{kj}\right)}{b_j} \quad j,k = 1,\ldots,s.$$

The Eqs. (14)–(18) can be transformed back to the original conditions with the multiplier η_i^j, using the relation (13), i.e., both systems are equivalent (cf. [18, Proposition 3.1]). In (14)–(18) the adjoint variables v_i^j and μ_i^j can be viewed as stage approximations for λ and intermediate adjoint algebraic states for μ, respectively. However, the Eqs. (5)–(8) and (14)–(18) are not a Runge–Kutta scheme applied to the KKT-system (1)–(3), (9)–(12) since the coefficients a_{jk} and $\tilde{a}_{jk}$ are not equal in general. Hence, further analysis is required to obtain error estimates.

3 Assumptions and main theorem

Before formulating the main result of this paper, we introduce the assumptions required to prove it. To conduct a convergence analysis in optimal control, it is typically presumed that the problem has a sufficiently smooth solution and that the system satisfies certain regularity properties. Moreover, in the nonlinear case, second order sufficient conditions are exploited since they imply stability of the problem. For the rest of the paper we assume the following:

(*Smoothness*) (OCP) has a local solution $(\hat{x}, \hat{y}, \hat{u}) \in W^{n_x}_{\kappa,\infty} \times W^{n_y}_{\kappa-1,\infty} \times W^{n_u}_{\kappa-1,\infty}$ for $\kappa \in \{2, 3\}$. For an open set $\mathcal{M} \subset \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u}$ and $\rho > 0$ such that $\mathfrak{B}_\rho(\hat{x}(t), \hat{y}(t), \hat{u}(t)) \subset \mathcal{M}$ for all $t \in [0, 1]$ the first κ derivatives of f and g exist and are Lipschitz continuous on $\mathcal{M}$. Furthermore, the first κ derivatives of φ exist and are Lipschitz continuous on $\mathfrak{B}_\rho(\hat{x}(1))$.

(*Index 1*) The matrix $g'_y(\hat{x}(t), \hat{y}(t), \hat{u}(t))$ is non-singular for all $t \in [0, 1]$.

(*Coercivity*) There exists $\gamma > 0$ such that the quadratic form

$$\mathcal{P}(x, y, u) := \frac{1}{2}\left(x(1)^\top \nabla^2 \varphi[1] x(1) + \int_0^1 \begin{pmatrix} x(t) \\ y(t) \\ u(t) \end{pmatrix}^\top \nabla^2_{(x,y,u)(x,y,u)} \mathcal{H}[t] \begin{pmatrix} x(t) \\ y(t) \\ u(t) \end{pmatrix} dt \right).$$

satisfies

$$\mathcal{P}(x, y, u) \geq \gamma \, \|u\|^2_{L_2} \tag{19}$$

for all $(x, y, u) \in W^{n_x}_{1,2} \times L^{n_y}_2 \times L^{n_u}_2$ such that

$$\begin{aligned} \dot{x}(t) &= f'_x[t]x(t) + f'_y[t]y(t) + f'_u[t]u(t), && \text{a.e. in } [0, 1], \\ 0 &= g'_x[t]x(t) + g'_y[t]y(t) + g'_u[t]u(t), && \text{a.e. in } [0, 1], \\ x(0) &= 0. \end{aligned}$$

Remark 2 (i) If smoothness for $\kappa \in \{2, 3\}$ and the index 1 condition are satisfied, then the Lagrange multipliers $(\hat{\lambda}, \hat{\mu})$ associated with the local solution $(\hat{x}, \hat{y}, \hat{u})$ are in the space $W^{n_x}_{\kappa,\infty} \times W^{n_y}_{\kappa-1,\infty}$. This can be seen by solving the adjoint algebraic equation (10) for $\hat{\mu}$ which yields $\hat{\mu} \in W^{n_y}_{1,\infty}$. Then, the adjoint differential equation (9) implies $W^{n_y}_{2,\infty}$. The process is repeated until the suggested smoothness is reached.

(ii) If the assumptions are satisfied, then there exists some $\beta > 0$ such that the Legendre-Clebsch condition

$$(w^\top, v^\top)\nabla^2_{(y,u)(y,u)}\mathcal{H}[t]\begin{pmatrix} w \\ v \end{pmatrix} \geq \beta(|w|^2 + |v|^2)$$

holds for all $(w, v) \in \ker(g'_y[t], g'_u[t])$ and $t \in [0, 1]$ (cf. [11, Lemma 2], [15], [27, Lemma 5.3.3]). In addition, this implies that the matrix

$$\begin{pmatrix} \nabla^2_{yy}\mathcal{H}[t] & \nabla^2_{yu}\mathcal{H}[t] & g'_y[t]^\top \\ \nabla^2_{uy}\mathcal{H}[t] & \nabla^2_{uu}\mathcal{H}[t] & g'_u[t]^\top \\ g'_y[t] & g'_u[t] & 0 \end{pmatrix} \tag{20}$$

is non-singular for all $t \in [0, 1]$.

(iii) Smoothness and the index 1 condition imply that $g'_y[t]$ and its inverse are continuous and uniformly bounded. Thus, we can solve the linear algebraic equation in the coercivity assumption for y and insert it into the differential equation to obtain

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{21}$$
$$x(0) = 0 \tag{22}$$

with the abbreviations

$$A(t) := f'_x[t] - f'_y[t]g'_y[t]^{-1}g'_x[t], \quad B(t) := f'_u[t] - f'_y[t]g'_y[t]^{-1}g'_u[t].$$

The quadratic form then reduces to

$$\widetilde{\mathcal{P}}(x, u) := \frac{1}{2}x(1)^\top\nabla^2\varphi[1]x(1) + \frac{1}{2}\int_0^1 x(t)^\top P(t)x(t) + 2x(t)^\top S(t)u(t) + u(t)^\top Q(t)u(t)dt$$

with the matrix functions

$$P(t) := \nabla^2_{xx}\mathcal{H}[t] - 2\nabla^2_{xy}\mathcal{H}[t]g'_y[t]^{-1}g'_x[t] + (g'_y[t]^{-1}g'_x[t])^\top\nabla^2_{yy}\mathcal{H}[t]g'_y[t]^{-1}g'_x[t],$$

Table 1 Conditions for Runge–Kutta schemes of order one to three for optimal control

Order	Conditions		
1	$\sum_{j=1}^{s} b_j = 1$		
2	$\sum_{j=1}^{s} d_j = \frac{1}{2}$		
3	$\sum_{j=1}^{s} c_j d_j = \frac{1}{6}$	$\sum_{j=1}^{s} b_j c_j^2 = \frac{1}{3}$	$\sum_{j=1}^{s} \frac{d_j^2}{b_j} = \frac{1}{3}$

$$
\begin{aligned}
S(t) &:= \nabla^2_{xu}\mathcal{H}[t] - \nabla^2_{xy}\mathcal{H}[t]g'_y[t]^{-1}g'_u[t] - (g'_y[t]^{-1}g'_x[t])^\top \nabla^2_{yu}\mathcal{H}[t] \\
&\quad + (g'_y[t]^{-1}g'_x[t])^\top \nabla^2_{yy}\mathcal{H}[t]g'_y[t]^{-1}g'_u[t], \\
Q(t) &:= \nabla^2_{uu}\mathcal{H}[t] - 2(g'_y[t]^{-1}g'_u[t])^\top \nabla^2_{yu}\mathcal{H}[t] \\
&\quad + (g'_y[t]^{-1}g'_u[t])^\top \nabla^2_{yy}\mathcal{H}[t]g'_y[t]^{-1}g'_u[t].
\end{aligned}
$$

For this reduced form we also have the coercivity condition $\widetilde{\mathcal{P}}(x,u) \geq \gamma \|u\|^2_{L_2}$ for all $(x,u) \in W^{n_x}_{1,2} \times L^{n_u}_2$ satisfying (21), (22). Furthermore, the Legendre Clebsch condition $v^\top Q(t)v \geq \gamma |v|^2$ holds for all $t \in [0,1]$ (cf. [11, Lemma 2], [15]).

Next, with the abbreviations $c_j := \sum_{k=1}^{s} a_{jk}$ and $d_j := \sum_{k=1}^{s} b_k a_{kj}$ we introduce the conditions required to get Runge–Kutta methods of order one to three in Table 1.

Herein, we have the additional condition $\sum_{j=1}^{s} \frac{d_j^2}{b_j} = \frac{1}{3}$ for third order, which is not needed for Runge–Kutta methods applied to DAEs. But in the context of optimal control, some extra conditions for the coefficients arise since a_{jk} and $\tilde{a}_{jk}$ are not equal in general.

In the following sections we will show that we can solve the Eqs. (7), (16), (18) for (y,u,μ) depending on (z,ν) near the continuous differential state and multiplier $(\hat{x},\hat{\lambda})$. Then, we get a discrete solution $(\hat{x}_h,\hat{\lambda}_h)$ of a reduced system, which satisfies error estimates with respect to $(\hat{x},\hat{\lambda})$. This further implies that the discrete problem (DOCP) has a local solution $(\hat{x}_h,\hat{z}_h,\hat{y}_h,\hat{u}_h)$ associated with multipliers $(\hat{\lambda}_h,\hat{\nu}_h,\hat{\mu}_h)$. However, the algebraic states $\hat{y}_h,\hat{\mu}_h$ and the control $\hat{u}_h$ might not converge at the same rate as the differential states, which we will later confirm with numerical experiments (compare Sect. 6). Though it is possible to obtain discrete algebraic states and a control that satisfy the same order of error estimates as the differential states by solving the algebraic equations (7), (16), (18) for (y,u,μ) with respect to $(x,\lambda) = (\hat{x}_h,\hat{\lambda}_h)$, i.e., $(y(\hat{x}_h,\hat{\lambda}_h), u(\hat{x}_h,\hat{\lambda}_h), \mu(\hat{x}_h,\hat{\lambda}_h))$. This allows us to formulate the main result of this paper:

Theorem 1 *For $\kappa \in \{2,3\}$ let smoothness, the index 1 condition, coercivity, $b_j > 0$ for $j = 1,\dots,s$, and the conditions in Table 1 up to order κ be satisfied. Then, (DOCP) has a local solution $(\hat{x}_h,\hat{z}_h,\hat{y}_h,\hat{u}_h)$ associated with the multipliers $(\hat{\lambda}_h,\hat{\nu}_h,\hat{\mu}_h)$ such that we have the error estimates*

$$\left\|\hat{x}-\hat{x}_h\right\|_{1,\infty,h}+\left\|\hat{y}-y(\hat{x}_h,\hat{\lambda}_h)\right\|_{\infty,h}+\left\|\hat{u}-u(\hat{x}_h,\hat{\lambda}_h)\right\|_{\infty,h}$$
$$+\left\|\hat{\lambda}-\hat{\lambda}_h\right\|_{1,\infty,h}+\left\|\hat{\mu}-\mu(\hat{x}_h,\hat{\lambda}_h)\right\|_{\infty,h}\leq\Gamma h^{\kappa}$$

for some constant $\Gamma\geq 0$ *and sufficiently small* h. *Herein,* $(y(\hat{x}_h,\hat{\lambda}_h), u(\hat{x}_h,\hat{\lambda}_h), \mu(\hat{x}_h,\hat{\lambda}_h))$ *is obtained by solving the algebraic constraints* (7), (16), (18) *for* (y,u,μ) *with respect to* $(x,\lambda)=(\hat{x}_h,\hat{\lambda}_h)$.

Note that the order κ of the error estimates in Theorem 1 is closely related to the assumed smoothness of the functions in (OCP). To derive error estimates, we exploit appropriate Taylor expansions, which contain higher order derivatives of the functions, i.e., they have to be sufficiently smooth (see Appendix 1 and Lemma 1).

4 Convergence theorem and abstract setting

Before we prove Theorem 1 in Sect. 5, we first transform the discrete KKT-system (5)–(8), (14)–(18) to obtain an equation $0=\mathcal{T}(\omega)$ and show that this system has a solution, which satisfies certain error estimates. To that end, we require the following result (cf. [14, Theorem 3.1], [12, Proposition 3.1], [18, Proposition 5.1]):

Proposition 2 *Let* Ω *be a Banach space and* Π *a linear, normed space. For some* $\bar{\omega}\in\Omega$ *and* $r>0$ *let the function* $\mathcal{T}:\mathfrak{B}_r(\bar{\omega})\subset\Omega\to\Pi$ *be continuously Frechét differentiable and let* $\mathcal{L}:\Omega\to\Pi$ *be a linear, bounded operator. Suppose there exist* $\theta,\vartheta,\sigma>0$ *such that*

- $\left\|\mathcal{T}'(\omega)-\mathcal{L}\right\|\leq\theta$ *for all* $\omega\in\mathfrak{B}_r(\bar{\omega})$.
- *The mapping* $\mathcal{L}^{-1}:\mathfrak{B}_\sigma(\hat{\pi})\to\Omega$ *for* $\hat{\pi}=\mathcal{T}(\bar{\omega})-\mathcal{L}(\bar{\omega})$ *is single valued and Lipschitz continuous with constant* ϑ.

If $\theta\vartheta<1$, $\theta r\leq\sigma$, *and* $\|\mathcal{T}(\bar{\omega})\|_\Pi\leq\min\left\{\sigma,\frac{(1-\vartheta\theta)r}{\vartheta}\right\}$, *then there exists a unique solution* $\omega\in\mathfrak{B}_r(\bar{\omega})$ *of* $0=\mathcal{T}(\omega)$ *satisfying the bound*

$$\|\omega-\bar{\omega}\|_\Omega\leq\frac{\vartheta}{1-\vartheta\theta}\|\mathcal{T}(\bar{\omega})\|_\Pi.$$

Our goal now is to transform the discrete KKT-system (5)–(8), (14)–(18) into a discretization of a boundary value problem such that we can apply Proposition 2. For that purpose, we introduce the abbreviations

$$X=(x,\lambda),\quad Y=(y,u,\mu),\quad Z=(z,\nu)$$
$$\mathbf{X}=(\underbrace{X,\ldots,X}_{s-\text{times}}),\quad \mathbf{Y}=(Y^1,\ldots,Y^s),\quad \mathbf{Z}=(Z^1,\ldots,Z^s)$$
$$F(X,Y)=\begin{pmatrix}f(x,y,u)\\-\nabla_x\mathcal{H}(x,y,u,\lambda,\mu)\end{pmatrix},\quad G(X,Y)=\begin{pmatrix}\nabla_y\mathcal{H}(x,y,u,\lambda,\mu)\\\nabla_u\mathcal{H}(x,y,u,\lambda,\mu)\\g(x,y,u)\end{pmatrix},$$

$$\Phi(X_0, X_N) = \begin{pmatrix} x_0 - x^0 \\ \lambda_N - \nabla\varphi(x_N) \end{pmatrix},$$

$$\Lambda_{jk} = \begin{pmatrix} a_{jk} I_{n_x} & 0 \\ 0 & \tilde{a}_{jk} I_{n_x} \end{pmatrix}, \quad \Upsilon_j := \sum_{k=1}^{s} \Lambda_{jk}, \quad j,k = 1,\dots,s.$$

Then, we can write the discrete system (5)–(8), (14)–(18) as

$$X_i' = \sum_{j=1}^{s} b_j F(Z_i^j, Y_i^j), \qquad i = 0,\dots,N-1, \tag{23}$$

$$Z_i^j = X_i + h\sum_{k=1}^{s} \Lambda_{jk} F(Z_i^k, Y_i^k), \qquad i = 0,\dots,N-1, \quad j = 1,\dots,s, \tag{24}$$

$$0 = G(Z_i^j, Y_i^j), \qquad i = 0,\dots,N-1, \quad j = 1,\dots,s, \tag{25}$$

$$0 = \Phi(X_0, X_N), \tag{26}$$

which is an approximation of the DAE boundary value problem

$$\begin{aligned} \dot{X}(t) &= F(X(t), Y(t)), && \text{a.e. in } [0,1], \\ 0 &= G(X(t), Y(t)), && \text{a.e. in } [0,1], \\ 0 &= \Phi(X(0), X(1)). \end{aligned}$$

According to (20), the matrix

$$G_Y' = \begin{pmatrix} \nabla_{yy}^2 \mathcal{H} & \nabla_{yu}^2 \mathcal{H} & g_y'^{\top} \\ \nabla_{uy}^2 \mathcal{H} & \nabla_{uu}^2 \mathcal{H} & g_u'^{\top} \\ g_y' & g_u' & 0 \end{pmatrix}$$

is non-singular along the trajectory $(\hat{X}, \hat{Y}) = (\hat{x}, \hat{\lambda}, \hat{y}, \hat{u}, \hat{\mu})$, i.e., the DAE has index 1. Since Λ_{jk} contains a_{jk} and $\tilde{a}_{jk}$, which are not equal in general, the discretization (23)–(26) with stage approximations Z_i^j is not a classic Runge–Kutta scheme and existing convergence results cannot be applied. Hence, further examination is required. We proceed by reducing the system (23)–(26) such that we obtain a discretization of an ODE boundary value problem. We first consider the algebraic equations

$$0 = G(Z^j, Y^j), \quad j = 1,\dots,s,$$

which are satisfied for $\bar{Z}^j = \hat{X}(t_i), \bar{Y}^j = \hat{Y}(t_i), j = 1,\dots,s$ for any $i = 0,\dots,N-1$. Moreover, the matrix $G_Y'[t_i]^{-1}$ exists for $i = 0,\dots,N-1$ (compare (20)). Thus, by the implicit function theorem (cf. [22, p. 29]), there exist some $\epsilon, \delta > 0$ such that the mappings $Y^j : \mathfrak{B}_\epsilon(\bar{\mathbf{Z}}) \to \mathfrak{B}_\delta(\bar{Y}^j)$, $j = 1\dots,s$ are continuously differentiable and satisfy

$$Y^j(\bar{\mathbf{Z}}) = \hat{Y}(t_i), \quad 0 = G(Z^j, Y^j(\mathbf{Z})) \qquad j = 1,\dots,s$$

$$\frac{\partial Y^j(\mathbf{Z})}{\partial Z^j} = -G'_Y(Z^j, Y^j(\mathbf{Z}))^{-1} G'_X(Z^j, Y^j(\mathbf{Z})), \qquad j = 1, \dots, s, \tag{27}$$

$$\frac{\partial Y^j(\mathbf{Z})}{\partial Z^\ell} = 0, \qquad j \neq \ell, \tag{28}$$

for $\mathbf{Z} \in \mathfrak{B}_\epsilon(\bar{\mathbf{Z}})$. Now, we insert the functions $Y^j(\cdot)$, $j = 1, \dots, s$ into the stage approximation (24) and consider the equations

$$0 = Z^j - P^j - h \sum_{k=1}^{s} \Lambda_{jk} F(Z^k, Y^k(\mathbf{Z})), \quad j = 1, \dots, s, \quad \mathbf{Z} \in \mathfrak{B}_\epsilon(\bar{\mathbf{Z}}) \tag{29}$$

with some parameters $\mathbf{P} = (P^1, \dots, P^s)$ replacing X_i. These equations are satisfied for $\bar{Z}^j = \hat{X}(t_i)$ and $\bar{P}^j = \hat{X}(t_i) - h\Upsilon_j F[t_i]$, $j = 1, \dots, s$ for any $i = 0, \dots, N-1$. Differentiating the right hand side with respect to $\mathbf{Z}$ yields a matrix of the form $I_{2sn_x} + O(h)$. Hence, this matrix is non-singular for sufficiently small h. Applying the implicit function theorem once more gives us continuously differentiable mappings $Z^j : \mathfrak{B}_{\tilde{\epsilon}}(\bar{\mathbf{P}}) \to \mathfrak{B}_{\tilde{\delta}}(\bar{Z}^j)$ for $j = 1, \dots, s$ and some $\tilde{\epsilon}, \tilde{\delta} > 0$. Then, for all $\mathbf{P} \in \mathfrak{B}_{\tilde{\epsilon}}(\bar{\mathbf{P}})$ we have

$$Z^j(\bar{\mathbf{P}}) = \hat{X}(t_i), \quad Z^j(\mathbf{P}) = P^j + h \sum_{k=1}^{s} \Lambda_{jk} F(Z^k(\mathbf{P}), Y^k(\mathbf{Z}(\mathbf{P}))), \quad j = 1, \dots, s,$$

$$\frac{\partial Z^j(\mathbf{P})}{\partial P^j} = I_{2n_x} + O(h), \qquad j = 1, \dots, s, \tag{30}$$

$$\frac{\partial Z^j(\mathbf{P})}{\partial P^\ell} = O(h), \qquad j \neq \ell. \tag{31}$$

Furthermore, since $\bar{P}^j - \hat{X}(t_i) = O(h)$ we get $\bar{\mathbf{Z}} = (\underbrace{\hat{X}(t_i), \dots, \hat{X}(t_i)}_{s-\text{times}}) \in \mathfrak{B}_{\tilde{\epsilon}}(\bar{\mathbf{P}})$ for sufficiently small h. Thus, the function $Z^j(\cdot)$ is also continuously differentiable on $(\mathfrak{B}_{\hat{\epsilon}}(\hat{X}(t_i)))^s$ for some $0 < \hat{\epsilon} < \tilde{\epsilon}$ and satisfies the stage approximations (29) for $P^j = \hat{X}(t_i)$. Finally, we insert $Z^j(\cdot)$ and $Y^j(\cdot)$ into the difference equation (23) to obtain

$$X'_i = \sum_{j=1}^{s} b_j F(Z^j(\mathbf{X}_i), Y^j(\mathbf{Z}(\mathbf{X}_i))), \quad i = 0, \dots, N-1,$$
$$0 = \Phi(X_0, X_N)$$

for $X_i \in \mathfrak{B}_{\hat{\epsilon}}(\hat{X}(t_i))$, $i = 0, \dots, N$. This system only depends on $X = (x, \lambda)$ and is an approximation of the ODE boundary value problem we get by solving (2), (10), (12) for (y, u, μ) with respect to (x, λ) and inserting the result into the differential equations (1) and (9). For this discrete equation we intent to apply Proposition 2 to get error estimates for x and λ. Therefore, we introduce the Banach space $\Omega := W^{2n_x}_{1,\infty,h}$

and the linear, normed space $\Pi := L_{1,h}^{2n_x} \times \mathbb{R}^{2n_x}$. Finally, with $\bar{X} \in \Omega$ defined by $\bar{X}_i := \hat{X}(t_i)$ for $i = 0, \ldots, N$ and some sufficiently small $r \in (0, \hat{\epsilon})$ we have the continuously Frechét differentiable function $\mathcal{T} : \mathfrak{B}_r(\bar{X}) \subset \Omega \to \Pi$ and the linear operator $\mathcal{L} : \Omega \to \Pi$

$$\mathcal{T}(X) := \begin{pmatrix} X_i' - \sum_{j=1}^{s} b_j F(Z^j(\mathbf{X}_i), Y^j(\mathbf{Z}(\mathbf{X}_i))), \quad i = 0, \ldots, N-1 \\ \Phi(X_0, X_N) \end{pmatrix},$$

$$\mathcal{L}(X) := \begin{pmatrix} X_i' - \sum_{j=1}^{s} b_j (F_X'[t_i] - F_Y'[t_i] G_Y'[t_i]^{-1} G_X'[t_i]) \mathbb{X}_i, \quad i = 0, \ldots, N-1 \\ \Phi_{X_0}'(\hat{X}(0), \hat{X}(1)) X_0 + \Phi_{X_1}'(\hat{X}(0), \hat{X}(1)) X_N \end{pmatrix}$$

with $\mathbb{X}_i := (x_i, \lambda_{i+1})$ for $i = 0, \ldots, N-1$.

5 Proof of the main theorem

Before proving Theorem 1, we derive an upper bound for the consistency error $\|\mathcal{T}(\bar{X})\|_\Pi$ with respect to h, which gives us error estimates for the discrete solution of $0 = \mathcal{T}(X)$ with respect to $\bar{X}$ after applying Proposition 2.

Lemma 1 *If smoothness for $\kappa \in \{2, 3\}$, the index 1 condition, $b_j > 0$ for $j = 1, \ldots, s$, and the conditions in Table 1 up to order κ are satisfied, then we have*

$$\|\mathcal{T}(\bar{X})\|_\Pi \leq \Gamma h^\kappa \tag{32}$$

for some constant $\Gamma \geq 0$ and sufficiently small h.

Proof The second component of $\mathcal{T}(\bar{X})$ is zero. Thus, it remains to estimate

$$\frac{\bar{X}_{i+1} - \bar{X}_i}{h} - \sum_{j=1}^{s} b_j F(Z^j(\bar{\mathbf{X}}_i), Y^j(\mathbf{Z}(\bar{\mathbf{X}}_i))), \quad i = 0, \ldots, N-1.$$

In Appendix 1 we derive Taylor expansions for $\frac{\bar{X}_{i+1} - \bar{X}_i}{h}$ and $F(Z^j(\bar{\mathbf{X}}_i), Y^j(\mathbf{Z}(\bar{\mathbf{X}}_i)))$ for $i = 0, \ldots, N-1$ such that the remainder terms are of order $O(h^3)$. Herein, we omit arguments, i.e., we write $\hat{F} = F[t_i]$ etc. for $i = 0, \ldots, N-1$. Now, we compare the terms of order $O(1)$, $O(h)$, and $O(h^2)$ for both expansions. The $O(1)$ terms $\hat{F}$ and $\sum_{j=1}^{s} b_j \hat{F}$ are equal if $\sum_{j=1}^{s} b_j = 1$, which corresponds to the order 1 condition in Table 1. For order $O(h)$ terms we have

$$\frac{1}{2}(\hat{F}_X' - \hat{F}_Y'(\hat{G}_Y')^{-1}\hat{G}_X')\hat{F} \quad \text{and} \quad \sum_{j=1}^{s} b_j (\hat{F}_X' - \hat{F}_Y'(\hat{G}_Y')^{-1}\hat{G}_X')\Upsilon_j \hat{F}.$$

Since $\Upsilon_j = \sum_{k=1}^{s} \Lambda_{jk} = \begin{pmatrix} c_j I_{nx} & 0 \\ 0 & \tilde{c}_j I_{nx} \end{pmatrix}$ with $\tilde{c}_j := \sum_{k=1}^{s} \tilde{a}_{jk}$, we can write

$$\Upsilon_j \hat{F} = c_j \hat{F}_f + \tilde{c}_j \hat{F}_{\mathcal{H}_x}, \quad \hat{F}_f := \begin{pmatrix} \hat{f} \\ 0 \end{pmatrix}, \quad \hat{F}_{\mathcal{H}_x} := \begin{pmatrix} 0 \\ -\nabla_x \hat{\mathcal{H}} \end{pmatrix}.$$

According to Table 1, for second order schemes we assume $\sum_{j=1}^{s} b_j c_j = \sum_{j=1}^{s} d_j = \frac{1}{2}$. Furthermore, we get

$$\sum_{j=1}^{s} b_j \tilde{c}_j = \sum_{j=1}^{s} \sum_{k=1}^{s} b_k (b_j - a_{kj}) = 1 - \sum_{j=1}^{s} d_j = \frac{1}{2}.$$

Thus, the terms are equal if the second order condition in Table 1 is satisfied. For third order we first consider linear terms of order $O(h^2)$ such as

$$\frac{1}{6} \hat{F}'_X (\hat{F}'_X - \hat{F}'_Y (\hat{G}'_Y)^{-1} \hat{G}'_X) \hat{F}$$

and

$$\begin{aligned}
&\sum_{j=1}^{s} b_j \hat{F}'_X \sum_{k=1}^{s} \Lambda_{jk} (\hat{F}'_X - \hat{F}'_Y (\hat{G}'_Y)^{-1} \hat{G}'_X) \Upsilon_k \hat{F} \\
&\quad = \hat{F}'_X \sum_{j=1}^{s} \sum_{k=1}^{s} b_j \Lambda_{jk} (\hat{F}'_X - \hat{F}'_Y (\hat{G}'_Y)^{-1} \hat{G}'_X)(c_k \hat{F}_f + \tilde{c}_k \hat{F}_{\mathcal{H}_x}) \\
&\quad = \hat{F}'_X \begin{pmatrix} \sum_{j=1}^{s} \sum_{k=1}^{s} b_j a_{jk} c_k I_{n_x} & 0 \\ 0 & \sum_{j=1}^{s} \sum_{k=1}^{s} b_j \tilde{a}_{jk} c_k I_{n_x} \end{pmatrix} (\hat{F}'_X - \hat{F}'_Y (\hat{G}'_Y)^{-1} \hat{G}'_X) \hat{F}_f \\
&\qquad + \hat{F}'_X \begin{pmatrix} \sum_{j=1}^{s} \sum_{k=1}^{s} b_j a_{jk} \tilde{c}_k I_{n_x} & 0 \\ 0 & \sum_{j=1}^{s} \sum_{k=1}^{s} b_j \tilde{a}_{jk} \tilde{c}_k I_{n_x} \end{pmatrix} (\hat{F}'_X - \hat{F}'_Y (\hat{G}'_Y)^{-1} \hat{G}'_X) \hat{F}_{\mathcal{H}_x}.
\end{aligned}$$

These terms are equal if

$$\frac{1}{6} = \sum_{j=1}^{s} \sum_{k=1}^{s} b_j a_{jk} c_k = \sum_{j=1}^{s} \sum_{k=1}^{s} b_j \tilde{a}_{jk} c_k = \sum_{j=1}^{s} \sum_{k=1}^{s} b_j a_{jk} \tilde{c}_k = \sum_{j=1}^{s} \sum_{k=1}^{s} b_j \tilde{a}_{jk} \tilde{c}_k,$$

which is satisfied if the third order conditions in Table 1 hold (cf. [18, Theorem 4.1]). For the quadratic terms we have, e.g., $\frac{1}{6}\hat{F}''_{XX}(\hat{F},\hat{F})$ and

$$\begin{aligned}\frac{1}{2}\sum_{j=1}^{s} b_j \hat{F}''_{XX}(\Upsilon_j\hat{F},\Upsilon_j\hat{F}) &= \frac{1}{2}\sum_{j=1}^{s} b_j \hat{F}''_{XX}(c_j\hat{F}_f+\tilde{c}_j\hat{F}_{\mathcal{H}_x}, c_j\hat{F}_f+\tilde{c}_j\hat{F}_{\mathcal{H}_x})\\ &= \frac{1}{2}\sum_{j=1}^{s} b_j\left[c_j^2\hat{F}''_{XX}(\hat{F}_f,\hat{F}_f)+2c_j\tilde{c}_j\hat{F}''_{XX}(\hat{F}_f,\hat{F}_{\mathcal{H}_x})+\tilde{c}_j^2\hat{F}''_{XX}(\hat{F}_{\mathcal{H}_x},\hat{F}_{\mathcal{H}_x})\right],\end{aligned}$$

which are equal if

$$\frac{1}{3}=\sum_{j=1}^{s} b_j c_j^2=\sum_{j=1}^{s} b_j c_j\tilde{c}_j=\sum_{j=1}^{s} b_j\tilde{c}_j^2.$$

According to [18, Theorem 4.1], this holds if the third order conditions in Table 1 are satisfied. Thus, we obtain

$$\left|\frac{\bar{X}_{i+1}-\bar{X}_i}{h}-\sum_{j=1}^{s} b_j F(\bar{Z}^j,\bar{Y}^j)\right| = O(h^3),\quad i=0,\dots,N-1$$

for Runge–Kutta schemes satisfying all conditions in Table 1. In summary, we have

$$\left\|\mathcal{T}(\bar{X})\right\|_{\Pi}\le \Gamma h^{\kappa},\quad \kappa\in\{2,3\}$$

for some $\Gamma\ge 0$ if the conditions in Table 1 up to order κ are satisfied. □

Next, we verify that the conditions of Proposition 2 are satisfied for our abstract setting:

Lemma 2 *Let the assumptions of Theorem 1 be satisfied. Then, for arbitrary $\theta>0$ there exists some sufficiently small $r>0$ and $h>0$ such that $\left\|\mathcal{T}'(W)-\mathcal{L}\right\|\le\theta$ for all $W\in\mathfrak{B}_r(\bar{X})$. Furthermore, for $\hat{\pi}=\mathcal{T}(\bar{\omega})-\mathcal{L}(\bar{X})$ and some constant $\sigma>0$ the mapping $\mathcal{L}^{-1}:\mathfrak{B}_\sigma(\hat{\pi})\to\Omega$ is single valued and Lipschitz continuous with constant $\vartheta\ge 0$.*

Proof First, we estimate the norm for the linear operator $\mathcal{T}'(W)-\mathcal{L}:\Omega=W^{2n_x}_{1,\infty,h}\to\Pi$, i.e.,

$$\left\|\mathcal{T}'(W)-\mathcal{L}\right\| := \sup_{X\neq 0}\frac{\left\|(\mathcal{T}'(W)-\mathcal{L})X\right\|_{\Pi}}{\|X\|_{1,\infty,h}}.$$

To this end, we abbreviate

$$E(X,Y) := F'_X(X,Y)-F'_Y(X,Y)G'_Y(X,Y)^{-1}G'_X(X,Y),$$

which is Lipschitz continuous with respect to (X, Y) close to $(\hat{X}(t), \hat{Y}(t))$ for some $t \in [0, 1]$. Using the sensitivities (27), (28), (30), (31) we get

$$\begin{aligned}\frac{dF(Z^j(\mathbf{X}), Y^j(\mathbf{Z}(\mathbf{X})))}{dX} &= F'_X(Z^j(\mathbf{X}), Y^j(\mathbf{Z}(\mathbf{X})))\frac{dZ^j(\mathbf{X})}{dX} \\ &\quad + F'_Y(Z^j(\mathbf{X}), Y^j(\mathbf{Z}(\mathbf{X})))\frac{\partial Y^j(\mathbf{Z}(\mathbf{X}))}{\partial \mathbf{Z}}\frac{\partial \mathbf{Z}(\mathbf{X})}{\partial \mathbf{X}} \\ &= E(Z^j(\mathbf{X}), Y^j(\mathbf{Z}(\mathbf{X})))(I_{2n_x} + O(h)).\end{aligned}$$

Then, the linearization of $\mathcal{T}$ in some $W \in \mathfrak{B}_r(\bar{X})$ yields

$$\mathcal{T}'(W)X = \begin{pmatrix} X'_i - \sum_{j=1}^{s} b_j(E(Z^j(\mathbf{W}_i), Y^j(\mathbf{Z}(\mathbf{W}_i))) + O(h))X_i, \quad i=0,\ldots,N-1 \\ \Phi'_{X_0}(W_0, W_N)X_0 + \Phi'_{X_1}(W_0, W_N)X_N \end{pmatrix}$$

and we can write the linear Operator $\mathcal{L}$ as

$$\mathcal{L}(X) = \begin{pmatrix} X'_i - \sum_{j=1}^{s} b_j E(\hat{X}(t_i), \hat{Y}(t_i))\mathbb{X}_i, \quad i=0,\ldots,N-1 \\ \Phi'_{X_0}(\hat{X}(0), \hat{X}(1))X_0 + \Phi'_{X_1}(\hat{X}(0), \hat{X}(1))X_N \end{pmatrix}, \quad \mathbb{X}_i = \begin{pmatrix} x_i \\ \lambda_{i+1} \end{pmatrix}.$$

For the first component of $\mathcal{T}'(W)X - \mathcal{L}(X)$ we get

$$\begin{aligned}&\left| \sum_{j=1}^{s} b_j \left[(E(Z^j(\mathbf{W}_i), Y^j(\mathbf{Z}(\mathbf{W}_i))) + O(h))X_i - E(\hat{X}(t_i), \hat{Y}(t_i))\mathbb{X}_i \right] \right| \\ &\quad \le \sum_{j=1}^{s} b_j \left[\left| E(Z^j(\mathbf{W}_i), Y^j(\mathbf{Z}(\mathbf{W}_i))) + O(h) - E(\hat{X}(t_i), \hat{Y}(t_i)) \right| |X_i| \right. \\ &\qquad \left. + \left| E(\hat{X}(t_i), \hat{Y}(t_i)) \right| |X_i - \mathbb{X}_i| \right] \\ &\quad \le \Gamma_1(r+h)\,\|X\|_{\infty,h} + \Gamma_2 h\,\|X\|_{1,\infty,h} \\ &\quad \le \Gamma_3(r+h)\,\|X\|_{1,\infty,h}\end{aligned}$$

for $i = 0, \ldots, N-1$ and some generic constants $\Gamma_1, \Gamma_2, \Gamma_3 \ge 0$. For the second component we exploit the Lipschitz continuity of Φ' to obtain the bound $r\,\|X\|_{1,\infty,h}$. Thus, we have $\left\|\mathcal{T}'(W) - \mathcal{L}\right\| \le \theta$ for arbitrary $\theta > 0$ if r and h are sufficiently small.

Next, we need to verify that $\mathcal{L}^{-1} : \mathfrak{B}_\sigma(\hat{\pi}) \to \Omega$ exists and is bounded. To this end, we consider the linear, perturbed system

$$\begin{aligned}x'_i &= \sum_{j=1}^{s} b_j(f'_x[t_i]x_i + f'_y[t_i]y_i^j + f'_u[t_i]u_i^j + \pi_{f,i}), \\ 0 &= b_j(g'_x[t_i]x_i + g'_y[t_i]y_i^j + g'_u[t_i]u_i^j + \pi_{g,i}^j),\end{aligned}$$

$$
\begin{aligned}
x_0 &= \pi_0,\\
\lambda_i' &= -\sum_{j=1}^{s} b_j \Big[\nabla_{xx}^2 \mathcal{H}[t_i] x_i + \nabla_{xy}^2 \mathcal{H}[t_i] y_i^j + \nabla_{xu}^2 \mathcal{H}[t_i] u_i^j\\
&\qquad + f_x'[t_i]^\top \lambda_{i+1} + g_x'[t_i]^\top \mu_i^j + \pi_{\mathcal{H}_x,i}\Big],\\
0 &= b_j(\nabla_{yx}^2 \mathcal{H}[t_i] x_i + \nabla_{yy}^2 \mathcal{H}[t_i] y_i^j + \nabla_{yu}^2 \mathcal{H}[t_i] u_i^j\\
&\qquad + f_y'[t_i]^\top \lambda_{i+1} + g_y'[t_i]^\top \mu_i^j + \pi_{\mathcal{H}_y,i}^j),\\
\lambda_N &= \nabla^2 \varphi[t_N] x_N + \pi_\varphi,\\
0 &= b_j(\nabla_{ux}^2 \mathcal{H}[t_i] x_i + \nabla_{uy}^2 \mathcal{H}[t_i] y_i^j + \nabla_{uu}^2 \mathcal{H}[t_i] u_i^j\\
&\qquad + f_u'[t_i]^\top \lambda_{i+1} + g_u'[t_i]^\top \mu_i^j + \pi_{\mathcal{H}_u,i}^j)
\end{aligned}
\tag{33}
$$

for $i = 0, \ldots, N-1$, $j = 1, \ldots, s$, and perturbations

$$
\begin{aligned}
&(\pi_f, \pi_g^j, \pi_0, \pi_{\mathcal{H}_x}, \pi_{\mathcal{H}_y}^j, \pi_\varphi, \pi_{\mathcal{H}_u}^j)\\
&\quad \in L_{1,h}^{n_x} \times L_{\infty,h}^{n_y} \times \mathbb{R}^{n_x} \times L_{1,h}^{n_x} \times L_{\infty,h}^{n_y} \times \mathbb{R}^{n_x} \times L_{\infty,h}^{n_u}.
\end{aligned}
$$

These are the KKT-conditions associated with the linear quadratic program

$$
\text{Minimize} \quad \mathcal{P}_h(x, y, u) + x_N^\top \pi_\varphi + h \sum_{i=0}^{N-1} \sum_{j=1}^{s} b_j \begin{pmatrix} x_i \\ y_i^j \\ u_i^j \end{pmatrix}^\top \begin{pmatrix} \pi_{\mathcal{H}_x,i} \\ \pi_{\mathcal{H}_y,i}^j \\ \pi_{\mathcal{H}_u,i}^j \end{pmatrix} \tag{LQP}
$$

$$
\begin{aligned}
\text{subject to} \quad x_i' &= \sum_{j=1}^{s} b_j (f_x'[t_i] x_i + f_y'[t_i] y_i^j + f_u'[t_i] u_i^j + \pi_{f,i}),\\
0 &= b_j (g_x'[t_i] x_i + g_y'[t_i] y_i^j + g_u'[t_i] u_i^j + \pi_{g,i}^j),\\
x_0 &= \pi_0.
\end{aligned}
\tag{34}
$$

for $i = 0, \ldots, N-1$, $j = 1, \ldots, s$, and with the discrete quadratic form

$$
\mathcal{P}_h(x, y, u) := \frac{1}{2} x_N^\top \nabla^2 \varphi[t_N] x_N + \frac{h}{2} \sum_{i=0}^{N-1} \sum_{j=1}^{s} b_j \begin{pmatrix} x_i \\ y_i^j \\ u_i^j \end{pmatrix}^\top \nabla_{(x,y,u)(x,y,u)}^2 \mathcal{H}[t_i] \begin{pmatrix} x_i \\ y_i^j \\ u_i^j \end{pmatrix}.
$$

Since the matrix $g_y'[t_i]$ is non-singular for all $i = 0, \dots, N$ by the index 1 assumption, we can solve (34) for y_i^j and obtain the reduced system

$$\begin{aligned} &\text{Minimize} \quad \widetilde{\mathcal{P}}_h(x,u) + x_N^\top \pi_\varphi + h \sum_{i=0}^{N-1} \sum_{j=1}^{s} b_j (x_i^\top \tilde{\pi}^j_{\mathcal{H}_x,i} + (u_i^j)^\top \tilde{\pi}^j_{\mathcal{H}_u,i}) \qquad \text{(RLQP)} \\ &\text{subject to} \quad x_i' = \sum_{j=1}^{s} b_j (A(t_i)x_i + B(t_i)u_i^j + \tilde{\pi}^j_{f,i}) \quad i = 0, \dots, N-1, \\ &\qquad\qquad x_0 = \pi_0, \end{aligned}$$

where the reduced quadratic form is defined by

$$\begin{aligned} \widetilde{\mathcal{P}}_h(x,u) := &\frac{1}{2} x_N^\top \nabla^2 \varphi[t_N] x_N \\ &+ \frac{h}{2} \sum_{i=0}^{N-1} \sum_{j=1}^{s} b_j \left(x_i^\top P(t_i) x_i + 2 x_i^\top S(t_i) u_i^j + (u_i^j)^\top Q(t_i) u_i^j \right), \end{aligned}$$

and the perturbations $(\tilde{\pi}^j_f, \tilde{\pi}^j_{\mathcal{H}_x}, \tilde{\pi}^j_{\mathcal{H}_u}) \in L^{n_x}_{1,h} \times L^{n_x}_{1,h} \times L^{n_u}_{\infty,h}$ by

$$\begin{aligned} \tilde{\pi}^j_{f,i} &:= \pi_{f,i} - f_y'[t_i] g_y'[t_i]^{-1} \pi^j_{g,i}, \quad \tilde{\pi}^j_{\mathcal{H}_x,i} := \pi_{\mathcal{H}_x,i} - (g_y'[t_i]^{-1} g_x'[t_i])^\top \pi^j_{\mathcal{H}_y,i}, \\ \tilde{\pi}^j_{\mathcal{H}_u,i} &:= \pi^j_{\mathcal{H}_u,i} - (g_y'[t_i]^{-1} g_u'[t_i])^\top \pi^j_{\mathcal{H}_y,i}, \quad i = 0, \dots, N-1, \quad j = 1, \dots, s. \end{aligned}$$

The matrix functions $A, \dots, Q$ are defined in Remark 2 (iii). Furthermore, the matrix $Q(t_i)$ is uniformly positive definite for all $i = 0, \dots, N-1$ according to Remark 2 (iii). Hence, we can apply [13, Lemma 6.1] which yields a Lipschitz continuous solution of (RLQP) with respect to the perturbations $(\tilde{\pi}^j_f, \pi_0, \tilde{\pi}^j_{\mathcal{H}_x}, \pi_\varphi, \tilde{\pi}^j_{\mathcal{H}_u})$, $j = 1, \dots, s$. This in turn implies that the linear quadratic program (LQP) and therefore the KKT-system (33) have a solution for any perturbation. Moreover, we can write (33) as

$$\begin{aligned} X_i' &= \sum_{j=1}^{s} b_j (F_X'[t_i] \mathbb{X}_i + F_Y'[t_i] Y_i^j + \pi_{F,i}) \quad i = 0, \dots, N-1, \\ 0 &= b_j (G_X'[t_i] \mathbb{X}_i + G_Y'[t_i] Y_i^j + \pi^j_{G,i}) \qquad i = 0, \dots, N-1, \quad j = 1, \dots, s, \\ 0 &= \Phi'_{X_0}(\hat{X}(0), \hat{X}(1)) X_0 + \Phi'_{X_1}(\hat{X}(0), \hat{X}(1)) X_N + \pi_\Phi \end{aligned}$$

with $\pi_{F,i} := (\pi_{f,i}, -\pi_{\mathcal{H}_x,i})$, $\pi^j_{G,i} := (\pi^j_{\mathcal{H}_y,i}, \pi^j_{\mathcal{H}_u,i}, \pi^j_{g,i})$, and $\pi_\Phi := (\pi_0, \pi_\varphi)$. Since the matrix $G_Y'[t_i]$ is invertible by (20), we can solve the second equation for Y_i^j and insert the result into the first equation. This yields the perturbed equation $\mathcal{L}(X) = \pi$, which has a unique Lipschitz continuous solution with respect to $\pi \in \Pi$ and some Lipschitz constant $\vartheta \geq 0$ as we have verified. Thus, the single valued mapping $\mathcal{L}^{-1} : \mathfrak{B}_\sigma(\hat{\pi}) \to \Omega$ exists and is Lipschitz continuous with constant ϑ. □

With the results of Lemmas 1 and 2 we are finally able to prove the main theorem of this paper:

Proof of Theorem 1 In Lemma 1 and 2 we derived an upper bound for $\left\|\mathcal{T}(\bar{X})\right\|_\Pi$ (compare (32)) and verified the stability conditions of Proposition 2 for some $r, \theta, \vartheta, \sigma > 0$, respectively. We can choose θ and r sufficiently small such that $\theta\vartheta < 1$ and $\theta r \leq \sigma$. Additionally, since $\left\|\mathcal{T}(\bar{X})\right\|_\Pi = O(h)$, for sufficiently small h we have $\left\|\mathcal{T}(\bar{X})\right\|_\Pi \leq \min\left\{\sigma, \frac{(1-\vartheta\theta)r}{\vartheta}\right\}$. Then, according to Proposition 2, the equation $0 = \mathcal{T}(X)$ has a solution $\hat{X}_h =: (\hat{x}_h, \hat{\lambda}_h)$ satisfying the bound

$$\left\|\hat{X}_h - \bar{X}\right\|_{1,\infty,h} \leq \frac{\vartheta}{1-\vartheta\theta}\left\|\mathcal{T}(\bar{X})\right\|_\Pi = O(h^\kappa) < \hat{\epsilon}, \quad \kappa \in \{2, 3\}$$

for sufficiently small h. Recalling $\bar{P}_i^j = \hat{X}(t_i) - h\sum_{k=1}^{s} \Lambda_{jk} F[t_i]$, we conclude

$$\left|\hat{X}_{h,i} - \bar{P}_i^j\right| = \left|\hat{X}_{h,i} - \hat{X}(t_i) + h\sum_{k=1}^{s} \Lambda_{jk} F[t_i]\right| = O(h)$$

for $i = 0, \ldots, N$ and $j = 1, \ldots, s$. This implies $\hat{\mathbf{X}}_h \in \mathfrak{B}_{\tilde{\epsilon}}(\bar{\mathbf{P}})$ for sufficiently small h and therefore $Z^j(\hat{\mathbf{X}}_h) =: (\hat{z}_h^j, \hat{v}_h^j)$ exists for $j = 1, \ldots, s$. In addition, we have

$$\left\|Z^j(\hat{\mathbf{X}}_h) - \hat{X}\right\|_{\infty,h} = \left\|Z^j(\hat{\mathbf{X}}_h) - Z^j(\bar{\mathbf{P}})\right\|_{\infty,h} \leq \Gamma_1\left\|\hat{\mathbf{X}}_h - \bar{\mathbf{P}}\right\|_{\infty,h} = O(h)$$

for $\Gamma_1 \geq 0$. Since $\bar{Z}_i^j = \hat{X}(t_i)$, we get $\mathbf{Z}(\hat{\mathbf{X}}_h) \in \mathfrak{B}_\epsilon(\bar{\mathbf{Z}})$ for sufficiently small h. Thus, $Y^j(\mathbf{Z}(\hat{\mathbf{X}}_h)) =: (\hat{y}_h^j, \hat{u}_h^j, \hat{\mu}_h^j)$ exists for $j = 1, \ldots, s$. Moreover, the bound

$$\begin{aligned}\left\|Y^j(\mathbf{Z}(\hat{\mathbf{X}}_h)) - \hat{Y}\right\|_{\infty,h} &= \left\|Y^j(\mathbf{Z}(\hat{\mathbf{X}}_h)) - Y^j(\bar{\mathbf{Z}})\right\|_{\infty,h} \\ &\leq \Gamma_2\left\|\mathbf{Z}(\hat{\mathbf{X}}_h) - \bar{\mathbf{Z}}\right\|_{\infty,h} = O(h)\end{aligned}$$

holds for some $\Gamma_2 \geq 0$. Therefore, $(\hat{x}_h, \hat{z}_h, \hat{y}_h, \hat{u}_h)$ is feasible for (DOCP) and a local minimizer for sufficiently small h according to [13, Lemma 7.2]. Finally, we verify the (improved) error estimates for $Y^j(\hat{\mathbf{X}}_h) =: (y^j(\hat{x}_h, \hat{\lambda}_h), u^j(\hat{x}_h, \hat{\lambda}_h), \mu^j(\hat{x}_h, \hat{\lambda}_h))$, which are obtained by solving the algebraic equations (7), (16), (18) for (y, u, μ) with respect to $(x, \lambda) = (\hat{x}_h, \hat{\lambda}_h)$. These exist for sufficiently small h since $\left\|\hat{\mathbf{X}}_h - \bar{\mathbf{Z}}\right\|_{\infty,h} = \left\|\hat{\mathbf{X}}_h - \hat{\mathbf{X}}\right\|_{\infty,h} = O(h^\kappa) < \epsilon$. Using the Lipschitz continuity of Y^j we get

$$\begin{aligned}\left\|Y^j(\hat{\mathbf{X}}_h) - \hat{Y}(t_i)\right\|_{\infty,h} &= \left\|Y^j(\hat{\mathbf{X}}_h) - Y^j(\bar{\mathbf{X}})\right\|_{\infty,h} \leq \Gamma\left\|\hat{X}_h - \bar{X}\right\|_{\infty,h} \\ &= O(h^\kappa), \quad \kappa \in \{2, 3\}\end{aligned}$$

for some $\Gamma \geq 0$, which proves the assertion. □

6 Numerical example

In order to verify the results of Theorem 1, we consider the following optimal control problem with a parameter $\alpha \neq 0$:

$$\begin{aligned}
&\text{Minimize} \quad x_2(1)\\
&\text{subject to} \quad \dot{x}_1(t) = x_1(t) + 2y(t) - u(t), && x_1(0) = 1,\\
&\qquad\qquad \dot{x}_2(t) = \frac{1}{2}(x_1(t)^2 + \alpha y(t)^2 + u(t)^2), && x_2(0) = 0,\\
&\qquad\qquad 0 \;\; = \frac{1}{2\alpha}x_1(t) - y(t) + \frac{1}{2}u(t).
\end{aligned}$$

Utilizing the normalized necessary conditions associated with this problem

$$\begin{aligned}
&\dot{\lambda}_1(t) = -\lambda_1(t) - x_1(t)\lambda_2(t) - \frac{1}{2\alpha}\mu(t), && \lambda_1(1) = 0,\\
&\dot{\lambda}_1(t) = 0, && \lambda_2(1) = 1,\\
&\qquad 0 = 2\lambda_1(t) + \alpha y(t)\lambda_2(t) - \mu(t),\\
&\qquad 0 = -\lambda_1(t) + u(t)\lambda_2(t) + \frac{1}{2}\mu(t)
\end{aligned}$$

yields the solution

$$\begin{aligned}
&\hat{x}_1(t) = \exp\left(\frac{\alpha+1}{\alpha}t\right), \quad \hat{x}_2(t) = \frac{\alpha^2+4\alpha+1}{4(\alpha+1)(\alpha+4)}\left(\exp\left(2\frac{\alpha+1}{\alpha}t\right) - 1\right),\\
&\hat{y}(t) = \frac{2}{\alpha(\alpha+4)}\exp\left(\frac{\alpha+1}{\alpha}t\right), \quad \hat{u}(t) = -\frac{1}{\alpha+4}\exp\left(\frac{\alpha+1}{\alpha}t\right),\\
&\hat{\lambda}_1(t) = \frac{\alpha^2+4\alpha+1}{(\alpha+1)(\alpha+4)}\exp\left(\frac{\alpha+1}{\alpha}\right)\sinh\left(\frac{\alpha+1}{\alpha}(1-t)\right), \quad \hat{\lambda}_2(t) = 1,\\
&\hat{\mu}(t) = -\frac{\alpha^2+2\alpha-1}{(\alpha+1)(\alpha+4)}\exp\left(\frac{\alpha+1}{\alpha}t\right) + \frac{\alpha^2+4\alpha+1}{(\alpha+1)(\alpha+4)}\exp\left(\frac{\alpha+1}{\alpha}(2-t)\right).
\end{aligned}$$

For $\alpha \in \mathbb{R} \setminus \{-4, -1, 0\}$ the smoothness assumptions are satisfied. Furthermore, we can (explicitly) solve the algebraic equation for y, i.e., the DAE has index 1. Moreover, we have the quadratic form

$$\mathcal{P}(x_1, x_2, y, u) := \frac{1}{2}\int_0^1 x_1(t)^2 + \alpha y(t)^2 + u(t)^2 dt,$$

Table 2 Examined Runge–Kutta schemes for the example

Heun (second order)			Radau IA (third order)		
0	0	0	0	$\frac{1}{4}$	$-\frac{1}{4}$
1	1	0	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{12}$
	$\frac{1}{2}$	$\frac{1}{2}$		$\frac{1}{4}$	$\frac{3}{4}$

and the linearized dynamic

$$\begin{aligned} \dot{x}_1(t) &= x_1(t) + 2y(t) - u(t), & x_1(0) &= 0, \\ \dot{x}_2(t) &= \hat{x}_1(t)x_1(t) + \alpha\hat{y}(t)y(t) + \hat{u}(t)u(t), & x_2(0) &= 0, \\ 0 &= \frac{1}{2\alpha}x_1(t) - y(t) + \frac{1}{2}u(t). \end{aligned}$$

If $\alpha > 0$, then the coercivity condition (19) is obviously satisfied. For negative α the linearized dynamic implies $x_1(t) = 0$ and $y(t) = \frac{1}{2}u(t)$ for all $t \in [0, 1]$. Hence, we obtain

$$\mathcal{P}(x_1, x_2, y, u) = \frac{1}{2}\int_0^1 (\frac{\alpha}{4} + 1)u(t)^2 dt = \frac{\alpha + 4}{8}\|u\|_2^2,$$

which is positive for $\alpha \in (-4, 0)$. However, the coercivity assumption does not hold for $\alpha < -4$. Therefore, we did numerical experiments for the parameter values $\alpha = 1$ and $\alpha = -4.2$ as well as the Runge–Kutta schemes in Table 2.

The convergence order κ for $\hat{w} = \hat{x}, \hat{y}, \ldots$ can be approximated with the formula

$$\kappa \approx \log_2\left(\frac{\|\hat{w}_h - \hat{w}\|_{\infty,h}}{\|\hat{w}_{h/2} - \hat{w}\|_{\infty,h}}\right) = -\log_2\left(\|\hat{w}_{h/2} - \hat{w}\|_{\infty,h}\right) + \log_2\left(\|\hat{w}_h - \hat{w}\|_{\infty,h}\right).$$

Therefore, in Figs. 1 and 2 we plotted $-\log_2\left(\|\hat{w}_h - \hat{w}\|_{\infty,h}\right)$ for the different errors and the values $N = 20, 40, 80, 160, 320, 640$. Then, the convergence order is indicated by the vertical distance between two consecutive points.

The Heun scheme is displayed in Fig. 1 with the aforementioned values of α. Solving the algebraic constraints for (y, μ, u) with respect to $(x, \lambda) = (\hat{x}_h, \hat{\lambda}_h)$ improves the accuracy for the algebraic states and control but not the convergence order of 2. Note that even if the coercivity condition is not satisfied (for $\alpha = -4.2$) we still get second order error estimates. For the third order scheme Radau IA and $\alpha = 1$ we get third order error estimates for the differential states but only second order error estimates for the algebraic states and control (compare Fig. 2). The order of convergence is improved by solving the algebraic constraints for (y, μ, u) with respect to $(x, \lambda) = (\hat{x}_h, \hat{\lambda}_h)$. For the parameter value $\alpha = -4.2$ we do not get third order error estimates for the differential states but still second order error estimates for the algebraic variables

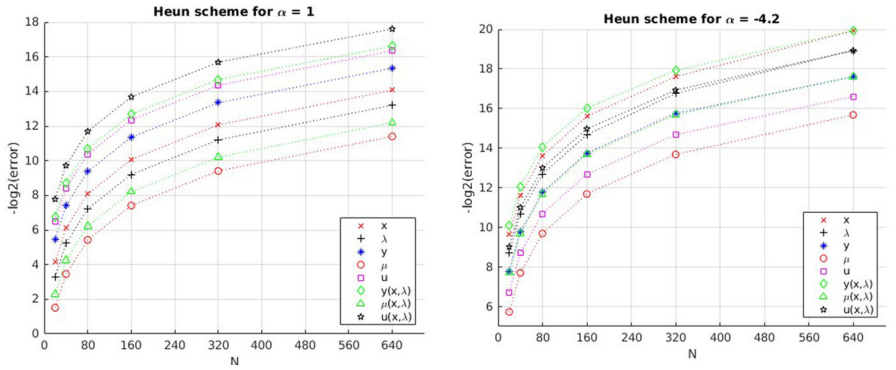

Fig. 1 Errors for the Heun scheme with the values $\alpha = 1$ and $\alpha = -4.2$

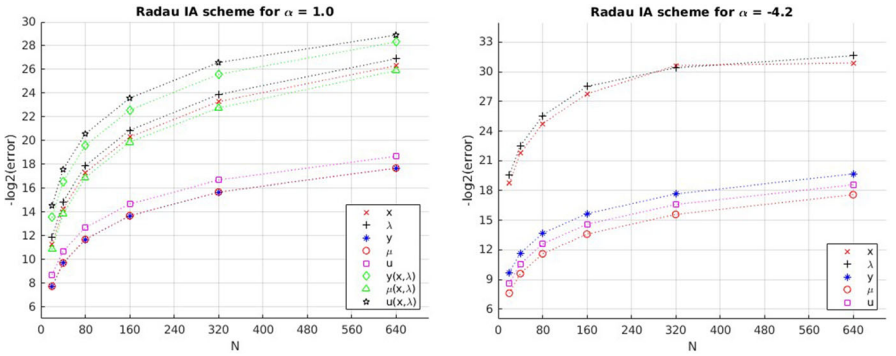

Fig. 2 Errors for the Radau IA scheme with the values $\alpha = 1$ and $\alpha = -4.2$

and control (compare Fig. 2). Since the differential states do not satisfy third order estimates, solving the algebraic constraints will not improve the convergence order for the algebraic states and control. Therefore, these errors were omitted for the $\alpha = -4.2$ case.

7 Conclusions

In this paper we proposed a new type of Runge–Kutta scheme for optimal control problems with DAEs. Instead of approximating the algebraic variable with stage derivatives, we treated the algebraic state like a control. For this discretization scheme we derived necessary conditions that are consistent with the continuous conditions and we were able to establish error estimate for Runge–Kutta schemes applied to optimal control problems with index 1 DAEs. The next step is to apply this method to problems with index 2 DAEs and derive error estimates. However, for optimal control problems with index 2 DAEs some new difficulties occur. There exists a discrepancy between the continuous and discrete necessary conditions, i.e., the adjoint continuous DAE has index 1 while the discrete system approximates an index 2 DAE. In [28] we

were able to overcome this problem by performing an index reduction for the discrete system. In that case we used the implicit Euler discretization. To obtain higher order error estimates, the idea is to perform such a discrete index reduction in the context of Runge–Kutta schemes. Unfortunately, so far we have not been able to find a way to derive consistent necessary conditions.

Acknowledgements This work was supported by the German Research Foundation DFG under the contract GE 1163/8-2.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Statements and Declarations The author has no relevant financial or non-financial interests to disclose. All data generated or analyzed during this study are included in this published article.

Appendix 1: Taylor expansions

In order to estimate the consistency error (32), we apply the Taylor expansion to $\frac{\bar{X}_{i+1}-\bar{X}_i}{h}$ and $F(Z^j(\bar{\mathbf{X}}_i), Y^j(\mathbf{Z}(\bar{\mathbf{X}}_i)))$ with $\bar{X}_i = \hat{X}(t_i)$ for $i = 0, \dots, N-1$. This gives us

$$\begin{aligned}
\frac{\bar{X}_{i+1}-\bar{X}_i}{h} &= \dot{\hat{X}}(t_i) + \frac{h}{2}\ddot{\hat{X}}(t_i) + \frac{h^2}{6}\hat{X}^{(3)}(t_i) + O(h^3) \\
&= \hat{F} + \frac{h}{2}(\hat{F}'_X - \hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\hat{F} \\
&\quad + \frac{h^2}{6}\Big[\hat{F}''_{XX}(\hat{F},\hat{F}) - 2\hat{F}''_{XY}(\hat{F},(\hat{G}'_Y)^{-1}\hat{G}'_X\hat{F}) \\
&\quad + \hat{F}''_{YY}((\hat{G}'_Y)^{-1}\hat{G}'_X\hat{F},(\hat{G}'_Y)^{-1}\hat{G}'_X\hat{F}) + \hat{F}'_X(\hat{F}'_X - \hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\hat{F} \\
&\quad - \hat{F}'_Y(\hat{G}'_Y)^{-1}\Big[\hat{G}''_{XX}(\hat{F},\hat{F}) - 2\hat{G}''_{XY}(\hat{F},(\hat{G}'_Y)^{-1}\hat{G}'_X\hat{F}) \\
&\quad + \hat{G}''_{YY}((\hat{G}'_Y)^{-1}\hat{G}'_X\hat{F},(\hat{G}'_Y)^{-1}\hat{G}'_X\hat{F}) \\
&\quad + \hat{G}'_X(\hat{F}'_X - \hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\hat{F}\Big]\Big] + O(h^3),
\end{aligned}$$

where the right hand side is evaluated at $t = t_i$, i.e., $\hat{F} = F[t_i]$ etc.. We abbreviate $\widetilde{Z}^j = Z^j(\bar{\mathbf{X}}_i)$ and $\widetilde{Y}^j = Y^j(\mathbf{Z}(\bar{\mathbf{X}}_i))$. Then, we have

$$\widetilde{Z}^j - \hat{X}(t_i) = h\sum_{k=1}^{s} \Lambda_{jk} F(\widetilde{Z}^k, \widetilde{Y}^k) = O(h),$$
$$\widetilde{Y}^j - \hat{Y}(t_i) = Y^j((\mathbf{Z}(\bar{\mathbf{X}}_i)) - Y^j(\bar{\mathbf{X}}_i) = O(h)$$

since $Y^j(\cdot)$ is Lipschitz continuous. Furthermore, expanding $F(\widetilde{Z}^j, \widetilde{Y}^j)$ and $G(\widetilde{Z}^j, \widetilde{Y}^j)$ in $(\hat{X}(t_i), \hat{Y}(t_i)) =: (\hat{X}, \hat{Y})$ yields

$$\begin{aligned}
\widetilde{Z}^j - \hat{X} &= h\sum_{k=1}^{s} \Lambda_{jk}\left[\hat{F} + \hat{F}'_X(\widetilde{Z}^k - \hat{X}) + \hat{F}'_Y(\widetilde{Y}^k - \hat{Y})\right] + O(h^3),\\
0 &= G(\widetilde{Z}^j, \widetilde{Y}^j)\\
&= \hat{G} + \hat{G}'_X(\widetilde{Z}^j - \hat{X}) + \hat{G}'_Y(\widetilde{Y}^j - \hat{Y})\\
&\quad + \frac{1}{2}\Big[\hat{G}''_{XX}(\widetilde{Z}^j - \hat{X}, \widetilde{Z}^j - \hat{X}) + 2\hat{G}''_{XY}(\widetilde{Z}^j - \hat{X}, \widetilde{Y}^j - \hat{Y})\\
&\quad + \hat{G}''_{YY}(\widetilde{Y}^j - \hat{Y}, \widetilde{Y}^j - \hat{Y})\Big] + O(h^3).
\end{aligned}$$

Hence, we get

$$\begin{aligned}
\widetilde{Z}^j - \hat{X} &= h\Upsilon_j\hat{F} + h^2\sum_{k=1}^{s}\Lambda_{jk}(\hat{F}'_X - \hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\Upsilon_k\hat{F} + O(h^3)\\
(\widetilde{Y}^j - \hat{Y}) &= -(\hat{G}'_Y)^{-1}\hat{G}'_X\left[h\Upsilon_j\hat{F} + h^2\sum_{k=1}^{s}\Lambda_{jk}(\hat{F}'_X - \hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\Upsilon_k\hat{F}\right]\\
&\quad - \frac{h^2}{2}(\hat{G}'_Y)^{-1}\Big[\hat{G}''_{XX}(\Upsilon_j\hat{F}, \Upsilon_j\hat{F}) - 2\hat{G}''_{XY}(\Upsilon_j\hat{F}, (\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F})\\
&\quad + \hat{G}''_{YY}((\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F}, (\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F})\Big] + O(h^3)
\end{aligned}$$

with $\Upsilon_j = \sum_{k=1}^{s}\Lambda_{jk}$. This implies

$$\begin{aligned}
F(\widetilde{Z}^j, \widetilde{Y}^j) &= \hat{F} + h(\hat{F}'_X - \hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\Upsilon_j\hat{F}\\
&\quad + h^2\Big[\frac{1}{2}\hat{F}''_{XX}(\Upsilon_j\hat{F}, \Upsilon_j\hat{F}) - \hat{F}''_{XY}(\Upsilon_j\hat{F}, (\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F})\\
&\quad + \frac{1}{2}\hat{F}''_{YY}((\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F}, (\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F})\\
&\quad + \hat{F}'_X\sum_{k=1}^{s}\Lambda_{jk}(\hat{F}'_X - \hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\Upsilon_k\hat{F}\\
&\quad - \hat{F}'_Y(\hat{G}'_Y)^{-1}\Big[\frac{1}{2}\hat{G}''_{XX}(\Upsilon_j\hat{F}, \Upsilon_j\hat{F}) - \hat{G}''_{XY}(\Upsilon_j\hat{F}, (\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F})
\end{aligned}$$

$$+\frac{1}{2}\hat{G}''_{YY}((\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F},(\hat{G}'_Y)^{-1}\hat{G}'_X\Upsilon_j\hat{F})$$
$$+\hat{G}'_X\sum_{k=1}^{s}\Lambda_{jk}(\hat{F}'_X-\hat{F}'_Y(\hat{G}'_Y)^{-1}\hat{G}'_X)\Upsilon_k\hat{F}\Bigg]\Bigg]+O(h^3).$$

References

1. Alt, W., Baier, R., Lempio, F., Gerdts, M.: Approximations of linear control problems with bang–bang solutions. Optimization **62**(1), 9–32 (2011). https://doi.org/10.1080/02331934.2011.568619
2. Alt, W., Baier, R., Gerdts, M., Lempio, F.: Error bounds for Euler approximation of linear-quadratic control problems with bang–bang solutions. Numer. Algebra Contr. Optim. **2**(3), 547–570 (2012). https://doi.org/10.3934/naco.2012.2.547
3. Alt, W., Seydenschwanz, M.: An implicit discretization scheme for linear-quadratic control problems with bang–bang solutions. Optim. Methods Softw. **29**(3), 535–560 (2014). https://doi.org/10.1080/10556788.2013.821612
4. Alt, W., Schneider, C.: Linear-quadratic control problems with L^1-control cost. Optim. Contr. Appl. Methods **36**(4), 512–534 (2015). https://doi.org/10.1002/oca.2126
5. Alt, W., Schneider, C., Seydenschwanz, M.: Regularization and implicit Euler discretization of linear-quadratic optimal control problems with bang–bang solutions. Appl. Math. Comput. **287–288**(5), 104–124 (2016). https://doi.org/10.1016/j.amc.2016.04.028
6. Alt, W., Felgenhauer, U., Seydenschwanz, M.: Euler discretization for a class of nonlinear optimal control problems with control appearing linearly. Comput. Optim. Appl. **69**(3), 825–856 (2018). https://doi.org/10.1007/s10589-017-9969-7
7. Betts, J. T.: Practical Methods for Optimal Control and Estimation Using Nonlinear Programming, 2nd edn. Advances in Design and Control. SIAM (2010). https://doi.org/10.1137/1.9780898718577
8. Bonnans, J.F., Festa, A.: Error estimates for the Euler discretization of an optimal control problem with first-order state constraints. SIAM J. Numer. Anal. **55**(2), 445–471 (2017). https://doi.org/10.1137/140999621
9. Brennan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential–Algebraic Equations. volume 14 of Classics in Applied Mathematics. SIAM (1996). https://doi.org/10.1137/1.9781611971224
10. Burger, M., Gerdts, M.: A survey on numerical methods for the simulation of initial value problems with sDAEs. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential–Algebraic Equations IV. Differential–Algebraic Equations Forum. Springer, Berlin (2017). https://doi.org/10.1007/978-3-319-46618-7_5
11. Dontchev, A.L., Hager, W.W., Poore, A.B., Yang, B.: Optimality, stability, and convergence in nonlinear control. Appl. Math. Optim. **31**, 297–326 (1995). https://doi.org/10.1007/BF01215994
12. Dontchev, A.L., Hager, W.W., Malanowski, K.: Error bounds for Euler approximation of a state and control constrained optimal control problem. Numer. Funct. Anal. Optim. **21**(5–6), 653–682 (2000). https://doi.org/10.1080/01630560008816979
13. Dontchev, A.L., Hager, W.W., Veliov, V.M.: Second-order Runge–Kutta approximations in control constrained optimal control. SIAM J. Numer. Anal. **38**(1), 202–226 (2000). https://doi.org/10.1137/S0036142999351765
14. Dontchev, A.L., Hager, W.W.: The Euler approximation in state constrained optimal control. Math. Comput. **70**(233), 173–203 (2001). https://doi.org/10.1090/S0025-5718-00-01184-4
15. Dunn, J.C., Tian, T.: Variants of the Kuhn–Tucker sufficient conditions in cones of nonnegative functions. SIAM J. Contr. Optim. **30**(6), 1361–1384 (1992). https://doi.org/10.1137/0330072
16. Gerdts, M.: Optimal Control of ODEs and DAEs. De Gruyter, Berlin (2012). https://doi.org/10.1515/9783110249996
17. Gerdts, M., Kunkel, M.: Convergence analysis of Euler discretization of control-state constrained optimal control problems with controls of bounded variation. J. Ind. Manag. Optim. **10**(1), 311–336 (2014). https://doi.org/10.3934/JIMO.2014.10.311
18. Hager, W.W.: Runge–Kutta methods in optimal control and the transformed adjoint system. Numer. Math. **87**(2), 247–282 (2000). https://doi.org/10.1007/s002110000178

19. Hairer, E., Lubich, C., Roche, M.: The Numerical Solution of Differential–Algebraic Systems by Runge–Kutta Methods, vol. 1409. Springer, Berlin (1989). https://doi.org/10.1007/BFb0093947
20. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems, Springer Series in Computational Mathematics, vol. 14, 2nd edn. Springer, Berlin (1996). https://doi.org/10.1007/978-3-662-09947-6
21. Haunschmied, J.L., Pietrus, A., Veliov, V.M.: The Euler method for linear control systems revisited. In: Proceedings of the 9th International Conference on Large-Scale Scientific Computations, Sozopol, pp. 90–97 (2013). https://doi.org/10.1007/978-3-662-43880-0_9
22. Ioffe, A.D., Tihomirov, V.M.: Theory of Extremal Problems. Studies in Mathematics and Its Applications, vol. 6. North-Holland Publishing Company, Amsterdam (1979)
23. Kraft, D.: FORTRAN-Programme zur numerischen Lösung optimaler Steuerungsprobleme. DFVLR-Mitteilung, vol. 80. DFVLR (1980)
24. Kunkel, P., Mehrmann, V.: Differential–Algebraic Equations: Analysis and Numerical Solution. EMS Textbooks in Mathematics. European Mathematical Society (2006). https://doi.org/10.4171/017
25. Malanowski, K., Büskens, C., Maurer, H.: Convergence of Approximations to Nonlinear Optimal Control Problems. Lecture Notes in Pure and Applied Mathematics. In: Fiacco, A.V. (ed.) Mathematical programming with data perturbations. CRC Press, Boca Raton (1997). https://doi.org/10.1201/9781003072119-12
26. Martens, B., Gerdts, M.: Convergence analysis of the implicit Euler-discretization and sufficient conditions for optimal control problems subject to index-one differential–algebraic equations. Set-Valued Var Anal **27**, 405–431 (2019). https://doi.org/10.1007/S11228-018-0471-X
27. Martens, B.: Necessary conditions, sufficient conditions, and convergence analysis for optimal control problems with differential–algebraic equations. PhD thesis, Universität der Bundeswehr München (2019) https://doi.org/10.1007/978-3-030-53905-4_10
28. Martens, B., Gerdts, M.: Convergence analysis for approximations of optimal control problems subject to higher index differential–algebraic equations and mixed control-state constraints. SIAM J. Contr. Optim. **58**(1), 1–33 (2020). https://doi.org/10.1137/18m1219382
29. Martens, B., Gerdts, M.: Error analysis for the implicit Euler discretization of linear-quadratic control problems with higher index DAEs and bang-bang solutions. In: Reis, T., Grundel, S., Schöps, S. (eds.) Progress in differential–algebraic Equations II. Differential–algebraic equations forum. Springer, Berlin (2020). https://doi.org/10.1007/978-3-030-53905-4_10
30. Martens, B., Gerdts, M.: Convergence analysis for approximations of optimal control problems subject to higher index differential–algebraic equations and pure state constraints. SIAM J. Contr. Optim. **59**(3), 1903–1926 (2021). https://doi.org/10.1137/20M1353952
31. Martens, B., Schneider, C.: Error analysis for the implicit Euler discretization of affine optimal control problems with index two DAEs. Pure Appl. Funct. Anal. **6**(6), 1383–1414 (2021)
32. Osmolovskii, N.P., Veliov, V.M.: Metric sub-regularity in optimal control of affine problems with free end state. ESAIM: COCV (2020). https://doi.org/10.1051/COCV/2019046
33. Pietrus, A., Scarinci, T., Veliov, V.M.: High order discrete approximations to Mayer's problems for linear systems. SIAM J. Contr. Optim. **56**(1), 102–119 (2018). https://doi.org/10.1137/16M1079142
34. Scarinci, T., Veliov, V.M.: Higher-order numerical scheme for linear-quadratic problems with bang-bang controls. Comput. Optim. Appl. **69**, 403–422 (2018). https://doi.org/10.1007/s10589-017-9948-z
35. Schneider, C., Wachsmuth, G.: Regularization and discretization error estimates for optimal control of ODEs with group sparsity. ESAIM: COCV **24**(2), 811–834 (2018). https://doi.org/10.1051/COCV/2017049
36. Seydenschwanz, M.: Convergence results for the discrete regularization of linear-quadratic control problems with bang–bang solutions. Comput. Optim. Appl. **61**, 731–760 (2015). https://doi.org/10.1007/s10589-015-9730-z
37. von Stryk, O.: Numerische Lösung optimaler Steuerungsprobleme: Diskretisierung, Parameteroptimierung und Berechnung der adjungierten Variablen. Ph.D. thesis, Technische Universität München (1994)
38. Veliov, V.M.: Error analysis of discrete approximations to bang–bang optimal control problems: the linear case. Contr. Cybern. **34**(3), 967–982 (2005)