Jeffrey Hellrung
Numerical Analysis Qualifying Exam, Winter 2003

1. Let $f(0)$, $f(h)$, and $f(2h)$ be the values of a real-valued function at $x = 0$, $x = h$, and $x = 2h$.

   (a) Derive the coefficients c_0, c_1, and c_2 so that

   $$Df_h = c_0 f(0) + c_1 f(h) + c_2 f(2h)$$

   is as accurate an approximation to $f'(0)$ as possible.

   (b) Derive the leading term of a truncation error estimate for the formula you derived in (a).

   Solution

   (a) By Taylor's Theorem,

   $$\begin{aligned} f(h) &= f(0) + f'(0)h + \frac{1}{2}f''(0)h^2 + \frac{1}{6}f^{(3)}(\alpha_1)h^3 \\ f(2h) &= f(0) + 2f'(0)h + 2f''(0)h^2 + \frac{4}{3}f^{(3)}(\alpha_2)h^3 \end{aligned}$$

   for some $\alpha_1, \alpha_2 \in [0, 2h]$. Thus,

   $$\begin{aligned} Df_h &= c_0 f(0) + c_1 f(h) + c_2 f(2h) \\ &= (c_0 + c_1 + c_2)f(0) + (c_1 + 2c_2)f'(0)h \\ &+ \left(\frac{1}{2}c_1 + 2c_2\right) f''(0)h^2 + \left(\frac{1}{6}c_1 f^{(3)}(\alpha_1) + \frac{4}{3}c_2 f^{(3)}(\alpha_2)\right) h^3. \end{aligned}$$

   We require that

   $$\begin{aligned} c_0 + c_1 + c_2 &= 0; \\ c_1 + 2c_2 &= 1; \\ \frac{1}{2}c_1 + 2c_2 &= 0; \end{aligned}$$

   which solves to

   $$\begin{aligned} c_0 &= -\frac{3}{2} \\ c_1 &= 2 \\ c_2 &= -\frac{1}{2}. \end{aligned}$$

   (b) The leading term of the truncation error is

   $$\left(\frac{1}{6}c_1 f^{(3)}(\alpha_1) + \frac{4}{3}c_2 f^{(3)}(\alpha_2)\right) h^3 = \left(\frac{1}{3}f^{(3)}(\alpha_1) - \frac{2}{3}f^{(3)}(\alpha_2)\right) h^3.$$

2. (a) Find and solve the normal equations used to determine the coefficients for a straight line that fits the following data in the least squares sense.

x_i	$f(x_i)$
-1	2
0	3
1	3
2	4

(b) Let A be an $m \times n$ matrix, with $m > n$, and the columns of A being linearly independent. Given the QR factorization of A, where Q's columns are orthonormal and R is upper triangular, what equations must you solve to find the least squares solution of the over-determined system of equations $Ax = b$?

(c) Show that the Gram-Schmidt orthogonalization process applied to the columns of A leads to a QR factorization of the matrix A. (Specifically, give the elements of Q and R when the Gram-Schmidt process is written in matrix form.)

Solution

(a) Given $f(x) = ax + b$, the data gives the linear least-squares problem

$$\begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 2 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 3 \\ 4 \end{pmatrix},$$

which has the solution

$$\begin{aligned} \begin{pmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 0 & 1 \\ 2 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} &= \begin{pmatrix} -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 3 \\ 4 \end{pmatrix} \\ \Rightarrow \begin{pmatrix} 6 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} &= \begin{pmatrix} 9 \\ 12 \end{pmatrix} \\ \Rightarrow \begin{pmatrix} a \\ b \end{pmatrix} &= \begin{pmatrix} 3/5 \\ 27/10 \end{pmatrix}. \end{aligned}$$

(b) The linear least-squares solution to $Ax = b$ is

$$A^T A x = A^T b.$$

If $A = QR$ is the QR factorization of A, and A's columns are independent, then R is nonsingular, hence

$$\begin{aligned} & A^T A x = A^T b \\ \Rightarrow \quad & (R^T Q^T) Q R x = (R^T Q^T) b \\ \Rightarrow \quad & R x = Q^T b. \end{aligned}$$

(c) Let

$$A = \begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix};$$

then the Gram-Schmidt process yields vectors $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ as follows:

$$\begin{array}{rclrcl} p_1 &=& a_1, & q_1 &=& p_1/\|p_1\| \\ p_2 &=& a_2 - (q_1^T a_2) q_1, & q_2 &=& p_2/\|p_2\| \\ p_3 &=& a_3 - (q_1^T a_3) q_1 - (q_2^T a_3) q_2, & q_3 &=& p_3/\|p_3\| \\ \vdots & & & \vdots & & \end{array}$$

Note that by construction, the set of vectors $\{p_1, \ldots, p_n\}$ are orthogonal and $\{q_1, \ldots, q_n\}$ are orthonormal. Thus,

$$\begin{pmatrix} a_1 & a_2 & a_3 & \cdots \end{pmatrix} = \begin{pmatrix} q_1 & q_2 & q_3 & \cdots \end{pmatrix} \begin{pmatrix} \|p_1\| & (q_1^T a_2) & (q_1^T a_3) & \cdots \\ 0 & \|p_2\| & (q_2^T a_3) & \cdots \\ 0 & 0 & \|p_3\| & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = QR,$$

as desired.

3. Consider the scalar function $f : \mathbb{R} \to \mathbb{R}$. Let x^* be a root of f and x^n be an approximation to that root.

   (a) Derive the formula for getting a "better" approximation to the root by setting x^{n+1} to be the root of the linear approximation to f obtained from the first two terms of the Taylor Series approximation to f at x^n.
   (b) What is the common name for the method you have derived?
   (c) Consider $F : \mathbb{R}^n \to \mathbb{R}^n$. Using the approach in (a), derive a vector iteration for solving $F(x) = 0$.

   Solution

   (a) The linear approximation to f at x^n is $\ell(x) = f(x^n) + f'(x^n)(x - x^n)$, so we wish to satisfy $\ell(x^{n+1}) = 0$, giving $x^{n+1} = x^n - f'(x^n)^{-1} f(x^n)$.
   (b) The common name for this method is Newton's method.
   (c) All the notation from (a) stands, so we obtain $x^{n+1} = x^n - DF(x_n)^{-1} F(x^n)$.

4. Consider the theta method

$$y_{i+1} = y_i + h\left(\theta f(t_i, y_i) + (1-\theta) f(t_{i+1}, y_{i+1})\right)$$

   to approximate the solution of the ordinary differential equation $y'(t) = f(t, y(t))$.

   (a) Find the order of the method as a function of the values of the parameter θ.
   (b) Determine all values of θ such that the theta method is A-stable.
   (c) What particular method is obtained for $\theta = 1$? Prove convergence of the method in this case $\theta = 1$ and state the necessary assumptions.

   Solution

   (a) (W05.4(a))
   (b) (W05.4(c))
   (c) The theta method for $\theta = 1$ is simply the forward Euler method.
   Fix some time interval $[0, T]$, and assume that f is smooth and Lipschitz continuous in both t and y, i.e., there exists $L > 0$ such that

$$|f(t_1, y_1) - f(t_2, y_2)| \le L\left(|t_1 - t_2| + |y_1 - y_2|\right)$$

   for all $t_1, t_2 \in [0, T]$ and $y_1, y_2 \in \mathbb{R}$. Further, let $M = \sup_{t \in [0,T]} |f(t, y(t))| < \infty$.
   By Taylor's thereom, the local truncation error τ_{i+1} is

$$\begin{aligned}
\tau_{i+1} &= y(t_{i+1}) - (y(t_i) + h f(t_1, y(t_i))) \\
&= \frac{1}{2} h^2 \left(f_t(\alpha_{i+1}, y(\alpha_{i+1})) + f_y(\alpha_{i+1}, y(\alpha_{i+1})) f(\alpha_{i+1}, y(\alpha_{i+1}))\right)
\end{aligned}$$

   for some $\alpha_{i+1} \in [t_i, t_{i+1}]$, so that

$$|\tau_{i+1}| \le \frac{1}{2} L(M+1) h^2.$$

   It follows that the global error e_{i+1} is

$$\begin{aligned}
e_{i+1} &= y(t_{i+1}) - y_{i+1} \\
&= y(t_{i+1}) - (y(t_i) + h f(t_i, y(t_i))) + (y(t_i) + h f(t_i, y(t_i))) - (y_i + h f(t_i, y_i)) \\
&= \tau_{i+1} + h\left(f(t_i, y(t_i)) - f(t_i, y_i)\right) + y(t_i) - y_i \\
&= \tau_{i+1} + h\left(f(t_i, y(t_i)) - f(t_i, y_i)\right) + e_i,
\end{aligned}$$

so that

$$\begin{aligned}
|e_{i+1}| &\leq |\tau_{i+1}| + Lh|y(t_i) - y_i| + |e_i| \\
&= |\tau_{i+1}| + (1 + Lh)|e_i|.
\end{aligned}$$

Expanding the recurrence relation and summing the geometric series results in

$$\begin{aligned}
|e_i| &\leq (1 + Lh)^i|e_0| + \sum_{j=1}^{i} |\tau_j|(1 + Lh)^{i-j} \\
&\leq (1 + Lh)^i|e_0| + \frac{1}{2}L(M + 1)h^2 \sum_{j=1}^{i}(1 + Lh)^{i-j} \\
&= (1 + Lh)^i|e_0| + \frac{1}{2}L(M + 1)h^2 \sum_{j=0}^{i-1}(1 + Lh)^{j} \\
&= (1 + Lh)^i|e_0| + \frac{1}{2}L(M + 1)h^2 \frac{(1 + Lh)^i - 1}{(1 + Lh) - 1}.
\end{aligned}$$

To obtain a cleaner estimate, we note that $1 + Lh < e^{Lh}$, hence

$$\begin{aligned}
|e_i| &< \left(e^{Lh}\right)^i |e_0| + \frac{1}{2}L(M + 1)h^2 \frac{\left(e^{Lh}\right)^i - 1}{Lh} \\
&= e^{Lih}|e_0| + \frac{1}{2}(M + 1)h\left(e^{Lih} - 1\right) \\
&\leq e^{Lih}|e_0| + \frac{1}{2}(M + 1)Lihe^{Lih}h.
\end{aligned}$$

Now since $0 \leq i \leq N$, where $Nh = T$, we finally get

$$|e_i| \leq e^{LT}\left(|e_0| + \frac{1}{2}(M + 1)LTh\right) \to e^{LT}|e_0|$$

as $h \to 0$. Thus, assuming $e_0 = 0$, we get $e_i \to 0$ as $h \to 0$, proving convergence.

5. To solve

$$u_t + au_x = 0 \text{ for } t > 0,\ 0 \leq x \leq 1;$$

$u(x, 0) = \phi(x)$ smooth; u periodic in x, i.e., $u(x + 1, t) = u(x, t)$; we use

$$\frac{1}{2\Delta t}\left(\left(v_j^{n+1} + v_{j+1}^{n+1}\right) - \left(v_j^n + v_{j+1}^n\right)\right) + \frac{a}{2\Delta x}\left(v_{j+1}^{n+1} - v_j^{n+1} + v_{j+1}^n - v_j^n\right) = 0.$$

For what values of $\frac{\Delta t}{\Delta x}$, if any, does this converge? At what rate? Explain your answers.

Solution

The symbol $p(s, \xi)$ of the differential operator $P = \partial_t + a\partial_x$ is

$$\begin{aligned}
p(s, \xi) &= P\left(e^{st}e^{i\xi x}\right)/\, e^{st}e^{i\xi x} \\
&= s + ia\xi,
\end{aligned}$$

while the symbol $p_{\Delta t,\Delta x}(s, \xi)$ of the difference operator $P_{\Delta t,\Delta x}$ corresponding to the scheme is

$$\begin{aligned}
p_{\Delta t,\Delta x}(s, \xi) &= P_{\Delta t,\Delta x}\left(e^{s\Delta tn}e^{i\xi\Delta xj}\right)/\, e^{s\Delta tn}e^{i\xi\Delta xj} \\
&= \frac{1}{2\Delta t}\left(e^{s\Delta t} - 1\right)\left(e^{i\xi\Delta x} + 1\right) + \frac{a}{2\Delta x}\left(e^{s\Delta t} + 1\right)\left(e^{i\xi\Delta x} - 1\right) \\
&= \frac{1}{2\Delta t}\left(s\Delta t + O(\Delta t^2)\right)(2 + O(\Delta x)) + \frac{a}{2\Delta x}(2 + O(\Delta t))\left(i\xi\Delta x + O(\Delta x^2)\right) \\
&= s + ia\xi + O(\Delta t) + O(\Delta x) \\
&= p(s, \xi) + O(\Delta t) + O(\Delta x),
\end{aligned}$$

showing first-order accuracy.

By the Lax-Richtmyer Equivalence Theorem, the scheme is convergent if and only if it is stable. We thus replace $g = e^{s\Delta t}$ in $p_{\Delta t,\Delta x} = 0$ and solve for g to determine the roots of the amplification polynomial:

$$\begin{aligned}
&\frac{1}{2\Delta t}(g-1)\left(e^{i\xi\Delta x}+1\right)+\frac{a}{2\Delta x}(g+1)\left(e^{i\xi\Delta x}-1\right)=0\\
\Rightarrow\quad &(g-1)\left(e^{i\theta}+1\right)+a\lambda(g+1)\left(e^{i\theta}-1\right)=0\\
\Rightarrow\quad &\left((1+a\lambda)e^{i\theta}+(1-a\lambda)\right)g=(1-a\lambda)e^{i\theta}+(1+a\lambda)\\
\Rightarrow\quad &g=\frac{(1+a\lambda)e^{i\theta/2}+(1-a\lambda)e^{-i\theta/2}}{(1+a\lambda)e^{-i\theta/2}+(1-a\lambda)e^{i\theta/2}},
\end{aligned}$$

from which we see immediately that $|g| = 1$ for all choices of θ, λ, and so the scheme is unconditionally stable, hence convergent.

6. Consider the differential equation

$$u_t = u_{xx} + u_{yy} + bu_{xy} \text{ for } t > 0,\ 0 < x < 1,\ 0 < y < 1;$$

with $u = 0$ on the boundary; and $u(x, y, 0) = \phi(x, y)$ a smooth function.

(a) For what values of b can you obtain a convergent, unconditionally stable finite difference scheme?

(b) Construct such a scheme. Explain your answers.

Solution

(a) (W05.6(a))

(b) We consider using Crank-Nicolson:

$$\begin{aligned}
P_{k,h_x,h_y}u^n_{\ell,m} &= D_{t+}u^n_{\ell,m}-\frac{1}{2}\left(D_x^2u^{n+1}_{\ell,m}+D_x^2u^n_{\ell,m}\right)-\frac{1}{2}\left(D_y^2u^{n+1}_{\ell,m}+D_y^2u^n_{\ell,m}\right)\\
&- \frac{b}{2}\left(D_{x0}D_{y0}u^{n+1}_{\ell,m}+D_{x0}D_{y0}u^n_{\ell,m}\right)\\
&= \frac{u^{n+1}_{\ell,m}-u^n_{\ell,m}}{k}\\
&- \frac{1}{2}\left(\frac{u^{n+1}_{\ell+1,m}-2u^{n+1}_{\ell,m}+u^{n+1}_{\ell-1,m}}{h_x^2}+\frac{u^n_{\ell+1,m}-2u^n_{\ell,m}+u^{n+1}_{\ell-1,m}}{h_x^2}\right)\\
&- \frac{1}{2}\left(\frac{u^{n+1}_{\ell,m+1}-2u^{n+1}_{\ell,m}+u^{n+1}_{\ell,m-1}}{h_y^2}+\frac{u^n_{\ell,m+1}-2u^n_{\ell,m}+u^{n+1}_{\ell,m-1}}{h_y^2}\right)\\
&- \frac{b}{2}\left(\frac{u^{n+1}_{\ell+1,m+1}-u^{n+1}_{\ell+1,m-1}-u^{n+1}_{\ell-1,m+1}+u^{n+1}_{\ell-1,m-1}}{4h_xh_y}\right.\\
&\quad\left.+\ \frac{u^n_{\ell+1,m+1}-u^n_{\ell+1,m-1}-u^n_{\ell-1,m+1}+u^n_{\ell-1,m-1}}{4h_xh_y}\right);\\
R_{k,h_x,h_y}f^n_{\ell,m} &= \frac{1}{2}\left(f^{n+1}_{\ell,m}+f^n_{\ell,m}\right).
\end{aligned}$$

The symbols $p_{k,h_x,h_y}(s,\xi,\eta)$ and $r_{k,h_x,h_y}(s,\xi,\eta)$ of the difference operators P_{k,h_x,h_y} and R_{k,h_x,h_y}, respectively, are

$$
\begin{aligned}
p_{k,h_x,h_y}(s,\xi,\eta) &= P_{k,h_x,h_y}\left(e^{skn}e^{i(\xi h_X\ell+\eta h_y)}\right)\Big/ e^{skn}e^{i(\xi h_X\ell+\eta h_y)}\\
&= \frac{1}{k}\left(e^{sk}-1\right)\\
&- \frac{1}{2h_x^2}\left(e^{sk}+1\right)\left(e^{i\xi h_x}-2+e^{-i\xi h_x}\right)\\
&- \frac{1}{2h_y^2}\left(e^{sk}+1\right)\left(e^{i\eta h_y}-2+e^{-i\eta h_y}\right)\\
&- \frac{b}{8h_xh_y}\left(e^{sk}+1\right)\left(e^{i\xi h_x}-e^{-i\xi h_x}\right)\left(e^{i\eta h_y}-e^{-i\eta h_y}\right)\\
&= \frac{1}{k}\left(e^{sk}-1\right)\\
&+ \left(e^{sk}+1\right)\left(\frac{1}{h_x^2}(1-\cos\xi h_x)+\frac{1}{h_y^2}(1-\cos\eta h_y)+\frac{b}{2h_xh_y}\sin\xi h_x\sin\eta h_y\right);\\
r_{k,h_x,h_y}(s,\xi,\eta) &= R_{k,h_x,h_y}\left(e^{skn}e^{i(\xi h_X\ell+\eta h_y)}\right)\Big/ e^{skn}e^{i(\xi h_X\ell+\eta h_y)}\\
&= \frac{1}{2}\left(e^{sk}+1\right).
\end{aligned}
$$

We now note that

$$\frac{1}{k}\left(e^{sk}-1\right)=\frac{1}{2}\left(e^{sk}+1\right)s+O(k^2),$$

so that, by expanding p_{k,h_x,h_y} out via Taylor's Theorem,

$$p_{k,h_x,h_y}(s,\xi,\eta)=\frac{1}{2}\left(e^{sk}+1\right)\left(s+\xi^2+\eta^2+b\xi\eta\right)+O(k^2)+O(h_x^2)+O(h_y^2).$$

It is now easy to see that this agrees with $r_{k,h_x,h_y}(s,\xi,\eta)p(s,\xi,\eta)$ to $O(k^2)+O(h_x^2)+O(h_y^2)$, where $p(s,\xi,\eta)$ is the symbol of the differential operator $P=\partial_t-\partial_x^2-\partial_y^2-b\partial_{xy}$ (refer to (a)). This shows that the scheme is second-order.

For the stability analysis, we replace $g=e^{sk}$ in $p_{k,h_x,h_y}(s,\xi,\eta)=0$ and solve for g to determine the roots of the amplification polynomial:

$$
\begin{aligned}
&\frac{1}{k}(g-1)+(g+1)\left(\frac{1}{h_x^2}(1-\cos\xi h_x)+\frac{1}{h_y^2}(1-\cos\eta h_y)+\frac{b}{2h_xh_y}\sin\xi h_x\sin\eta h_y\right)=0\\
&\Rightarrow\quad g-1+(g+1)\left(\mu_x(1-\cos\theta)+\mu_y(1-\cos\phi)+\frac{1}{2}b\sqrt{\mu_x}\sin\theta\sqrt{\mu_y}\sin\phi\right)=0.
\end{aligned}
$$

Let $c=\mu_x(1-\ldots\sqrt{\mu_y}\sin\phi$ to simplify the notation. Then

$$g=\frac{1-c}{1+c},$$

hence $|g|\le 1$ if and only if $c\ge 0$. Indeed,

$$
\begin{aligned}
c &= \mu_x(1-\cos\theta)+\mu_y(1-\cos\phi)+\frac{1}{2}b\sqrt{\mu_x}\sin\theta\sqrt{\mu_y}\sin\phi\\
&= \left(\sqrt{\mu_x}\sin\theta\right)^2\frac{1-\cos\theta}{\sin^2\theta}+\left(\sqrt{\mu_y}\sin\phi\right)^2\frac{1-\cos\phi}{\sin^2\phi}+\frac{1}{2}b\left(\sqrt{\mu_x}\sin\theta\right)\left(\sqrt{\mu_y}\sin\phi\right)\\
&= \left(\sqrt{\mu_x}\sin\theta\right)^2\frac{1}{1+\cos\theta}+\left(\sqrt{\mu_y}\sin\phi\right)^2\frac{1}{1+\cos\phi}+\frac{1}{2}b\left(\sqrt{\mu_x}\sin\theta\right)\left(\sqrt{\mu_y}\sin\phi\right)\\
&\ge \frac{1}{2}\left(\left(\sqrt{\mu_x}\sin\theta\right)^2+\left(\sqrt{\mu_y}\sin\phi\right)^2+b\left(\sqrt{\mu_x}\sin\theta\right)\left(\sqrt{\mu_y}\sin\phi\right)\right)\\
&\ge 0
\end{aligned}
$$

if $-2 \leq b \leq 2$, as required for well-posedness. Therefore, the scheme is unconditionally stable.

7. (a)

   (b)

   Solution

   (a)

   (b)