Stiff Systems

1 Introduction. What is stiffness?

The main purposes of the following lectures is twofold. Firstly, to discuss from a user´s point of view the concept of stiff problems, which appear so often in practical situations. Secondly, to give recipes that help to choose the right method to solve the stiff problem at hand. Following Lambert (1991) one should consider stiffness as a phenomenon exhibited by a system, rather than a property of it, because the word property is associated to the existence of a definition which is both comprehensive and precise, whereas it is difficult to come out with a satisfactory definition for the concept of "stiffness". A convenient and constructive way of introducing the discussion about stiffness is in relation with the concept of linear stabilty of a numerical method used to calculate the numerical solution of the well-posed initial value problem (IVP)

$$\begin{array}{c} y' = f(t,y),\ t \in (0,T], \\ y(0) = a,\ prescribed,. \end{array} \tag{1.1}$$

where $y \in \mathbf{C}^m,\ m$ being the dimension of the system. Generalizing the technique of the linear stability analysis of a numerical method, we substitute (1.1) by the problem

$$\begin{array}{c} y' = A(y - p(t)) + p'(t),\ t \in (0,T], \\ y(0) = a,\ prescribed, \end{array} \tag{1.2}$$

where $A = diag(a_{ii}),\ i = 1 \to m,\ a_{ii} \in \mathbf{C}$. Note that (1.2) is theoretically more relevant that apparently seems to be, because if we wish to study the solution of (1.1) near a particular solution $g(t)$ we should apply a Taylor series expansion and get

$$\begin{array}{c} y' = J(t,g(t))(y - g(t)) + f(t,g(t)) \\ = J(t,g(t))(y - g(t)) + g'(t), \end{array} \tag{1.3}$$

where $J(t, g(t))$ is the Jacobian matrix, which is assumed to be slowly varying in t, so that, locally, $J(t, g(t))$ can be taken as a constant. For the ease of the exposition, we shall further assume that J is diagonalizable; hence, after application of the diagonalization procedure to (1.3), we obtain system (1.2). We remark that the latter hypothesis on J is not restrictive, because in the general case J admits a Jordan canonical form, so that from (1.3) can be obtained (1.2) with the matrix A being now the Jordan canonical form of J.

The solution of (1.2) is

$$y(t) = (a - p(0))e^{At} + p(t) \tag{1.4}$$

The qualitative behavior of $y(t)$ depends mainly on A. Thus, we can distinguish three cases. i) If for all i, $\mathrm{Re}(a_{ii}) > 0$ and large, then the solution curves for various a will fan out as t increases and the problem will be difficult for any numerical method. We say that the problems is *unstable.* ii) If for all i, $\mathrm{Re}(a_{ii}) > 0$, but small, the problem can be easily handled for any conventional numerical method since the solution curves are more or less parallel. The IVP (1,1) is *neutrally stable.* iii) If for all i, $\mathrm{Re}(a_{ii}) < 0$ and there are i and j, $j \neq i$, such that

$$\frac{\mathrm{Re}(a_{jj})}{\mathrm{Re}(a_{ii})} \textit{ is small.}$$

then the solution will tend to $p(t)$ after a given time t called the *initial transient.* In this case the IVP is *stable.*

The characteristics that distinguish a stiff problem from a non-stiff one are encapsulated in the following statement (SG)

Statement 1 We say that a problem is stiff if the following conditions are fulfilled. A) No solution component is unstable, or equivalently, no eigenvalue of the Jacobian matrix has a real part which is at all large and positive, and at least some component is very stable, that is, at least one eigenvalue has a negative part which is negative and large. B) The solution is slowly varying with respect to the negative real part of the eigenvalues.

Some comments on the meaning of this statement are now in order.

1) Roughly speaking, condition B) of this statement means that the solution is smooth and the norm of its derivatives is much smaller than the norm of the derivatives of $\exp(At)$.

2) Implicitly recognized in Statement 1 is the fact that a problem may be stiff in some intervals of t and not in others. For example, if for all

i, $\mathrm{Re}(a_{ii}) << 0$ ($very$ negative) the problem will be stiff after the initial transient $\exp(At)$ has died out, but the problem is not stiff during the transient interval. This also applies to a linear problem with a constant Jacobian such that $\mathrm{Re}(a_{ii}) << 0$.

3) A stiff problem can exhibit several periods of rapid change. (or can have various rapid transients at different time intervals) because the term $p(t)$ may suddenly change.

4) A symptom of the potential presence of stiffness in an stable IVP (1.1) is the existence of components which change much faster than others, although we must point out that such a symptom is not necessarily an indication of stiffness, because the stiffness depends on the differential equation rather than the behavior of the solution itself.

1.1 Examples of stiff problems

The areas of chemical engineering, nonlinear mechanics, biochemistry and life sciences are sources of stiff problems.

1.1.1 Chemical reactions systems.

A famous chemical reaction is the **Oregenator** reaction between $HBrO_2$, Br^-, and $Ce(IV)$ described by Field and Noyes in 1984. The Oregenator is expressed mathematically by the following IVP

$$\begin{aligned} y_1' &= 77.27\left(y_2 + y_1\left(1 - 8.375 \times 10^{-6} y_1 - y_2\right)\right), \\ y_2' &= \frac{1}{77.27}\left(y_3 - \left(\left(1 + y_1\right) y_2\right), \\ y_3' &= 0.16(y_1 - y_3). \end{aligned} \tag{1.5}$$

The stiffness o (1.5) is due to the fast variation of components y_1 and y_3 as compared to y_2. If T is sufficiently large, the stiffness phenomena may appear several times in the interval $[0, T]$.

1.1.2 Reaction-diffusion systems

Problems in which the diffusion is modeled via the Laplace operator may become stiff as they are discretized in space by finite differences or finite

elements. A typical example of such systems which appear so often in mathematical biology is the following

$$\begin{array}{c} u_t = \nu u_{xx} + a + u(uv - (b+1)) \\ v_t = \nu v_{xx} + u(b - uv),\ 0 \le x \le 1,\ t \in (0, T], \end{array} \tag{1.6a}$$

with initial and boundary conditions

$$\begin{array}{c} u(x,0) = 1 + \sin(2\pi x),\ v(x,0) = c \\ u(0,t) = u(1,t) = 1;\ v(0,t) = v(1,t) = c, \end{array} \tag{1.6b}$$

where a, b and c are real constants, and ν is another positive constant which is called the diffusion coefficient. If we discretize the spatial derivatives by second order finite difference on a grid of I points $x_i = \dfrac{i}{I+1},\ 1 \le i \le I$, with spatial discretization parameter $h = \dfrac{1}{I+1}$, we obtain the following IVP

$$\begin{array}{c} u_i' = \dfrac{\nu}{h^2}(u_{i+1} - 2u_i + u_{i-1}) + a + u_i(u_i v_i - (b+1)), \\ v_i' = \dfrac{\nu}{h^2}(v_{i+1} - 2v_i + v_{i-1}) + u_i(b - u_i v_i) \\ u_i(0) = 1 + \sin(2\pi x_i),\ v_i(0) = c,\ i = 1 \to I, \\ u_0(t) = u_{I+1}(t) = d;\ v_0(t) = v_{I+i}(t) = f \end{array} \tag{1.7}$$

Setting $u := (u_1, ..., u_I)^T,\ v := (v_1, ..., v_I)^T$ and $y := (u, v)$, it follows from (1.7) that

$$y' = f(y),$$

where

$$f(y) := (f_1(u,v), f_2(u,v)) = \frac{\nu}{h^2} \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} R_1 \\ R_2 \end{bmatrix},$$

with

$$T = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & . & . & \\ & & . & -2 & 1 \\ & & & 1 & -2 \end{bmatrix},$$

$R_1 = diag(a + u_i(u_i v_i - (b+1))$ and $R_2 = diag(u_i(b - u_i v_i))$.

Thus, the Jacobian matrix $J = \dfrac{\partial(f_1, f_2)}{\partial(u, v)}$ is the sum of a diffusion matrix and a reaction matrix as

$$J = \frac{\nu}{h^2} \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix} + \begin{bmatrix} diag(2u_i v_i - b - 1) & diag(u_i^2) \\ diag(b - 2u_i v_i) & diag(-u_i^2) \end{bmatrix}$$

In many problems of interest the reaction matrix represents a perturbation to the real symmetric diffusion matrix whose eigenvalues are given by (Thomas)

$$\lambda_k = -4\frac{\nu}{h^2}(\sin\frac{k\pi}{2h})^2, 1 \leq k \leq I + 1.$$

Notice that λ_k takes values between 0 and $\dfrac{-4\nu}{h^2}$; so that, for h smaller than ν the stiffness of this problem is caused by the diffusion term.

2 Why is difficult to solve stiff problems?

In this section we shall analyze some of the reasons that make stiff problems be difficult for conventional explicit numerical methods. A first approach to measure the efficiency of a numerical method is to look at the accuracy and stability properties; so that, we shall examine the consequences of these two attributes as we solve stiff problems by conventional explicit methods. First, **accuracy**. To carry out our analysis, we consider the model problem (1.1) which, for ease of the exposition , is solved by the Euler explicit scheme with a prescribed **tolerance** ε. Assuming the solution is sufficiently smooth, we implement our computer code in such a way that it can select the time step length Δt in order to give a numerical solution with the prescribed accuracy. Therefore,

$$y_{n+1} = y_n + \Delta t f(t, y_n) = y_n + \Delta t y'_n. \tag{2.1}$$

Taking $y_n = y(t_n)$, the local truncation error (LTE) is expressed by

$$LTE = \frac{\Delta t^2}{2} y''(t_n) + O(\Delta t^3). \tag{2.2}$$

We further assume that the tolerance ε is such that

$$\varepsilon = \frac{\Delta t^2}{2} \parallel y''(t_n) \parallel,$$

hence

$$\Delta t_n \simeq \left(\frac{2\varepsilon}{\parallel y''(t_n) \parallel} \right)^{\frac{1}{2}}. \tag{2.3}$$

Clearly, using (1.4) in (2.3) we can distinguish the following two cases:
1) For t small, (2.3) yields

$$\Delta t_n \simeq \left(\frac{2\varepsilon}{\parallel (a - p(0))A^2 \parallel} \right)^{\frac{1}{2}}. \tag{2.4a}$$

2) For t large,(2.3) yields

$$\Delta t_n \simeq \left(\frac{2\varepsilon}{\parallel p''(t_n) \parallel} \right)^{\frac{1}{2}}, \tag{2.4b}$$

because the exponential becomes very small. Since p'' is small and $\parallel A^2 \parallel$ is large, (2.4) is telling us that in stiff problems we can achieve the prescribed accuracy using small time steps Δt during the initial transient, and large time steps after the initial transients have died out. Next, let us examine **stability.** We recall that stability is the property of a numerical method to keep the errors bounded as the calculation advances. Therefore, if we consider the global error at time t_n, $e_n := y_n - y(t_n)$, it follows for the Euler method that

$$e_{n+1} = (1 + \Delta t_n A)e_n + LTE.$$

Hence, in order to keep $\parallel e_n \parallel$ bounded we have that

$$\parallel 1 + \Delta t_n A \parallel \leq 1, \quad or \quad -2 \leq \Delta t_n \parallel A \parallel \leq 0. \tag{2.5}$$

This means that for stiff problems, for which $\parallel A \parallel$ is too large, the selection of the length of the time step in an explicit method is made by the stability restriction, which, outside the transient region, imposes the use of a Δt that is clearly inefficient. But, recalling the definitions of the region of absolute stability $\mathbf{S}$ of a method and its associated stability function $R(z)$,i.e.,

$$\mathbf{S} := \{z \in \mathbf{C} :\mid R(z) \mid \leq 1\}, \tag{2.6}$$

we see that (2.5) means that Δt_n is chosen in such way that $z = \Delta t_n \max \mid \lambda_i \mid$, $\lambda_i \in \mathbf{C}$ being the eigenvalues of A, is in $\mathbf{S}$. Since $\mathbf{S}$ is small for explicit methods and large for the implicit ones, then another heuristic way to characterize stiff problems is the following

Statement 2. Stiff problems are those for which explicit method are inefficient.

Hence, the following recommendation makes sense

Recommendation 1. In general, for a stiff problem it is better to use implicit schemes with time step selection

3 Linear stability definitions pertinent to stiffness

From the previous analysis on linear stability and stiff problems, we come to the conclusion that if the method employed to integrate a stiff problem has an absolute stability region **S** extending the whole left-half plane, then there will not be any stability restriction on the time step, and therefore, we can select the time steplength based on accuracy considerations. At this point, it is convenient to characterize the linear stabilty requirements for stiff problems. Following the linear stability analysis of previous lectures, we consider the model problem (Dahlquist, 1963)

$$y' = \lambda y,\ \lambda \in \mathbf{C},\ \mathrm{Re}(\lambda) \leq 0. \tag{3.1}$$

and set $z = \Delta t \lambda$.

A-STABILITY

A method is said to be **A-stable** if $\mathbf{S} \supseteq C^- := \{z : \mathrm{Re}(z) \leq 0\}$

A-stability is a strong requirement, in particular for linear multi-step methods, so that it is natural to restrict the class of problems in some way and seek alternative definitions of stability that remove the restriction on the time steplength for that class of problems.

A(α)-STABILITY

A method is said to be $\mathbf{A}(\alpha)-$**stable**, $\alpha \in (0, \pi/2)$, *if* $\mathbf{S} \supseteq \{z : -\alpha < \pi - \arg z < \alpha\}$; *it is said to be* **A(0)-stable** *if it is A(α)-stable for some* $\alpha \in (0, \pi/2)$.

$\mathbf{A_0}$-STABLE

A method is said to be $\mathbf{A_0}-$**stable** *if* $\mathbf{S} \supseteq \{z : \mathrm{Re}(z) < 0,\ \mathrm{Im}(z) = 0\}$.

Under the observation that for many problems the eigenvalues responsible of the fastest transients all lie to the left of a line $\mathrm{Re}(z) = -a$, where a is positive, and the remaining eigenvalues, responsible of the slower transients, will have small negative real parts and be clustered close to the origin , Gear (1969) gives the following definition

STIFFLY STABLE

Let R_1 and R_2 be two regions of the complex plane defined as $R_1 := \{z : \mathrm{Re}(z) < -a\}$, $R_2 := \{z : -a \leq \mathrm{Re}(z) < 0,\ -c \leq \mathrm{Im}(z) \leq c\}$, where a and c are positive constants. A method is said to be **stiffly stable** *if* $\mathbf{S} \supseteq R_1 \cup R_2$.

There are methods with a rational stability function $R(z)$, for which A-stability is not as desirable a property as it seems to be, because for rational stability functions $R(z)$ we have that for z real and very negative, $\mid R(z) \mid$ is less than 1 but very close to 1, so that the stiff components are damped out very slowly. This motivates the following definition

L-STABILITY

A method is said to be **L-stable** *if is A-stable and if in addition*

$$\lim R(z) \to 0,\ as\ z \to \infty$$

This property is sometimes called *stiff A-stability* or *strong A-stability.* It is worth noting the stability hierarchy

$$\boxed{\begin{array}{c} L - stability \to A - stability \to stiff\ stability \\ \to A(\alpha) - stability \to A(0) - stability \to A_0 - stability \end{array}} \tag{3.2}$$

In the next sections, we shall examine some relevant implications of the stability hierarchy on members of the families of Runge-Kutta and multi-step methods. This will help to establish practical criteria to choose the good method for the integration of stiff problems

4 A-stability and Runge-Kutta methods

The important fact we must point out is given in the following statement (HW II)

Statement 3. No explicit Runge-Kutta method is A-stable

Let us examine then some interesting relations between implicit methods and A-stability. First, we recall the general formula of implicit Runge-Kutta methods of $s-stages$.

$$\begin{array}{c} Z_i = y_n + \Delta t \sum_{j=1}^{s} a_{ij} f(t_n + c_J \Delta t, Z_j), \;\; 1 \le j \le s, \\ y_{n+1} = y_n + \Delta t \sum_{i=1}^{s} b_i f(t_n + c_i \Delta t, Z_i), \end{array} \tag{4.1a}$$

or using the Butcher tableau

$$\begin{array}{|c|c|}\hline c & A \\ \hline & b^T \\ \hline \end{array} := \begin{array}{|c|c|c|c|c|c|}\hline c_1 & a_{11} & a_{12} & \cdot & \cdot & a_{1s} \\ \hline c_2 & a_{21} & a_{22} & \cdot & \cdot & a_{2s} \\ \hline \cdot & \cdot & \cdot & \cdot & & \cdot \\ \hline \cdot & \cdot & \cdot & & \cdot & \cdot \\ \hline c_s & a_{s1} & a_{s2} & \cdot & \cdot & a_{ss} \\ \hline & b_1 & b_2 & \cdot & \cdot & b_s \\ \hline \end{array} \tag{4.1b}$$

To construct IRK the following assumptions are used (HWN I)

$$\begin{array}{c} B(p) : \sum_{i=1}^{s} b_i c_i^{q-1} = \dfrac{1}{q}, \;\; 1 \le q \le p, \\ C(\eta) : \sum_{j=1}^{s} a_{ij} c_j^{q-1} = \dfrac{c_i^q}{q}, \;\; 1 \le i \le s, \;\; 1 \le q \le \eta, \\ D(\zeta) : \sum_{j=1}^{s} b_i c_i^{q-1} a_{ij} = \dfrac{b_j}{q}(1 - c_j^q), \;\; 1 \le j \le s, \;\; 1 \le q \le \zeta \end{array} \tag{4.1c}$$

The relevance of these conditions is established in a theorem due to Butcher (1965) which says if the coefficients $b_i, \;\; c_i, \;\; a_{ij}$ of a RK method satisfy $B(p), \;\; C(\eta)$ and $D(\zeta)$ with $p \le \eta + \zeta + 1$ and $p \le 2\eta + 2$, then the method is of order p

It is easy to see that the stability function of the IRK methods of $s-stages$ is

$$R(z) = 1 + z b^T (I - zA)^{-1} \mathbf{1}, \;\; \mathbf{1} = (1, ..^{(s)}., 1)^T,$$

which can be written as a rational function (HW II)

$$R(z) = \frac{\det(I - zA + z\mathbf{1}b^T)}{\det(I - zA)} \tag{4.2a}$$

It is usual to write $R(z)$ as

$$R(z) = \frac{P(z)}{Q(z)}, \tag{4.2b}$$

with $P(z)$ and $Q(z)$ polynomials of degree $\leq s$, $\deg P = k$ and $\deg Q = j$. We have the following properties (HW II)

Proposition i) The IRK (4.1) is $A-stable$ if and only if

$$\begin{array}{l} for\ all real\ y, \mid R(iy) \mid \leq 1,\ and \\ R(z)\ is\ analytic\ for\ R(z) < 0. \end{array} \tag{4.3}$$

ii) If the matrix A of the IRK (4.1) is non singular and

$$a_{sj} = b_j,\ j = 1, ..., s, \tag{4.4a}$$

and

$$a_{i1} = b_1,\ i = 1, ..., s, \tag{4.4b}$$

then the method is $L-stable$, that is, $\lim R(z) \rightarrow 0$ as $z \rightarrow 0$. The method is $stiffly\ accurate$ if (4.4b) is satisfied..

Next, we formulate the most frequently used IRK methods in stiff system codes

LOW ORDER IRK METHODS

1) *Implicit Euler (s=1). Order 1*

$$y_{n+1} = y_n + \Delta f(t_{n+1}, y_{n+1}) \tag{4.5a}$$

$$R(z) = \frac{1}{1-z} \tag{4.5b}$$

Note that this method is also $L-stable$.

2) *Mid-point rule (s=1). Order 2*

$$\begin{array}{l} Z_1 = y_n + \dfrac{\Delta t}{2} f(t_n + \dfrac{\Delta t}{2}, Z_1), \\ y_{n+1} = y_n + \Delta t f(t_n + \dfrac{\Delta t}{2}, Z_1) \end{array} \tag{4.6a}$$

$$R(z) = \frac{1 + z/2}{1 - z/2} \tag{4.6b}$$

This method is $A-stable$. Note that the method is not $L-stable$.

HIGHER ORDER METHODS

There are many IRK methods of order higher than 2, the so called **Gauss and Radau methods** are frequently used in codes designed for stiff problems due to the following property (HW-II)

Theorem *i) The s-stage Gauss method is A-stable and of order 2s, with stability function R(z) given by the (s,s) Padé approximation to exp(z). ii) The s-stage Radau IA and Radau IIA methods are A-stable and of order 2s-1, with stability function given by the (s-1,s) subdiagonal Padé approximation to exp(z).*

The interesting fact of this theorem is that Gauss and Radau methods represent a good balance between computational effort and order of the method, for with a low number of stages the methods may achieve a high order accurate numerical solution. We shall write down the Butcher tableau of some of these methods which are used in canned codes for stiff problems. But before doing so, we give a brief idea of how these methods are constructed. For this purpose, we recall that if we apply a general Runge-Kutta method to the scalar equation $y' = f(t)$, we obtain the quadrature formula

$$\int_{t_n}^{t_{n+1}} f(t)dt \simeq y_{n+1} - y_n = \Delta t \sum_{j=1}^{s} b_j f(t_n + c_j \Delta t),$$

which in essence is a Gaussian quadrature formula with weights b_j and integration points (abscissae) $t_n + c_j \Delta t$. Computing the coefficients c_j, $1 \leq j \leq s$, as the zeros of the shifted Legendre polynomial of degree s

$$\frac{d^s}{dt^s}(t^s(t-1)^s)$$

the result is *the Gauss method of order s*. But if the coefficients c_j are computed as the zeros of

$$\frac{d^s - 1}{dt^{s-1}}(t^s(t-1)^{s-1}),\ (Radau\ I),\ or$$
$$\frac{d^{s-1}}{dt^{s-1}}(t^{s-1}(t-1)^s),\ (Radau\ II),$$

and the weights $b_1,\ 1 \le i \le s,$ are chosen such that condition $B(2s-1)$ in (4.1c) is satisfied, the result that follows is *the methods Radau I and Radau II.*

1) *Gauss method of s=2 and order 4.*
The Butcher tableau of this method is:

$\frac{3-\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{3-2\sqrt{3}}{12}$
$\frac{3+\sqrt{3}}{6}$	$\frac{3+2\sqrt{3}}{12}$	$\frac{1}{4}$
	$\frac{1}{2}$	$\frac{1}{2}$

(4.7a)

with

$$R(z) = \frac{1 + \frac{1}{2}z + \frac{1}{6}\frac{z^2}{2_{\mathrm{i}}}}{1 - \frac{1}{2}z + \frac{1}{6}\frac{z^2}{2_{\mathrm{i}}}} \tag{4.7b}$$

2) *Gauss method of s=3 and order 6*
The Butcher tableau of this method is

$\frac{5-\sqrt{15}}{10}$	$\frac{5}{36}$	$\frac{10-3\sqrt{15}}{45}$	$\frac{25-6\sqrt{15}}{180}$
$\frac{1}{2}$	$\frac{10+3\sqrt{15}}{72}$	$\frac{2}{9}$	$\frac{10-3\sqrt{15}}{72}$
$\frac{5+\sqrt{15}}{10}$	$\frac{25+6\sqrt{15}}{180}$	$\frac{10+3\sqrt{15}}{45}$	$\frac{5}{36}$
	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

(4.8a)

with

$$R(z) = \frac{1 + \frac{1}{2}z + \frac{1}{5}\frac{z^2}{2_{\mathrm{i}}} + \frac{1}{20}\frac{z^3}{3_{\mathrm{i}}}}{1 - \frac{1}{2}z + \frac{1}{5}\frac{z^2}{2_{\mathrm{i}}} - \frac{1}{20}\frac{z^3}{3_{\mathrm{i}}}} \tag{4.8b}$$

As a remark, note that the mid-point rule is a Gauss method with s=1.

Radau I and II of s=2 and order 3

$$\begin{array}{|c|cc|}\hline 0 & \frac{1}{4} & -\frac{1}{4} \\ \hline \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \\ \hline \end{array} \ (Radau\ I), \qquad \begin{array}{|c|cc|}\hline \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ \hline 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \\ \hline \end{array} \ (Radau\ II) \tag{4.9}$$

ROSENBROCK METHODS

These methods are very competitive for stiff systems of moderate size, for example, systems of dimension ≤ 10, and when the required accuracy is not

too high, say $10^{-4} - 10^{-5}$. Of course, one can use Rosenbrock methods with conditions that are more demanding, but in such cases the methods will be less efficient. The most attractive feature of these methods is that they are as easy to implement as explicit RK methods. We give Kaps and Rentrop formulation of an $s - stage$ Rosenbrock method to integrate (1.1) Thus an $s-stage$ *Rosenbrock method* is given by the formula

$$\begin{aligned} k_i = \Delta t f(t_n + \alpha_i \Delta t, y_n + \textstyle\sum_{j=1}^{i-1} \alpha_{ij} k_j) + \gamma_i \Delta t^2 \dfrac{\partial f}{\partial t}(t_n, y_n) \\ + \Delta t \dfrac{\partial f}{\partial y}(t_n, y_n) \textstyle\sum_{j=1}^{i} \gamma_{ij} k_j,\ i = 1, ..., s \\ y_{n+1} = y_n + \textstyle\sum_{j=1}^{s} b_j k_j., \end{aligned} \tag{4.10a}$$

where the coefficients α_i and γ_i satisfy the relations

$$\alpha_i = \sum_{j=1}^{i-1} \alpha_{ij} \ and \ \gamma_i = \sum_{j=1}^{i} \gamma_{ij} \tag{4.10b}$$

Note that at each stage of these methods a linear system of equations as

$$\left(I - \Delta t \gamma_{ii} \frac{\partial f}{\partial y}(t_n, y_n) \right) k_i = R_i \tag{4.10c}$$

has to be solved. Of special interest are those method which chose the coefficients $\gamma_{ii} = \gamma$ for all i, because in such case only one LU decomposition is necessary per step.

Kaps and Rentrop implementation (4.10a)-(4.10c) is via an stepsize control-embedded formula with $s = 4$ for the high order solution and $s = 3$ for the low order solution. For further details on the order conditions, stability and coding of Rosenbrock methods see HW II.

5 A-stability and multistep methods

We recall that a general multistep method of *k-steps* is formulated as

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \Delta t \sum_{j=0}^{k} \beta_j f_{n+j}, \tag{5.1}$$

where $f_m := f(t_m, y_m)$. The linear stability analysis of the method is performed by applying (5.1) to the test problem

$$y' = \lambda y,$$

yielding

$$\sum_{j=0}^{k} (\alpha_j - \mu\beta_j) y_{n+j} = 0, \quad \mu := \lambda \Delta t. \tag{5.2}$$

To solve (5.2) we use the Lagrange method, setting $y_j = \zeta^j$, dividing by ζ^n and obtaining

$$\sum_{j=0}^{k} (\alpha_j - \mu\beta_j)\zeta^j \equiv \rho(\zeta) - \mu\sigma(\zeta) = 0 \tag{5.3a}$$

with (the already familiar polynomials)

$$\begin{array}{l} \rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \\ \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j. \end{array} \tag{5.3b}$$

The difference equation (5.2) has stable solutions for arbitrary starting values, if and only if all roots of (5.3a) are ≤ 1 in modulus. So that, we define the absolute stability region $\mathbf{S}$ of (5.1) as

$$\mathbf{S} = \left\{ \mu \in \mathbf{C} : \begin{array}{c} all\ roots\ \zeta_i\ of\ (5.3a)\ satisfy\ \mid \zeta_i(\mu) \mid \leq 1, \\ multiple\ roots\ satisfy\ \mid \zeta_i(\mu) \mid < 1 \end{array} \right\} \tag{5.4}$$

Considering (5.4), we particularize the general definition of A-stability to multi-step methods as follows

Lemma *The multistep method (5.1) is A-stable if for all $\mu \in C^-$ it holds*
i) all roots of (5.3a) satisfy

$$\mid \zeta_i(\mu) \mid \leq 1,$$

ii) and multiple roots satisfy

$$\mid \zeta_i(\mu) \mid < 1.$$

The Adams and the Predictor-Corrector families of multistep methods are the most popular multistep methods, but they possess the sad property that

none of them is A-stable (so that, they are no convenient for stiff problems), except the implicit Adams method of order 2,also known as **trapezoidal rule**

$$y_{n+1} = y_n + \frac{\Delta t}{2}[f_n + f_{n+1}] \tag{5.5}$$

This observation was taken to the category of theorem by Dahlquist in 1963. (See HW II)

Theorem *An A-stable multi-step method must be of order $p \leq 2$. If the order is 2, then the order constant satisfies*

$$C \leq -\frac{1}{12}.$$

The trapezoidal rule is the only A-stable multi-step method of order 2 with $C = \frac{1}{12}$.

This theorem is also known with the descriptive name of **Dahlquist second barrier**, and it seems to convey the rather deceptive message that multistep methods are inferior to RK methods as far as A-stability is concerned, and therefore, one may arrive to the conclusion that multistep methods are not good for stiff problems. However, this conclusion is not entirely true, because one can "break" the barrier either i) by weaken the condition of A-stabilty, or ii) by strengthen the method. Weaken the condition of A-stability has led us to introduce different definitions of stability in section 3, which are relevant to many stiff problems. This opens the possibility of using a family of multistep methods, known as **BDF** methods, for stiff problems.

THE BACKWARD DIFFERENTIATION FORMULAE (BDF)

The general expression of these formulae is (HWN I)

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \Delta t \beta_k f_{n+k} \tag{5.6}$$

Two important properties of BDF are given in the following propositions

Proposition 1 *The BDF with $k \leq 6$ are all of them stiffly stable, with the following values of a (in the stiffly stability definition) and $\alpha($ in the*

$A(\alpha)$-stability definition)

k	1	2	3	4	5	6
α	90^o	90^o	86.03^o	73.35^o	51.84^o	17.84^o
a	0	0	0.083	0.667	2.327	6.075

(5.7)

Note that for $k = 1$ and 2, the BDF are also $A - stable$, because $a = 0$ and $\alpha = \frac{\pi}{2}$ imply that $\mathbf{S} \supseteq \mathbf{C}^-$

Proposition 2 *.For $k \geq 7$, all BDF are unstable.*

For completeness of this notes we give next the table with the coefficients and the order constant of the stable BDF.

k	α_6	α_5	α_4	α_3	α_2	α_1	α_0	β_1	p	C_p
1						1	-1	1	1	$-\frac{1}{2}$
2					1	$-\frac{4}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	2	$-\frac{2}{9}$
3				1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$	$\frac{6}{11}$	3	$-\frac{3}{22}$
4			1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$	$\frac{12}{25}$	4	$-\frac{12}{125}$
5		1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$	$\frac{60}{137}$	5	$-\frac{10}{137}$
6	1	$-\frac{360}{147}$	$-\frac{72}{147}$	$-\frac{450}{147}$	$-\frac{400}{147}$	$-\frac{225}{147}$	$-\frac{10}{147}$	$\frac{60}{147}$	6	$-\frac{20}{343}$

(5.8)

6 Characteristics of the solution methods

In previous sections, we have introduced a list of methods which are good to compute the numerical solution of stiff systems. All those methods share the property of being **implicit**, this means that an equation of the form

$$w = \Delta t g(w) + \Psi(y_{n-s}; t_{n-s}, \Delta t), \tag{6.1}$$

where the vectors $g(w)$ and $\Psi(y_{n-s}; t_{n-s}, \Delta t)$ are known and g is also differentiable. If the problem is linear (that is, if g is linear), then a linear equation must be solved, but if the problem is nonlinear a nonlinear equation must be solved at each step. In the latter case, we must consider wether the problem is stiff or not. For non stiff problems an efficient method to solve (6.1) may be the functional iteration

$$w^{(k+1)} = \Delta t g(w^{(k)}) + \Psi \tag{6.2a}$$

that converges if and only if

$$\| \Delta t \frac{\partial g}{\partial w} \| < 1. \tag{6.2b}$$

This condition, which is equivalent to the absolute stability condition of the numerical method since $\frac{\partial g}{\partial w}$ is the Jacobian matrix, is not restrictive for non-stiff problems because the norm of the Jacobian matrix is not large; however, this argument shows that for stiff problems this procedure is inefficient because $\| \frac{\partial g}{\partial w} \|$ is very large. So that when (6.1) represents a stiff problem the Newton-Raphson algorithm or some of its variants is currently used to calculate a numerical solution of (6.1). The Newton-Raphson algorithm as used in many codes is

$$[I - \Delta t J(w^{(k)})] \widetilde{\Delta} w^{(k)} = -w^{(k)} + \Delta t g(w^{(k)}) + \Psi, \tag{6.3}$$

where

$$I \; is \; the \; unit \; matrix, \quad J(w^{(k)}) = \frac{\partial g(w^{(k)})}{\partial w},$$
$$and \; \widetilde{\Delta} w^{(k)} = w^{(k+1)} - w^{(k)}.$$

Note that in terms of CPU time and computer storage, (6.3) is much more expensive to carry out than (6.2a), because at each iteration we have a) to compute and storage the Jacobian matrix, and b) to solve a linear system of equations whose matrix is, in general, nonsymmetric and full. Despite these unavoidable tasks, the Newton-Raphson algorithm is in general (for stiff problems) more efficient than functional iteration for the following reasons: (i) as we mention above, (see (6.2a)), the size of Δt has to be much smaller in the functional iteration than in the Newton-Raphson algorithm; and (ii) the rate of convergence of the Newton Raphson algorithm is higher than that of the functional iteration if a good starting is available. In fact, it can be proved (IK) that as long as Δt is sufficiently small the rate of convergence of the Newton-Raphson algorithm is quadratic, that is, there exists a positive constant K such that if $\widetilde{w}$ is the exact solution of (6.1) then

$$\| w^{(k+1)} - \widetilde{w} \| \leq K \| w^{(k+1)} - \widetilde{w} \|^2, \; for \; k \; sufficiently \; large,$$

whereas the rate of convergence of the functional iteration is linear, that is, there exists a positive constant K such that

$$\| w^{(k+1)} - \widetilde{w} \| \leq K \| w^{(k+1)} - \widetilde{w} \|, \; for \; k \; sufficiently \; large.$$

(i) implies that the Newton-Raphson algorithm needs take less time steps to compute the solution at time, say t=T, whereas (ii) means that the Newton-Raphson algorithm needs less iterations per step to compute the solution with a prescribed accuracy ε. So that, from (i) and (ii) we can conclude that the Newton-Raphson algorithm is more efficient than the functional iteration for a large number of strong stiff problems. Unfortunately, one can come across stiff problems for which the use of the Newton-Raphson algorithm may represent in several respects a heavy burden due, in particular, to the computation and storage the Jacobian matrix at each iteration. In such cases, a variant of the Newton-Raphson algorithm is employed, usually a linear one that consists in replacing at each step the Jacobian $J(w^{(k)})$ by $J(w^{(0)})$, this means that the Jacobian is calculated once per step (or may be for several steps). In doing so, we can see after some algebraic manipulation and setting $J_0 := J(w^{(0)})$ that (6.3) becomes

$$w^{(k+1)}) = (I - \Delta t J_0)^{-1} \Delta t (g(w^{(k)}) - J_0 w^{(k)}), \tag{6.4a}$$

which is a functional iteration with

$$\widetilde{g}(w) = (I - \Delta t J_0)^{-1} (g(w^{(k)}) - J_0 w^{(k)}). \tag{6.4b}$$

Of course, (6.4a) does not possess the quadratic convergence of the true Newton-Raphson algorithm (6.3), but considering that $g(w)$ is sufficiently smooth and applying the mean value theorem it is shown that the rate of convergence of (6.4a) is

$$\| w^{(k+1)} - \widetilde{w} \| \leq \frac{\| J(z) - J_0 \|}{\| (I - \Delta t J_0)^{-1} \|} \| w^{(k)} - \widetilde{w} \|,$$

where $z \in (w^{(0)}, w^{(k)})$. Unless the Jacobian matrix varies strongly with time, the term $\| J(z) - J_0 \|$ is likely to be small. On the other hand, for stiff problems, $\| (I - \Delta t J_0)^{-1} \|$ is likely also to be small, so that stiffness will help to make $\dfrac{\| J(z) - J_0 \|}{\| (I - \Delta t J_0)^{-1} \|}$ small, contributing in this way to accelerate the rate of convergence. There are still one more point to comment on. which is related with the fact that in the Newton-Raphson algorithm or any of its variants a linear system has to be solved. The thumb rule is that if such a system is not too large LU decomposition is a good choice as a solver, but for midsize or large systems the GMRES iterative method is becoming the favorite solver nowadays.

7 Basic Bibliography

[HNW] Hairer, E., S.P. Norsett and G. Wanner (1987): *Solving Ordinary Differential Equations I (Nonstiff Problems)* Springer- Verlag

[HW] Hairer, E. and G. Wanner (1991): *Solving Ordinary Differential Equations II (Stiff and Differential-Algebraic Problems)* Springer- Verlag

[IK] Isaccson, E. and H.Keller (1966) *Analysis of Numerical Methods.* Dover.

[L] Lambert, J. (1991) *Numrerical Methods for Ordinary Differential Systems.* John Wiley and Sons. Chichester.

[SG] Shampine, L.F. and W. Gear. (1979) *A user´s view of solving stiff ordinary differential equations.* SIAM Review. 21, 1, 1-17.