

# Accessible criteria for the local existence and uniqueness of DAE solutions

Steffen Voigtmann\*

Institut für Mathematik, Technische Universität Berlin  
Straße des 17. Juni 136, 10623 Berlin, Germany  
*voigtmann@math.tu-berlin.de*

**Abstract.** Classical results about the local existence and uniqueness of DAE solutions are based on the derivative array [2] or on a geometrical approach [13]. Thus these results can't be applied to equations with non-smooth coefficients. Also, sufficient conditions that guarantee solvability are hard to check in general [6, 13]. In this paper a new approach to proving local existence and uniqueness of DAE solutions is presented. The main tool is a decoupling procedure that makes it possible to split DAE solutions into their characteristic parts. Thus it is possible to weaken the smoothness requirements considerably. In order for the decoupling procedure to work we require a certain structural condition to hold. In contrast to results already known, this condition can be easily verified.

## 1 Introduction

When developing realistic models for a large variety of industrial applications one is often confronted to deal with systems of differential algebraic equations (DAEs) that have to be solved numerically. Thus questions of existence and uniqueness of solutions are of key importance for the successful application of numerical methods.

Most of the work on numerical analysis of DAEs has focused on the computation of a solution that is assumed to exist [2]. However, the solvability of DAEs is, to some extent, still an open question. While the case of linear time-varying DAEs is well understood [1, 9, 10], there are only partial results for nonlinear equations. We will focus on quasilinear DAEs

$$A(t)(d(x(t), t))' + b(x(t), t) = 0 \tag{1}$$

with index 1 or 2 which are most relevant for applications<sup>1</sup>. In [13, Theorem 24.1] results concerning existence and uniqueness were obtained using a geometric approach. There it is assumed that the coefficients are in the class  $C^\infty$ . The sufficient conditions given in [1, 2] and [6] use derivative arrays

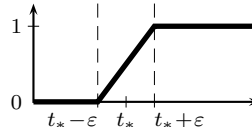
---

\*Supported by the DFG research center MATHEON – *Mathematics for key technologies*

<sup>1</sup>In particular we have the charge oriented modified nodal analysis (MNA) in mind.

and index reduction techniques that lead to higher derivatives and smooth coefficients have to be assumed. When simulating electrical circuits, but also for many other applications, these smoothness requirements are generally too strong. For instance mappings such as

$$g_{\varepsilon, t_*}(t) = \begin{cases} 0 & , & 0 \leq t < t_* - \varepsilon \\ \frac{t-t_*}{2\varepsilon} + \frac{1}{2} & , & t_* - \varepsilon \leq t < t_* + \varepsilon \\ 1 & , & t_* + \varepsilon \leq t \end{cases}$$



with small  $\varepsilon > 0$  are often used to model an independent source that is switched on at  $t = t_*$ . As  $g_{\varepsilon, t_*}$  is continuous but not continuously differentiable, the example

$$\begin{aligned} x_1' + x_2^2 - x_3(1 + x_3) &= g_{\varepsilon, t_*}(t) \\ x_2' - x_3 &= g_{\varepsilon, t_*}(t) \\ x_1 - x_2 &= 0 \end{aligned} \tag{2}$$

and similar equations can't be treated using the concepts mentioned above. Also, a treatment on the subintervals  $[0, t_*)$  and  $[t_*, \infty)$  is not feasible in general, as the switching point  $t_*$  may not be known in advance. Nevertheless, there is a unique solution of the initial value problem (2),  $x(0) = x_0$ . The solution will be constructed in section 3.

The object of this paper is to widen the class of differential algebraic equations for which existence and uniqueness of solutions can be proved. We focus on low smoothness requirements in order to be able to include examples such as (2). Thus we use the concept of the tractability index for our investigations [9, 11]. In section 2 we impose a structural condition (\*) on the DAE (1) that makes it possible to generalise the decoupling procedure for linear DAEs [16] to quasilinear ones. In section 3 this decoupling procedure is used to give sufficient conditions that guarantee the existence of unique local solutions.

Recall that the geometric index is sometimes hard to check in practice [13]. Also the hypothesis from [6] is rather complicated to verify. Our approach has the advantage that the sufficient conditions given here are easily accessible. In particular the structural condition (\*) is valid for all DAEs modelling electrical circuits via the modified nodal analysis. Thus, apart from the smoothness requirements, no additional effort is necessary to guarantee the local existence and uniqueness of solutions.

## 2 Preliminaries

We consider quasilinear differential algebraic equations (DAEs)

$$A(t)(d(x(t), t))' + b(x(t), t) = 0 \tag{1}$$

with continuous coefficients. Let  $\mathcal{D} \subset \mathbb{R}^m$  be a domain and  $\mathcal{I} \subset \mathbb{R}$  an interval,  $A(t) \in L(\mathbb{R}^n, \mathbb{R}^m)$ ,  $d(x, t) \in \mathbb{R}^n$ ,  $b(x, t) \in \mathbb{R}^m$  for  $(x, t) \in \mathcal{D} \times \mathcal{I}$ . We assume

that continuous partial derivatives  $d'_x$  and  $b'_x$  exist. Let the leading term of (1) be properly stated [8], i.e.

$$\ker A(t) \oplus \operatorname{im} d'_x(x, t) = \mathbb{R}^n \quad \forall (x, t) \in \mathcal{D} \times \mathcal{I},$$

and there is a smooth projector function  $R \in C^1(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^n))$  such that  $\ker R(t) = \ker A(t)$ ,  $\operatorname{im} R(t) = \operatorname{im} d'_x(x, t)$  and  $d(x, t) = R(t)d(x, t)$  for every  $(x, t) \in \mathcal{D} \times \mathcal{I}$ . In particular,  $\operatorname{im} d'_x$  does not depend on  $x$ .

A function  $x \in C(\mathcal{I}_x, \mathbb{R}^m)$ ,  $\mathcal{I}_x \subset \mathcal{I}$ , is said to be a solution of (1) if  $x(t) \in \mathcal{D}$ ,  $t \in \mathcal{I}_x$ ,  $d(x(\cdot), \cdot) \in C^1(\mathcal{I}_x, \mathbb{R}^m)$  and  $x$  satisfies the DAE pointwise for  $t \in \mathcal{I}_x$ . Unfortunately this notion does not lead to a linear function space. Thus we consider DAEs of the form

$$A(t)(D(t)x(t))' + b(x(t), t) = 0 \quad (3)$$

where solutions lie in the linear space

$$C_D^1(\mathcal{I}, \mathbb{R}^m) := \{ z \in C(\mathcal{I}, \mathbb{R}^m) \mid Dz \in C^1(\mathcal{I}, \mathbb{R}^n) \}.$$

Note that (1) can be transformed into (3) by considering the enlarged system

$$A(t)(R(t)y(t))' + b(x(t), t) = 0, \quad (4a)$$

$$y(t) - d(x(t), t) = 0. \quad (4b)$$

With  $\hat{x} = \begin{pmatrix} x \\ y \end{pmatrix}$ ,  $\hat{A} = \begin{pmatrix} A \\ 0 \end{pmatrix}$ ,  $\hat{D} = \begin{pmatrix} 0 & R \end{pmatrix}$  and  $\hat{b}(\hat{x}, t) = \begin{pmatrix} b(x, t) \\ y - d(x, t) \end{pmatrix}$  (4) is seen to be of type (3). Indeed, in [8] it is shown that (1) and (4) are equivalent and we restrict ourselves to DAEs of type (3).

When analysing DAEs with properly stated leading terms it is advantageous to introduce a certain sequence of matrix functions and subspaces. This sequence not only provides means for defining the tractability index but also allows a refined analysis of (3). Here we summarise the results necessary for our later investigations. Details can be found in [9, 11, 14].

Pointwise for  $t \in \mathcal{I}$ ,  $x \in \mathcal{D}$  we introduce

$$\begin{aligned} G_0(t) &= A(t)D(t), & B(x, t) &= b'_x(x, t) \\ N_0(t) &= \ker G_0(t), & S_0(x, t) &= \{ z \in \mathbb{R}^m \mid B(x, t)z \in \operatorname{im} G_0(t) \} \end{aligned} \quad (5a)$$

and a projector function  $Q_0 \in C(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^m))$  onto  $N_0$ . We define

$$\begin{aligned} G_1(x, t) &= G_0(t) + B(x, t)Q_0(t), \\ N_1(x, t) &= \ker G_1(x, t), \quad S_1(x, t) = \{ z \in \mathbb{R}^m \mid B(x, t)z \in \operatorname{im} G_1(x, t) \}. \end{aligned} \quad (5b)$$

Let  $Q_1 \in C(\mathcal{D} \times \mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^m))$  be a projector function onto  $N_1$  and define  $P_0(t) = I - Q_0(t)$ ,  $P_1(x, t) = I - Q_1(x, t)$ . Finally consider

$$\begin{aligned} C(x^1, x, t) &= (DP_1D^-)'_x(x, t)x^1 + (DP_1D^-)'_t(x, t), \\ B_1(x^1, x, t) &= B(x, t)P_0(t) - G_1(x, t)D^-(t)C(x^1, x, t)D(t), \\ G_2(x^1, x, t) &= G_1(x, t) + B_1(x^1, x, t)Q_1(x, t) \end{aligned} \quad (5c)$$

for  $t \in \mathcal{I}$ ,  $x \in \mathcal{D}$  and  $x^1 \in \mathbb{R}^m$ .  $D^-(t)$  is the generalised reflexive inverse of  $D(t)$  defined by

$$DD^-D = D, \quad D^-DD^- = D^-, \quad D^-D = P_0, \quad DD^- = R.$$

When defining  $C$  we have to make sure that the partial derivatives of  $DP_1D^-$  exist. Note that  $C(\bar{x}'(t), \bar{x}(t), t) = \frac{d}{dt}[(DP_1D^-)(\bar{x}(t), t)]$  follows for arbitrary functions  $\bar{x} \in C^1(\mathcal{I}, \mathbb{R}^m)$ .

**Definition 2.1** ([11]) *Let (3) be a DAE with a properly stated leading term.*

- (i) *The DAE is called regular with (tractability) index 1 on  $\mathcal{D} \times \mathcal{I}$ , if there is a sequence (5) such that  $G_0$  and  $G_1$  have constant rank  $r_0$  and  $r_1$ , respectively, for  $(x, t) \in \mathcal{D} \times \mathcal{I}$  and  $r_0 < r_1 = m$ .*
- (ii) *The DAE is called regular with (tractability) index 2 on  $\mathcal{D} \times \mathcal{I}$ , if there is a sequence (5) such that*
  - (a)  *$Q_0$  and  $Q_1$  are continuous but  $DP_1D^-$  is continuously differentiable,*
  - (b)  *$N_0 \subset \ker Q_1$  pointwise for  $(x, t) \in \mathcal{D} \times \mathcal{I}$ ,*
  - (c)  *$G_i$  has constant rank  $r_i$  for  $0 \leq i \leq 2$ ,  $t \in \mathcal{I}$ ,  $x \in \mathcal{D}$ ,  $x^1 \in \mathbb{R}^m$  and  $r_0 \leq r_1 < r_2 = m$ .*

This definition is a generalisation of the index notion given for linear DAEs in [10]. In this case we have  $C(t) = (DP_1D^-)'(t)$ .

For index-2 DAEs definition 2.1 implies that  $G_2(x^1, x, t)$  remains nonsingular on  $\mathbb{R}^m \times \mathcal{D} \times \mathcal{I}$  and we have

$$N_1(x, t) \oplus S_1(x, t) = \mathbb{R}^m.$$

In the following we will always choose  $Q_1$  to be the canonical projector onto  $N_1$  along  $S_1$ . Due to  $N_0 \subset S_1$  the property (ii) is always valid for the canonical projector  $Q_1$ .

The space  $N_0(t) \cap S_0(x, t)$  is of vital importance for index-2 DAEs. Following [16] we introduce a projector function  $T \in C(\mathcal{D} \times \mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^m))$  such that

$$\text{im } T(x, t) = N_0(t) \cap S_0(x, t)$$

and define  $U(x, t) = I - T(x, t)$ . Due to  $\text{im } P_0(t) \cap (N_0(t) \cap S_0(x, t))$  we can always assume that

$$Q_0T = T = TQ_0, \quad P_0U = P_0 = UP_0.$$

It is well known that in order to prove existence of solutions, the DAE has to satisfy certain structural conditions. In [6] the DAE is assumed to satisfy a hypothesis based on the derivative array. In contrast to that, [12] requires  $Q_1G_2^{-1}(b(x, t) - b(P_0x, t)) = 0$ . Unfortunately the latter requirement is too

restrictive in the sense that there are DAEs arising from the modified nodal analysis in circuit simulation that do not satisfy this condition [16]. Hence in [3, 16] the generalised structural condition

$$N_0(t) \cap S_0(x, t) \quad \text{does not depend on } x \quad (*)$$

was introduced. The space  $N_0 \cap S_0$  describes the so-called index-2 components, i.e. the particular part of  $x$  that can be calculated only by performing an inherent differentiation process. Since the circuit's layout determines the subspace  $N_0(t) \cap S_0(x, t)$ , it is indeed independent of  $x$  and  $(*)$  holds for all DAEs obtained from the modified nodal analysis (MNA) [4].

We will show that  $(*)$  is already sufficient for the local existence and uniqueness of solutions. Thus in circuit simulation there is no need to check complicated conditions that guarantee the existence of solutions. For DAEs from other application backgrounds,  $(*)$  can be checked using linear algebra tools in practice [7].

As we will always assume  $(*)$  to hold, we choose  $T$  to be independent of  $x$ .

Finally let us remark that condition  $(*)$  is the same for (1) and the enlarged system (4) since the corresponding subspaces are related by

$$\hat{N}_0(t) \cap \hat{S}_0(\hat{x}, t) = (N_0(t) \cap S_0(x, t)) \times \{0\}.$$

More details are given in [8].

### 3 The decoupling procedure for index-2 DAEs

We assume that (3) is a regular DAE with a properly stated leading term that has index 2 on  $\mathcal{D} \times \mathcal{I}$  and that  $(y^0, x^0, t_0)$  is a point in  $\text{im } D(t_0) \times \mathcal{D} \times \mathcal{I}$  such that

$$(A1) \quad A(t_0)y^0 + b(x^0, t_0) = 0.$$

The initialisation  $(y^0, x^0, t_0)$  doesn't need to be consistent, i.e. apriori we do not require that there is a solution passing through  $x^0$ . However, we will use  $(y^0, x^0, t_0)$  for the construction of a consistent initialisation  $(y_0, x_0, t_0)$ . This process can be compared with the step-by-step construction of consistent initial values in [3].

Let  $\bar{x} \in C^1(\mathcal{I}, \mathbb{R}^m)$  be an arbitrary function satisfying

$$(A2) \quad \bar{x}(t_0) = x^0, \quad (\bar{x}(t), t) \in \mathcal{D} \times \mathcal{I} \quad \forall t \in \mathcal{I}.$$

Some of the matrix functions defined above depend not only on  $t$  but also on the arguments  $x$  and  $x^1$ . We will evaluate these functions in  $\bar{x}$ , i.e. we consider the functions

$$\begin{aligned} \bar{Q}_1(t) &= Q_1(\bar{x}(t), t), & \bar{P}_1(t) &= P_1(\bar{x}(t), t), \\ \bar{G}_1(t) &= G_1(\bar{x}(t), t), & \bar{B}(t) &= B(\bar{x}(t), t), \\ & & \bar{G}_2(t) &= G_2(\bar{x}'(t), \bar{x}(t), t) \end{aligned} \quad (6)$$

defined for  $t \in \mathcal{I}$ . Remember that due to the index-2 condition,  $\bar{G}_2(t)$  remains nonsingular on  $\mathcal{I}$ . Also, recall from [5] the representation  $\bar{Q}_1 = \bar{Q}_1 \bar{G}_2^{-1} \bar{B} P_0$ , as  $Q_1$  was chosen to be the canonical projector onto  $N_1$  along  $S_1$ .

### 3.1 Splitting of DAE solutions

For the moment let us assume that there is a solution  $x_*(\cdot) \in C_D^1(\mathcal{I}, \mathbb{R}^m)$  of (3) with

$$(y^0, x^0, t_0) = \left( (Dx_*)'(t_0), x_*(t_0), t_0 \right). \quad (7)$$

The results obtained when assuming the existence of a solution will lead the way to constructing one in the more general setting. Observe that in this section (and only in this section)  $x^0$  is a consistent initial value due to (7).

We define functions

$$u(t) = D(t)\bar{P}_1(t)x_*(t), \quad w(t) = T(t)x_*(t), \quad z(t) = \bar{Z}(t)x_*(t) \quad (8)$$

for  $t \in \mathcal{I}$ , where  $\bar{Z}(t) = P_0(t)\bar{Q}_1(t) + U(t)Q_0(t)$  is again a projector function. Figure 1 shows how these functions are obtained from  $x_*$  by successive splitting and hence the solution  $x_*(\cdot)$  itself can be written as

$$x_*(t) = D^-(t)u(t) + z(t) + w(t). \quad (9)$$

We think of  $u(\cdot)$ ,  $z(\cdot)$  and  $w(\cdot)$  as the dynamical, algebraic and differential part, respectively. The motivation for defining these functions comes from the study of linear index-2 equations where  $u(\cdot)$  is determined by the inherent regular ODE,  $z(\cdot)$  is given by a purely algebraic equation but in order to determine  $w(\cdot)$  one has to carry out a differentiation of certain parts of the right-hand side [16]. Since the solution  $x_*$  belongs to  $C_D^1(\mathcal{I}, \mathbb{R}^m)$  we have  $z \in C_D^1(\mathcal{I}, \mathbb{R}^m)$ . Additionally we consider  $v(t) = D(t)\bar{Q}_1(t)x_*(t) = D(t)z(t)$ , such that

$$(Dx_*)' = (D\bar{P}_1x_* + D\bar{Q}_1x_*)' = u' + v'.$$

We will now rewrite the DAE (3) in terms of the new variables introduced above. Using the notation

$$f(y, x, t) = A(t)y + b(x, t)$$

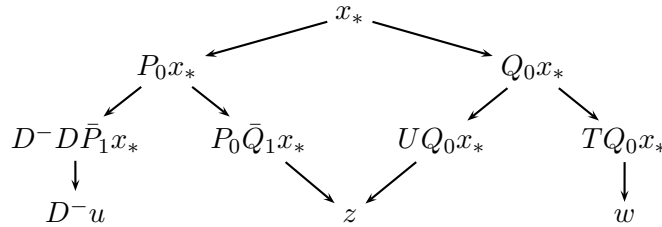


Figure 1: The relation of  $u$ ,  $z$  and  $w$  to  $x_*$ .

(3) can be written as

$$\begin{aligned}
0 &= f\left((Dx_*)'(t), x_*(t), t\right) \\
&= f\left(u'(t) + v'(t), (D^-u)(t) + z(t) + w(t), t\right) \\
&= F\left(u(t), w(t), z(t), u'(t), v'(t), t\right), \quad t \in \mathcal{I}.
\end{aligned} \tag{10a}$$

The function  $F$  is defined by

$$F(u, w, z, \eta, \zeta, t) = f\left(\eta + \zeta, D^-(t)u + \bar{Z}(t)z + T(t)w, t\right) \tag{10b}$$

where  $u$ ,  $w$ ,  $z$ ,  $\eta$  and  $\zeta$  are considered to be parameters.  $\bar{Z}$  and  $T$  were introduced for convenience when defining  $F$ . Because of (8) they do not change (10a) at all but will be quite useful when calculating derivatives of  $F$ .

**Lemma 3.1** *Let (3) be a regular DAE with index 2 on  $\mathcal{D} \times \mathcal{I}$ . Assume that (7) holds for a solution  $x_* \in C^1(\mathcal{I}, \mathbb{R}^m)$ . Choose  $\bar{x} = x_*$  and define*

$$u_0 = u(t_0), \quad w_0 = w(t_0), \quad z_0 = z(t_0), \quad \eta_0 = u'(t_0), \quad \zeta_0 = v'(t_0)$$

for the functions  $u$ ,  $w$  and  $z$  from (8). If the structural condition (\*) holds, then locally around  $(u_0, w_0, z_0, \eta_0, \zeta_0, t_0)$  equation (10) is equivalent to

$$z(t) = \mathbf{z}(u(t), t), \quad u'(t) = \mathbb{f}(u(t), w(t), t), \quad w(t) = \mathbf{w}(u(t), v'(t), t) \tag{11}$$

with continuous functions  $\mathbf{z}$ ,  $\mathbb{f}$  and  $\mathbf{w}$  being defined on neighbourhoods of  $(u_0, t_0)$ ,  $(u_0, w_0, t_0)$  and  $(u_0, \zeta_0, t_0)$ , respectively.

To prove this result one splits the function  $F$  from (10b) using an approach similar to the one depicted in figure 1. Two applications of the implicit function theorem yield the mappings  $\mathbf{z}$  and  $\mathbf{w}$ . The function

$$\begin{aligned}
\mathbb{f}(u, w, t) &= (D\bar{P}_1 D^-)'(t) \left( u + \mathbf{v}(u, t) \right) \\
&\quad - (D\bar{P}_1 \bar{G}_2^{-1})(t) b \left( D^-(t)u + \mathbf{z}(u, t) + T(t)w, t \right)
\end{aligned} \tag{12}$$

can be written in terms of the original data. The mapping  $\mathbf{v}(u, t) = D(t)\mathbf{z}(u, t)$  is given once  $\mathbf{z}$  is known. The detailed proof of lemma 3.1 will be carried out in section 4.

### 3.2 Local existence and uniqueness of DAE solutions

From now on we drop the assumption that there is a solution of (3). However, we assume that  $(y^0, x^0, t_0) \in \text{im } D(t_0) \times \mathcal{D} \times \mathcal{I}$  and  $\bar{x} \in C^1(\mathcal{I}, \mathbb{R}^m)$  are given such that (A1) and (A2) are valid. Using the matrix functions defined in (6) we introduce

$$u_0 = D(t_0)\bar{P}_1(t_0)x^0, \quad w_0 = T(t_0)x^0, \quad z_0 = \bar{Z}(t_0)x^0. \tag{13}$$

Notice that  $x^0 = D^-(t_0)u_0 + z_0 + w_0$ . We will now study the equation

$$F(u, w, z, \eta, \zeta, t) = 0 \quad (14)$$

without assuming that there is a solution of the original DAE. Observe that the parameters  $\eta$  and  $\zeta$  replace the derivatives  $u'$  and  $v'$ , respectively.

The results obtained in this section are based on the proof of lemma 3.1. Hence the results will only be quoted here. Full proofs are given in section 4. The first statement provides a function  $\mathbf{z}$  similar to the one obtained in lemma 3.1.

**Lemma 3.2** *Let (3) be a regular DAE with index 2 on  $\mathcal{D} \times \mathcal{I}$ . Assume that (A1), (A2) and the structural condition (\*) hold. Then the function*

$$\bar{Z}(t)\bar{G}_2^{-1}(t)F(u, w, z, \eta, \zeta, t) + (I - \bar{Z}(t))z =: \hat{F}_1(u, z, t)$$

*is independent of  $w$ ,  $\eta$ ,  $\zeta$  and there is  $r_z > 0$  and a continuous function*

$$\mathbf{z} : B_{r_z}(u_0, t_0) \rightarrow \mathbb{R}^m, \quad \mathbf{z}(u_0, t_0) = z_0,$$

*such that  $\hat{F}_1(u, \mathbf{z}(u, t), t) = 0$  for every  $(u, t) \in B_{r_z}(u_0, t_0)$ .*

Using the function  $\mathbf{z}$  from this lemma we introduce  $\mathbf{v}(u, t) = D(t)\mathbf{z}(u, t)$  and define  $\mathfrak{f}(u, w, t)$  as in (12). Recall that we had  $u'(t) = \mathfrak{f}(u(t), w(t), t)$  in lemma 3.1. The function  $\mathfrak{f}$  obtained here will have the same significance. Thus in (14) we replace  $\eta$  by  $\mathfrak{f}$  and  $z$  by  $\mathbf{z}$ , respectively. The resulting equation is studied in the next lemma.

**Lemma 3.3** *Let (3) be a regular DAE with index 2 on  $\mathcal{D} \times \mathcal{I}$ . Assume that (A1), (A2) and the structural condition (\*) hold. Consider*

$$\hat{F}_2(u, w, \zeta, t) = T(t)\bar{G}_2^{-1}(t)F(u, w, \mathbf{z}(u, t), \mathfrak{f}(u, w, t), \zeta, t) + (I - T(t))w$$

*where  $\mathbf{z}$  is the mapping from lemma 3.2. Let  $\zeta_0 = y^0 - \mathfrak{f}(u_0, w_0, t_0)$ . Then there is  $r_w > 0$  and a continuous function*

$$\mathfrak{w} : B_{r_w}(u_0, \zeta_0, t_0) \rightarrow \mathbb{R}^m, \quad \mathfrak{w}(u_0, \zeta_0, t_0) = w_0,$$

*such that  $\hat{F}_2(u, \mathfrak{w}(u, \zeta, t), \zeta, t) = 0$  for every  $(u, \zeta, t) \in B_{r_w}(u_0, \zeta_0, t_0)$ .*

The mappings  $\mathbf{z}$ ,  $\mathbf{v}$ ,  $\mathfrak{f}$  and  $\mathfrak{w}$  introduced above allow the construction of a solution. We need to consider the following system of differential algebraic equations

$$\begin{aligned} z &= \mathbf{z}(u, t), & v &= \mathbf{v}(u, t) = D(t)\mathbf{z}(u, t), \\ u' &= \mathfrak{f}(u, w, t), & w &= \mathfrak{w}(u, v', t). \end{aligned}$$

Inserting  $\mathfrak{w}$  into  $\mathfrak{f}$ , it turns out that we have to deal with the implicit DAE

$$u' = \mathfrak{f}(u, \mathfrak{w}(u, v', t), t) =: f(u, v', t) \quad (15a)$$

$$v = D(t)\mathbf{z}(u, t) =: g(u, t). \quad (15b)$$

first. Once  $u$  and  $v$  are known, we obtain the remaining components via  $z = \mathbf{z}(u, t)$ ,  $w = \mathfrak{w}(u, v', t)$ .



**Lemma 3.4** *Let (3) be a regular DAE with index 2 on  $\mathcal{D} \times \mathcal{I}$ . Assume that (A1), (A2) and the structural condition (\*) hold. Let  $\mathbf{z}$ ,  $\mathfrak{f}$  and  $\mathfrak{w}$  be the functions obtained from lemma 3.2 and 3.3.*

- (i) *The implicit DAE (15) has (differentiation) index 1.*
- (ii) *For every consistent initial condition  $(u(t_0), v(t_0)) = (u_0, g(u_0, t_0))$  there is a unique solution of (15).*
- (iii) *If  $u_0 \in \text{im } D(t_0)\bar{P}_1(t_0)$ , then  $u(t) \in \text{im } D(t)\bar{P}_1(t)$  for every  $t$  where the solution exists.*

**Proof:** To see (i) it suffices to note that  $I - f'_v g'_u$  is nonsingular in a neighbourhood of  $(u_0, \zeta_0, t_0)$  (see remark 4.2).

In order to prove (ii), differentiate (15b) and insert the result into (15a). This yields  $u' = f(u, g'_u(u, t)u' + g'_t(u, t), t) = \hat{f}(u, u', t)$  and due to the index-1 condition we can solve for  $u'$ . Thus (15b) is equivalent to an ordinary differential equation  $u' = \mathcal{F}(u, t)$ . Now solve the initial value problem  $u' = \mathcal{F}(u, t)$ ,  $u(t_0) = u_0$  to see that  $(u(t), g(u(t), t))$  is the unique solution.

(iii) can be proved similar to the case of linear DAEs [9]. Let  $(u, v)$  be a solution of (15) with  $u(t_0) \in \text{im } D(t_0)\bar{P}_1(t_0)$ . Multiplication of (15a) by  $I - D\bar{P}_1D^-$  yields

$$(I - D\bar{P}_1D^-)(t) u'(t) = -(I - D\bar{P}_1D^-)'(t) (D\bar{P}_1D^-)(t) u(t)$$

since  $D\bar{P}_1D^-v(u, \cdot) = 0$ . This means that  $\hat{u} = (I - D\bar{P}_1D^-)u$  satisfies the linear ODE

$$\hat{u}' = (I - D\bar{P}_1D^-)' \hat{u}. \quad (16)$$

Now  $u(t_0) \in \text{im } D(t_0)\bar{P}_1(t_0)$  implies  $\hat{u}(t_0) = 0$  and the solution of (16) is identically zero, i.e.  $u(t) \in \text{im } D(t)\bar{P}_1(t) \forall t$ .  $\square$

Starting from lemma 3.4 it is now straight forward to construct a solution of the original index-2 system (3). We collect the result in the following theorem.

**Theorem 3.5** *Let (3) be a regular index-2 DAE on  $\mathcal{D} \times \mathcal{I}$  with a properly stated leading term. Let the structural condition (\*) hold and additionally require*

(A1)  $\exists (y^0, x^0, t_0) \in \text{im } D(t_0) \times \mathcal{D} \times \mathcal{I}$  such that  $A(t_0)y^0 + b(x^0, t_0) = 0$ ,

(A2)  $\exists \bar{x} \in C^1(\mathcal{I}, \mathbb{R}^m)$  such that  $\bar{x}(t_0) = x^0$  and  $(\bar{x}(t), t) \in \mathcal{D} \times \mathcal{I} \forall t \in \mathcal{I}$ ,

(A3) the derivatives  $b_x$ ,  $D'$  and  $\frac{\partial}{\partial t}(\bar{Z}\bar{G}_2^{-1}b)$  exist and are continuous.

Then there is a unique local solution of the initial value problem

$$A(t)(D(t)x(t))' + b(x(t), t) = 0, \quad D(t_0)P_1(x^0, t_0)(x(t_0) - x^0) = 0. \quad (17)$$

**Proof:** Use lemma 3.2 to obtain the mappings  $\mathbf{z}$  and  $\mathbf{v}(u, t) = D(t)\mathbf{z}(u, t)$ . Let  $\mathbf{w}$  be the mapping defined in lemma 3.3. Then due to lemma 3.4 there is a local solution of the implicit index-1 system

$$\begin{aligned} u' &= \mathbb{f}(u, \mathbf{w}(u, v', t), t), & u(t_0) &= u_0 := D(t_0)P_1(x^0, t_0)x^0, \\ v &= D(t)\mathbf{z}(u, t), & v(t_0) &= g(u_0, t_0) \end{aligned}$$

existing for  $t \in \mathcal{I}_\varepsilon = (t_0 - \varepsilon, t_0 + \varepsilon) \cap \mathcal{I}$  for some  $\varepsilon > 0$ . We define

$$x_*(t) = D^-(t)u(t) + \mathbf{z}(u(t), t) + \mathbf{w}(u(t), v'(t), t). \quad (18)$$

It remains to check that (18) is indeed a solution.

Due to  $u_0 \in \text{im}(D\bar{P}_1)(t_0)$  and lemma 3.4 we have  $u(t) \in \text{im}(D\bar{P}_1)(t)$  for every  $t \in \mathcal{I}_\varepsilon$ . Recall that  $R(t) = (DD^-)(t)$  is the projector function related to the properly stated leading term. Therefore  $u(t) \in \text{im}R(t)$  and  $Dx_* = Ru + D\bar{Z}\mathbf{z}(u, \cdot) + DT\mathbf{w}(u, \cdot) = u + \mathbf{v}(u, \cdot)$  is a  $C^1$  mapping<sup>2</sup> due to (A3). In particular we have  $D\bar{P}_1x_* = u$ ,  $\bar{Z}x_* = \mathbf{z}(u, \cdot)$ ,  $Tx_* = \mathbf{w}(u, v', \cdot)$ . Thus we get

$$\begin{aligned} \bar{Z}\bar{G}_2^{-1} [A(Dx_*)' + b(x_*, \cdot)] &= \hat{F}_1(u, \mathbf{z}(u, \cdot), \cdot) &= 0, \\ T\bar{G}_2^{-1} [A(Dx_*)' + b(x_*, \cdot)] &= \hat{F}_2(u, \mathbf{w}(u, \zeta, \cdot), \zeta, \cdot) &= 0, \\ P_0\bar{P}_1\bar{G}_2^{-1} [A(Dx_*)' + b(x_*, \cdot)] &= D^-(u' - \mathbb{f}(u, \mathbf{w}(u, v', \cdot), \cdot)) &= 0 \end{aligned}$$

(see remark 4.3). Since  $I = \bar{Z} + T + P_0\bar{P}_1$  we conclude that  $x_*$  is indeed a solution of (3). Due to  $D(t_0)P_1(x^0, t_0)x_*(t_0) = u(t_0) = u_0 = D(t_0)P_1(x^0, t_0)x^0$  this solution satisfies the initial value problem (17).

If there was another solution, say  $\hat{x}_*$ , then we could decouple  $x_*$  and  $\hat{x}_*$  as described in section 3.1. Therefore the corresponding  $D\bar{P}_1$  parts  $u$  and  $\hat{u}$  solve the same inherent index-1 system (15) and are therefore equal. Because of lemma 3.1  $z$  and  $\hat{z}$  as well as  $w$  and  $\hat{w}$  are also equal, respectively, and  $x_*$  and  $\hat{x}_*$  coincide.  $\square$

The smoothness that is required in order to be able to construct the solution, is given when the function  $\mathbf{v}$  is differentiable with respect to  $t$ . The conditions on  $D$  and  $\bar{Z}\bar{G}_2^{-1}b$  in theorem 3.5 guarantee this fact but they are unnecessary strong in general.

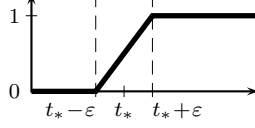
The theorem is given for index-2 DAEs. Using  $Q_1 = 0$  and  $P_1 = I$  it turns out that the result contains index-1 equations as well. In this case  $\mathbf{v}$  is zero and no additional smoothness is required.

**Example 3.6** We want to employ the decoupling procedure described above for explicitly constructing a solution of the DAE

$$\begin{aligned} x_1' + x_2^2 - x_3(1+x_3) &= g_{\varepsilon, t_*}(t) \\ x_2' - x_3 &= g_{\varepsilon, t_*}(t) \\ x_1 - x_2 &= 0 \end{aligned} \Leftrightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \left( \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right)' + \begin{bmatrix} x_2^2 - x_3(1+x_3) - g_{\varepsilon, t_*}(t) \\ -x_3 - g_{\varepsilon, t_*}(t) \\ x_1 - x_2 \end{bmatrix} = 0.$$

<sup>2</sup>The special construction of  $\mathbf{z}$  ensures that the partial derivative  $\mathbf{v}'_u$  always exists. Since  $\bar{Z}\bar{G}_2^{-1}b$  is smooth,  $\phi_u(t) = \mathbf{z}(u, t)$  is a  $C^1$ -mapping for every fixed  $u$ . As  $D$  is a  $C^1$  mapping, too, we have  $\phi_u \in C_D^1$ . Thus the partial derivative  $\mathbf{v}'_t$  exists and is continuous.

This equation was already considered in the introduction. Recall that the piecewise linear function

$$g_{\varepsilon, t_*}(t) = \begin{cases} 0 & , \quad 0 \leq t < t_* - \varepsilon \\ \frac{t-t_*}{2\varepsilon} + \frac{1}{2} & , \quad t_* - \varepsilon \leq t < t_* + \varepsilon \\ 1 & , \quad t_* + \varepsilon \leq t \end{cases}$$


may represent an independent source that is switched on at  $t = t_*$ . Let  $t_* > 0$  and  $0 < \varepsilon < t_*$ .

We will use the initialisation  $(y^0, x^0, t_0) = \left( [-\frac{5}{4} \ -\frac{1}{2}]^T, [1 \ 1 \ -\frac{1}{2}]^T, 0 \right)$  since  $Ay^0 + b(x^0, t_0) = 0$ . We start by calculating the matrix sequence

$$\begin{aligned} G_0 &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, & B_0 &= \begin{bmatrix} 0 & 2x_2 & -2x_3-1 \\ 0 & 0 & -1 \\ 1 & -1 & 0 \end{bmatrix}, & N_0 &= \text{span} \left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}, & Q_0 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ G_1 &= G_0 + B_0 Q_0 = \begin{bmatrix} 1 & 0 & -2x_3-1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}, & N_1 &= \text{span} \left\{ \begin{bmatrix} 2x_3+1 \\ 1 \\ 1 \end{bmatrix} \right\} \\ & & S_0 &= \text{span} \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\} = S_1 \end{aligned}$$

As  $N_0 \cap S_0 = N_0$  is independent of  $x$ , the structural condition (\*) holds. Also, due to  $N_1 \cap S_1 = \{0\}$  for  $x_3 \neq 0$  the index is 2. Calculating the canonical projector  $Q_1 = \frac{1}{2x_3} \begin{bmatrix} 2x_3+1 & -2x_3-1 & 0 \\ 1 & -1 & 0 \end{bmatrix}$  onto  $N_1$  along  $S_1$  we find that

$$G_2(x^1, x, t) = \frac{1}{2x_3} \begin{bmatrix} 2x_3(x_2+x_3)-x_3^1 & -2x_2x_3+x_3^1 & -2x_3(2x_3+1) \\ -x_3^1 & 2x_3+x_3^1 & -1 \\ 2x_3 & -2x_3 & 0 \end{bmatrix}.$$

This matrix depends on  $x, t$  and the auxiliary variable  $x^1$ . Choosing  $\bar{x}(t) \equiv x^0$  we find

$$\bar{G}_2(t) = \begin{bmatrix} -1 & 2 & 0 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix}, \quad \bar{Q}_1 = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 1 & 0 \end{bmatrix}, \quad T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \bar{Z} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Instead of the original DAE we turn to investigate (14) written in terms of the new variables  $u, w, z, \eta, \zeta$  and  $t$ ,

$$F(u, w, z, \eta, \zeta, t) = \begin{bmatrix} \eta_1 + \zeta_1 - w_3^2 - w_3 + (u_2 - z_1 + z_2)^2 - g_{\varepsilon, t_*}(t) \\ \eta_2 + \zeta_2 - w_3 - g_{\varepsilon, t_*}(t) \\ -u_2 + z_1 - z_2 + u_1 \end{bmatrix} = 0.$$

The mapping  $\hat{F}_1(u, z, t) = [z_1, u_2 - u_1 + z_2, z_3]^T = 0$  defined in lemma 3.2 allows the determination of  $z = \mathbf{z}(u, t) = [0, u_1 - u_2, 0]^T$  and therefore we get  $\mathbf{v}(u, t) = D(t)\mathbf{z}(u, t) = [0, u_1 - u_2]^T$ . Notice that  $\mathbf{v}(u, t)$  is continuously differentiable with respect to both arguments.

Now we can define the mapping  $\mathfrak{f}(u, w, t) = \begin{bmatrix} w_3^2 + w_3 - u_1^2 + g_{\varepsilon, t_*}(t) \\ w_3^2 + w_3 - u_1^2 + g_{\varepsilon, t_*}(t) \end{bmatrix}$  according to (12) and  $\hat{F}_2(u, w, \zeta, t) = [w_1, w_2, \zeta_1 - \zeta_2 - w_3^2 + u_1^2]^T = 0$  from lemma 3.3 fixes the  $w$  component  $w = \mathfrak{w}(u, \zeta, t) = [0, 0, -\sqrt{\zeta_1 - \zeta_2 + u_1^2}]^T$ . Note that  $u_0 = [1 \ 1]^T$ ,  $w_0 = [0 \ 0 \ -\frac{1}{2}]^T$  and thus  $\zeta_0 = y^0 - \mathfrak{f}(u_0, w_0, t_0) = [0, \frac{3}{4}]^T$ . We needed to choose the negative sign for the root in order to guarantee  $w_0 = \mathfrak{w}(u_0, \zeta_0, t_0)$ .

Finally we arrive at the implicit index-1 DAE

$$\begin{aligned} u' &= \mathbb{f}(u, \mathbb{w}(u, v', \cdot), \cdot) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} [g_{\varepsilon, t_*} + v'_1 - v'_2 - \sqrt{v'_1 - v'_2 + u_1^2}], & u(t_0) &= u_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ v &= D\mathbf{z}(u, \cdot) = \begin{bmatrix} 0 \\ u_1 - u_2 \end{bmatrix}, & v(t_0) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned}$$

$t$  arguments were omitted for better readability. Obviously  $v(t) \equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix}$  is uniquely determined by the initial data and we have to consider the ordinary differential equation  $u'(t) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} [g_{\varepsilon, t_*}(t) - u_1(t)]$ ,  $u(t_0) = u_0$ . The unique solution is given by

$$u_1(t) = u_2(t) = \begin{cases} \alpha(t) & , \quad 0 \leq t < t_* - \varepsilon \\ \alpha(t) + \beta(t) & , \quad t_* - \varepsilon \leq t < t_* + \varepsilon \\ \alpha(t) + \beta(t) + \gamma(t) & , \quad t_* + \varepsilon \leq t \end{cases}$$

where the functions  $\alpha$ ,  $\beta$  and  $\gamma$  are defined by

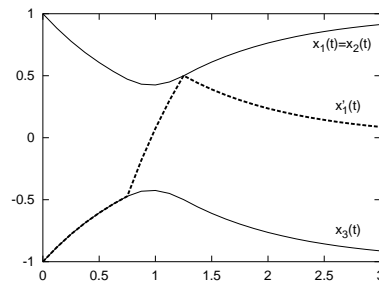
$$\alpha(t) = e^{-t}, \quad \beta(t) = \frac{1}{2} + \frac{1}{2\varepsilon} [t - t_* - 1 + e^{t_* - \varepsilon - t}], \quad \gamma(t) = 1 + \frac{1}{2\varepsilon} [e^{t_* - \varepsilon - t} - e^{t_* + \varepsilon - t}].$$

Following (18) we find

$$x_*(t) = D^-(t)u(t) + \mathbf{z}(u(t), t) + \mathbf{w}(u(t), v'(t), t) = u_1(t) \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}.$$

Direct computation shows that  $x_*$  is indeed a solution. Notice that there was no need for the initialisation to be consistent, i.e. we have  $x^0 \neq x_*(t_0) = [1, 1, -1]^T$ . However,  $x_*$  satisfies the initial condition  $D(t_0)\bar{P}_1(t_0)(x_*(t_0) - x_0) = 0$  as stated in theorem 3.5.

The solution obtained for  $t_* = 1$  and  $\varepsilon = \frac{1}{4}$  is plotted on the right.  $\square$



### 3.3 Some remarks about the decoupling procedure

In contrast to the case of linear DAEs [9, 10] we did not obtain an inherent ordinary differential equation for the nonlinear DAE (3). In fact, we derived the implicit DAE system (15) that governs the dynamical behaviour. Using the concept of the differentiation index it turned out that (15) has index 1.

Of course, the transformation of higher index DAEs to equivalent ones having index 1 is a well-known theme in the theory of differential algebraic equations. But in contrast to classical approaches [1, 2, 6], we didn't use the derivative array at all. The concept of the tractability index made a more refined analysis of index-2 DAEs possible leading to lower smoothness requirements. This is of vital importance for applications.

Indeed, (15) can also be used to analyse numerical methods for index-2 DAEs. The decoupling procedure introduced here is a theoretical device and there won't be any explicit decoupling when doing serious computations. However,

one has to ensure that a given method, when applied to (3), behaves as if it was integrating the index-1 system (15). This will guarantee that numerical results behave as expected.

We remark that (15) is neither in Hessenberg form nor formulated with a properly stated leading term. It is easily seen that reformulations that fit into these classes of equations will have index 2 again. Thus one has to be careful when reformulating the implicit system and it turns out to be advantageous to consider (15) directly.

Finally observe that theorem 3.5 is indeed a generalisation of corresponding statements for linear DAEs. If the DAE (3) was linear,  $A(Dx)' + Bx = q$ , then  $DP_1G_2^{-1}BZ = (DP_1D^-)'DQ_1$  shows that (15) reduces to

$$\begin{aligned} u' &= \mathfrak{f}(u, w, t) = (DP_1D^-)'(t)u - (DP_1G_2^{-1}BD^-)(t)u + (DP_1G_2^{-1}q)(t), \\ v(t) &= (DQ_1G_2^{-1}q)(t). \end{aligned}$$

This is precisely the inherent regular ODE from [9].

But still, for nonlinear equations we can't expect that there is a full decoupling leading to an inherent ODE. This is clear as in [15] it was already observed that errors in the algebraic component  $v$  may influence the dynamical component  $u$ . The system (15) clearly reveals this interconnection.

## 4 Proofs of the results

In this section we prove lemma 3.1, 3.2 and 3.3. Unfortunately, the proofs given here are rather technical. The structure of our approach is depicted in figure 2. We rewrite the DAE (3) in terms of new variables and consider the equation (14),  $F(u, w, z, \eta, \zeta, t) = 0$ . The derivatives  $u'$  and  $v'$  are replaced by parameters  $\eta$  and  $\zeta$ , respectively. Using the fact that  $I = \bar{Z} + P_0\bar{P}_1 + T$ , this equation is split into three parts that can be dealt with one after the other.

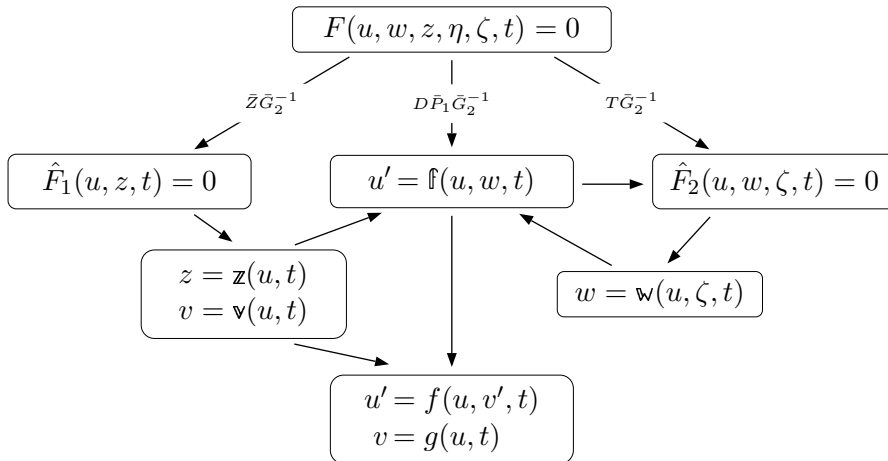


Figure 2: roadmap to the proofs

- ① The first part,  $\hat{F}_1(u, z, t) = 0$ , together with the implicit function theorem yields the mapping  $z$ . Here the structural condition (\*) is crucial.
- ② Inserting  $z = z(u, t)$  into the second part then provides an explicit representation  $u' = \mathfrak{f}(u, w, t)$ .
- ③ Using the information about  $z$  and  $\eta = u'$ , the third part then reads  $\hat{F}_2(u, w, \zeta, t) = 0$  and another application of the implicit function theorem yields  $w = \mathfrak{w}(u, \zeta, t)$ .

To derive the implicit DAE (15), we need to plug  $w$  back into  $\mathfrak{f}$  keeping in mind that  $\zeta$  was representing  $v'$ .

This procedure will now be carried out in detail. We start by giving a preliminary result.

**Lemma 4.1** *Let (3) be a regular DAE with a properly stated leading term that has index 2 on  $\mathcal{D} \times \mathcal{I}$ . Let (A2) and (A1) hold. Then*

$$\bar{G}_2^{-1}(t)A(t)D(t) = \bar{P}_1(t)P_0(t), \quad G_1(\xi, t)P_0(t) = (\bar{G}_2\bar{P}_1P_0)(t),$$

hold for every  $(\xi, t) \in \mathcal{D} \times \mathcal{I}$ . Furthermore we have

$$(T\bar{G}_2^{-1})(t_0)B(x^0, t_0)T(t_0) = T(t_0), \quad (\bar{Z}\bar{G}_2^{-1})(t_0)B(x^0, t_0)\bar{Z}(t_0) = \bar{Z}(t_0).$$

If, in addition, the structural condition (\*) is valid, then

$$\text{im } B(\xi, t)T(t) \subset \text{im } (\bar{G}_2\bar{P}_1P_0)(t) \quad \forall (\xi, t) \in \mathcal{D} \times \mathcal{I}.$$

**Proof:** Since  $\bar{G}_2\bar{P}_1P_0 = AD = G_1(\xi, \cdot)P_0$  for every  $\xi$ , the first two equations hold. The second two equations follow from

$$\begin{aligned} \bar{B}T &= \bar{B}Q_0T = \bar{G}_1Q_0T = \bar{G}_2\bar{P}_1T, \\ \bar{Z}\bar{G}_2^{-1}\bar{B}\bar{Z} &= P_0\bar{Q}_1\bar{G}_2^{-1}\bar{B}P_0\bar{Q}_1 + UQ_0\bar{G}_2^{-1}\bar{B}P_0\bar{Q}_1 + \bar{Z}\bar{G}_2^{-1}\bar{B}Q_0UQ_0 \\ &= P_0\bar{Q}_1 + UQ_0 = \bar{Z}, \end{aligned}$$

respectively<sup>3</sup>. The argument  $t_0$  was dropped for better readability. Notice that for  $t \neq t_0$  we get  $\bar{G}_2^{-1}(t)G_2(x^1, \xi, t) \neq I$  in general and the second two equations need not hold for  $t \neq t_0$ .

Given (\*), we have  $\text{im } T(t) \subset S_0(\xi, t)$  since  $N_0(t) \cap S_0(x, t)$  is independent of  $x$ . We find  $\text{im } B(\xi, t)T(t) \subset \text{im } G_0(t)$  which proves the last assertion of the lemma.  $\square$

**Proof of lemma 3.1:** We assume that  $x_*$  is a solution of (3), define the functions  $u, w, z$  according to (8),

$$u(t) = D(t)\bar{P}_1(t)x_*(t), \quad w(t) = T(t)x_*(t), \quad z(t) = \bar{Z}(t)x_*(t), \quad (8)$$

---

<sup>3</sup>Recall that  $\bar{Q}_1 = \bar{Q}_1\bar{G}_2^{-1}\bar{B}P_0$ ,  $\bar{B}P_0 = \bar{B}_1\bar{Q}_1 + \bar{G}_2\bar{P}_1D^-(D\bar{P}_1D^-)'D\bar{Q}_1$  and  $\bar{B}Q_0 = \bar{G}_2\bar{P}_1Q_0$ .

and consider the mapping

$$0 = f\left((Dx_*)'(t), x_*(t), t\right) = F\left(u(t), w(t), z(t), u'(t), v'(t), t\right), \quad (10a)$$

$$F(u, w, z, \eta, \zeta, t) = f\left(\eta + \zeta, D^-(t)u + \bar{Z}(t)z + T(t)w, t\right). \quad (10b)$$

Using the identity  $I = P_0\bar{P}_1 + \bar{Z} + T = D^-D\bar{P}_1 + \bar{Z} + T$  as motivation we split (10b) into

$$F_1(u, w, z, \eta, \zeta, t) = \bar{Z}(t)\bar{G}_2^{-1}(t)F(u, w, z, \eta, \zeta, t) + (I - \bar{Z}(t))z,$$

$$F_2(u, w, z, \eta, \zeta, t) = T(t)\bar{G}_2^{-1}(t)F(u, w, z, \eta, \zeta, t) + (I - T(t))w,$$

$$F_3(u, w, z, \eta, \zeta, t) = D(t)\bar{P}_1(t)\bar{G}_2^{-1}(t)F(u, w, z, \eta, \zeta, t).$$

Observe that (8) and (10a) imply

$$F_i(u(t), w(t), z(t), u'(t), v'(t), t) = 0, \quad i = 1, 2, 3, \quad (20)$$

for  $t \in \mathcal{I}$ . We study these functions around  $(u_0, w_0, z_0, \eta_0, \zeta_0, t_0)$  where

$$u_0 = u(t_0), \quad w_0 = w(t_0), \quad z_0 = z(t_0), \quad \eta_0 = u'(t_0), \quad \zeta_0 = v'(t_0).$$

As in (9) we have  $x^0 = D^-(t_0)u_0 + z_0 + w_0$ .

① Lemma 4.1 shows that (dropping the  $t$  argument)

$$F_1(u, w, z, \eta, \zeta, \cdot) = \bar{Z}\bar{G}_2^{-1}b(D^-u + \bar{Z}z + Tw, \cdot) + (I - \bar{Z})z$$

does not depend on  $\eta$  nor  $\zeta$ . Due to the structural condition (\*)  $F_1$  is even independent of  $w$  as

$$F'_{1,w}(u, w, z, \eta, \zeta, t) = (\bar{Z}\bar{G}_2^{-1})(t)B(\xi(t), t)T(t) = 0 \quad (21)$$

with  $\xi(t) = D^-(t)u + \bar{Z}(t)z + T(t)w$  (see lemma 4.1). Now we may redefine  $F_1$  using the proper argument list:

$$\begin{aligned} \hat{F}_1(u, z, \cdot) &= \bar{Z}\bar{G}_2^{-1}F(u, 0, z, 0, 0, \cdot) + (I - \bar{Z})z \\ &= \bar{Z}\bar{G}_2^{-1}b(D^-u + \bar{Z}z, \cdot) + (I - \bar{Z})z. \end{aligned} \quad (22)$$

Keep in mind that due to (21)

$$\hat{F}_1(u, z, \cdot) = \bar{Z}\bar{G}_2^{-1}b(D^-u + \bar{Z}z + w_0, \cdot) + (I - \bar{Z})z \quad (23)$$

is also valid. Using lemma 4.1 again, we calculate

$$\hat{F}'_{1,z}(u_0, z_0, t_0) = \bar{Z}(t_0)\bar{G}_2^{-1}(t_0)B(x^0, t_0)\bar{Z}(t_0) + (I - \bar{Z}) = I. \quad (24)$$

(20,  $i=1$ ) and (24) allow the application of the implicit function theorem and  $z(t) = \mathbf{z}(u(t), t)$  is given as a function of  $u(t)$  and  $t$ . The mapping  $\mathbf{z}$  is defined

locally around  $(u_0, t_0)$  and satisfies  $\hat{F}_1(u, \mathbf{z}(u, t), t) = 0$  in a neighbourhood of  $(u_0, t_0)$ . Thus

$$\mathbf{z}(u, t) = \bar{Z}(t)\mathbf{z}(u, t) \quad (25)$$

is also valid since  $0 = (I - \bar{Z}(t))\hat{F}_1(u, \mathbf{z}(u, t), t) = (I - \bar{Z}(t))\mathbf{z}(u, t)$ . Due to  $D(t)\mathbf{z}(u, t) = D(t)\bar{Z}\mathbf{z}(u, t) = D(t)\bar{Q}_1(t)\mathbf{z}(u, t)$  we arrive at

$$\begin{aligned} v(t) &= D(t)\bar{Q}_1(t)z(t) = D(t)\bar{Q}_1(t)\bar{Z}(t)\mathbf{z}(u(t), t) \\ &= D(t)\mathbf{z}(u(t), t) = \mathbf{v}(u(t), t) \end{aligned}$$

with the function  $\mathbf{v}(u, t) = D(t)\mathbf{z}(u, t)$ .

② Noting that (dropping  $t$  arguments)

$$\begin{aligned} D\bar{P}_1\bar{G}_2^{-1}A(u+v)' &= D\bar{P}_1\bar{G}_2^{-1}ADD^-(u+v)' = D\bar{P}_1D^-(u+v)' \\ &= u' - (D\bar{P}_1D^-)'(u+v) \end{aligned}$$

it turns out that  $F_3$  provides an explicit representation of  $u'(\cdot)$  in terms of  $u(\cdot)$  and  $w(\cdot)$ . In particular, (20,  $i=3$ ) is equivalent to

$$u'(t) = \mathbb{F}(u(t), w(t), t) \quad (26a)$$

with

$$\begin{aligned} \mathbb{F}(u, w, t) &= (D\bar{P}_1D^-)'(t)(u + \mathbf{v}(u, t)) \\ &\quad - (D\bar{P}_1\bar{G}_2^{-1})(t)b(D^-(t)u + \mathbf{z}(u, t) + T(t)w, t). \end{aligned} \quad (26b)$$

③ Combining  $F_2$  with the results obtained so far we get

$$\hat{F}_2(u(t), w(t), v'(t), t) = 0 \quad (27a)$$

where

$$\begin{aligned} \hat{F}_2(u, w, \zeta, t) &= F_2(u, w, \mathbf{z}(u, t), \mathbb{F}(u, w, t), \zeta, t) \\ &= (T\bar{G}_2^{-1}A)(t)(\mathbb{F}(u, w, t) + \zeta) + (I - T(t))w \\ &\quad + (T\bar{G}_2^{-1})(t)b(D^-(t)u + \mathbf{z}(u, t) + T(t)w, t) \end{aligned} \quad (27b)$$

is defined on a neighbourhood of  $(u_0, w_0, \zeta_0, t_0)$ . In order to apply the implicit function theorem once again, we need to show that  $\hat{F}'_{2,w}(u_0, w_0, \zeta_0, t_0)$  is nonsingular. After calculating

$$\mathbb{F}'_w(u_0, w_0, t_0) = -(D\bar{P}_1\bar{G}_2^{-1})(t_0)B(x^0, t_0)T(t_0)$$

we obtain from lemma 4.1

$$\begin{aligned} \hat{F}'_{2,w}(u_0, w_0, \zeta_0, t_0) &= (T\bar{G}_2^{-1}A)(t_0)\mathbb{F}'_w(u_0, w_0, t_0) + I - T(t_0) + (T\bar{G}_2^{-1})(t_0)B(x^0, t_0)T(t_0) \\ &= -(T\bar{P}_1P_0\bar{P}_1\bar{G}_2^{-1})(t_0)B(x^0, t_0)T(t_0) + I - T(t_0) + T(t_0). \end{aligned}$$



This proves  $\hat{F}'_{2,w}(u_0, w_0, t_0) = I$ , since  $T\bar{P}_1 P_0 \bar{P}_1 = 0$ . Thus the implicit function theorem shows that locally around  $(u_0, w_0, \zeta_0, t_0)$  (27a) is equivalent to

$$w(t) = \mathfrak{w}(u(t), v'(t), t)$$

with a continuous mapping  $\mathfrak{w}$ . Similar to (25)  $\mathfrak{w}(u, \zeta, t) = T(t)\mathfrak{w}(u, \zeta, t)$  holds.

The results of ①–③ show that (10) implies (11) from lemma 3.1, i.e.

$$z(t) = \mathfrak{z}(u(t), t), \quad u'(t) = \mathfrak{f}(u(t), w(t), t), \quad w(t) = \mathfrak{w}(u(t), v'(t), t).$$

To finish the proof we note that these mappings imply (dropping  $t$  arguments)

$$\begin{aligned} & \bar{G}_2^{-1} F(u, w, z, u', v', \cdot) \\ &= \hat{F}_1(u, \mathfrak{z}(u, \cdot), \cdot) + \hat{F}_2(u, \mathfrak{w}(u, v', \cdot), v', \cdot) + D^-(u' - \mathfrak{f}(u, \mathfrak{w}(u, v', \cdot), \cdot)) = 0. \quad \square \end{aligned}$$

**Proof of lemma 3.2:** Again we drop the assumption that there is a solution. We require that (A1), (A2) and  $(*)$  hold. Exactly as in (22) we define the mapping  $\hat{F}_1 = \hat{F}_1(u, z, t)$  where  $u$  and  $z$  are considered to be parameters chosen in a neighbourhood of  $(u_0, z_0)$ . Recall that  $u_0, w_0, z_0$  are defined in (13). We have

$$\hat{F}_1(u_0, z_0, t_0) = (\bar{Z}\bar{G}_2^{-1})(t_0) [A(t_0)y^0 + b(x^0, t_0)] = 0$$

due to (23) and (A1). As in (24) we find  $\hat{F}'_{1,z}(u_0, z_0, t_0) = I$  and the implicit function theorem provides the function  $\mathfrak{z}$  satisfying  $\hat{F}_1(u, \mathfrak{z}(u, t), t) = 0$  in a neighbourhood of  $(u_0, t_0)$ . Notice that  $\mathfrak{z}$  satisfies (25).  $\square$

**Proof of lemma 3.3:** Having  $\mathfrak{z}$  at our disposal we introduce  $\mathfrak{v}(u, t) = D(t)\mathfrak{z}(u, t)$  and consider the mapping  $\mathfrak{f} = \mathfrak{f}(u, w, t)$  defined in (26b). Now the function  $\hat{F}_2 = \hat{F}_2(u, w, \zeta, t)$  can be defined as in (27b). Let  $\zeta_0 = y^0 - \mathfrak{f}(u_0, w_0, t_0)$  such that

$$\hat{F}_2(u_0, w_0, \zeta_0, t_0) = (T\bar{G}_2^{-1})(t_0) [A(t_0)y^0 + b(x^0, t_0)] = 0.$$

We already calculated  $\hat{F}'_{2,w}(u_0, w_0, t_0) = I$  and therefore the implicit function theorem yields the mapping  $\mathfrak{w}$  with  $\hat{F}_2(u, \mathfrak{w}(u, \zeta, t), \zeta, t) = 0$  in a neighbourhood of  $(u_0, \zeta_0, t_0)$ . Again  $\mathfrak{w}(u, \zeta, t) = T(t)\mathfrak{w}(u, \zeta, t)$  is satisfied.  $\square$

**Remark 4.2** As in lemma 3.4 the decoupling procedure discussed above leads to the implicit DAE

$$u' = \mathfrak{f}(u, \mathfrak{w}(u, v', t), t) =: f(u, v', t) \tag{28a}$$

$$v = D(t)\mathfrak{z}(u, t) =: g(u, t). \tag{28b}$$

It was remarked earlier that (28) has (differentiation) index 1. This follows from the fact that  $M(u, \zeta, t) = I - f'_{v'}(u, \zeta, t)g'_u(u, t)$  with

$$\begin{aligned} M(u_0, \zeta_0, t_0) &= I - (\mathfrak{f}'_w \mathfrak{w}'_{\zeta} \mathfrak{v}'_u)(u_0, \zeta_0, t_0) \\ &= I + (D\bar{P}_1 \bar{G}_2^{-1} \bar{B}T)(t_0) (T\bar{G}_2^{-1} A)(t_0) (D\bar{Q}_1 D^-)(t_0) = I \end{aligned}$$

remains nonsingular locally around  $(u_0, \zeta_0, t_0)$ .  $\square$

**Remark 4.3** In theorem 3.5 a solution  $x_*$  was constructed by first solving (28) and then defining  $x_*(t) = D^-(t)u(t) + \mathbf{z}(u(t), t) + \mathbf{w}(u(t), v'(t), t)$  as in (18).  $x_*$  is indeed a solution since

$$\bar{Z}\bar{G}_2^{-1} [A(Dx_*)' + b(x_*, \cdot)] = \hat{F}_1(u, \mathbf{z}(u, \cdot), \cdot) = 0, \quad (29a)$$

$$T\bar{G}_2^{-1} [A(Dx_*)' + b(x_*, \cdot)] = \hat{F}_2(u, \mathbf{w}(u, v', \cdot), v', \cdot) = 0, \quad (29b)$$

$$P_0\bar{P}_1\bar{G}_2^{-1} [A(Dx_*)' + b(x_*, \cdot)] = D^-(u' - \mathfrak{f}(u, \mathbf{w}(u, v', \cdot), \cdot)) = 0. \quad (29c)$$

Here we want to remark that in order to see (29a) recall that  $\bar{Z}\bar{G}_2^{-1}AD = 0$  and  $\bar{Z}\bar{G}_2^{-1}b(\xi, \cdot) = \bar{Z}\bar{G}_2^{-1}b(U\xi, \cdot)$ . The latter relation follows from the structural condition (\*) as was seen in (21). Also the property (25) is used. Similarly,  $\mathbf{w}(u, \zeta, t) = T(t)\mathbf{w}(u, \zeta, t)$  and (27b) imply (29b). Finally, (29c) follows from (28a),  $v(t) = \mathbf{v}(u(t), t)$  and the definition of  $\mathfrak{f}$  in (26b).

## References

- [1] K.E. Brenan, S.L. Campbell, and L.R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations*, Classics in Applied Mathematics, vol. 14, Society for Industrial and Applied Mathematics (SIAM), 1996.
- [2] S.L. Campbell and E. Griepentrog, *Solvability of general differential algebraic equations*, SIAM Journal on Scientific Computing 16 (1995), no. 2, 257–270.
- [3] D. Estévez Schwarz, *Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation*, Ph.D. thesis, Humboldt Universität zu Berlin, 2000.
- [4] D. Estévez Schwarz and C. Tischendorf, *Structural analysis of electric circuits and consequences for MNA*, International Journal of Circuit Theory and Applications 28 (2000), 131–162.
- [5] E. Griepentrog and R. März, *Differential-algebraic equations and their numerical treatment*, Teubner, Leipzig, 1986.
- [6] P. Kunkel and V. Mehrmann, *Regular solutions of nonlinear differential-algebraic equations and their numerical determination*, Numer. Math. 79 (1998), no. 4, 581–600.
- [7] R. Lamour, *Index determination for DAEs*, Tech. Report 02-19, Humboldt Universität zu Berlin, 2001.
- [8] R. März, *Nichtlineare Algebro-Differentialgleichungen mit proper formuliertem Hauptterm*, Tech. Report 01-3, Humboldt Universität zu Berlin, 2001.
- [9] ———, *The index of linear differential algebraic equations with properly stated leading terms*, Results in Mathematics 42 (2002), 308–338.
- [10] ———, *Solvability of linear differential algebraic equations with properly stated leading terms*, Tech. Report 02-12, Humboldt Universität zu Berlin, 2002, to appear in Results in Mathematics.
- [11] ———, *Differential algebraic systems with properly stated leading term and MNA equations*, Modeling, simulation, and optimization of integrated circuits (Oberwolfach, 2001), Internat. Ser. Numer. Math., vol. 146, Birkhäuser, Basel, 2003, pp. 135–151.

- [12] R. März and C. Tischendorf, *Solving more general index-2 differential algebraic equations*, Computers and Mathematics with Applications 28 (1994), no. 10–12, 77–105.
- [13] P. Rabier and W. Rheinboldt, *Theoretical and numerical analysis of differential-algebraic equations*, vol. VIII, Elsevier Science B.V., 2002, edited by P.G. Ciarlet and J.L. Lions.
- [14] I. Schumilina, *Charakterisierung von DAEs mit index 3*, Ph.D. thesis, Humboldt Universität zu Berlin, 2004.
- [15] C. Tischendorf, *Feasibility and stability behaviour of the BDF applied to index-2 differential algebraic equations*, Z. Angew. Math. Mech. 75 (1995), no. 12, 927–946.
- [16] ———, *Solution of index-2 differential algebraic equations and its application in circuit simulation*, Ph.D. thesis, Humboldt Universität zu Berlin, 1996.