The properties of differential-algebraic equations representing optimal control problems

Roland England[1]
Susana Gómez[2]
René Lamour[3]

1 Introduction

The purpose of this paper is to express an optimal control problem in terms of a system of Differential-Algebraic Equations (DAEs) and to investigate their properties. This system is obtained using calculus of variations to get the Kuhn-Tucker conditions. The inequalities associated with the complementarity conditions are converted to equalities by the addition of a new variable, combining the slack variable and the corresponding Lagrange multiplier. The sign of this variable indicates whether the constraint is active or not.

The well-conditioning of the problem can be expressed in terms of the index of the resulting system of DAEs, which is a measure of the difficulty involved in obtaining a numerical solution. The concept of the *tractability index* is introduced as a general purpose way of determining the index. But a projector related to the tractability index makes it possible, in the case of higher index, to determine exactly which equations must be differentiated in order to reduce the index.

Other ways of solving the optimal control problem involve the discretization of the original problem to convert it to a finite dimensional constrained optimization problem. In each of the following references ([1, 3, 4, 16, 20, 21, 22, 23, 26]) all the variables are discretized in one way or another. The discretization may be carried out only on the control variables ([2, 8, 9]), and any inequality constraints on the state variables might be approximated by a penalty function ([19, 24, 25]), whereas in the method given in this paper they are treated exactly.

The examples used here are the minimization of the time to travel a fixed distance, subject to bounds on the acceleration and on the velocity, and the maximization of the yield of a component on a packed bed reactor. These problems have index varying from 1 to 3. The first two examples have simple analytic solutions; the third example appeared to be more complicated, but an analytic solution is presented.

[1]Dept. of Applied Maths., The Open University, Milton Keynes, MK7 6AA Great Britain.
E-mail, r.england@open.ac.uk
[2]IIMAS, National University of Mexico, Apdo. Postal 20-726, Mexico D.F..
E-mail, susanag@servidor.unam.mx
[3]Humboldt-University of Berlin, I. of Mathematics, D-10099, Berlin, Germany.
E-mail, lamour@math.hu-berlin.de

2 General transformation process

2.1 Formulation of an optimal control problem

Consider an optimal control problem, expressed as a dynamical system of ordinary differential equations subject to a number of initial and terminal conditions, and to a number of inequalities on the state variables and the control variables, and with some unknown constant parameters. The objective function has the form of an integral of some function of the same state and control variables and parameters.

$$\begin{aligned} \text{minimize} \quad J(\underline{u}) &= \int_0^b h(\underline{y}, \underline{u}, \underline{c})\, dx && (1) \\ \text{subject to :} \quad \underline{y}' &= \underline{f}(\underline{y}, \underline{u}, \underline{c}), \quad y_i(0) = y_{i0} \ (i \in \mathcal{I}), \quad y_j(b) = y_{j1} \ (j \in \mathcal{F}), && (2) \\ \underline{0} &\leq \underline{g}(\underline{y}, \underline{u}, \underline{c}). && (3) \end{aligned}$$

Here, $\mathcal{I}$ and $\mathcal{F}$ are subsets of the state variables $\underline{y}$ for which initial and terminal values, respectively, are specified.

2.2 Calculus of variations

As we wish to transform this problem to a system of DAEs, we use the variational formulation. Introducing the perturbations $\underline{\delta y}(x)$, $\underline{\delta u}(x)$, $\underline{\delta c}$ constant, and the Lagrange multipliers $\underline{v}(x)$ for the differential equations, and $\underline{w}(x)$ for the inequalities,

$$\begin{aligned} \int_0^b (h_y^T \underline{\delta y} + h_u^T \underline{\delta u} + h_c^T \underline{\delta c})\, dx = & \int_0^b \underline{v}^T (\underline{\delta y}' - f_y \underline{\delta y} - f_u \underline{\delta u} - f_c \underline{\delta c})\, dx \\ + & \int_0^b \underline{w}^T (g_y \underline{\delta y} + g_u \underline{\delta u} + g_c \underline{\delta c})\, dx, \qquad (4) \end{aligned}$$

$$\delta y_i(0) = 0 \ (i \in \mathcal{I}), \qquad \delta y_j(b) = 0 \ (j \in \mathcal{F}), \qquad (5)$$

$$w_i g_i(\underline{y}(x), \underline{u}(x), \underline{c}) = 0 \ (\forall x, \forall i), \qquad \underline{0} \leq \underline{w}(x). \qquad (6)$$

To eliminate the term $\underline{\delta y}'$ using integration by parts, under the assumption that both $\underline{\delta y}(x)$ and $\underline{v}(x)$ are continuous and piecewise differentiable,

$$\int_0^b \underline{v}^T \underline{\delta y}'\, dx = -\int_0^b \underline{v}'^T \underline{\delta y}\, dx,$$

where

$$v_i(0) = 0 \ (i \notin \mathcal{I}), \quad v_j(b) = 0 \ (j \notin \mathcal{F}). \qquad (7)$$

The perturbations $\underline{\delta y}(x)$, $\underline{\delta u}(x)$, $\underline{\delta c}$ are independent, and therefore:

$$\begin{aligned}
\int_0^b h_y^T \underline{\delta y}\, dx &= \int_0^b \underline{v}^T(\underline{\delta y}' - f_y \underline{\delta y})\, dx + \int_0^b \underline{w}^T g_y \underline{\delta y}\, dx = \int_0^b (-\underline{v}'^T - \underline{v}^T f_y + \underline{w}^T g_y)\underline{\delta y}\, dx, \\
\int_0^b h_u^T \underline{\delta u}\, dx &= \int_0^b (-\underline{v}^T f_u + \underline{w}^T g_u)\underline{\delta u}\, dx, \\
\int_0^b h_c^T \underline{\delta c}\, dx &= \int_0^b (-\underline{v}^T f_c + \underline{w}^T g_c)\underline{\delta c}\, dx.
\end{aligned}$$

Apart from the continuity condition on $\underline{\delta y}(x)$, the perturbations $\underline{\delta y}(x), \underline{\delta u}(x), \underline{\delta c}$ are also arbitrary, and so:

$$\underline{v}'^T = -\underline{v}^T f_y + \underline{w}^T g_y - h_y^T, \tag{8}$$

$$\underline{0}^T = -\underline{v}^T f_u + \underline{w}^T g_u - h_u^T, \tag{9}$$

$$\underline{0}^T = \int_0^b (-\underline{v}^T f_c + \underline{w}^T g_c - h_c^T) dx. \tag{10}$$

The original differential equations (and boundary conditions) (2), together with the adjoint equations (8–10), boundary conditions (7), inequalities (3), and complementarity conditions (6), form the Kuhn-Tucker necessary conditions for $(\underline{y}, \underline{u}, \underline{c})$ to be a minimizer of the functional $J(\underline{u})$ in equation (1).

In order to express the integral equation (10) as a differential equation, new variables $\underline{r}(x)$ may be introduced, corresponding to the constants $\underline{c}$, and satisfying

$$\underline{r}'^T = -\underline{v}^T f_c + \underline{w}^T g_c - h_c^T, \quad \underline{r}(0) = \underline{0}, \quad \underline{r}(b) = \underline{0}. \tag{11}$$

Introducing a Hamilton function as

$$H(\underline{y}, \underline{v}, \underline{c}, \underline{u}, \underline{w}) := -\underline{f}^T(\underline{y}, \underline{u}, \underline{c})\underline{v} + \underline{g}^T(\underline{y}, \underline{u}, \underline{c})\underline{w} - h(\underline{y}, \underline{u}, \underline{c})$$

we can express the right hand side of (8-11) in terms of H.

2.3 Elimination of inequalities

In order to eliminate the inequalities on $\underline{g}$ and $\underline{w}$ in the complementarity conditions (3) and (6), new variables $\underline{p}$ may be introduced, such that

$$\underline{p} = \underline{g} - \underline{w}, \quad \underline{g} = \max(\underline{0}, \underline{p}) := \underline{p}^+, \quad \underline{w} = \max(-\underline{p}, \underline{0}) := \underline{p}^-,$$

and so the Kuhn-Tucker conditions (2),(3),(6–9) and (11) may be expressed in the form of the following system of DAEs subject to initial and terminal conditions:

$$\begin{array}{llllll}
\underline{y}' = & \underline{f}(\underline{y}, \underline{u}, \underline{c}), & y_i(0) = & y_{i0}\ (i \in \mathcal{I}), & y_j(b) = & y_{j1}\ (j \in \mathcal{F}), \\
\underline{v}' = & H_y(\underline{y}, \underline{v}, \underline{c}, \underline{u}, \underline{p}^-), & v_i(0) = & 0\ (i \notin \mathcal{I}), & v_j(b) = & 0\ (j \notin \mathcal{F}), \\
\underline{r}' = & H_c(\underline{y}, \underline{v}, \underline{c}, \underline{u}, \underline{p}^-), & \underline{r}(0) = & \underline{0}, & \underline{r}(b) = & \underline{0}, \\
\underline{0} = & H_u(\underline{y}, \underline{v}, \underline{c}, \underline{u}, \underline{p}^-), & & & & \\
\underline{0} = & \underline{p}^+ - \underline{g}(\underline{y}, \underline{u}, \underline{c}). & & & &
\end{array} \tag{12}$$

3 Examples

To demonstrate this transformation we apply (12) to three known examples.

3.1 Problem 1 (see [4, 16])

A simple example of such a problem is that of
Minimum time to cover a fixed distance.

3.1.1 Problem statement

Let the time taken to cover the distance (300 units) be $t_f > 0$. Then the problem is to

$$\begin{aligned}
\text{minimize} \quad & t_f \\
\text{subject to :} \quad & \frac{dx_1}{dt} = x_2, \qquad x_1(0) = 0, \quad x_1(t_f) = 300, \\
& \frac{dx_2}{dt} = u, \qquad x_2(0) = 0, \quad x_2(t_f) = 0, \\
& -2 \le u \le 1,
\end{aligned}$$

where u is the acceleration.

3.1.2 Conversion to the General Formulation

In order to express this problem in the form given in (1–3), we define variables and constants as follows:

$$x_1 = y_1,\ x_2 = y_2,\ t = t_f s,\ u = z,\ t_f = c > 0,$$

and so obtain

$$\begin{aligned}
\text{minimize} \quad & t_f = \int_0^1 c\, ds \\
\text{subject to :} \quad & \frac{dy_1}{ds} = y_2 c, \qquad y_1(0) = 0, \qquad y_1(1) = 300, \\
& \frac{dy_2}{ds} = zc, \qquad y_2(0) = 0, \qquad y_2(1) = 0, \\
& 0 \le 2 + z, \\
& 0 \le 1 - z.
\end{aligned}$$

The exact solution of this problem is:
$c = 30$, and
for $0 \le s \le \frac{2}{3}$

$$\begin{aligned}
y_1 &= \tfrac{1}{2} c^2 s^2, \\
y_2 &= cs, \\
z &= 1;
\end{aligned} \tag{13}$$

for $\frac{2}{3} \leq s \leq 1$

$$\begin{aligned} y_1 &= 300 - c^2(1-s)^2, \\ y_2 &= 2c(1-s), \\ z &= -2. \end{aligned} \tag{14}$$

The Hamiltonian function is

$$H = -y_2 c v_1 - z c v_2 + (2+z)p_1^- + (1-z)p_2^- - c.$$

3.1.3 System of DAEs

Using the procedure outlined above, this gives rise to the following system of DAEs (12) without inequalities:

$$\begin{aligned} \frac{dy_1}{ds} &= y_2 c, & y_1(0) = 0, \quad y_1(1) = 300, \\ \frac{dy_2}{ds} &= zc, & y_2(0) = 0, \quad y_2(1) = 0, \\ \frac{dv_1}{ds} &= 0, \\ \frac{dv_2}{ds} &= -cv_1, \\ \frac{dr}{ds} &= -y_2 v_1 - z v_2 - 1, & r(0) = 0, \quad r(1) = 0, \\ 0 &= -cv_2 + p_1^- - p_2^-, \\ 0 &= p_1^+ - 2 - z, \\ 0 &= p_2^+ - 1 + z. \end{aligned}$$

This system has 8 variables and 1 unknown constant which must satisfy 5 equations with derivatives and 3 algebraic equations, and has 6 boundary conditions for the 5 differentiated variables and the constant.

3.2 Problem 2 (see [16])

A slightly more complicated problem is given by imposing a *Speed limit.*

3.2.1 Problem statement

Let the speed limit be k, where the other variables have the same meaning as before. The problem is to

$$\begin{aligned}
\text{minimize} \quad & t_f \\
\text{subject to :} \quad & \frac{dx_1}{dt} = x_2, \qquad x_1(0) = 0, \quad x_1(t_f) = 300, \\
& \frac{dx_2}{dt} = u, \qquad x_2(0) = 0, \quad x_2(t_f) = 0, \\
& -2 \le u \le 1, \quad x_2 \le k.
\end{aligned}$$

3.2.2 Conversion to the General Formulation

We define variables and constants as before

$$x_1 = y_1,\ x_2 = y_2,\ t = t_f s,\ u = z,\ t_f = c > 0,$$

and so obtain

$$\begin{aligned}
\text{minimize} \quad & t_f = \int_0^1 c\, ds \\
\text{subject to :} \quad & \frac{dy_1}{ds} = y_2 c, \qquad y_1(0) = 0, \qquad y_1(1) = 300, \\
& \frac{dy_2}{ds} = zc, \qquad y_2(0) = 0, \qquad y_2(1) = 0, \\
& 0 \le 2 + z, \\
& 0 \le 1 - z, \\
& 0 \le k - y_2.
\end{aligned}$$

If $k \ge 20$ the solution of this problem is identically to that of Problem 1.
If $k \le 20$ the exact solution is:
$c = 30 + \frac{3}{4k}(20 - k)^2$, and
for $0 \le s \le \frac{k}{c}$

$$\begin{aligned}
y_1 &= \tfrac{1}{2}c^2 s^2, \\
y_2 &= cs, \\
z &= 1;
\end{aligned} \tag{15}$$

for $\frac{k}{c} \le s \le 1 - \frac{k}{2c}$

$$\begin{aligned}
y_1 &= kcs - \tfrac{1}{2}k^2, \\
y_2 &= k, \\
z &= 0;
\end{aligned} \tag{16}$$

for $1 - \frac{k}{2c} \leq s \leq 1$

$$\begin{aligned} y_1 &= 300 - c^2(1-s)^2, \\ y_2 &= 2c(1-s), \\ z &= -2. \end{aligned} \tag{17}$$

The Hamiltonian function is

$$H = -y_2 c v_1 - z c v_2 + (2+z)p_1^- + (1-z)p_2^- + (k-y_2)p_3^- - c.$$

3.2.3 System of DAEs

This gives rise to the system of DAEs (12) without inequalities:

$$\begin{aligned} \frac{dy_1}{ds} &= y_2 c, && y_1(0) = 0, \quad y_1(1) = 300, \\ \frac{dy_2}{ds} &= zc, && y_2(0) = 0, \quad y_2(1) = 0, \\ \frac{dv_1}{ds} &= 0, \\ \frac{dv_2}{ds} &= -cv_1 - p_3^-, \\ \frac{dr}{ds} &= -y_2 v_1 - z v_2 - 1, && r(0) = 0, \quad r(1) = 0, \\ 0 &= -cv_2 + p_1^- - p_2^-, \\ 0 &= p_1^+ - 2 - z, \\ 0 &= p_2^+ - 1 + z, \\ 0 &= p_3^+ - k + y_2. \end{aligned}$$

This system has 9 variables and 1 unknown constant which must satisfy 5 equations with derivatives and 4 algebraic equations, and has 6 boundary conditions for the 5 differentiated variables and the constant.

3.3 Problem 3 (see [16], [13])

An example of a higher index problem: *Catalyst mixing for packed bed reactor.*

3.3.1 Problem statement

In [16], this problem is given as one of maximizing the concentration $\big(1 - z^a(t_f) - z^b(t_f)\big)$. The statement of the problem is as follows:

$$
\begin{aligned}
\max_{F} \quad & \big(1 - z^a(t_f) - z^b(t_f)\big) \\
\text{subject to :} \quad & \frac{dz^a}{dt} = F(10z^b - z^a), && z^a(0) = 1, \\
& \frac{dz^b}{dt} = F(z^a - 10z^b) - (1-F)z^b, && z^b(0) = 0, \\
& 0 \le F \le 1.
\end{aligned}
$$

3.3.2 Conversion to the General Formulation

In order to express the problem in the form given in (1–3), the revised objective function must be rewritten as

$$\min_F \big(z^a(t_f) + z^b(t_f) - 1\big) = \min_F \int_0^{t_f} (F-1)z^b \, dt$$

and the inequality constraints as $0 \le F,\ 0 \le 1 - F$.
Note that, in this problem, t is a distance and t_f is a given, constant, reactor length.
Extending [13],[5], it is possible to give the exact solution.
If $t_f \le \frac{1}{11}\ln(1+\sqrt{12.1}) + \ln(1+\sqrt{0.1}) =: d_c \approx 0.4111$ and t_c satisfies

$$e^{t_c}(e^{11t_c} + 10) = 11e^{t_f},$$

the solution is:
for $0 \le t \le t_c$

$$
\begin{aligned}
z^a &= \tfrac{1}{11}(10 + e^{-11t}), \\
z^b &= \tfrac{1}{11}(1 - e^{-11t}), \\
F &= 1;
\end{aligned}
\tag{18}
$$

for $t_c \le t \le t_f$

$$
\begin{aligned}
z^a &= \tfrac{1}{11}(10 + e^{-11t}), \\
z^b &= \tfrac{1}{11}(1 - e^{-11t})e^{-(t-t_c)}, \\
F &= 0.
\end{aligned}
\tag{19}
$$

If $t_f \ge d_c$ the solution consists of three parts.
Let $t_a := \frac{1}{11}\ln(1+\sqrt{12.1}) \approx 0.1363$ and $t_b := t_f - \ln(1+\sqrt{0.1}) \approx t_f - 0.2748$. Then
for $0 \le t \le t_a$

$$
\begin{aligned}
z^a &= \tfrac{1}{11}(10 + e^{-11t}), \\
z^b &= \tfrac{1}{11}(1 - e^{-11t}), \\
F &= 1;
\end{aligned}
\tag{20}
$$

for $t_a \le t \le t_b$

$$\begin{aligned} z^a &= \tfrac{1}{111}(100+\sqrt{10})e^{\frac{1}{52}(-6+\sqrt{10})(t-t_a)}, \\ z^b &= \tfrac{1}{111}(11-\sqrt{10})e^{\frac{1}{52}(-6+\sqrt{10})(t-t_a)}, \\ F &= \frac{5\sqrt{10}-4}{52} \approx 0.2271; \end{aligned} \tag{21}$$

for $t_b \le t \le t_f$

$$\begin{aligned} z^a &= \tfrac{1}{111}(100+\sqrt{10})e^{\frac{1}{52}(-6+\sqrt{10})(t_b-t_a)}, \\ z^b &= \tfrac{1}{111}(11-\sqrt{10})e^{-(t-t_b)+\frac{1}{52}(-6+\sqrt{10})(t_b-t_a)}, \\ F &= 0. \end{aligned} \tag{22}$$

The Hamiltonian is

$$H = -F(10z^b - z^a)(v^a - v^b) + (1-F)z^b(v^b+1) + Fp_1^- + (1-F)p_2^-.$$

3.3.3 System of DAEs

The problem gives rise to the system of DAEs (12) without inequalities:

$$\begin{aligned} \frac{dz^a}{dt} &= (10z^b - z^a)F, & z^a(0) &= 1, \\ \frac{dz^b}{dt} &= (z^a - 10z^b)F - (1-F)z_b, & z^b(0) &= 0, \\ \frac{dv^a}{dt} &= (v^a - v^b)F, & v^a(t_f) &= 0, \\ \frac{dv^b}{dt} &= 10(v^b - v^a)F + (v^b+1)(1-F), & v^b(t_f) &= 0, \\ 0 &= (z^a - 10z^b)(v^a - v^b) - z^b(v^b+1) + p_1^- - p_2^-, \\ 0 &= p_1^+ - F, \\ 0 &= p_2^+ - 1 + F. \end{aligned}$$

We now have a system with 7 variables which must satisfy 4 equations with derivatives and 3 algebraic equations but no inequalities, and which has 4 boundary conditions for the 4 variables which are differentiated.

4 The tractability Index Concept

4.1 Short Introduction

In the case of linear DAEs, the index indicates how often we have to differentiate parts of the right hand side of the DAE to obtain an expression for the solution. Therefore the index

describes the difficulty involved in solving a system numerically.
A way of determining the index of a system of DAEs is given by the tractability index concept (see also [17]). The motivation of tractability index comes from an equivalent transformation of a DAE without differentiation. This is important e.g. if the data of the DAE have low smoothness properties.
The definition of the tractability index is based on a matrix chain G_i, $i \geq 0$ in the following way. Consider a DAE in quasilinear form

$$\underline{F}((D(t)\underline{x}(t))', \underline{x}(t), t) := A(\underline{x}, t)(D(t)\underline{x})' + \underline{b}(\underline{x}, t) = \underline{0}, \tag{23}$$

where $\underline{F}(\underline{y}, \underline{x}, t) : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n, A(\underline{x}, t) \in \mathbb{R}^{n\times m}, D(t) \in \mathbb{R}^{m\times n}$ and $\underline{b}(\underline{x}, t) \in \mathbb{R}^n$. We prefer systems of DAEs with properly stated leading term, because of their clearer description and their better numerical properties (see [11],[12]). Properly stated leading term means that $\ker A(\underline{x}, t) \oplus \mathrm{im}\, D(t) = \mathbb{R}^m$ and the projector realizing this splitting is continuously differentiable (see [18]). With

$$B(\underline{y}, \underline{x}, t) := F_x(\underline{y}, \underline{x}, t)$$

(we will drop the arguments) a matrix chain is defined by

$$\begin{aligned} G_0 &:= AD, \quad B_0 := B, \\ G_{i+1} &:= G_i + B_i Q_i, \\ B_{i+1} &:= (B_i - G_{i+1} D^- (DP_0 ... P_{i+1} D^-)' DP_0 ... P_{i-1}) P_i, \end{aligned} \tag{24}$$

where Q_i denotes a projector onto $N_i := \ker G_i$, $P_i := I - Q_i$ and D^- describes a reflexive generalized inverse of D, i.e. $D = DD^-D$, $D^- = D^-DD^-$ and additionally $D^-D = P_0$.

Definition 4.1 *(See [18]) An equation (23) with properly stated leading term is said to be a DAE with tractability index μ on the interval I, if there is a continuous matrix function sequence (24) such that*

$$\left.\begin{array}{l} \text{(a) } G_i \text{ has constant } \operatorname{rank} r_i \text{ on } I, \\ \text{(b) } N_0 \oplus N_1 \oplus \cdots \oplus N_{i-1} \subseteq \ker Q_i, \\ \text{(c) } Q_i \in C(I, \mathbb{R}^{n\times n}),\ DP_0 \cdots P_i D^- \in C^1(I, \mathbb{R}^{m\times m}) \\ \text{(d) } 0 \leq r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = n. \end{array}\right\} 0 \leq i \leq \mu,$$

To check the index of a DAE we have to check the ranks of the matrices $G_i, 0 \leq i \leq \mu$.

Remark: The ranks and therefore the index are independent of linear transformations.

By means of the tractability index concept, it is also possible to get a cheap way to reduce the index of a higher index system of DAEs.

If we consider a system of DAEs of semiexplicit structure (23)

$$A(\underline{x}, t)(D(t)\underline{x})' + \underline{b}(\underline{x}, t) = \underline{0}$$

of index k (i.e. G_k remains nonsingular) the system of DAEs

$$A(\underline{x}, t)(D(t)\underline{x})' + (I - W_{k-1})\underline{b}(\underline{x}, t) + W_{k-1}\frac{d}{dt}(W_{k-1}\underline{b}(\underline{x}, t)) = \underline{0} \tag{25}$$

has, for a wide class of DAEs, index $k-1$, where W_{k-1} denotes a projector along $\operatorname{im} G_{k-1}$. This is proved for linear equations and for index-2 equations of structure (23) (see [7]). For index-3 equations we prove the following theorem.

Theorem 4.2 *Let $A(D\underline{x})' + \underline{b}(\underline{x},t) = \underline{0}$ be an index-3 system of DAEs with constant matrices A and D. Let W_2 be a constant projector along $\operatorname{im} G_2$, $(W_2\underline{b})(\underline{x},t) = (W_2\underline{b})(P_0\underline{x},t)$ and $I + Q_2G_3^{-1}(B(\underline{y},\underline{x},.)D^-\underline{y})_xP_0Q_1Q_2$ nonsingular for arbitrary $\underline{y}$. Then the system of DAEs*

$$A(D\underline{x})' + (I-W_2)\underline{b}(\underline{x},t) + W_2\frac{d}{dt}(W_2\underline{b}(\underline{x},t)) = \underline{0} \tag{26}$$

has index 2.

Proof:
Equation (26) can be written in greater detail as

$$A(D\underline{x})' + (I-W_2)\underline{b}(\underline{x},t) + W_2(W_2BD^-(D\underline{x})' + (W_2\underline{b})_t) = \underline{0} \tag{27}$$

The matrix chain of (26) with matrices linearized in $(\underline{y},\underline{x})$ is given by the following

$$\begin{aligned} \tilde{G}_0 &= G_0 + W_2BP_0, \quad \text{with } \tilde{Q}_0 = Q_0 \\ \tilde{B}_0 &= (I-W_2)B + W_2(W_2BD^-\underline{y} + (W_2\underline{b})_t)_xP_0, \end{aligned}$$

where $B = \underline{b}_x(\underline{x},t)$. The next chain elements are given by

$$\begin{aligned} \tilde{G}_1 = \tilde{G}_0 + \tilde{B}_0\tilde{Q}_0 &= G_0 + W_2BP_0 + (I-W_2)BQ_0 \\ &= G_1 + W_2BP_0. \end{aligned}$$

From $0 = W_{i+1}G_{i+1} = W_i(G_i + B_iQ_i)$ we derive $W_{i+1}W_i = W_{i+1}$ and $W_{i+1}B_iQ_i = W_{i+1}B_0Q_i = 0$. Therefore $W_2BP_0 = W_2BP_0P_1$ and $\tilde{G}_1$ and G_1 have the same nullspace, i.e. we can choose $\tilde{Q}_1 = Q_1$. Then

$$\begin{aligned} \tilde{B}_1 &= \tilde{B}_0P_0 - \tilde{G}_1D^-(DP_1D^-)'D \\ &= ((I-W_2)B + W_2(W_2BD^-\underline{y} + (W_2\underline{b})_t)_xP_0 - \tilde{G}_1D^-(DP_1D^-)'D. \end{aligned}$$

The next step gives

$$\begin{aligned} \tilde{G}_2 &= \tilde{G}_1 + \tilde{B}_1\tilde{Q}_1 \\ &= G_1 + W_2BP_0 + (I-W_2)BP_0Q_1 \\ &\qquad + W_2(W_2BD^-\underline{y})_xP_0Q_1 + \underbrace{W_2(W_2\underline{b})_t)_xP_0Q_1}_{=W_2BD^-(DP_1D^-)'DQ_1} - \tilde{G}_1D^-(DP_1D^-)'DQ_1 \\ &= (G_1 + BP_0Q_1)(I - P_1D^-(DP_1D^-)'DQ_1) + W_2BP_0P_1 + W_2(W_2B\underline{D^-}\underline{y})_xP_0Q_1 \\ &= G_2 + W_2BP_0P_1 + W_2(W_2B\underline{D^-}\underline{y})_xP_0Q_1. \end{aligned}$$

Consider $\tilde{G}_2\underline{z} = \underline{0}$. Multiplying

$$(G_2 + W_2BP_0P_1 + W_2(W_2BD^-\underline{y})_xP_0Q_1)\underline{z} = \underline{0}$$

by $(I - W_2)$ we get $G_2\underline{z} = \underline{0}$, which leads to $\underline{z} = Q_2\underline{z}$, and by W_2 we get

$$(\underbrace{W_2BP_0P_1Q_2}_{=W_2B_2Q_2} + W_2(W_2BD^-\underline{y})_xP_0Q_1Q_2)\underline{z} = \underline{0}.$$

Using the special projector $W_2 := G_3Q_2G_3^{-1}$ we obtain

$$(W_2B_2Q_2 + W_2(BD^-\underline{y})_xP_0Q_1Q_2)\underline{z} = (G_3Q_2\underbrace{G_3^{-1}B_2Q_2}_{=Q_2} + G_3Q_2G_3^{-1}(BD^-\underline{y})_xP_0Q_1Q_2)\underline{z} = 0. \tag{28}$$

Multiplying (28) by G_3 leads to

$$Q_2z + Q_2G_3^{-1}(BD^-\underline{y})_xP_0Q_1Q_2)z = (I + Q_2G_3^{-1}(BD^-\underline{y})_xP_0Q_1Q_2)Q_2\underline{z}$$

and hence $Q_2\underline{z} = \underline{0}$.
This means that $\tilde{G}_2$ is nonsingular and (26) has index 2. □
Remark: A check of the nonsingularity condition of $I + Q_2G_3^{-1}(BD^-\underline{y})_xP_0Q_1Q_2$ is not trivial, because it needs e.g. G_3^{-1}. But it can be seen immediately that the condition is fulfilled for linear DAEs. The complicated theoretical computation also makes a direct numerical algorithm necessary.

4.2 The Tractability Index of the DAEs

We will investigate the index of the DAE (12) in general form, applying the tractability index concept. To get a DAE, which has as many equations as unknowns, we introduce an extra ODE for $\underline{c}$. The DAE is given by

$$\begin{array}{rcl} \underline{y}' = & \underline{f}(\underline{y}, \underline{u}, \underline{c}), \\ \underline{c}' = & \underline{0} \\ \underline{v}' = & H_y(\underline{y}, \underline{v}, \underline{c}, \underline{u}, \underline{p}^-), \\ \underline{r}' = & H_c(\underline{y}, \underline{v}, \underline{c}, \underline{u}, \underline{p}^-), \\ \underline{0} = & H_u(\underline{y}, \underline{v}, \underline{c}, \underline{u}, \underline{p}^-), \\ \underline{0} = & \underline{p}^+ - \underline{g}(\underline{y}, \underline{u}, \underline{c}). \end{array} \tag{29}$$

The matrices A, D and B are

$$A = \begin{pmatrix} I & & & & \\ & I & & & \\ & & I & & \\ & & & I & \\ & & & 0 & \\ & & & 0 & \end{pmatrix}, \quad D = \begin{pmatrix} I & & & & & \\ & I & & & & \\ & & I & & & \\ & & & I & 0 & 0 \end{pmatrix}$$

and with the unknown vector $x = \left(y^T, c^T, v^T, r^T, u^T, p^T\right)^T$ we obtain

$$B = b'_x = \left(\begin{array}{cccc|cc} -f_y & -f_c & 0 & 0 & -f_u & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -H_{yy} & -H_{yc} & -H_{yv} & 0 & -H_{yu} & -H_{yp} \\ -H_{cy} & -H_{cc} & -H_{cv} & 0 & -H_{cu} & -H_{cp} \\ \hline H_{uy} & H_{uc} & H_{uv} & 0 & H_{uu} & H_{up} \\ -g_y & -g_c & 0 & 0 & -g_u & \underline{p}_{\underline{p}}^{+} \end{array}\right) =: \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}.$$

where $\underline{p}_{\underline{p}}^{+} := \frac{\partial \underline{p}^+}{\partial \underline{p}} = \left(\frac{\partial p_i^+}{\partial p_j}\right)$. The first matrix chain element is

$$G_0 = AD = \begin{pmatrix} I & & & & & \\ & I & & & & \\ & & I & & & \\ & & & I & & \\ & & & & 0 & \\ & & & & & 0 \end{pmatrix} \text{ and } Q_0 = \begin{pmatrix} 0 & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & 0 & & \\ & & & & I & \\ & & & & & I \end{pmatrix}$$

is a nullspace projector of G_0. The next chain element G_1 will be calculated as $G_1 = G_0 + BQ_0$. We find

$$G_1 = \begin{pmatrix} I & B_{12} \\ & B_{22} \end{pmatrix}.$$

It is easy to see that G_1 is nonsingular iff B_{22} remains nonsingular, i.e. we have an index 1 DAE. If B_{22} is singular and we know a nullspace projector of B_{22}, we can construct a nullspace projector Q_1 of G_1. For DAEs with tractability index we know that $N_k \cap N_{k+1} = \{0\}$ (see [18]),in particular for $k = 0, \{0\} = N_0 \cap N_1 = \ker G_0 \cap (\ker G_0 \cap \ker BQ_0) = \ker G_0 \cap \ker BQ_0$. Therefore $\begin{pmatrix} B_{12} \\ B_{22} \end{pmatrix}$ must have full rank.

Let $\bar{Q}_1$ be a nullspace projector of B_{22}; then if $R = \bar{Q}_1 S_{B_2}^{-1} B_{12}^T$ and $S_{B_2} := \begin{pmatrix} B_{12}^T & B_{22}^T \end{pmatrix} \begin{pmatrix} B_{12} \\ B_{22} \end{pmatrix}$,

$$Q_1 = \begin{pmatrix} B_{12}R & 0 \\ -R & 0 \end{pmatrix} \tag{30}$$

represents a nullspace projector of G_1 with $Q_1 Q_0 = 0$. If we know Q_1 we can calculate the next matrix chain element

$$G_2 = G_1 + B_1 Q_1 = (\underbrace{G_1 + B_0 P_0 Q_1}_{=:\mathcal{G}_2})(I - P_1 D^-(DP_1 D^-)' DQ_1). \tag{31}$$

To investigate the singularity of G_2 it is sufficient to investigate the singularity of $\mathcal{G}_2$, because the second factor in the representation (31) of G_2 remains nonsingular.

In order to construct a nullspace projector $\bar{Q}_1$ of B_{22} the structure of the given problem is sometimes useful. Very often the objective function and the right hand sides f of the ODE and g of the inequality depend only linearly on the control u. In that case $H_{uu} \equiv 0$. If additionally g_u has full rank the following Lemma is valid.

Lemma 4.3 *1. The matrix* $M = \begin{pmatrix} 0 & g_u^T \underline{p}_p^- \\ -g_u & \underline{p}_p^+ \end{pmatrix}$ *with full rank* g_u *is nonsingular iff* $Z = (\underline{p}_p^+ - g_u(g_u^T g_u)^{-1} g_u^T)$ *is nonsingular and*

2. if M *is singular a nullspace projector onto* $\ker M$ *is given by* $\bar{Q} = \begin{pmatrix} 0 & (g_u^T g_u)^{-1} g_u^T \tilde{Q} \\ 0 & \tilde{Q} \end{pmatrix}$, *where* $\tilde{Q}$ *describes a nullspace projector onto* $\ker Z$.

Proof: 1. From $p = \underline{p}^+ - \underline{p}^-$ we obtain $I = \underline{p}_p^+ - \underline{p}_p^-$. Using $\underline{p}_p^- = \underline{p}_p^+ - I$ we obtain from $g_u^T Z = g_u^T \underline{p}_p^+ - g_u^T = g_u^T \underline{p}_p^-$. Multiplying M with a nonsingular matrix

$$M \begin{pmatrix} I & (g_u^T g_u)^{-1} g_u^T \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & g_u^T \\ -g_u & I \end{pmatrix} \begin{pmatrix} I & \\ & Z \end{pmatrix}.$$

We obtain a factorization into two matrices. The first factor remains nonsingular for full rank g_u and it is shown that M is nonsingular iff Z is nonsingular.

2. Let $\tilde{Q}$ be a nullspace projector onto $\ker Z$. From $g_u^T Z \tilde{Q} = 0$ we obtain, using $I = \underline{p}_p^+ - \underline{p}_p^-$, that $g_u^T \underline{p}_p^- \tilde{Q} = 0$. Then it is easy to see that $M\bar{Q} = 0$. □

4.3 Application to the Examples

Let us now consider the examples given in section 3. We will study the examples using matrix chains.

4.3.1 Problem 1

The vector of dependent variables is given by $\underline{x} = (y_1, y_2, c, v_1, v_2, r, z, p_1, p_2)$.

$$G_0 = \begin{pmatrix} I_6 & \\ & 0_3 \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 0_6 & \\ & I_3 \end{pmatrix}.$$

The matrix B is given by

$$B = \left(\begin{array}{cccccc|ccc} 0 & -c & -y_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & z & 0 & 0 & 0 & -c & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & v_1 & c & 0 & 0 & 0 & 0 & 0 \\ 0 & v_1 & 0 & y_2 & z & 0 & v_2 & 0 & 0 \\ \hline 0 & 0 & -v_2 & 0 & -c & 0 & 0 & p_{1_{p_1}}^- & -p_{2_{p_2}}^- \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & p_{1_{p_1}}^+ & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & p_{2_{p_2}}^+ \end{array}\right),$$

and the next chain matrix is calculated as

$$G_1 = \left(\begin{array}{cccccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -c & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & v_2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & p^-_{1_{p_1}} & -p^-_{2_{p_2}} \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & p^+_{1_{p_1}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & p^+_{2_{p_2}} \end{array}\right).$$

The nonsingularity of G_1 depends on the nonsingularity of $\begin{pmatrix} 0 & p^-_{1_{p_1}} & -p^-_{2_{p_2}} \\ -1 & p^+_{1_{p_1}} & 0 \\ 1 & 0 & p^+_{2_{p_2}} \end{pmatrix}$. This matrix has exactly the structure of matrix M of Lemma 4.3. The relevant matrix Z is given by

$$Z = \begin{pmatrix} p^+_{1_{p_1}} - \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & p^+_{2_{p_2}} - \frac{1}{2} \end{pmatrix}.$$

Here two cases are possible: either p_1 and p_2 have the same sign or they do not.
For different signs of p_1 and p_2, $\det Z = -\frac{1}{2}$, which means that G_1 is nonsingular and the DAE has index 1. If both p_1 and p_2 are negative the last two equations create a contradiction, because each of them gives a fixed value of z, but they are different (-2 and 1); in terms of the original problem statement both constraints are active simultaneously. The DAE has no tractability index in that case, because $\begin{pmatrix} B_{12} \\ B_{22} \end{pmatrix}$ does not have full rank. Therefore the pencil $(\lambda G_0 + B)$ is singular.
If both p_1 and p_2 are positive the last two equations do not determine the control z. The algebraic equation $0 = -cv_2$ implies that $v_2 = 0$ and from the fourth equation $v_1 = 0$. The acceleration z appears in the second equation but, together with the first equation, that provides insufficient information to determine z. The DAE has no tractability index in that case (see Section 5). The pencil $(\lambda G_0 + B)$ is singular.

4.3.2 Problem 2

The vector of dependent variables is given by $\underline{x} = (y_1, y_2, c, v_1, v_2, r, z, p_1, p_2, p_3)$.

$$G_0 = \begin{pmatrix} I_6 & \\ & 0_4 \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 0_6 & \\ & I_4 \end{pmatrix}.$$

The matrix B is given by

$$B = \left(\begin{array}{cccccc|cccc} 0 & -c & -y_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -z & 0 & 0 & 0 & -c & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & v_1 & c & 0 & 0 & 0 & 0 & 0 & p^-_{3_{p_3}} \\ 0 & v_1 & 0 & y_2 & z & 0 & v_2 & 0 & 0 & 0 \\ \hline 0 & 0 & -v_2 & 0 & -c & 0 & 0 & p^-_{1_{p_1}} & -p^-_{2_{p_2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & p^+_{1_{p_1}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & p^+_{2_{p_2}} & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p^+_{3_{p_3}} \end{array}\right),$$

and the next chain matrix is calculated as

$$G_1 = \left(\begin{array}{cccccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -c & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & p^-_{3_{p_3}} \\ 0 & 0 & 0 & 0 & 0 & 1 & v_2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & p^-_{1_{p_1}} & -p^-_{2_{p_2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & p^+_{1_{p_1}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & p^+_{2_{p_2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p^+_{3_{p_3}} \end{array}\right).$$

The matrix in the lower right corner, which determines the singularity of G_1, has the structure of M in Lemma 4.3 and the associated matrix Z has the structure

$$Z = \begin{pmatrix} -p^+_{1_{p_1}} & -\frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & p^+_{2_{p_2}} & -\frac{1}{2} & 0 \\ 0 & 0 & p^+_{3_{p_3}} \end{pmatrix}.$$

If $p_3 > 0$ we discover the same cases as in Problem 1: if p_1 and p_2 have different signs then Z and therefore M is nonsingular and the DAE has index 1; if p_1 and p_2 have the same sign no index is defined.

If $p_3 < 0$, from the last equation, $y_2 = k$, and from the second equation $z = 0$. The equations for p_1^+ and p_2^+ gives $p_1^+ = 2$ and $p_2^+ = 1$ and both p_1 and p_2 are positive.

If $p_3 < 0$ and both p_1 and p_2 are negative, $\begin{pmatrix} B_{12} \\ B_{22} \end{pmatrix}$ does not have full rank (all three constraints are active). If $p_3 < 0$ and p_1 and p_2 have different signs the matrix pencil $\lambda G_0 + B$ is singular.

If $p_3 < 0$, $p_1 > 0$ and $p_2 > 0$, Z looks like $Z = \frac{1}{2}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ and a nullspace projector is $\tilde{Q} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Using the nullspace projector of M constructed in Lemma 4.3 we obtain a nullspace projector of G_1 by (30) as

$$Q_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c^2\mu & 0 & 0 & 0 & -cv_2\mu & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -cv_2\mu & 0 & 0 & 0 & v_2^2\mu & 0 & 0 & 0 & 0 \\ 0 & c\mu & 0 & 0 & 0 & -v_2\mu & 0 & 0 & 0 & 0 \\ 0 & c\mu & 0 & 0 & 0 & -v_2\mu & 0 & 0 & 0 & 0 \\ 0 & -c\mu & 0 & 0 & 0 & v_2\mu & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

with $\mu = \frac{1}{c^2+v_2^2}$ and we obtain $\mathcal{G}_2$ by (31) as

$$\mathcal{G}_2 = \begin{pmatrix} 1 & -c^3\mu & 0 & 0 & 0 & c^2v_2\mu & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -c & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & c^2v_1\mu & 0 & 0 & z & 1 - cv_1v_2\mu & v_2 & 0 & 0 & 0 \\ 0 & c & 0 & 0 & -c & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & c^2\mu & 0 & 0 & 0 & -cv_2\mu & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrix $\mathcal{G}_2$ remains nonsingular ($\det(\mathcal{G}_2) = c^2$) and we have an index-2 DAE. The projector

$$W_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

shows that we have to differentiate the sixth and the ninth equations to reduce the index. This index-2 DAE corresponds to the middle interval of the solution (16).

4.3.3 Problem 3

The vector of dependent variables is given by $\underline{x} = (z^a, z^b, v^a, v^b, F, p_1, p_2)$.

$$G_0 = \begin{pmatrix} I_4 & \\ & 0_3 \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 0_4 & \\ & I_3 \end{pmatrix}.$$

For the matrix B we have

$$B = \left(\begin{array}{cccc|ccc} F & -10F & 0 & 0 & z^a - 10z^b & 0 & 0 \\ -F & 1+9F & 0 & 0 & -z^a + 9z^b & 0 & 0 \\ 0 & 0 & -F & F & -v^a + v^b & 0 & 0 \\ 0 & 0 & 10F & -1-9F & 1 + 10v^a - 9v^b & 0 & 0 \\ \hline v^a - v^b & -(1+10v^a - 9v^b) & z^a - 10z^b & -z^a + 9z^b & 0 & p^-_{1_{p_1}} & p^-_{2_{p_2}} \\ 0 & 0 & 0 & 0 & -1 & p^+_{1_{p_1}} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & p^+_{2_{p_2}} \end{array}\right),$$

and the first chain element is given by

$$G_1 = \left(\begin{array}{cccc|ccc} 1 & 0 & 0 & 0 & z^a - 10z^b & 0 & 0 \\ 0 & 1 & 0 & 0 & -z^a + 9z^b & 0 & 0 \\ 0 & 0 & 1 & 0 & -v^a + v^b & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 + 10v^a - 9v^b & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & p^-_{1_{p_1}} & p^-_{2_{p_2}} \\ 0 & 0 & 0 & 0 & -1 & p^+_{1_{p_1}} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & p^+_{2_{p_2}} \end{array}\right) =: \begin{pmatrix} G^1_{11} & G^1_{12} \\ G^1_{21} & G^1_{22} \end{pmatrix}.$$

The submatrix G^1_{22} of G_1, which determines the singularity is exactly the same as in Problem 1. We have a nonsingular matrix G_1 if the signs of p_1 and p_2 are different. If p_1 and p_2 are negative we have a nonregular DAE (see Problem 1). Lastly we have to investigate the case where both p_1 and p_2 are positive.

If $p_1 > 0$ and $p_2 > 0$ then $Z = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and a nullspace projector $\tilde{Q} = \begin{pmatrix} 0 & -1 \\ 0 & 1 \end{pmatrix}$.

The gradient of the constraint vector g is $g_u = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and from Lemma 4.3 the projector $\bar{Q}_1 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{pmatrix}$. From (31)

$$\mathcal{G}_2 = G_1 + BP_0Q_1 = \begin{pmatrix} I & B_{12} \\ 0 & B_{22} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} I & \\ & 0 \end{pmatrix} \begin{pmatrix} B_{12}R & 0 \\ -R & 0 \end{pmatrix} = \begin{pmatrix} I + B_{11}B_{12}R & B_{12} \\ B_{21}B_{12}R & B_{22} \end{pmatrix}.$$

By examination of B, it may be seen that $B_{21}B_{12} \equiv 0$ and $B_{22} = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ therefore $\mathcal{G}_2$ has a zero row and is singular. The singularity of $\mathcal{G}_2$ leads to a singular matrix G_2, which means that the DAE has index at least 3 (if it exists). To investigate that, theoretically we have to construct a projector Q_2 (if at all possible) with $Q_2Q_0 = 0$ and $Q_2Q_1 = 0$. It is a very complex problem and we will investigate that case numerically (see Section 5). We will see that the DAE has index 3.

We can apply Theorem 4.2. The assumptions are fulfilled, but to check it we used a formula manipulation system.

The image of $\mathcal{G}_2$ is given by $\{y : y = \mathcal{G}_2 z, z \in \mathbb{R}^n\}$. Setting $z = \begin{pmatrix} (I - B_{12}R)u \\ Ru + (I - \bar{Q}_1)\begin{pmatrix} 0 \\ v \end{pmatrix} \end{pmatrix}$ with arbitrary vectors $u \in \mathbb{R}^4$ and $v \in \mathbb{R}^2$ we get $\mathcal{G}_2 z = \begin{pmatrix} u \\ 0 \\ v \end{pmatrix}$. This shows that the projector along $\operatorname{im} G_2 = \operatorname{im} \mathcal{G}_2$ is given by

$$W_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We have to differentiate the fifth equation. The resulting equation (26) which may be written

$$(A + W_2 B D^-)(D\underline{x})' + (I - W_2)b(\underline{x}) = \underline{0} \tag{32}$$

has index 2 with

$$A + W_2 B D^- = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ v^a - v^b & -(1 + 10v^a - 9v^b) & z^a - 10z^b & -z^a + 9z^b & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Multiplying the system of DAEs on the left by a scaling matrix

$$T(\underline{x}) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ -(v^a - v^b) & 1 + 10v^a - 9v^b & -(z^a - 10z^b) & z^a - 9z^b & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

with $T(\underline{x}) = T(P\underline{x})$, will not change the index of (32) and it achieves the substitution of the derivatives. The new matrix chain elements corresponding to the equation

$$T(\underline{x})(A + W_2 B D^-)(D\underline{x})' + T(\underline{x})(I - W_2)\underline{b}(\underline{x}) = \underline{0} \tag{33}$$

are given by

$$\hat{G}_0 = T(AD + W_2 B D^- D) = G_0, \quad \hat{B} = (T(\underline{x})(I - W_2)\underline{b}(\underline{x}))_x,$$

and

$$\hat{G}_1 = G_0 + (T(\underline{x})(I - W_2)\underline{b}(\underline{x}))_x Q_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & z^a - 10z^b & 0 & 0 \\ 0 & 1 & 0 & 0 & -z^a + 9z^b & 0 & 0 \\ 0 & 0 & 1 & 0 & -v^a + v^b & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 + 10v^a - 9v^b & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} = G_1$$

because $\hat{Q}_0 = Q_0$.
Since equation (32) has index 2, the matrix $\hat{G}_2$ remains nonsingular and the projector $\hat{W}_1 (= W_2)$ along im $\hat{G}_1$ shows that for an additional index reduction we have to differentiate the fifth equation once more, as was found in the ad-hoc study of the index of this example in [5], [6]. The index-3 DAE for the case where $p_1 > 0$ and $p_2 > 0$ corresponds to the middle interval of the solution (21).

5 The numerical tests

The matrix chain (24) is realized numerically (see [15]) as a Matlab code. We will use it to test the problems numerically. Numerically means that we check the properties of the investigated DAEs pointwise for particular values of the variables. The tractability index works with linearizations of the DAE along a function, with appropriate smoothness properties. Here in our experiments we linearize the DAE along a linear function through $\underline{x}(\bar{t})$ with derivative of every component equal to 1.
Under "remarks" we put the answer of the algorithm and the last calculated matrix chain element, or we indicate to which part of the solution it corresponds.

Problem 4.3.1:
We use $\underline{x}(\bar{t}) = (y_1, y_2, c, v_1, v_2, r, u, p_1, p_2)^T = (0, 0, 1, 1, 1, 0, 1, \pm 1, \pm 1)^T$.

p_1	p_2	index	remarks
> 0	> 0	-	singular pencil - G_4
> 0	< 0	1	refers to (13)
< 0	> 0	1	refers to (14)
< 0	< 0	-	singular pencil - G_1

Problem 4.3.2:
We use $\underline{x}(\bar{t}) = (y_1, y_2, c, v_1, v_2, r, u, p_1, p_2, p_3)^T = (0, 0, 1, 1, 1, 0, 1, \pm 1, \pm 1, \pm 1)^T$.

p_1	p_2	p_3	index	remarks
>0	>0	>0	-	singular pencil - G_4
>0	<0	>0	1	refers to (15)
<0	>0	>0	1	refers to (17)
<0	<0	>0	-	singular pencil - G_1
>0	>0	<0	2	refers to (16)
>0	<0	<0	-	singular pencil - G_3
<0	>0	<0	-	singular pencil - G_3
<0	<0	<0	-	singular pencil - G_1

Problem 4.3.3:
We use $\underline{x}(\bar{t}) = (z_a, z_b, v_a, v_b, F, p_1, p_2)^T = (0, 0, 1, 1, 1, \pm 1, \pm 1)^T$.

p_1	p_2	index	remarks
>0	>0	3	refers to (21)
>0	<0	1	refers to (20)
<0	>0	1	refers to (22)
<0	<0	-	singular pencil - G_0

6 Conclusions

We have outlined here a procedure for transforming a general optimal control problem to a system of DAEs. The Kuhn-Tucker conditions consist of differential equations, complementarity conditions and corresponding inequalities. These latter are converted to equalities by the addition of a new variable combining the slack variable and the corresponding Lagrange multipliers. The sign of this variable indicates whether the constraint is active or not.

We have introduced the *tractability index* concept as a general purpose tool for determining the index of a general system of DAEs by checking for the nonsingularity of the elements of the matrix chain. This is helpful in determining the well-conditioning of the problem, and an appropriate method for solving it numerically.

In the examples used here, the solution of all the differential equations could be performed analytically. The given examples are tested by the numerical determination of the tractability index chain, and the results confirm the previously known properties of the examples.

References

[1] L.T. Biegler, "Solution of dynamic optimization problems by successive quadratic programming and orthogonal collocation", Comput. Chem. Eng., Vol. 8, pp 243-248, 1984.

[2] T. Binder, A. Cruse, C.A. Cruz Villar and W. Marquardt, "Dynamic optimization using a wavelet based adaptive control vector parameterization strategy", Comput. Chem. Eng., Vol. 24, pp 1201-1207, 2000.

[3] M.D. Canon, C.D. Cullum and E. Polak, "Theory of optimal control and mathematical programming", Mc.Graw-Hill, NY,1970.

[4] J.E. Cuthrell and L.T. Biegler, "On the optimization of differential-algebraic process systems", AIChE J. Vol.33, No. 8, pp 1257-1270, 1987.

[5] R. England, S. Gómez, R. Lamour, "A study of the index of differential-algebraic equations for optimal control problems", Reportes de Investigación, IIMAS-UNAM, Vol. 12, no. 62, 2004.

[6] R. England, S. Gómez, R. Lamour, "Expressing optimal control problems as differential algebraic equations", Comput. Chem. Eng., Vol. 29, no. 8, pp 1720-1730, 2005.

[7] D. Estévez Schwarz and R. Lamour, "The computation of consistent initial values for nonlinear index-2 differential-algebraic equations", Numerical Algorithms, Vol. 26, No. 1, pp 49-75, 2001.

[8] W.F. Feehery and P.I. Barton, "Dynamic optimization with state variable path constraints", Comput. Chem. Eng., Vol. 22, no. 9, pp 1241-1256, 1998.

[9] W.F. Feehery and P.I. Barton, "Dynamic optimization with equality path constraints", Ind. Eng. Chem. Res., Vol. 38, no. 6, pp 2350-2363, 1999.

[10] C.W. Gear, B. Leimkuhler and G.K. Gupta, "Automatic integration of Euler-Lagrange equations with constraints", J. Comp. Appl. Math. Vol.12/13, pp 77-90, 1985.

[11] I. Higueras, R. März, C. Tischendorf, "Stability preserving integration of index-1 DAE's", Appl. Numer. Math., Vol. 45, pp 175-200, 2003.

[12] I. Higueras, R. März, C. Tischendorf, "Stability preserving integration of index-2 DAEs", Appl. Numer. Math., Vol. 45, pp 201-229, 2003.

[13] R. Jackson, "Optimal use of mixed catalyst for two successive chemical reactions", J. Optim. Theory Appl., Vol. 2, no. 1, 1968.

[14] R. Lamour, "A shooting method for fully implicit index-2 differential algebraic equations", SIAM J. Sci. Comput., Vol. 18, no. 1, pp 94-114, 1997.

[15] R. Lamour, "Index determination and calculation of consistent initial values for DAEs", Comp. Math. with Appl., Vol. 50, pp 1125-1140, 2005.

[16] J.S. Logsdon and L.T. Biegler , "Accurate solution of differential-algebraic optimization problems", Int. Eng. Chem. Res. Vol.28, No. 11, pp 1628-1639, 1989.

[17] R. März, "Numerical methods for differential-algebraic equations", Acta Numerica 1992, pp. 141-198.

[18] R. März, "The index of differential algebraic systems with properly stated leading term", Results in Math., Vol. 42, no 3/4, pp 308-338, 2002.

[19] C.C. Pantelides, R.W.H. Sargent and V.S. Vassiliadis, "Optimal control of multistage systems described by high-index differential-algebraic equations", Computational Optimal Control, Eds. R. Bulirsch and D. Kraft, International Series of Numerical Mathematics, Birkhäuser-Verlag, Basel, Vol. 115, pp 177-191, 1994.

[20] J.G. Rentro, "Computational studies in the optimization of systems described by differential/algebraic equations", Ph.D. Thesis, University of Houston, 1986.

[21] J.G. Rentro, A.M. Moshedi and O.A. Asbjornsen, "Simultaneous optimization and solution of systems described by differential/algebraic equations", Comput. Chem. Eng., Vol. 11, pp 503-517, 1987.

[22] T.H. Tsang, D.M. Himmelblau and T.F. Edgar, "Optimal control via collocation and non-linear programming", Int. J. Control, Vol. 21, pp 763-768, 1975.

[23] S. Vasantharajan and L.T. Biegler, "Simultaneous strategies for the optimization of differential-algebraic systems with enforcement of error criteria", Comput. Chem. Eng., Vol. 14, pp 1083-1100, 1990.

[24] V.S. Vassiliadis, R.W.H. Sargent and C.C. Pantelides, "Solution of a class of multistage dynamic optimization problems. I. Problems without path constraints", Ind. Eng. Chem. Res., Vol. 33, no. 9, pp 2111-2122, 1994.

[25] V.S. Vassiliadis, R.W.H. Sargent and C.C. Pantelides, "Solution of a class of multistage dynamic optimization problems. II. Problems with path constraints", Ind. Eng. Chem. Res., Vol. 33, no. 9, pp 2123-2133, 1994.

[26] J. Vlassenbroeck and R.A. van Dooren, "Chebyshev technique for solving non-linear optimal control problems", IEEE Trans. autom. control, Vol. 33, pp 333-340, 1988.