# UW–Madison Math/CS 714

# Methods of Computational Mathematics I

## Boundary value problems II

Instructor: Yue Sun (yue.sun@wisc.edu)

September 11, 2025

# Stability

This is not completely convincing. We are assuming that solving the difference equations gives a reasonable approximation of the underlying differential equation.

Instead, let us focus on the discrete system

$$A^h E^h = -\tau^h$$

where the $h$ superscripts indicate that the mesh width is $h$. (Here, $A \in \mathbb{R}^{m \times m}$.)

Solving the system gives

$$E^h = -(A^h)^{-1} \tau^h.$$

# Stability

Taking norms[1] gives

$$\|E^h\| = \|(A^h)^{-1}\tau^h\| \leq \|(A^h)^{-1}\|\|\tau^h\|.$$

We know that $\|\tau^h\| = O(h^2)$. Thus if $\|(A^h)^{-1}\|$ is bounded as $h \to 0$, then we will obtain $\|E^h\| = O(h^2)$ as desired.

We want $\|(A^h)^{-1}\| \leq C$ for all sufficiently small $h$. This motivates our definition of stability.

[1] This expression is using the matrix norm induced by a vector norm. See AM205 video 0.3 and associated notes.

# Stability

Suppose a finite difference method for a linear BVP gives a sequence of equations of the form $A^h U^h = F^h$ where $h$ is the mesh width. Then the method is stable if $(A^h)^{-1}$ exists for all $h$, and there are constants $C > 0$ and $h_0 > 0$ such that

$$\|(A^h)^{-1}\| \leq C \qquad \text{for all } h < h_0.$$

# Consistency

A method is consistent with the differential equation and boundary conditions if

$$\|\tau^h\| \to 0 \qquad \text{as } h \to 0.$$

Usually $\|\tau^h\| = O(h^p)$ for some integer $p > 0$, which implies that the method is consistent.

# Convergence

A method is convergent if $\|E^h\| \to 0$ as $h \to 0$. Using the previous ideas,

$$\text{(consistency)} + \text{(stability)} = \text{(convergence)}.$$

While the derivation focused on a linear BVP, the same principles can be applied to most finite difference approximations of differential equations.

The statement can be strengthened to

$$(O(h^p) \text{ local trunc. error}) + \text{(stability)} = (O(h^p) \text{ global error}).$$

# Proving stability

We know how to check the local truncation error. But it is less obvious how to check stability, even for a linear BVP.

For more complicated problems, the notion of stability may need to change.

# Stability in the 2-norm

The methods we use for stability analysis will depend on the choice of norm. For now, we focus on the 2-norm of the linear BVP.

Since $A$ is symmetric, the 2-norm is equal to its spectral radius,

$$\|A\|_2 = \rho(A) = \max_{1 \le p \le m} |\lambda_p|,$$

where $\lambda_p$ is the $p$th eigenvalue of $A$. Since $A^{-1}$ is also symmetric,

$$\|A^{-1}\|_2 = \rho(A^{-1}) = \max_{1 \le p \le m} |\lambda_p^{-1}| = \left( \min_{1 \le p \le m} |\lambda_p| \right)^{-1}.$$

Hence, we need to compute the eigenvalues of $A$ and show they are bounded away from zero as $h \to 0$.

# Stability in the 2-norm

The derivation shows that the smallest eigenvalue of $A$ is

$$\lambda_1 = \frac{2(\cos \pi h - 1)}{h^2} = -\pi^2 + O(h^2)$$

This is bounded away from zero as $h \to 0$. Furthermore,

$$\|E^h\|_2 \leq \|(A^h)^{-1}\|_2 \|\tau^h\|_2 \approx \frac{1}{\pi^2} \|\tau^h\|_2.$$

Since $\tau_j^h \approx \frac{h^2}{12} u^{(4)}(x_j)$, then

$$\|\tau^h\|_2 \approx \frac{h^2}{12} \|u^{(4)}\|_2 = \frac{h^2}{12} \|f''\|_2,$$

which shows how the LTE will depend on the function $f$.

# Stability in the 2-norm

The eigenvectors that we derived for the discrete system are related to the eigenfunctions of the differential operator $\partial^2/\partial x^2$. Consider

$$u^p = \sin p\pi x$$

for $p = 1, 2, \ldots$. They satisfy

$$\frac{\partial^2}{\partial x^2} u^p = \mu_p u^p$$

where $\mu_p = -p^2\pi^2$. They also satisfy the homogenous boundary conditions $u^p(0) = u^p(1) = 0$.

# Stability in other norms

Examining stability in other norms requires a different approach.

In particular, proving that $\|E\|_\infty = O(h^2)$ would be useful, because $\|E\|_\infty = \max_j |E_j|$ so this would bound the maximum pointwise error.

# A bound on $\|E\|_\infty$

We can use the bound on $\|E\|_2$ to obtain a bound on $\|E\|_\infty$. Recall $E = (E_1, E_2, \ldots, E_m)$ are let $E_j$ be the component with largest magnitude. Then

$$\begin{aligned}\|E\|_2 &= \sqrt{\frac{1}{h}\sum_{i=1}^{m} |E_i|^2} \\ &\geq \sqrt{\frac{1}{h}|E_j|^2} = \frac{1}{\sqrt{h}}|E_j| = \frac{\|E\|_\infty}{\sqrt{h}}.\end{aligned}$$

Since $\|E\|_2 = O(h^2)$, it follows that $\|E\|_\infty = O(h^{3/2})$.

This is useful, although a direct analysis can do better,[2] and show that $\|E\|_\infty = O(h^2)$.

[2] See Sec. 2.11 in the finite-difference textbook, which uses Green's function solutions.

# Neumann boundary conditions

A Neumann boundary condition specifies the derivative of the function instead of its value. For the example linear BVP, we could use

$$u'(0) = \sigma, \qquad u(1) = \beta,$$

with one Neumann condition at $x = 0$.

We could also use two Neumann conditions,

$$u'(0) = \sigma, \qquad u'(1) = \eta,$$

although this would be ill-posed by itself, since if $u(x)$ is a solution then $u(x) + C$ for a constant $C$ is also a solution. We would need an additional constraint to obtain a unique solution.

# Neumann boundary conditions (approach 1)

We now need to determine $U_0$ as one of the unknowns, and we need an equation for it. One approach is to use a first-order discretization, so that

$$\frac{U_1 - U_0}{h} = \sigma.$$

For convenience, and symmetry, we could also build in the equation

$$U_{m+1} = \beta$$

into our linear system, to set the Dirichlet condition as well.

# Neumann boundary conditions (approach 1)

This results in the following linear system

$$\underbrace{\frac{1}{h^2}\begin{pmatrix} -h & h & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & 1 & -2 & 1 \\ & & & & & 0 & h^2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{pmatrix}}_{U} = \underbrace{\begin{pmatrix} \sigma \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ \beta \end{pmatrix}}_{F}$$

Solving this system only results in a first-order accurate solution.

# Neumann boundary conditions (approach 2)

An alternative approach to handle the boundary condition is to use the centered difference approximation

$$\frac{U_1 - U_{-1}}{2h} = \sigma.$$

This makes use of the solution $U_{-1}$ at a ghost node, outside of the interval. We can obtain a second equation for $U_{-1}$ by discretizing $u'' = f$ at $x = 0$, so that

$$\frac{U_{-1} - 2U_0 + U_1}{h^2} = f(x_0).$$

Combining the two equations allows the ghost node term to be eliminated, so that

$$\frac{-U_0 + U_1}{h} = \sigma + \frac{hf(x_0)}{2}.$$

Will give second-order accuracy overall.

# Neumann boundary conditions (approach 3)

A third approach would be to use a second-order one-sided derivative

$$\frac{-3U_0 + 4U_1 - U_2}{2h} = \sigma.$$

This would result in the linear system

$$\underbrace{\frac{1}{h^2}\begin{pmatrix} -3h/2 & 2h & -h/2 & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & 1 & -2 & 1 \\ & & & & & 0 & h^2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} U_0 \\ U_1 \\ U_2 \\ \vdots \\ U_m \\ U_{m+1} \end{pmatrix}}_{U} = \underbrace{\begin{pmatrix} \sigma \\ f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \\ \beta \end{pmatrix}}_{F}$$

A possible disadvantage with this approach is that the matrix is no longer tridiagonal.[3]

[3] Hence tridiagonal matrix solvers (*e.g.* the Thomas algorithm) could no longer be used.

# Pendulum: nonlinear boundary value problems

Nonlinear BVPs can use many of the same approaches, but complications arise. As an example, consider a pendulum of length $L$ with time-dependent angle $\theta(t)$ from the vertical. It will follow the equation

$$\theta''(t) = -\frac{g}{L}\sin\theta(t)$$

where $g$ is the gravitational acceleration. Rescaling time so that $g/L = 1$ gives

$$\theta''(t) = -\sin\theta(t).$$

For small angles $\sin\theta \approx \theta$, so this can be approximated by simple harmonic motion

$$\theta''(t) = -\theta(t),$$

which has general solution

$$\theta(t) = A\cos t + B\sin t.$$

# Pendulum: large angles

Suppose now that the pendulum has large oscillations, so that the linear approximation $\sin\theta \approx \theta$ does not hold. We could also search for solutions where the pendulum reaches two specific angles at two specific times $t = 0$ and $t = T$.

Then we have a two point nonlinear BVP

$$\theta''(t) = -\sin\theta(t),$$

with

$$\theta(0) = \alpha, \qquad \theta(T) = \beta.$$

# Pendulum: nonlinear discretization

This problem can be discretized using the same methods as before, writing $h = T/(m+1)$, defining $\theta_0 = \alpha$, $\theta_{m+1} = \beta$, and

$$\frac{\theta_{i-1} - 2\theta_i + \theta_{i+1}}{h^2} + \sin\theta_i = 0$$

for $i = 1, \ldots, m$.

This forms a nonlinear system with $m$ equations for $m$ unknowns. It can be written as

$$G(\theta) = 0$$

where $G : \mathbb{R}^m \to \mathbb{R}^m$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$.

We can no longer solve the system with linear algebra alone, and we need a new approach.

# Nonlinear systems of equations

Nonlinear systems of equations are more complicated to solve than linear systems. There may be zero, one, or any number of possible solutions, and it is difficult to know *a priori* how many solutions there will be.

Whereas linear systems can be solved in a single step, nonlinear systems are often solved using iteration. Starting from an initial guess, the solution is iteratively improved until it converges on a solution.

We will look at the Newton's method, which finds a root to the system $G(\theta) = 0$ by using a sequence of linear approximations of $G$.

# Newton's method in one dimension

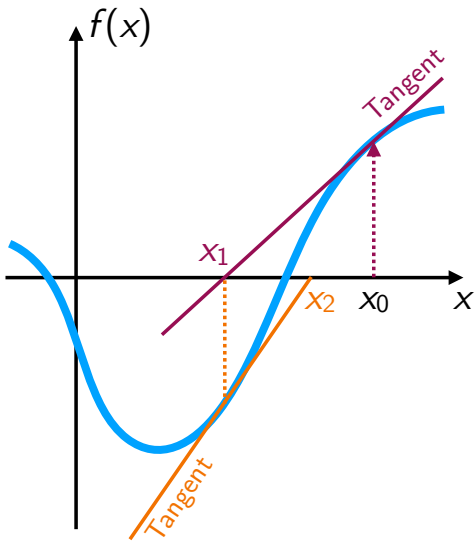
To begin, first consider a scalar function $f : \mathbb{R} \to \mathbb{R}$, and let $x_0$ be an approximation for its root.

Draw the tangent line from $(x_0, f(x_0))$. The place where it crosses the $x$ axis is

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Assuming that the function is nearly linear near the root, this will give a better approximation.

Apply iteratively to obtain a sequence $x_0, x_1, x_2, \ldots$ that converges to root (under certain conditions).



# Newton's method

Alternatively, Newton's method can be thought of as making a sequence of linear Taylor series approximations of the function. Write $x = x_k + \Delta x_k$.

The linear Taylor series is

$$f_{\text{lin}}(x) = f(x_k) + \Delta x_k \, f'(x_k).$$

Setting $f_{\text{lin}}(x) = 0$ gives

$$\Delta x_k = -\frac{f(x_k)}{f'(x_k)}$$

and then the next iterate is

$$x_{k+1} = x_k + \Delta x_k.$$

# Multidimensional Newton's method

This directly extends to finding the root of a function $f(x)$ where $f : \mathbb{R}^m \to \mathbb{R}^m$. At a given iterate $x_k \in \mathbb{R}^m$, write $x = x_k + \Delta x_k$. Then the linear Taylor series is

$$f_{\text{lin}}(x) = f(x_k) + J_f(x_k)\Delta x_k$$

where $J_f(x_k) \in \mathbb{R}^{m \times m}$ is the Jacobian of $f$ evaluated at $x_k$. Setting $f_{\text{lin}}(x) = 0$ gives

$$J_f(x_k)\Delta x_k = -f(x_k),$$

which is a linear system that can be solved for $\Delta x_k$. Then

$$x_{k+1} = x_k + \Delta x_k$$

as before.

# Returning to pendulum problem

The nonlinear equations describing the pendulum problem are

$$\frac{\theta_{i-1} - 2\theta_i + \theta_{i+1}}{h^2} + \sin\theta_i = 0 \tag{1}$$

for $i = 1, \ldots, m$, with the boundary conditions that $\theta_0 = \theta_{m+1} = 0$. We therefore perform nonlinear root-finding on

$$G(\theta) = \begin{pmatrix} h^{-2}(-2\theta_1 + \theta_2) + \sin\theta_1 \\ h^{-2}(\theta_1 - 2\theta_2 + \theta_3) + \sin\theta_2 \\ h^{-2}(\theta_2 - 2\theta_3 + \theta_4) + \sin\theta_3 \\ \vdots \\ h^{-2}(\theta_{m-2} - 2\theta_{m-1} + \theta_m) + \sin\theta_{m-1} \\ h^{-2}(\theta_{m-1} - 2\theta_m) + \sin\theta_m \end{pmatrix}. \tag{2}$$

# Jacobian for the pendulum problem

Hence the Jacobian has components

$$J_{ij}(\theta) = \begin{cases} h^{-2} & \text{if } |i-j| = 1, \\ -2h^{-2} + \cos\theta_i & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases}$$

and can be written as

$$J(\theta) = \frac{1}{h^2} \begin{pmatrix} -2 + h^2\cos\theta_1 & 1 & & & \\ 1 & -2 + h^2\cos\theta_2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 + h^2\cos\theta_{m-1} & 1 \\ & & & 1 & -2 + h^2\cos\theta_m \end{pmatrix}.$$

See Homework 1 Question 3 for a similar example.