University of Saskatchewan
Department of Mathematics and Statistics

Numerical Analysis III
(MATH 314)
Instructor: Dr. Raymond J. Spiteri

ASSIGNMENT 03
Due: 10:00 a.m. Tuesday, November 10, 2009

1. [**15 marks**]

   (a) Derive the AB2 and AM3 methods and their local truncation errors for constant step size Δt.

   (b) Use the above results to derive the local truncation errors of AB2 and AM3 *for the equation* $\dot{y} = -y$.

   (c) Use the result from the previous question to derive the corresponding local truncation error for the AB2-AM3 predictor-corrector pair.

   Is the accuracy of a predictor-corrector pair different from that of the corrector evaluated as an implicit formula?

2. [**25 marks**] A simple model for ozone kinetics is

$$\begin{aligned}\dot{y}_1 &= -q_1 y_1 y_3 - q_2 y_1 y_2 + 2q_3(t) y_3 + q_4(t) y_2,\\ \dot{y}_2 &= \quad q_1 y_1 y_3 - q_2 y_1 y_2 - q_4(t) y_2,\end{aligned}$$

where $y_1 = y_1(t)$ is atomic oxygen, $y_2 = y_2(t)$ is ozone, and $y_3 = 3.7 \times 10^{16}$ is molecular oxygen. Two of the reactions ($i = 1, 2$) have constant rate coefficients

$$q_1 = 1.63 \times 10^{-16}, \qquad q_2 = 4.66 \times 10^{-16},$$

and two ($i = 3, 4$) have diurnal rate coefficients

$$q_i(t) = \begin{cases} \exp(-c_i/\sin(\omega t)), & \sin(\omega t) > 0, \\ 0 & \sin(\omega t) \leq 0, \end{cases}$$

where

$$c_3 = 22.62, \qquad c_4 = 7.601, \qquad \omega = \frac{\pi}{43200},$$

for t measured in seconds. The initial conditions are

$$y_1(0) = 10^6, \qquad y_2(0) = 10^{12}.$$

   (a) Plot the diurnal rate coefficients and interpret their meaning.

   (b) Solve the initial-value problem to a final time of 10.25 days using `ode45` and `ode15s`. You should experiment with different tolerance values until you have some confidence that a semi-logarithmic plot of the solution looks correct. Do you think the problem is stiff? Explain.

(c) Plot the step-sizes used as a function of time and give an analysis of the behaviour.

(d) Explain why the presence of negative values for any of the solution components would be problematic. Did you encounter any negative values in your simulation? If so, how did you deal with them so they did not undermine the credibility of your results?

3. **[15 marks]** To help understand the results of Question 2, compute the eigenvalues of the Jacobian $\frac{\partial \mathbf{f}}{\partial \mathbf{y}}$ at the initial point.

(a) Explain how the eigenvalues support or contradict your conclusion about the problem stiffness.

(b) Consider the eigenvalues of the Jacobian near the point $y_1 = 0$, $y_2 = 10^{12}$. What can you say about the stability of the problem if y_1 is small and negative? Explain how this supports or contradicts the behaviour of the system in Question 2 if it encountered small and negative values of y_1.

4. **[10 marks]**

(a) Derive the stability region S for BDF3.

(b) Show that BDF3 is not A-stable.

Bonus (3 marks): Plot S.

5. **[25 marks]** Consider the following Runge–Kutta method:

$$\begin{array}{c|ccccc}
\frac{1}{4} & \frac{1}{4} & & & & \\
\frac{3}{4} & \frac{1}{2} & \frac{1}{4} & & & \\
\frac{11}{20} & \frac{17}{50} & -\frac{1}{25} & \frac{1}{4} & & \\
\frac{1}{2} & \frac{371}{1360} & -\frac{137}{2720} & \frac{15}{544} & \frac{1}{4} & \\
1 & \frac{25}{24} & -\frac{49}{48} & \frac{125}{16} & -\frac{85}{12} & \frac{1}{4} \\
\hline
 & \frac{25}{24} & -\frac{49}{48} & \frac{125}{16} & -\frac{85}{12} & \frac{1}{4}
\end{array}$$

(a) Classify the method as completely as possible; i.e., specify as precisely as you can what kind of implicit method it is. What are the advantages of such a method?

(b) What is the order of the method?

(c) Test the method for A-stability and L-stability.

(d) **Bonus: [10 marks]** Without using any extra stages, derive a continuous extension formula for the method that has the same order as the method.

Hint: The main theoretical result that we need for the Bonus question is the following. Let the interpolant of the Runge–Kutta method be given by

$$u(\theta) = y_n + \Delta t \sum_{i=1}^{s^*} b_i(\theta) k_i,$$

where $b_i(\theta) = \sum_{j=1}^{p^*} b_{ij}\theta^j$, are polynomials to be determined such that

$$u(\theta) - y(t_n + \theta\Delta t) = \mathcal{O}(\Delta t^{p^*+1}). \tag{1}$$

The error in $u(\theta)$ is of order p^* (i.e., its local interpolation error satisfies (1)) if and only if the so-called *continuous order conditions* are satisfied. If you use the Albrecht form of the classical Runge–Kutta order conditions, all you have to do is replace $\mathbf{b}$ with $\mathbf{b}(\theta)$ and θ^p/p in the quadrature condition at order p; e.g., $\mathbf{b}^T\mathbf{e} = 1$ becomes $\mathbf{b}(\theta)^T\mathbf{e} = \theta$ at order 1, $\mathbf{b}^T\mathbf{c} = 1/2$ becomes $\mathbf{b}(\theta)^T\mathbf{c} = \theta^2/2$, etc.

You then form and solve a system of simultaneous linear equations for the unknowns b_{ij}, $i = 1, 2, \ldots, s^*$, $j = 1, 2, \ldots, p^*$.

6. **[10 marks]** Consider a perturbed version of the Robertson problem:

$$\begin{aligned}
\dot{y}_1 &= -0.04y_1 + 10^4 y_2 y_3, & y_1(0) &= 1 - 2\epsilon_{\text{machine}},\\
\dot{y}_2 &= 0.04y_1 - 10^4 y_2 y_3 - 3\times 10^7 y_2^2, & y_2(0) &= \epsilon_{\text{machine}},\\
\dot{y}_3 &= 3\times 10^7 y_2^2, & y_3(0) &= \epsilon_{\text{machine}}.
\end{aligned}$$

(a) Re-write the Robertson problem under the transformation $y_i = e^{u_i}$, $i = 1, 2, 3$.

(b) Give an important advantage of solving the transformed problem.

(c) List 3 distinct disadvantages of solving the transformed problem.

(d) Compare the performance of solving the original problem with `ode15s` to solving the transformed problem with `ode45` and `ode15s` to a final time of $t_f = 4 \times 10^{11}$s. Do you recommend solving the original problem or the transformed one?