Convergence of Nonconvergent IRK Discretizations of Optimal Control Problems

J. T. Betts* N. Biehn† S. L. Campbell‡ W. P. Huffman§

Abstract

It has been observed that optimization codes are sometimes able to solve state constrained optimal control problems with discretizations which do not converge when used as an integrator on the constrained dynamics. Understanding this phenomenon could lead to a more robust design for direct transcription codes as well as better test problems. This paper examines how this phenomena can occur. The difference between solving index three Differential Algebraic Equations (DAEs) using the trapezoid method in the context of direct transcription for optimal control problems and a straight forward Implicit Runge-Kutta formulation of the same trapezoidal discretization is analyzed. It is shown through numerical experience and theory that the two can differ as much as $O(1/h^3)$ in the control. Moreover, a small sacrifice in the accuracy of the states in the early stages of the trapezoidal method allows better accuracy in the control, where more precise solutions converge to an incorrect solution. The theoretical results are used to explain computational observations.

Key words: Optimal control, Implicit runge-kutta.

AMS subject classifications: 65K10.

1 Introduction

This paper studies the difference between solving index 3 differential algebraic equations (DAEs) using the trapezoid method (TR) in the direct transcription of optimal control problems versus a straight forward implementation of the implicit Runge-Kutta formulation (IRK) of the trapezoid rule. Many optimal control problems contain inequality path constraints on the states and/or controls. Along the constraints, the resulting system becomes a DAE. Depending on which set of variables are constrained, a low or high index DAE results. For more on DAEs and their indices, see [4].

The direct transcription software SOCS (Sparse Optimal Control Software) developed by Boeing [1, 2], is used throughout this paper. The results obtained are relevant to other direct transcription algorithms with similar strategies. In particular, SOCS passes the discretized problem directly to the NLP solver and does not first integrate the dynamics to reduce the dimension of the problem. The discretization formulas used by SOCS are mathematically equivalent to the Lobatto IIIA formulas TR and Hermite-Simpson (HS). These methods are symmetric, stiffly stable and produce desirable sparsity patterns in the Jacobians of the defect constraints in the associated Non-Linear Program (NLP). They also can be implemented as collocation methods. This paper is concerned primarily with TR. Computational experience and a more technical analysis shows the same phenomenon occurring when using HS.

TR implemented as an IRK converges for DAEs of index 0,1, and 2. For problems with index greater than 2, examples can be constructed for which TR converges to an incorrect solution [3, 7]. Numerical experience has shown

*Mathematics and Engineering Analysis, The Boeing Company, P.O. Box 3707, MS 7L-21, Seattle, Washington 98124-2207. john.t.betts@boeing.com.

†North Carolina State University, Operations Research Program, Raleigh, NC 27695-7913. ndbiehn@unity.ncsu.edu.

‡North Carolina State University, Department of Mathematics, Raleigh, NC 27695-8205. slc@math.ncsu.edu. Research supported in part by NSF Grants DMS-9714811 and DMS-9802259.

§Mathematics and Engineering Analysis, The Boeing Company, P.O. Box 3707, MS 7L-21, Seattle, Washington 98124-2207. whuffman@redwood.rt.cs.boeing.com.

This space left blank for copyright notice.

that when used in SOCS, TR can converge to the true optimal solution while the straight forward IRK formulation does not [3]. There are two questions to be resolved. One is how SOCS is able to converge. The second is why does SOCS converge. This paper begins the task of understanding this phenomena by answering the "how" question. Considerable research exists on the convergence of discretizations of optimization problems. We cite only [5, 6]. However, previous work on the state constrained problem always assumes that the discretization converges on a problem with the constraint holding identically. This analysis does not apply to the phenomena investigated here.

The next section presents the main result of this paper which theoretically predicts the difference between the direct transcription and IRK solutions. Numerical experiments are then given in Section 3. In order to avoid confusion we use SOCS TR for the SOCS solution using the TR discretization and IRK TR for the TR solution implemented as an IRK when constraints are active. Finally, our investigations are motivated by the optimization of tool paths. In these problems, complexity and problem size sometimes forces the termination of the iterations in SOCS at coarser grids than those required for termination according to the termination criteria. For this reason we are very interested not only in behavior on sufficiently fine grids but also on "coarser" grids. A full version of this paper with proofs is available on third author's web site.

2 The Control Problem & Main Theorem

The optimal control problem studied here contains an objective function, state equations, and constraints;

$$J(u) = \phi(t_f) + \int_{t_0}^{t_f} L(y,u,t)\,dt \tag{1a}$$

$$y'(t) = f(y(t),u(t),t), \quad y(t_0) = y_0 \tag{1b}$$

$$y_L \le y(t) \le y_U, \quad u_L \le u(t) \le u_U \tag{1c}$$

$$g_L \le g(y(t),u(t),t) \le g_U \tag{1d}$$

The continuous optimal control problem is then transcribed to a finite dimensional Non-Linear Program (NLP). Let $y(t_k), u(t_k)$ be y_k, u_k. Then $[y_0, u_0, y_1, u_1, \ldots, y_N, u_N]^T$ are the NLP variables with $N+1$ grid points. Using TR, we approximate the state equations using the defect constraints $0 = y_k - y_{k-1} - \frac{h_k}{2}(f(y_k,u_k,t_k) + f(y_{k-1},u_{k-1},t_{k-1}))$. The objective function (1a) is approximated using a trapezoidal quadrature. The NLP is then solved using a sparse sequential quadratic programming algorithm whose solution is a discrete approximation to the continuous optimal control problem [2].

Consider the following optimal control problem,

$$J(u) = \min_u \int_0^4 x^2(t) + 10^{-3}u^2(t)\,dt \tag{2a}$$

$$x' = v, \quad v' = u \tag{2b}$$

$$0 \le x - 15 + (t-4)^4 \tag{2c}$$

$$x(0) = 5, \qquad v(0) = 1 \tag{2d}$$

The constraint (2c) is active on the interval $\left[\frac{34}{15},\ 4\right]$ giving the index 3 DAE (3) in (x,v,u) on $[t_0,t_f] = \left[\frac{34}{15},\ 4\right]$;

$$x' = v, \quad v' = u \tag{3a}$$

$$x = 15 - (t-4)^4 \tag{3b}$$

$$x(t_0) = 15 - (t_0-4)^4, \qquad v(t_0) = -4(t_0-4)^3, \quad u(t_0) = -12(t_0-4)^2 \tag{3c}$$

Figure 1 shows u obtained by IRK TR and SOCS TR with $N = 10$. Notice that SOCS TR is much closer to the true solution. In fact, the solution given by SOCS seems to converge with N, while the IRK solution does not. If this convergence can be fully and rigorously understood, then it might be possible to safely use direct transcription codes with such discretizations when solving some classes of higher index DAEs. This paper will discuss how these two similar methods can differ.

Since the numerical solutions of the control differ by large amounts, there must be a fundamental difference between using TR directly on the DAE and in a direct transcription formulation. The NLP solution is known to satisfy the given constraints up to a tolerance which is looser than the accuracy to which the IRK equations are usually solved. It was suggsted in [3] that this NLP "slop" provides one possible explanation for the large differences in the computed control

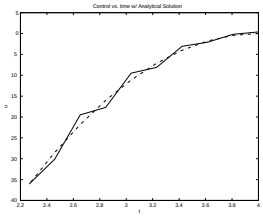

Figure 1a: SOCS TR control and true solution

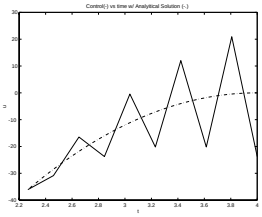

Figure 1b: IRK TR control and true solution

between IRK TR and SOCS TR. We will see that the "slop conjecture" can account for computational observations on fine grids when SOCS is close to convergence but connot fully explain what is observed on coarser grids. A careful analysis of an index three problem proves enlightening.

Consider the state equations with constraints (4),

$$x' = Ax + Bu \tag{4a}$$

$$0 \leq Cx - \widehat{g}(t) \tag{4b}$$

where A is $m \times m$, B is $m \times q$, C is $r \times m$ and $\widehat{g}(t)$ is $r \times 1$. Suppose that the optimal solution has (4b) as an equality constraint over an interval $[a, b]$. This gives us a DAE (4) in (x, u) along this interval which can be written as

$$Ey' = \widehat{A}y - g(t) \tag{5}$$

with $E = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}$, $\widehat{A} = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$, $g(t) = \begin{bmatrix} 0 \\ \widehat{g}(t) \end{bmatrix}, y = \begin{bmatrix} x \\ u \end{bmatrix}$. Then TR for (5) implemented as an IRK is

$$EY_1' - \widehat{A}y_{n-1} = -g(t_{n-1}) \tag{6a}$$

$$EY_2' - \widehat{A}y_{n-1} - \tfrac{h}{2}\widehat{A}(Y_1' + Y_2') = -g(t_n) \tag{6b}$$

$$y_n = y_{n-1} + \tfrac{h}{2}(Y_1' + Y_2') \tag{6c}$$

where the Y_i', are the stage derivatives for $i = 1, 2$. Let $Y_i' = \begin{bmatrix} X_i' \\ U_i' \end{bmatrix}$, $i = 1, 2$. Then (6) can be rewritten as

$$X_1' = Ax_{n-1} + Bu_{n-1} \tag{7a}$$

$$X_2' = A(x_{n-1} + \frac{h}{2}(X_1' + X_2')) + Bu_n \tag{7b}$$

$$0 = Cx_n - \widehat{g}(t_n) \tag{7c}$$

$$x_n = x_{n-1} + \frac{h}{2}(X_1' + X_2') \tag{7d}$$

In SOCS TR the numerical solution $(\widetilde{x}, \widetilde{u})$ must satisfy (8) at every mesh point to be considered feasible;

$$\widetilde{x}_n - \widetilde{x}_{n-1} - \frac{h}{2}(A\widetilde{x}_{n-1} + B\widetilde{u}_{n-1} + A\widetilde{x}_n + B\widetilde{u}_n) + \delta_{n-1} = 0 \tag{8a}$$

$$C\widetilde{x}_n - \widehat{g}_n + \epsilon_{n-1} = 0 \tag{8b}$$

δ, ϵ are tolerances for the accuracy to which we solve (8a), (8b). We initially view these as the tolerances in the NLP solve. Define $\widetilde{X}_1' = A\widetilde{x}_{n-1} + B\widetilde{u}_{n-1}$, $\widetilde{X}_2' = A\widetilde{x}_n + B\widetilde{u}_n$. Then a calculation shows that we may write (8) as

$$\widetilde{X}_1' = A\widetilde{x}_{n-1} + B\widetilde{u}_{n-1} \tag{9a}$$

$$\widetilde{X}_2' = A(\widetilde{x}_{n-1} + \frac{h}{2}(\widetilde{X}_1' + \widetilde{X}_2')) + B\widetilde{u}_n + A\delta_{n-1} \tag{9b}$$

$$\widetilde{x}_n = \widetilde{x}_{n-1} + \frac{h}{2}(\widetilde{X}_1' + \widetilde{X}_2') + \delta_{n-1} \tag{9c}$$

$$C\widetilde{x}_n = \widehat{g}_n + \epsilon_{n-1} \tag{9d}$$

Comparing (7) to (9) we see that we may consider the SOCS solution as a perturbation of the IRK solution. If we subtract (7) from (8) we get a system for the difference between the SOCS TR and the IRK TR solutions in the form,

$$Gz_n = Hz_{n-1} + \gamma_{n-1} \tag{10}$$

$$G = \begin{bmatrix} I & 0 & -\frac{h}{2}I & -\frac{h}{2}I \\ 0 & -B & -\frac{h}{2}A & I - \frac{h}{2}A \\ 0 & 0 & I & 0 \\ C & 0 & 0 & 0 \end{bmatrix},\ H = \begin{bmatrix} I & 0 & 0 & 0 \\ A & 0 & 0 & 0 \\ A & B & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \gamma_n = \begin{bmatrix} \delta_n \\ A\delta_n \\ 0 \\ \epsilon_n \end{bmatrix}, z_n = \begin{bmatrix} \widetilde{x}_n - x_n \\ \widetilde{u}_n - u_n \\ \widetilde{X}_1' - X_1' \\ \widetilde{X}_2' - X_2' \end{bmatrix} \tag{11}$$

Suppose that G is invertible and solve (10) for z_n,

$$z_{n+1} = G^{-1}Hz_n + G^{-1}\gamma_n \tag{12}$$

The solution of (12) is $z_k = (G^{-1}H)^k z_0 + \sum_{j=0}^{k-1}(G^{-1}H)^{k-j-1}G^{-1}\gamma_j$. We assume that $z_0 = 0$, and thus obtain the difference between the IRK and SOCS TR solutions starting from the same initial conditions,

$$z_k = \sum_{j=0}^{k-1}(G^{-1}H)^{k-j-1}G^{-1}\gamma_j \tag{13}$$

The analysis and discussion that follows differs from the usual error analysis of IRK methods. It is not enough to convert (13) to an order estimate. We need to know the actual size of the perturbation z_k and we need to consider mesh widths that may not be small. These results are given in Theorem 1.

(13) holds for all DAEs of index 1 or higher. We know that IRK TR converges for DAEs of index 1 and 2. However, for DAEs whose index is greater than 2, IRK TR may not converge [3]. We assume the DAE (5) is an index 3 problem. Then CB must be singular. For time invariant problems we can decouple the different index subsystems using linear time invariant coordinate changes. Thus we may assume that $CB = 0$. Without loss of generality we may further assume that C is full row rank, B is full column rank and there exist matrices $\widehat{P}$ and $\widehat{Q}$ so that $\widehat{P}C\widehat{Q}^{-1} = \begin{bmatrix} I & 0 \end{bmatrix}$, $\widehat{Q}B = \begin{bmatrix} 0 \\ I \end{bmatrix}$. Thus we may assume that

$$C = \begin{bmatrix} I & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ I \end{bmatrix} \tag{14}$$

A linear time invariant index 3 DAE with $CB = 0$ must also have CAB invertible. Define $Q = I - \frac{h}{2}A$ and let $Q = \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix}$, $A = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}$where the partitioning conforms to C and B. From the special form given for C and B we know that CAB is invertible if and only if A_2 is invertible. Q is invertible for all but a finite number of nonzero h. The value of $Q_2 = C(I - \frac{h}{2}A)^{-1}B$ is given analytically in terms of A by $Q_2 = -2hA_2(-4I + 2hA_4 + 2hA_1 - h^2A_1A_4 + h^2A_2A_3)^{-1}$ which is invertible for small nonzero h since A_2 is invertible. Then Q_2 is invertible for all but a finite number of h. With these assumptions we may now state the following theorem.

Theorem 2.1 *Suppose that the DAE (5) is index three and (7), (9), (14) hold. Assume that CAB and $C(I-\frac{h}{2}A)^{-1}B$ are invertible. Then the difference in the states between the IRK TR and SOCS TR solutions is*

$$\bar{x}_k - x_k = \sum_{j=0}^{k-1}(-1)^{k-1-j}\left(\begin{bmatrix} 0 & 0 \\ -Q_4Q_2^{-1} & I \end{bmatrix}\delta_j + \begin{bmatrix} 0 & 0 \\ -\frac{h}{2}W & 0 \end{bmatrix}A\delta_j + \begin{bmatrix} 0 \\ -Y + Q_4Q_2^{-1} \end{bmatrix}\epsilon_j \right) \tag{15}$$

The difference between the controls is

$$\begin{aligned} \bar{u}_k - u_k \quad = \quad & \sum_{j=0}^{k-1}(-1)^{k-1-j}\frac{2}{h}\Big\{[-2(k-j-1)Q_4Q_2^{-1} - Q_2^{-1} \quad 2(k-j-1)I]\delta_j \\ & + \frac{h}{2}[-2(k-j-1)W - Q_2^{-1}Q_1 \quad -I]A\delta_j + [-2(k-j-2)Y + 2Q_2^{-1}Q_1 + 2(k-j-1)Q_4Q_2^{-1}]\epsilon_k \Big\} \end{aligned} \tag{16}$$

where $W = Q_4Q_2^{-1}Q_1 - Q_3, Y = Q_4Q_2^{-1} - 2W$.

Page restrictions prohibits giving the proof. Note that the controls can differ considerably depending on the size of δ and ϵ. If we suppose that we are on the interval $[0, 1]$, then the value of k at the end of the interval would act like $1/h$. Q_2^{-1} may possess a $1/h$ factor as well. Thus a single δ_j will introduce a difference between the controls of at least $\bar{u}_k - u_k \sim O\left(\frac{1}{h^3}\right)$. A constant perturbation δ is also $O(h^{-3})$ due to cancelation. The worst case given by δ_j with alternating sign is $O(h^{-4})$. This is not possible on coarse grids because of the one sided nature of constraint chatter.

3 Numerical Experiments

The preceding analysis shows that the feasible solutions examined by SOCS can differ substantially from the IRK solutions. Once h is sufficiently small, approximations of the true optimal solution become feasible. In order to determine if this is the full story we conduct some computational tests on coarser grids. The numerical solutions on $[34/15, 4]$, $N = 10$, for (2) obtained from SOCS TR and IRK TR were subtracted The results are given in Table 1a.

Time	$\lvert\widetilde{x} - x\rvert$	$\lvert\widetilde{v} - v\rvert$	$\lvert\widetilde{u} - u\rvert$
2.2667	0.0000	0.0000	0.0000
2.4593	0.0070	0.0726	0.7540
2.6519	0.0000	0.1452	3.0159
2.8444	0.0000	0.1452	6.0318
3.0370	0.0000	0.1452	9.0477
3.2296	0.0000	0.1452	12.0637
3.4222	0.0000	0.1452	15.0796
3.6148	0.0000	0.1452	18.0955
3.8074	0.0000	0.1452	21.1114
4.0000	0	0.1452	24.1273

Table 1a: Absolute Difference of Solutions

$\lvert\widetilde{v} - v\rvert$		$\lvert\widetilde{u} - u\rvert$	
Actual Diff	Pred. Diff.	Actual Diff.	Pred. Diff.
0.0000	0	0.0000	0
0.0726	0.0727	0.7540	0.7548
0.1452	0.1454	3.0159	3.0193
0.1452	0.1454	6.0318	6.0449
0.1452	0.1454	9.0477	9.0673
0.1452	0.1454	12.0637	12.0897
0.1452	0.1454	15.0796	15.1121
0.1452	0.1454	18.0955	18.1346
0.1452	0.1454	21.1114	21.1570
0.1452	0.1454	24.1273	24.1794

Table 1b: Predicted and Absolute Differences of Solutions

SOCS solves (8a) up to the square root of machine precision. In addition, when the constraint is active, (8b) is also solved to the square root of machine precision. Theorem 1 gives absolute error bounds. After substituting $\epsilon_k = \delta_k = \sqrt{\varepsilon_{mach}}$ into the formulas of Theorem 1, we observe that the differences between the two solutions are much larger than the estimates allow. For (2) on a grid of size $N = 10$, Theorem 1 predicted a difference on the order of 10^{-5}. As seen in Figures 1a and 1b, the two solutions differ by much more than that. The NLP tolerances were not enough to explain what we are seeing computationally on coarse grids. Closer examination of the solution obtained from SOCS shows that not all variables $\widetilde{x}_k$ lie directly on the constraint. In fact, $\widetilde{x}_1$ is consistently above the constraint for grids larger than $N = 5$, as illustrated by the $|\widetilde{x} - x|$ column in the second line of Table 1a.

We again initially set the tolerances at the SOCS defaults

$$\delta_k = \begin{bmatrix} \sqrt{\varepsilon_{mach}} \\ \sqrt{\varepsilon_{mach}} \end{bmatrix} \tag{17}$$

However, we set ϵ_1 equal to the distance $\widetilde{x}_1$ is from the true solution. When $N = 10$ we have

$$\epsilon_1 = 0.007, \quad \epsilon_k = \sqrt{\varepsilon_{mach}}, \quad k \geq 2. \tag{18}$$

The results of applying Theorem 1 using these new tolerances and $N = 10$ are given in Table 1b. The small error created by the constraint chatter in the early stages of SOCS TR propagates down the numerical solution as seen in (13) to create large differences between IRK and SOCS later in time. In addition, as h gets smaller the distance from the constraint also gets smaller.

Table 1b suggests that for this test problem the perturbations (17) and (18) account for the ability of SOCS to find the optimum. As an additional check, rather than simply integrating the DAE (5) using IRK TR, we force the TR equations to include the addition of ϵ_1 at the first time step. That is, at t_1 we have, $0 = Cx_1 - g(t_1) - \epsilon_1$ as part of the IRK equations given in (7a)-(7d). This results in the much more accurate approximation of the control in Figure 2a. Note the perturbed IRK TR solution of Figure 2a is now very close to the SOCS TR solution of Figure 1a.

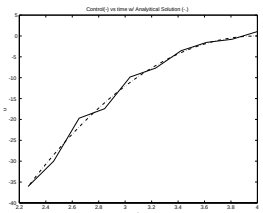

Figure 2a: TR on the DAE with perturbation

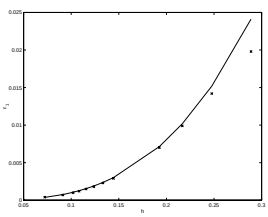

Figure 2b: Perturbation Size versus Step Size

The initial constraint chatter shown in Table 1 is present at a wide range of mesh sizes. Figure 2b plots the size of the constraint chatter (perturbation) versus the step size. The chatter is roughly $O(h^3)$. Theorem 2.1 says perturbations

can be enlarged by $1/(h^3)$. Thus an h^3 perturbation is reasonable for inducing an $O(1)$ difference between IRK and SOCS. Constraint chatter is not unique to problem (2). We have performed a number of numerical experiments on a variety of index 3 and index 4 problems. In each problem we saw chattering occuring on many levels. When solving an index 4 problem the constrained solution bounced on and off the curve with greater frequency than the index 3 case. This behavior also occurs with nonlinear problems. In addition, the chatter seems to occur at both the beginning and the end of the state constrained interval for boundary value problems. For example, consider the problem

$$\min \int_0^4 x_1^2 + x_2^2 + u^2 \, dt \tag{19a}$$

$$x_1' = x_3, \quad x_2' = x_4, \quad x_3' = x_2 x_1 - u + 1, \quad x_4' = x_1 u - x_2 \tag{19b}$$

$$\frac{1}{2} \geq x_2 - x_2 x_1 \tag{19c}$$

State and controls for (19) are given in Figure 3a. Examination of the constraint residual on the interval on which it is active shows that there is again chatter which goes to zero with mesh size. However, now the chatter is asymmetrical and changes shape with the mesh. Figure 3b gives an enlarged view of the plot of $x_2 - x_2 x_1$ which is equal to $\frac{1}{2}$ when the constraint is active.

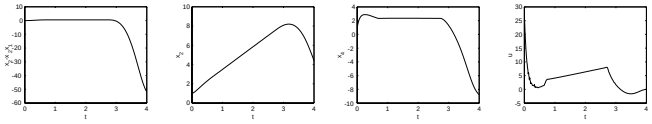

Figure 3a: State and Control Solutions

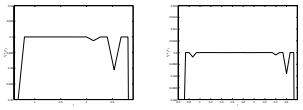

Figure 3b: Chatter along constraint, N=31 and N=61.

4 Conclusions

This paper examines how discretizations in a direct transcription method for optimal control can converge for DAEs where an application of the mathematically equivalent IRK formulation does not. The two approaches are viewed as perturbations of each other. A formula for calculating the maximum distance between the two methods was obtained for linear index three problems. These estimates were verified numerically and gave insight as to why such a large difference could be observed. A perturbation away from the constraint on the order of $O(h^3)$ propagated down the solution to allow for good approximations to the control. SOCS solves a boundary value problem while the IRK solves a DAE intial value problem. On a coarse grid SOCS will seek a global solution of the NLP problem no matter what the discretization error is. The chatter enables the NLP to find a cheaper solution than that provided by the IRK. A rigorous analysis of when and why SOCS is able to do this is under investigation.

References

[1] J. T. Betts and W. P. Huffman, *Mesh Refinement in Direct Transcription Methods for Optimal Control*, Optimal Control Applications & Methods, 19 (1998), 1-21.

[2] J.T. Betts and W.P. Huffman, *Application of sparse nonlinear programming to trajectory optimization*, J. Guidance, Control Dyn., 15, 198-206, 1992.

[3] N. Biehn, S.L. Campbell, L. Jay and T. Westbrook, *Some comments on DAE theory for IRK methods and trajectory optimization*, J. Comp. Appl. Math, to appear..

[4] K. Brenan, S.L. Campbell, and L. Petzold, *Numerical Solution of Initial Value Problems in Differential Algebraic Equations*, SIAM, 1996.

[5] A. L. Dontchev and W. W. Hager, *A new approach to Lipschitz continuity in state constrained optimal control*, Systems & Control Letters, 35 (1998), 137-143.

[6] P. J. Enright and B. A. Conway, *Discrete approximations to optimal trajectories using direct transcription and nonlinear programming*, AIAA Journal of Guidance Control and Dynamics, 15 (1992), 994-1002.

[7] L. Jay, *Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2*, BIT, 33 (1993), 137-150.