Department of Applied Mathematics
Preliminary Examination in Numerical Analysis
August 2024

Instructions

You have three hours to complete this exam. Submit solutions to four (and no more) of the following six problems. Please start each problem on a new page. You MUST prove your conclusions or show a counter-example for all problems unless otherwise noted. Write your student ID number (not your name!) on your exam.

Problem 1: Root Finding (25 points)

Consider the task of finding the fixed point of the vector function

$$\vec{G}(\vec{x}) = \begin{bmatrix} g_1(\vec{x}) \\ g_2(\vec{x}) \end{bmatrix}$$

where $\vec{x} = (x, y)$ and $g_1(\vec{x})$, $g_2(\vec{x})$ are in $\mathcal{C}^\infty$. Let $\vec{\alpha} = (\alpha_1, \alpha_2)$ denote the fixed point of $\vec{G}(\vec{x})$.

(a) Derive conditions on the function $\vec{G}(\vec{x})$ that gaurantee that the fixed point iteration

$$\vec{x}_{k+1} = \vec{G}(\vec{x}_k)$$

converges to the fixed point $\vec{\alpha}$ for all initial guesses $\vec{x}_0$ in a neighborhood D of the fixed point.

(b) Prove that when the condition you found in part (a) is satisified the fixed point iteration converges linearly.

Solution:

(a) By definition of fixed point and the mean value theorem, we know that

$$\begin{aligned} \alpha_1 - x_{k+1} &= g_1(\alpha_1, \alpha_2) - g_1(x_k, y_k) \\ &= \frac{\partial g_1(\hat{x}_k, \hat{y}_k)}{\partial x}(\alpha_1 - x_k) + \frac{\partial g_1(\hat{x}_k, \hat{y}_k)}{\partial y}(\alpha_2 - y_k) \end{aligned}$$

for some $(\hat{x}_k, \hat{y}_k)$ on the line connecting (α_1, α_2) and (x_k, y_k).

Likewise,

$$\begin{aligned} \alpha_2 - y_{k+1} &= g_2(\alpha_1, \alpha_2) - g_2(x_k, y_k) \\ &= \frac{\partial g_2(\bar{x}_k, \bar{y}_k)}{\partial x}(\alpha_1 - x_k) + \frac{\partial g_2(\bar{x}_k, \bar{y}_k)}{\partial y}(\alpha_2 - y_k) \end{aligned}$$

for some $(\bar{x}_k, \bar{y}_k)$ on the line connecting (α_1, α_2) and (x_k, y_k).

This means that

$$\vec{\alpha} - \vec{x}_{k+1} = \mathbf{J}_k(\vec{\alpha} - \vec{x}_k)$$

where $\mathbf{J}_k$ is the Jacobian defined by

$$\begin{bmatrix} \frac{\partial g_1(\hat{x}_k,\hat{y}_k)}{\partial x} & \frac{\partial g_1(\hat{x}_k,\hat{y}_k)}{\partial y} \\ \frac{\partial g_2(\bar{x}_k,\bar{y}_k)}{\partial x} & \frac{\partial g_2(\bar{x}_k,\bar{y}_k)}{\partial y} \end{bmatrix}$$

This means that

$$\|\vec{\alpha} - \vec{x}_{k+1}\| \leq \|\mathbf{J}_k\|\|(\vec{\alpha} - \vec{x}_k)\|.$$

Let $M = max_{\vec{x}\in D}\|\mathbf{J}\|$, then

$$\|\vec{\alpha} - \vec{x}_{k+1}\| \leq M\|(\vec{\alpha} - \vec{x}_k)\|.$$

If $M < 1$, the fixed point iteration is a contraction and it will converge.

(b)

$$\lim_{k\to\infty} \frac{\|\vec{\alpha} - \vec{x}_{k+1}\|}{\|(\vec{\alpha} - \vec{x}_k)\|} \leq M$$

since M is a constant independent of k, the method is first order convergent.

Problem 2: Interpolation/Approximation (25 points)

We denote the weighted L^2 inner-product of two functions u and v by:

$$(u, v) = \int_a^b w(x)u(x)v(x)dx$$

where w is a positive weight function. We associate to this inner-product the norm

$$\|u\|_{L^2(a,b)} = \left(\int_a^b w(x)(u(x))^2 dx\right)^{1/2}$$

(a) Let $\{\Psi_j\}_{0\leq j\leq n}$ be a set of nonzero, orthogonal (with respect to $(\cdot,\cdot)$) polynomials of degree less than or equal to n.

The space of polynomials of degree less than or equal to n is denoted by $\mathbb{P}_n$. Prove that the $(\Psi_0, \Psi_1, \ldots, \Psi_n)$ forms a basis for $\mathbb{P}_n$.

(b) Let Φ_j's be a set of polynomials defined by

$$\Phi_0(x) = 1, \quad \Phi_1(x) = x - \frac{(x,1)}{(1,1)}$$
$$\Phi_j(x) = (x-a)\Phi_{j-1}(x) - b\Phi_{j-2}(x), \quad j \geq 2$$

with

$$a = \frac{(x\Phi_{j-1}(x), \Phi_{j-1}(x))}{(\Phi_{j-1}(x), \Phi_{j-1}(x))} \quad b = \frac{(x\Phi_{j-1}(x), \Phi_{j-2}(x))}{(\Phi_{j-2}(x), \Phi_{j-2}(x))}$$

Show that the Φ_j's form a set of orthogonal polynomials.

(c) Let $f \in \mathcal{C}(a, b)$. Use orthogonal polynomials to derive a general solution to the following problem

$$\min_{p\in\mathbb{P}_n} \|f - p\|_{L^2(a,b)}$$

(d) Find the line that best approximates $\sqrt{x}$ in the weighted L^2 norm on the interval $(0, 1)$. The weight function is chosen to be $w = 1$.

Solution:

(a) Let $p(x) = \sum_{j=1}^n \alpha_j \Phi_j(x) \in \mathbb{P}_n$. We need to show that the only way $p(x) = 0$ is if $\alpha_j = 0$ for all j.

We do this by taking an inner produce with Φ_l for $l = 1, \ldots, n$.

$$0 = (0, \Phi_l) = \sum_{j=1}^n \alpha_j(\Phi_j(x), \Phi_l(x))$$
$$\alpha_l(\Phi_l(x)\Phi_l(x))$$

since $(\Phi_j(x), \Phi_l(x)) = 0$ for $j = l$ for $j \neq l$. Thus $\alpha_l = 0$ for all l.

(b) First check that Φ_0 and Φ_1 are orthogonal.

$$\begin{aligned}(\Phi_0, \Phi_1) &= (1, x - \frac{(x,1)}{(1,1)}) \\ &= (1,x) - \frac{(1,1)(x,1)}{(1,1)} \\ &= (1,x) - (x,1)\end{aligned}$$

since the inner product is symmetric.

Now we will prove by induction for $j >= 2$ Assume that Φ_j form an orthogonal set of polynomials for $j < n$. Now we must show that Φ_{n+1} is orthogonal to Φ_j for all $j <= n$. From the recursion we know that

$$\Phi_{n+1}(x) = (x-a)\Phi_n(x) - b\Phi_{n-1}(x)$$

so

$$\begin{aligned}(\Phi_{n+1}(x), \Phi_j(x)) &= ((x-a)\Phi_n(x) - b\Phi_{n-1}(x), \Phi_j(x)) \\ &= ((x-a)\Phi_n(x), \Phi_j(x)) - (b\Phi_{n-1}(x), \Phi_j(x)) \\ &= (x\Phi_n(x), \Phi_j(x)) - a(x\Phi_n(x), \Phi_j(x)) - (b\Phi_{n-1}(x), \Phi_j(x))\end{aligned}$$

If $j \leq n-2$ all the inner products are equal to zero due to assumed orthogonality.

If $j = n-1$, $(\Phi_n, \Phi_{n-1}) = 0$ due to assumed orthogonality and the remainder items add to 0.
If $j = n$, $(\Phi_{n-1}, \Phi_n) = 0$ due to assumed orthogonality and the remainder items add to 0.

(c) We will write our approximation as follows

$$p(x) = \sum_{j=1}^{n} \alpha_j \Phi_j(x)$$

. Plugging this into the inner product we get the following expression

$$(f-p, f-p) = \|f\|^2 - 2\sum_{j=1}^{n} \alpha_j (f, \Phi_j) + \sum_{j=1}^{n} \alpha_j^2 (\Phi_j, \Phi_j)$$

This is an upward facing quadratic of α_j. We want to find the minimum of this fuction so we will take the deriviative wrt α_j and set it equal to 0. In the end we the that $\alpha_j = \frac{-(f,\Phi_j)}{(\Phi_j,\Phi_j)}$.

(d) We can use our polynomials given in part (b) and the coefficients in part (c) to get our approximation. $p(x) = \frac{(\sqrt{x},\Phi_0)}{(\Phi_0,\Phi_0)} - \frac{(\sqrt{x},\Phi_1)}{(\Phi_1,\Phi_1)}\Phi_1(x) = 2/3 - 1/3x$

Problem 3: Quadrature (25 points)

Consider the task of numerically approximating

$$I(f) = \int_a^b f(x)dx$$

where $f \in C^\infty[a,b]$.

(a) Derive the trapezoidal rule and corresponding error for approximating $I(f)$.
Useful information: $\int_a^b (x-a)(x-b)dx = -\frac{1}{6}(b-a)^3$

(b) Find the formula for the composite trapezoidal rule using uniform intervals of size $h = \frac{b-a}{n}$ where $n+1$ is the number of quadrature points. i.e. the quadrature points are $x_j = a + j * h$ for $j = 0, \ldots, n$

(c) Derive the error for the composite trapezoidal rule.

Solution:

(a) The trapezoidal rule is based on integrating the linear interpolation with interpolation points $x_0 = a$ and $x_1 = b$. Using this information, we use Taylor's Theorem with a modified Lagrange remainder term to rewrite $f(x)$ as

$$f(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a} + \frac{f''(\eta_x)}{2}(x-a)(x-b)$$

for some $\eta_x \in [a,b]$.
Integrating the linear approximation we find that the trapezoidal rule is given by

$$I_1(f) = \int_a^b \left[f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a} \right] = \frac{f(a)+f(b)}{2}(b-a).$$

The error in approximating the integral using the trapezoidal rule has an upper bound given by

$$\begin{aligned} E_1(f) = I(f) - I_1(f) &= \int_a^b \frac{f''(\eta_x)}{2}(x-a)(x-b)dx \\ &= f''(\eta)\frac{(b-a)^3}{12} \end{aligned}$$

for some $\eta \in [a,b]$ by the mean value theorem. (See for example Chapter 1 Thm 1.3 of Atkinson Numerical Analysis text.) The trapezoidal rule is given by $I_1(f) = \frac{f(a)+f(b)}{2}(b-a)$ and the error term is $E_1(f) = f''(\eta)\frac{(b-a)^3}{12}$.

(b) $I_n(f) = h\left[\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{n-1}) + \frac{f(x_n)}{2}\right]$

(c)

$$E_n(f) \leq I(f) - I_n(f) = \frac{-h^3}{12} \sum_{j=1}^{n} f''(\eta_j)$$
$$= \frac{-h^3 n}{12} \left(\frac{1}{n} \sum_{j=1}^{n} f''(\eta_j) \right)$$
$$= \frac{-h^3 n(b-a)}{12(b-a)} \left(\frac{1}{n} \sum_{j=1}^{n} f''(\eta_j) \right)$$

We know that

$$\min_{x \in [a,b]} f''(x) \leq 1/n \sum_{j=1}^{n} f''(\eta_j) \leq \max_{x \in [a,b]} f''(x).$$

Since $f''(x)$ is continuous in $[a, b]$, there exists an $\eta \in [a, b]$ such that

$$1/n \sum_{j=1}^{n} f''(\eta_j) = f''(\eta).$$

Thus $E_n(f) = \frac{-h^2(b-a)}{12} f''(\eta)$ for some $\eta \in [a, b]$.

Problem 4: Linear Algebra (25 points)

Let $\mathtt{A} \in \mathbb{R}^{n\times n}$ a normal matrix with eigenvalues λ_i and corresponding eigenvectors $\vec{u}_i$ (forming an orthonormal basis of $\mathbb{C}^n$). Further, assume that $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Note that this assumption implies $(\lambda_1, \vec{u}_1)$ are real. We consider the power method to obtain the dominant eigenpair $(\lambda_1, \vec{u}_1)$, given an initial guess for the eigenvector $\vec{z}_0$:

$$\begin{aligned}\vec{w}_{k+1} &= \mathtt{A}\vec{z}_k \\ \vec{z}_{k+1} &= \frac{\vec{w}_{k+1}}{||\vec{w}_{k+1}||_\infty} \\ \lambda_{k+1} &= \frac{\vec{z}_k^* \vec{w}_{k+1}}{\vec{z}_k{}^* \vec{z}_k}\end{aligned}$$

(a) Consider the Rayleigh quotient function $R_{\mathtt{A}}(\vec{z}) = \frac{\vec{z}^* \mathtt{A}\vec{z}}{\vec{z}^*\vec{z}}$ for $\vec{z} \in \mathbb{C}^n$. Find a formula for $R_{\mathtt{A}}(\vec{z})$ that does not involve any vectors. Use this formula to prove that the approximation for λ_{k+1} in the power method converges to λ_1 as $k \to \infty$.

(b) Let μ be an estimate for the simple, real eigenvalue λ_q. State a necessary and sufficient condition for μ such that the power method applied to $(\mathtt{A} - \mu\mathtt{I})^{-1}$ converge linearly to $\vec{u}_q$ (up to a scalar factor). What is the linear rate of convergence?

(c) Given $\mu_0 = \mu$ and $\vec{z}_0$, consider the following algorithm:

$$\begin{aligned}\vec{w}_{k+1} &= (\mathtt{A} - \mu_k\mathtt{I})^{-1}\vec{z}_k \\ \vec{z}_{k+1} &= \frac{\vec{w}_{k+1}}{||\vec{w}_{k+1}||_\infty} \\ \lambda_{k+1} &= \frac{\vec{z}_k^* \mathtt{A}\vec{z}_k}{\vec{z}_k{}^* \vec{z}_k} \\ \mu_{k+1} &= \lambda_{k+1}\end{aligned}$$

Explain in what sense this algorithm is an acceleration of the one proposed in (b) (Hint: does the order or rate improve?).

Solution:

(a) Given that the $\{u_j\}_{j=1}^n$ are an orthonormal basis of eigenvectors, the numerator and denominator both greatly simplify:

$$\vec{z}^* \mathtt{A}\vec{z} = \left(\sum_{i=1}^n \overline{\alpha_i}\vec{u}_i^*\right)\left(\sum_{j=1}^n \alpha_j \mathtt{A}\vec{u}_j\right) = \left(\sum_{i=1}^n \lambda_i |\alpha_i|^2\right) = \lambda_1 \left(\sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_1}\right)|\alpha_i|^2\right)$$

$$\vec{z}^* \vec{z} = \left(\sum_{i=1}^n \overline{\alpha_i}\vec{u}_i^*\right)\left(\sum_{j=1}^n \alpha_j \vec{u}_j\right) = \left(\sum_{i=1}^n |\alpha_i|^2\right)$$

Let $\vec{z}_0 = \sum_{i=1}^n \beta_i \vec{u}_i$. After k power method iterations, iterate $\vec{z}_k$ is proportional to $\sum_{i=1}^n \beta_i \lambda_i^k \vec{u}_i$ (constants in front of it cancel in the Rayleigh quotient). Plugging that into our formula gives us:

$$R(z_k) = \lambda_1 \frac{\left(\sum_{i=1}^n \left(\frac{\lambda_i |\lambda_i|^{2k}}{\lambda_1}\right) |\beta_i|^2\right)}{(\sum_{i=1}^n |\lambda_i|^{2k} |\beta_i|^2)} = \lambda_1 \frac{\left(\sum_{i=1}^n \left(\frac{\lambda_i |\lambda_i|^{2k}}{\lambda_1 |\lambda_1|^{2k}}\right) |\beta_i|^2\right)}{\left(\sum_{i=1}^n \frac{|\lambda_i|^{2k}}{|\lambda_1|^{2k}} |\beta_i|^2\right)}$$

where we divide both numerator and denominator by $|\lambda_1|^{2k}$. It is clear from this that as k goes to infinity, $R(z_k)$ converges to λ_1, and that the error (difference between them) is $O\left(\left(\frac{|\lambda_2|}{|\lambda_1|}\right)^2\right)$.

(b) The spectra of $(\mathtt{A} - \mu\mathtt{I})^{-1}$ is $\{\frac{1}{\lambda_i - \mu}\}_{i=1}^n$. Given that we want the power method to hone into λ_q, we need our estimate μ to make $\frac{1}{\lambda_q - \mu}$ the dominant eigenvalue. So, a necessary and sufficient condition is that $|\lambda_q - \mu| = \text{argmin}_j |\lambda_j - \mu|$, that is, that μ is closer to λ_q than it is to any other eigenvalue of $\mathtt{A}$.

We know from class and from the analysis above that the power method converges linearly (to the leading eigenvector) with linear rate $\frac{|\lambda_2|}{|\lambda_1|}$. So, the linear rate of convergence for the inverse, shifted power method proposed above will be the ratio of the second to the first largest eigenvalues of $(\mathtt{A} - \mu\mathtt{I})^{-1}$. We could denote that as:

$$\frac{|\lambda_q - \mu|}{\text{argmin}_{j \neq q} |\lambda_j - \mu|}$$

how good this rate is will depend on the distance of μ to λ_q relative to its distance to other nearby eigenvalues (the closer it is to λ_q, the faster the algorithm will converge).

(c) Given our analysis in (a) and (b), this algorithm is now improving upon our estimate of the leading eigenvalue *every iteration*, and the linear rate of convergence (the improvement on the error in one step) is now given by the formula

$$\frac{|\lambda_q - \mu_k|}{\text{argmin}_{j \neq q} |\lambda_j - \mu_k|}$$

This is enough for us to conclude that this new algorithm is *superlinearly* convergent. (If we do a bit more analysis using the formula in (a), we could conclude it is, in fact, cubically convergent).

Problem 5: Numerical ODE (25 points)

We wish to numerically solve the IVP for a system of N first order ODEs $\{\vec{y}'(t) = f(t, \vec{y}),\ \vec{y}(0) = \vec{y}_0\}$. Consider the family of single-step methods

$$y_{n+1} = y_n + \Delta t\,(\theta f(t_n, y_n) + (1-\theta) f(t_{n+1}, y_{n+1}))$$

for a parameter $\theta \in [0, 1]$. This is sometimes known as the family of θ methods.

(a) Determine for which values of θ these methods are consistent. For the values of θ where the method is consistent, determine the order of the method.

(b) Derive an equation for the region of absolute stability R_θ that lies in the complex plane $\mathbb{C}$. For all values of $\theta \in [0, 1]$, describe geometrically the region of absolute stability. (It looks different for different values of θ.) Determine for what values of θ is the method A-stable.

(c) Determine for which values of θ these methods are explicit or implicit. In the case of θ where the method is implicit, explain what method (and what inputs or parameters) you would use to compute the next timestep and why.

Solution:

(a) To test for consistency and order for the truncation error, we apply the conditions for multistep methods

$$a_1 y_{n+1} + a_0 y_n = \Delta t\,(b_1 f(t_{n+1}, y_{n+1}) + b_0 f(t_n, y_n))$$

with $a_1 = 1, a_0 = -1, b_1 = 1-\theta, b_0 = \theta$. These conditions are:

$$\begin{aligned}
a_1 + a_0 &= 1 - 1 = 0 \\
\sum_{m=0}^{1} (m a_m - b_m) &= (0-\theta) + (1-(1-\theta)) = 0 \\
\sum_{m=0}^{1} (m^2 a_m - 2m b_m) &= (0-0) + (1-2(1-\theta)) = 2\theta - 1 \\
\sum_{m=0}^{1} (m^3 a_m - 3m^2 b_m) &= (0-0) + (1-3(1-\theta)) = 3\theta - 2
\end{aligned}$$

this tells us that all θ methods are consistent and at least first order, but only for $\theta = \frac{1}{2}$ (Implicit Trapezoidal) is the method order 2. No θ method is higher order ($3/2 - 2 \neq 0$).

(b) Applying the θ method to the model problem $y' = \lambda y, y(0) = 1$ gives us:

$$\begin{aligned}
y_{n+1} - y_n &= \Delta t \lambda\,(\theta y_n + (1-\theta) y_{n+1}) \\
y_{n+1} &= \frac{1 + (\Delta t \lambda)\theta}{1 - (\Delta t \lambda)(1-\theta)} y_n
\end{aligned}$$

So, the region of absolute stability is the region in the left complex plane for which this factor has modulus less than or equal to 1. That is:

$$R_\theta = \left\{ z \in \mathbb{C} | Re(z) \leq 0 \text{ and } \left| \frac{1+\theta z}{1-(1-\theta)z} \right| \leq 1 \right\}$$

Writing $z = x + iy$, we can square both sides of the inequality obtained from the linear equation to find:

$$\begin{aligned}
\left| \frac{1+\theta z}{1-(1-\theta)z} \right|^2 = \frac{(1-\theta x)^2 + \theta^2 y^2}{(1+(\theta-1)x)^2 + (\theta-1)^2 y^2} &<= 1 \\
\frac{1 + 2\theta x + \theta^2(x^2+y^2)}{[1 + 2\theta x + \theta^2(x^2+y^2)] - 2x + (1-2\theta)(x^2+y^2)} &<= 1 \\
(2\theta - 1)(x^2+y^2) + 2x &\leq 0 \\
(2\theta - 1)(x^2 + \frac{2}{2\theta-1}x + y^2) &\leq 0 \\
(2\theta - 1)\left(\left(x + \frac{1}{2\theta-1} \right)^2 - \frac{1}{(2\theta-1)^2} + y^2 \right) &\leq 0
\end{aligned}$$

If $2\theta - 1 > 0$, meaning $\theta > \frac{1}{2}$, then this is a circle of radius $\frac{1}{(2\theta-1)}$ centered at $x = -\frac{1}{2\theta-1}, y = 0$, which is entirely located in the left $\mathbb{C}$ plane. We know, for example, that $\theta = 1$ (Forward Euler) gives us the inside of a disk of radius 1 centered at $x = -1$.

If $2\theta - 1 < 0$, meaning $\theta < \frac{1}{2}$, the sign of the inequality changes when we divide by $2\theta - 1$, and we get the exterior of a disk of radius $\frac{1}{(1-2\theta)}$ centered at $x = \frac{1}{(2\theta-1)}, y = 0$. This disk is entirely contained in the right complex plane, so these methods are A-stable. Finally, $\theta = \frac{1}{2}$ means the inequality is automatically satisfied, and once again, the method is A-stable. So, θ methods are A-stable for $\theta \in [0, \frac{1}{2}]$.

(c) All θ methods are implicit except for $\theta = 1$ (Forward Euler), which is explicit (this is the only value for which the coefficient for $f(t_{n+1}, y_{n+1})$ is zero). For a given θ for which the method is implicit, say $\theta = \frac{1}{2}$, we have to solve the potentially non-linear system of equations:

$$y - \Delta t(1-\theta) f(t_{n+1}, y) = y_n + \Delta t \theta f(t_n, y_n)$$

given our previous guess y_n at time t_n. We propose to use the Newton method with initial guess $y_0 = y_n$ (we expect this to be close to y_{n+1} for small Δt) and a target accuracy that is at least proportional to the local truncation error for the method being used (second order for Implicit Trapezoidal, first order otherwise).

Problem 6: Numerical PDE (25 points)

Consider the following IVP for an advection-diffusion PDE with periodic boundary conditions:

$$\begin{aligned} u_t(x,t) = bu_{xx}(x,t) - au_x(x,t) + f(x) \quad & (x,t) \in (-\pi,\pi) \times (0,T) \\ u(x,0) = \phi(x) \quad & x \in (-\pi,\pi) \\ u(-\pi,t) = u(\pi,t) \quad & t \in [0,T] \end{aligned}$$

with $a \in \mathbb{R}, b \geq 0$ and ϕ, f smooth and 2π periodic.

(a) For a regular grid in x with n points spaced by $\Delta x = \frac{2\pi}{n}$, we can write $U_j(t) \simeq u(x_j, t)$. Using the forward difference for u_x and centered second difference for u_{xx}, write down a system of n first-order ODEs in time for $U_j(t)$.

(b) Note that this system can be written as $\vec{U}(t)' = \mathtt{M}\vec{U}(t) + \vec{F}$. Describe the entries and structure of matrix $\mathtt{M}$.

(c) Use the implicit trapezoidal method to discretize the system above in time, and write down an equation to compute $\vec{U}(t_{k+1})$ in terms of $\vec{U}(t_k)$.

(d) If you are given a formula for the eigenvalues $\{\lambda_j(\Delta x)\}_{j=1}^n$ of matrix $\mathtt{M}$, explain how you could use them to perform a stability analysis on the finite difference scheme described above.

Solution:

(a) We discretize the first and second derivatives in the advection-diffusion PDE, and obtain the following:

$$\begin{aligned} U_j'(t) &= \frac{b}{\Delta x^2}\left(U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)\right) - \frac{a}{\Delta x}\left(U_{j+1}(t) - U_j(t)\right) + F(x_j) \\ U_j(0) &= \phi(x_j) \ , \ U_{n+1}(t) = U_1(t) \end{aligned}$$

for $j = 1, \ldots, n$, and periodicity is set by $U_{n+1}(t) = U_1(t)$.

(b) Note that this system can be written as $\vec{U}(t)' = \mathtt{M}\vec{U}(t) + \vec{F}$. Describe the entries and structure of matrix $\mathtt{M}$.

Matrix $\mathtt{M}$ is sparse, circulant with only three non-zero diagonals (it would be tridiagonal except for two entries on the top right and bottom left corners). That is,

$$\mathtt{M} = \begin{bmatrix} -\frac{2b}{\Delta x^2} + \frac{a}{\Delta x} & \frac{b}{\Delta x^2} - \frac{a}{\Delta x} & 0 & \cdots & 0 & \frac{b}{\Delta x^2} \\ \frac{b}{\Delta x^2} & -\frac{2b}{\Delta x^2} + \frac{a}{\Delta x} & \frac{b}{\Delta x^2} - \frac{a}{\Delta x} & \cdots & 0 & 0 \\ 0 & \frac{b}{\Delta x^2} & -\frac{2b}{\Delta x^2} + \frac{a}{\Delta x} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \cdots & 0 & 0 \\ \frac{b}{\Delta x^2} - \frac{a}{\Delta x} & 0 & 0 & \cdots & \frac{b}{\Delta x^2} & -\frac{2b}{\Delta x^2} + \frac{a}{\Delta x} \end{bmatrix}$$

this means that it can be applied and operated with using fast algorithms for sparse matrices, and that it is diagonalized by the Fourier basis, which means we can use FFT-based algorithms (recall that we are solving a PDE with periodic boundary conditions).

(c) Implicit Trapezoidal applied to the system of ODEs we are working with gives us the following:

$$\begin{aligned}
\vec{U}(t_{k+1}) &= \vec{U}(t_k) + \frac{\Delta t}{2}\left(\mathtt{M}\vec{U}(t_{k+1}) + \vec{F} + \mathtt{M}\vec{U}(t_k) + \vec{F}\right) \\
\left(\mathtt{I} - \frac{\Delta t}{2}\mathtt{M}\right)\vec{U}(t_{k+1}) &= \left(\mathtt{I} + \frac{\Delta t}{2}\mathtt{M}\right)\vec{U}(t_k) + \Delta t\vec{F} \\
\vec{U}(t_{k+1}) &= \left(\mathtt{I} - \frac{\Delta t}{2}\mathtt{M}\right)^{-1}\left(\left(\mathtt{I} + \frac{\Delta t}{2}\mathtt{M}\right)\vec{U}(t_k) + \Delta t\vec{F}\right) \\
\vec{U}(t_{k+1}) &= \left(\mathtt{I} - \frac{\Delta t}{2}\mathtt{M}\right)^{-1}\left(\mathtt{I} + \frac{\Delta t}{2}\mathtt{M}\right)\vec{U}(t_k) + \Delta t\left(\mathtt{I} - \frac{\Delta t}{2}\mathtt{M}\right)^{-1}\vec{F}
\end{aligned}$$

(d) Let $\{\lambda_j(\Delta x), \vec{v}_j(\Delta x)\}$ be an eigenpair of $\mathtt{M}$, and denote $\mathtt{L}(\Delta t, \Delta x) = \left(\mathtt{I} - \frac{\Delta t}{2}\mathtt{M}\right)^{-1}\left(\mathtt{I} + \frac{\Delta t}{2}\mathtt{M}\right)$. Then, we know that

$$\mathtt{L}(\Delta t, \Delta x)\vec{v}_j(\Delta x) = \left(\frac{1 + \frac{\Delta t}{2}\lambda_j(\Delta x)}{1 - \frac{\Delta t}{2}\lambda_j(\Delta x)}\right)\vec{v}_j(\Delta x)$$

That is, we have a formula for the eigenvalues of the matrix being applied repeatedly (via a linear system solve) to take each time-step. If we have two numerical solutions $\vec{V}_n$ and $\vec{W}_n$ that start from different initial data (that differ in norm by less than some small ε), then we have that:

$$\begin{aligned}
\vec{V}_k - \vec{W}_k &= \mathtt{L}(\Delta t, \Delta x)(\vec{V}_{k-1} - \vec{W}_{k-1}) \\
&= (\mathtt{L}(\Delta t, \Delta x))^k\,(\vec{V}_0 - \vec{W}_0) \\
||\vec{V}_k - \vec{W}_k|| &\leq ||\mathtt{L}(\Delta t, \Delta x)||^k||\vec{V}_0 - \vec{W}_0|| \leq ||\mathtt{L}(\Delta t, \Delta x)||^k\varepsilon
\end{aligned}$$

If we write $(\vec{V}_0 - \vec{W}_0)$ in the basis of eigenvectors of $\mathtt{M}$, what we need for stability is that $||\mathtt{L}(\Delta t, \Delta x)||^k$ stays bounded for $k = 1, 2, \ldots, T/\Delta t$ (we don't want the error to blow up in any direction). One way to ensure that is to ask that the spectral radius of $\mathtt{L}$ stays within the complex unit disk. That is:

$$\left|\frac{1 + \frac{\Delta t}{2}\lambda_j(\Delta x)}{1 - \frac{\Delta t}{2}\lambda_j(\Delta x)}\right| \leq 1 \quad \forall j = 1, \cdots, n$$

In general, this will impose a condition bounding Δt by some function of Δx and/or Δx^2 (e.g. CFL condition), unless the finite difference scheme is unconditionally stable (e.g. Crank Nicholson for the diffusion PDE). Note that the analysis we just went through is closely related to the Von Neumann stability analysis for the same scheme, which performs Fourier analysis and finds an amplification factor for each Fourier basis term.