Department of Mathematics, University of Utah

Analysis of Numerical Methods, II

MATH 6620 – Section 001 – Spring 2024

Homework 2

Time-stepping methods, I

Due Friday, February 9, 2024

Submit your solutions online through Gradescope.

1. (Quadrature rules)

a. Show that

$$\frac{1}{h}\int_0^h u(x)\,\mathrm{d}x \approx u\left(\frac{h}{2}\right),$$

is a second-order approximation in h.

b. Compute (polynomial) interpolatory quadrature weights for the rule,

$$\int_0^1 f(x)\,\mathrm{d}x \approx w_1 f(0) + w_2 f\left(\frac{1}{3}\right) + w_3 f(1)$$

What is the polynomial degree of exactness for this rule?

c. Compute quadrature weights for the rule,

$$f'(0) + \int_{-1}^{1} f(x)\,\mathrm{d}x \approx w_{-1}f(-1) + w_0 f(0) + w_1 f(1).$$

that is exact for all quadratic polynomials.

<u>**Solution**</u>:

a. We'll use Taylor series, expanding around $x = h/2$ to confirm this:

$$\begin{aligned}\frac{1}{h}\int_0^h u(x)\,\mathrm{d}x &= \frac{1}{h}\int_0^h u(h/2) + (x-h/2)u'(h/2) + \frac{1}{2}(x-h/2)^2 u''(h/2) + \mathcal{O}(h^3)\,\mathrm{d}x \\ &= u(h/2) + \frac{1}{u'(h/2)}\int_0^h (x-h/2)\,\mathrm{d}x + \frac{1}{2h}u''(h/2)\int_0^h (x-h/2)^2\,\mathrm{d}x + \mathcal{O}(h^3) \\ &= u(h/2) + \frac{h^2}{24}u''(h/2) + \mathcal{O}(h^3) = u(h/2) + \mathcal{O}(h^2),\end{aligned}$$

as desired.

b. We'll do this using interpolatory approaches, but moment matching or Taylor series approaches would also work. On the nodal set $(x_1, x_2, x_3) = (0, 1/3, 1)$, the cardinal Lagrange functions as,

$$\begin{aligned}\ell_1(x) &= \frac{(x-1/3)(x-1)}{1/3} & \ell_2(x) &= \frac{(x-0)(x-1)}{-2/9}, & \ell_3(x) &= \frac{(x-0)(x-1/3)}{2/3} \\ &= (3x-1)(x-1) & &= -\frac{9}{2}(x^2-x), & &= \frac{1}{2}x(3x-1).\end{aligned}$$

Then integrating these yields the weights:

$$w_1 = \int_0^1 \ell_1(x) = 0, \qquad w_2 = \frac{3}{4}, \qquad w_3 = \frac{1}{4}.$$

c. We'll perform this exercise using moment matching, but again something like an interpolatory approach would work. We will assert the equality,

$$\left[\frac{\mathrm{d}}{\mathrm{d}x} x^j\right]\Bigg|_{x=0} + \int_{-1}^{1} x^j \,\mathrm{d}x = w_{-1}(-1)^j + w_0(0)^j + w_1 1^j,$$

for $j = 0, 1, 2$. This yields the 3 equations:

$$\begin{aligned} w_{-1} + w_0 + w_1 &= 2 \\ -w_{-1} + w_1 &= 1 \\ w_{-1} + w_1 &= \frac{2}{3}, \end{aligned}$$

which has the solution,

$$(w_{-1}, w_0, w_1) = \left(-\frac{1}{6}, \frac{4}{3}, \frac{5}{6}\right)$$

2. (FD convergence)
For the boundary value problem,

$$-u''(x) = f(x), \qquad\qquad u(0) = u(1) = 0,$$

with f given, consider the standard 3-point stencil finite-difference approximation,

$$-D_+D_-u_j = f_j, \qquad\qquad j \in [M],$$

where for a given $M \in \mathbb{N}$,

$$u_j \approx u(x_j), \qquad\qquad x_j \coloneqq jh, \qquad\qquad h \coloneqq \frac{1}{M+1},$$

with $u_0 = u_{M+1} = 0$. Prove that in the scaled ℓ^1 norm,

$$\|\boldsymbol{u}\|_{1,h} \coloneqq h \sum_{j\in[M]} |u_j| \approx \int_0^1 |u(x)| \,\mathrm{d}x,$$

that this scheme is convergent to second order in h. You may cite without proof any results given in class/on the slides.

__**Solution**__: Fixing M, if $\boldsymbol{e} \in \mathbb{R}^M$ has entries $(\boldsymbol{e})_j = u_j - u(x_j)$, then we have shown that, e.g., on slide `D03-S15(c)` that,

$$\|\boldsymbol{e}\|_{2,h} \coloneqq \sqrt{h}\|\boldsymbol{e}\|_2 = \mathcal{O}(h^2),$$

where $\|\cdot\|_2$ is the standard (unscaled) ℓ^2 norm on vectors. (Note that the notation we use here differs from the slides.) If $\|\cdot\|_1$ is the standard (unscaled) norm on size-M vectors, then we have the following equivalence of norms for any $\boldsymbol{w} \in \mathbb{R}^M$:

$$\|\boldsymbol{w}\|_1 \leq \sqrt{M}\|\boldsymbol{w}\|_2.$$

This bound is a result of the Cauchy-Schwarz inequality: $\|\boldsymbol{w}\|_1 = |\langle \boldsymbol{w}, \mathbf{1}\rangle| \leq \|\mathbf{1}\|_2 \|\boldsymbol{w}\|_2 = \sqrt{M}\|\boldsymbol{w}\|_2$. Using these facts, we have:

$$\|\boldsymbol{e}\|_{1,h} = h\|\boldsymbol{e}\|_1 \leq h\sqrt{M}\|\boldsymbol{e}\|_2 = \sqrt{h}\sqrt{M}\|\boldsymbol{e}\|_{2,h} = \sqrt{hM}\mathcal{O}(h^2) = \mathcal{O}(h^2),$$

which is what was desired. Note that this result could also be proven by rederiving convergence through the Lax Equivalence theorem, e.g., redefining and establishing consistency in terms of the norm $\|\cdot\|_{1,h}$, and also defining and establishing stability in terms of the matrix norm induced by the vector norm $\|\cdot\|_{1,h}$.

3. (ODE convergence)
For the ODE system $\boldsymbol{u}'(t) = \boldsymbol{f}(t, \boldsymbol{u})$ with initial condition $\boldsymbol{u}(0)$ given and $\boldsymbol{f}$ globally Lipschitz continuous in $\boldsymbol{u} \in \mathbb{R}^M$ uniformly in t, show that backward Euler with the initial state $\boldsymbol{u}_0 = \boldsymbol{u}(0)$ is *convergent* to first order as defined on slide `D06-S07`.

Solution: $\boldsymbol{f}(\cdot, \boldsymbol{u})$ globally Lipschitz in $\boldsymbol{u}$ uniformly in t means that there exists some $L \geq 0$ such that for all $t \geq 0$ and any $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^M$,

$$\|\boldsymbol{f}(t, \boldsymbol{v}) - \boldsymbol{f}(t, \boldsymbol{w})\| \leq L\|\boldsymbol{v} - \boldsymbol{w}\|.$$

Given a terminal time $T > 0$, we seek to show that the implicit scheme,

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + k\boldsymbol{f}(t_n, \boldsymbol{u}_{n+1})$$

converges to first order in k, i.e., that for k sufficiently small,

$$\max_{n\in[N]} \|\boldsymbol{e}_n\| \leq Ck = \mathcal{O}(k),$$

where $\boldsymbol{e}_n = \boldsymbol{u}(t_n) - \boldsymbol{u}_n$, and $N = T/k$. Note that since $\boldsymbol{u}_0 = \boldsymbol{u}(0)$, then $\boldsymbol{e}_0 = \mathbf{0}$. We'll show convergence by first showing both consistency and 0-stability. Consistency requires us to compute the local truncation error, which is given by,

$$LTE_n \coloneqq \frac{\boldsymbol{u}(t_{n+1}) - \boldsymbol{u}(t_n)}{k} - \boldsymbol{f}(t_n, \boldsymbol{u}(t_{n+1})) = \mathcal{O}(k),$$

where we have Taylor expanded $\boldsymbol{u}(t_{n+1})$ around $t = t_n$. The second ingredient is 0-stability, defined by the condition,

$$\max_{n\in[N]} \|\boldsymbol{e}_n\| \leq C\left(\|\boldsymbol{e}_0\| + \max_{n\in[N]} \|R_n \boldsymbol{u}(t_n)\|\right) = C\max_{n\in[N]} \|R_n \boldsymbol{u}(t_n)\|.$$

where $R_n(\cdot)$ is the scheme residual, which operating on the exact solution is identical to the local truncation error:

$$R_n \boldsymbol{u}(t_n) \coloneqq D^+\boldsymbol{u}(t_n) - \boldsymbol{f}\left(t_{n+1}, \boldsymbol{u}(t_{n+1})\right) = LTE_n. \tag{1}$$

In particular, by definition of the local truncation error and consistency, we have that,

$$\max_{n\in[N]} \|R_n \boldsymbol{u}(t_n)\| = \max_{n\in[N]} \|LTE_n\| =: \omega = \mathcal{O}(k).$$

(Strictly speaking, we require u'' to be bounded in order to make the last equality valid.) To show 0-stability, we compute:

$$\begin{aligned}\boldsymbol{e}_n &= \boldsymbol{u}(t_n) - \boldsymbol{u}_n \\ &= \boldsymbol{u}(t_n) - \boldsymbol{u}(t_{n-1}) + \boldsymbol{u}(t_{n-1}) - \boldsymbol{u}_{n-1} + \boldsymbol{u}_{n-1} - \boldsymbol{u}_n \\ &= (\boldsymbol{u}(t_n) - \boldsymbol{u}(t_{n-1})) + \boldsymbol{e}_{n-1} - k\boldsymbol{f}(t_n, \boldsymbol{u}_n) \end{aligned} \tag{2}$$

where the last equality has used the scheme definition. Via (1), the first term on the right-hand side above can be written as,

$$\boldsymbol{u}(t_n) - \boldsymbol{u}(t_{n-1}) = kR_{n-1}\boldsymbol{u}(t_{n-1}) + k\boldsymbol{f}(t_n, \boldsymbol{u}(t_n)).$$

Using this inequality in (2) results in:

$$\begin{aligned}\boldsymbol{e}_n &= kR_{n-1}\boldsymbol{u}(t_{n-1}) + k\boldsymbol{f}(t_n, \boldsymbol{u}(t_n)) + \boldsymbol{e}_{n-1} - k\boldsymbol{f}(t_n, \boldsymbol{u}_n) \\ &= \boldsymbol{e}_{n-1} + kR_{n-1}\boldsymbol{u}(t_{n-1}) + k\left(\boldsymbol{f}(t_n, \boldsymbol{u}(t_n)) - \boldsymbol{f}(t_n, \boldsymbol{u}_n)\right).\end{aligned}$$

Employing the triangle inequality and Lipschitz continuity of $\boldsymbol{f}$ yields,

$$\|\boldsymbol{e}_n\| \leq \|\boldsymbol{e}_{n-1}\| + k\omega + kL\|\boldsymbol{u}(t_n) - \boldsymbol{u}_n\| = \|\boldsymbol{e}_{n-1}\| + k\omega + kL\|\boldsymbol{e}_n\|.$$

So for $k < L$, we have,

$$\|\boldsymbol{e}_n\| \leq \frac{1}{1-kL}\|\boldsymbol{e}_{n-1}\| + \frac{k\omega}{1-kL}$$

Iterating this inequality yields,

$$\begin{aligned}\|\boldsymbol{e}_n\| &\leq (1-kL)^{-n}\|\boldsymbol{e}_0\| + k\omega\sum_{j=0}^{n-1}(1-kL)^{-j} \\ &\overset{\boldsymbol{e}_0=\boldsymbol{0}}{=} k\omega\sum_{j=0}^{n-1}(1-kL)^{-j}\end{aligned} \tag{3}$$

Using the truncated geometric series, we have,

$$\begin{aligned}\sum_{j=0}^{n-1}(1-kL)^{-j} \leq \sum_{j=0}^{N}(1-kL)^{-j} &= \frac{\frac{1}{(1-kL)^{N+1}} - 1}{\frac{1}{1-kL} - 1} \\ &= 1 + \frac{1}{kL}\left(\frac{1}{(1-kL)^N} - 1\right) \\ &\leq 1 + \frac{1}{kL}\frac{1}{(1-kL)^N} \\ &\leq 1 + \frac{1}{kL}\frac{1}{(1-TL/N)^N} \\ &\overset{k\ll 1}{\leq} 1 + 2\frac{e^{LT}}{kL} \overset{kL<1}{\leq} 3\frac{e^{LT}}{kL}.\end{aligned}$$

Using this in (3) yields,

$$\|\boldsymbol{e}_n\| \leq k\omega 3\frac{e^{LT}}{kL} = \frac{3e^{LT}}{L}\omega = C\omega,$$

which verifies 0-stability. To prove convergence, we combine 0-stability and consistency:

$$\max_{n\in[N]} \|\boldsymbol{e}_n\| \overset{0-\text{stability}}{\leq} C\omega \overset{\text{consistency}}{=} C\mathcal{O}(k).$$

4. (Regions of stability)

a. Compute the stability/amplification factor for Crank-Nicolson and show that it is the left half-plane.

b. Compute the stability/amplification factor for the following scheme,

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n + \frac{k}{4}\boldsymbol{f}(t_n, \boldsymbol{u}_n) + \frac{3k}{4}\boldsymbol{f}\left(t_n + \frac{2}{3}k, \boldsymbol{u}_n + \frac{2}{3}k\boldsymbol{f}(t_n, \boldsymbol{u}_n)\right),$$

and plot the corresponding stability region, using software if desired. (This scheme is called Ralston's method.)

<u>**Solution**</u>:

a. The Crank-Nicolson scheme for a scalar problem $u' = \lambda u$ is:

$$\begin{aligned} u_{n+1} &= u_n + \frac{k}{2}f(t_n, u_n) + \frac{k}{2}f(t_{n+1}, u_{n+1}) \\ &= u_n + \frac{k\lambda}{2}u_n + \frac{k\lambda}{2}u_{n+1} \end{aligned}$$

Then solving for u_{n+1} yields,

$$\begin{aligned} u_{n+1} &= u_n \left(\frac{1 + \frac{k\lambda}{2}}{1 - \frac{k\lambda}{2}}\right) \\ &\overset{z:=k\lambda}{=} u_n \left(\frac{2+z}{2-z}\right), \end{aligned}$$

and hence the stability factor is,

$$\phi(z) = \frac{2+z}{2-z}.$$

Ensuring stability requires $|\phi(z)| \leq 1$, which in turn requires,

$$|z+2| \leq |z-2|.$$

This implies that z is inside the region of stability if and only if the complex plane distance from z to -2 is less than or equal to the complex plane distance from z to +2. I.e., the only restriction is that the real part of z is non-positive. Therefore, the region of stability is precisely the left half-plane.

b. Carrying out a similar computation here for the scalar problem $u' = \lambda u$, we have,

$$\begin{aligned} u_{n+1} &= \frac{k\lambda}{4}u_n + \frac{3k\lambda}{4}\left(u_n + \frac{2k\lambda}{3}u_n\right) \\ &\overset{z:=k\lambda}{=} u_n\left(1 + \frac{z}{4} + \frac{3z}{4}\left(1 + \frac{2z}{3}\right)\right) \\ &= u_n\left(1 + z + \frac{1}{2}z^2\right). \end{aligned}$$

Therefore, the stability factor is,

$$\phi(z) = 1 + z + \frac{1}{2}z^2 = \frac{1}{2}(z+1)^2 + \frac{1}{2}.$$

Stability requires that $|\phi(z)| \leq 1$. Hence, a point z is on the boundary of this region if $|\phi(z)| = 1$, i.e., if $\phi(z) = e^{i\theta}$ for some $\theta \in [0, 2\pi)$. Solving this equation for z yields.

$$z = -1 \pm \sqrt{2e^{i\theta} - 1},$$

which parameterizes points on the boundary of the region of stability in the complex plane. The region itself for this problem is shown in Figure 1.

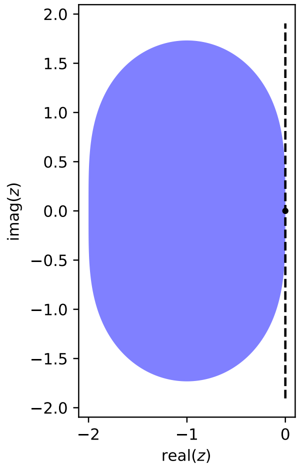

Figure 1: Problem 4: Region of stability for Ralston's method (shaded blue region). Also shown is the origin (black dot) and the imaginary axis (dotted black line).

5. (ODE solvers)
Consider the IVP,

$$\boldsymbol{u}'(t) = \boldsymbol{A}\boldsymbol{u}, \qquad (\boldsymbol{u}(0))_j = j(-1)^j, \qquad \boldsymbol{A} = C^2 \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{10\times 10} \tag{4}$$

where $j \in [10]$, $\boldsymbol{A}$ is symmetric and tridiagonal, and $C \in \mathbb{R}$. We'll compute numerical solutions in this problem up to a terminal time $T = 1$.

a. Take $C = 3$. Implement both Forward Euler and Crank-Nicolson solvers for this IVP, and demonstrate that the schemes exhibit the expected orders of convergence (with respect to the time discretization parameter k) for each method. (You may use any norm on $\boldsymbol{u}$.) Briefly discuss the advantages and disadvantages of each scheme.

b. Take $C = 10$. Again discuss the advantages and disadvantages of each scheme, using numerical results to support your conclusions.

<u>**Solution**</u>:

a. For various values of $k = \Delta t$, numerical results are shown in Figure 2 (left, middle); we use the max norm $\|\cdot\|$ for vectors to compute error. The exact solution at each value of t was computed using the matrix exponential of $\boldsymbol{A}$. We see that both schemes achieve their expected orders of convergence: Forward Euler converges to first-order in k, and Crank-Nicolson converges to second order. In this case, even without considering the order of convergence, Crank-Nicolson achieves a much smaller value of accuracy than Forward Euler. The accuracy results suggest that Crank-Nicolson is always preferred to Forward Euler, but Crank-Nicolson is an implicit scheme; in this case this requires a linear solve at every time step. For more complicated problems it may require a nonlinear solve. Since we expect that a linear solve at every time step iteration is a considerable extra cost, Figure 2 (right) also plots the accuracy as a function of the total wall clock time required to compute the numerical solution. We see that indeed Crank-Nicolson requires more computational time. For this relatively small (size-10) system, the linear solve is still not very expensive unless a very low accuracy is required. For larger size problems, this tradeoff calculus will change.

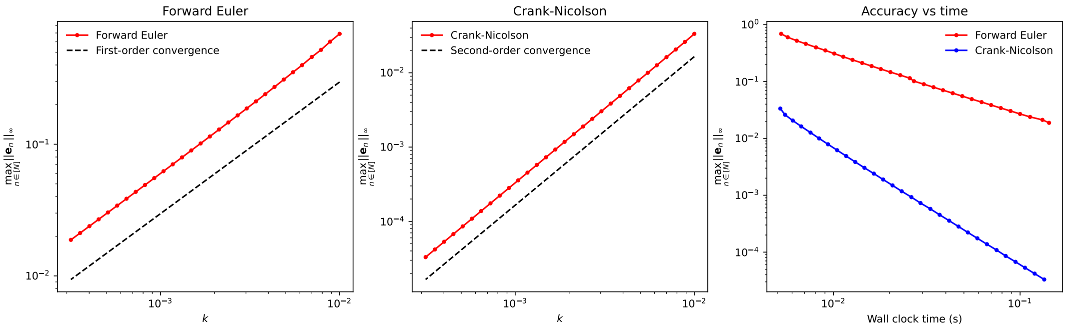

Figure 2: Numerical results for problem 5(a) with $C = 3$. Left and middle: Forward Euler and Crank-Nicolson accuracy vs timestep k for the ODE system (4). Right: Forward Euler and Crank-Nicolson accuracy versus wall clock time required to compute the solution.

b. The results for this problem are shown in Figure 3. For this larger value of C, the problem is much stiffer, and Forward Euler is actually unstable until time step $k \sim 5 \times 10^{-3}$. However, Crank-Nicolson remains very stable over a wide range of timesteps. The (asymptotic) accuracy observations are the same as in part (a).

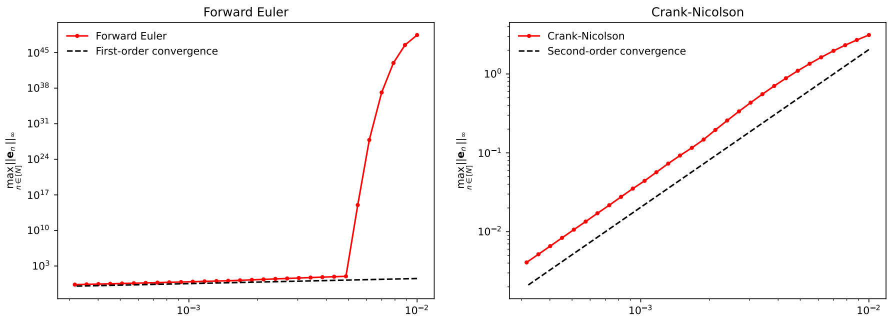

Figure 3: Numerical results for problem 5(b) with $C = 10$. Forward Euler and Crank-Nicolson accuracy vs timestep k for the ODE system (4).