Department of Mathematics, University of Utah

Analysis of Numerical Methods, II
MATH 6620 – Section 001 – Spring 2024
Homework 6 Solutions
Fourier Series for time-dependent problems

Due Monday, April 22, 2024

Note: In this assignment, you must complete problems 1-3. However, you need only choose one of problems 4, 5, or 6 to complete and submit. (I.e., you must complete a total of 4 problems in this assignment.)

Corrections: corrections to an earlier version of this assignment are shown in blue boldface. You may choose to use the earlier, incorrect assignment, in which case you should justify what can be accomplished for those problems.

Submit your solutions online through Gradescope.

1. (Operator splitting)
Consider the following semi-discrete form of a PDE:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = A(\boldsymbol{u}) + B(\boldsymbol{u}), \qquad \boldsymbol{u}(x,0) = \boldsymbol{u_0}(x).$$

where both A and B may be arbitrary operators.

a. Suppose A and B are linear operators (and hence $A(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$ and $B(\boldsymbol{u}) = \boldsymbol{B}\boldsymbol{u}$). Show that the numerical scheme,

$$\begin{aligned} \boldsymbol{u}_* &= e^{k\boldsymbol{A}}\boldsymbol{u}_n \\ \boldsymbol{u}_{n+1} &= e^{k\boldsymbol{B}}\boldsymbol{u}_* \end{aligned}$$

is in general first-order accurate. Also give a sufficient condition on $\boldsymbol{A}$ and $\boldsymbol{B}$ to ensure that the scheme is second-order accurate.

b. Again assuming A and B are linear operators, show that the scheme

$$\begin{aligned} \boldsymbol{u}^* &= e^{k\boldsymbol{A}/2}\boldsymbol{u}_n \\ \boldsymbol{u}^{**} &= e^{k\boldsymbol{B}}\boldsymbol{u}_* \\ \boldsymbol{u}_{n+1} &= e^{k\boldsymbol{A}/2}\boldsymbol{u}_{**} \end{aligned}$$

is in general second-order accurate. (This scheme is called *Strang splitting*.)

c. Repeat the above procedures in the general nonlinear case, confirming that both schemes are first- and second-order accurate. I.e., replace $\boldsymbol{A}\boldsymbol{u}$ with $A(\boldsymbol{u})$, and similarly for the operator B, **and also replace exponential solution operators with exact solution operators for nonlinear problems**. (You need not give sufficient conditions on A and B to ensure that the first-order scheme is actually second-order accurate.)

Solution:

a. First we note that the exact solution is given by,

$$\boldsymbol{u}(t_{n+1}) = e^{k(\boldsymbol{A}+\boldsymbol{B})}\boldsymbol{u}(t_n) = \left(\sum_{j=0}^{\infty} k^j \frac{(\boldsymbol{A}+\boldsymbol{B})^j}{j!}\right)\boldsymbol{u}(t_n) = \boldsymbol{u}(t_n) + k(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{u}(t_n) + \frac{k^2}{2}(\boldsymbol{A}+\boldsymbol{B})^2\boldsymbol{u}(t_n) + \mathcal{O}($$

To compare this against the splitting method, consider the Taylor approximations of the individual matrix exponentials,

$$e^{k\boldsymbol{A}} = \sum_{j=0}^{\infty} k^j \frac{\boldsymbol{A}^j}{j!}, \qquad e^{k\boldsymbol{B}} = \sum_{j=0}^{\infty} k^j \frac{\boldsymbol{B}^j}{j!},$$

Therefore,

$$\begin{aligned} e^{k\boldsymbol{B}}e^{k\boldsymbol{A}}\boldsymbol{u}(t_n) &= \sum_{j=0}^{\infty} k^j \left(\sum_{\ell=0}^{j} \frac{\boldsymbol{A}^\ell \boldsymbol{B}^{j-\ell}}{\ell!(j-\ell)!}\right)\boldsymbol{u}(t_n) \\ &= \boldsymbol{u}(t_n) + k(\boldsymbol{B}+\boldsymbol{A})\boldsymbol{u}(t_n) + k^2\left(\frac{\boldsymbol{B}^2}{2} + \boldsymbol{A}\boldsymbol{B} + \frac{\boldsymbol{A}^2}{2}\right)\boldsymbol{u}(t_n) + \mathcal{O}(k^3). \end{aligned}$$

Hence, we have,

$$\boldsymbol{u}(t_{n+1}) - e^{k\boldsymbol{B}}e^{k\boldsymbol{A}}\boldsymbol{u}(t_n) = \frac{k^2}{2}\left(\boldsymbol{B}\boldsymbol{A} - \boldsymbol{A}\boldsymbol{B}\right)\boldsymbol{u}(t_n) + \mathcal{O}(k^3). \tag{1}$$

Hence, the local truncation error is,

$$\frac{\boldsymbol{u}(t_{n+1}) - e^{k\boldsymbol{B}}e^{k\boldsymbol{A}}\boldsymbol{u}(t_n)}{k} = \mathcal{O}(k),$$

and so this is a first-order accurate scheme. However the expression in (1) shows that the scheme is second-order accurate if,

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}.$$

Hence, this condition is sufficient to yield second-order consistency. In actuality, this commutation condition is sufficient to ensure that the splitting scheme is *exact*: If $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$, then note that we can use the binomial theorem to expand the binomial matrix expression,

$$\frac{1}{j!}(\boldsymbol{A}+\boldsymbol{B})^j = \sum_{\ell=0}^{j} \binom{j}{\ell} \frac{\boldsymbol{A}^j\boldsymbol{B}^{j-\ell}}{j!} = \sum_{\ell=0}^{j} \frac{\boldsymbol{A}^j\boldsymbol{B}^{\ell-j}}{\ell!(j-\ell)!}.$$

And hence,

$$\begin{aligned} \boldsymbol{u}(t_{n+1}) - e^{k\boldsymbol{B}}e^{k\boldsymbol{A}}\boldsymbol{u}(t_n) &= \left(\sum_{j=0}^{\infty} k^j \frac{(\boldsymbol{A}+\boldsymbol{B})^j}{j!}\right)\boldsymbol{u}(t_n) - \sum_{j=0}^{\infty} k^j \left(\sum_{\ell=0}^{j} \frac{\boldsymbol{A}^\ell\boldsymbol{B}^{j-\ell}}{\ell!(j-\ell)!}\right)\boldsymbol{u}(t_n) \\ &= \sum_{j=0}^{\infty} k^j \left[\frac{(\boldsymbol{A}+\boldsymbol{B})^j}{j!} - \sum_{\ell=0}^{j} \frac{\boldsymbol{A}^\ell\boldsymbol{B}^{j-\ell}}{\ell!(j-\ell)!}\right]\boldsymbol{u}(t_n) \\ &= \sum_{j=0}^{\infty} k^j \boldsymbol{0}\boldsymbol{u}_n = \boldsymbol{0}. \end{aligned}$$

b. To analyze Strang splitting, we perform a similar computation as in the previous part, explicitly computing products of matrix exponentials:

$$\begin{aligned} e^{k\boldsymbol{A}/2}e^{k\boldsymbol{B}}e^{k\boldsymbol{A}/2}\boldsymbol{u} &= \left(\boldsymbol{I} + \frac{k\boldsymbol{A}}{2} + \frac{k^2\boldsymbol{A}^2}{8} + \mathcal{O}(k^3)\right) \\ &\quad \left(\boldsymbol{I} + k\boldsymbol{B} + \frac{k^2\boldsymbol{B}^2}{2} + \mathcal{O}(k^3)\right) \\ &\quad \left(\boldsymbol{I} + \frac{k\boldsymbol{A}}{2} + \frac{k^2\boldsymbol{A}^2}{8} + \mathcal{O}(k^3)\right)\boldsymbol{u} \\ &= \boldsymbol{u} + k\left(\boldsymbol{A}+\boldsymbol{B}\right)\boldsymbol{u} + k^2\Bigg(\left[\frac{1}{8}+\frac{1}{8}+\frac{1}{4}\right]\boldsymbol{A}^2 + \frac{1}{2}\boldsymbol{A}\boldsymbol{B} + \frac{1}{2}\boldsymbol{B}\boldsymbol{A} + \\ &\qquad\qquad \left[\frac{1}{8}+\frac{1}{8}+\frac{1}{4}\right]\boldsymbol{B}^2\Bigg)\boldsymbol{u} + \mathcal{O}(k^3) \\ &= \boldsymbol{u} + k\left(\boldsymbol{A}+\boldsymbol{B}\right)\boldsymbol{u} + \frac{k^2}{2}\left(\boldsymbol{A}+\boldsymbol{B}\right)^2\boldsymbol{u} + \mathcal{O}(k^3). \end{aligned}$$

Therefore, the local truncation error for this scheme is,

$$\frac{\boldsymbol{u}(t_{n+1}) - e^{k\boldsymbol{A}/2}e^{k\boldsymbol{B}}e^{k\boldsymbol{A}/2}\boldsymbol{u}(t_n)}{k} = \mathcal{O}(k^2).$$

c. For a nonlinear scheme, we replace the matrix exponential by the exact solution associated to the individual evolution equations. More precisely, if $\boldsymbol{u}(t)$ evolves according to,

$$\boldsymbol{u}'(t) = \boldsymbol{C}(\boldsymbol{u}),$$

where $\boldsymbol{C}$ is an arbitrary vector-valued function, then by Taylor's theorem,

$$\begin{aligned} \boldsymbol{u}(t_{n+1}) &= \boldsymbol{u}(t_n) + k\boldsymbol{u}'(t_n) + \frac{k^2}{2}\boldsymbol{u}''(t_n) + \mathcal{O}(k^3) \\ &= \boldsymbol{u}(t_n) + k\boldsymbol{C}(\boldsymbol{u}(t_n)) + \frac{k^2}{2}\boldsymbol{C}'(\boldsymbol{u}(t_n))\boldsymbol{C}(\boldsymbol{u}(t_n)) + \mathcal{O}(k^3), \end{aligned} \tag{2}$$

where we have used the fact that $\boldsymbol{u}' = \boldsymbol{C}$, and that $\boldsymbol{u}''$ can be computed through the chain rule, which involves the Jacobian $\boldsymbol{C}'(\boldsymbol{u}) \coloneqq \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{u}}(\boldsymbol{u})$ of $\boldsymbol{C}$. If we apply this to the exact solution $\boldsymbol{u}$ with $\boldsymbol{C}(\boldsymbol{u}) = \boldsymbol{A}(\boldsymbol{u}) + \boldsymbol{B}(\boldsymbol{u})$, this yields,

$$\boldsymbol{u}(t_{n+1}) = \boldsymbol{u}(t_n) + k\left(\boldsymbol{A}_n + \boldsymbol{B}_n\right) + \frac{k^2}{2}\left(\boldsymbol{A}_n' + \boldsymbol{B}_n'\right)\left(\boldsymbol{A}_n + \boldsymbol{B}_n\right) + \mathcal{O}(k^3), \tag{3}$$

where we have introduced the abbreviations,

$$\boldsymbol{A}_n = \boldsymbol{A}(\boldsymbol{u}(t_n)), \qquad \boldsymbol{A}_n' = \frac{\partial \boldsymbol{A}}{\partial \boldsymbol{u}}(\boldsymbol{u}(t_n)), \qquad \boldsymbol{B}_n = \boldsymbol{B}(\boldsymbol{u}(t_n)), \qquad \boldsymbol{B}_n' = \frac{\partial \boldsymbol{B}}{\partial \boldsymbol{u}}(\boldsymbol{u}(t_n)).$$

Hence, we seek to show that Strang splitting recovers the expression above up to $\mathcal{O}(k^2)$. To begin, we characterize $\boldsymbol{u}^*$ as follows: We consider $\boldsymbol{w}(t)$ to satisfy,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{w}(t) = \boldsymbol{A}(\boldsymbol{w}), \qquad\qquad \boldsymbol{w}(t_n) = \boldsymbol{u}(t_n), \tag{4}$$

with $\boldsymbol{u}^* \coloneqq \boldsymbol{w}(t_n + k/2)$. Then by the argument in (2),

$$\boldsymbol{u}^* = \boldsymbol{u}(t_n) + \frac{k}{2}\boldsymbol{A}_n + \frac{k^2}{8}\boldsymbol{A}_n'\boldsymbol{A}_n + \mathcal{O}(k^3).$$

Similarly, we have,

$$\begin{aligned}\boldsymbol{u}^{**} &= \boldsymbol{u}^* + k\boldsymbol{B}_* + \frac{k^2}{2}\boldsymbol{B}_*'\boldsymbol{B}_* + \mathcal{O}(k^3)\\ \boldsymbol{u}_{n+1} &= \boldsymbol{u}^{**} + \frac{k}{2}\boldsymbol{A}_{**} + \frac{k^2}{8}\boldsymbol{A}_{**}'\boldsymbol{A}_{**} + \mathcal{O}(k^3),\end{aligned}$$

where we have used the abbreviations,

$$\begin{aligned}\boldsymbol{B}_* &= \boldsymbol{B}(\boldsymbol{u}^*), & \boldsymbol{B}_*' &= \frac{\partial \boldsymbol{B}}{\partial \boldsymbol{u}}(\boldsymbol{u}^*),\\ \boldsymbol{A}_{**} &= \boldsymbol{A}(\boldsymbol{u}^{**}), & \boldsymbol{A}_{**}' &= \frac{\partial \boldsymbol{A}}{\partial \boldsymbol{u}}(\boldsymbol{u}^{**})\end{aligned}$$

To determine leading order behavior of these expressions, we note that,

$$\boldsymbol{B}_* = \boldsymbol{B}(\boldsymbol{u}^*) = \boldsymbol{B}\left(\boldsymbol{u}_n + \frac{k}{2}\boldsymbol{A}_n + \mathcal{O}(k^2)\right) = \boldsymbol{B}_n + \frac{k}{2}\boldsymbol{B}_n'\boldsymbol{A}_n + \mathcal{O}(k^2).$$

There are two ways to see the validity of the above expression: either one can Taylor expand $\boldsymbol{u}^* \mapsto \boldsymbol{B}(\boldsymbol{u}^*)$ around $\boldsymbol{u}^* = \boldsymbol{u}(t_n)$, or one can Taylor expand $s \mapsto \boldsymbol{B}(\boldsymbol{w}(s))$ as a function of s around $s = t_n$, where $\boldsymbol{w}$ evolves according to (4). In a similar vein, we have,

$$\boldsymbol{B}_*' = \boldsymbol{B}_n' + \mathcal{O}(k).$$

Using the above Taylor expansions for $\boldsymbol{u}^*$, $\boldsymbol{B}_*$, and $\boldsymbol{B}_*'$ in the Taylor expansion for $\boldsymbol{u}^{**}$ yields,

$$\boldsymbol{u}^{**} = \boldsymbol{u}(t_n) + k\left(\frac{1}{2}\boldsymbol{A}_n + \boldsymbol{B}_n\right) + k^2\left(\frac{1}{8}\boldsymbol{A}_n'\boldsymbol{A}_n + \frac{1}{2}\boldsymbol{B}_n'\boldsymbol{A}_n + \frac{1}{2}\boldsymbol{B}_n'\boldsymbol{B}_n\right) + \mathcal{O}(k^3)$$

This then implies that,

$$\begin{aligned}\boldsymbol{A}_{**} &= \boldsymbol{A}(\boldsymbol{u}_n + k/2\boldsymbol{A}_n + k\boldsymbol{B}_n + \mathcal{O}(k^2)) = \boldsymbol{A}_n + \frac{k}{2}\boldsymbol{A}_n'\boldsymbol{A}_n + k\boldsymbol{A}_n'\boldsymbol{B}_n + \mathcal{O}(k^2)\\ \boldsymbol{A}_{**}' &= \boldsymbol{A}'(\boldsymbol{u}_n + \mathcal{O}(k)) = \boldsymbol{A}_n' + \mathcal{O}(k)\end{aligned}$$

Finally, using the Taylor expansions of $\boldsymbol{u}^{**}$, $\boldsymbol{A}_{**}$, and $\boldsymbol{A}_{**}'$ in the Taylor expansion of $\boldsymbol{u}_{n+1}$ yields,

$$\begin{aligned}\boldsymbol{u}_{n+1} &= \boldsymbol{u}(t_n) + k\left(\frac{1}{2}\boldsymbol{A}_n + \boldsymbol{B}_n + \frac{1}{2}\boldsymbol{A}_n\right) + k^2\Bigg[\left(\frac{1}{8} + \frac{1}{4} + \frac{1}{8}\right)\boldsymbol{A}_n'\boldsymbol{A}_n + \frac{1}{2}\boldsymbol{B}_n'\boldsymbol{A}_n\\ &\qquad + \frac{1}{2}\boldsymbol{B}_n'\boldsymbol{B}_n + \frac{1}{2}\boldsymbol{A}_n'\boldsymbol{B}_n\Bigg] + \mathcal{O}(k^3)\\ &= \boldsymbol{u}(t_n) + k\left(\boldsymbol{A}_n + \boldsymbol{B}_n\right) + \frac{k^2}{2}\left((\boldsymbol{A}_n' + \boldsymbol{B}_n')(\boldsymbol{A}_n + \boldsymbol{B}_n)\right) + \mathcal{O}(k^3).\end{aligned}$$

Therefore, combining this with (3) the local truncation error is,

$$\text{LTE} = \frac{\boldsymbol{u}(t_{n+1}) - \boldsymbol{u}_{n+1}}{k} = \mathcal{O}(k^2),$$

as desired.

2. (Fourier spectral methods)
Consider the partial differential equation,

$$u_t + (\sin x)\, u_x = \frac{1}{2} u_{xx}, \qquad u(x,0) = \sin x \tag{5}$$

for $u = u(x,t)$, $x \in [0, 2\pi)$, $t > 0$, and periodic boundary conditions.

- **a.** Write out the semi-discrete form for the Fourier-Galerkin scheme for this problem.
- **b.** If using Forward Euler for fully discretizing the semi-discrete form, what is the timestep restriction (say as a function of N) to ensure that the scheme is stable in the sense of absolute stability? (You may investigate this analytically or empirically via the region of stability.) What aspect of the equation (5) is the dominant contributor to the timestep restriction? How does this make the total computational time of the solver scale with N?
- **c.** Implement both implicit (e.g., Crank-Nicolson) and explicit time-stepping methods (say RK2) for solving the above problem up to time $T = 1$. Visualize the simulated results and discuss the advantages and disadvantages of each approach.

Solution:

a. We seek to perform the approximation,

$$u(x,t) \simeq u_N(x,t) \coloneqq \sum_{|k| \leq N} \widehat{u}_k(t) \phi_k(x), \qquad \phi_k(x) = \frac{1}{\sqrt{2\pi}} e^{ikx}.$$

and the Fourier-Galerkin scheme is the system of $(2N+1)$ coupled ODEs defining the evolution of the $\widehat{u}_k(t)$ functions:

$$\left\langle \frac{\partial}{\partial t} u_N + \sin x \frac{\partial}{\partial x} u_N, \phi_\ell \right\rangle = \left\langle \frac{1}{2} \frac{\partial^2}{\partial x^2} u_N, \phi_\ell \right\rangle, \qquad |\ell| \leq N,$$

where $\langle \cdot, \cdot \rangle$ is the standard (conjugate-bilinear) inner product on $L^2([0, 2\pi]; \mathbb{C})$. Using the fact that the ϕ_k are orthonormal functions, then the system above simplifies to,

$$\frac{\mathrm{d}}{\mathrm{d}t} \widehat{u}_\ell = -\frac{\ell^2}{2} \widehat{u}_\ell - \sum_{|k| \leq N} \widehat{u}_k(t) \left\langle \sin x\ \phi_k'(x), \phi_\ell(x) \right\rangle. \qquad |\ell| \leq N$$

As shown on slide D17-S05(B), we have,

$$\sin x \phi_k'(x) = \frac{k}{2} \left(\phi_{k+1} - \phi_{k-1} \right),$$

and hence the evolution for $\widehat{u}_\ell$ simplifies to,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \widehat{u}_\ell &= -\frac{\ell^2}{2} \widehat{u}_\ell + \frac{\ell+1}{2} \widehat{u}_{\ell+1}. && \ell = -N \\
\frac{\mathrm{d}}{\mathrm{d}t} \widehat{u}_\ell &= -\frac{\ell^2}{2} \widehat{u}_\ell - \frac{\ell-1}{2} \widehat{u}_{\ell-1} + \frac{\ell+1}{2} \widehat{u}_{\ell+1}. && |\ell| \leq N-1 \\
\frac{\mathrm{d}}{\mathrm{d}t} \widehat{u}_\ell &= -\frac{\ell^2}{2} \widehat{u}_\ell - \frac{\ell-1}{2} \widehat{u}_{\ell-1}. && \ell = N
\end{aligned}$$

with the initial condition $\widehat{\boldsymbol{u}}(0) = \widehat{\boldsymbol{u}}_0$ defined by,

$$(\widehat{u}_0)_\ell = \langle \sin x, \phi_\ell \rangle \implies (\widehat{u}_0)_1 = -i\sqrt{\frac{\pi}{2}},\ (\widehat{u}_0)_{-1} = i\sqrt{\frac{\pi}{2}},\ (\widehat{u}_0)_k = 0 \text{ otherwise}$$

The matrix-vector form of the semi-discrete scheme is given by,

$$\widehat{\boldsymbol{u}}'(t) = \boldsymbol{A}\widehat{\boldsymbol{u}}(t), \qquad \widehat{\boldsymbol{u}}(0) = \widehat{\boldsymbol{u}}_0 \tag{6}$$

where,

$$\widehat{\boldsymbol{u}}(t) = (\widehat{u}_{-N}(t),\ \widehat{u}_{-N+1}(t),\ \ldots\ \widehat{u}_N(t))^T$$

and

$$\boldsymbol{A} = \frac{1}{2}\begin{pmatrix} -N^2 & -N+1 & & & \\ N & -(N-1)^2 & -N+2 & & \\ & \ddots & \ddots & \ddots & \\ & & & & \\ & & & -(N-1) & -N^2 \end{pmatrix}$$

with missing entries denoting zeros.

b. For Forward Euler applied to a system of ODEs of the form (6), absolute stability requires that

$$|z+1| < 1, \qquad z = k\lambda(\boldsymbol{A}),$$

and note that if λ has real and imaginary parts $\lambda = \mu + \mathbf{i}\omega$, then the above condition is equivalent to,

$$k \leq \frac{-2\mu}{|\lambda|^2}, \tag{7}$$

and hence we must have (as expected) that the real part of the spectrum of $\boldsymbol{A}$ lies in the left half-plane in $\mathbb{C}$. Instead of directly computing the spectrum of $\boldsymbol{A}$, we opt to bound the spectrum using Gerschgorin's theorem. Recall that this theorem states any eigenvalue of $\boldsymbol{A}$ lies within at least one Gerschgorin disc in $\mathbb{C}$. The Gerschgorin discs are circular discs centered at a diagonal element of $\boldsymbol{A}$, with radius equal to the sum-of-magnitudes of every off-diagonal element in that row (or column). In this case, $\boldsymbol{A}$ is tridiagonal so that the discs and radii are explicitly computable. In particular, the Gerschgorin discs have centers c_j and radii r_j given by,

$$c_\ell = -\frac{1}{2}\ell^2, \qquad r_\ell = \frac{\max\{2\ell, 2\}}{2} \sim \ell$$

where ℓ satisfies $|\ell| \leq N$. Then for ℓ satisfying $|\ell| \geq 3$, any eigenvalue $\lambda = \mu + i\omega$ lying in the the ℓth Gerschgorin disc satisfies,

$$\frac{-\mu}{|\lambda|} \geq \frac{\ell^2 - 2\ell - 1}{\ell^2 + 2\ell + 1} \geq C,$$

for a universal constant $C > 0$. (For $|\ell| \leq 2$, the quantity above cannot be bounded below by a positive number.) Hence, we expect that

$$k \leq C\frac{1}{|\lambda|} \leq \frac{-\mu}{|\lambda|}\ \frac{2}{|\lambda|},$$

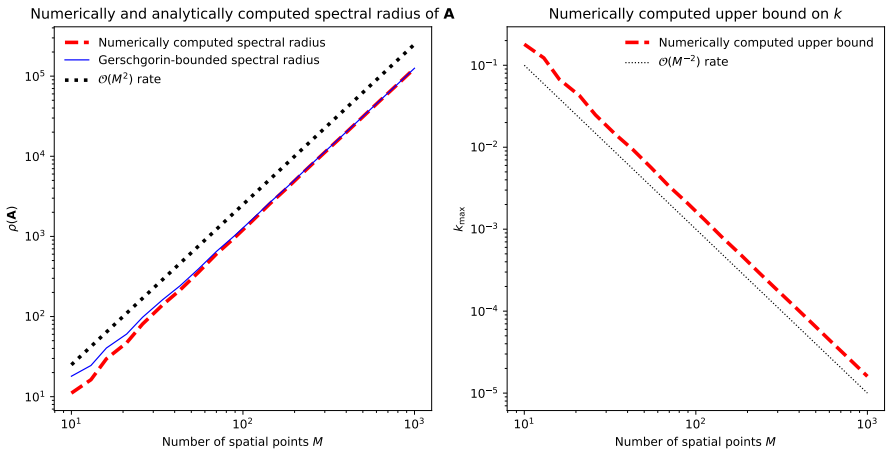

Figure 1: Problem 2b. Left: Analytically computed spectral radius of the matrix $\boldsymbol{A}$ versus the Gerschgorin upper bound (8) and the asymptotically-expected $\mathcal{O}(M^2)$ behavior. Right: the right-hand side of the exact upper bound (7) versus the expected $\mathcal{O}(M^{-2})$ behavior.

can guarantee absolute stability, i.e., taking

$$k \lesssim \frac{1}{\rho(\boldsymbol{A})},$$

should satisfy absolute stability, where $\rho(\boldsymbol{A})$ is the spectral radius of $\boldsymbol{A}$. Again using Gerschgorin discs, we can explicitly bound the spectral radius:

$$\rho(\boldsymbol{A}) \leq \max_j |\lambda_j(\boldsymbol{A})| \leq \max_\ell |c_\ell| + r_\ell \leq \frac{1}{2}(N+1)^2, \tag{8}$$

where $N = (M-1)/2$, for some (other) universal constant $c > 0$. Hence, we expect that $k \lesssim 1/M^2$ is the requisite stability condition. We verify all of this in Figure 1. In the left pane we show that $\rho(\boldsymbol{A}) \sim M^2$ as expected, and in the right pane we show that the exactly computed upper bound (7) behaves like $k \sim M^{-2}$. Note that one of the eigenvalues of $\boldsymbol{A}$ is zero (corresponding to the constant function), and so that eigenvalue is not computed in the bound. (The zero eigenvalue already satisfies absolute stability.) Hence, we have verified that $k \lesssim M^{-2}$ is the time step restriction. This M^{-2} behavior is a direct result of the second derivative u_{xx} term, which contributes the $-\ell^2$ diagonal entries of $\boldsymbol{A}$. Note that this is essentially the same restriction that the u_{xx} term place on time-stepping for finite-difference methods. Like the finite-difference case, this increases the computational cost of using explicit methods since $k \lesssim M^{-2}$ is the time step restriction. In particular, doubling M means that we must take 4 times as many time steps. Therefore, the cost of numerically solving this PDE using this scheme requires $\mathcal{O}(N^3)$ complexity: each time step requires $\mathcal{O}(N)$ operations, and we must take approximately $T/k \simeq N^2$ time steps.

c. We implement both Runge Kutta 2 (explicit midpoint) and the Trapezoid rule (Runge Kutta 2). The results are shown in figure 2. We use $M = 401$ (corresponding to $N = 200$), and a Runge Kutta time step of $k = 2.5 \times 10^{-5} \approx N^{-2}$, and a Crank-Nicholson timestep of $k = 0.01$. The results in figure 2 show that both solutions behave qualitatively the same, with the diffusion being the dominant characteristic, although some advection is present

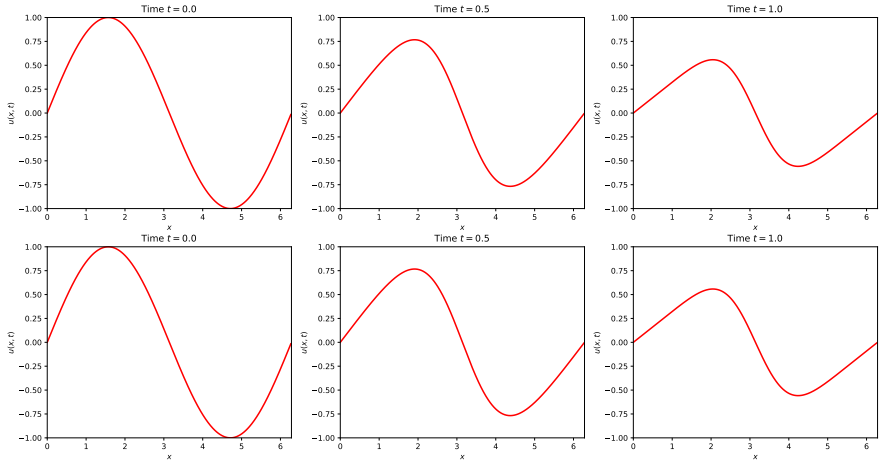

Figure 2: Problem 2c. Top: Fourier-Galerkin solution at various times using Runge-Kutta 2. Bottom: Fourier-Galerkin solution at various times using the trapezoid rule (Crank-Nicholson).

that makes the slope around $x = \pi$ more steep for later times. However, because we implement this approach using sparse matrix-vector multiplies with sparse linear solves, the Trapezoid rule (implicit) method is approximately 100 times faster than the explicit Runge-Kutta based method. This demonstrates the power of implicit methods, assuming one is able to efficiently solve large linear (or even nonlinear) systems.

3. (Viscous Burgers' equation)
Consider the viscous Burgers' partial differential equation,

$$u_t + f(u)_x = \nu u_{xx}, \quad x \in [0, 2\pi) \qquad f(u) = u^2, \qquad u(x, 0) = \sin x.$$

with periodic boundary conditions. Implement both a Fourier-Galerkin and Fourier-collocation solver for this equation. You may use an explicit time-stepping method.

- **a.** For *viscous* Burgers', $\nu > 0$, show and discuss results for $t > 0$ when ν is small, and when ν is large.
- **b.** For *inviscid* Burgers', $\nu = 0$, what differences do you observe between the collocation and Galerkin methods? Do the solutions appear to be accurate for large t?

Solution:

- **a.** We choose $\nu = 0.5$ as a "large" value of ν, and $\nu = 0.1$ as a "small" value of ν. With $M = 101$ degrees of freedom (corresponding to a maximum frequency of $N = 50$), we implement both Fourier-Galerkin and Fourier-collocation solvers, using Runge Kutta 4 for the time-stepping method. The timestep k is chosen as $k = N^{-2}$, consistent with the discussion in problem 2. We show results at times $t = 0, 0.5, 1.0$, and 1.5 in Figure 3 for large viscosity $\nu = 0.5$, and in Figure 4 for small viscosity $\nu = 0.1$. We observe that both solutions yield consistent (and correct) solutions; in these cases, the viscosity ν is still "large" enough so that any onset of discontinuities is reversed by diffusive processes that smooth out the solution. Here, both solutions are reasonable and of comparable computational cost.

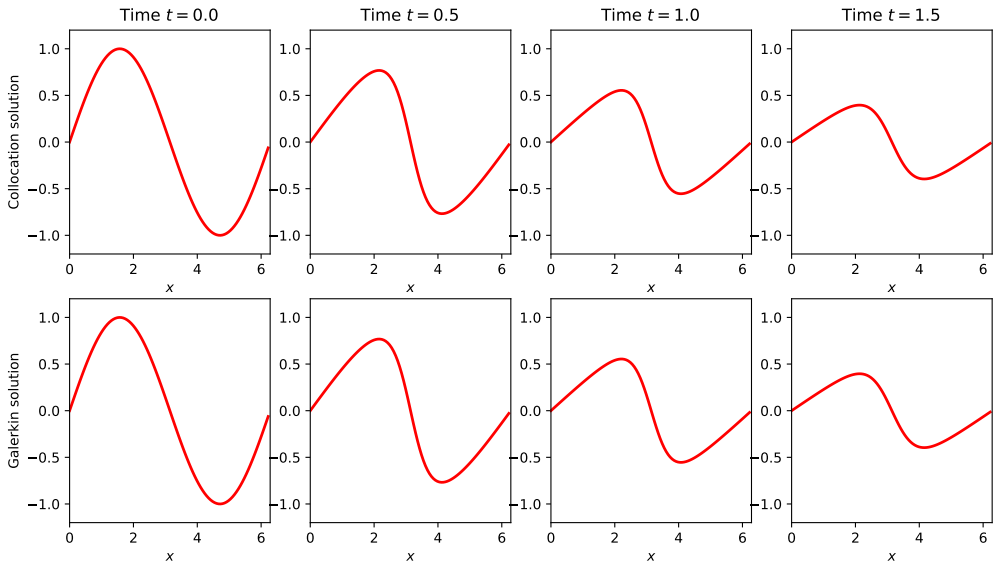

Figure 3: Problem 3a. "Large" viscosity $\nu = 0.5$ Top: Fourier-collocation solution at various times. Bottom: Fourier-Galerkin solution at various times.

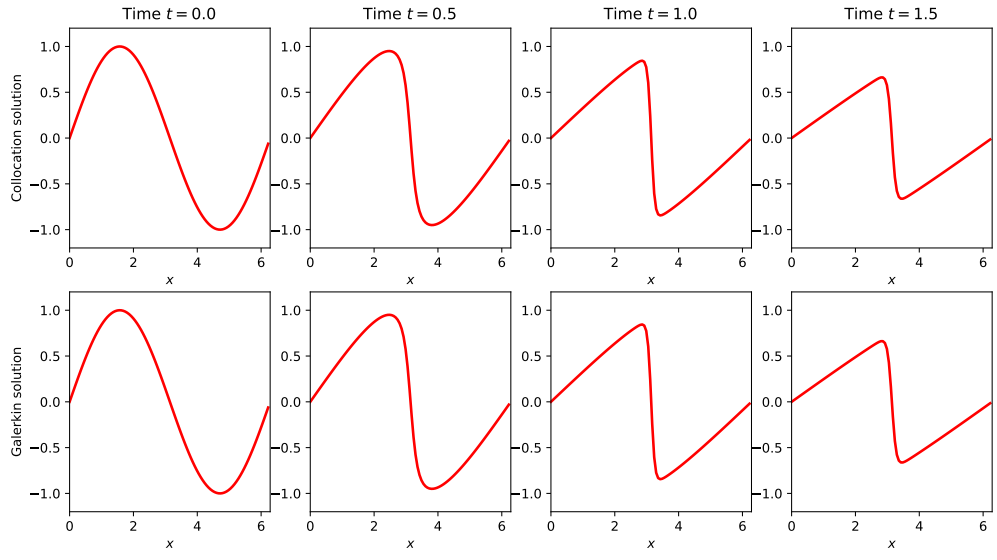

Figure 4: Problem 3a. "Small" viscosity $\nu = 0.1$ Top: Fourier-collocation solution at various times. Bottom: Fourier-Galerkin solution at various times.

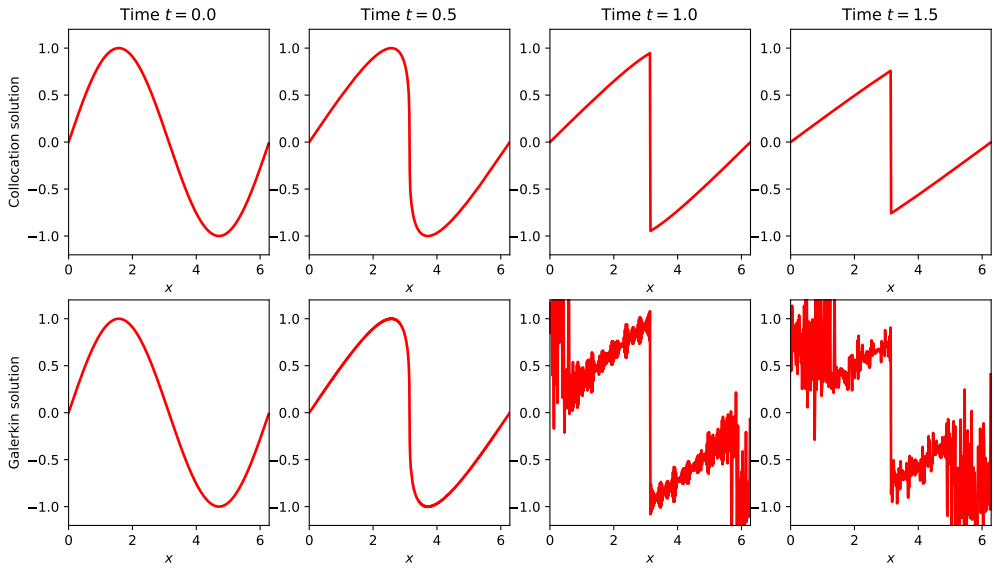

Figure 5: Problem 3b. Inviscid Burgers' equation, $\nu = 0$. Top: Fourier-collocation solution at various times. Bottom: Fourier-Galerkin solution at various times.

b. We set $M = 401$ for this setup and set $\nu = 0$. The results are shown in Figure 5. In this case we see that without viscosity a high-gradient discontinuity develops. The collocation solution is able to resolve this discontinuity, but the Galerkin solution results in considerable spurious oscillations after the discontinuity forms. Here, the collocation solution looks reasonable for larger t, but the Galerkin solution does not.

Complete only one of problems 4-6, of your choice. For each problem: describe and write down a scheme (from the general types that we've covered in this class) to numerically solve the provided PDE. To whatever extent you are able, discuss consistency, stability, convergence, and computational cost of the scheme. Implement the scheme and discuss the numerical results (say with snapshots of the solution), investigating convergence when applicable. The terminal time T is not given; you should investigate the numerical solution for a reasonable range of t.

4. (Korteveg-De Vries equation) Consider the Korteweg-De Vries equation:

$$u_t + u_{xxx} - 6uu_x = 0, \qquad u(x,0) = -2\text{sech}^2(x), \tag{9}$$

over the domain $x \in \mathbb{R}$. There are self-similar solitary wave solutions ("solitons") to this PDE. One such solution is:

$$u(x,t) = -2\text{sech}^2(x - 4t - 5)$$

There are also interacting soliton solutions. The following initial condition corresponds to one such solution:

$$u(x,0) = -2\text{sech}^2(x-5) - \frac{1}{2}\text{sech}^2\left(\frac{1}{2}(x-1)\right).$$

5. (Multidimensional wave equation) Consider the partial differential equation

$$u_t + \boldsymbol{c} \cdot \nabla u(x,y) = 0,$$

with periodic boundary conditions on $(x,y) \in [0,2\pi)^2$. Let the wavespeed be

$$\boldsymbol{c}(x,y) = (\sin(x+y), \sin(x-y)).$$

Let the initial condition be,

$$u(x,y,0) = \sin^6 x \ \sin^6 y$$

6. (1D Euler equations) Consider the one-dimensional Euler equations of gas dynamics. These equations, in conservation form, are written as

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \qquad\qquad t > 0,$$

where the conserved variable $\boldsymbol{u}(x,t) \in \mathbb{R}^3$ and flux $\boldsymbol{f}$ are functions of the physical gas density $\rho(x,t)$, momentum $m(x,t)$, and energy $E(x,t)$:

$$\boldsymbol{u} = \begin{bmatrix} \rho \\ m \\ E \end{bmatrix}, \qquad\qquad \boldsymbol{f} = \begin{bmatrix} m \\ p + \frac{m^2}{\rho} \\ (E+p)\frac{m}{\rho} \end{bmatrix},$$

where the pressure p and energy E are related by

$$p = (\gamma - 1)\left(E - \frac{m^2}{2\rho}\right),$$

where γ is the heat capacity ratio; for this problem set $\gamma = \frac{7}{5}$. The initial condition is,

$$\rho(x,0) = \begin{cases} \frac{5}{8} - \frac{3}{8}\sin\left(x - \frac{1}{2}\right), & \left|x - \frac{1}{2}\right| \le \frac{\pi}{2} \\ 1, & x < \frac{1}{2} - \frac{\pi}{2} \\ \frac{1}{4}, & x > \frac{1}{2} + \frac{\pi}{2} \end{cases} \qquad\qquad m(x,0) = 0,$$

$$E(x,0) = \frac{1}{\gamma - 1}\begin{cases} \frac{3}{5} - \frac{2}{5}\sin\left(x - \frac{1}{2}\right), & \left|x - \frac{1}{2}\right| \le \frac{\pi}{2} \\ 1, & x < \frac{1}{2} - \frac{\pi}{2} \\ \frac{1}{5}, & x > \frac{1}{2} + \frac{\pi}{2} \end{cases}$$

Consider the physical domain to be $x \in [-5,5]$. The boundary conditions are Dirichlet conditions with values given by the value of the initial data at the boundaries.
The following may/not be useful: With $v \coloneqq m/\rho$ the velocity variable, the flux Jacobian is,

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}} = \begin{bmatrix} 0 & 1 & 0 \\ -\frac{3-\gamma}{2}v^2 & (3-\gamma)v & \gamma - 1 \\ -\frac{\gamma E v}{\rho} + (\gamma-1)v^3 & \frac{\gamma E}{\rho} - \frac{3(\gamma-1)v^2}{2} & \gamma v \end{bmatrix},$$

and is diagonalizable with real eigenvalues:

$$\lambda\left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\right) = \{v + c,\ v,\ v - c\},$$

where c is the (local) sound speed:

$$c = \sqrt{\frac{\gamma p}{\rho}}.$$