Norwegian University of Science and Technology
Department of Mathematical Sciences

Contact during the exam:
Eirik Hoel Høiseth (40 40 75 39)

Numerical Methods (MA2501)

Monday 12 August 2013
Time: 09:15 – 13:00
Grading: Monday 2 September 2013

Permitted Aids:

- Cheney & Kincaid, *Numerical Mathematics and Computing*, 6. or 7. edition.
- Rottmann, *Mathematical Formulae.*
- Note on fixed point iterations.
- Approved calculator.

General:

- All subproblems carry the same weight when grading.
- All answers should include your reasoning.
- All answers should include enough details to make it clear which methods and results have been used.

Problem 1 Let $\mathbf{x} = [x_1, x_2]^T$. Perform 1 iteration of Newton's method to determine an approximate solution $\mathbf{x}^{(1)}$ to the nonlinear system

$$\begin{aligned} 4x_1^2 - 20x_1 + \frac{1}{4}x_2^2 + 10 = 0, \\ \frac{1}{2}x_1 x_2^2 - 5x_2 + 5 = 0, \end{aligned}$$

starting with $\mathbf{x}^{(0)} = [0, 0]^T$.

Solution: The Jacobian matrix for this system is

$$\mathbf{F}'(\mathbf{x}) = \begin{bmatrix} 8x_1 - 20 & \frac{1}{2}x_2 \\ \frac{1}{2}x_2^2 & x_1 x_2 - 5 \end{bmatrix}.$$

$\mathbf{F}'\left(\mathbf{x}^{(0)}\right)$ is a diagonal matrix, so it is trivial to compute the inverse

$$\left[\mathbf{F}'\left(\mathbf{x}^{(0)}\right)\right]^{-1} = \begin{bmatrix} -20 & 0 \\ 0 & -5 \end{bmatrix}^{-1} = \begin{bmatrix} -\frac{1}{20} & 0 \\ 0 & -\frac{1}{5} \end{bmatrix}.$$

One Newton iteration for this system yields

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \left[\mathbf{F}'\left(\mathbf{x}^{(0)}\right)\right]^{-1} \mathbf{F}\left(\mathbf{x}^{(0)}\right) = -\begin{bmatrix} -\frac{1}{20} & 0 \\ 0 & -\frac{1}{5} \end{bmatrix} \begin{bmatrix} 10 \\ 5 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}.$$

Problem 2 Consider the following five point approximation formula for $f^{(4)}(x)$ of a smooth function f

$$f^{(4)}(x) \approx D_h(f)(x) = \frac{1}{h^4}\left(f(x-2h) - 4f(x-h) + 6f(x) - 4f(x+h) + f(x+2h)\right), \qquad (1)$$

a) Use (1) to estimate the fourth derivative of $f(x) = e^{x/2}$ at $x = 2$ with $h = 0.1$.

Solution: Insertion into the stated formula gives

$$D_{0.1}(f)(2) = \frac{1}{0.1^4}\left(e^{1.8/2} - 4e^{1.9/2} + 6e^{2/2} - 4e^{2.1/2} + e^{2.2/2}\right) = 0.1699634162.$$

Though not required, we can easily compare this against the actual fourth derivative at $x = 2$

$$f^{(4)}(2) = \frac{e^{2/2}}{2^4} = 0.1698926143.$$

b) Show that

$$D_h(f)(x) = f^{(4)}(x) + K_2h^2 + K_4h^4 + K_6h^6 + \dots$$

i.e. the error series consist only of terms with even powers of h. Determine the formula for K_2.

Solution: We first verify that the formula is second order and find K_2. Expanding in Taylor polynomials we have

$$\begin{aligned}
f(x-2h) + f(x+2h) &= 2f(x) + 2\frac{(2h)^2}{2!}f''(x) + 2\frac{(2h)^4}{4!}f^{(4)}(x) + 2\frac{(2h)^6}{6!}f^{(6)}(x) + \dots \\
&= 2f(x) + 4h^2f''(x) + \frac{4}{3}h^4f^{(4)}(x) + \frac{8}{45}h^6f^{(6)}(x) + \dots, \\
-4\left[f(x-h) + f(x+h)\right] &= -8f(x) - 8\frac{h^2}{2!}f''(x) - 8\frac{h^4}{4!}f^{(4)}(x) - 8\frac{h^6}{6!}f^{(6)}(x) + \dots \\
&= -8f(x) - 4h^2f''(x) - \frac{1}{3}h^4f^{(4)}(x) - \frac{1}{90}h^6f^{(6)}(x) + \dots
\end{aligned}$$

Inserting this into the formula for $D_h(f)(x)$ we find that

$$\begin{aligned}
D_h(f)(x) &= \frac{1}{h^4}\left[f(x-2h) - 4f(x-h) + 6f(x) - 4f(x+h) + f(x+2h)\right] \\
&= \frac{1}{h^4}\Bigg[(6+2-8)f(x) + (4-4)h^2f''(x) + \left(\frac{4}{3} - \frac{1}{3}\right)h^4f^{(4)}(x) + \\
&\quad + \left(\frac{8}{45} - \frac{1}{90}\right)h^6f^{(6)}(x) + \dots\Bigg] \\
&= f^{(4)}(x) + \frac{1}{6}f^{(6)}(x)h^2 + \dots
\end{aligned}$$

so the method is second order accurate and

$$K_2 = \frac{1}{6}f^{(6)}(x)$$

That the error series only contains terms with even powers of h can be argued directly from the previous computations. We can also show this to be an immediate consequence of the fact that the formula is even with respect to h, i.e. $D_h(f)(x) = D_{-h}(f)(x)$. If we write

$$D_h(f)(x) = \sum_{i=0}^{\infty} K_ih^i,$$

then because the formula is even with respect to h

$$D_h(f)(x) = \frac{1}{2}\left[D_h(f)(x) + D_{-h}(f)(x)\right] = \sum_{i=0}^{\infty}\frac{1}{2}\left(1 + (-1)^i\right)K_ih^i = \sum_{m=0}^{\infty}K_{2m}h^{2m}$$

c) In the table below the formula (1) has been used to compute approximations of $f^{(4)}(x)$ for some fixed x-value and smooth function f.

h	0.2	0.1	0.05
$D_h(f)(x)$	-0.3879069414	-0.3781498103	-0.3757827913

Compute the highest precision approximation you can for $f^{(4)}(x)$ from these values. Solution: Based on the information, Richardson extrapolation should immediately come to mind. Since the error series is even we can use the normal extrapolation formula. Fixing $h_0 = 0.2$ and setting

$$h_n = h_0/2^n \quad \text{and} \quad D(n,0) = D_{h_n}(f)(x) \quad n \geq 0,$$

we have the extrapolation formula

$$D(n,m) = \frac{4^m}{4^m - 1} D(n, m-1) - \frac{1}{4^m - 1} D(n-1, m-1) \quad 1 \leq m \leq n.$$

The given approximations make up the first column in the array. We compute the rest from the extrapolation formula: Our best approximation is $D(2,2) = -0.3750002084$,

-0.3879069414		
-0.3781498103	-0.3748974333	
-0.3757827913	-0.3749937850	-0.3750002084

which is of order six. The true value turns out to be $f^{(4)}(x) = -0.375$, so our improved approximation turns out to be far more accurate than any of the original approximations in the table.

Problem 3 Find coefficients a and b such that the expression

$$\int_{-1}^{1} \left[ax^2 + b \sin x - e^x\right]^2 \, dx$$

is as small as possible.

Solution: This is a least squares problem. Differentiating with respect to a and b gives the following equations in matrix form upon diving by 2

$$\begin{bmatrix} \int_{-1}^{1} x^4 \, dx & \int_{-1}^{1} x^2 \sin x \, dx \\ \int_{-1}^{1} x^2 \sin x \, dx & \int_{-1}^{1} \sin^2 x \, dx \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \int_{-1}^{1} e^x x^2 \, dx \\ \int_{-1}^{1} e^x \sin x \, dx \end{bmatrix}.$$

The off-diagonal elements are 0, since we're integrating an odd function over a symmetric interval with respect to 0. The remaining integrals are straightforward to solve by elementary techniques (trigonometric identities or partial integration). The resulting diagonal system

$$\begin{bmatrix} 2/5 & 0 \\ 0 & 1-(\sin 2)/2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} e^1 - 5e^{-1} \\ \sin 1 \cosh 1 - \cos 1 \sinh 1 \end{bmatrix},$$

is trivially solved for

$$a = \frac{5}{2}\left[e^1 - 5e^{-1}\right], \qquad b = \frac{\sin 1 \cosh 1 - \cos 1 \sinh 1}{1 - (\sin 2)/2}.$$

Problem 4 Consider the linear system

$$\begin{bmatrix} 2 & 1 & 3 \\ 4 & 6 & 8 \\ 6 & \alpha & 10 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$$

Which of the following values of α require no row interchange when solving the system using Gaussian elimination with scaled partial pivoting?

1. $\alpha = 6$
2. $\alpha = 9$
3. $\alpha = -3$

Solution: Since we have $|\alpha| < 10$ the scale vector is

$$s = [3, 8, 10]$$

and since

$$\max\left(\frac{2}{3}, \frac{4}{8}, \frac{6}{10}\right) = \frac{2}{3}$$

no row interchange is required for the first forward elimination step. After the first step the coefficient matrix is

$$\begin{bmatrix} 2 & 1 & 3 \\ 0 & 4 & 2 \\ 0 & \alpha - 3 & 1 \end{bmatrix}.$$

We will not have to perform a row interchange in the second and last elimination step iff

$$\left|\frac{\alpha - 3}{10}\right| < \frac{4}{8} = \frac{1}{2},$$

or equivalently iff

$$|\alpha - 3| < 5.$$

This only holds for option 1, $\alpha = 6$. The other two α-values require a row interchange in the second step.

Problem 5 Estimate the value of the integral

$$\int_1^3 x \ln x \, dx,$$

using the composite Simpson's rule. Choose the number of subintervals n such that the absolute integration error is guaranteed to not exceed 10^{-4}.

Solution: The error term for composite Simpson with f on [a,b] and n subintervals is

$$e = -\frac{(b-a)^5}{180n^4} f^{(4)}(\xi).$$

For our function $f = x \ln x$ we easily compute

$$f^{(4)}(x) = \frac{2}{x^3}$$

which has absolute maximum $|f^{(4)}(1)| = 2$ on the interval. After inserting values and rearranging, we get the inequality

$$|e| \leq \frac{2^5}{180n^4} \cdot 2 = \frac{16}{45n^4}, \quad \Longleftrightarrow \quad n \geq \sqrt[4]{\frac{16}{45|e|}}.$$

Now, for $|e| = 10^{-4}$ we find a lower limit for the necessary number of subintervals

$$n \geq \sqrt[4]{\frac{16}{45 \cdot 10^{-4}}} \approx 7.72.$$

Since n must be an even integer we choose $n = 8$. Composite Simpson for this intergal with 8 subintervals, i.e. $h = (3-1)/8 = 0.25$, becomes

$$\begin{aligned} S = &\frac{1}{12}(1 \ln 1 + 3 \ln 3) + \frac{1}{3}(1.25 \ln 1.25 + 1.75 \ln 1.75 + 2.25 \ln 2.25 + 2.75 \ln 2.75) + \\ &+ \frac{1}{6}(1.5 \ln 1.5 + 2 \ln 2 + 2.5 \ln 2.5) = 2.9437737349 \end{aligned}$$

Problem 6 Consider the second order differential equation for $y(t)$

$$\ddot{y} + \dot{y}\sin(y) = 0, \quad \text{where } \dot{y} = \frac{dy}{dt}, \; \ddot{y} = \frac{d^2y}{dt^2}, \tag{2}$$

with initial conditions

$$y(0) = 1, \quad \dot{y}(0) = 2. \tag{3}$$

a) We introduce the new variables $x_1 = y$, $x_2 = \dot{y}$. Rewrite the initial value problem (2) and (3) into a system of first-order differential equations in the variables $\mathbf{X} = [x_1, x_2]^T$.

Denote by $\mathbf{X}_i = [x_{1i}, x_{2i}]^T$ the approximation from a numerical method to $\mathbf{X}(t_i)$ with $t_i = t_0 + ih$ for $i = 0, 1, 2, \ldots$. Approximate $\mathbf{X}(0.2)$ for this initial value problem, by taking two steps with Euler's method and stepsize $h = 0.1$, i.e. by computing $\mathbf{X}_2$.

Solution: We follow the standard procedure detailed in C & K to replace the old variables with the new ones, and write up the new system of first-order differential equations:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \dot{\mathbf{X}} = \mathbf{F}(\mathbf{X}) = \begin{bmatrix} x_2 \\ -x_2 \sin x_1 \end{bmatrix},$$

with

$$\mathbf{X}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \mathbf{S} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

With the prescribed notation, Euler's method for this system has the component form

$$\begin{bmatrix} x_{1,i+1} \\ x_{2,i+1} \end{bmatrix} = \begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} + h \begin{bmatrix} x_{2i} \\ -x_{2i} \sin x_1 i \end{bmatrix}, \quad i \geq 0.$$

We take two steps of size $h = 0.1$ to find $\mathbf{X}_2 \approx \mathbf{X}(0.2)$. Starting with $\mathbf{X}_0 = \mathbf{X}(0)$, we get

$$\mathbf{X}_1 = \begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0.1 \begin{bmatrix} 2 \\ -2 \sin 1 \end{bmatrix} = \begin{bmatrix} 1.2 \\ 1.8317058030 \end{bmatrix},$$

and then

$$\mathbf{X}_2 = \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} = \begin{bmatrix} 1.2 \\ 1.8317058030 \end{bmatrix} + 0.1 \begin{bmatrix} 1.8317058030 \\ -1.8317058030 \sin 1.2 \end{bmatrix} = \begin{bmatrix} 1.3831705803 \\ 1.6609836627 \end{bmatrix}.$$

b) Consider here a general autonomous system of first-order differential equations

$$\dot{\mathbf{X}} = \mathbf{F}(\mathbf{X}). \tag{4}$$

When applied to (4), both Euler's method and the implicit Euler method

$$\mathbf{X}_{n+1} = \mathbf{X}_n + h\mathbf{F}(\mathbf{X}_{n+1}),$$

are only first order.

We try to generate a higher order method for (4) by combining these two methods in the following way:

1. Take a step of size $h/2$ with Euler's method to get from $\mathbf{X}_n$ to $\mathbf{X}_{n+1/2}$.
2. Take a step of size $h/2$ with the implicit Euler method to get from $\mathbf{X}_{n+1/2}$ to $\mathbf{X}_{n+1}$.

Show that the resulting method is a Runge-Kutta method. Write down its Butcher tableau. Determine the order of our new method.

Hint: A general Runge-Kutta method with s-stages can be written for (4) as

$$\begin{aligned}\mathbf{K}_i &= \mathbf{F}(\mathbf{X}_n + h\sum_{j=1}^{s} a_{ij}\mathbf{K}_j) \quad i = 1, \dots, s,\\ \mathbf{X}_{n+1} &= \mathbf{X}_n + h\sum_{i=1}^{s} b_i\mathbf{K}_i.\end{aligned}$$

The parameters a_{ij}, $c_i = \sum_{j=1}^{s} a_{ij}$ and b_i that specify the method are commonly stated in a *Butcher tableau*

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

Order conditions for a Runge-Kutta method can be given as algebraic conditions for the coefficients

$$\begin{aligned} & \sum_{i=1}^{s} b_i = 1 && \text{for order 1,} \\ \text{in addition} \quad & \sum_{i=1}^{s} b_i c_i = 1/2 && \text{for order 2,} \\ \text{in addition} \quad & \sum_{i=1}^{s} b_i c_i^2 = 1/3, && \\ \text{and} \quad & \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{i,j} c_j = 1/6 && \text{for order 3.} \end{aligned}$$

Higher order than 3 requires additional algebraic conditions.

Solution: We first write out the described method explicitly:

$$\begin{aligned}\mathbf{X}_{n+1/2} &= \mathbf{X}_n + \frac{h}{2}\mathbf{F}(\mathbf{X}_n),\\ \mathbf{X}_{n+1} &= \mathbf{X}_{n+1/2} + \frac{h}{2}\mathbf{F}(\mathbf{X}_{n+1}).\end{aligned}$$

We can merge the two parts, by inserting the expression for $\mathbf{X}_{n+1/2}$ from the first line into the second

$$\mathbf{X}_{n+1} = \mathbf{X}_n + h\frac{\mathbf{F}(\mathbf{X}_n) + \mathbf{F}(\mathbf{X}_{n+1})}{2}.$$

Comparing this to the Runge-Kutta expression, we observe we can write the method in this form by setting

$$\begin{aligned}
&\mathbf{K}_1 = \mathbf{F}(\mathbf{X}_n), \\
&\mathbf{K}_2 = \mathbf{F}\left(\mathbf{X}_{n+1}\right) = \mathbf{F}\left(\mathbf{X}_n + h\frac{\mathbf{F}(\mathbf{X}_n) + \mathbf{F}(\mathbf{X}_{n+1})}{2}\right) = \mathbf{F}\left(\mathbf{X}_n + h\left[\frac{1}{2}\mathbf{K}_1 + \frac{1}{2}\mathbf{K}_2\right]\right),
\end{aligned}$$

which gives

$$\mathbf{X}_{n+1} = \mathbf{X}_n + h\left(\frac{1}{2}\mathbf{K}_1 + \frac{1}{2}\mathbf{K}_2\right).$$

Thus the method is a Runge-Kutta method with $s = 2$ stages. In fact it is just the well known trapezoidal method. Reading out the coefficients, the Butcher tableau of the method is

$$\begin{array}{c|cc}
0 & 0 & 0 \\
1 & \frac{1}{2} & \frac{1}{2} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}$$

Finally we check the algebraic conditions to determine the order of the method

$$\begin{aligned}
\sum_{i=1}^{2} b_i &= \frac{1}{2} + \frac{1}{2} = 1, \quad \text{order 1 satisfied,} \\
\sum_{i=1}^{2} b_i c_i &= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2}, \quad \text{order 2 satisfied,} \\
\sum_{i=1}^{2} b_i c_i^2 &= \frac{1}{2}\cdot 0^2 + \frac{1}{2}\cdot 1^2 = \frac{1}{2} \neq \frac{1}{3}, \quad \text{order 3 NOT satisfied.}
\end{aligned}$$

The method is therefore second order.