Numerical Analysis Comprehensive Exam Questions

1. Let $f(x) = (x-\alpha)^m g(x)$ where $m \geq 2$ is an integer and $g(x) \in \mathcal{C}^2(\mathrm{R}), g(\alpha) \neq 0$. Write down the Newton's method for finding the root α of $f(x)$, and study the order of convergence for the method. Can you propose a method that converges faster?

Solution: Newton's method is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} := h(x_n)$$

For the fixed-point iteration $x_{n+1} = h(x_n)$, we can investigate $h'(\alpha)$.

If $0 < |h'(\alpha)| < 1$, the iteration converges linearly with the rate of conv. tending to $|h'(\alpha)|$; if $h'(\alpha) = 0$, then if converges at least quadratically.

In fact,

$$h'(x) = \frac{f(x)f''(x)}{[f'(x)]^2} \to \frac{m-1}{m}, \quad \text{as} \quad x \to \alpha.$$

$$\therefore 0 < \frac{m-1}{m} < 1 \quad \text{if} \quad m \geq 2.$$

The Newton's method only converges linearly if x_o is chosen close enough to α.

To improve the order of convergence, once could use

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)} := H(x_n).$$

It's easy to show that $H'(x) \to 0$ as $x \to \alpha$. Therefore, the new method converges at least quadratically.

2. Let $x_0, x_1, \ldots, x_n$ be distinct real numbers and $l_k(x)$ be the Lagrange's basis function. Let $\Psi_n(x) = \Pi_{k=0}^n (x - x_k)$. Prove that

(1) For any polynomial $p(x)$ of degree $n+1$,

$$p(x) - \sum_{k=0}^{n} p(x_k) l_k(x) = \frac{1}{(n+1)!} p^{(n+1)}(x) \Psi_n(x).$$

(2) If further $x_0, \ldots, x_n$ are Gauss-Legendre points in the interval $[-1, 1]$, then

$$\int_{-1}^{1} l_i(x) l_j(x) dx = 0 \quad \text{for} \quad i \neq j.$$

Solution: (1) Note that $p(x) - \sum_{k=0}^{n} p(x_k) l_k(x)$ has zeros at $x = x_0, \dots, x_n$, therefore, $p(x) - \sum_{k=0}^{n} p(x_k) l_k(x) = C\Psi_n(x)$ for some constant C.

Since $\sum_{k=0}^{n} p(x_k) l_k(x)$ is of degree n and $p(x)$ is of degree $n+1$, C must be the highest degree coeff. of $p(x)$, thus $C = \frac{1}{(n+1)!} p^{(n+1)}(x)$. □

(2) Since $x_0, \dots, x_n$ are Gauss-Legendre points,

$$\int_{-1}^{1} p(x)dx = \sum_{k=0}^{n} w_k p(x_k)$$

for any polynomial p of degree $\leq 2n+1$. Therefore,

$$\int_{-1}^{1} l_i(x) l_j(x) dx = \sum_{k=0}^{n} w_k l_i(x_k) l_j(x_k) = 0, \quad \text{if} \quad i \neq j. \quad \square$$

3. For the initial value problem $y'(t) = f(t, y)$, $y(0) = y_0$, $t \geq 0$, consider the θ-method

$$y_{n+1} = y_n + h[\theta f(t_n, y_n) + (1-\theta) f(t_{n+1}, y_{n+1})],$$

where time has been discretized such that $t_n = nh$, and y_n is the numerical approximation of $y(t_n)$.

(a) Is this method consistent? What is the order? Is this method zero-stable? How does the result differ for different θ?

(b) What is the region of absolute stability? What is the region when $\theta = 0, \frac{1}{2}$, or 1? For what θ values is this method $A-$stable?

(c) Does this method have stiff decay? Show why or why not.

Solution: (a) Consider the local truncation error around t_n,

$$\begin{aligned} y + hy' + \tfrac{h^2}{2!}y'' + \cdots - y - h(\theta y' + (1-\theta)(y' + hy'' + \tfrac{h^2}{2!}y''' + \cdots) \\ = h^2 y''(\tfrac{1}{2} - (1-\theta)) + h^3(-\tfrac{1}{3} + \tfrac{1}{2}\theta) y''' + \cdots \end{aligned}$$

The local truncation error is at least 2nd order, i.e. globally at least first order method, so this method is consistent. The order of the method becomes 2nd order if $\theta = \frac{1}{2}$, otherwise, it is a first order method. The first characteristic polynomial is $\rho(z) = z - 1$, i.e. $z = 1$, so this method is zero-stable for any θ. □

(b) Consider $y' = \lambda y$, then $y_{n+1} = y_n + h[\theta\lambda y_n + (1-\theta)\lambda y_{n+1}]$,

$$y_{n+1} = \frac{1 + h\theta\lambda}{1 - h(1-\theta)\lambda} y_n, \quad |R(h\theta)| := \left| \frac{1 + h\theta\lambda}{1 - h(1-\theta)\lambda} \right| \leq 1 \text{ gives the stability region.}$$

When $\theta = 0$, the region is outside of the unit circle centered at $z = 1$ in the complex plane, when $\theta = \frac{1}{2}$, the negative half plane of the complex plane, and when $\theta = 1$, inside of the unit circle centered at $z = -1$ in the complex plane. Therefore, this method is A-stable for $0 \le \theta \le \frac{1}{2}$, since the stability region includes the negative half plane. □

(c) To have stiff decay, one must show $\lim_{h\theta \to -\infty} R(h\theta) \to 0$.

$$\lim_{h\theta \to -\infty} R(h\theta) = \lim_{h\theta \to -\infty} \frac{1 + h\theta\lambda}{1 - h(1-\theta)\lambda} = \frac{\theta}{1-\theta}$$

i.e. goes to 0 only when $\theta = 0$. Only when $\theta = 0$ (which is the backward Euler method), this method has stiff decay. □

4. Consider the initial boundary values problem

$$\begin{cases} u_t = u_{xx}, & (x,t) \in (0,1) \times (0,T] = D \\ u(0,t) = g_1(t), u(1,t) = g_2(t), & t \in [0,T] \\ u(x,0) = f(x), & x \in [0,1] \end{cases}$$

Let (x_i, t_n) be a grid point in a uniform rectangular grid, s.t. $x_i = i\Delta x, t_n = n\Delta t$ for $i = 0, 1, \ldots, I$ and $n = 0, 1, \ldots, N$ where $I\Delta x = 1$ and $N\Delta t = T$, and let U_i^n be a numerical approximation of $u(x_i, t_n)$. Assuming the exact solution is sufficiently smooth, show that the scheme

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} = \frac{U_{i+1}^{n+1} - 2U_i^{n+1} + U_{i-1}^{n+1}}{\Delta x^2}$$

is unconditionally stable and $|U_i^n - u(x_i, t_n)| = \mathcal{O}(\Delta t + \Delta x^2)$ as $\Delta t, \Delta x \to 0$.

Solution: Given some perturbation of $f(x)$, $g_1(t)$, and $g_2(t)$, the difference between the perturbed Solution and U_i^n, say ρ_i^n will also satisfy

$$\frac{\rho_i^{n+1} - \rho_i^n}{\Delta t} = \frac{\rho_{i+1}^{n+1} - 2\rho_i^{n+1} + \rho_{i-1}^{n+1}}{\Delta x^2}.$$

Let $\lambda = \frac{\Delta t}{\Delta x^2}$, then $\rho_i^{n+1} = \frac{\lambda}{1+2\lambda}\rho_{i+1}^{n+1} + \frac{1}{1+2\lambda}\rho_i^n + \frac{\lambda}{1+2\lambda}\rho_{i-1}^{n+1}$. Therefore, ρ_i^{n+1} is a strict convex combination of ρ_{i+1}^{n+1}, ρ_i^n and ρ_{i-1}^{n+1} and satisfy the maximum principle, which proves the stability.

Note that the exact Solution u satisfies

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{\Delta x^2} + \mathcal{O}(\Delta t + \Delta x^2).$$

Let $e_i^n = u_i^n - U_i^n$, then $e_i^{n+1} = \frac{\lambda}{1+2\lambda}e_{i+1}^{n+1} + \frac{1}{1+2\lambda}e_i^n + \frac{\lambda}{1+2\lambda}e_{i-1}^{n+1} + \mathcal{O}(\Delta t^2 + \Delta t \Delta x^2)$.

$$\begin{gathered}\therefore |e_i^{n+1}| \le \tfrac{\lambda}{1+2\lambda}\|e^{n+1}\|_\infty + \tfrac{1}{1+2\lambda}\|e^n\|_\infty + \tfrac{\lambda}{1+2\lambda}\|e^{n+1}\|_\infty + \mathcal{O}(\Delta t^2 + \Delta t\Delta x^2)\\ \therefore \|e^{n+1}\|_\infty \le \tfrac{\lambda}{1+2\lambda}\|e^{n+1}\|_\infty + \tfrac{1}{1+2\lambda}\|e^n\|_\infty + \tfrac{\lambda}{1+2\lambda}\|e^{n+1}\|_\infty + \mathcal{O}(\Delta t^2 + \Delta t\Delta x^2)\\ \therefore \tfrac{\lambda}{1+2\lambda}\|e^{n+1}\|_\infty \le \tfrac{1}{1+2\lambda}\|e^n\|_\infty + C(\Delta t^2 + \Delta t\Delta x^2)\end{gathered}$$

This implies

$$\|e^n\|_\infty \le \|e^0\|_\infty + C(n\Delta t^2 + n\Delta t\Delta x^2) \le C(\Delta t + \Delta x^2)$$

since $\|e^0\|_\infty \to 0$. $\square$

5. Consider the boundary-value problem

$$\begin{cases} -u_{xx} + u = f(x), & x \in (0,1)\\ u(0) = a \\ u_x(1) = b \end{cases}$$

Given a partition $0 = x_0 < x_1 < \cdots < x_n = 1$, please formulate a piecewise linear continuous finite element method to solve the problem. Show that your method has a unique solution.

Solution:

Let $\phi_i(x)$, $i = 0, \ldots, n$ be continuous piecewise linear functions defined on $[0,1]$ s.t. $\phi_i(x_j) = \gamma_{ij}$ for $j = 0, \ldots, n$. Let

$$V_h = \text{span}\{\phi_i(x) : i = 1, 2, \ldots, n\} \text{ and } W_h = a\phi_0 + V_h.$$

For any $\phi \in V_h$,

$$\int_0^1 (-u_{xx} + u)\phi dx = -(u_x\phi)|_0^1 + \int_0^1 u_x\phi_x dx + \int_0^1 u\phi dx = \int_0^1 f\phi dx$$

here $-(u_x\phi)|_0^1 = -u_x(1)\phi(1) = -b\phi(1)$.

The FEM is to find $U \in W_h$ s.t.

$$\int_0^1 U_x\phi_x dx + \int_0^1 U\phi dx = \int_0^1 f\phi dx + b\phi(1),$$

for any $\phi \in V_h$.

This system has a unique solution if and only if the associated homogeneous system

$$\int_0^1 U_x\phi_x dx + \int_0^1 U\phi dx = 0$$

has a unique 0 Solution, assuming $a, b = 0$.

In fact, let $\phi = U$,then

$$\int_0^1 |U_x|^2 dx + \int_0^1 |U|^2 dx = 0 \quad \Longrightarrow \quad U \equiv 0. \quad \Box$$

6. Consider the following matrix A and solving the linear system $A\vec{x} = \vec{b}$ by iterative methods,

$$A = \begin{pmatrix} 1 & \alpha & \beta \\ -\alpha & 1 & -\gamma \\ \beta & \gamma & 1 \end{pmatrix}.$$

(a) What are the conditions on the variables α, β, and γ for Jacobi's method and Gauss-Seidel method to converge?

(b) Describe the Jacobi's method and Gauss-Seidel method.

(c) Find a set of values (if any exist) of α, β, and γ for which the Jacobi method converges but Gauss-Seidel does not, and vice versa.

Solution: (a) The matrix A should be strictly diagonally dominant:

$$|\alpha| + |\beta| < 1, |\alpha| + |\gamma| < 1, \text{ and } |\beta| + |\gamma| < 1. \quad \Box$$

(b) $A = D - L - U$, where D is the diagonal matrix, L lower triangular matrix and U upper triangular matrix.

$$\begin{gathered} (D - L - U)\vec{x} = \vec{b} \\ D\vec{x} = (L + U)\vec{x} + \vec{b} \\ \vec{x}_{n+1} = D^{-1}(L + U)\vec{x}_n + D^{-1}\vec{b} \end{gathered}$$

This is Jacobi iteration, and for Gauss-Seidel,

$$\begin{gathered} (D - L)\vec{x} = U\vec{x} + \vec{b} \\ \vec{x}_{n+1} = (D - L)^{-1}U\vec{x}_n + (D - L)^{-1}\vec{b} \end{gathered}$$

Algorithm: With any initial condition $\vec{x}_0$, iterate $\vec{x}_{n+1} = D^{-1}(L + U)\vec{x}_n + D^{-1}\vec{b}$ (Jacobi) or $\vec{x}_{n+1} = (D - L)^{-1}U\vec{x}_n + (D - L)^{-1}\vec{b}$ (Gauss-Seidel) until it converges. □

(c) (Outline of solution) The standard convergence condition is when the spectral radius of the iteration matrix is less than 1. Compute the eigenvalues of $D^{-1}(L+U)$, let's call them λ_Js, and the eigenvalues of $(D - L)^{-1}U$ to be λ_Gs, then find the condition on α, β, and γ which corresponds to $|\lambda_J| < 1$ and $|\lambda_G| > 1$ and vice versa. □

7. Describe an algorithm to compute least squares solution by singular value decomposition. Prove the solution obtained is the least squares solution and estimate the leading computation cost for your algorithm. Assume the least squares system is

$$A\vec{x} = \vec{b}, A \in \mathrm{R}^{m\times n}, \vec{b} \in \mathrm{R}^m, \text{ and } \vec{x} \in \mathrm{R}^n.$$

Solution: Algorithm:

(a) Compute the reduced SVD: $A = U\Sigma V^*$.

(b) Compute the vector $\vec{y} = U^*\vec{b}$.

(c) Solve the diagonal system $\Sigma\vec{w} = \vec{y}$.

(d) Set $\vec{x} = V\vec{w}$.

Proof: Least squares solution

$$\begin{aligned} & A^*A\vec{x} = A^*\vec{b} \\ \Leftrightarrow\ & V\Sigma^*U^*U\Sigma V^*\vec{x} = V\Sigma^*U^*\vec{b} \\ \Leftrightarrow\ & \Sigma^*\Sigma V^*\vec{x} = \Sigma^*U^*\vec{b} \\ \Leftrightarrow\ & \Sigma V^*\vec{x} = \vec{y} \\ \Leftrightarrow\ & \Sigma\vec{w} = \vec{y}. \end{aligned}$$

Leading cost: dominated by SVD $\sim 2mn^2 + 11n^3$ flops. □

8. Describe the Conjugate Gradient Iteration method for solving a linear system

$$A\vec{x} = \vec{b}.$$

Prove that the residuals are orthogonal.

Solution: Algorithm:

(a) $\vec{x_o} = 0, \vec{r_o} = \vec{b}$, $\vec{p_o} = \vec{r_o}$

(b) for $n = 1, 2, 3, \ldots$

   i. $\alpha_n = (\vec{r}_{n-1}^T\vec{r}_{n-1})/(\vec{p}_{n-1}^T A\vec{p}_{n-1})$

   ii. $\vec{x}_n = \vec{x}_{n-1} + \alpha_n\vec{p}_{n-1}$.

   iii. $\vec{r}_n = \vec{r}_{n-1} - \alpha_n A\vec{p}_{n-1}$.

```
    iv. $\beta_n = (\vec{r}_n^T\vec{r}_n)/(\vec{r}_{n-1}^T\vec{r}_{n-1})$
    v. $\vec{p}_n = \vec{r}_n + \beta_n\vec{p}_{n-1}$.
  end.
```

Proof for $\vec{r}_n^T\vec{r}_j = 0$ for $j < n$.

By induction on n,

$$\vec{r}_n^T\vec{r}_j = \vec{r}_{n-1}^T\vec{r}_j - \alpha_n\vec{p}_{n-1}^T A\vec{r}_j.$$

If $j < n - 1$, it is true by induction,

If $j = n - 1$,

$$\vec{r}_n^T\vec{r}_{n-1} = 0 \quad\Leftrightarrow\quad \alpha_n = (\vec{r}_n^T\vec{r}_{n-1})/\vec{p}_n^T A\vec{r}_{n-1}.$$

□