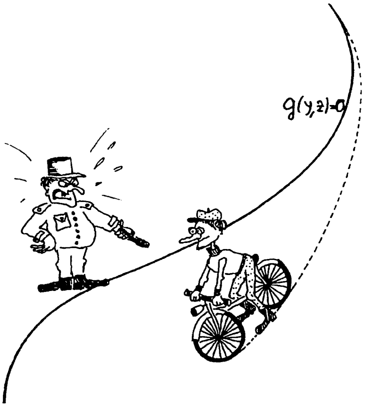
Chapter VII. Differential-Algebraic Equations of Higher Index



(Drawing by K. Wanner)

In the preceding chapter we considered the simplest special case of differential-algebraic equations – the so-called index 1 problem. Many problems of practical interest are, however, of higher index, which makes them more and more difficult for their numerical treatment.

We start by classifying differential-algebraic equations (DAE) by the index (index of nilpotency for linear problems with constant coefficients; differentiation and perturbation index for general nonlinear problems) and present some examples arising in applications (Sect. VII.1). Several different approaches for solving numerically higher index problems are discussed in Sect. VII.2: index reduction by differentiation combined with suitable projections, state space form methods, and treatment as overdetermined or unstructured systems. Sections VII.3 and VII.4 study the convergence properties of multistep methods and Runge-Kutta methods when they are applied directly to index 2 systems. It may happen that the order of convergence is lower than for ordinary differential equations ("order reduction"). The study of conditions which guarantee a certain order is the subject of Sect. VII.5. Half-explicit methods for index 2 problems are especially suited for constrained mechanical systems (Sect. VII.6). A multibody mechanism and its numerical treatment are detailed in Sect. VII.7. Finally, we discuss symplectic methods for constrained Hamiltonian systems (Sect. VII.8), and explain their long-term behaviour by a backward error analysis for differential equations on manifolds.

VII.1 The Index and Various Examples

The most general form of a differential-algebraic system is that of an implicit differential equation

$$F(u',u)=0 \tag{1.1}$$

where F and u have the same dimension. We always assume F to be sufficiently differentiable. A non-autonomous system is brought to the form (1.1) by appending x to the vector u, and by adding the equation $x'=1$.

If $\partial F/\partial u'$ is invertible we can formally solve (1.1) for u' to obtain an ordinary differential equation. In this chapter we are interested in problems (1.1) where $\partial F/\partial u'$ is singular.

Linear Equations with Constant Coefficients

> Uebrigens kann ich die Meinung des Hrn. *Jordan* nicht theilen, dass es ziemlich schwer sei, der *Weierstrass*-schen Analyse zu folgen; sie scheint mir im Gegentheil vollkommen durchsichtig zu sein, ... (L. Kronecker 1874)

The simplest and best understood problems of the form (1.1) are linear differential equations with constant coefficients

$$Bu'+Au=d(x). \tag{1.2}$$

In looking for solutions of the form $e^{\lambda x}u_0$ (if $d(x)\equiv 0$) we are led to consider the "matrix pencil" $A+\lambda B$. When $A+\lambda B$ is singular for all values of λ, then (1.2) has either no solution or infinitely many solutions for a given initial value (Exercise 1). We shall therefore deal only with *regular matrix pencils*, i.e., with problems where the polynomial $\det(A+\lambda B)$ does not vanish identically. The key to the solution of (1.2) is the following simultaneous transformation of A and B to canonical form.

Theorem 1.1 (Weierstrass 1868, Kronecker 1890). *Let $A+\lambda B$ be a regular matrix pencil. Then there exist nonsingular matrices P and Q such that*

$$PAQ=\begin{pmatrix} C & 0\\ 0 & I\end{pmatrix},\qquad PBQ=\begin{pmatrix} I & 0\\ 0 & N\end{pmatrix} \tag{1.3}$$

where $N = \text{blockdiag}\,(N_1, \dots, N_k)$, *each* N_i *is of the form*

$$N_i = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & & & 0 \end{pmatrix}, \qquad \textit{of dimension } m_i\,, \tag{1.4}$$

and C *can be assumed to be in Jordan canonical form.*

Proof (Gantmacher 1954 (Chapter XII), see also Exercises 2 and 3). We fix some c such that $A + cB$ is invertible. If we multiply

$$A + \lambda B = A + cB + (\lambda - c)B$$

by the inverse of $A + cB$ and then transform $(A + cB)^{-1}B$ to Jordan canonical form (Theorem I.12.2) we obtain

$$\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + (\lambda - c)\begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}. \tag{1.5}$$

Here, J_1 contains the Jordan blocks with non-zero eigenvalues, J_2 those with zero eigenvalues (the dimension of J_1 is just the degree of the polynomial $\det(A + \lambda B)$). Consequently, J_1 and $I - cJ_2$ are both invertible and multiplying (1.5) from the left by $\text{blockdiag}\,(J_1^{-1}, (I - cJ_2)^{-1})$ gives

$$\begin{pmatrix} J_1^{-1}(I - cJ_1) & 0 \\ 0 & I \end{pmatrix} + \lambda \begin{pmatrix} I & 0 \\ 0 & (I - cJ_2)^{-1}J_2 \end{pmatrix}.$$

The matrices $J_1^{-1}(I - cJ_1)$ and $(I - cJ_2)^{-1}J_2$ can then be brought to Jordan canonical form. Since all eigenvalues of $(I - cJ_2)^{-1}J_2$ are zero, we obtain the desired decomposition (1.3). □

Theorem 1.1 allows us to solve (1.2) as follows: we premultiply (1.2) by P and use the transformation

$$u = Q\begin{pmatrix} y \\ z \end{pmatrix}, \qquad Pd(x) = \begin{pmatrix} \eta(x) \\ \delta(x) \end{pmatrix}.$$

This decouples the differential-algebraic system (1.2) into

$$y' + Cy = \eta(x), \qquad Nz' + z = \delta(x). \tag{1.6}$$

The equation for y is just an ordinary differential equation. The relation for z decouples again into k subsystems, each of the form (with $m = m_i$)

$$\begin{aligned} z_2' + z_1 &= \delta_1(x) \\ &\vdots \\ z_m' + z_{m-1} &= \delta_{m-1}(x) \\ z_m &= \delta_m(x). \end{aligned} \tag{1.7}$$

Here z_m is determined by the last equation, and the other components are obtained recursively by repeated differentiation. Thus z_1 depends on the $(m-1)$-th derivative of $\delta_m(x)$. Since numerical differentiation is an unstable procedure, the largest m_i appearing in (1.4) is a measure of numerical difficulty for solving problem (1.2). This integer $(\max m_i)$ is called the *index of nilpotency* of the matrix pencil $A+\lambda B$. It does not depend on the particular transformation used to get (1.3) (see Exercise 4).

Linear Equations with Variable Coefficients. In the case, where the matrices A and B in (1.2) depend on x, the study of the solutions is much more complicated. Multiplying the equation by $P(x)$ and substituting $u = Q(x)v$, yields the system

$$PBQv' + (PAQ + PBQ')v = 0, \tag{1.8}$$

which shows that the transformation (1.3) is no longer relevant. With the use of transformations of the form (1.8), Kunkel & Mehrmann (1995) derive a canonical form for linear systems with variable coefficients.

Differentiation Index

A lot of English cars have steering wheels.
(*Fawlty Towers*, Cleese and Booth 1979)

Let us start with the following example:

$$\begin{aligned} y_1' &= 0.7\cdot y_2 + \sin(2.5\cdot z) = f_1(y,z) \\ y_2' &= 1.4\cdot y_1 + \cos(2.5\cdot z) = f_2(y,z) \end{aligned} \tag{1.9a}$$

$$0 = y_1^2 + y_2^2 - 1 = g(y). \tag{1.9b}$$

The "control variable" z in (1.9a) can be interpreted as the position of a "steering wheel" keeping the vector field (y_1', y_2') tangent to the circle $y_1^2 + y_2^2 = 1$, so that condition (1.9b) remains continually satisfied (see Fig. 1.1a). By differentiating (1.9b) and substituting (1.9a) we therefore must have

$$g_y(y)f(y,z) = 0. \tag{1.9c}$$

This defines a "hidden" submanifold of the cylinder, on which all solutions of (1.9a,b) must lie (see Fig. 1.1b). We still do not know how, with increasing x, the variable z changes. This is obtained by differentiating (1.9c) with respect to x: $g_{yy}(f,f) + g_y f_y f + g_y f_z z' = 0$. From this relation we can extract

$$z' = -(g_y f_z)^{-1}\Big(g_{yy}(f,f) + g_y f_y f\Big) \tag{1.9d}$$

if

$$g_y(y)f_z(y,z) \qquad \text{is invertible.} \tag{1.10}$$

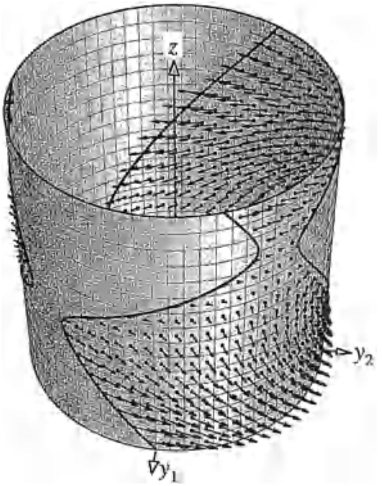


Fig. 1.1a. The vector field (1.9a,d)

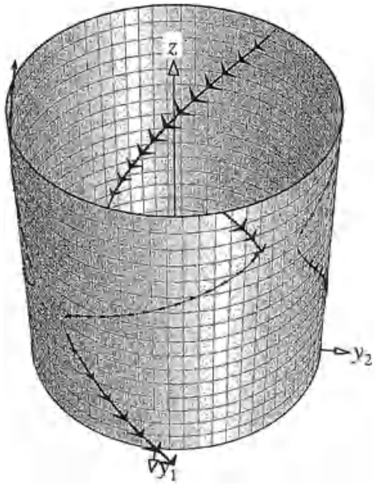


Fig. 1.1b. The hidden submanifold

We have been able to transform the above differential-algebraic equation (1.9a,b) into an ordinary differential system (1.9a,d) by *two analytic differentiations* of the constraint (1.9c). This fact is used for the following definition, which has been developed in several papers (Gear & Petzold 1983, 1984; Gear, Gupta & Leimkuhler 1985, Gear 1990, Campbell & Gear 1995).

Definition 1.2. Equation (1.1) has *differentiation index* $di = m$ if m is the minimal number of analytical differentiations

$$F(u',u)=0,\quad \frac{dF(u',u)}{dx}=0\,,\quad \ldots\,,\quad \frac{d^mF(u',u)}{dx^m}=0 \tag{1.11}$$

such that equations (1.11) allow us to extract by algebraic manipulations an explicit ordinary differential system $u'=\varphi(u)$ (which is called the "*underlying ODE*").

Examples. *Linear Equations with Constant Coefficients.* The following problem

$$\begin{array}{rcrcl} z_2'+z_1=\delta_1 & & z_2''+z_1'=\delta_1' & & \\ z_3'+z_2=\delta_2 & \Rightarrow & z_3'''+z_2''=\delta_2'' & \Rightarrow & z_1'=\delta_1'-\delta_2''+\delta_3''' \\ z_3=\delta_3 & & z_3'''=\delta_3''' & & \end{array} \tag{1.12}$$

can be seen to have differentiation index 3. For linear equations with constant coefficients the differentiation index and the index of nilpotency are therefore the same.

Systems of Index 1. The differential-algebraic systems already seen in Chapter VI

$$y'=f(y,z) \tag{1.13a}$$

$$0=g(y,z) \tag{1.13b}$$

have no z'. We therefore differentiate (1.13b) to obtain

$$z' = -g_z^{-1}(y,z)g_y(y,z)f(y,z) \tag{1.13c}$$

which is possible if g_z is invertible in a neighbourhood of the solution. The problem (1.13a,b), for invertible g_z, is thus of differentiation index 1.

Systems of Index 2. In the system (see example (1.9))

$$\begin{aligned} y' &= f(y,z) && (1.14a)\\ 0 &= g(y), && (1.14b) \end{aligned}$$

where the variable z is absent in the algebraic constraint, we obtain by differentiation of (1.14b) the "hidden constraint"

$$0 = g_y(y)f(y,z). \tag{1.14c}$$

If (1.10) is satisfied in a neighbourhood of the solution, then (1.14a) and (1.14c) constitute an index 1 problem. Differentiation of (1.14c) yields the missing differential equation for z, so that the problem (1.14a,b) is of differentiation index 2. If the initial values satisfy $0 = g(y_0)$ and $0 = g_y(y_0)f(y_0,z_0)$, we call them *consistent.* In this case, and only in this case, the system (1.14a,b) possesses a (locally) unique solution.

System (1.14a,b) is a representative of the larger class of problems of type (1.13a,b) with *singular* g_z. If we assume that g_z has constant rank in a neighbourhood of the solution, we can eliminate certain algebraic variables from $0 = g(y,z)$ until the system is of the form (1.14). This can be done as follows: from the constant rank assumption it follows that either there exists a component of g such that $\partial g_i/\partial z_1 \neq 0$ locally, or $\partial g/\partial z_1$ vanishes identically so that g is already independent of z_1. In the first case we can express z_1 as a function of y and the remaining components of z, and then we can eliminate z_1 from the system. Repeating this procedure with z_2, z_3, etc., will lead to a system of the form (1.14). This transformation does not change the index. Moreover, most numerical methods are invariant under this transformation. Therefore, theoretical work done for systems of the form (1.14) will also be valid for more general problems.

Systems of Index 3. Problems of the form

$$\begin{aligned} y' &= f(y,z) && (1.15a)\\ z' &= k(y,z,u) && (1.15b)\\ 0 &= g(y) && (1.15c) \end{aligned}$$

are of differentiation index 3, if

$$g_y f_z k_u \qquad \text{is invertible} \tag{1.16}$$

in a neighbourhood of the solution. Differentiating (1.15c) twice gives

$$\begin{aligned} 0 &= g_y f && (1.15d)\\ 0 &= g_{yy}(f,f) + g_y f_y f + g_y f_z k. && (1.15e) \end{aligned}$$

Equations (1.15a,b), (1.15e) together with Condition (1.16) are of the index 1 form (1.13a,b). Consistent inital values must satisfy the three conditions (1.15c,d,e).

An extensive study of the solution space of general differential-algebraic systems is done by Griepentrog & März (1986), März (1989, 1990). These authors try to avoid assumptions on the smoothness on the problem as far as possible and replace the above differentiations by a careful study of suitable projections depending only on the first derivatives of F.

Differential Equations on Manifolds

In the language of differentiable manifolds, whose use in DAE theory was urged by Rheinboldt (1984), a constraint (such as $g(y)=0$) represents a manifold, which we denote by

$$\mathcal{M}=\{y\in\mathbb{R}^n \mid g(y)=0\}. \tag{1.17}$$

We assume that $g:\mathbb{R}^n\to\mathbb{R}^m$ (with $m<n$) is a sufficiently differentiable function whose Jacobian $g_y(y)$ has full rank for $y\in\mathcal{M}$. For a fixed $y\in\mathcal{M}$ we denote by

$$T_y\mathcal{M}=\{v\in\mathbb{R}^n \mid g_y(y)v=0\}, \tag{1.18}$$

the tangent space of $\mathcal{M}$ at y. This is a linear space and has the same dimension $n-m$ as the manifold $\mathcal{M}$.

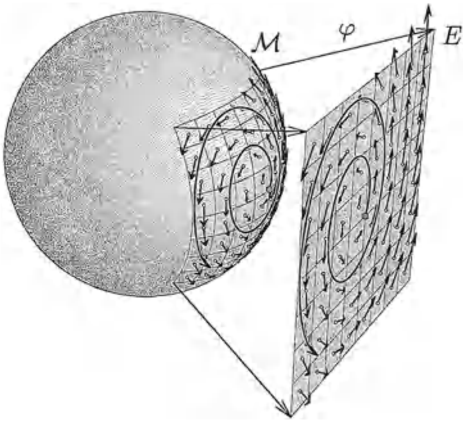


Fig. 1.2. A manifold with a tangent vector field, a chart, and a solution curve

A *vector field on* $\mathcal{M}$ is a mapping $v:\mathcal{M}\to\mathbb{R}^n$, which satisfies $v(y)\in T_y\mathcal{M}$ for all $y\in\mathcal{M}$. For such a vector field we call

$$y'=v(y), \qquad y\in\mathcal{M} \tag{1.19}$$

a *differential equation on the manifold* $\mathcal{M}$. Differentiation on an $(n-m)$-dimensional manifold is described by so-called *charts* $\varphi_i:U_i\to E_i$, where the U_i cover the manifold $\mathcal{M}$ and the E_i are open subsets of $\mathbb{R}^{n-m}$ (Fig. 1.2; see also Lang (1962), Chap. II and Abraham, Marsden & Ratiu (1983), Chap. III). The local theory of ordinary differential equations can be extended to vector fields on manifolds in a straightforward manner:

> Project the vectors $v(y)$ onto E_i via a chart φ_i by multiplying $v(y)$ with the Jacobian of φ_i at y. Then apply standard results to the projected vector field in $\mathbb{R}^{n-m}$, and pull the solution back to $\mathcal{M}$.

(see Fig. 1.2). The local existence of solutions of (1.19) can be shown in this way. The obtained solution is independent of the chosen chart. Where the solution leaves the domain of a chart, the integration must be continued via another one.

Index 2 Problems. Consider the system (1.14a,b) and suppose that (1.10) is satisfied. This condition implies that $g_y(y)$ is of full rank, so that (1.17) is a smooth manifold. Moreover, the Implicit Function Theorem implies that the differentiated constraint (1.14c) can be solved for z (in a neighbourhood of the solution), i.e., there exists a smooth function $h(y)$ such that

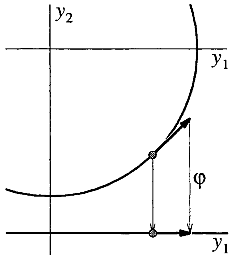
$$g_y(y)f(y,z)=0 \qquad \Longleftrightarrow \qquad z=h(y). \tag{1.20}$$

Inserting this relation into (1.10a) yields

$$y'=f\big(y,h(y)\big), \qquad y\in\mathcal{M} \tag{1.21}$$

which is a differential equation on the manifold (1.17), because $f(y,h(y))\in T_y\mathcal{M}$ by (1.20). The differential equation (1.21) is equivalent to (1.14a,b).

Example. The manifold $\mathcal{M}$ for problem (1.9) is one-dimensional (circle). In points, where $y_1\neq\pm1$, we can solve (1.9b) to obtain locally $y_2=\pm\sqrt{1-y_1^2}$. The map $(y_1,y_2)\mapsto y_1$ consitutes a chart φ, which is bijective in a neighbourhood of the considered point. Inserting z from (1.9c) and the above y_2 into (1.9a), yields an equation $y_1'=G(y_1)$, which is the projected vector field in $\mathbb{R}^1$.



Index 3 Problems. For the system (1.15a,b,c) the solutions lie on the manifold

$$\mathcal{M}=\{(y,z)\mid g(y)=0,\;\; g_y(y)f(y,z)=0\}. \tag{1.22}$$

The assumption (1.16) implies that $g_y(y)$ and $g_y(y)f_z(y,z)$ have full rank, so that $\mathcal{M}$ is a manifold. Its tangent space at (y,z) is

$$\begin{aligned} T_{(y,z)}\mathcal{M}=\{(v,w)\mid g_y(y)v=0,\;\; & g_{yy}(y)\big(f(y,z),v\big)\\ & +g_y(y)\big(f_y(y,z)v+f_z(y,z)w\big)=0\}. \end{aligned} \tag{1.23}$$

Solving Eq. (1.15e) for u and inserting the result into (1.15b) yields a differential equation on the manifold $\mathcal{M}$. Because of (1.15d,e), the obtained vector field lies in the tangent space $T_{(y,z)}\mathcal{M}$ for all $(y,z)\in\mathcal{M}$.

The Perturbation Index

Now fills thy sleep with perturbations.
(The *Ghost of Anne* in Shakespeare's *Richard III*, act V, sc. III)

A second concept of index, due to HLR89 [1], interprets the index as a measure of sensitivity of the solutions with respect to perturbations of the given problem.

Definition 1.3. Equation (1.1) has *perturbation index* $pi = m$ along a solution $u(x)$ on $[0, \overline{x}]$, if m is the smallest integer such that, for all functions $\widehat{u}(x)$ having a defect

$$F(\widehat{u}', \widehat{u}) = \delta(x), \tag{1.24}$$

there exists on $[0, \overline{x}]$ an estimate

$$\|\widehat{u}(x) - u(x)\| \le C\Big(\|\widehat{u}(0) - u(0)\| + \max_{0\le\xi\le x} \|\delta(\xi)\| + \ldots + \max_{0\le\xi\le x} \|\delta^{(m-1)}(\xi)\|\Big) \tag{1.25}$$

whenever the expression on the right-hand side is sufficiently small.

Remark. We deliberately do not write "Let $\widehat{u}(x)$ be the solution of $F(\widehat{u}', \widehat{u}) = \delta(x)$ ..." in this definition, because the existence of such a solution $\widehat{u}(x)$ for an arbitrarily given $\delta(x)$ is not assured. We start with $\widehat{u}$ and then compute δ as defect of (1.1).

Systems of Index 1. For the computation of the perturbation index of (1.13a,b) we consider the perturbed system

$$\widehat{y}' = f(\widehat{y}, \widehat{z}) + \delta_1(x) \tag{1.26a}$$

$$0 = g(\widehat{y}, \widehat{z}) + \delta_2(x). \tag{1.26b}$$

The essential observation is that the difference $\widehat{z} - z$ can be estimated with the help of the Implicit Function Theorem, without any differentiation of the equation. Since g_z is invertible by hypothesis, this theorem gives from (1.26b) compared to (1.13b)

$$\|\widehat{z}(x) - z(x)\| \le C_1\big(\|\widehat{y}(x) - y(x)\| + \|\delta_2(x)\|\big) \tag{1.27}$$

as long as the right-hand side of (1.27) is sufficiently small. We now subtract (1.26a) from (1.13a), integrate from 0 to x, use a Lipschitz condition for f and the above estimate for $\widehat{z}(x) - z(x)$. This gives for $e(x) = \|\widehat{y}(x) - y(x)\|$:

$$e(x) \le e(0) + C_2 \int_0^x e(t)dt + C_3 \int_0^x \|\delta_2(t)\|dt + \Big\| \int_0^x \delta_1(t)dt \Big\|.$$

In this estimate the norm is *inside* the integral for δ_2, but *outside* the integral for δ_1. This is due to the fact that perturbations of the algebraic equation (1.13b) are more

[1] The "Lecture Notes" of Hairer, Lubich & Roche (1989) will be cited frequently in the subsequent sections. Reference to this publication will henceforth be denoted by HLR89.

serious than perturbations of the differential equation (1.13a). We finally apply Gronwall's Lemma (Exercise I.10.2) to obtain on a bounded interval $[0,\overline{x}]$

$$\begin{aligned}\|\widehat{y}(x)-y(x)\| &\le C_4\Big(\|\widehat{y}(0)-y(0)\|+\int_0^x\|\delta_2(t)\|dt+\max_{0\le\xi\le x}\Big\|\int_0^\xi\delta_1(t)dt\Big\|\Big)\\ &\le C_5\Big(\|\widehat{y}(0)-y(0)\|+\max_{0\le\xi\le x}\|\delta_2(\xi)\|+\max_{0\le\xi\le x}\|\delta_1(\xi)\|\Big).\end{aligned}$$

This inequality, together with (1.27), shows that the perturbation index of the problem is 1.

Systems of Index 2. We consider the following perturbation of system (1.14a,b)

$$\widehat{y}'=f(\widehat{y},\widehat{z})+\delta(x) \qquad (1.28a)$$

$$0=g(\widehat{y})+\theta(x). \qquad (1.28b)$$

Differentiation of (1.28b) gives

$$0=g_y(\widehat{y})f(\widehat{y},\widehat{z})+g_y(\widehat{y})\delta(x)+\theta'(x). \qquad (1.29)$$

Under the assumption (1.10) we can use the estimates of the index 1 case (with $\delta_2(x)$ replaced by $g_y(\widehat{y}(x))\delta(x)+\theta'(x)$) to obtain

$$\begin{aligned}\|\widehat{y}(x)-y(x)\| &\le C\Big(\|\widehat{y}(0)-y(0)\|+\int_0^x\big(\|\delta(\xi)\|+\|\theta'(\xi)\|\big)d\xi\Big)\\ \|\widehat{z}(x)-z(x)\| &\le C\Big(\|\widehat{y}(0)-y(0)\|+\max_{0\le\xi\le x}\|\delta(\xi)\|+\max_{0\le\xi\le x}\|\theta'(\xi)\|\Big).\end{aligned} \qquad (1.30)$$

Since these estimates depend on the first derivative of θ, the perturbation index of this problem is 2. A sharper estimate for the y-component is given in Exercise 6.

Example. Fig. 1.3 presents an illustration for the index 2 problem (1.9a,b). Small perturbations of $g(y)$, once discontinuous in the first derivative (left), the other of oscillatory type (right), results in discontinuities or violent oscillations of z, respectively.

The above examples might give the impression that the differentiation index and the perturbation index are always equal. The following counter-examples show that this is not true.

Counterexamples. The first counterexample of type $M(y)y'=f(y)$ is given by Lubich (1989):

$$\begin{aligned} y_1'-y_3y_2'+y_2y_3'&=0 \qquad & \widehat{y}_1'-\widehat{y}_3\widehat{y}_2'+\widehat{y}_2\widehat{y}_3'&=0\\ y_2&=0 & \widehat{y}_2&=\varepsilon\sin\omega x \qquad (1.31)\\ y_3&=0 & \widehat{y}_3&=\varepsilon\cos\omega x\end{aligned}$$

with $y_i(0)=0$ $(i=1,2,3)$. Inserting $\widehat{y}_2=\varepsilon\sin\omega x$ and $\widehat{y}_3=\varepsilon\cos\omega x$ into the first equation gives $\widehat{y}_1'=\varepsilon^2\omega$ which makes, for ε fixed and $\omega\to\infty$, an estimate

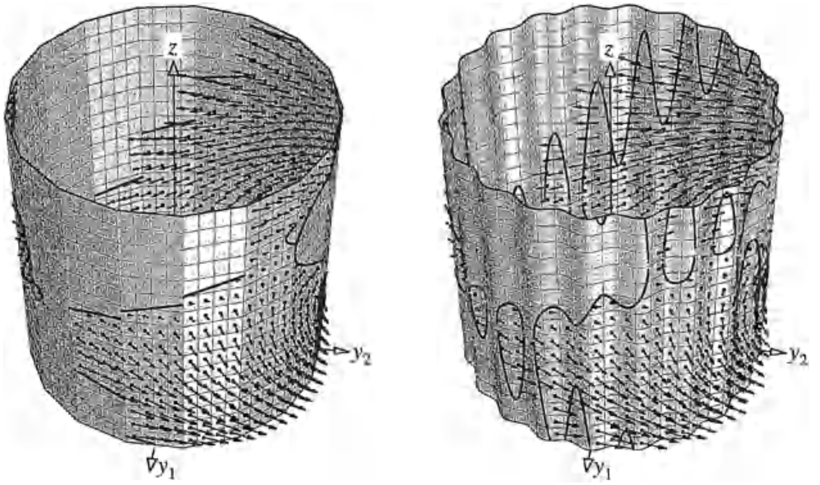


Fig. 1.3. Perturbations of an index 2 problem

(1.25) with $m = 1$ impossible. However, for $m = 2$ the estimate (1.25) is clearly satisfied. This problem, which is obviously of differentiation index 1, is thus of perturbation index 2.

It was believed for some time (see the first edition, p. 479), that the differentiation and perturbation indices can differ at most by 1. The following example, due to Campbell & Gear (1995), was therefore a big surprise:

$$y_m N y' + y = 0, \tag{1.32}$$

where N is a $m \times m$ upper triangular nilpotent Jordan block. Since the last row of N is zero, we have $y_m = 0$, and the differentiation index is 1. On the other hand, adding a perturbation makes y_m different from zero. This is the reason why the perturbation index of (1.32) is m.

Control Problems

Many problems of control theory lead to ordinary differential equations of the form $y' = f(y, u)$, where u represents a set of controls. Similar as in example (1.9) above, these controls must be applied so that the solution satisfies some constraints $0 = g(y, u)$. For numerical examples of such control problems we refer to Brenan (1983) (space shuttle simulation) and Brenan, Campbell & Petzold (1989).

Optimal Control Problems are differential equations $y' = f(y, u)$ formulated in such a way that the control $u(x)$ has to minimize some cost functional. The Euler–Lagrange equation then often becomes a differential-algebraic system (Pontryagin, Boltyanskij, Gamkrelidze & Mishchenko 1961, Athans & Falb 1966, Campbell 1982). We demonstrate this on the problem

$$y' = f(y, u), \qquad y(0) = y_0 \tag{1.33a}$$

with cost functional

$$J(u)=\int_0^1 \varphi\big(y(x),u(x)\big)\,dx. \tag{1.33b}$$

For a given function $u(x)$ the solution $y(x)$ is determined by (1.33a). In order to find conditions for $u(x)$ which minimize $J(u)$ of (1.33b), we consider the perturbed control $u(x)+\varepsilon\delta u(x)$ where $\delta u(x)$ is an arbitrary function and ε a small number. To this control there corresponds a solution $y(x)+\varepsilon\delta y(x)+\mathcal{O}(\varepsilon^2)$ of (1.33a); hence (by comparing powers of ε)

$$\delta y'(x)=f_y(x)\delta y(x)+f_u(x)\delta u(x), \qquad \delta y(0)=0, \tag{1.34}$$

where, as usual, $f_y(x)=f_y(y(x),u(x))$, etc. Linearization of (1.33b) shows that

$$J(u+\varepsilon\delta u)-J(u)=\varepsilon\int_0^1\Big(\varphi_y(x)\delta y(x)+\varphi_u(x)\delta u(x)\Big)dx+\mathcal{O}(\varepsilon^2)$$

so that

$$\int_0^1\Big(\varphi_y(x)\delta y(x)+\varphi_u(x)\delta u(x)\Big)dx=0 \tag{1.35}$$

is a necessary condition for $u(x)$ to be an optimal solution of our problem. In order to express δy in terms of δu in (1.35), we introduce the adjoint differential equation

$$v'=-f_y(x)^Tv-\varphi_y(x)^T, \qquad v(1)=0 \tag{1.36}$$

with inhomogeneity $\varphi_y(x)^T$. Hence we have (see Exercise 7)

$$\int_0^1 \varphi_y(x)\delta y(x)dx=\int_0^1 v^T(x)f_u(x)\delta u(x)dx. \tag{1.37}$$

Inserted into (1.35) this gives the necessary condition

$$\int_0^1\Big(v^T(x)f_u(x)+\varphi_u(x)\Big)\delta u(x)dx=0. \tag{1.38}$$

Since this relation has to be satisfied for all δu we obtain the necessary relation $v^T(x)f_u(x)+\varphi_u(x)=0$ by the so-called "fundamental lemma of variational calculus".

In summary, we have proved that a solution of the above optimal control problem has to satisfy the system

$$\begin{aligned}
y'&=f(y,u), & y(0)&=y_0\\
v'&=-f_y(y,u)^Tv-\varphi_y(y,u)^T, & v(1)&=0\\
0&=v^Tf_u(y,u)+\varphi_u(y,u).
\end{aligned} \tag{1.39}$$

This is a boundary value differential-algebraic problem. It can also be obtained directly from the Pontryagin minimum principle (see Pontryagin et al. 1961, Athans & Falb 1966).

Differentiation of the algebraic relation in (1.39) shows that the system (1.39) has index 1 if the matrix

$$\sum_{i=1}^{n} v_i \frac{\partial^2 f_i}{\partial u^2}(y,u) + \frac{\partial^2 \varphi}{\partial u^2}(y,u) \tag{1.40}$$

is invertible along the solution. A situation where the system (1.39) has index 3 is presented in Exercise 8. An index 5 problem of this type is given in "Example 3.1" of Clark (1988). Other control problems with a large index are discussed in Campbell (1995).

Mechanical Systems

> ... berechnen wir T, V, L. Mehr brauchen wir von der Geometrie und Mechanik unseres Systems nicht zu wissen. Alles übrige besorgt ohne unser Zutun der Formalismus von LAGRANGE.
>
> (Sommerfeld 1942, §35)

An interesting class of differential-algebraic systems appears in mechanical modeling of constrained systems. A choice method for deriving the equations of motion of mechanical systems is the Lagrange-Hamilton principle, whose long history goes back to merely theological ideas of Leibniz and Maupertuis. Let $q_1, \ldots, q_n$ be position coordinates of a system and $u_i = \dot{q}_i$ the velocities. Suppose a function $L(q, \dot{q})$ is given; then the Euler equations of the variational problem

$$\int_{t_1}^{t_2} L(q, \dot{q})dt = \min! \tag{1.41}$$

are given by

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0, \qquad k = 1, \ldots, n \tag{1.42}$$

or

$$\sum_{\ell=1}^{n} L_{\dot{q}_k \dot{q}_\ell} \ddot{q}_\ell = L_{q_k} - \sum_{\ell=1}^{n} L_{\dot{q}_k q_\ell} \dot{q}_\ell . \tag{1.43}$$

The great discovery of Lagrange (1788) is that for $L = T - U$, where T is the *kinetic energy* and U the *potential energy*, the differential equations (1.43) describe the movement of the corresponding "conservative system". For a proof and various generalizations, consult any book on mechanics e.g., Sommerfeld (1942), vol. I, §§ 33–37, or Arnol'd (1979), part II.

Example 1. For the mathematical pendulum of length ℓ we choose as position coordinate the angle $\theta = q_1$ such that $T = m\ell^2\dot{\theta}^2/2$ and $U = -\ell mg \cos\theta$. Then (1.43) becomes $\ell\ddot{\theta} = -g \sin\theta$, the well-known pendulum equation.

Movement with Constraints. Suppose now that we have some constraints $g_1(q) = 0, \ldots, g_m(q) = 0$ on our movement. Another great idea of Lagrange is to vary the "Lagrange function" as follows in this case

$$L = T - U - \lambda_1 g_1(q) - \ldots - \lambda_m g_m(q) \tag{1.44}$$

where the "Lagrange multipliers" λ_i are appended to the coordinates. The important fact is that, since L is independent of $\dot{\lambda}_i$, the equation (1.43), for the derivatives with respect to λ_k, just becomes $0 = g_k(q)$, the desired side conditions.

Example 2. We now describe the pendulum in Cartesian coordinates x, y with constraint $x^2 + y^2 - \ell^2 = 0$. This gives for (1.44)

$$L = \frac{m}{2}(\dot{x}^2 + \dot{y}^2) - mgy - \lambda(x^2 + y^2 - \ell^2)$$

and (1.43) becomes

$$\begin{aligned} m\ddot{x} &= -2x\lambda \\ m\ddot{y} &= -mg - 2y\lambda \\ 0 &= x^2 + y^2 - \ell^2. \end{aligned} \tag{1.45}$$

In this example the physical meaning of λ is the tension in the rod which maintains the mass point on the desired orbit.

The general form of a constrained mechanical system (1.43) is in vector notation (after replacing dots by primes)

$$q' = u \tag{1.46a}$$
$$M(q)u' = f(q,u) - G^T(q)\lambda \tag{1.46b}$$
$$0 = g(q) \tag{1.46c}$$

where $M(q) = T_{\dot{q}\dot{q}} = T_{uu}$ is a positive definite matrix, $G(q) = \partial g / \partial q$ and $q = (q_1, \ldots, q_n)^T$, $u = (\dot{q}_1, \ldots, \dot{q}_n)^T$, $\lambda = (\lambda_1, \ldots, \lambda_m)^T$. Various formulations are possible for such a problem, each of which leads to a different numerical approach.

Index 3 Formulation (position level, descriptor form). If we formally multiply (1.46b) by M^{-1}, the system (1.46) becomes of the form (1.15) with (q, u, λ) in the roles of (y, z, u). The condition (1.16), written out for (1.46), is

$$GM^{-1}G^T \qquad \text{is invertible} . \tag{1.47}$$

This is satisfied, if the constraints (1.46c) are independent, i.e., if the rows of the matrix G are linearly independent. Under this assumption, the system (1.46a,b,c) is thus an index 3 problem.

Index 2 Formulation (velocity level). Differentiation of (1.46c) gives

$$0 = G(q)u. \tag{1.46d}$$

If we replace (1.46c) by (1.46d) we obtain a system of the form (1.14a,b) with (q, u) in the role of y and λ that of z. One verifies that Condition (1.10) is equivalent to (1.47), so that (1.46a,b,d) represents a problem of index 2.

Index 1 Formulation (acceleration level). If we differentiate twice the constraint (1.46c), the resulting equation together with (1.46b) yield

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} u' \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q,u) \\ -g_{qq}(q)(u,u) \end{pmatrix}. \tag{1.46e}$$

This allows us to express u' and λ as functions of q, u, provided that the matrix in (1.46e) is invertible. Hence, (1.46a,e) consitute an index 1 problem. The assumption on the matrix in Eq. (1.46e) is weaker than (1.47), because $M(q)$ need not be regular.

All these formulations are mathematically equivalent, if the initial values are consistent, i.e., if (1.46c,d,e) are satisfied. However, if for example the index 2 system (1.46a,b,d) is integrated numerically, the constraints of the original problem will no longer be exactly satisfied. For this reason Gear, Gupta & Leimkuhler (1985) introduced another index 2 formulation ("... an interesting way of reducing the problem to index two and adding variables so that the constraint continues to be satisfied".).

GGL Formulation. The idea is to add the constraint (1.46d) to the original system and to introduce an additional Lagrange multiplier μ in (1.46a). For the sake of symmetry we also multiply (1.46a) by $M(q)$, so that the whole system becomes

$$\begin{aligned} M(q)q' &= M(q)u - G^T(q)\mu \\ M(q)u' &= f(q,u) - G^T(q)\lambda \\ 0 &= g(q) \\ 0 &= G(q)u. \end{aligned} \tag{1.48}$$

Here the differential variables are (q, u) and the algebraic variables are (μ, λ). System (1.48) is of the form (1.14a,b) and the index 2 assumption is satisfied if (1.47) holds.

A concrete mechanical system is described in detail, together with numerical results for all the above formulations, in Sect. VII.7.

Exercises

1. Prove that the initial value problem

$$Bu' + Au = 0, \qquad u(0) = u_0 \tag{1.49}$$

has a unique solution if and only if the matrix pencil $A + \lambda B$ is regular.

Hint for the "only if" part. If n is the dimension of u, choose arbitrarily $n+1$ distinct λ_i and vectors $v_i \neq 0$ satisfying $(A+\lambda_i B)v_i = 0$. Then take a linear combination, such that $\sum \alpha_i v_i = 0$, but $\sum \alpha_i e^{\lambda_i x} v_i \not\equiv 0$.

2. (Stewart 1972). Let $A+\lambda B$ be a regular matrix pencil. Show that there exist unitary matrices Q and Z such that

$$QAZ = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \qquad QBZ = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \tag{1.50}$$

are both triangular. Further, the diagonal elements of A_{22} and B_{11} are all 1, those of B_{22} are all 0.

Hint (Compare with the Schur decomposition of Theorem I.12.1). Let λ_1 be a zero of $\det(A+\lambda B)$ and $v_1 \neq 0$ be such that $(A+\lambda_1 B)v_1 = 0$. Verify that $Bv_1 \neq 0$ and that

$$AZ_1 = Q_1 \begin{pmatrix} -\lambda_1 & * \\ 0 & \widetilde{A} \end{pmatrix}, \qquad BZ_1 = Q_1 \begin{pmatrix} 1 & * \\ 0 & \widetilde{B} \end{pmatrix}$$

where Q_1, Z_1 are unitary matrices whose first columns are Bv_1 and v_1, respectively. The matrix pencil $\widetilde{A}+\lambda\widetilde{B}$ is again regular and this procedure can be continued until $\det(\widetilde{A}+\lambda\widetilde{B}) = \mathit{Const}$ which implies that $\det\widetilde{B} = 0$. In this case we take a vector $v_2 \neq 0$ such that $\widetilde{B}v_2 = 0$ and transform $\widetilde{A}+\lambda\widetilde{B}$ with unitary matrices Q_2, Z_2, whose first columns are $\widetilde{A}v_2$ and v_2, respectively. For a practical computation of the decomposition (1.50) see Golub & Van Loan (1989), Sect. 7.7.

3. Under the assumptions of Exercise 2 show that there exist matrices S and T such that

$$\begin{pmatrix} I & S \\ 0 & I \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\begin{pmatrix} I & T \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix},$$

$$\begin{pmatrix} I & S \\ 0 & I \end{pmatrix}\begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}\begin{pmatrix} I & T \\ 0 & I \end{pmatrix} = \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix}.$$

Hint. These matrices have to satisfy

$$A_{11}T + A_{12} + SA_{22} = 0 \tag{1.51a}$$

$$B_{11}T + B_{12} + SB_{22} = 0 \tag{1.51b}$$

and can be computed as follows: the first column of T is obtained from (1.51b) because B_{11} is invertible and the first column of SB_{22} vanishes; then the first column of S is given by (1.51a) because A_{22} is invertible; the second column of SB_{22} is then known and we can compute the second column of T from (1.51b), etc.

4. Prove that the index of nilpotency of a regular matrix pencil $A+\lambda B$ does not depend on the choice of P and Q in (1.3).

Hint. Consider two different decompositions of the form (1.3) and denote the matrices which appear by C_1, N_1 and C_2, N_2, respectively. Show the existence of a regular matrix T such that $N_2 = T^{-1} N_1 T$.

5. Prove that the system (VI.3.4a,b) has index 2 (it is of the form (1.14a,b) and satisfies (1.10)). The full system (VI.3.4) has perturbation index k.

6. (Arnold 1993). Consider the index 2 problem (1.14) and its perturbation (1.28). Prove that the difference $\Delta y(x) = \widehat{y}(x) - y(x)$ satisfies

$$\|\Delta y(x)\| \le C\Big(\|\Delta y(0)\| + \max_{0\le\xi\le x}\Big(\Big\|\int_0^\xi P(t)\delta(t)\,dt\Big\| \\ + \|\theta(\xi)\| + \big(\|\delta(\xi)\| + \|\theta'(x)\|\big)^2\Big)\Big)$$

with the projector $P(t) = I - \big(f_z(g_y f_z)^{-1} g_y\big)\big(y(t), z(t)\big)$, provided that the right hand side is sufficiently small.

Hint. Linearize Eq. (1.29) around (y, z), extract $\widehat{z} - z$, and insert it into the difference of (1.28a) and (1.14a). The term $\big(f_z(g_y f_z)^{-1}\big)\big(y(x), z(x)\big)\,\theta'(x)$ can be replaced by $\frac{d}{dx}\big(f_z(g_y f_z)^{-1}\big(y(x), z(x)\big)\theta(x)\big) + \mathcal{O}(\|\theta(x)\|)$ before integration.

7. For the linear initial value problem

$$y' = A(x)y + f(x), \qquad y(0) = 0$$

consider the *adjoint* problem

$$v' = -A(x)^T v - g(x), \qquad v(1) = 0.$$

Prove that $\displaystyle\int_0^1 g(x)^T y(x)dx = \int_0^1 v(x)^T f(x)dx.$

8. Consider a linear optimal control problem with quadratic cost functional

$$y' = Ay + Bu + f(x), \qquad y(0) = y_0$$
$$J(u) = \frac{1}{2}\int_0^1 \Big(y(x)^T C y(x) + u(x)^T D u(x)\Big)dx,$$

where C and D are assumed to be positive semi-definite.

a) Prove that $J(u)$ is minimal if and only if

$$\begin{aligned} y' &= Ay + Bu + f(x), & y(0) &= y_0 \\ v' &= -A^T v - Cy, & v(1) &= 0 \\ 0 &= B^T v + Du. \end{aligned} \tag{1.52}$$

b) If D is positive definite, then (1.52) has index 1.

c) If $D = 0$ and $B^T C B$ is positive definite, then (1.52) has index 3.

VII.2 Index Reduction Methods

We have seen in Sect. VI.1 that the numerical treatment of problems of index 1, which are either in the half-explicit form (1.13) or in the form $Mu' = \varphi(u)$, is not much more difficult than that of ordinary differential equations. For higher index problems the situation changes completely. This section is devoted to the study of several approaches that are all based on the idea of modifying the problem in such a way that the index is reduced.

Index Reduction by Differentiation

The most apparent way of reducing the index is to differentiate repeatedly the algebraic constraints (see Definition 1.2). In general, it is recommended to differentiate until having obtained an index 1 problem. For example, the index 2 problem (1.14a,b) is replaced by (1.14a,c), or the constrained mechanical system (1.46a,b,c) by (1.46a,b,e). The resulting problem is then solved by the methods of Chapter VI.

We illustrate this approach at the "pendulum example"

$$x' = u, \qquad u' = -x\lambda \tag{2.1a}$$

$$y' = v, \qquad v' = -1 - y\lambda \tag{2.1b}$$

$$0 = x^2 + y^2 - 1. \tag{2.1c}$$

In this form it has index 3. Differentiating the algebraic constraint twice yields

$$0 = xu + yv, \tag{2.2}$$

$$0 = -\lambda(x^2 + y^2) - y + u^2 + v^2. \tag{2.3}$$

Equations (2.1a,b) together with (2.3) represent an index 1 problem. We can extract λ from (2.3) and insert it into (2.1a,b) to get a differential equation for x, y, u, v, which can be solved by standard methods.

Drift-off Phenomenon. As an example we apply the code DOPRI5 to the index 1 problem (2.1a,b), (2.3) with initial values $x_0 = 1$, $y_0 = 0$, $u_0 = 0$, $v_0 = 0$. We are interested, how well the constraints (2.1c) and (2.2) are preserved by the numerical solution. The result presented in Fig. 2.1 shows that the error in the constraint (2.2) grows linearly, that in (2.1c) grows even quadratically. This phenomenon is explained as follows:

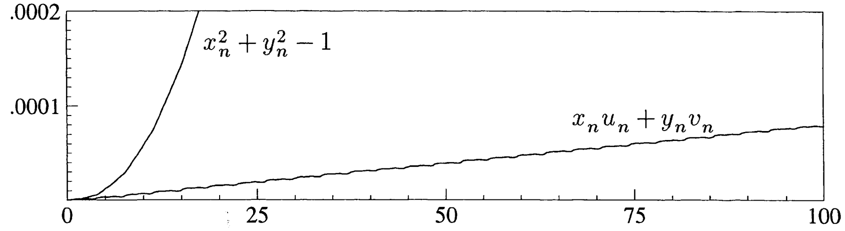


Fig. 2.1. Error in the constraints for DOPRI5 ($Atol = Rtol = 10^{-6}$)

Consider a constrained mechanical system (see (1.46))

$$\begin{aligned} q' &= u && \text{(2.4a)}\\ M(q)u' &= f(q,u) - G^T(q)\lambda && \text{(2.4b)}\\ 0 &= g(q). && \text{(2.4c)} \end{aligned}$$

Differentiating (2.4c) twice we get

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix}\begin{pmatrix} u' \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q,u) \\ -g_{qq}(q)(u,u) \end{pmatrix} \tag{2.5}$$

which, together with (2.4a), is the corresponding index 1 problem. The important observation is now that the index 1 problem possesses a solution for arbitrary initial values q_0 and u_0. Due to the fact that the second derivative of $g(q(t))$ vanishes (this is a consequence of the lower relation of (2.5)), the solution of the index 1 problem satisfies

$$\begin{aligned} g(q(t)) &= g(q_0) + (t-t_0)G(q_0)u_0, && \text{(2.6a)}\\ G(q(t))u(t) &= G(q_0)u_0. && \text{(2.6b)} \end{aligned}$$

Theorem 2.1. *If we apply a pth order numerical method to the index 1 problem (2.4a), (2.5) with consistent initial values at $t_0 = 0$, then the numerical solution (q_n, u_n) at time t_n satisfies (for $t_n - t_0 \le Const$)*

$$\|g(q_n)\| \le h^p(At_n + Bt_n^2), \qquad \|G(q_n)u_n\| \le h^p C t_n.$$

The value h represents the maximal step size used.

Proof. Denote by $q(t, t_0, q_0, u_0)$ the solution of the index 1 problem with initial value (q_0, u_0) at $t = t_0$. Since the local error $q_{j+1} - q(t_{j+1}, t_j, q_j, u_j)$ is of size $\mathcal{O}(h_j^{p+1})$ (and similarly for the u-component), it follows from (2.6a) that

$$\big\|g\big(q(t_n, t_{j+1}, q_{j+1}, u_{j+1})\big) - g\big(q(t_n, t_j, q_j, u_j)\big)\big\| \le h_j^{p+1}\big(A + 2B(t_n - t_{j+1})\big).$$

Adding up these inequalities from $j = 0$ to $j = n-1$ gives the desired bound for $g(q_n)$, because the initial values are consistent, i.e., $g(q(t_n, t_0, q_0, u_0)) = 0$. The second estimate of Theorem 2.1 is proved in the same way. □

Baumgarte Stabilization. The historically first remedy for this drift-off is due to Baumgarte (1972). Instead of replacing the constraint (2.4c) by its second time derivative, he proposes to replace (2.4c) by the linear combination

$$0 = \ddot{g} + 2\alpha\dot{g} + \beta^2 g, \tag{2.7}$$

where $\dot{g}$, $\ddot{g}$ are time derivatives of (2.4c), i.e.,

$$g = g(q), \qquad \dot{g} = G(q)u, \qquad \ddot{g} = g_{qq}(q)(u,u) + G(q)\big(f(q,u) - G^T(q)\lambda\big).$$

Eq. (2.7) together with (2.4b) determines u' and λ as functions of (q,u), and the resulting differential equation can be solved numerically. The idea is now to choose the free parameters α and β in such a way that (2.7) is an asymptotically stable differential equation, e.g., $\beta = \alpha$ and $\alpha > 0$. Consequently, the functions $g(q(t))$ and $G(q(t))u(t)$ are exponentially decreasing, in contrast to (2.6). The difficulty of this approach lies in a good choice of α. For small values of α the damping will not be sufficiently strong, whereas for large α the resulting differential equation becomes stiff and explicit methods are no longer efficient. A careful investigation on the choice of α can be found in Ascher, Chin & Reich (1994).

Stabilization by Projection

We shall now analyze another possibility for avoiding the instability of the preceding example, namely the repeated projection of the numerical solution onto the solution manifold.

Index 2 Problems. Consider the system (1.14a,b). Suppose that (y_{n-1}, z_{n-1}) is an approximation to the solution at time t_{n-1} which satisfies $g(y_{n-1}) = 0$ and $g_y(y_{n-1})f(y_{n-1}, z_{n-1}) = 0$. Applying a numerical one-step method (state space form method of Sect. VI.1) with these values to the index 1 system (1.14a,c) yields an approximation $\widetilde{y}_n, \widetilde{z}_n$ that, in general, does not satisfy the constraint (1.14b). A natural way of projecting the approximation $\widetilde{y}_n$ to the solution manifold $\mathcal{M}$ of Eq. (1.17) is along the image of f_z (see also the projected Runge-Kutta methods of Sect. VII.4). We therefore define y_n as the solution of

$$y - \widetilde{y}_n = f_z(\widetilde{y}_n, \widetilde{z}_n)\mu, \qquad g(y) = 0, \tag{2.8}$$

and then we adjust z_n by solving the equation $g_y(y_n)f(y_n, z_n) = 0$. Applying simplified Newton iterations to the nonlinear system (2.8) requires the decomposition of the matrix

$$\begin{pmatrix} I & f_z(\widetilde{y}_n, \widetilde{z}_n) \\ g_y(\widetilde{y}_n) & 0 \end{pmatrix}. \tag{2.9}$$

Block elimination shows that the invertibility of (2.9) is a consequence of (1.10), and that only the matrix $g_y f_z$ has to be decomposed. Such a decomposition is usually already available from the application of the numerical method, so that the projection (2.8) is very cheap.

It is now natural to ask, whether this projection procedure can distroy the convergence properties of the underlying method. For a pth order one-step method the local error is of size $\mathcal{O}(h^{p+1})$. Since the solution of (1.14a,c) passing through (y_{n-1}, z_{n-1}) satisfies $g(y(t)) = 0$, it holds $g(\widetilde{y}_n) = \mathcal{O}(h^{p+1})$. Hence, the solution of (2.8) satisfies $\mu = \mathcal{O}(h^{p+1})$, $y_n - \widetilde{y}_n = \mathcal{O}(h^{p+1})$, and $z_n - \widetilde{z}_n = \mathcal{O}(h^{p+1})$. By the Implicit Function Theorem this solution depends smoothly on $(\widetilde{y}_n, \widetilde{z}_n)$, so that the mapping $(y_{n-1}, z_{n-1}) \mapsto (y_n, z_n)$ represents a pth order one-step method for (1.14a,c). Convergence of order p thus follows from the standard theory (see Sects. VI.1 and II.3). This proof also applies to multistep methods.

Constrained Mechanical Systems. For the index 3 system (2.4a,b,c) the situation is slightly more complex. We assume consistent values $(q_{n-1}, u_{n-1}, \lambda_{n-1})$ at time t_{n-1} and apply a one-step method to the index 1 system (2.4a), (2.5) to obtain $(\widetilde{q}_n, \widetilde{u}_n)$. Since the position constraint (2.4c) only depends on q, the projections for q and u can be done sequentially.

Projection on Position Constraint. We define q_n as solution of the nonlinear system

$$\begin{aligned} M(\widetilde{q}_n)(q_n - \widetilde{q}_n) + G^T(\widetilde{q}_n)\mu &= 0 \\ g(q_n) &= 0. \end{aligned} \tag{2.10}$$

Projection on Velocity Constraint. With the value q_n obtained from the above projection we let u_n be the solution of

$$\begin{aligned} M(q_n)(u_n - \widetilde{u}_n) + G^T(q_n)\mu &= 0 \\ G(q_n)u_n \qquad\qquad &= 0. \end{aligned} \tag{2.11}$$

Lubich (1991) introduced this kind of projection, because “it is invariant under affine transformations of coordinates”. We remark that the system (2.11) is linear, whereas (2.10) is nonlinear and has to be solved by (simplified) Newton iterations. The index 3 assumption that the matrix in Eq. (2.5) is invertible, implies the existence of the projected values q_n and u_n (at least for sufficiently small step size). It is possible to alter slightly the arguments of M and G^T in the upper lines of (2.10) and (2.11) or to solve the system (2.11) iteratively, if this is computationally advantageous. Convergence of this method is proved in the same way as in the index 2 case.

Velocity Stabilization. It can be seen from (2.6) that errors in the velocity constraint $G(q)u = 0$ are more critical for the numerical solution than errors in the position constraint $g(q) = 0$. It is therefore interesting to study the method, where the numerical solution is projected only to the velocity constraint. Alishenas & Ólafsson (1994) come to the conclusion that “*velocity projection* is the most efficient projection with regard to improvement of the numerical integration”.

We have applied the code DOPRI5 in four different variants to the index 1 formulation of the pendulum equation (2.1): (i) standard application without any projection, (ii) only projection on the position constraint, (iii) only projection on the velocity constraint, (iv) sequential position and velocity projections. The the global

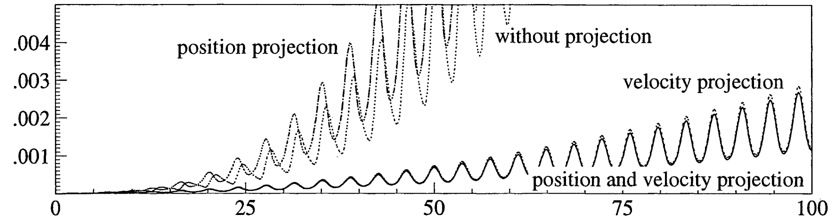


Fig. 2.2. Global error of DOPRI5 with various projections ($Atol = Rtol = 10^{-6}$)

error (in position and velocity) during integration is shown in Fig. 2.2. We conclude that a projection on the position constraint without projection on the velocity constraint does not improve the global error (it makes it even worse in our example). On the other hand, velocity stabilization is as efficient as the complete projection (position and velocity). Nearly no difference can be observed in Fig. 2.2.

Differential Equations with Invariants

Closely related to the above techniques is the numerical treatment of differential equations with invariants. Consider the initial value problem

$$y' = f(y), \qquad y(0) = y_0, \tag{2.12}$$

and suppose that the solution is known to have the invariant

$$\varphi(y) = 0. \tag{2.13}$$

For example, the differential equation (1.46a,e) for (q, u) has the invariants (1.46c) and (1.46d). Conservation laws (total energy,...) may also be written in the form (2.13). The invariant (2.13) is called a *first integral*, if $\varphi_y(y)f(y) \equiv 0$ in a neighbourhood of the solution.

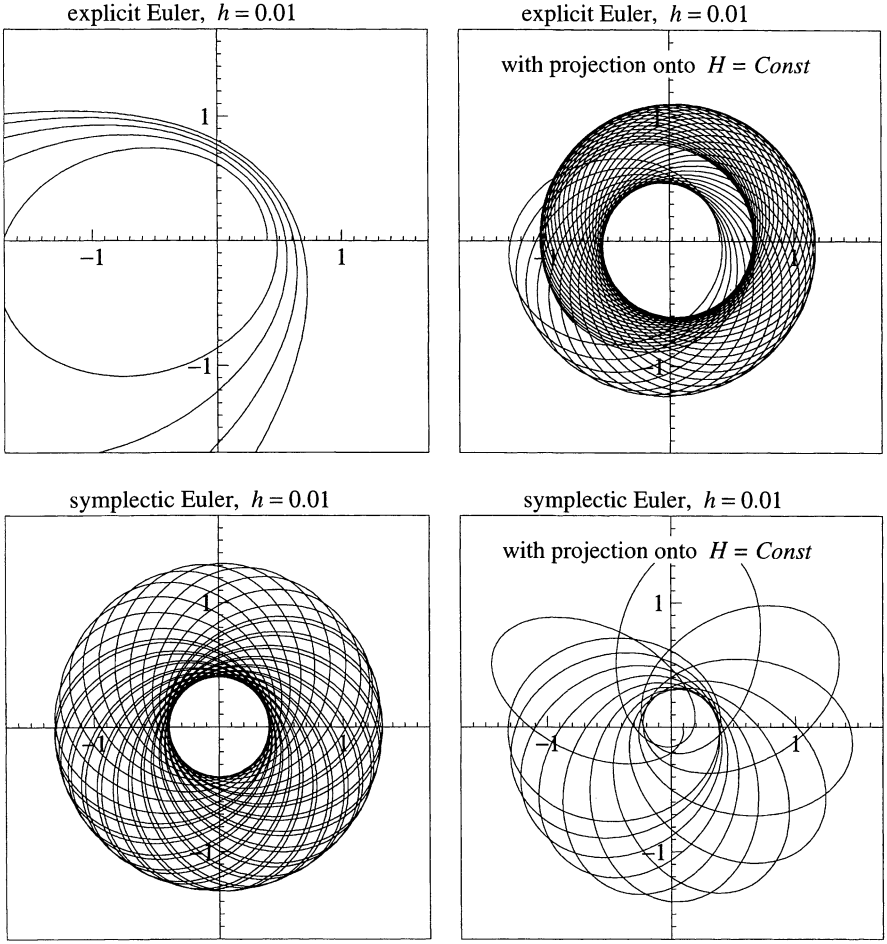
Linear first integrals of the form $\varphi(y) = c + d^T y$ are preserved exactly by most integration methods (e.g., Runge-Kutta and multistep methods). Quadratic first integrals are preserved exactly by symplectic Runge-Kutta methods (see Theorem II.16.7). More complicated invariants are in general not preserved.

The above projection techniques can be adapted to the treatment of the problem (2.12-13) (see Shampine (1986), Eich (1993), Ascher, Chin & Reich (1994)). We apply a numerical method to (2.12) and project (orthogonally or somehow else) the numerical solution onto the manifold defined by (2.13). As discussed above, this precedure retains the order of convergence of the basic method.

Hamiltonian Systems. Differential equations of the form

$$p_i' = -\frac{\partial H}{\partial q_i}(p, q), \qquad q_i' = \frac{\partial H}{\partial p_i}(p, q), \qquad i = 1, \ldots, n, \tag{2.14}$$

where $H : \mathbb{R}^{2n} \to \mathbb{R}$ is a smooth function, always have $H(p, q) = Const$ as first integral. It is tempting to exploit this information and project the numerical solution


explicit Euler, $h = 0.01$

explicit Euler, $h = 0.01$

with projection onto $H = Const$

symplectic Euler, $h = 0.01$

symplectic Euler, $h = 0.01$

with projection onto $H = Const$

Fig. 2.3. Study of the projection onto the manifold $H(p, q) = H(p_0, q_0)$

onto the manifold $H(p, q) = H(p_0, q_0)$. Consider for example the perturbed Kepler problem with Hamiltonian

$$H(p, q) = \frac{p_1^2 + p_2^2}{2} - \frac{1}{\sqrt{q_1^2 + q_2^2}} - \frac{0.005}{\sqrt{(q_1^2 + q_2^2)^3}} \tag{2.15}$$

and initial values $q_1(0) = 1 - e$, $q_2(0) = 0$, $p_1(0) = 0$, $p_2(0) = \sqrt{(1 + e)/(1 - e)}$ (eccentricity $e = 0.6$). The upper pictures of Fig. 2.3 show the numerical solution obtained by the explicit Euler method with step size $h = 0.01$; to the left without any projection, and to the right with projection onto $H = Const$. An improvement can be observed, but the numerical solution still does not reflect the geometric structure of the exact solution (invariant torus). We also have applied the symplectic Euler method (see Eq. (16.54) of Sect. II.16). Here we see that the numerical solution (without projection) shows the correct qualitative behaviour (this can be explained by a backward error analysis, see Sect. II.16), whereas the projection onto $H = Const$ destroys this property. A remedy could be the following: apply a symplectic method to the problem, project the numerical solution to $H = Const$, but continue the integration with the unprojected values.

Methods Based on Local State Space Forms

This method is also called *differential-geometric approach* by Potra & Rheinboldt (1990). The idea is to regard the differential-algebraic system as a differential equation on a manifold (see Sect. VII.1) and to solve the equation in this manifold by introducing suitable local coordinates.

Let us illustrate this approach at the pendulum example. The equations, formulated in cartesian coordinates, are given in the beginning of this section. The solution manifold is (compare with Eq. (1.22))

$$\mathcal{M} = \big\{(x,y,u,v) \mid x^2+y^2=1,\ \ xu+yv=0\big\}.$$

This is a 2-dimensional manifold in $\mathbb{R}^4$ and can be parametrized by (φ,η) as follows:

$$\begin{aligned} x &= \cos\varphi, \qquad & u &= -\eta\sin\varphi, \\ y &= \sin\varphi, \qquad & v &= \eta\cos\varphi. \end{aligned} \tag{2.16}$$

A short calculation shows that the system (2.1a,b), (2.3), written in the new coordinates, leads to the well-known equation

$$\varphi' = \eta, \qquad \eta' = -\cos\varphi. \tag{2.17}$$

This differential equation can be solved numerically without any difficulties. The numerical approximation in the original coordinates is then obtained via (2.16). Obviously, the position and velocity constraints are satisfied exactly.

Although this example nicely illustrates the main ideas, it may be misleading. First of all, in typical applications it is not possible to use one and the same parametrization throughout the whole integration. Secondly, the choice of coordinates is usually not obvious and the transformed differential equation can be much more complicated than the original one (see for example Alishenas (1992)).

Local State Space Form. Suppose that the differential-algebraic system, which we want to solve, can be written as a differential equation

$$y' = v(y), \qquad y \in \mathcal{M} \tag{2.18}$$

on a smooth d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^n$. Consider a coordinate function $\omega : U \to V$ (sufficiently differentiable, bijective, and $\omega'(\eta)$ of full rank) between the open set $U \subset \mathbb{R}^d$ and $V \subset \mathcal{M}$, and denote the coordinates in U by $\eta \in \mathbb{R}^d$. Under the transformation $y = \omega(\eta)$ the equation (2.18) becomes

$$\omega'(\eta)\eta' = v\big(\omega(\eta)\big). \tag{2.19}$$

Since $v(y) \in T_y\mathcal{M}$ for all $y \in \mathcal{M}$ (see Eq. (1.19)), there exists η' such that (2.19) holds. Moreover η' is unique, because $\omega'(\eta)$ is of full rank. Using the notation $\omega'(\eta)^+ = \big(\omega'(\eta)^T\omega'(\eta)\big)^{-1}\omega'(\eta)^T$ for the pseudo-inverse of $\omega'(\eta)$ we therefore obtain

$$\eta' = \omega'(\eta)^+ v\big(\omega(\eta)\big), \tag{2.20}$$

which is an ordinary differential equation in $\mathbb{R}^d$ and is called *local state space form* of (2.18). Observe that different coordinate functions lead to different state space forms.

The *numerical procedure* for solving (2.18) is the following: suppose that an approximation $y_k \in \mathcal{M}$ of $y(t_k)$ is given. We then choose a coordinate function and apply a standard method (e.g., Runge-Kutta) with initial value $\eta_k = \omega^{-1}(y_k)$ to the state space form (2.20). This yields an approximation η_{k+1} at time t_{k+1}. Finally, we put $y_{k+1} = \omega(\eta_{k+1})$. By definition of this procedure, the numerical approximation y_{k+1} again lies in $\mathcal{M}$.

If one uses one and the same local state space form for the whole integration (as it is the case for the pendulum example, Eq. (2.17)), the convergence properties for (2.20) carry immediately over to (2.18) via the coordinate function $y = \omega(\eta)$. In more complex situations it may be necessary to change the coordinates several times, and from a computational point of view it may even be more advantageous to change them in every integration step.

Theorem 2.2. *Consider the above procedure for the numerical solution of (2.18), and denote by $y = \omega_k(\eta)$ the coordinate transformation of the kth step. If, in a neighbourhood of $\omega_k^{-1}(y_k)$, the matrices $\omega_k'(\eta)$ and $\omega_k'(\eta)^+$ are uniformly bounded in k, then the convergence properties for standard ordinary differential equations carry over to the problem (2.18) on a manifold $\mathcal{M}$.*

Proof. In the case of one-step methods we have

$$y_{k+1} = \omega_k\Big(\omega_k^{-1}(y_k) + h\Phi_k\big(\omega_k^{-1}(y_k), h\big)\Big),$$

where $\Phi_k(\eta, h)$ is the increment function of the method when applied to (2.20) with ω replaced by ω_k. Due to the regularity assumptions on $\omega_k(\eta)$, this formula can be written as

$$y_{k+1} = y_k + h\Psi_k(y_k, h)$$

and takes the form of a standard one-step method. The assumptions guarantee that the functions Ψ_k have a uniform Lipschitz constant with respect to the first argument. Therefore the convergence proofs of Sect. II.3 apply. For multistep methods the situation is analogous. □

Choice of Local Coordinates. Let us explain two choices for the constrained mechanical system (2.4), whose solution manifold is given by

$$\mathcal{M} = \{(q,u) \mid g(q) = 0,\ \ G(q)u = 0\}. \tag{2.21}$$

Here $q, u \in \mathbb{R}^n$ are generalized coordinates, $g(q) \in \mathbb{R}^m$ and $G(q) = g_q(q)$. The adaptation to other differential-algebraic systems with known solution manifold is more or less straightforward.

Generalized Coordinate Partitioning (Wehage & Haug 1982). Assuming that the Jacobian $G(q)$ has full row rank, there exists a partitioning $q = (\eta, \widehat{\eta})$ such that $g_{\widehat{\eta}}(\eta, \widehat{\eta})$ is invertible ($\eta \in \mathbb{R}^{n-m}$, $\widehat{\eta} \in \mathbb{R}^m$). By the Implicit Function Theorem the constraint $g(q) = 0$ can be solved for $\widehat{\eta}$ in a neighbourhood of a consistent value $q_0 = (\eta_0, \widehat{\eta}_0)$. Hence, there exists a function $\widehat{\eta} = h(\eta)$ (defined for η close to η_0) such that $g(\eta, h(\eta)) = 0$. With a corresponding partitioning $u = (\nu, \widehat{\nu})$ the velocity constraint becomes $g_\eta(\eta, \widehat{\eta})\nu + g_{\widehat{\eta}}(\eta, \widehat{\eta})\widehat{\nu} = 0$ and allows us to express $\widehat{\nu}$ in terms of η, ν as $\widehat{\nu} = k(\eta, \nu)$. A coordinate function is thus given by $\omega(\eta, \nu) = \big((\eta, h(\eta)), (\nu, k(\eta, \nu))\big)$, and the differential equation in these local coordinates is

$$\eta' = \nu, \qquad \nu' = \nu'\big(\omega(\eta, \nu)\big), \tag{2.22}$$

where $\nu'(q, u)$ collects the ν-components of the solution $u'(q, u)$ of the linear system (1.38e). We emphasize that for a numerical implementation the differential equation (2.22) need not be known analytically. However, a nonlinear system has to be solved each time when the right-hand side of (2.22) has to be evaluated.

Tangent Space Parametrization (Potra & Rheinboldt 1991, Yen 1993). Instead of partitioning the components of q and u we split the vectors $q - q_0$ and $u - u_0$ according to

$$q - q_0 = Q_0\eta + Q_1\widehat{\eta}, \qquad u - u_0 = Q_0\nu + Q_1\widehat{\nu}, \tag{2.23}$$

where the columns of Q_0 form a basis of the tangent space $\{v \mid G(q_0)v = 0\}$ to the manifold $q(q) = 0$, which is completed by the columns of Q_1 to a basis of the whole space. The condition $g(q) = 0$ together with the first relation of (2.23) define (locally) q and $\widehat{\eta}$ as functions of η. Similarly, $G(q)u = 0$ and the second relation of (2.23) define u and $\widehat{\nu}$ as functions of ν and q. Denoting these relationships by $\widehat{\eta} = h(\eta)$, $\widehat{\nu} = k(\eta, \nu)$, we get formally the same coordinate function as in the previous example, and the state space form is given by

$$\eta' = \nu, \qquad \nu' = Q_0^+ u'\big(\omega(\eta, \nu)\big), \tag{2.24}$$

where $Q_0^+ = (Q_0^T Q_0)^{-1} Q_0^T$ is the pseudo-inverse of Q_0, and $u'(q, u)$ denotes the solution of the linear system (2.5).

The evaluation of $h(\eta)$ requires the solution of a nonlinear system, whose Jacobian is

$$\begin{pmatrix} I & -Q_1 \\ G(q_0) & 0 \end{pmatrix}.$$

This suggests to take $-Q_1 = G^T(q_0)$ or better $-Q_1 = M^{-1}(q_0)G^T(q_0)$, so that simplified Newton iterations lead to linear systems with a matrix that already appears in (2.5). The linear system for the computation of $k(\eta, \nu)$ has the same structure.

Due to the fact that the evaluation of the right-hand side of (2.24) requires the solution of a nonlinear system, the authors of this approach prefer the use of multistep methods which, in general, use less function evaluations than one-step methods. In connection with Runge-Kutta methods, Potra (1995) suggests the use of certain predicted values instead of the exact solutions of these nonlinear systems, and requires that only the approximation at the end of every step lies on the manifold $\mathcal{M}$. The resulting algorithm is then equivalent to solving the index 1 problem combined with projections onto $\mathcal{M}$ at the end of each step.

Overdetermined Differential-Algebraic Equations

In contrast to the approach at the beginning of this section, where the constraint is replaced by one of its derivatives, we consider the original system and one or more derivatives of the constraints as a unity. For example, the equations of motion of a constrained mechanical system become

$$\begin{aligned} q' &= u && (2.25a)\\ M(q)u' &= f(q,u) - G^T(q)\lambda && (2.25b)\\ 0 &= g(q) && (2.25c)\\ 0 &= G(q)u && (2.25d)\\ 0 &= g_{qq}(q)(u,u) + G(q)M(q)^{-1}\big(f(q,u) - G^T(q)\lambda\big). && (2.25e) \end{aligned}$$

This system is overdetermined, because we are concerned with more equations than unknowns. Nevertheless, it possesses a unique solution, if (1.47) is satisfied and consistent initial values are prescribed.

We illustrate the numerical solution of (2.25) with the BDF method. A formal application (see Sect. VI.2) gives

$$\begin{aligned} & q_k - \widehat{q} - h\gamma u_k = 0 && (2.26a)\\ & M(q_k)(u_k - \widehat{u}) - h\gamma\big(f(q_k,u_k) - G^T(q_k)\lambda_k\big) = 0 && (2.26b)\\ & g(q_k) = 0 && (2.26c)\\ & G(q_k)u_k = 0 && (2.26d)\\ & g_{qq}(q_k)(u_k,u_k) + G(q_k)M(q_k)^{-1}\big(f(q_k,u_k) - G^T(q_k)\lambda_k\big) = 0, && (2.26e) \end{aligned}$$

where $\gamma = \beta_k/\alpha_k$, $\widehat{q} = \big(\sum_{i=0}^{k-1}\alpha_i q_i\big)/\alpha_k$, and $\widehat{u} = \big(\sum_{i=0}^{k-1}\alpha_i u_i\big)/\alpha_k$ are known quantities. The system (2.26) is overdetermined and does not have a solution, in general. A natural idea (Führer 1988) is to search for a least square solution of (2.26). There are several ways to do this. One can consider different norms, or one can require some of the equations to be exactly satisfied and the remaining ones in a least square sense. Führer & Leimkuhler (1991) impose all constraints (2.26c,d,e), and treat the remaining equations by the use of a special pseudoinverse. This can be achieved by introducing Lagrange multipliers μ_k, η_k in the first two equations of (2.26) as follows:

$$M(q_k)(q_k - \widehat{q} - h\gamma u_k) + h\gamma\big(G^T(q_k)\mu_k + (G_q(q_k)u_k)^T\eta_k\big) = 0 \qquad (2.27a)$$

$$M(q_k)(u_k - \widehat{u}) - h\gamma\big(f(q_k,u_k) - G^T(q_k)\lambda_k\big) + h\gamma G^T(q_k)\eta_k = 0. \qquad (2.27b)$$

For sufficiently small h, the system (2.27a,b), (2.26c,d,e) has a locally unique solution, if (1.47) is satisfied.

Connection with GGL-Formulation. If we omit the acceleration constraint (2.26e), there is no need for two Lagrange multipliers, and we can put $\eta_k = 0$. The resulting system (2.27a,b), (2.26c,d) is then nothing else than the standard BDF discretization of the system (1.48).

Unstructured Higher Index Problems

We consider a general differential-algebraic system

$$F(u',u) = 0. \qquad (2.28)$$

For its numerical solution we shall construct an 'underlying ODE' (see Definition 1.2) and solve it by any integration method. This approach has been developed in several papers by Campbell (1989, 1993). We shall explain the main ideas following the presentation of Campbell & Moore (1995).

Inspired by the definition of the differentiation index we consider the *derivative array equations*

$$F(u',u) = 0, \quad \frac{dF(u',u)}{dx} = 0\ , \ \ldots\ , \ \frac{d^m F(u',u)}{dx^m} = 0$$

which we write in compact form as

$$G(u',w,u) = 0, \qquad (2.29)$$

where $w = (u'', u''', \ldots, u^{(m+1)})$ collects the higher derivatives of u. In Eq. (2.29) we consider w, u, and also u' as independent variables. Besides the usual differentiability assumptions we assume that

(A1) the matrix $(G_{u'}, G_w)$ is 1-full with respect to u'; this means that the relation $G_{u'}\Delta u' + G_w \Delta w = 0$ implies $\Delta u' = 0$;

(A2) the matrix $(G_{u'}, G_w)$ has constant rank;

(A3) the matrix $(G_{u'}, G_w, G_u)$ has full row rank.

These assumptions are required to hold in a neighbourhood of a particular solution of (2.28). The construction of the underlying ODE is based on the following lemma and on its proof.

Lemma 2.3 (Campbell & Moore 1995). *Consider a sufficiently smooth problem (2.28) and assume that (A1), (A2), and (A3) hold. Then there exist coordinate partitions* $w = (w_a, w_b)$, $u = (u_a, u_b)$ *(and also* $u' = (u'_a, u'_b)$ *with the same partition* *as for* u *), such that the derivative array equations (2.29) are equivalent to*

$$\begin{aligned} u_a' &= f_a(u_b), \qquad & w_a &= \varphi_2(w_b, u_b) \\ u_b' &= f_b(u_b), \qquad & u_a &= \varphi_3(u_b) \end{aligned} \tag{2.30}$$

in a neighbourhood of the consistent initial value (u_0', w_0, u_0).

Proof. We consider the matrix $(G_{u'}, G_w, G_u)$ evaluated at (u_0', w_0, u_0) and perform a QR factorization, where column permutations are restricted to components within the vectors u', w, and u. This yields

$$Q^T (G_{u'}, G_w, G_u) P = \left(\begin{array}{c|cc|cc} B_1 & C_1 & C_2 & D_1 & D_2 \\ 0 & C_3 & C_4 & D_3 & D_4 \\ 0 & 0 & 0 & D_5 & D_6 \end{array}\right), \tag{2.31}$$

where B_1, C_3, D_5 are nonsingular by Assumption (A3), Q is an orthogonal matrix, and $P = \operatorname{diag}(P_1, P_2, P_3)$ with suitable permutation matrices P_1, P_2, P_3. Fixing the permutation P, we apply the above factorization also to $(G_{u'}, G_w, G_u)$ evaluated at an arbitrary point (u', w, u) close to (u_0', w_0, u_0). Because of Assumption (A2) this gives (2.31) with smooth matrices Q, B_i, C_i, and D_i. The decomposition (2.31) defines the partitions $w = (w_a, w_b)$ and $u = (u_a, u_b)$. The first, second and fourth block-columns in (2.31) form an invertible matrix. The Implicit Function Theorem thus implies that (2.29) can be solved for u', w_a, u_a, and we obtain the equivalent system

$$u' = \varphi_1(w_b, u_b), \qquad w_a = \varphi_2(w_b, u_b), \qquad u_a = \varphi_3(w_b, u_b).$$

We still have to show that the functions φ_1 and φ_3 are independent of w_b. By definition of the φ_i we have

$$G\Big(\varphi_1(w_b, u_b), (\varphi_2(w_b, u_b), w_b), (\varphi_3(w_b, u_b), u_b)\Big) = 0.$$

Differentiating with respect to w_b yields

$$G_{u'} \cdot \frac{\partial \varphi_1}{\partial w_b} + G_{w_a} \cdot \frac{\partial \varphi_2}{\partial w_b} + G_{w_b} + G_{u_a} \cdot \frac{\partial \varphi_3}{\partial w_b} = 0. \tag{2.32}$$

Multiplying this relation by Q^T, we see from Eq. (2.31) that $D_5(\partial\varphi_3/\partial w_b) = 0$. Since D_5 is nonsingular, this implies $(\partial\varphi_3/\partial w_b) = 0$, so that φ_3 is independent of w_b. Assumption (A1) now implies from (2.32) that also $(\partial\varphi_1/\partial w_b)$ vanishes. This completes the proof of the lemma. □

Suppose that we know how to compute $f_a(u_b)$, $f_b(u_b)$ and $\varphi_3(u_b)$ for a given value u_b. From (2.30) we then have an ordinary differential equation for u_b, which can be solved by any integration method (Runge-Kutta or multistep, explicit or implicit, ...), and the remaining components are given by $u_a = \varphi_3(u_b)$. The numerical solution of this method thus preserves all constraints (also the hidden ones).

Computation of the Values $f_a(u_b)$, $f_b(u_b)$ *and* $\varphi_3(u_b)$. It follows from Assumption (A3) that $(G_{u'}, G_w, G_u)^T G = 0$ is equivalent to $G = 0$. Thus, for given u_b, any method of finding the minimum (u', w, u_a) of the function $G^T G$ may be used. Campbell & Moore (1995) propose the use of Gauss-Newton iterations.

Remark. A closely related algorithm has been proposed by Kunkel & Mehrmann (1996). Instead of extracting from the derivative array equations an ordinary differential equation for all variables, they extract an equivalent index 1 problem and solve it by standard integration methods. This modification usually requires one differentiation less of the original system (2.28).

Exercises

1. Repeat the experiment of Fig. 2.1 with other numerical methods (explicit Euler method, multistep methods, constant and variable step sizes, ...). You will observe that in some situations the error in $g(q_n)$ grows only linearly, and the error in $G(q_n)u_n$ remains bounded. Try to explain this observation.

2. a) Prove that the matrix in (2.5) is 1-full with respect to u' if and only if the restriction of M to the kernel of G is injective (this is exactly the condition that is needed in order to be able to apply the methods of this section).

   b) Show by examples that neither M needs to be nonsingular nor G has to be of full rank in order that the condition of part (a) is satisfied.