Homework 2

due 30/1-2012

Task 1 : Machine Epsilon

The following code can be used in MATLAB to determine the machine accuracy ε.

```
numprec=double(1.0); % Define 1.0 with double precision
numprec=single(1.0); % Define 1.0 with single precision
while(1 < 1 + numprec)
    numprec=numprec*0.5;
end
numprec=numprec*2
```

a) Determine ε using the above program, both for single and double precision.

   Note: The implementation of single/double precision arithmetics differs between versions of MATLAB. If runs with both single and double precision give the same answer, please try another computer/version of MATLAB if possible. Otherwise, write down your MATLAB version and move on. The above code is working properly on release 2009a on Linux, for instance.

b) Give a definition of the machine accuracy based on the code above. Try to use words and not mathematical expressions.

Task 2 : Round-off Error

In this exercise, the errors involved in numerically calculating derivatives are examined. For example, the derivative of a function f can be approximated with central differences:

$$f'_{num}(x) = \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x} \tag{1}$$

a) Determine the relative error ϵ of the derivative of the function $f(x) = \dfrac{1}{1+x} + x$ when using the central difference approximation defined above:

$$\epsilon = \frac{|f'(x) - f'_{num}(x)|}{|f'(x)|}$$

   at the location $x = 2$. In the calculation use the stepsizes $\Delta x = 10^{-20} \ldots 10^{0}$. Use both single and double precision for the calculation, and present the results in a double logarithmic plot (ϵ vs. Δx). In MATLAB double logarithmic plots are obtained by the function `loglog()`. Remember that all variables used here should be defined as double or single precision as in Task 1.

b) The general formula for the propagation error, for a function $h(x_j)$ with n variables x_j is given by:

$$\epsilon_h = \sum_{j=1}^{n} \left| \frac{x_j}{h} \frac{\partial h}{\partial x_j} \right| \varepsilon_{x_j} \, ,$$

where ε_{x_j} is the relative error. Based on that, show that the propagation error ϵ_h, when adding two numbers x and y, is given by:

$$\epsilon_h = \frac{|x|}{|x+y|}\varepsilon_x + \frac{|y|}{|x+y|}\varepsilon_y$$

where ε_x and ε_y are the corresponding errors for each number.

c) Show that the relative discretisation error of using equation (1) is given by:

$$\epsilon_d = \frac{\Delta x^2 |f'''(x)|}{6|f'(x)|}$$

(Hint: Taylor expansion)
and the propagation error is given by (round-off error):

$$\epsilon_r = \frac{\varepsilon \cdot |f(x)|}{\Delta x |f'(x)|}$$

(Hint: Use equation from part b)
with the machine accuracy ε. Find the value of Δx that minimises the total error:

$$\epsilon_g = \epsilon_r + \epsilon_d$$

Plot the results for $\epsilon_r, \epsilon_d, \epsilon_g$ together with the results from part a).

Task 3 : Integration of differential equation

In this problem the stability and convergence order of some simple integration methods is examined. The first order, ordinary, linear differential equation with constant coefficient is considered (Dahlquist equation)

$$\frac{\mathrm{d}u}{\mathrm{d}t} = A(u) = \lambda u \qquad , \qquad u(0) = 1$$

where $0 \leq t \leq T$ and $\lambda = \mathit{const} \in \mathbb{C}$. The time interval $[0, T]$ is split into N parts with the same length Δt. The following integration methods should be used:

- explicit Euler

$$u^{n+1} - u^n = \Delta t A(u^n)$$

- implicit Euler

$$u^{n+1} - u^n = \Delta t A(u^{n+1})$$

- Crank-Nicolson

$$u^{n+1} - u^n = \frac{1}{2}\Delta t(A(u^{n+1}) + A(u^n))$$

where $n = 0, ..., N$. Calculate until $T = 16$ and use the discretization with $N = 20, 40, 50, 100, 200$ steps.

a) Derive the analytical solution u_{ex}.

b) For $\lambda = -0.2 + i$, calculate the numerical solution with the given discretisations N and the three integration methods. Plot the real part of the analytical solution and the three numerical solutions for each value of N.

c) Discuss the usefulness and accuracy of the methods.

d) For $\lambda = -0.2 + 0.1i$, do as in b) and calculate the numerical and analytical solutions. Show also the error $|u_{ex} - u_{num}|$ at the time $t = 16$ as a function of N in a double logarithmic plot. Explain the differences between the methods.

Note (for all tasks): Together with your solutions, hand in the MATLAB-codes that you have written yourself.