Norwegian University of Science and Technology
Department of Mathematical Sciences

Contact during the exam:
Elena Celledoni, tlf. 73593541, cell phone 48238584

Exam in TMA4215
December 7th 2012

Allowed aids code C: Textbook Endre Süli and David Mayers, An introduction to Numerical Analysis. TMA4215 lecture notes (61 pages). Rottman. Photocopies from the textbook are also allowed instead of the book itself, but they should be kept separate from the note of the course to allow control.

Itemized description of the learning outcome

- L1 approximation of functions;
- L2 numerical quadrature;
- L3 odes;
- L4 Linear and nonlinear equations;
- L5 error analysis in general;
- L6 analysis of algorithms and methods;
- L7 implementation;
- L8 design of numerical experiments (tested in the project work);
- L9 interpretation of the numerical results (tested in the project work);
- L10 usage of precise mathematical language to describe solution to the problems and findings in the project.

Problem 1 Consider $f \in C^{(4)}([-a,a])$ and let $p_3(x)$ be the interpolation polynomial of degree 3 satisfying

$$p_3(-a) = f(-a), \quad p_3(a) = f(a), \quad p_3'(-a) = f'(-a), \quad p_3'(a) = f'(a).$$

Show that if $M_4 = \max\limits_{-a \le x \le a} |f^{(4)}(x)|$, then

$$|f(x) - p_3(x)| \le \frac{a^4}{24} M_4.$$

Solution This is Hermite interpolation with $n = 1$, from the theorem about the error of Hermite interpolation (page 190 in Süli and Mayers) we see that the exact expression for the error is

$$f(x) - p_3(x) = \frac{f^4(\xi)}{4!}[(x-a)(x+a)]^2.$$

We get

$$|f(x) - p_3(x)| \leq \frac{M_4}{24} \max_{x\in[-a,a]} [x^2 - a^2]^2$$

finding the maximum on $[-a, a]$ of the polynomial $[x^2 - a^2]^2$ we obtain the result.

Tested learning outcome: L1, L5, L10.

Problem 2 Given the distinct absissae x_i, $i = 0, 1, \dots, n+1$, and the values y_i, $i = 0, 1, \dots, n+1$, let q be the interpolation polynomial of degree n for the set of points $\{(x_i, y_i) : i = 0, 1, \dots, n\}$ and let r be the interpolation polynomial of degree n for the points $\{(x_i, y_i) : i = 1, 2, \dots, n+1\}$. Define

$$p(x) = \frac{(x - x_0)r(x) - (x - x_{n+1})q(x)}{x_{n+1} - x_0}.$$

Show that p is the interpolation polynomial of degree $n+1$ for the points $\{(x_i, y_i) : i = 0, 1, \dots, n+1\}$.

Solution We verify that: $p(x_0) = q(x_0) = y_0$, $p(x_{n+1}) = r(x_{n+1}) = y_{n+1}$ and for x_i with $i = 1, \dots, n$

$$p(x_i) = \frac{(x_i - x_0)r(x_i) - (x_i - x_{n+1})q(x_i)}{x_{n+1} - x_0} = \frac{(x_i - x_0)y_i - (x_i - x_{n+1})y_i}{x_{n+1} - x_0} = y_i.$$

So obviously p interpolates $y_0, \dots, y_{n+1}$ on $x_0, \dots, x_{n+1}$, and since the interpolation polynomial through $n+2$ distinct points is unique, from the theorem of existence and uniqueness of the interpolation polynomial, p must be a polynomial of degree $n+1$.

Tested learning outcome: L1, L5, L10

Problem 3 Write down the errors in the approximation of

$$\int_0^1 x^4 dx \quad \text{and} \quad \int_0^1 x^5 dx$$

by the trapezium rule and the Simpson's rule (page 202 and 203 in the textbook). Use the exact values of the two integrals. Hence find the value of the constant C for which the trapezium rule gives the correct result for the calculation of

$$\int_0^1 (x^5 - Cx^4) dx,$$

and show that the trapezium rule gives a more accurate result than the Simpson's rule when $\frac{15}{14} < C < \frac{85}{74}$.

Solution The values of the two integrals are respectively 1/5 and 1/6. If we approximate both the two integrals with the trapezium rule we get in both cases the value 1/2 as approximation. So for the trapezium rule we get the two errors

$$|\frac{1}{5}-\frac{1}{2}|, \quad |\frac{1}{6}-\frac{1}{2}|,$$

and one proceeds similarly for the Simpson rule. We also have

$$\int_0^1 (x^5 - Cx^4)dx = \frac{5-6C}{30},$$

and approximating with the trapezium rule the same integral we get

$$\int_0^1 (x^5 - Cx^4)dx \approx \frac{1}{2} - C\frac{1}{2}.$$

So we get

$$\frac{5-6C}{30} = \frac{1}{2} - C\frac{1}{2},$$

when $C = \frac{10}{9}$. Using Simpson to approximate the same integral we obtain

$$\int_0^1 (x^5 - Cx^4)dx \approx \frac{1}{6}\left(\frac{9-10C}{8}\right).$$

Let us call I the exact value of the integral, T the approximation due to the trapezium rule and S the one due to the Simpson rule, then we have

$$I - T = \frac{-10+9C}{30}$$

and the trapezium formula gives the exact value of the integral when $C = \frac{10}{9}$.

We also have

$$I - S = \frac{-5+2C}{240},$$

and both $I - T$ and $I - S$ are linear functions of C. We have to find the values of C such that $|I - T| \leq |I - S|$.

The two functions are plotted in figure 1: $|I - T|$ as a function of C decreases for values of $C \leq \frac{10}{9}$, and increases for $C > \frac{10}{9}$. $|I - S|$ has a similar behaviour, and is zero in $C = \frac{5}{2}$. It suffices to find the points of intersection of the two graphs. It turns out that the graph of $|I - T|$ intersects $|I - S| = S - I$ for $C < \frac{5}{2}$, and $S - I$ coincides with the line through the two points $(-5/240, 0)$ and $(5/2, 0)$ for $C < \frac{5}{2}$. This line intersects $|I - T|$ in two points corresponding to the values $C = \frac{15}{14}$ and $C = \frac{85}{74}$. So $|I - T| \leq |I - S|$ for $\frac{15}{14} \leq C \leq \frac{85}{74}$.

Tested learning outcome: L2, L5, L6, L10.

Problem 4 Apply the implicit Runge-Kutta method

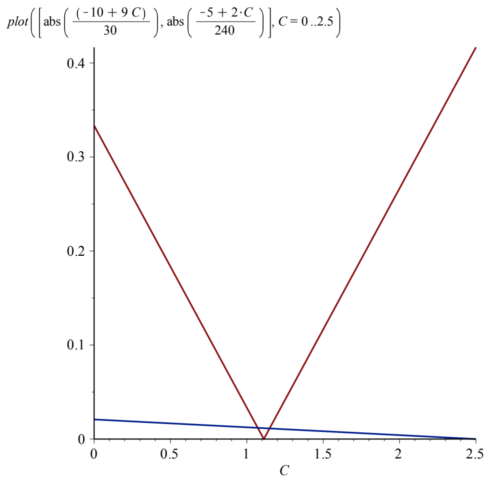

Figure 1: The two functions $|I - T|$ (in red) and $|I - S|$ (in blue) as functions of C.

$$\begin{array}{c|cc} \frac{1}{6}(3-\sqrt{3}) & \frac{1}{4} & \frac{1}{12}(3-2\sqrt{3}) \\ \frac{1}{6}(3+\sqrt{3}) & \frac{1}{12}(3+2\sqrt{3}) & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

to the initial value problem

$$y' = f(t, y), \quad y(t_0) = y_0,$$

with time step Δt. Derive the equations giving rise to the method and discuss the implementation tasks to be performed at each time-step.

Solution The Runge-Kutta method has two stages Y_1 and Y_2 and they are obtained as the solution of the equations[1]:

$$\begin{aligned} Y_1 &= y_0 + \Delta t \left(\frac{1}{4} f\big(t_0 + \frac{3-\sqrt{3}}{6}\Delta t, Y_1\big) + \frac{1}{12}(3-2\sqrt{3}) f\big(t_0 + \frac{3+\sqrt{3}}{6}\Delta t, Y_2\big) \right) \\ Y_2 &= y_0 + \Delta t \left(\frac{1}{12}(3+2\sqrt{3}) f\big(t_0 + \frac{3-\sqrt{3}}{6}\Delta t, Y_1\big) + \frac{1}{4} f\big(t_0 + \frac{3+\sqrt{3}}{6}\Delta t, Y_2\big) \right). \end{aligned}$$

To solve these equations we can use a fixed point iteration or a Newton method. With a fixed point iteration the procedure becomes:

Initialization

$Y_1^0 = y_0$, $Y_2^0 = y_0$,

$k = 0$

Iteration

while ($\varepsilon \leq TOL$ and $k \leq 100$)

$Y_1^{old} = Y_1^k$

$Y_2^{old} = Y_2^k$

$$\begin{aligned} Y_1^{k+1} &= y_0 + \Delta t \left(\frac{1}{4} f\big(t_0 + \frac{3-\sqrt{3}}{6}\Delta t, Y_1^k\big) + \frac{1}{12}(3-2\sqrt{3}) f\big(t_0 + \frac{3+\sqrt{3}}{6}\Delta t, Y_2^k\big) \right) \\ Y_2^{k+1} &= y_0 + \Delta t \left(\frac{1}{12}(3+2\sqrt{3}) f\big(t_0 + \frac{3-\sqrt{3}}{6}\Delta t, Y_1^k\big) + \frac{1}{4} f\big(t_0 + \frac{3+\sqrt{3}}{6}\Delta t, Y_2^k\big) \right). \end{aligned}$$

$k = k + 1$

$\varepsilon = \|Y_1^k - Y_1^{old}\|_2 + \|Y_2^k - Y_2^{old}\|_2$

end while

[1] The RK-method and the corresponding equations can be also formulated by means of the unknowns $K_i = f\big(t_0 + c_i\Delta t, y_0 + \Delta t \sum_{j=1}^{s} a_{i,j} K_j\big)$.

$Y_1 = Y_1^k$
$Y_2 = Y_2^k$
$y_1 = y_0 + \Delta t \frac{1}{2} \left(f\left(t_0 + \frac{3-\sqrt{3}}{6}\Delta t, Y_1\right) + f\left(t_0 + \frac{3+\sqrt{3}}{6}\Delta t, Y_2\right) \right).$

Tested learning outcome: L3, L4, L7, L10.

Problem 5

a) Consider the θ-method

$$y_{n+1} = y_n + h[(1-\theta)f_n + \theta f_{n+1}],$$

for $\theta \in [0,1]$, for the initial value problem

$$y' = f(t,y), \quad y(t_0) = y_0,$$

where $f_n := f(t_n, y_n)$, $t_n = t_0 + nh$, $y_n \approx y(t_n)$ and h the time step.

Write the θ-method as a Runge-Kutta method by finding the Butcher tableau of this method.

Solution

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 0 & 1 \\ \hline & 1-\theta & \theta \end{array} \quad \text{or} \quad \begin{array}{c|cc} 0 & 0 & \\ 1 & 1-\theta & \theta \\ \hline & 1-\theta & \theta \end{array}$$

Tested learning outcome: L3.

b) Determine and draw the region of A-stability for the method obtained for $\theta = 1$ and for $\theta = \frac{1}{2}$.

Solution

For $\theta = 1$ we have the backward Euler method whose region of absolute stability is

$$S_A = \{z \in \mathbf{C} | |z-1| \geq 1\}.$$

For $\theta = \frac{1}{2}$ we have the trapezoidal rule whose region of absolute stability is the negative half complex plane.

Tested learning outcome: L3, L10.

c) Show that the method is A-stable if and only if $\theta \geq \frac{1}{2}$.

Solution We consider the scalar test equation

$$y' = \lambda y, y(0) = y_0,$$

where the real part of λ is non positive. The stability function of the θ-method is

$$R(z) = \frac{1 + (1-\theta)z}{1-\theta z}.$$

The method is A-stable if

$$\mathcal{R}e(z) \leq 0 \Rightarrow |R(z)| \leq 1.$$

We assume then that $\mathcal{R}e(z) \leq 0$ and explore for which values of θ we have that $|R(z)| \leq 1$ for all such z.

$$\left| \frac{1 + (1-\theta)z}{1-\theta z} \right| \leq 1 \Leftrightarrow |1 + (1-\theta)z| \leq |1 - \theta z|,$$
$$\sqrt{(1 + \mathcal{R}e((1-\theta)z)^2 + (1-\theta)^2 \mathcal{I}m(z)^2} \leq \sqrt{1 - 2\theta \mathcal{R}e(z) + (1-\theta)^2 |z|^2}.$$

Taking squares on both sides and simplifying we get

$$1 - 2\theta \leq 0 \Leftrightarrow \theta \geq \frac{1}{2}.$$

Tested learning outcome: L3, L6, L10.

Problem 6 Let $a \in \mathbb{R}$ and consider the matrix

$$A = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

a)
- For which values of a is A positive definite?
- For which values of a is Gauss-Seidel method convergent?

Solution The eigenvalues of A are $\lambda_1 = 2a + 1$, $\lambda_2 = \lambda_3 = 1 - a$. Therefore all eigenvalues are positive if $-\frac{1}{2} < a < 1$.

Consider $A = M - N$ where M is the lower triangular part of A including the diagonal then $M^{-1}N$ has eigenvalues: 0 and

$$\frac{1}{2} a \left(3a - a^2 \pm (a-1)\sqrt{a(a-4)} \right)$$

and the spectral radius is

$$\rho(M^{-1}N) = \left| \frac{1}{2} a \left(3a - a^2 + (a-1)\sqrt{a(a-4)} \right) \right|$$

and it remains less than 1 for $-\frac{1}{2} < a < 1$ (these are the values for which the Gauss-Seidel method converges).

Tested learning outcome: L4, L6, L10.

b)
- For which values of a is the Jacobi iterative method convergent?
- For which values of a is the Gauss-Seidel iterative method converging faster than the Jacobi iteration?

Solution Consider $A = M - N$ where M is the identity matrix, then $M^{-1}N$ has eigenvalues: $-2a$, a and a, so the spectral radius of this matrix is $2|a|$ and Jacobi method converges if and only if $|a| < \frac{1}{2}$.

For $|a| < \frac{1}{2}$ and $a \neq 0$, the inequality

$$\left| \frac{1}{2} a \left(3a - a^2 + (a-1)\sqrt{a(a-4)} \right) \right| < 2|a|,$$

is always satisfied (we have equality for $a = 0$). Therefore for $\frac{1}{2} < a < 1$ Gauss-Seidel converges while Jacobi doesn't and for $|a| < \frac{1}{2}$ and $a \neq 0$ Gauss-Seidel converges faster than Jacobi.

Tested learning outcome: L4, L6, L10.

Problem 7 Reformulate the following equations into fix-point equations leading to convergent fix-point iterations on some interval $[a, b]$:

$$x^2 - x + 1 = 0, \qquad e^{-x} - \sin(x) = 0.$$

Find a and b. Justify your answers.

Solution The second equation has a zero in the interval $(0, \frac{\Pi}{2}]$, and can be transformed to the fixed point equation

$$x = x \frac{e^{-x}}{\sin(x)},$$

by dividing by $\sin(x)$ and multiplying by x on both sides. The function $g(x) = x\frac{e^{-x}}{\sin(x)}$ maps $(0, \frac{\Pi}{2}]$ into itself and, by the mean value theorem (since g is continuous and differentiable on $(0, \frac{\Pi}{2}]$) ,

$$|g(x) - g(y)| \leq \max_{\xi \in (0, \frac{\Pi}{2}]} |g'(\xi)|\, |x - y|.$$

Computing the derivative of g we observe that it is bounded by 1 on the interval $(0, \frac{\Pi}{2}]$ so g is a contraction on this interval. This suffices to conclude that the fixed point iteration

$$x^{(k)} = x^{(k-1)} \frac{e^{-x^{(k-1)}}}{\sin(x^{(k-1)})}$$

converges for any starting value $x_0 \in (0, \frac{\Pi}{2}]$, by the contraction mapping theorem.

The equation $x^2 - x + 1 = 0$ has two complex conjugate roots. We consider

$$x^2 = x - 1$$

take square roots on both sides and add x on both sides and, after dividing by 2 we obtain

$$x = \frac{1}{2}x + \frac{1}{2}\sqrt{x-1}.$$

Such fixed-point equation for $x_0 < 1$ guarantees that $\sqrt{x_0 - 1}$ is pure imaginary and x_1 is complex. So we can then continue analyzing the iteration in the complex plane. The iteration converges to the root $\frac{1}{2}(1 + i\sqrt{3})$.

Tested learning outcome: L4, L5, L6, L10.