Chapter 16
Differential-Algebraic Equations: Theory and Simulation

Peter Kunkel

Abstract We give an overview of the theory of unstructured nonlinear DAEs of arbitrary index. The approach is extended to overdetermined consistent DAEs in order to be able to include known first integrals. We then discuss various computational issues for the numerical solution of corresponding DAE problems. These include the design of special Gauß-Newton techniques as well as the treatment of parametrized nonlinear systems in the context of DAEs. Examples demonstrate their applicability and performance.

16.1 Preface

It was in the year 1988. My contract at the *Sonderforschungsbereich 123* of the University of Heidelberg as research assistant was about to expire and could not be prolonged. My supervisor at that time was W. Jäger. While I was searching a new position, aiming at the possibility to earn a habilitation, he organized a fellowship at the research center of IBM in Heidelberg. I seized this opportunity and signed a contract for 9 months. On the first day of the new job, I met two other colleagues starting at the same day. One of them had a permanent contract. The other one was Volker Mehrmann who was on leave from the University of Bielefeld to spend the same 9 months at the research center of IBM. Our common head R. Janßen put us three into the same office. This was the beginning of Volker's and my joint venture. I therefore want to express my sincere thanks to W. Jäger and R. Janßen for their support which brought me in contact with Volker.

P. Kunkel (✉)
Fakultät für Mathematik und Informatik, Mathematisches Institut, Universität Leipzig, Augustusplatz 10, D-04109 Leipzig, Germany
e-mail: kunkel@math.uni-leipzig.de

P. Benner et al. (eds.), *Numerical Algebra, Matrix Theory, Differential-Algebraic Equations and Control Theory*, DOI 10.1007/978-3-319-15260-8_16

16.2 Introduction

Differential-algebraic equations (DAEs) arise if physical systems are modeled that contain constraints restricting the possible states of the systems. Moreover, in modern hierarchical modeling tools like [5], even if the submodels are ordinary differential equations (ODEs), the equations describing how the submodels are linked yield DAEs as overall models.

The general form of a DAE is given by

$$F(t, x, \dot{x}) = 0, \tag{16.1}$$

with $F \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}, \mathbb{R}^m)$ sufficiently smooth, $\mathbb{I} \subseteq \mathbb{R}$ (compact) interval, and $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ open. In this paper, we will not assume any further structure of the equations. It should, however, be emphasized that additional structure should, if possible, be utilized in the numerical treatment when efficiency is an issue. On the other hand, a general approach is of advantage when it is desirable to have no restrictions in the applicability of the numerical procedure.

It is the aim of the present paper to give an overview of the relevant theory of general unstructured nonlinear DAEs with arbitrary index and its impact on the design of numerical techniques for their approximate solution. We will concentrate mainly on the quadratic case, i.e., on the case $m = n$, but also address the overdetermined case $m \geq n$ assuming consistency of the equations. The attractivity of the latter case lies in the fact that we may add known properties of the solution like first integrals to the system, thus enforcing that the generated numerical solution will respect these properties as well. In the discussion of numerical techniques, we focus on two families of Runge-Kutta type one-step methods and the development of appropriate techniques for the solution of the arising nonlinear systems. Besides the mentioned issues on DAE techniques for treating first integrals, we include a discussion on numerical path following and turning point determination in the area of parametrized nonlinear equations, which can also be treated in the context of DAEs combined with root finding. Several examples demonstrate the performance of the presented numerical approaches.

The paper is organized as follows. In Sect. 16.3, we give an overview of the analysis of unstructured regular nonlinear DAEs of arbitrary index. In particular, we present existence and uniqueness results. We discuss how these results can be extended to overdetermined consistent DAEs, thus allowing for the treatment of known first integrals. Section 16.4 is then dedicated to various computational issues. We first present possible one-step methods, develop Gauß-Newton like processes for the treatment of the arising nonlinear systems, which includes a modification to stabilize the numerical solution. After some remarks on the use of automatic differentiation, we show how problems with first integrals and parametrized nonlinear equations can be treated in the context of DAEs. We close with some conclusions in Sect. 16.5.

16.3 Theory of Nonlinear DAEs

Dealing with nonlinear problems, the first step is to require a suitable kind of regularity. In the special case of an ODE $\dot{x} = f(t, x)$, obviously no additional properties besides smoothness must be required to obtain (local) existence and uniqueness of solutions for the corresponding initial value problem. In the special case of a pure algebraic (parametrized) system $F(t, x) = 0$, the typical requirement is given by assuming that $F_x(t, x)$, denoting the Jacobian of F with respect to x, is nonsingular for all relevant arguments. The regularity then corresponds to the applicability of the implicit function theorem allowing to (locally) solve for x in terms of t. In the general case of DAEs, we of course want to include these extreme cases into the definition of a regular problem. Moreover, we want to keep the conditions as weak as possible. The following example gives an idea, how the conditions for regularity should look like.

Example 1 The system

$$\begin{aligned}
\dot{x}_1 &= x_4, \quad & \dot{x}_4 &= 2x_1x_7,\\
\dot{x}_2 &= x_5, \quad & \dot{x}_5 &= 2x_2x_7,\\
\dot{x}_3 &= x_6, \quad & \dot{x}_6 &= -1 - x_7,\\
0 &= x_3 - x_1^2 - x_2^2,
\end{aligned}$$

see [16], describes the movement of a mass point on a paraboloid under the influence of gravity.

Differentiating the constraint twice and eliminating the arising derivatives of the unknowns yields

$$\begin{aligned}
0 &= x_6 - 2x_1x_4 - 2x_2x_5,\\
0 &= -1 - x_7 - 2x_4^2 - 4x_1^2x_7 - 2x_5^2 - 4x_2^2x_7.
\end{aligned}$$

In particular, the so collected three constraints can be solved for x_3, x_6, and x_7 in terms of the other unknowns, leaving, if eliminated, ODEs for these other unknowns. Hence, we may replace the original problem by

$$\begin{aligned}
\dot{x}_1 &= x_4, \quad & \dot{x}_4 &= 2x_1x_7,\\
\dot{x}_2 &= x_5, \quad & \dot{x}_5 &= 2x_2x_7,\\
0 &= x_3 - x_1^2 - x_2^2,\\
0 &= x_6 - 2x_1x_4 - 2x_2x_5,\\
0 &= -1 - x_7 - 2x_4^2 - 4x_1^2x_7 - 2x_5^2 - 4x_2^2x_7.
\end{aligned}$$

◇

From this example, we deduce the following. The solution process may require to differentiate part of the equations such that the solution may depend on the

derivatives of the data. Without assuming structure, it is not known in advance which equations should be differentiated. By the differentiation process, we obtain additional constraints that must be satisfied by a solution.

16.3.1 A Hypothesis

In order to include differentiated data, we follow an idea of Campbell, see [1], and define so-called derivative array equations

$$F_\ell(t, x, \dot{x}, \ddot{x}, \ldots, x^{(\ell+1)}) = 0, \tag{16.2}$$

where the functions $F_\ell \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n, \mathbb{R}^{(l+1)m})$ are defined by stacking the original function F together with its formal time derivatives up to order ℓ, i.e.,

$$F_\ell(t, x, \dot{x}, \ddot{x}, \ldots, x^{(\ell+1)}) = \begin{bmatrix} F(t, x, \dot{x}) \\ \frac{d}{dt} F(t, x, \dot{x}) \\ \vdots \\ (\frac{d}{dt})^\ell F(t, x, \dot{x}) \end{bmatrix}. \tag{16.3}$$

Jacobians of F_l with respect to the selected variables x, y will be denoted by $F_{l;x,y}$ in the following. A similar notation will be used for other functions.

The desired regularity condition should include that the original DAE implies a certain number of constraints, that these constraints should be independent, and that given an initial value satisfying these constraints can always be extended to a local solution. In the case $m = n$, this leads to the following hypothesis.

Hypothesis 1 *There exist (nonnegative) integers μ, a, and d such that the set*

$$\mathbb{L}_\mu = \{(t, x, y) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t, x, y) = 0\} \tag{16.4}$$

associated with F is nonempty and such that for every point $(t_0, x_0, y_0) \in \mathbb{L}_\mu$, there exists a (sufficiently small) neighborhood $\mathbb{V}$ in which the following properties hold:

1. *We have* rank $F_{\mu;y} = (\mu + 1)n - a$ *on $\mathbb{L}_\mu \cap \mathbb{V}$ such that there exists a smooth matrix function Z_2 of size $((\mu + 1)n, a)$ and pointwise maximal rank, satisfying $Z_2^T F_{\mu;y} = 0$ on $\mathbb{L}_\mu \cap \mathbb{V}$.*
2. *We have* rank $Z_2^T F_{\mu;x} = a$ *on $\mathbb{V}$ such that there exists a smooth matrix function T_2 of size (n, d), $d = n - a$, and pointwise maximal rank, satisfying $Z_2^T F_{\mu;x} T_2 = 0$.*
3. *We have* rank $F_{\dot{x}} T_2 = d$ *on $\mathbb{V}$ such that there exists a smooth matrix function Z_1 of size (n, d) and pointwise maximal rank, satisfying* rank $Z_1^T F_{\dot{x}} T_2 = d$.

Note that the local existence of functions Z_2, T_2, Z_1 is guaranteed by the following theorem, see, e.g., [13, Theorem 4.3]. Moreover, it shows that we may assume that they possess (pointwise) orthonormal columns.

Theorem 1 *Let $E \in C^\ell(\mathbb{D}, \mathbb{R}^{m,n})$, $\ell \in \mathbb{N}_0 \cup \{\infty\}$, and assume that* $\operatorname{rank} E(x) = r$ *for all $x \in \mathbb{M} \subseteq \mathbb{D}$, $\mathbb{D} \subseteq \mathbb{R}^k$ open. For every $\hat{x} \in \mathbb{M}$ there exists a sufficiently small neighborhood $\mathbb{V} \subseteq \mathbb{D}$ of $\hat{x}$ and matrix functions $T \in C^\ell(\mathbb{V}, \mathbb{R}^{n,n-r})$, $Z \in C^\ell(\mathbb{V}, \mathbb{R}^{m,m-r})$, with pointwise orthonormal columns such that*

$$ET = 0, \quad Z^T E = 0 \tag{16.5}$$

on $\mathbb{M}$.

The quantity μ denotes how often we must differentiate the original DAE in order to be able to make conclusions about existence and uniqueness of solutions. Typically, such a quantity is called index. To distinguish it from other indices, the quantity μ, if chosen minimally, is called strangeness index of the given DAE.

For linear DAEs, the above hypothesis is equivalent (for sufficiently smooth data) to the assumption of a well-defined differentiation index and thus to regularity of the given linear DAE, see [13]. In the nonlinear case, the hypothesis, of course, should imply some kind of regularity of the given problems.

In the following, we say that F satisfies Hypothesis 1 with (μ, a, d), if Hypothesis 1 holds with the choice μ, a, and d for the required integers.

16.3.2 Implications

In order to show that Hypothesis 1 implies a certain kind of regularity for the given DAE, we revise the approach first given in [12], see also [13].

Let $(t_0, x_0, y_0) \in \mathbb{L}_\mu$ and

$$T_{2,0} = T_2(t_0, x_0, y_0), \quad Z_{1,0} = Z_1(t_0, x_0, y_0), \quad Z_{2,0} = Z_2(t_0, x_0, y_0).$$

Furthermore, let $Z'_{2,0}$ be chosen such that $[Z'_{2,0}\, Z_{2,0}]$ is orthogonal. By Hypothesis 1, the matrices $Z_{2,0}^T F_{\mu;x}(t_0, x_0, y_0)$ and $Z_{2,0}'^T F_{\mu;y}(t_0, x_0, y_0)$ have full row rank. Thus, we can split the variables x and y, without loss of generalization according to $x = (x_1, x_2)$ and $y = (y_1, y_2)$, such that $Z_{2,0}^T F_{\mu;x_2}(t_0, x_0, y_0)$ and $Z_{2,0}'^T F_{\mu;y_2}(t_0, x_0, y_0)$ are nonsingular. Because of

$$\operatorname{rank} F_{\mu;x_2,y_2} = \operatorname{rank} \begin{bmatrix} Z_{2,0}'^T F_{\mu;x_2} & Z_{2,0}'^T F_{\mu;y_2} \\ Z_{2,0}^T F_{\mu;x_2} & Z_{2,0}^T F_{\mu;y_2} \end{bmatrix}$$

and $Z_{2,0}^T F_{\mu;y_2}(t_0, x_0, y_0) = 0$, this implies that $F_{\mu;x_2,y_2}(t_0, x_0, y_0)$ is nonsingular. The implicit function theorem then yields that the equation $F_\mu(t, x_1, x_2, y_1, y_2) = 0$

is locally solvable for x_2 and y_2. Hence, there are locally defined functions $\mathscr{G}$ and $\mathscr{H}$ with

$$F_\mu(t, x_1, \mathscr{G}(t, x_1, y_1), y_1, \mathscr{H}(t, x_1, y_1)) \equiv 0, \tag{16.6}$$

implying the following structure of $\mathbb{L}_\mu$.

Theorem 2 *The set $\mathbb{L}_\mu$ forms a manifold of dimension $n+1$ that can be locally parametrized by variables (t, x_1, y_1), where x_1 consists of d variables from x and y_1 consists of a variables from y.*

In order to examine the implicitly defined functions in more detail, we consider the system of nonlinear equations $H(t, x, y, \alpha) = 0$ with $\alpha \in \mathbb{R}^a$ given by

$$H(t, x, y, \alpha) = \begin{bmatrix} F_\mu(t, x, y) - Z_{2,0}\alpha \\ T_{1,0}^T(y - y_0) \end{bmatrix}, \tag{16.7}$$

where the columns of $T_{1,0}$ form an orthonormal basis of kernel $F_{\mu;y}(t_0, x_0, y_0)$. Obviously, we have that $H(t_0, x_0, y_0, 0) = 0$. Choosing $T'_{1,0}$ such that $[\, T'_{1,0} \;\; T_{1,0} \,]$ is orthogonal, we get

$$\operatorname{rank} H_{y,\alpha} = \operatorname{rank} \begin{bmatrix} F_{\mu;y} & -Z_{2,0} \\ T_{1,0}^T & 0 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} Z_{2,0}'^T F_{\mu;y} T'_{1,0} & Z_{2,0}'^T F_{\mu;y} T_{1,0} & * \\ Z_{2,0}^T F_{\mu;y} T'_{1,0} & Z_{2,0}^T F_{\mu;y} T_{1,0} & -I_a \\ * & I_d & 0 \end{bmatrix},$$

where here and in the following I_k denotes the identity matrix in $\mathbb{R}^{k,k}$ and its counterpart as constant matrix function. It follows that

$$\operatorname{rank} H_{y,\alpha}(t_0, x_0, y_0, 0) = \operatorname{rank} \begin{bmatrix} Z_{2,0}'^T F_{\mu;y}(t_0, x_0, y_0) T'_{1,0} & 0 & 0 \\ 0 & 0 & -I_a \\ 0 & I_d & 0 \end{bmatrix}$$

and $H_{y,\alpha}(t_0, x_0, y_0, 0)$ is nonsingular because $Z_{2,0}'^T F_{\mu;y}(t_0, x_0, y_0) T'_{1,0}$, representing the linear map obtained by the restriction of $F_{\mu;y}(t_0, x_0, y_0)$ to the linear map from its cokernel onto its range, is nonsingular. Thus, the nonlinear equation (16.7) is locally solvable with respect to (y, α), i.e., there are locally defined functions $\hat{F}_2$ and $\mathscr{Y}$ such that

$$F_\mu(t, x, \mathscr{Y}(t, x)) - Z_{2,0}\hat{F}_2(t, x) \equiv 0, \quad T_{1,0}^T(\mathscr{Y}(t, x) - y_0) \equiv 0. \tag{16.8}$$

If we then define $\hat{F}_1$ by

$$\hat{F}_1(t, x, \dot{x}) = Z_{1,0}^T F(t, x, \dot{x}), \tag{16.9}$$

we obtain a DAE

$$\begin{aligned} \hat{F}_1(t, x, \dot{x}) &= 0, \ (d \text{ differential equations}) \\ \hat{F}_2(t, x) &= 0, \quad (a \text{ algebraic equations}) \end{aligned} \tag{16.10}$$

whose properties shall be investigated.

Differentiating (16.8) with respect to x gives

$$F_{\mu;x} + F_{\mu;y}\mathscr{Y}_x - Z_{2,0}\hat{F}_{2;x} = 0.$$

Multiplying with $Z_{2,0}^T$ form the left and evaluating at (t_0, x_0) then yields

$$\hat{F}_{2;x}(t_0, x_0) = Z_{2,0}^T F_{\mu;x}(t_0, x_0, y_0).$$

With the above splitting for x, we have that $\hat{F}_2(t_0, x_0) = 0$ due to the construction of $\hat{F}_2$ and $\hat{F}_{2;x_2}(t_0, x_0)$ being nonsingular due to the choice of the splitting. Hence, we can apply the implicit function theorem once more to obtain a locally defined function $\mathscr{R}$ satisfying

$$\hat{F}_2(t, x_1, \mathscr{R}(t, x_1)) \equiv 0. \tag{16.11}$$

In particular, the set $\mathbb{M} = \hat{F}_2^{-1}(\{0\})$ forms a manifold of dimension $d + 1$.

Lemma 1 *Let $(t_0, x_0, y_0) \in \mathbb{L}_\mu$. Then there is a neighborhood of (t_0, x_0, y_0) such that*

$$\mathscr{R}(t, x_1) = \mathscr{G}(t, x_1, y_1) \tag{16.12}$$

for all (t, x, y) in this neighborhood.

Proof We choose the neighborhood of (t_0, x_0, y_0) to be a ball with center (t_0, x_0, y_0) and sufficiently small radius. In particular, we assume that all implicitly defined functions can be evaluated for the stated arguments.

Differentiating (16.6) with respect to y_1 gives

$$F_{\mu;x_2}\mathscr{G}_{y_1} + F_{\mu;y_1} + F_{\mu;y_2}\mathscr{H}_{y_1} = 0,$$

where we omitted the argument $(t_1, x_1, \mathscr{G}(t, x_1, y_1), y_1, \mathscr{H}(t, x_1, y_1))$. If we multiply this with $Z_2(t_1, x_1, \mathscr{G}(t, x_1, y_1), y_1, \mathscr{H}(t, x_1, y_1))^T$, defined according to Hypothesis 1, we get $Z_2^T F_{\mu;x_2}\mathscr{G}_{y_1} = 0$. Since $Z_2^T F_{\mu;x_2}$ is nonsingular for a sufficiently small radius of the neighborhood, it follows that $\mathscr{G}_{y_1}(t, x_1, y_1) = 0$.

Inserting $x_2 = \mathscr{R}(t, x_1)$ into the first relation of (16.8) and splitting $\mathscr{Y}$ according to y, we obtain

$$F_\mu(t, x_1, \mathscr{R}(t, x_1), \mathscr{Y}_1(t, x_1, \mathscr{R}(t, x_1)), \mathscr{Y}_2(t, x_1, \mathscr{R}(t, x_1))) = 0.$$

Comparing with (16.6), this yields

$$\mathscr{R}(t,x_1) = \mathscr{G}(t,x_1,\mathscr{Y}_1(t,x_1,\mathscr{R}(t,x_1))).$$

With this, we further obtain, setting $\tilde{y}_1 = \mathscr{Y}_1(t,x_1,\mathscr{R}(t,x_1))$ for short, that

$$\begin{aligned}\mathscr{G}(t,x_1,y_1) - \mathscr{R}(t,x_1) &= \mathscr{G}(t,x_1,y_1) - \mathscr{G}(t,x_1,\tilde{y}_1)\\ &= \mathscr{G}(t,x_1,\tilde{y}_1 + s(y_1-\tilde{y}_1))|_0^1\\ &= \textstyle\int_0^1 \mathscr{G}_{y_1}(t,x_1,\tilde{y}_1 + s(y_1-\tilde{y}_1))(y_1-\tilde{y}_1)\,ds = 0.\end{aligned}$$

□

With the help of Lemma 1, we can simplify the relation (16.6) to

$$F_\mu(t,x_1,\mathscr{R}(t,x_1),y_1,\mathscr{H}(t,x_1,y_1)) \equiv 0. \tag{16.13}$$

Theorem 3 *Consider a sufficiently small neighborhood of* $(t_0,x_0,y_0) \in \mathbb{L}_\mu$. *Let* $\hat{F}_2$ *and* $\mathscr{R}$ *be well-defined according to the above construction and let* (t,x) *with* $x = (x_1,x_2)$ *be given such that* (t,x) *is in the domain of* $\hat{F}_2$ *and* (t,x_1) *is in the domain of* $\mathscr{R}$. *Then the following statements are equivalent:*

(a) There exists y *such that* $F_\mu(t,x,y) = 0$.
(b) $\hat{F}_2(t,x) = 0$.
(c) $x_2 = \mathscr{R}(t,x_1)$.

Proof The statements (b) and (c) are equivalent due to the implicit function theorem defining $\mathscr{R}$. Assuming (a), let there be y such that $F_\mu(t,x,y) = 0$. Then, $x_2 = \mathscr{G}(t,x_1,y_1) = \mathscr{R}(t,x_1)$ due to the implicit function theorem defining $\mathscr{G}$ and Lemma 1. Assuming (c), we set $y = \mathscr{Y}(t,x)$. With $\hat{F}_2(t,x) = 0$, the relation (16.8) yields $F_\mu(t,x,y) = 0$. □

Theorem 4 *Let* F *from (16.1) satisfy Hypothesis 1 with* (μ,a,d). *Then,* $\hat{F} = (\hat{F}_1,\hat{F}_2)$ *satisfies Hypothesis 1 with* $(0,a,d)$.

Proof Let $\hat{\mathbb{L}}_0 = \hat{F}^{-1}(\{0\})$ and let $\hat{Z}_2,\hat{T}_2,\hat{Z}_1$ denote the matrix functions belonging to $\hat{F}$ as addressed by Hypothesis 1.

For $(t_0,x_0,y_0) \in F_\mu^{-1}(\{0\})$, the above construction yields $\hat{F}_2(t_0,x_0) = 0$. If $\dot{x}_0$ denotes the first n components of y_0, then $F(t_0,x_0,\dot{x}_0) = 0$ holds as first block of $F_\mu(t_0,x_0,y_0) = 0$ implying $\hat{F}_1(t_0,x_0,\dot{x}_0) = 0$. Hence, $(t_0,x_0,\dot{x}_0) \in \hat{\mathbb{L}}_0$ and $\hat{\mathbb{L}}_0$ is not empty.

Since $Z_{1,0}^T F_{\dot{x}}(t_0,x_0,\dot{x}_0)$ possesses full row rank due to Hypothesis 1, we may choose $\hat{Z}_2^T = [\,0\ \ I_a\,]$. Differentiating (16.8) with respect to x yields

$$F_{\mu;x} + F_{\mu;y}\mathscr{Y}_x - Z_{2,0}\hat{F}_{2;x} = 0.$$

Multiplying with Z_2^T from the left, we get $Z_2^T Z_{2,0} \hat{F}_{2;x} = Z_2^T F_{\mu;x}$, where $Z_2^T Z_{2,0}$ is nonsingular in a neighborhhood of (t_0, x_0, y_0). Hence, we have

$$\text{kernel } \hat{F}_{2;x} = \text{kernel } Z_2^T F_{\mu;x}$$

such that we can choose $\hat{T}_2 = T_2$. The claim then follows since $\hat{F}_{1;\dot{x}} T_2 = Z_{1,0}^T F_{\dot{x}} T_2$ possesses full column rank due to Hypothesis 1. □

Since (16.10) has vanishing strangeness index, it is called a reduced DAE belonging to the original possibly higher index DAE (16.1). Note that a reduced DAE is defined in a neighborhood of every $(t_0, x_0, y_0) \in \mathbb{L}_\mu$, but also that it is not uniquely determined by the original DAE even for a fixed $(t_0, x_0, y_0) \in \mathbb{L}_\mu$. What is uniquely determined for a fixed $(t_0, x_0, y_0) \in \mathbb{L}_\mu$ is (at least when treating it as a function germ) the function $\mathscr{R}$.

Every continuously differentiable solution of (16.10) will satisfy $x_2 = \mathscr{R}(t, x_1)$ pointwise. Thus, it will also satisfy $\dot{x}_2 = \mathscr{R}_t(t, x_1) + \mathscr{R}_{x_1}(t, x_1)\dot{x}_1$ pointwise. Using these two relations, we can reduce the relation $\hat{F}_1(t, x_1, x_2, \dot{x}_1, \dot{x}_2) = 0$ of (16.10) to

$$\hat{F}_1(t, x_1, \mathscr{R}(t, x_1), \dot{x}_1, \mathscr{R}_t(t, x_1) + \mathscr{R}_{x_1}(t, x_1)\dot{x}_1) = 0. \tag{16.14}$$

If we now insert $x_2 = \mathscr{R}(t, x_1)$ into (16.8), we obtain

$$F_\mu(t, x_1, \mathscr{R}(t, x_1), \mathscr{Y}(t, x_1, \mathscr{R}(t, x_1))) = 0. \tag{16.15}$$

Differentiating this with respect to x_1 yields

$$F_{\mu;x_1} + F_{\mu;x_2}\mathscr{R}_{x_1} + F_{\mu;y}(\mathscr{Y}_{x_1} + \mathscr{Y}_{x_2}\mathscr{R}_{x_1}) = 0.$$

Multiplying with Z_2^T from the left, we get

$$Z_2^T \begin{bmatrix} F_{\mu;x_1} & F_{\mu;x_2} \end{bmatrix} \begin{bmatrix} I_d \\ \mathscr{R}_{x_1} \end{bmatrix} = 0.$$

Comparing with Hypothesis 1, we see that we may choose

$$T_2 = \begin{bmatrix} I_d \\ \mathscr{R}_{x_1} \end{bmatrix}. \tag{16.16}$$

Differentiating now (16.14) with respect to $\dot{x}_1$ and using the definition of $\hat{F}_1$, we find

$$Z_{1,0}^T F_{\dot{x}_1} + Z_{1,0}^T F_{\dot{x}_2}\mathscr{R}_{x_1} = Z_{1,0}^T F_{\dot{x}} T_2,$$

which is nonsingular due to Hypothesis 1. In order to apply the implicit function theorem, we need to require that $(t_0, x_{10}, \dot{x}_{10})$ solves (16.14). Note that this is not

a consequence of $(t_0, x_0, y_0) \in \mathbb{L}_\mu$. Under this additional requirement, the implicit function theorem implies the local existence of a function $\mathscr{L}$ satisfying

$$\hat{F}_1(t, x_1, \mathscr{R}(t, x_1), \mathscr{L}(t, x_1), \mathscr{R}_t(t, x_1) + \mathscr{R}_{x_1}(t, x_1)\mathscr{L}(t, x_1)) \equiv 0. \tag{16.17}$$

With the help of the functions $\mathscr{L}$ and $\mathscr{R}$, we can formulate a further DAE of the form

$$\begin{aligned} \dot{x}_1 &= \mathscr{L}(t, x_1), \ (d \text{ differential equations}) \\ x_2 &= \mathscr{R}(t, x_1). \ (a \text{ algebraic equations}) \end{aligned} \tag{16.18}$$

Note that this DAE consists of a decoupled ODE for x_1, where we can freely impose an initial condition as long as we remain in the domain of $\mathscr{L}$. Having so fixed x_1, the part x_2 follows directly from the second relation. In this sense, (16.18) can be seen as a prototype for a regular DAE.

The further discussion is now dedicated to the relation between (16.18) and the original DAE.

We start with the assumption that the original DAE (16.1) possesses a smooth local solution x^* in the sense that there is a continuous path $(t, x^*(t), \mathscr{P}(t)) \in \mathbb{L}_\mu$ defined on a neighborhood of t_0, where the first block of $\mathscr{P}$ coincides with $\dot{x}^*$. Note that if x^* is $(\mu+1)$-times continuously differentiable we can just take the path given by $\mathscr{P} = (\dot{x}^*, \ddot{x}^*, \ldots, (d/dt)^{\mu+1}x^*)$. Setting $(t_0, x_0, y_0) = (t_0, x^*(t_0), \mathscr{P}(t_0))$, Theorem 3 yields that $x_2^*(t) = \mathscr{R}(t, x_1^*(t))$. Hence, $\dot{x}_2^*(t) = \mathscr{R}_t(t, x_1^*(t)) + \mathscr{R}_{x_1}(t, x_1^*(t))\dot{x}_1^*(t)$. In particular, Eq. (16.14) is solved by $(t, x_1, \dot{x}_1) = (t, x_1^*, \dot{x}_1^*)$. Thus, it follows also that $\dot{x}_1^*(t) = \mathscr{L}(t, x_1^*(t))$. In this way, we have proven the following theorem.

Theorem 5 *Let F from (16.1) satisfy Hypothesis 1 with (μ, a, d). Then every local solution x^* of (16.1) in the sense that it extends to a continuous local path $(t, x^*(t), \mathscr{P}(t)) \in \mathbb{L}_\mu$, where the first block of $\mathscr{P}$ coincides with $\dot{x}^*$, also solves the reduced problems (16.10) and (16.18).*

16.3.3 The Way Back

To show a converse result to Theorem 5, we need to require the solvability of (16.14) for the local existence of the function $\mathscr{L}$. For this, we assume that F not only satisfies Hypothesis 1 with (μ, a, d), but also with $(\mu+1, a, d)$. Let now $(t_0, x_0, y_0, z_0) \in \mathbb{L}_{\mu+1}$. Due to the construction of F_ℓ, we have

$$F_{\mu+1} = \begin{bmatrix} F_\mu \\ (\frac{d}{dt})^{\mu+1}F \end{bmatrix}, \quad F_{\mu+1;y,z} = \begin{bmatrix} F_{\mu;y} & 0 \\ ((\frac{d}{dt})^{\mu+1}F)_y & ((\frac{d}{dt})^{\mu+1}F)_z \end{bmatrix}, \tag{16.19}$$

where the independent variable z is a short-hand notation for $x^{(\mu+2)}$. Since $F_{\mu;y}$ and $F_{\mu+1;y,z}$ are assumed to have the same rank drop, we find that Z_2 belonging to F_μ satisfies

$$[\, Z_2^T \;\; 0\,]F_{\mu+1;y,z} = [\, Z_2^T \;\; 0\,]\begin{bmatrix} F_{\mu;y} & 0 \\ ((\frac{d}{dt})^{\mu+1}F)_y & ((\frac{d}{dt})^{\mu+1}F)_z \end{bmatrix} = [\,0 \;\; 0\,].$$

Consequently, in Hypothesis 1 considered for $F_{\mu+1}$, we may choose $[\, Z_2^T \;\; 0\,]$ describing the left nullspace of $F_{\mu+1;y,z}$ such that the same choices are possible for T_2 and Z_1.

Observing that we may write the independent variables (t, x, y, z) also as $(t, x, \dot{x}, \dot{y})$ by simply changing the partitioning of the blocks, and that the equation $F_{\mu+1} = 0$ contains $F_\mu = 0$ as well as $\frac{d}{dt}F_\mu = 0$, which has the form

$$\tfrac{d}{dt}F_\mu = F_{\mu;t} + F_{\mu;x}\dot{x} + F_{\mu;y}\dot{y} = 0,$$

we get

$$Z_2^T F_{\mu;t} + Z_2^T F_{\mu;x}\dot{x} = 0.$$

Using the same splitting $x = (x_1, x_2)$ as above and $\dot{x} = (\dot{x}_1, \dot{x}_2)$ accordingly, we obtain

$$Z_2^T F_{\mu;t} + Z_2^T F_{\mu;x_1}\dot{x}_1 + Z_2^T F_{\mu;x_2}\dot{x}_2 = 0,$$

which yields

$$\dot{x}_2 = -(Z_2^T F_{\mu;x_2})^{-1}(Z_2^T F_{\mu;t} + Z_2^T F_{\mu;x_1}\dot{x}_1). \tag{16.20}$$

On the other hand, differentiation of (16.13) with respect to t yields

$$F_{\mu;t} + F_{\mu;x_1}\dot{x}_1 + F_{\mu;x_2}(\mathscr{R}_t + \mathscr{R}_{x_1}\dot{x}_1) + F_{\mu;y_1}\dot{y}_1 + F_{\mu;y_2}(\mathscr{H}_t + \mathscr{H}_{x_1}\dot{x}_1 + \mathscr{H}_{y_1}\dot{y}_1) = 0$$

and thus

$$Z_2^T F_{\mu;t} + Z_2^T F_{\mu;x_1}\dot{x}_1 = -Z_2^T F_{\mu;x_2}(\mathscr{R}_t + \mathscr{R}_{x_1}\dot{x}_1).$$

Inserting this into (16.20) yields

$$\dot{x}_2 = \mathscr{R}_t + \mathscr{R}_{x_1}\dot{x}_1. \tag{16.21}$$

Hence, the given point $(t_0, x_0, \dot{x}_0, \dot{y}_0)$ satisfies

$$\dot{x}_{20} = \mathscr{R}_t(t_0, x_{10}) + \mathscr{R}_{x_1}(t_0, x_{10})\dot{x}_{10}.$$

It then follows that $(t_0, x_{10}, \dot{x}_{10})$ solves (16.14). In particular, this guarantees that the implicit function theorem is applicable to (16.14) leading to a locally defined $\mathscr{L}$. Thus, the reduced system (16.18) is locally well-defined. Moreover, for every initial value for x_1 near x_{10}, the initial value problem for x_1 in (16.18) possesses a solution x_1^*. The second equation in (16.18) then yields a locally defined x_2^* such that $x^* = (x_1^*, x_2^*)$ forms a solution of (16.18).

For the same reasons as for $\mathbb{L}_\mu$, the set $\mathbb{L}_{\mu+1}$ can be locally parametrized by $n+1$ variables. Among these variables are again t and x_1. But since x_2, $\dot{x}_1$, and $\dot{x}_2$ are all functions of (t, x_1), the remaining variables, say p, are now from $\dot{y}$. In particular, there is a locally defined function $\mathscr{Z}$ satisfying

$$F_{\mu+1}(t, x_1, \mathscr{R}(t, x_1), \mathscr{L}(t, x_1), \mathscr{R}_t(t, x_1) + \mathscr{R}_{x_1}(t, x_1)\mathscr{L}(t, x_1), \mathscr{Z}(t, x_1, p)) \equiv 0.$$

Choosing now $x_1^*(t)$ for x_1 and $p^*(t)$ arbitrarily within the domain of $\mathscr{Z}$, for example $p^*(t) = p_0$, where p_0 is the matching part of $\dot{y}_0$, yields

$$F_{\mu+1}(t, x_1^*(t), x_2^*(t), \dot{x}_1^*(t), \dot{x}_2^*(t), \mathscr{Z}(t, x_1^*(t), p^*(t))) \equiv 0,$$

which contains

$$F(t, x_1^*(t), x_2^*(t), \dot{x}_1^*(t), \dot{x}_2^*(t)) \equiv 0 \tag{16.22}$$

in the first block. But this means nothing else than that $x^* = (x_1^*, x_2^*)$ locally solves the original problem. Moreover, locally there is a continuous function $\mathscr{P}$ such that its first block coincides with $\dot{x}^*$ and $(t, x^*(t), \mathscr{P}(t)) \in \mathbb{L}_\mu$. Summarizing, we have proven the following statement.

Theorem 6 *If F satisfies Hypothesis 1 with (μ, a, d) and $(\mu+1, a, d)$ then every local solution x^* of the reduced DAE (16.18) is also a local solution of the original DAE. Moreover, it extends to a continuous local path $(t, x^*(t), \mathscr{P}(t)) \in \mathbb{L}_\mu$, where the first block of $\mathscr{P}$ coincides with $\dot{x}^*$.*

The numerical treatment of DAEs is usually based on the assumption that there is a solution to be computed. In view of Theorem 5 it is therefore sufficient to work with the derivative array F_μ. However, we must assume in addition that the given point $(t_0, x_0, y_0) \in \mathbb{L}_\mu$ provides suitable starting values for the nonlinear system solvers being part of the numerical procedure. Note that this corresponds to the assumption that we may apply the implicit function theorem for the definition of $\mathscr{L}$.

16.3.4 Overdetermined Consistent DAEs

Hypothesis 1 can be generalized in various ways. For example, we may include underdetermined problems which would cover control problems by treating states and controls as indistinguishable parts of the unknown. We may also allow

overdetermined problems or problems with redundant equations. The main problem in the formulation of corresponding hypotheses is for which points to require properties of the Jacobians of the derivative array equation. Note that the restriction in Hypothesis 1 to points in the solution set of the derivative array equation leads to better covariance properties of the hypothesis, see [13], but it excludes problems where this set is empty, e.g., linear least-squares problems. In the following, we want to present a generalization to overdetermined, but consistent (i.e., solvable) DAEs. Such DAEs may arise by extending a given DAE by some or all hidden constraints, i.e., relations contained in $\hat{F}_2(t,x) = 0$ that require the differentiation of the original DAE, or by extending a given DAE or even an ODE by known first integrals.

Hypothesis 2 *There exist (nonnegative) integers μ, a, d, and v such that the set*

$$\mathbb{L}_\mu = \{(t,x,y) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t,x,y) = 0\} \tag{16.23}$$

associated with F is nonempty and such that for every point $(t_0,x_0,y_0) \in \mathbb{L}_\mu$, there exists a (sufficiently small) neighborhood $\mathbb{V}$ in which the following properties hold:

1. *We have* rank $F_{\mu;y} = (\mu+1)m - v$ *on $\mathbb{L}_\mu \cap \mathbb{V}$ such that there exists a smooth matrix function Z_2 of size $((\mu+1)m, v)$ and pointwise maximal rank, satisfying $Z_2^T F_{\mu;y} = 0$ on $\mathbb{L}_\mu \cap \mathbb{V}$.*
2. *We have* rank $Z_2^T F_{\mu;x} = a$ *on $\mathbb{V}$ such that there exists a smooth matrix function T_2 of size (n,d), $d = n - a$, and pointwise maximal rank, satisfying $Z_2^T F_{\mu;x} T_2 = 0$.*
3. *We have* rank $F_{\dot{x}} T_2 = d$ *on $\mathbb{V}$ such that there exists a smooth matrix function Z_1 of size (m,d) and pointwise maximal rank, satisfying* rank $Z_1^T F_{\dot{x}} T_2 = d$.

A corresponding construction as for Hypothesis 1 shows that Hypothesis 2 implies a reduced DAE of the form (16.10) with the same properties as stated there. In particular, a result similar to Theorem 5 holds. Due to the assumed consistency, the omitted relations (the reduced DAEs are $m - n$ scalar relations short) do not contradict these equations. Thus, the solutions fixed by the reduced DAE will be solutions of the original overdetermined DAE under assumptions similar to those of Theorem 6. Since the arguments are along the same lines as presented above, we omit details here.

An example for a problem covered by Hypothesis 2 is given by Example 1 when we just add the two equations obtained by differentiation and elimination to the original DAE leading to a problem consisting of 9 equations in 7 unknowns. A second example, which we will also address in the numerical experiments, consists of an ODE with known first integral.

Example 2 A simple predator/prey model is described by the so-called Lotka/Volterra system

$$\dot{x}_1 = x_1(1 - x_2), \quad \dot{x}_2 = -c\, x_2(1 - x_1),$$

where $c > 0$ is some given constant, see, e.g., [14]. It is well-known that

$$H(x_1, x_2) = c(x_1 - \log x_1) + (x_2 - \log x_2)$$

is a first integral of this system implying that the positive solutions are periodic. The combined overdetermined system

$$\begin{aligned} \dot{x}_1 &= x_1(1 - x_2), \\ \dot{x}_2 &= -c\, x_2(1 - x_1), \\ c(x_1 - \log x_1) &+ (x_2 - \log x_2) = H_0, \end{aligned}$$

where $H_0 = H(x_{10}, x_{20})$ for given initial values $x_1(t_0) = x_{10}$, $x_2(t_0) = x_{20}$, is therefore consistent. Moreover, it can be shown to satisfy Hypothesis 2 with $\mu = 0$, $a = 1$, $d = 1$, and $v = 1$. In contrast to Example 1, we cannot decide in advance which of the two differential equations should be used together with the algebraic constraint. For stability reasons, we should rather use an appropriate linear combination of the two differential equations. But this just describes the role of Z_1 in Hypothesis 2. ◇

16.4 Integration of Nonlinear DAEs

In this section, we discuss several issues that play a role when one wants to integrate DAE systems numerically in an efficient way.

16.4.1 Discretizations

The idea for developing methods for the numerical solution of unstructured DAEs is to discretize not the original DAE (16.1) but the reduced DAE (16.10) because of its property that it does not contain hidden constraints, i.e., that we do not need to differentiate the functions in the reduced DAE. Of course, the functions in the reduced DAE are themselves defined by relations that contain differentiations. But these are differentiations of the original function F which may be obtained by hand or by means of automatic differentiation.

A well-known discretization of DAEs are the BDF methods, see, e.g., [6]. We want to concentrate here on two families of one-step methods that are suitable for the integration of DAEs of the form (16.10). In the following, we denote the initial value at t_0 by x_0 and the stepsize by h. The discretization should then fix an approximate solution x_1 at the point $t_1 = t_0 + h$.

The first family of methods are the Radau IIa methods, which are collocation methods based on the Radau nodes

$$0 < \gamma_1 < \cdots < \gamma_s = 1, \tag{16.24}$$

where $s \in \mathbb{N}$ denotes the number of stages, see, e.g., [10]. The discretization of (16.10) then reads

$$\begin{aligned} &\hat{F}_1(t_0 + \gamma_j h, X_j, \tfrac{1}{h}(v_{j0} x_0 + \textstyle\sum_{l=1}^{s} v_{jl} X_l)) = 0, \\ &\hat{F}_2(t_0 + \gamma_j h, X_j) = 0, \qquad\qquad j = 1, \ldots, s, \end{aligned} \tag{16.25}$$

together with $x_1 = X_s$, where X_j, $j = 1, \ldots, s$, denote the stage values of the Runge-Kutta scheme. The coefficients v_{jl} are determined by the nodes (16.24). For details and the proof of the following convergence result, see, e.g., [13].

Theorem 7 *The Radau IIa methods (16.25) applied to a reduced DAE (16.10) are convergent of order* $p = 2s - 1$.

Note that the Radau IIa methods exhibit the same convergence order as in the special case of an ODE. The produced new value x_1 satisfies all the constraints due to the included relation $\hat{F}_2(t_1, x_1) = 0$.

The second family of methods consists of partitioned collocation methods, which use Gauß nodes for the differential equations and Lobatto nodes for the algebraic equations given by

$$0 < \rho_1 < \cdots < \rho_k < 1, \quad 0 = \sigma_0 < \cdots < \sigma_k = 1, \tag{16.26}$$

with $k \in \mathbb{N}$. Observe that we use one more Lobatto node equating thus the order of the corresponding collocation methods for ODEs. The discretization of (16.10) then reads

$$\begin{aligned} &\hat{F}_1(t_0 + \rho_j h, u_{j0} x_0 + \textstyle\sum_{l=1}^{k} u_{jl} X_l, \tfrac{1}{h}(v_{j0} x_0 + \textstyle\sum_{l=1}^{k} v_{jl} X_l)) = 0, \\ &\hat{F}_2(t_0 + \sigma_j h, X_j) = 0, \qquad\qquad j = 1, \ldots, k, \end{aligned} \tag{16.27}$$

together with $x_1 = X_k$. The coefficients u_{jl} and v_{jl} are determined by the nodes (16.26). For details and the proof of the following convergence result, see again [13].

Theorem 8 *The Gauß-Lobatto methods (16.27) applied to a reduced DAE (16.10) are convergent of order* $p = 2k$.

Note that in contrast to the Radau IIa methods, the Gauß-Lobatto methods are symmetric. Thus, they may be prefered when symmetry of the method is an issue, e.g., in the solution of boundary value problems. In the case of an ODE, the Gauß-Lobatto methods reduce to the corresponding Gauß collocation methods. As for the Radau IIa methods, the produced new value x_1 satisfies all the constraints due to the included relation $\hat{F}_2(t_1, x_1) = 0$.

For the actual computation, we lift the discretization from the reduced DAE to the original DAE by using Theorem 3. In particular, we replace every relation of the form $\hat{F}_2(t,x) = 0$ by $F_\mu(t,x,y) = 0$ with the help of an additional unknown y. Note that by this process the system describing the discretization becomes underdetermined. Nevertheless, the desired value x_1 will still (at least locally) be uniquely fixed. The Radau IIa methods then read

$$\begin{aligned} &Z_{1,0}^T F(t_0 + \gamma_j h, X_j, \tfrac{1}{h}(v_{j0}x_0 + \textstyle\sum_{l=1}^s v_{jl} X_l)) = 0, \\ &F_\mu(t_0 + \gamma_j h, X_j, Y_j) = 0, \qquad\qquad j = 1, \dots, s, \end{aligned} \tag{16.28}$$

and the Gauß-Lobatto methods then read

$$\begin{aligned} &Z_{1,0}^T F(t_0 + \rho_j h, u_{j0}x_0 + \textstyle\sum_{l=1}^k u_{jl} X_l, \tfrac{1}{h}(v_{j0}x_0 + \textstyle\sum_{l=1}^k v_{jl} X_l)) = 0, \\ &F_\mu(t_0 + \sigma_j h, X_j, Y_j) = 0, \qquad\qquad j = 1, \dots, k. \end{aligned} \tag{16.29}$$

In the case of overdetermined DAEs governed by Hypothesis 2, the discretizations look the same.

In order to perform a step with the above one-step methods given an initial value $(t_0, x_0, y_0) \in \mathbb{L}_\mu$, we can determine $Z_{1,0}$ along the lines of the above hypotheses. We then must provide starting values for a suitable nonlinear system solver for the solution of the nonlinear systems describing the discretization, typically the Gauß-Newton method or a variant of it. Upon convergence, we obtain a final value (t_1, x_1, y_1) as part of the overall solution (which includes the internal stages), which will then be the initial value for the next step. Note that for performing a Gauß-Newton-like method for these problems, which we will write as $\mathscr{F}(z) = 0$ for short in the following, we must be able to evaluate the function $\mathscr{F}$ and its Jacobian $\mathscr{F}_z$ at given points. Thus, we must be able to evaluate F and F_μ and their Jacobians, which can be done by using automatic differentiation, see below.

16.4.2 Gauß-Newton-Like Processes

The design of the Gauß-Newton-like method is crucial for the efficiency of the approach. Note that we had to replace $\hat{F}_2$ by F_μ thus increasing the number of equations and unknowns significantly. However, there is some structure in the equations that can be utilized in order to improve the efficiency. We will sketch this approach in the following for the case of the Radau IIa discretization. Similar techniques can be applied to the case of the Gauß-Lobatto discretization.

Linearizing the equation $\mathscr{F}(z) = 0$ around some given z yields the linear problem $\mathscr{F}(z) + \mathscr{F}_z(z)\Delta z = 0$ for the correction Δz. The ordinary Gauß-Newton method is then characterized by solving for Δz by means of the Moore-Penrose pseudoinverse $\mathscr{F}_z(z)^+$ of $\mathscr{F}_z(z)$, i.e.,

$$\Delta z = -\mathscr{F}_z(z)^+ \mathscr{F}(z). \tag{16.30}$$

Instead of the Moore-Penrose pseudoinverse, we are allowed to use any other equation-solving generalized inverse of $\mathscr{F}_z(z)$. Due to the consistency of the nonlinear problem to be solved, we are also allowed to perturb the Jacobian as long as the perturbation is sufficiently small or even tends to zero during the iteration.

In the case (16.28), linearization leads to

$$\begin{aligned} &Z_{1,0}^T F_x^j \Delta X_j + Z_{1,0}^T F_{\dot{x}}^j \tfrac{1}{h} \textstyle\sum_{l=1}^s v_{jl} \Delta X_l = -Z_{1,0}^T F^j, \\ &F_{\mu;x}^j \Delta X_j + F_{\mu;y}^j \Delta Y_j = -F_\mu^j, \qquad\qquad j = 1, \ldots, s. \end{aligned} \tag{16.31}$$

which is to be solved for $(\Delta X_j, \Delta Y_j)$, $j = 1, \ldots, s$. The superscript j indicates, that the corresponding function is evaluated at the argument occurring in the j-th equation, i.e., at $(t_0 + \gamma_j h, X_j, \frac{1}{h}(v_{j0} x_0 + \sum_{l=1}^s v_{jl} X_l))$ in the case of F and $(t_0 + \gamma_j h, X_j, Y_j)$ in the case of F_μ. Since (16.28) contains $F_\mu^j = 0$, we will have rank $F_{\mu;y}^j = (\mu + 1)n - a$ at a solution of (16.28) due to Hypothesis 1. Near the solution, the matrix $F_{\mu;y}^j$ is thus a perturbation of a matrix with rank drop a. The idea therefore is to perturb $F_{\mu;y}^j$ to a matrix M_j with rank $M_j = (\mu + 1)n - a$. Such a perturbation can be obtained by rank revealing QR decomposition or by singular value decomposition, see, e.g., [7]. The second part of (16.31) then consists of equations of the form

$$F_{\mu;x}^j \Delta X_j + M_j \Delta Y_j = -F_\mu^j. \tag{16.32}$$

With the help of an orthogonal matrix $[\, Z'_{2,j} \;\; Z_{2,j} \,]$, where the columns of $Z_{2,j}$ form an orthonormal basis of the left nullspace of M_j, we can split (16.32) into

$$Z_{2,j}'^T F_{\mu;x}^j \Delta X_j + Z_{2,j}'^T M_j \Delta Y_j = -Z_{2,j}'^T F_\mu^j, \quad Z_{2,j}^T F_{\mu;x}^j \Delta X_j = -Z_{2,j}^T F_\mu^j. \tag{16.33}$$

The first part can be solved for ΔY_j via the Moore-Penrose pseudoinverse

$$\Delta Y_j = -(Z_{2,j}'^T M_j)^+ Z_{2,j}'^T (F_\mu^j + F_{\mu;x}^j \Delta X_j) \tag{16.34}$$

in terms of ΔX_j, thus fixing a special equation-solving pseudoinverse of the Jacobian under consideration. In order to determine the corrections ΔX_j, we take an orthogonal matrix $[\, T'_{2,j} \;\; T_{2,j} \,]$, where the columns of $T_{2,j}$ form an orthonormal basis of the right nullspace of $Z_{2,j}'^T F_{\mu;x}^j$, which is of full row rank near the solution due to Hypothesis 1. Defining the transformed corrections

$$\Delta V'_j = T_{2,j}'^T \Delta X_j, \quad \Delta V_j = T_{2,j}^T \Delta X_j, \tag{16.35}$$

we have $\Delta X_j = T'_{2,j} \Delta V'_j + T_{2,j} \Delta V_j$ and the second part of (16.33) becomes

$$Z_{2,j}^T F_{\mu;x}^j T'_{2,j} \Delta V'_j = -Z_{2,j}^T F_\mu^j. \tag{16.36}$$

Due to Hypothesis 1, the square matrix $Z_{2,j}^T F_{\mu;x}^j T_{2,j}'$ is nonsingular near a solution such that we can solve for $\Delta V_j'$ to get

$$\Delta V_j' = -(Z_{2,j}^T F_{\mu;x}^j T_{2,j}')^{-1} Z_{2,j}^T F_\mu^j. \tag{16.37}$$

Finally, transforming the equation in the first part of (16.31) to the variables $(\Delta V_j', \Delta V_j)$ and eliminating the terms $\Delta V_j'$ leaves a system in the unknowns ΔV_j, which is of the same size and form as if we would discretize an ODE of d equations by means of the Radau IIa method. This means that we actually have reduced the complexity to that of solving an ODE of the size of the differential part. Solving this system for the quantities ΔV_j and combining these with the already obtained values $\Delta V_j'$ then yields the corrections ΔX_j.

The overall Gauß-Newton-like process, which can be written as

$$\Delta z = -\mathscr{J}(z)^+ \mathscr{F}(z) \tag{16.38}$$

with $\mathscr{J}(z) \to \mathscr{F}_z(z)$ when z converges to a solution, can be shown to be locally and quadratically convergent, see again [13]. Using such a process is indispensable for the efficient numerical solution of unstructured DAEs.

16.4.3 Minimal-Norm-Corrected Gauß-Newton Method

We have implemented the approach of the previous section both for the Radau IIa methods and for the Gauß-Lobatto methods. Experiments show that one can successfully solve nonlinear DAEs even for larger values of μ without having to assume a special structure. Applying it to the problem of Example 1, however, reveals a drawback of the approach described so far. In particular, we observe the following. Trying to solve the problem of Example 1 on a larger time interval starting at $t = 0$, one realizes that the integration terminates at about $t = 14.5$ because the nonlinear system solver fails, cp. Fig. 16.1. A closer look shows that the reason for this is that the undetermined components y, which are not relevant for the solution one is interested in, run out of scale. Scaling techniques cannot avoid the effect. They can only help to make use of the whole range provided by the floating point arithmetic. Using diagonal scaling, the iteration terminates then at about $t = 71.4$, cp. again Fig. 16.1.

Actually, proceeding from numerical approximations (x_i, y_i) at t_i to numerical approximations (x_{i+1}, y_{i+1}) at t_{i+1} consists of two mechanisms. First, we must provide a starting value z for the nonlinear system solver. We call this predictor and write

$$z = \mathfrak{P}(x_i, y_i). \tag{16.39}$$

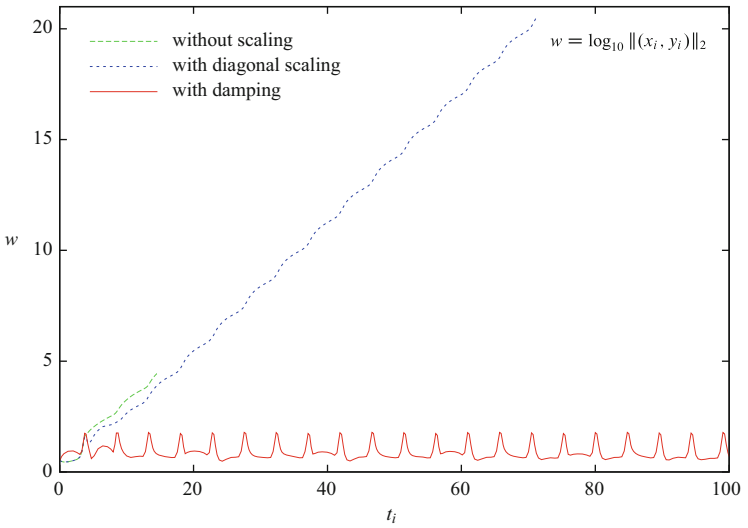

Fig. 16.1 Decadic logarithm of the Euclidean norm of the generated numerical solution (x_i, y_i)

Then, the nonlinear system solver, called corrector in this context, yields the new approximation according to

$$(x_{i+1}, y_{i+1}) = \mathfrak{C}(z). \tag{16.40}$$

Thus the numerical flow Φ of our method effectively has the form

$$(x_{i+1}, y_{i+1}) = \Phi(x_i, y_i), \quad \Phi = \mathfrak{C} \circ \mathfrak{P}. \tag{16.41}$$

The problem can then be described as follows. Even if the actual solution and the numerical approximations x_i are bounded, there is no guaranty that the overall numerical solutions (x_i, y_i) stay bounded.

In [3], it was examined how different predictors $\mathfrak{P}$, in particular extrapolation of some order, influence the overall behavior of the process. The result was that linear extrapolation should be prefered to higher order extrapolation. However, even linear extrapolation cannot avoid the blow-up.

The idea here is to modify the corrector $\mathfrak{C}$, in particular to introduce damping into the nonlinear system solver. Recall that the nonlinear system to be solved does in general not have a unique solution but that the part one is interested in, namely x_{i+1}, is unique. Consider the iteration given by

$$\Delta z = -\alpha z - \mathscr{F}_z(z)^+ (\mathscr{F}(z) - \alpha \mathscr{F}_z(z) z) \tag{16.42}$$

with $\alpha \in [0, 1]$ replacing (16.30). For $\alpha = 0$, we rediscover (16.30). For $\alpha = 1$, we have

$$z + \Delta z = \mathscr{F}_z(z)^+(\mathscr{F}_z(z)z - \mathscr{F}(z)),$$

which in the linear case $\mathscr{F}(z) = \mathbf{A}z - \mathbf{b}$ leads to $z + \Delta z = \mathbf{A}^+\mathbf{b}$ and thus to the shortest solution with respect to the Euclidean norm. In this sense, the process defined by (16.42) contains some damping. Moreover, if $\alpha \to 0$ quadratically during the iteration, we maintain the quadratic convergence of the Gauß-Newton process. The following result is due to [2].

Theorem 9 *Consider the problem $\mathscr{F}(z) = 0$ and assume that the Jacobians $\mathscr{F}_z(z)$ have full row rank. Furthermore, consider the iteration defined by (16.42) and assume that $\alpha \to 0$ quadratically during the iteration. Then the so defined process yields iterates that converge locally and quadratically to a solution of the given problem.*

Observe that replacing (16.30) by (16.42) only consists of a slight modification of the original process. The main computational effort, namely the representation of $\mathscr{F}_z(z)^+$, stays the same. Moreover, using a perturbed Jacobian $\mathscr{J}(z)$ instead of $\mathscr{F}_z(z)$ is still possible and does not influence the convergence behavior. Figure 16.1 shows that with this modified nonlinear system solver we are now able to produce bounded overall solutions in the case of Example 1.

16.4.4 Automatic Differentiation

In order to integrate (unstructured) DAEs, we must provide procedures for the evaluation of F and F_μ together with their Jacobians. As already mentioned this can be done by exploiting techniques from automatic differentiation, see, e.g., [9].

The simplest approach is to evaluate the functions on the fly, i.e., by using special classes and overloaded operators, a call of a template function which implements F can produce the needed evaluations just by changing the class of the variables. The drawback in this approach is that there may be a lot of trivial computations when the derivatives are actually zero. Moreover, no code optimization is possible.

An alternative approach consists of two phases. First, one uses automatic differentiation to produce code for the evaluation of the needed functions. This code can then be easily compiled using optimization. The drawback here is that one has to adapt the automatic differentiation process or the produced code to the form one needs for the following integration of the DAE. Nevertheless, one can expect this approach to be more efficient for the actual integration of the DAE, especially for larger values of μ. Actually, one would prefer the first approach while a model is developed. If the model is finalized, one would then prefer the second approach.

As an example, we have run the problem from Example 1 with both approaches on the interval $[0, 100]$ using the Gauß-Lobatto method for $k = 3$ and the minimal-norm-corrected Gauß-Newton-like method starting with $\alpha = 0.1$ and

using successive squaring. The computing time in the first case exploiting automatic differentiation on the fly was 2.8 s. The computing time in the second case exploiting optimized code produced by automatic differentiation was 0.6 s.

16.4.5 Exploiting First Integrals

If for a given ODE or DAE model first integrals are known, they should be included into the model thus enforcing the produced numerical approximations to obey these first integrals. The enlarged system is of course overdetermined but consistent. In general, it is not clear how to deduce a square system from the overdetermined one in order to apply standard integration procedures, cp. Example 2.

In Example 1, there are two hidden constraints which were found by differentiation. As already mentioned there, it is in this case possible to reduce the problem consisting of the original equations and the two additional constraints to a square system by just omitting two equations of the original system. Sticking to automatic differentiation and using the same setting as above, we can solve the overdetermined system in 0.9 s and the reduced square system in 0.7 s.

For Example 2, such a beforehand reduction is not so obvious, but still possible due to the simple structure of this specific problem. We solved the overdetermined problem by means of the implicit Euler method (which is the Radau IIa method for $s = 1$) as well as the original ODE by means of the explicit and implicit Euler method performing 1,000 steps with stepsize $h = 0.02$. The results are shown in Fig. 16.2. As one would expect, the numerical solution for the ODE

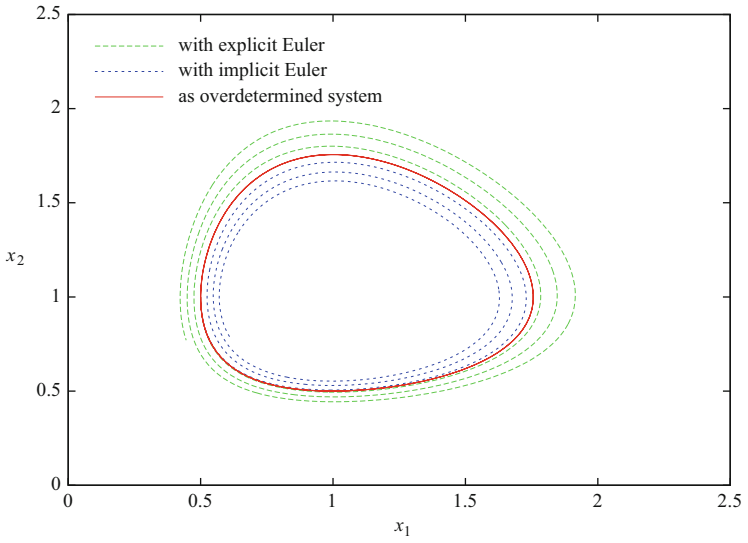

Fig. 16.2 Numerical solutions for the Lotka/Volterra model

produced by the explicit Euler method spirals outwards thus increasing the energy while the numerical solution for the ODE produced by the implicit Euler method spirals inwards thus decreasing energy. The numerical solution obtained from the overdetermined system, of course, conserves the energy by construction.

16.4.6 Path Following by Arclength Parametrization

There are two extreme cases of DAEs, the case of ODEs $\dot{x} = f(t,x)$ on the one hand and the case of nonlinear equations $f(x) = 0$ on the other hand. For $F(t,x,\dot{x}) = \dot{x} - f(t,x)$, Hypothesis 1 is trivially satisfied with $\mu = 0$, $a = 0$, and $d = n$. For $F(t,x,\dot{x}) = f(x)$, Hypothesis 1 is satisfied with $\mu = 0$, $a = n$, and $d = 0$, provided $f_x(x)$ is nonsingular for all $x \in \mathbb{L}_0$. Since t does neither occur as an argument nor via differentiated variables, the solutions are constant in time and thus, as solutions of a DAE, not so interesting. This changes if one considers parameter dependent nonlinear equations $f(x,\tau) = 0$, where τ shall be a scalar parameter. The problem is now underdetermined. Thus, it cannot satisfy one of the above hypotheses. Under the assumption that $[\, f_x \;\; f_\tau \,]$ has full row rank for all $(x,\tau) \in \mathbb{M} = f^{-1}(\{0\}) \neq \emptyset$, the solution set forms a one-dimensional manifold. If one is interested in tracing this manifold, one can use path following techniques, see, e.g., [4, 17]. However, it is also possible to treat such problems with solution techniques for DAEs. A first choice would be to interpret the parameter τ as time t of the DAE. This would, however, imply that the parameter τ is strictly monotone along the one-dimensional manifold. But there are applications, where this is not the case. It may even happen that the points where the parameter τ is extremal are of special interest. In order to treat such problems, we are in need of defining a special type of time which is monotone in any case. Such a quantity is given as the arclength of the one-dimensional manifold, measured say from the initial point we start off. Since the arclength parametrization of a path is characterized by the property that the derivative with respect to the parametrization has Euclidean length one, we consider the DAE

$$f(x,\tau) = 0, \quad \|\dot{x}\|_2^2 + |\dot{\tau}|^2 = 1 \tag{16.43}$$

for the unknown (x,τ). If $(x_0,\tau_0) \in \mathbb{M}$ and $[\, f_x \;\; f_\tau \,]$ is of full row rank on $\mathbb{M}$, the implicit function theorem yields that there is a local solution path $(\hat{x}(t),\hat{\tau}(t))$ passing through (x_0,τ_0). Moreover, $\|\dot{\hat{x}}(t)\|_2^2 + |\dot{\hat{\tau}}(t)|^2 = 1$, when we parametrize by arclength. Hence, the DAE (16.43) possesses a solution. Moreover, writing (16.43) as $F(z,\dot{z}) = 0$ with $z = (x,\tau)$, we have

$$\mathbb{L}_0 = \{(z,\dot{z}) \mid z = (\hat{x}(t),\hat{\tau}(t)),\ \dot{z} = (\dot{\hat{x}}(t),\dot{\hat{\tau}}(t))\}$$

in Hypothesis 1. Because of

$$F_{0;\dot{z}} = \begin{bmatrix} 0 & 0 \\ 2\dot{x}^T & 2\dot{\tau} \end{bmatrix}, \quad F_{0;z} = \begin{bmatrix} f_x & f_\tau \\ 0 & 0 \end{bmatrix},$$

we may choose

$$Z_2 = \begin{bmatrix} I_n \\ 0 \end{bmatrix}.$$

By assumption, $Z_2^T F_{0;z} = [\ f_x \ \ f_\tau\ ]$ has full row rank and we may choose T_2 as a normalized vector in kernel$[\ f_x \ \ f_\tau\ ]$, which is one-dimensional. In particular, we may choose

$$T_2 = \begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{\tau}} \end{bmatrix}$$

on $\mathbb{L}_0$. Finally, we observe that

$$F_{\dot{z}} T_2 = \begin{bmatrix} 0 & 0 \\ 2\dot{\hat{x}}^T & 2\dot{\hat{\tau}} \end{bmatrix} \begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{\tau}} \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

has full column rank at the solution and thus in a neighborhood of it. Hence, the DAE (16.43) satisfies Hypothesis 1 with $\mu = 0$, $a = n$, and $d = 1$, where n denotes the size of x. We can then use DAE solution techniques to solve (16.43) thus tracing the solution path of the original parametrized system of nonlinear equations.

In order to determine points along the path, where the parameter τ is extremal, we may combine the DAE (16.43) with a root finding procedure, e.g., along the lines of [18] or the references therein. The points of interests are characterized by the condition $\dot{\tau} = 0$. We therefore augment the DAE (16.43) according to

$$f(x, \tau) = 0, \quad \|\dot{x}\|_2^2 + |\dot{\tau}|^2 = 1, \quad w - \dot{\tau} = 0, \tag{16.44}$$

and try to locate points along the solution satisfying $w = 0$. Writing the DAE (16.44) again as $F(z) = 0$, where now $z = (x, \tau, w)$, we have

$$\mathbb{L}_0 = \{(z, \dot{z}) \mid z = (\hat{x}(t), \hat{\tau}(t), \dot{\hat{\tau}}(t)),\ \dot{z} = (\dot{\hat{x}}(t), \dot{\hat{\tau}}(t), \ddot{\hat{\tau}}(t))\}$$

in Hypothesis 1. Because of

$$F_{0;\dot{z}} = \begin{bmatrix} 0 & 0 & 0 \\ 2\dot{x}^T & 2\dot{\tau} & 0 \\ 0 & -1 & 0 \end{bmatrix}, \quad F_{0;z} = \begin{bmatrix} f_x & f_\tau & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

we may choose

$$Z_2 = \begin{bmatrix} I_n \\ 0 \\ 0 \end{bmatrix}.$$

Along the same lines as above, we may now choose

$$T_2 = \begin{bmatrix} \dot{\hat{x}} & 0 \\ \dot{\hat{\tau}} & 0 \\ 0 & 1 \end{bmatrix}$$

on $\mathbb{L}_0$. We then observe that

$$F_{\dot{z}} T_2 = \begin{bmatrix} 0 & 0 & 0 \\ 2\dot{\hat{x}}^T & 2\dot{\hat{\tau}} & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{x}} & 0 \\ \dot{\hat{\tau}} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 2 & 0 \\ -\dot{\hat{\tau}} & 0 \end{bmatrix}$$

fails to have full column rank at the solution. Thus, Hypothesis 1 cannot hold with $\mu = 0$. We therefore consider Hypothesis 1 for $\mu = 1$. Starting from

$$F_{1;\dot{z},\ddot{z}} = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & & & \\ 2\dot{x}^T & 2\dot{\tau} & 0 & & & \\ 0 & -1 & 0 & & & \\ \hline f_x & f_\tau & 0 & 0 & 0 & 0 \\ * & * & 0 & 2\dot{x}^T & 2\dot{\tau} & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{array}\right], \quad F_{1;z} = \left[\begin{array}{ccc} f_x & f_\tau & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ \hline * & * & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right],$$

we use the fact that $0 \neq (\dot{x}^T, \tau)^T \in \text{kernel}[\, f_x \;\; f_\tau \,]$ at a solution and therefore

$$\begin{bmatrix} f_x & f_\tau \\ \dot{x}^T & \dot{\tau} \end{bmatrix} \text{ nonsingular}$$

near the solution to deduce that rank $F_{1;\dot{z},\ddot{z}} = n + 3$. Choosing

$$Z_2 = \left[\begin{array}{cc} I_n & 0 \\ 0 & * \\ 0 & 1 \\ \hline 0 & * \\ 0 & 0 \\ 0 & 0 \end{array}\right]$$

gives

$$Z_2^T F_{1;z} = \begin{bmatrix} f_x & f_\tau & 0 \\ * & * & 1 \end{bmatrix},$$

which has full row rank by assumption. Choosing

$$T_2 = \begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{\tau}} \\ * \end{bmatrix}$$

at the solution then yields

$$F_{\dot{z}} T_2 = \begin{bmatrix} 0 & 0 & 0 \\ 2\dot{\hat{x}}^T & 2\dot{\hat{\tau}} & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{\tau}} \\ * \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -\dot{\hat{\tau}} \end{bmatrix}.$$

Hence, the DAE (16.44) satisfies Hypothesis 1 with $\mu = 1$, $a = n + 1$, and $d = 1$, and we can treat (16.44) by the usual techniques. The location of points $\hat{t}$ with $\dot{\tau}(\hat{t}) = 0$ can now be seen as a root finding problem along solutions of (16.44) for the function g defined by

$$g(x, \tau, w) = w \tag{16.45}$$

In particualar, it can be treated by standard means of root finding techniques.

In order to be able to determine a root $\hat{t}$ of g, we need that this root is simple, i.e., that

$$\tfrac{d}{dt} g(\hat{x}(t), \hat{\tau}(t), \hat{w}(t))|_{t=\hat{t}} \neq 0, \quad \hat{w}(t) = \dot{\hat{\tau}}(t). \tag{16.46}$$

In the case of (16.45), this condition simply reads

$$\ddot{\hat{\tau}}(\hat{t}) \neq 0. \tag{16.47}$$

In order to determine $\ddot{\hat{\tau}}(\hat{t})$, we start with $f(\hat{x}(t), \hat{\tau}(t)) = 0$ along the solution. Differentiating twice yields (omitting arguments)

$$f_x \dot{\hat{x}} + f_\tau \dot{\hat{\tau}} = 0 \tag{16.48}$$

and

$$f_{xx}(\dot{\hat{x}}, \dot{\hat{x}}) + f_{x\tau}(\dot{\hat{x}})(\dot{\hat{\tau}}) + f_x \ddot{\hat{x}} + f_{x\tau}(\dot{\hat{x}})(\dot{\hat{\tau}}) + f_{\tau\tau}(\dot{\hat{\tau}}, \dot{\hat{\tau}}) + f_\tau \ddot{\hat{\tau}} = 0. \tag{16.49}$$

Since $\dot{\hat{\tau}}(\hat{t}) = 0$, the relation (16.48) gives

$$f_x(x^*, \tau^*)v = 0, \quad v = \dot{\hat{x}}(\hat{t}) \neq 0, \tag{16.50}$$

with $x^* = \hat{x}(\hat{t})$ and $\tau^* = \hat{\tau}(\hat{t})$ for short. Thus, the square matrix $f_x(x^*, \tau^*)$ is rank-deficient such that there is a vector $u \neq 0$ with

$$u^T f_x(x^*, \tau^*) = 0. \tag{16.51}$$

Multiplying (16.49) with u^T from the left and evaluating at $\hat{t}$ yields

$$u^T f_{xx}(x^*, \tau^*)(v, v) + u^T f_\tau(x^*, \tau^*)\ddot{\hat{\tau}}(\hat{t}) = 0. \tag{16.52}$$

Assuming now that

$$u^T f_{xx}(x^*, \tau^*)(v, v) \neq 0, \quad u^T f_\tau(x^*, \tau^*) \neq 0 \tag{16.53}$$

guarantees

$$\ddot{\hat{\tau}}(\hat{t}) = -(u^T f_\tau(x^*, \tau^*))^{-1}(u^T f_{xx}(x^*, \tau^*)(v, v)) \neq 0. \tag{16.54}$$

Note that the assumptions for (x^*, τ^*) we have required here are just those that characterize a so-called simple turning point, see, e.g., [8, 15].

Example 3 Consider the example

$$\begin{aligned}
&\tau(1 - x_3)\exp(10x_1)/(1 + 0.01x_1) - x_3 = 0,\\
&22\tau(1 - x_3)\exp(10x_1)/(1 + 0.01x_1) - 30x_1 = 0,\\
&x_3 - x_4 + \tau(1 - x_3)\exp(10x_2)/(1 + 0.01x_2) = 0,\\
&10x_1 - 30x_2 + 22\tau(1 - x_4)\exp(10x_2)/(1 + 0.01x_2) = 0,
\end{aligned}$$

from [11]. Starting from the trivial solution into the positive cone, the solution path exhibits six turning points before the solution becomes nearly independent of τ, see Fig. 16.3, which has been produced by solving the corresponding DAE (16.44) by the implicit Euler method combined with standard root finding techniques. ◇

16.5 Conclusions

We revised the theory of regular nonlinear DAEs of arbitrary index and gave some extensions to overdetermined but consistent DAEs. We also discussed several computational issues in the numerical treatment of such DAEs, namely suitable

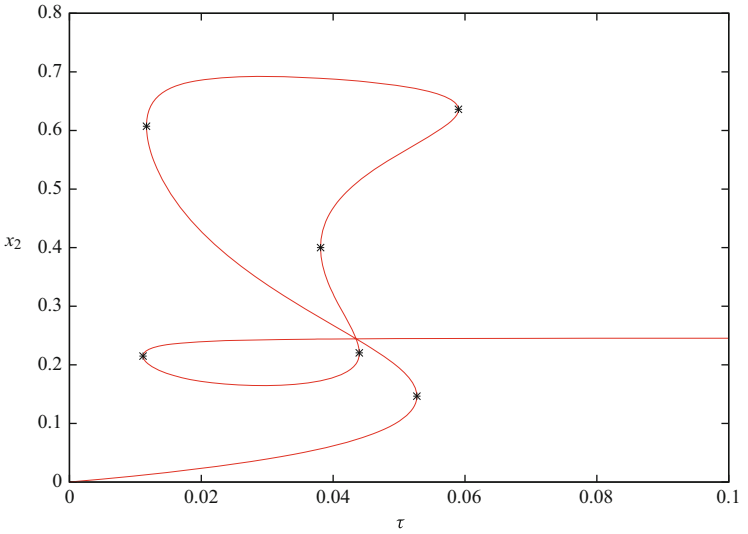

Fig. 16.3 Solution path for Example 3 projected into the (τ, x_2)-plane

discretizations, efficient nonlinear system solvers and their stabilization, as well as automatic differentiation. We finally presented a DAE approach for numerical path following for parametrized systems of nonlinear equations including the detection and determination of (simple) turning points.

References

1. Campbell, S.L.: A general form for solvable linear time varying singular systems of differential equations. SIAM J. Math. Anal. **18**, 1101–1115 (1987)
2. Campbell, S.L., Kunkel, P., Bobinyec, K.: A minimal norm corrected underdetermined Gauß-Newton procedure. Appl. Numer. Math. **62**, 592–605 (2012)
3. Campbell, S.L., Yeomans, K.D.: Behavior of the nonunique terms in general DAE integrators. Appl. Numer. Math. **28**, 209–226 (1998)
4. Deuflhard, P., Fiedler, B., Kunkel, P.: Efficient numerical path following beyond critical points. SIAM J. Numer. Anal. **24**, 912–927 (1987)
5. Fritzson, P.: Principles of Object-Oriented Modeling and Simulation with Modelica. Wiley/IEEE, Hoboken/Piscataway (2003)
6. Gear, C.W.: The simultaneous numerical solution of differential-algebraic equations. IEEE Trans. Circuit Theory **CT-18**, 89–95 (1971)
7. Golub, G.H., Van Loan, C.F.: Matrix Computations, 2nd edn. The Johns Hopkins University Press, Baltimore (1989)
8. Golubitsky, M., Schaeffer, D.: Singularities and Groups in Bifurcation Theory, vol. I. Springer, New York (1984)
9. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, 2nd edn. SIAM, Philadelphia (2008)

10. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II. Springer, Berlin (1991)
11. Kubiček, M.: Algorithm 502. Dependence of solutions of nonlinear systems on a parameter. ACM Trans. Math. Softw. **2**, 98–107 (1976)
12. Kunkel, P., Mehrmann, V.: Regular solutions of nonlinear differential-algebraic equations and their numerical determination. Numer. Math. **79**, 581–600 (1998)
13. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations – Analysis and Numerical Solution. EMS Publishing House, Zürich (2006)
14. Lotka, A.J.: Analytical note on certain rhythmic relations in organic systems. Proc. Natl. Acad. Sci. U.S.A. **6**, 410–415 (1920)
15. Pönisch, G., Schwetlick, H.: Computing turning points of curves implicitly defined by nonlinear equations depending on a parameter. Computing **26**, 107–121 (1981)
16. Rheinboldt, W.C.: Differential-algebraic systems as differential equations on manifolds. Math. Comput. **43**, 473–482 (1984)
17. Rheinboldt, W.C., Burkardt, J.V.: A locally parametrized continuation process. ACM Trans. Math. Softw. **9**, 236–246 (1983)
18. Shampine, L.F., Gladwell, I., Brankin, R.W.: Reliable solution of special event location problems for ODEs. ACM Trans. Math. Softw. **17**, 11–25 (1991)