

Solutions to Problem Set 5, Math 228A, Fall 2005

1. To find values for γ

Way 1(most straightforward):

$$R(z) = 1 + zb^T(I - zA)^{-1}e = \dots = \frac{1 + (1 - 2\gamma)z + (\frac{1}{2} - 2\gamma + \gamma^2)z^2}{(1 - \gamma z)^2}$$

As a result, $R(\infty) = 0$ implies that $\gamma^2 - 2\gamma + \frac{1}{2} = 0$

Way 2: Make use of the conclusion from Ex2, i.e. we want

$$b^T A^{-1}e = 1$$

Way 3:(much more detail will be provided as follows)

We need to check three things:

- (1) A-stability;
- (2) $R(-\infty) = 0$;
- (3) 2nd order.

We will use (2) to find γ first. Then check (1) and (3).

First of all, what is an SDIRK? An ERK (explicit) has a lower triangular A matrix with zeros on the diagonal. A DIRK (diagonally implicit) has a lower triangular A matrix with nonzero entries on the diagonal. A SIRK (singularly implicit) has an A matrix with all the eigenvalues equal. An SDIRK (singularly diagonally implicit) has a lower triangular A matrix with all the eigenvalues equal. So (since the eigenvalues of a lower triangular matrix are just the diagonal entries), an SDIRK has all its diagonal entries equal.

Actually, at this point it is useful to work out a formula for $R(z)$ for general RK methods (this approach due to Dekker and Verwer). First we work with an alternate formulation (equivalent to that whole k_i stuff):

$$y_{n+1} = y_n + h \sum_{i=1}^{i=s} b_i f(t_n + c_i h, Y_i),$$

where

$$Y_i = y_n + h \sum_{j=1}^{j=s} a_{ij} f(t_n + c_j h, Y_j)$$

(We can switch back and forth between formulations just by setting $k_i = f(x_n + c_i h, Y_i)$).

Now, with this new formalism, we can apply the RK method to the simple

equation $y' = \lambda y$:

$$Y_i = y_n + z \sum_{j=1}^{j=s} a_{ij} Y_j$$

$$y_{n+1} = y_n + z \sum_{i=1}^{i=s} b_i Y_i$$

We can rewrite this in block matrix form (Let $Y = (Y_1, Y_2, \dots, Y_s)^T$ be the s-vector composed of the Y_i , and let $e = (1, 1, \dots, 1)^T$ be the s-vector of all 1's):

$$\begin{pmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{pmatrix} \begin{pmatrix} Y \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} y_n e \\ y_n \end{pmatrix}$$

We can solve this for y_{n+1} in 2 ways.

One way involves inverting the matrix $(I_s - zA)$ and solving for the Y_i . This yields the familiar:

$$R(z) = 1 + zb^T(I_s - zA)^{-1}e$$

Of course we can then compute the inverse and evaluate $R(z)$.

Another way involves Cramer's Rule (that whole transpose of the cofactor matrix stuff). Let $(\cdot)^C$ mean 'taking cofactors' and let $(\cdot)^T$ mean 'transpose'. Now $M^{-1} = \frac{1}{|M|} M^{TC}$. Now when we solve for y_{n+1} using Cramer's Rule, we get:

$$\begin{pmatrix} Y \\ y_{n+1} \end{pmatrix} = \frac{1}{\begin{vmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{vmatrix}} \begin{pmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{pmatrix}^{TC} \begin{pmatrix} y_n e \\ y_n \end{pmatrix}$$

$$\begin{pmatrix} Y \\ y_{n+1} \end{pmatrix} = \frac{1}{\begin{vmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{vmatrix}} \begin{pmatrix} I_s - zA^T & -zb \\ 0 & 1 \end{pmatrix}^C \begin{pmatrix} y_n e \\ y_n \end{pmatrix}$$

And now, since we are interested in y_{n+1} , we only have to carry out matrix multiplication on the bottom row of the right hand side. Now since we have to compute a bunch of cofactors along the bottom row, multiply them by y_n , and add them all together, we might as well stick the y_n 's in the bottom row to begin with, and just take a determinant.

$$y_{n+1} = \frac{1}{\begin{vmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{vmatrix}} \begin{vmatrix} I_s - zA^T & -zb \\ y_n e^T & y_n \end{vmatrix}$$

Now, with determinants, we are free to carry out elementary row (column) operations (as long as we keep track of what we are doing). First, we can want to divide the bottom row by y_n (and multiply the determinant by y_n).

$$y_{n+1} = \frac{y_n}{\begin{vmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{vmatrix}} \begin{vmatrix} I_s - zA^T & -zb \\ e^T & 1 \end{vmatrix}$$

Now we want to subtract the last column from the other columns (without changing the determinant)

$$y_{n+1} = \frac{y_n}{\begin{vmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{vmatrix}} \begin{vmatrix} I_s - zA^T + zbe^T & -zb \\ 0 & 1 \end{vmatrix}$$

Now actually evaluating the determinant is easy, since we can just expand across the bottom row

$$y_{n+1} = \frac{y_n |I_s - zA + zeb^T|}{\begin{vmatrix} I_s - zA & 0 \\ -zb^T & 1 \end{vmatrix}}$$

Actually, evaluating the other determinant (the one in the denominator) is also easy. (I suppose I should have done that earlier). At any rate:

$$y_{n+1} = \frac{y_n |I_s - zA + zeb^T|}{|I_s - zA|}$$

And so we see that

$$R(z) = \frac{|I_s - zA + zeb^T|}{|I_s - zA|}$$

Now let's apply this formula for $R(z)$ to an SDIRK with diagonal entries $a_{ii} = \gamma$. First note that $(I_s - zA)$ is a lower triangular matrix with diagonal entries $1 - z\gamma$. And so

$$|I_s - zA| = (1 - z\gamma)^s = (-1)^s \gamma^s z^s + s(-1)^{s-1} \gamma^{s-1} z^{s-1} + \dots + 1$$

Now note that $(I_s - zA + zeb^T)$ is a full matrix. In order for the SDIRK to be L-stable, $R(-\infty)$ must equal 0. This means that the $|I_s - zA + zeb^T|$ had better be a polynomial of degree less than s .

(If $|I_s - zA + zeb^T|$ is a degree s polynomial, say $q_s z^s + z_{s-1} z^{s-1} + \dots + q_0$, then $R(-\infty)$ will equal $\frac{q_s}{(-1)^s \gamma^s}$). (Moreover, if $|I_s - zA + zeb^T|$ is a polynomial of

degree $s - 1$ or less, then $R(-\infty)$ will equal 0).

So if we are to have the SDIRK be L-stable, then the coefficient of z^s in $|I_s - zA + zeb^T|$ must be zero.

We have the equality

$$|I_s - zA + zeb^T| = (-z)^s |(A - eb^T) - \frac{1}{z}I_s|$$

And so the coefficient of z^s in $|I_s - zA + zeb^T|$ will equal zero iff the constant term of $|(A - be^T) - \frac{1}{z}I_s|$ is equal to zero.

However, this is true iff the characteristic polynomial of $(A - be^T)$ has zero constant term.

Which is equivalent to $(A - be^T)$ having a zero eigenvalue (which in turn implies $(A - be^T)$ is singular).

Whew. So, after all that rigamarole, we have whittled down the criterion for L-stability (of an SDIRK method) into the simple requirement that $(A - be^T)$ be singular.

So now (for the second part of the question) we have to apply this criterion to a 2-stage SDIRK that looks like

$$A = \begin{pmatrix} \gamma & 0 \\ 1 - 2\gamma & \gamma \end{pmatrix}$$

$$b^T = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$A - eb^T = \begin{pmatrix} \gamma - \frac{1}{2} & \frac{-1}{2} \\ \frac{1}{2} - 2\gamma & \gamma - \frac{1}{2} \end{pmatrix}$$

$$|A - be^T| = \gamma^2 - \gamma + \frac{1}{4} + \frac{1}{4} - \gamma = \gamma^2 - 2\gamma + \frac{1}{2} = 0$$

$$\gamma = 1 \pm \frac{1}{\sqrt{2}}$$

Next, a simple task, we should check the order conditions to make sure the method is of second order. This is trivial.

Finally we need to check if the method is A-stable(**unfortunately many people forgot this**) for $\gamma = 1 \pm \frac{1}{\sqrt{2}}$. Let $z = \lambda h = u + iv$.

$$R(z) = 1 + zb^T(I - zA)^{-1}e = \frac{1 + z(1 - 2\gamma)}{(1 - \gamma z)^2}.$$

We want $|R(z)| < 1$. Of course we can plot the RAS.

Or let's do the direct computations. For nonzero denominator $|R(z)| < 1$ iff

$$[1 + (1 - 2\gamma)u]^2 + [v(1 - 2\gamma)]^2 < [(1 - \gamma u)^2 - \gamma^2 v^2]^2 + [2\gamma v(1 - \gamma u)]^2$$

Now expand(tedious of course)

$$\begin{aligned} & u^2 + 2u - 4\gamma u^2 + (1 - 4\gamma + 2\gamma^2)v^2 \\ & < 2\gamma^2 u^2 + \gamma^4 u^4 - 4\gamma^3 u^3 + \gamma^4 v^4 - 4\gamma^2 uv^2 + 2\gamma^2 u^2 v^2. \end{aligned}$$

and fully make use of $2\gamma^2 - 4\gamma + 1 = 0$ on the LHS we have

$$0 < 4\gamma^2 u^2 - 2u + \gamma^4 u^4 - 4\gamma^3 u^3 + \gamma^4 v^4 - 4\gamma^2 uv^2 + 2\gamma^2 u^2 v^2.$$

Clearly when $u < 0$ every term on the RHS is nonnegative(most terms positive).

Thus the method is A-stable for $\gamma = 1 \pm \frac{1}{\sqrt{2}}$.

To summary, we got an L-stable SDIRK method.

2. Now for more RK goodness. Here we have to show that any implicit A-stable RK method with non-singular A satisfying $a_{sj} = b_j$ is L-stable.

So we can assume that the A matrix is nonsingular, and its bottom row is equal to b . We already know that the method is A-stable. So to show L-stability, we only need to show that $R(-\infty) = 0$. Here the first definition of $R(z)$ is more convenient:

$$R(z) = 1 + zb^T(I_s - zA)^{-1}e$$

First we have to find an expression for $(I_s - zA)^{-1}$. We are all familiar with the following:

$$(I_s - zA)^{-1} = 1 + zA + z^2A^2 + z^3A^3 + \dots$$

However, this is only useful when $\|zA\|$ is small. We want to consider the times when $\|zA\|$ is large. So we need another (less used) expansion:

$$\begin{aligned} (I_s - zA)^{-1} &= \frac{-1}{z}A^{-1} + \frac{-1}{z^2}A^{-2} + \frac{-1}{z^3}A^{-3} + \dots \\ &= \frac{-1}{z}A^{-1} + "O\left(\frac{1}{\|zA\|^2}\right)" \end{aligned}$$

Note here we use " $O\left(\frac{1}{\|zA\|^2}\right)$ " to denote the high order terms. As they are matrices and we generally don't write $O\left(\frac{1}{\|zA\|^2}\right)$. So we have

$$R(z) = 1 + zb^T\left(\frac{-1}{z}A^{-1}\right)e + O\left(\frac{1}{\|zA\|}\right)$$

Now note that $AA^{-1} = I_s$, which means that

$$(A)_{ik}(A^{-1})_{kj} = \delta_{ij}$$

And since $(A)_{sk} = b^T$, we know the last row of $AA^{-1} = I_s$ is

$$b^T(A^{-1}) = (0, \dots, 0, 1)$$

And so we have

$$\begin{aligned} R(z) &= 1 + \frac{-z}{z}(0, \dots, 0, 1)e + O\left(\frac{1}{\|zA\|}\right) \\ &= 1 - 1 + O\left(\frac{1}{\|zA\|}\right) = 0 + O\left(\frac{1}{\|zA\|}\right) \end{aligned}$$

So $R(-\infty)$ equals 0 (and our method is L-stable).

Another choice is to compute the limit $\lim_{z \rightarrow -\infty} R(z)$.

5. Obtain the coefficients $b_i(\theta)$ of the continuous output in Dormand-Prince (5,4) (Note: Dormand-Prince 4(5) is used in matlab function `ODE45`.)

The formula used for the continuous output is (vector form)

$$\begin{aligned} \text{CONTD5} = \text{CON} + \theta \cdot \{ & \text{CON1} + (1 - \theta) \cdot [\text{CON2} \\ & + \theta \cdot (\text{CON3} + (1 - \theta) \cdot \text{CON4}) \}, \end{aligned}$$

where

$$y = y_0 + h \sum_{i=1, i \neq 2}^6 A_{7i} k_i \quad (\text{note: no } k_2, k_7)$$

$$\text{CON} = y_0$$

$$\text{CON1} = y - y_0 = h \sum_{i=1, i \neq 2}^6 A_{7i} k_i$$

$$\text{CON2} = h k_1 - \text{CON1} = h \left(k_1 - \sum_{i=1, i \neq 2}^6 A_{7i} k_i \right)$$

$$\text{CON3} = -h k_2 + \text{CON1} - \text{CON2} = h \left(-k_1 - k_2 + 2 \sum_{i=1, i \neq 2}^6 A_{7i} k_i \right)$$

$$\text{CON4} = h \left(D_1 k_1 + \sum_{i=3}^6 D_i k_i + D_7 k_2 \right) \quad (\text{note: no } D_2, k_7)$$

where A_{7i} and D_i are given in the subroutine `cdopri`. Thus

$$\begin{aligned} & \text{CONTD5} \\ = & y_0 + \theta \cdot \left\{ h \sum_{i=1, i \neq 2}^6 A_{7i} k_i + (1 - \theta) \cdot \left[h \left(k_1 - \sum_{i=1, i \neq 2}^6 A_{7i} k_i \right) \right. \right. \\ & \left. \left. + \theta \cdot \left(h \left(-k_1 - k_2 + 2 \sum_{i=1, i \neq 2}^6 A_{7i} k_i \right) + (1 - \theta) \cdot h \left(D_1 k_1 + \sum_{i=1, i \neq 2}^6 D_i k_i + D_7 k_2 \right) \right) \right] \right\} \\ = & y_0 + h \{ \underbrace{[\theta A_{71} + \theta(1 - \theta)(1 - A_{71}) + \theta^2(1 - \theta)(-1 + 2A_{71}) + \theta^2(1 - \theta)^2 D_1]}_{\text{coefficient of } k_1} k_1 \\ & + \underbrace{[-\theta^2(1 - \theta) + \theta^2(1 - \theta)^2 D_7]}_{\text{coefficient of } k_2} k_2 \\ & + \sum_{i=3}^6 \underbrace{[\theta A_{7i} - \theta(1 - \theta)A_{7i} + 2\theta^2(1 - \theta)A_{7i} + \theta^2(1 - \theta)^2 D_i]}_{\text{coefficient of } k_i} k_i + \underline{0} \cdot k_7 \} \\ \equiv & y_0 + h \sum_{i=1}^7 \underline{b_i(\theta)} k_i \end{aligned}$$