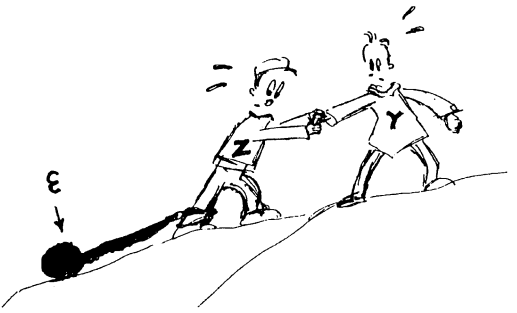
Chapter VI. Singular Perturbation Problems and Index 1 Problems



(Drawing by G. Di Marzo)

Singular perturbation problems (SPP) form a special class of problems containing a parameter ε. When this parameter is small, the corresponding differential equation is stiff; when ε tends to zero, the differential equation becomes differential algebraic. This chapter investigates the numerical solution of such singular perturbation problems. This allows us to understand many phenomena observed for very stiff problems. Much insight is obtained by studying the limit case $\varepsilon = 0$ ("the reduced system" or "problem of index 1") which is usually much easier to analyze.

We start by considering the limit case $\varepsilon = 0$. Two numerical approaches – the ε-embedding method and the state space form method – are investigated in Sect. VI.1. We then analyze multistep methods in Sect. VI.2, Runge-Kutta methods in Sect. VI.3, Rosenbrock methods in Sect. VI.4 and extrapolation methods in Sect. VI.5. Convergence is studied for singular perturbation problems and for semi-explicit differential-algebraic systems of "index 1".

VI.1 Solving Index 1 Problems

Singular perturbation problems (SPP) have several origins in applied mathematics. One comes from fluid dynamics and results in linear boundary value problems containing a small parameter ε (the coefficient of viscosity) such that for $\varepsilon \to 0$ the differential equation loses the highest derivative (see Exercise 1 below). Others originate in the study of nonlinear oscillations with *large* parameters (van der Pol 1926, Dorodnicyn 1947) or in the study of chemical kinetics with slow and fast reactions (see e.g., Example (IV.1.4)).

Asymptotic Solution of van der Pol's Equation

The classical paper of Dorodnicyn (1947) studied the van der Pol Equation (IV.1.5') with large μ, i.e., with small ε. The investigation becomes a little easier if we use Liénard's coordinates (see Exercise I.16.8). In Eq. (IV.1.5'), written here as

$$\varepsilon z'' + (z^2 - 1)z' + z = 0, \tag{1.1}$$

we insert the identity

$$\varepsilon z'' + (z^2 - 1)z' = \frac{d}{dx} \underbrace{\left(\varepsilon z' + \left(\frac{z^3}{3} - z\right)\right)}_{:= y}$$

so that (1.1) becomes

$$\begin{aligned} y' &= -z && =: f(y,z) \\ \varepsilon z' &= y - \left(\frac{z^3}{3} - z\right) && =: g(y,z). \end{aligned} \tag{1.2}$$

Fig. 1.1 shows solutions of Eq. (1.2) with $\varepsilon = 0.03$ in the (y,z)-plane. One observes rapid movements towards the manifold M defined by $y = z^3/3 - z$, close to which the solution becomes smooth. In order to approximate the solution for very small ε, we set $\varepsilon = 0$ in (1.2) and obtain the so-called *reduced* system

$$\begin{aligned} y' &= -z && = f(y,z) \\ 0 &= y - \left(\frac{z^3}{3} - z\right) && = g(y,z). \end{aligned} \tag{1.2'}$$

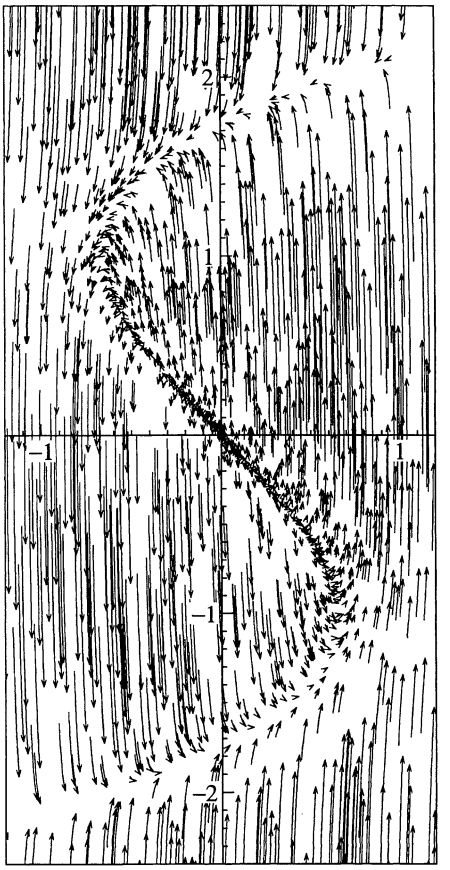


Fig. 1.1. Solutions of SPP (1.2)

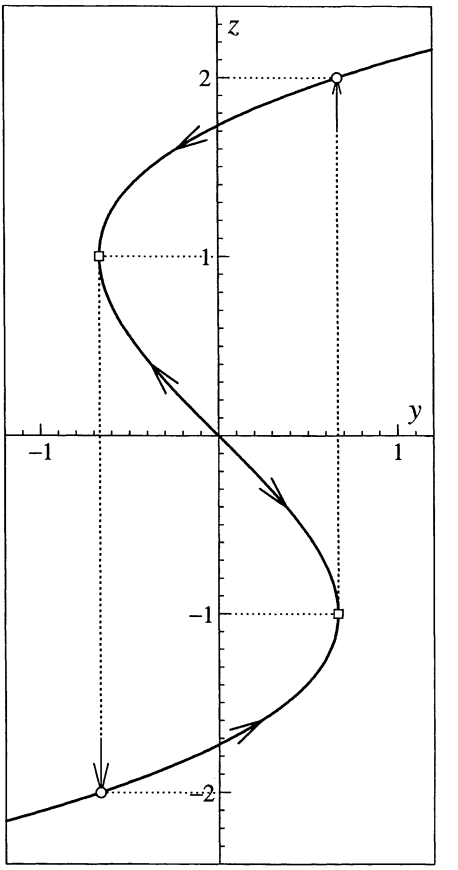


Fig. 1.2. Reduced problem (1.2')

While (1.2) has no analytic solution, (1.2') can easily be solved to give

$$y' = -z = (z^2 - 1)z' \quad \text{or} \quad \ln|z| - \frac{z^2}{2} = x + C. \tag{1.3}$$

Equation (1.2') is called a *differential algebraic equation* (DAE), since it combines a differential equation (first line) with an algebraic equation (second line). Such a problem only makes sense if the initial values are *consistent*, i.e., lie on the manifold M. The points of M with coordinates $y = \pm 2/3$, $z = \mp 1$ are of special interest (Fig. 1.2): at these points the partial derivative $g_z = \partial g/\partial z$ vanishes and the defining manifold is no longer "transversal" to the direction of the fast movement. Here the solutions of (1.2') cease to exist, while the solutions of the full problem (1.2) for $\varepsilon \to 0$ jump with "infinite" speed to the opposite manifold. For $-1 < z < 1$ the manifold M is *unstable* for the solution of (1.2) (here $g_z > 0$), otherwise M is *stable* ($g_z < 0$).

We demonstrate the power of the reduced equation by answering the question:

what is the period T of the limit cycle solution of van der Pol's equation for $\varepsilon \to 0$? Fig. 1.2 shows that the asymptotic value of T is just twice the time which $z(x)$ of (1.3) needs to advance from $z = -2$ to $z = -1$, i.e.,

$$T = 3 - 2\ln 2. \tag{1.4}$$

This is the first term of Dorodnicyn's asymptotic formula. We also see that $z(x)$ reaches its largest values (i.e., crosses the Poincaré cut $z' = 0$, see Fig. I.16.2) at $z = \pm 2$. We thus have the curious result that the limit cycle of van der Pol's equation (1.1) has the same asymptotic initial value $z = 2$ and $z' = 0$ for $\varepsilon \to 0$ and for $\varepsilon \to \infty$ (see Eq. (I.16.10)).

The ε-Embedding Method for Problems of Index 1

We now want to study the behaviour of the *numerical solution* for $\varepsilon \to 0$. This will give us insight into many phenomena encountered for very stiff equations and also suggest advantageous numerical procedures for stiff and differential-algebraic equations. Let an arbitrary singular perturbation problem be given,

$$y' = f(y, z) \tag{1.5a}$$

$$\varepsilon z' = g(y, z), \tag{1.5b}$$

where y and z are vectors; suppose that f and g are sufficiently often differentiable vector functions of the same dimensions as y and z, respectively. The corresponding *reduced* equation is the DAE

$$y' = f(y, z) \tag{1.6a}$$

$$0 = g(y, z), \tag{1.6b}$$

whose initial values are *consistent* if $0 = g(y_0, z_0)$. A general assumption of the present chapter will be that the Jacobian

$$g_z(y, z) \quad \text{is invertible} \tag{1.7}$$

in a neighbourhood of the solution of (1.6). Equation (1.6b) then possesses a locally unique solution $z = G(y)$ ("Implicit Function Theorem") which inserted into (1.6a) gives

$$y' = f\big(y, G(y)\big), \tag{1.8}$$

the so-called "state space form", an ordinary differential system. Under the assumption (1.7), Eq. (1.6) is said to be a differential-algebraic equation of *index 1*.

An interesting approach for solving (1.6) is to apply some numerical method to the SPP (1.5) and to put $\varepsilon = 0$ in the resulting formulas. Let us illustrate this approach for Runge-Kutta methods. Applied to the system (1.5) we obtain

$$Y_{ni} = y_n + h \sum_{j=1}^{s} a_{ij} f\,(Y_{nj}, Z_{nj}) \tag{1.9a}$$

$$\varepsilon Z_{ni} = \varepsilon z_n + h \sum_{j=1}^{s} a_{ij}\, g\,(Y_{nj}, Z_{nj}) \tag{1.9b}$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i\, f\,(Y_{ni}, Z_{ni}) \tag{1.9c}$$

$$\varepsilon z_{n+1} = \varepsilon z_n + h \sum_{i=1}^{s} b_i\, g\,(Y_{ni}, Z_{ni}). \tag{1.9d}$$

We now suppose that the RK matrix (a_{ij}) is invertible and obtain from (1.9b)

$$hg(Y_{ni}, Z_{ni}) = \varepsilon \sum_{j=1}^{s} \omega_{ij}(Z_{nj} - z_n), \tag{1.10}$$

where the ω_{ij} are the elements of the inverse of (a_{ij}). Inserting this into (1.9d) makes the definition of z_{n+1} independent of ε. We thus put without more ado $\varepsilon = 0$ and obtain

$$Y_{ni} = y_n + h \sum_{j=1}^{s} a_{ij}\, f\,(Y_{nj}, Z_{nj}) \tag{1.11a}$$

$$0 = g(Y_{ni}, Z_{ni}) \tag{1.11b}$$

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i\, f\,(Y_{ni}, Z_{ni}) \tag{1.11c}$$

$$z_{n+1} = \left(1 - \sum_{i,j=1}^{s} b_i\, \omega_{ij}\right) z_n + \sum_{i,j=1}^{s} b_i\, \omega_{ij}\, Z_{nj}. \tag{1.11d}$$

Here

$$1 - \sum_{i,j=1}^{s} b_i\, \omega_{ij} = R(\infty) \tag{1.11e}$$

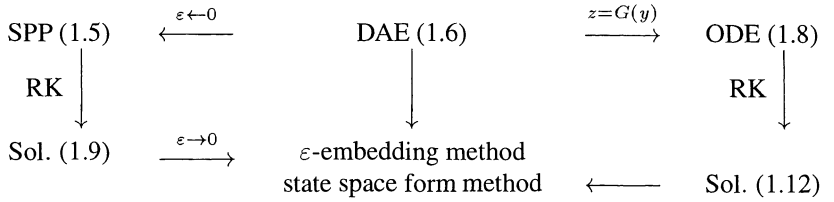
(see Eq. (IV.3.15)), where $R(z)$ is the stability function of the method.

State Space Form Method

The numerical solution (y_{n+1}, z_{n+1}) of the above approach will usually *not* lie on the manifold $g(y, z) = 0$. However, this can easily be repaired by replacing (1.11d) by the condition

$$0 = g(y_{n+1}, z_{n+1}). \tag{1.12}$$

Then, we do not only have $Z_{nj} = G(Y_{nj})$ (see (1.11b)), but also $z_{n+1} = G(y_{n+1})$. In this case the method (1.11a–c), (1.12) is *identical* to the solution of the state space form (1.8) with the same Runge-Kutta method. This will be called the *state space form method*. The whole situation is summarized in the following diagram:



Of special importance here are *stiffly accurate* methods, i.e., methods which satisfy

$$a_{si} = b_i \qquad \text{for} \quad i = 1, \ldots, s. \tag{1.13}$$

This means that $y_{n+1} = Y_{ns}$, $z_{n+1} = Z_{ns}$ and (1.12) is satisfied anyway. Hence for stiffly accurate methods the ε-embedding method and the state space form method are identical. For this reason, Griepentrog & März (1986) denote such methods IRK(DAE).

Both approaches have their own merits. Theoretical results for the ε-embedding method yield insight into the method when applied to singular perturbation problems. Moreover, this approach can easily be extended to more general situations, where the algebraic relation is not explicitly separated from the differential equation (see below). The state space form method, on the other hand, has the advantage that it is not restricted to implicit methods. Applying an explicit Runge-Kutta method or a multistep method to Eq. (1.8) is certainly a method of choice for semi-explicit index 1 equations. No new theory is necessary in this case.

A Transistor Amplifier

> ... auf eine merkwürdige Tatsache aufmerksam machen, das ist die außerordentlich grosse Zahl berühmter Mathematiker, die aus Königsberg stammen ...: Kant 1724, Richelot 1808, Hesse 1811, Kirchhoff 1824, Carl Neumann 1832, Clebsch 1833, Hilbert 1862.
> (F. Klein, Entw. der Math., p. 159)

Very often, differential-algebraic problems arising in practice are not at once in the semi-explicit form (1.6), but rather in the form $Mu' = \varphi(u)$ where M is a constant *singular* matrix.

As an example we compute the amplifier of Fig. 1.3, where $U_e(t)$ is the entry voltage, $U_b = 6$ the operating voltage, $U_i(t)$ $(i = 1, 2, 3, 4, 5)$ the voltages at the nodes $1, 2, 3, 4, 5$, and $U_5(t)$ the output voltage. The current through a resistor satisfies $I = U/R$ (Ohm 1827), the current through a capacitor $I = C \cdot dU/dt$, where R and C are constants and U the voltage. The transistor acts as amplifier in that the current from node 4 to node 3 is 99 times larger than that from node 2 to node 3 and depends on the voltage difference $U_3 - U_2$ in a nonlinear way. Kirchhoff's law (a Königsberg discovery) says that the sum of currents entering a node vanishes. This law applied to the 5 nodes of Fig. 1.3 leads to the following equations:

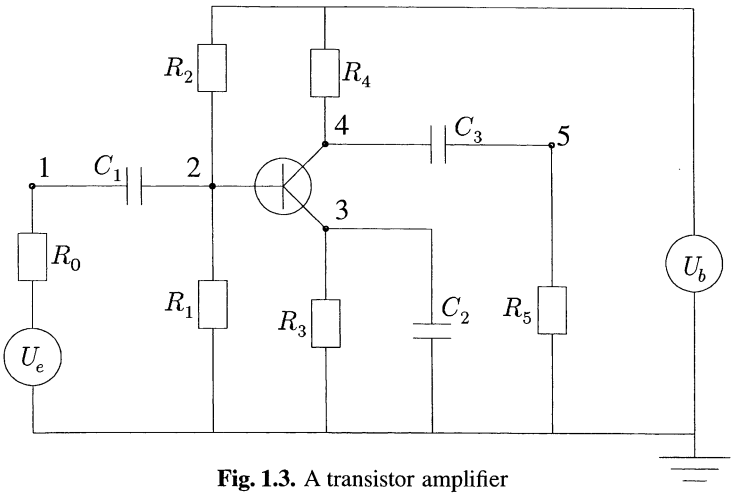


Fig. 1.3. A transistor amplifier

$$
\begin{aligned}
&\text{node 1:} && \frac{U_e(t)}{R_0} - \frac{U_1}{R_0} + C_1(U_2' - U_1') = 0 \\
&\text{node 2:} && \frac{U_b}{R_2} - U_2\Big(\frac{1}{R_1} + \frac{1}{R_2}\Big) + C_1(U_1' - U_2') - 0.01\, f(U_2 - U_3) = 0 \\
&\text{node 3:} && f(U_2 - U_3) - \frac{U_3}{R_3} - C_2 U_3' = 0 \\
&\text{node 4:} && \frac{U_b}{R_4} - \frac{U_4}{R_4} + C_3(U_5' - U_4') - 0.99 f(U_2 - U_3) = 0 \\
&\text{node 5:} && -\frac{U_5}{R_5} + C_3(U_4' - U_5') = 0.
\end{aligned}
\tag{1.14}
$$

As constants we adopt the values reported (for a similar problem) by Rentrop, Roche & Steinebach (1989)

$$
\begin{aligned}
&f(U) = 10^{-6}\Big(\exp\Big(\frac{U}{0.026}\Big) - 1\Big) \\
&R_0 = 1000, \quad R_1 = \ldots = R_5 = 9000 \\
&C_k = k \cdot 10^{-6}, \quad k = 1, 2, 3,
\end{aligned}
$$

and the initial signal is chosen as

$$
U_e(t) = 0.4 \cdot \sin(200\pi t). \tag{1.15}
$$

Equations (1.14) are of the form $Mu' = \varphi(u)$ where

$$
M = \begin{pmatrix}
-C_1 & C_1 & & & \\
C_1 & -C_1 & & & \\
 & & -C_2 & & \\
 & & & -C_3 & C_3 \\
 & & & C_3 & -C_3
\end{pmatrix}
$$

is obviously a singular matrix of rank 3. The sum of the first two and of the last two equations leads directly to two algebraic equations. Introducing e.g.,

$$U_1 - U_2 = y_1, \quad U_3 = y_2, \quad U_4 - U_5 = y_3, \quad U_1 = z_1, \quad U_4 = z_2,$$

transforms equations (1.14) to the form (1.6). *Consistent initial values* must thus satisfy $\varphi_1(u) + \varphi_2(u) = 0$ and $\varphi_4(u) + \varphi_5(u) = 0$. If we put $U_2(0) = U_3(0)$, we have $f(U_2(0) - U_3(0)) = 0$. Since $U_e(0) = 0$, we then easily find consistent initial values, e.g., as

$$U_1(0) = 0, \quad U_2(0) = U_3(0) = \frac{U_b R_1}{R_1 + R_2}, \quad U_4(0) = U_b, \quad U_5(0) = 0. \quad (1.16)$$

Problems of the Form $M u' = \varphi(u)$

Numerical methods for problems of the form

$$M u' = \varphi(u), \quad (1.17)$$

where M is a constant matrix, can be derived as follows: we assume that M is regular, apply an ODE method to $u' = M^{-1}\varphi(u)$ and multiply the resulting formulas by M. For Runge-Kutta methods we obtain in this way

$$M(U_{ni} - u_n) = h \sum_{j=1}^{s} a_{ij} \varphi(U_{nj}) \quad (1.18a)$$

$$u_{n+1} = \left(1 - \sum_{i,j=1}^{s} b_i \omega_{ij}\right) u_n + \sum_{i,j=1}^{s} b_i \omega_{ij} U_{nj}, \quad (1.18b)$$

where again (ω_{ij}) is the inverse of (a_{ij}). The second formula was obtained from

$$M(u_{n+1} - u_n) = h \sum_{i=1}^{s} b_i \varphi(U_{ni}) \quad (1.18c)$$

in exactly the same way as above (see (1.10)).

Formulas (1.18) also make sense formally when M is a *singular* matrix. In this case, problem (1.17) is mathematically equivalent to a semi-explicit system (1.6) and method (1.18) corresponds to method (1.11). This can be seen as follows: we decompose the matrix M (e.g., by Gaussian elimination with total pivoting) as

$$M = S \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} T, \quad (1.19)$$

where S and T are invertible matrices and the dimension of I represents the rank of M. Inserting this into (1.17), multiplying by S^{-1}, and using the transformed variables

$$Tu = \begin{pmatrix} y \\ z \end{pmatrix} \quad (1.20)$$

gives

$$\begin{pmatrix} y' \\ 0 \end{pmatrix} = S^{-1}\varphi\Big(T^{-1}\begin{pmatrix} y \\ z \end{pmatrix}\Big) =: \begin{pmatrix} f(y,z) \\ g(y,z) \end{pmatrix}, \tag{1.21}$$

a problem of type (1.6). An initial value u_0 is *consistent* if $\varphi(u_0)$ lies in the range of the matrix M.

Similarly, if (1.19) is inserted into (1.18), and the variables

$$TU_{nj} = \begin{pmatrix} Y_{nj} \\ Z_{nj} \end{pmatrix}, \qquad Tu_n = \begin{pmatrix} y_n \\ z_n \end{pmatrix} \tag{1.22}$$

are introduced, Eq. (1.18b) (for Z_{n+1}) and Eq. (1.18c) (for Y_{n+1}) lead precisely to equations (1.11). This means that the diagram

$$\begin{array}{ccc} \text{Problem (1.17)} & \xrightarrow{\textit{Transf. (1.20)}} & \text{Problem (1.6)} \\ \Big\downarrow \textit{Meth. (1.18)} & & \Big\downarrow \textit{Meth. (1.11)} \\ \{u_n\} & \xrightarrow{\textit{Transf. (1.22)}} & \{y_n\}, \{z_n\} \end{array} \tag{1.23}$$

commutes. An important consequence of this commutativity is that all results for semi-explicit systems (1.6) and the ε-embedding method (1.11) (existence of a numerical solution, convergence, asymptotic expansions, ...) also apply to implicit problems (1.17) with singular M and method (1.18).

All codes, such as RADAU5, which have an option for implicit differential equations (1.17) can thus be applied directly. This has been done for problem (1.14) with initial values (1.16), integration interval $0 \le x \le 0.2$, and $Tol = 10^{-4}$. The code computed the solution $U_5(t)$ displayed in Fig. 1.4 in 556 (accepted) steps. The comparison with the entry voltage $U_e(t)$ shows that our amplifier is working. See also Hairer, Lubich & Roche (1989), p. 108-111 for a more elaborate example.

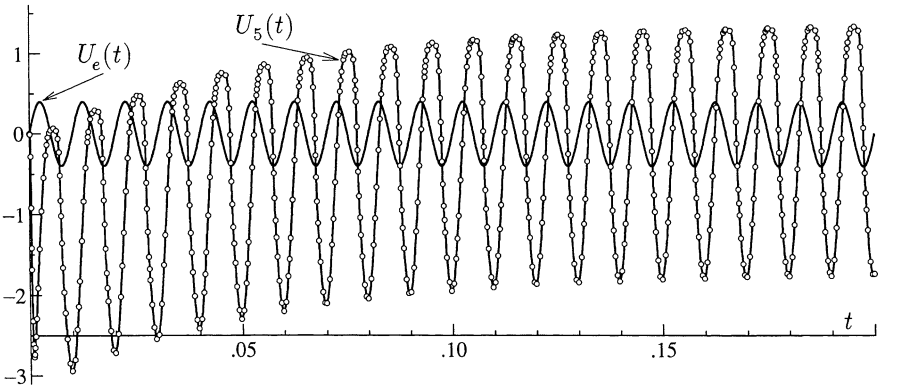


Fig. 1.4. Computed solution of amplifier problem (1.14)

Convergence of Runge-Kutta Methods

If the method is stiffly accurate, the numerical solutions (1.11) are equivalent to those of the *ordinary* equation (1.8). Therefore the convergence of the solutions is described by Theorems II.3.4 and II.3.6 as

$$y_n - y(x_n) = \mathcal{O}(h^p), \qquad z_n - z(x_n) = \mathcal{O}(h^p), \tag{1.24}$$

where p is the *classical* order of the method (the second formula follows from a Lipschitz condition for G). For *general* methods, the estimate (1.24) remains valid for y_n, because (1.11a,b,c) are independent of z_n and do not change if (1.11d) is replaced by (1.12). Thus we only have to prove a convergence result for z_n. An essential ingredient of the following theorem is the *stage order* q of the method, i.e., condition $C(q)$ of Sect. II.7 or IV.5.

Theorem 1.1. *Suppose that the system (1.6) satisfies (1.7) in a neighbourhood of the exact solution $(y(x), z(x))$ and assume the initial values are consistent. Consider a Runge-Kutta method of order p, stage order q and with invertible matrix A. Then the numerical solution of (1.11a–d) has global error*

$$z_n - z(x_n) = \mathcal{O}(h^r) \quad \text{for} \quad x_n - x_0 = nh \le \text{Const}, \tag{1.25}$$

where

a) $r = p$ *for stiffly accurate methods,*
b) $r = \min(p, q+1)$ *if the stability function satisfies* $-1 \le R(\infty) < 1$,
c) $r = \min(p-1, q)$ *if* $R(\infty) = +1$.
d) If $|R(\infty)| > 1$, *the numerical solution diverges.*

Proof. Part (a) has already been discussed. For the remaining cases we proceed as follows: we first observe that Condition $C(q)$ and order p imply

$$z(x_n + c_i h) = z(x_n) + h \sum_{j=1}^{s} a_{ij} z'(x_n + c_j h) + \mathcal{O}(h^{q+1}) \tag{1.26a}$$

$$z(x_{n+1}) = z(x_n) + h \sum_{i=1}^{s} b_i z'(x_n + c_i h) + \mathcal{O}(h^{p+1}) \,. \tag{1.26b}$$

Since A is invertible we can compute $z'(x_n + c_j h)$ from (1.26a) and insert it into (1.26b). This gives

$$z(x_{n+1}) = \varrho z(x_n) + b^T A^{-1} \widehat{Z}_n + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}) \tag{1.27}$$

where $\varrho = 1 - b^T A^{-1} \mathbb{1} = R(\infty)$ and $\widehat{Z}_n = (z(x_n + c_1 h), \ldots, z(x_n + c_s h))^T$. We then denote the global error by $\Delta z_n = z_n - z(x_n)$, and $\Delta Z_n = Z_n - \widehat{Z}_n$. Subtracting (1.27) from (1.11d) yields

$$\Delta z_{n+1} = \varrho \Delta z_n + b^T A^{-1} \Delta Z_n + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}). \tag{1.28}$$