CPSC520: Solutions to Assignment 2, 2012

1. (a) Obviously this is a one-step method with two stages. In our RK notation we have $K_1 = f(t_n, y_n)$ and $K_2 = f(t_{n+\theta}, y_n + k\theta K_1)$. In Tableau form

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

   (b) This method consists of two steps of forward Euler, so it is certainly first order accurate for any $0 \le \theta \le 1$. The condition (2.13) in the text for 2nd order accuracy holds only when $\theta = 1/2$.

   (c) For the test equation $y' = \lambda y$ with $z = k\lambda$ we have

$$y_{n+1} = y_n + z y_{n+\theta} = y_n + z(y_n + \theta z y_n) = (1 + z + \theta z^2) y_n.$$

   So, $R(z) = 1 + z + \theta z^2$. Next, let $z = \imath\epsilon$ for some real small ϵ, $|\epsilon| \ll 1$. Then $R(z) = (1 - \theta\epsilon^2) + \imath\epsilon$, hence we are searching for values ϵ such that

$$|R(z)|^2 = 1 - 2\theta\epsilon^2 + \theta^2\epsilon^4 + \epsilon^2 \le 1.$$

   If $\theta \neq 1/2$ then the $\mathcal{O}(\epsilon^4)$ is dominated by the $\mathcal{O}(\epsilon^2)$ term and may be ignored. We obtain absolute stability if $1 - 2\theta < 0$, i.e., $\theta > \frac{1}{2}$. If $\theta = 1/2$ then the $\mathcal{O}(\epsilon^2)$ term vanishes and we conclude no absolute stability because $1 + \epsilon^4 > 1$.

2. (a) Non-negativity implies that the magnitudes are not needed in the usual condition of absolute stability $|y_{n+1}| \le |y_n|$. This directly yields monotonicity.

   (b) For forward Euler $R(z) = 1 + z \ge 0$ implies $z \ge -1$.

   (c) For backward Euler $R(z) = \frac{1}{1-z} \ge 0$ for any $z \le 0$.

   (d) For trap $R(z) = \frac{1+z/2}{1-z/2} \ge 0$ implies the condition $1 + z/2 \ge 0$, which holds when $z \ge -2$.

   (e) At first $y_{n+1/2} = \frac{1+z/4}{1-z/4} y_n$. Plugging this into the expression for y_{n+1} the condition for non-negativity is seen to be $4\frac{1+z/4}{1-z/4} - 1 \ge 0$ which yields $z \ge -\frac{12}{5}$. This method is not unconditionally nonnegative although it is L-stable.

   (f) A symmetric implicit method based on quadratic instead of linear elements would have

$$R(z) = \frac{1 + z/2 + \alpha z^2}{1 - z/2 + \alpha z^2}.$$

If $\alpha = 0$, as in the trapezoidal and midpoint methods, then $R(\infty) = -1$. But if $\alpha > 0$ then we can have $R(\infty) = 1$, so at least in the limit R is positive.

For instance, Gaussian collocation at two points ($s = 2$) has

$$R(z) = \frac{1 + z/2 + z^2/8}{1 - z/2 + z^2/8}.$$

Like for the trapezoidal rule in item (d) we need investigate only whether the denominator remains nonnegative for all $z \leq 0$. Clearly that is so at $z = 0$ and $z = -\infty$. In between the minimum is where the derivative vanishes, which is at $1/2 + z/4 = 0$, i.e. $z = -2$. But there $1 + z/2 + z^2/8 = (-2)^2/8 > 0$. So this scheme is unconditionally nonnegative.

3. (a) This is standard: set $y_1 = u_1,\ y_2 = u_1',\ y_3 = u_2,\ y_4 = u_2'$. This gives the first order ODE system of size $m = 4$

$$\begin{aligned}
y_1' &= y_2, \\
y_2' &= y_1 + 2y_4 - \hat{\mu}\frac{y_1 + \mu}{D_1} - \mu\frac{y_1 - \hat{\mu}}{D_2}, \\
y_3' &= y_4, \\
y_4' &= y_3 - 2y_2 - \hat{\mu}\frac{y_3}{D_1} - \mu\frac{y_3}{D_2}, \quad \text{where} \\
D_1 &= ((y_1 + \mu)^2 + y_3^2)^{3/2}, \\
D_2 &= ((y_1 - \hat{\mu})^2 + y_3^2)^{3/2}.
\end{aligned}$$

   (b) Using `ode45` with default tolerances, this necessitated 309 time steps with $\max_n k_n = .1892$ and $\min_n k_n = 1.5921\text{e-}7$.

   (c) Using 1,000 steps yields nonsensical results. Even 5,000 uniform steps yield qualitatively incorrect results. Only 10,000 uniform steps yield a qualitatively correct figure, similar as far as the naked eye is concerned to Figure 1.

   Here, an adaptive step size selection obviously reaps great benefits. Indeed, notice how small is the smallest step size used by `ode45`.

4. (a) This claim follows straight from the definition of the composite quadrature rules of midpoint and trapezoidal, i.e., upon writing

$$T = \sum_{i=0}^{J-1} \int_{x_i}^{x_{i+1}} \left[a(u')^2 - 2uq\right] dx$$

   and applying the basic rules to each short integral.

(b) For each $j,\ j = 1, \ldots, J$, we set $\frac{\partial T_h}{\partial v_j} = 0$. (Note v_0 is not an unknown, but v_J is.) We get contributions from $i = j$ and from $i = j - 1$, and these add up as specified.

(c) Here $\mathbf{v} = (v_1, v_2, \ldots, v_J)^T$, and the right hand side is likewise $\mathbf{q} = (q_1, q_2, \ldots, q_J)^T$, where $q_j = \frac{h_j + h_{j-1}}{2} q(x_j)$.

Denoting the (k, l)th element of A by $a_{k,l}$, the matrix A is tridiagonal and symmetric, with the main diagonal elements

$$a_{j,j} = \frac{a(x_{j-1/2})}{h_{j-1}} + \frac{a(x_{j+1/2})}{h_j},\ j = 1, 2, \ldots, J-1, \quad a_{J,J} = \frac{a(x_{J-1/2})}{h_{J-1}},$$

and the super-diagonal elements

$$a_{j,j+1} = -\frac{a(x_{j+1/2})}{h_j},\ j = 1, 2, \ldots, J-1.$$

(The sub-diagonal elements are determined likewise by symmetry.)

Furthermore we have, since $a(x) > 0$, that $a_{i,i} > 0$, while $a_{i,i+1} = a_{i+1,i} < 0$, for all relevant i, and thus $a_{i,i} \geq \sum_{j \neq i} |a_{i,j}|,\ i = 1, 2, \ldots, J$. (In fact, $a_{i,i} = \sum_{j \neq i} |a_{i,j}|,\ i = 1, 2 \ldots, J$.) This yields the conclusion (by Gerschgorin's Theorem) that the eigenvalues are all nonnegative. To see that the matrix is nonsingular, we can consider directly solving $A\mathbf{v} = \mathbf{0}$. Back-substitution for all but the first row shows that $\mathbf{v}$ must be a constant vector, but this does not agree with the first row, hence there is no solution and 0 is not an eigenvalue. Therefore, all eigenvalues are positive. Thus, A is symmetric positive definite.

(d) The necessary conditions clearly approximate to 2nd order the differential equation $-(au')' = q$ on subintervals (x_{j-1}, x_{j+1}). So, the truncation error is $\mathcal{O}(h^2)$. Stability (which is trickier to prove) then leads to the conclusion that $\mathbf{v}$ is a second order approximation of u at mesh points as well.

For a numerical example I used $a(x) = (2+x)/(1+x)$, $q(x) = 2x + 1$, and hence $u(x) = x - x^3/3$. A nonuniform mesh is constructed using the script

```
x(1) = 0; count = 1;
while (x(count) < 1-hun)
    x(count+1) = x(count) + hun; count = count + 1;
    if (mod(count,2) == 0),
        x(count+1) = x(count) + hun/2; count = count + 1;
        x(count+1) = x(count) + hun/2; count = count + 1;
    end
end
if (x(count) < 1), x(count+1) = 1; end
```

Thus, every second subinterval (or element) is halved.

For $hun = .1$ the maximum error is 1.8e-3, whereas for $hun = .05$ the error is 4.27e-4, and for $hun = .025$ the error is 1.1e-4. The trend is clearly 2nd order.

5. (a) The hat functions are defined to be linear on each mesh subinterval (element), so they are each piecewise linear. Also, $\phi_i(x_j) = \delta_{i,j}$, i.e., $= 1$ when $j = i$ and $= 0$ otherwise. The function $w(x)$ is a sum of piecewise linear functions so it also is one. Furthermore, indeed using $\delta_{i,j}$, $w(x_j) = \sum w(x_i)\phi_i(x_j)$.

   (b) The Galerkin equations are

$$\begin{aligned} \sum_{j=1}^{J} b(\phi_i, \phi_j)v_j &= \sum_j a_{i,j}v_j = \sum_j \int_0^1 a(x)\phi_i(x)\phi_j(x)dx = \int_0^1 q(x)\phi_i(x) \\ &= (q, \phi_i), \quad i = 1, 2, \ldots, J. \end{aligned}$$

   This gives a matrix A with elements $a_{i,j}$. Note that $\phi_i(x)\phi_j(x) = 0, \forall x$, unless $j = i$ or $j = i \pm 1$. In fact, it is not difficult to see that A is tridiagonal and symmetric. Indeed, upon using the midpoint rule to approximate these integrals we obtain the matrix A of Exercise 4. Likewise, by applying the trapezoidal rule to the integrals of (q, ϕ_j) we obtain the vector $\mathbf{q}$ of the previous exercise.