Jeffrey Hellrung
Numerical Analysis Qualifying Exam, Fall 2004

1. (5 Pts.) Let $\overline{x}$ be a root of a continuously differentiable function $f(x) : \mathbb{R} \to \mathbb{R}$. If x^* is an approximate root, then

   (a) Derive an expression that relates the magnitude of the residual at x^* to the magnitude of the error of the root x^*.

   (b) Give an example of a function where the magnitude of the residual at x^* over-estimates the error of the root x^*.

   (c) Give an example of a function where the magnitude of the residual at x^* under-estimates the error of the root x^*.

   Solution

   (a) By Taylor's Theorem,
   $$f(x^*) = f(x) + f'(\alpha)(x^* - x) = f'(\alpha)(x^* - x)$$
   for some α between x^* and x, so
   $$|f(x^*)| = |f'(\alpha)||x^* - x|.$$

   (b) $f(x) = 2x$

   (c) $f(x) = \frac{1}{2}x$

2. (5 Pts.) Consider the integration formula
   $$\int_{-1}^{1} f(x)dx \approx f(\alpha_1)\beta + f(\alpha_2)\beta.$$

   (a) Determine α_1, α_2, and β so that this formula is exact for all quadratic polynomials.

   (b) What is the minimal degree polynomial for which the formula with the coefficients derived in (a) is not exact?

   (c) What is the expected order of a composite integration method based upon the formula with coefficients derived in (a)?

   Solution

   (a) Let $f(x) = ax^2 + bx + c$. Then, on the one hand,
   $$\int_{-1}^{1} f(x)dx = \frac{1}{3}ax^3 + \frac{1}{2}bx^2 + cx\Big|_{-1}^{1} = \frac{2}{3}a + 2c,$$
   and on the other hand,
   $$\beta\left(f(\alpha_1) + f(\alpha_2)\right) = \beta(\alpha_1^2 + \alpha_2^2)a + \beta(\alpha_1 + \alpha_2)b + 2\beta c.$$
   Equating coefficients on the above two expression, we find that $\alpha_1 = -\alpha_2 = \frac{1}{\sqrt{3}}$ and $\beta = 1$.

   (b) 3

   (c) Given that we wish to integrate f over $[-h, h]$, we would approximate f by a quadratic q to $O(h^3)$ over $[-h, h]$. The integral over $[-h, h]$ of f could then be approximated by the integral of q over $[-h, h]$ to $O(h^4)$. The composite integration scheme would then be accurate to $O(h^3)$.

3. (5 Pts.) Let $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$ with $m > n$. For $\sigma > 0$ consider the following minimization problem:

$$\min_{x\in\mathbb{R}^m} \left(\|Ax - b\|_2^2 + \sigma^2\|x\|_2^2\right).$$

Derive the equation that the optimal solution satisfies and explain why the optimal solution is unique.

Solution

Let

$$F(x) = \|Ax - b\|_2^2 + \sigma^2\|x\|_2^2$$

and suppose x^* minimizes F. Then for any $\epsilon \in \mathbb{R}^m$,

$$F(x^*) \le F(x^* + \epsilon).$$

In other words, if $g(\epsilon) = F(x^* + \epsilon)$, g has a minimum at $\epsilon = 0$, i.e., $(\nabla g)(0) = 0$. We thus compute

$$\begin{aligned} \nabla g &= \nabla\left(\|A(x^* + \epsilon) - b\|_2^2 + \sigma^2\|x^* + \epsilon\|_2^2\right) \\ &= 2A^T(A(x^* + \epsilon) - b) + 2\sigma^2(x^* + \epsilon), \end{aligned}$$

and at $\epsilon = 0$,

$$0 = (\nabla g)(0) = 2\left((A^TA + \sigma^2 I)x^* - A^Tb\right)$$

and we find that x^* satisfies $(A^TA + \sigma^2I)x^* = A^Tb$.

We will show that all eigenvalues of $A^TA+\sigma^2I$ are positive, hence $A^TA+\sigma^2I$ will be nonsingular, and it will follow that x^* is unique. To this end, let λ be an eigenvalue of $A^TA+\sigma^2I$ and x a corresponding eigenvector. Then

$$\begin{aligned} &(A^TA + \sigma^2I)x = \lambda x \\ &\quad\Rightarrow\quad x^TA^TAx + x^T\sigma^2x = x^T\lambda x \\ &\quad\Rightarrow\quad \|Ax\|_2^2 + \sigma^2\|x\|_2^2 = \lambda\|x\|_2^2 \\ &\quad\Rightarrow\quad \lambda = \sigma^2 + \frac{\|Ax\|_2^2}{\|x\|_2^2} > 0 \end{aligned}$$

since $\sigma > 0$. By the previous comments, we have that x^* is unique.

4. (10 Pts.) Show that the one-step method given by

$$\begin{aligned} k_1 &= f(t^n, y^n), \\ k_2 &= f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2}k_1\right), \\ k_3 &= f\left(t^n + h, y^n + h(-k_1 + 2k_2)\right), \\ y^{n+1} &= y^n + \frac{h}{6}\left(k_1 + 4k_2 + k_3\right) \end{aligned}$$

for solving $y'(t) = f(t, y(t))$ is third-order.

Solution

To simplify the notation, let f, f_t, f_y, etc.; denote $f(t^n, y^n)$, $f_t(t^n, y^n)$, $f_y(t^n, y^n)$, etc.; respectively. We use Taylor's Theorem to expand each intermediate variable out to $O(h^3)$:

$$k_1 = f$$

$$
\begin{aligned}
k_2 &= f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2}k_1\right) \\
&= f + f_t\frac{h}{2} + f_y\frac{h}{2}k_1 + \frac{1}{2}f_{tt}\frac{h^2}{4} + f_{ty}\frac{h^2}{4}k_1 + \frac{1}{2}f_{yy}\frac{h^2}{4}k_1^2 + O(h^3) \\
&= f + \left(\frac{1}{2}f_t + \frac{1}{2}f_yf\right)h + \left(\frac{1}{8}f_{tt} + \frac{1}{4}f_{ty}f + \frac{1}{8}f_{yy}f^2\right)h^2 + O(h^3)
\end{aligned}
$$

$$
\begin{aligned}
k_3 &= f\left(t^n + h, y^n + h(-k_1 + 2k_2)\right) \\
&= f + f_th + f_yh(-k_1 + 2k_2) + \frac{1}{2}f_{tt}h^2 + f_{ty}h^2(-k_1 + 2k_2) + \frac{1}{2}f_{yy}h^2(-k_1 + 2k_1)^2 + O(h^3).
\end{aligned}
$$

We work on simplifying out each term in k_3 with $-k_1 + 2k_2$ separately and only up to $O(h^3)$:

$$
\begin{aligned}
f_yh(-k_1 + 2k_2) &= f_yh\left(f + (f_t + f_yf)h + O(h^2)\right) \\
&= f_yfh + \left(f_tf_y + f_y^2f\right)h^2 + O(h^3) \\
f_{ty}h^2(-k_1 + 2k_2) &= f_{ty}h^2\left(f + O(h)\right) \\
&= f_{ty}fh^2 + O(h^3) \\
\frac{1}{2}f_{yy}h^2(-k_1 + 2k_1)^2 &= \frac{1}{2}f_{yy}h^2\left(f + O(h)\right)^2 \\
&= \frac{1}{2}f_{yy}f^2h^2 + O(h^3).
\end{aligned}
$$

The expression for k_3 now becomes

$$
\begin{aligned}
k_3 &= f + f_th + f_yh(-k_1 + 2k_2) + \frac{1}{2}f_{tt}h^2 + f_{ty}h^2(-k_1 + 2k_2) + \frac{1}{2}f_{yy}h^2(-k_1 + 2k_1)^2 + O(h^3) \\
&= f + f_th + \left(f_yfh + \left(f_tf_y + f_y^2f\right)h^2\right) + \frac{1}{2}f_{tt}h^2 + \left(f_{ty}fh^2\right) + \left(\frac{1}{2}f_{yy}f^2h^2\right) + O(h^3) \\
&= f + (f_t + f_yf)h + \left(f_tf_y + f_y^2f + \frac{1}{2}f_{tt} + f_{ty}f + \frac{1}{2}f_{yy}f^2\right)h^2 + O(h^3).
\end{aligned}
$$

We can now evaluate the quantity $k_1 + 4k_2 + k_3$:

$$
\begin{aligned}
k_1 + 4k_2 + k_3 &= f \\
&+ 4\left(f + \left(\frac{1}{2}f_t + \frac{1}{2}f_yf\right)h + \left(\frac{1}{8}f_{tt} + \frac{1}{4}f_{ty}f + \frac{1}{8}f_{yy}f^2\right)h^2 + O(h^3)\right) \\
&+ f + (f_t + f_yf)h + \left(f_tf_y + f_y^2f + \frac{1}{2}f_{tt} + f_{ty}f + \frac{1}{2}f_{yy}f^2\right)h^2 + O(h^3) \\
&= 6f + 3(f_t + f_yf)h + \left(f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_yf_t + f_y^2f\right)h^2 + O(h^3)
\end{aligned}
$$

and so

$$
\begin{aligned}
y^{n+1} &= y^n + \frac{h}{6}(k_1 + 4k_2 + k_3) \\
&= y^n + fh + \frac{1}{2}(f_t + f_yf)h^2 + \frac{1}{6}\left(f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_yf_t + f_y^2f\right)h^3 + O(h^4).
\end{aligned}
$$

It is easy to see that this agrees with $y(t^n + h)$ to $O(h^4)$:

$$
\begin{aligned}
y'(t^n) &= f \\
y''(t^n) &= f_t + f_yf \\
y^{(3)}(t^n) &= f_{tt} + f_{ty}f + f_{ty}f + f_{yy}f^2 + f_yf_t + f_y^2f;
\end{aligned}
$$

therefore the method is indeed third-order.

5. (10 Pts.) Given the second-order partial differential equation

$$u_{tt} + 2bu_{tx} = a^2 u_{xx} + cu_x + du_t + eu + f(t, x)$$

to be solved for $t > 0$, $0 \le x \le 2\pi$, with $u(x, t)$ periodic in x of period 2π:

(a) For what values of a, b is the initial value problem with initial data

$$\begin{aligned} u(x, 0) &= u_0(x) \\ u_t(x, 0) &= u_1(x) \end{aligned}$$

well-posed?

(b) Write a stable convergent finite difference approximation for this problem. Justify your answer.

Hint: You might consider making this into a first-order system of equations.

Solution

(a) The symbol $p(s, \xi)$ of the differential operator $P = \partial_t^2 + 2b\partial_{tx} - a^2\partial_x^2 - c\partial_x - d\partial_t - e$ is

$$\begin{aligned} p(s, \xi) &= P\left(e^{st}e^{i\xi x}\right) / e^{st}e^{i\xi x} \\ &= s^2 + 2ibs\xi + a^2\xi^2 - ic\xi - ds - e \\ &= s^2 + (2ib\xi - d)s + a^2\xi^2 - ic\xi - e. \end{aligned}$$

The roots of the symbol (as a function of s) are then

$$\begin{aligned} q_\pm(\xi) &= \frac{1}{2}\left(d - 2ib\xi \pm \sqrt{(2ib\xi - d)^2 - 4\left(a^2\xi^2 - ic\xi - e\right)}\right) \\ &= \frac{d}{2} - ib\xi \pm \sqrt{d^2 + 4e + 4i(c - bd)\xi - 4(a^2 + b^2)\xi^2}. \end{aligned}$$

Well-posedness requires that $\Re(q_\pm)$ be bounded above for all ξ. But this is indeed the case regardless of the values of $a, \ldots, e$ so long as $a^2 + b^2 > 0$. For certainly

$$\Re\left(\frac{d}{2} - ib\xi\right) = \frac{d}{2}$$

is bounded, while for large enough $|\xi|$,

$$4(a^2 + b^2)\xi^2 - d^2 - 4e > 4|c - bd||\xi|,$$

and hence the square root, for large enough $|\xi|$, has negative real part. It follows by continuity that

$$\Re\left(\sqrt{d^2 + 4e + 4i(c - bd)\xi - 4(a^2 + b^2)\xi^2}\right)$$

is bounded for all $\xi \in \mathbb{R}$, hence the problem is well-posed for any $a, \ldots, e$ so long as $a^2 + b^2 > 0$.
Alternatively, we can rewrite the equation as a system. Introduce

$$U(t, x) = \begin{pmatrix} u(x, t) \\ u_t(x, t) \\ u_x(x, t) \end{pmatrix};$$

then we find that

$$U_t = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -2b & a^2 \\ 0 & 1 & 0 \end{pmatrix} U_x + \begin{pmatrix} 0 & 1 & 0 \\ e & d & c \\ 0 & 0 & 0 \end{pmatrix} U + \begin{pmatrix} 0 \\ f(t, x) \\ 0 \end{pmatrix}.$$

For well-posedness, we can ignore the inhomogeneous and lower-order terms (U) (as long as not both a and b are 0), and thus consider the well-posedness of the system

$$U_t = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -2b & a^2 \\ 0 & 1 & 0 \end{pmatrix} U_x.$$

We then have, after a Fourier transform in space, $\widehat{U}_t = Q(\xi)\widehat{U}$, where

$$Q(\xi) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -2ib\xi & ia^2\xi \\ 0 & i\xi & 0 \end{pmatrix},$$

which has characteristic equation

$$\begin{aligned} \det(qI - Q(\xi)) &= \begin{vmatrix} q & 0 & 0 \\ 0 & q + 2ib\xi & -ia^2\xi \\ 0 & -i\xi & q \end{vmatrix} \\ &= q\left(q^2 + 2ib\xi q + a^2\xi^2\right) \end{aligned}$$

with roots

$$q_0(\xi) = 0, \; q_\pm(\xi) = i\left(-b \pm \sqrt{a^2 + b^2}\right)\xi,$$

which are both purely imaginary for all ξ (i.e., $\Re(q)$ is bounded for all roots q). Therefore, the system is well-posed so long as not both a and b are 0.

(b) We consider using Lax-Wendroff for the equivalent system for U. To simplify the notation, write

$$U_t = AU_x + BU + F.$$

By Taylor's Theorem (where U, U_t, etc.; denote $U(t,x)$, $U_t(t,x)$, etc.),

$$U(t+k,x) = U + U_t k + \frac{1}{2}U_{tt}k^2 + O(k^3).$$

We can substitute the time derivatives of U for space derivatives using $U_t = AU_x + BU + F$:

$$\begin{aligned} U_t &= AU_x + BU + F \\ U_{tt} &= (AU_x + BU + F)_t \\ &= A\left(U_t\right)_x + BU_t + F_t \\ &= A\left(AU_x + BU + F\right)_x + B\left(AU_x + BU + F\right) + F_t \\ &= A^2U_{xx} + (AB + BA)U_x + B^2U + BF + F_t + AF_x, \end{aligned}$$

so

$$\begin{aligned} U(t+k,x) &= U + \left(AU_x + BU + F\right)k \\ &+ \frac{1}{2}\left(A^2U_{xx} + (AB+BA)U_x + B^2U + BF + F_t + AF_x\right)k^2 + O(k^3) \\ &= \frac{1}{2}A^2k^2U_{xx} + \left(Ak + \frac{1}{2}(AB+BA)k^2\right)U_x + \left(I + Bk + \frac{1}{2}B^2k^2\right)U \\ &+ \left(Ik + \frac{1}{2}Bk^2\right)F + \frac{1}{2}k^2F_t + \frac{1}{2}Ak^2F_x + O(k^3). \end{aligned}$$

By using the approximations

$$
\begin{aligned}
U_{xx} &= D_x^2 U_m^n + O(h^2) \\
U_x &= D_{x0} U_m^n + O(h^2) \\
F_t &= D_{t0} F_m^n + O(k^2) \\
F_x &= D_{x0} F_m^n + O(h^2)
\end{aligned}
$$

we obtain a second-order accurate (explicit) scheme.

For the stability analysis, we can safely ignore the lower-order term BU, and thus just consider the system $U_t = AU_x$. The Lax-Wendroff scheme above then simplifies to

$$
\begin{aligned}
U_m^{n+1} &= \frac{1}{2} k^2 A^2 D_x^2 U_m^n + AkD_{x0} U_m^n + U_m^n \\
&= \frac{1}{2} k^2 A^2 \frac{1}{h^2} \left(U_{m+1}^n - 2U_m^n + U_{m-1}^n \right) + Ak \frac{1}{2h} \left(U_{m+1}^n - U_{m-1}^n \right) + U_m^n,
\end{aligned}
$$

and hence the amplification matrix G is (substituting $G^n e^{i\xi mh} = U_m^n$)

$$
\begin{aligned}
G &= \frac{k^2}{2h^2} A^2 \left(e^{i\xi h} - 2 + e^{-i\xi h} \right) + \frac{k}{2h} A \left(e^{i\xi h} - e^{-i\xi h} \right) + 1 \\
&= 1 - \lambda^2 A^2 (1 - \cos\theta) + i\lambda A \sin\theta.
\end{aligned}
$$

Since G is a polynomial in A, it shares precisely the same eigenvectors as A, and the eigenvalues are related by

$$
g = 1 - \lambda^2 q^2 (1 - \cos\theta) + i\lambda q \sin\theta,
$$

for q an eigenvalue of A and g an eigenvalue of G. Thus

$$
\begin{aligned}
|g| &= \left(1 - \lambda^2 q^2 (1 - \cos\theta)\right)^2 + \lambda^2 q^2 \sin^2\theta \\
&= \left(1 - 2\lambda^2 q^2 \sin^2 \frac{\theta}{2}\right)^2 + 4\lambda^2 q^2 \sin^2 \frac{\theta}{2} \cos^2 \frac{\theta}{2} \\
&= 1 - 4\lambda^2 q^2 \sin^2 \frac{\theta}{2} + 4\lambda^4 q^4 \sin^4 \frac{\theta}{2} + 4\lambda^2 q^2 \sin^2 \frac{\theta}{2} \cos^2 \frac{\theta}{2} \\
&= 1 - 4\lambda^2 q^2 \sin^2 \frac{\theta}{2} \left(1 - \lambda^2 q^2 \sin^2 \frac{\theta}{2} - \cos^2 \frac{\theta}{2}\right) \\
&= 1 - 4\lambda^2 q^2 \left(1 - \lambda^2 q^2\right) \sin^4 \frac{\theta}{2}
\end{aligned}
$$

, and we see that $|g| \le 1$ if and only if $|\lambda q| \le 1$. The eigenvalues of A were (effectively) found in (a):

$$
q = 0, \ -b \pm \sqrt{a^2 + b^2}.
$$

Stability of the scheme thus requires

$$
\lambda \left(|b| + \sqrt{a^2 + b^2} \right) \le 1.
$$

By the Lax-Richtmyer Equivalence Theorem, with this choice of λ, the scheme is convergent.

6. (10 Pts.) Consider the equation

$$
u_t = u_{xx} + u_x
$$

to be solved for $t > 0$, $0 \le x \le 2\pi$, with $u(x,t)$ periodic in x of period 2π, and initial data $u(x,0) = u_0(x)$.

Write an unconditionally stable convergent second-order accurate scheme for this equation and prove that your scheme satisfies these properties.

Solution

(W06.6)

7. **Solution**