AMSC 667 Section 0101 Spring 2011

NUMERICAL ANALYSIS II

HOMEWORK #3 (Pbs 1-4 due Mar 31, Pbs 5-7 due Apr 7)

1 (15 pts). *Finite differences*: Given a function f with three continuous derivatives in the interval $[x_0, x_2]$, let q be a quadratic polynomial that interpolates f at the nodes $x_0 < x_1 < x_2$, with variable spacing $h_1 = x_1 - x_0$ and $h_2 = x_2 - x_1$. Let $h = \max(h_1, h_2)$.
(a) Find an expression for $q(x)$ in terms of $\{x_i\}_{i=0}^2$ and $\{f(x_i)\}_{i=0}^2$. Using Taylor expansion, show the error estimates

$$|f'(x_1) - q'(x_1)| \le Ch^2 \max_{x_0 \le x \le x_2} |f^{(3)}(x)|,$$
$$|f''(x_1) - q''(x_1)| \le Ch \max_{x_0 \le x \le x_2} |f^{(3)}(x)|.$$

(b) Consider the case of equally spaced nodes, i.e. $h_1 = h_2 = h$. Derive the *centered* difference formulas

$$q'(x_1) = \frac{1}{2h}\big(f(x_2) - f(x_0)\big), \qquad q''(x_1) = \frac{1}{h^2}\big(f(x_0) - 2f(x_1) + f(x_2)\big).$$

(c) Use Taylor expansion again to derive the improved error bound

$$|f''(x_1) - q''(x_1)| \le Ch^2 \max_{x_0 \le x \le x_2} |f^{(4)}(x)|,$$

provided f has four continuous derivatives in $[x_0, x_2]$. Explain what would happen if f has only three continuous derivatives. Could you also get an improved error of order $O(h^3)$ for the first derivative?

2 (10 pts) *Strict diagonal dominance and M-matrices*: Let A be an $n \times n$ matrix such that

$$a_{ii} > 0, \qquad a_{ij} \le 0 \quad i \ne j, \qquad \sum_{j=1}^n a_{ij} > 0 \quad 1 \le i \le n.$$

(a) Show that $A\mathbf{x} = \mathbf{b} \ge 0$ implies $\mathbf{x} \ge 0$. To this end, assume by contradiction that $x_j = \min_{1 \le i \le n} x_i < 0$, and write the j-th equation as

$$\sum_{k \ne i} a_{jk}(x_k - x_j) + x_j \sum_{k=1}^n a_{jk} = b_j \ge 0. \tag{1}$$

(b) Show that if A satisfies (a), then A is nonsingular and $A^{-1} \ge 0$.

3 (10 pts) *Irreducibility and M-matrices*: Let A be an irreducible $n \times n$ matrix such that

$$a_{ii} > 0, \qquad a_{ij} \le 0 \quad i \ne j, \qquad \sum_{j=1}^n a_{ij} \ge 0 \quad 1 \le i \le n, \tag{2}$$

and for some $1 \le i_0 \le n$

$$\sum_{j=1}^n a_{i_0 j} > 0. \tag{3}$$

Show that $A\mathbf{x} = \mathbf{b} \ge 0$ implies $\mathbf{x} \ge 0$. To this end proceed by contradiction as follows:

(a) Let $J \ne \emptyset$ be the set of indices such that x_j satisfies the condition in Pb2(a). Conclude that $\sum_{k=1}^n a_{jk} = 0$ for all $j \in J$, and that the complement I of J is nonempty (in fact $i_0 \in I$).

(b) Deduce that $a_{ji} = 0$ for all $i \in I$ and $j \in J$, whence A is reducible.

4 (15 pts) *Advection-diffusion equation*: Let u be the solution of the following problem on $(0,1)$:

$$-\epsilon u'' + u' = 0 \qquad u(0) = 1, u(1) = 0. \tag{4}$$

Let $\mathcal{T}_h = \{x_i\}_{i=0}^{N+1}$ be a uniform partition of $(0,1)$ of size h. Let $\mathbf{U} = (U_i)_{i=0}^{N+1}$ be the discrete solution using centered differences.

(a) Write the i-th equation satisfied by $\mathbf{U}$ in the form

$$\beta_i U_{i-1} + \alpha_i U_i + \gamma_i U_{i+1} = 0 \quad 1 \le i \le N, \tag{5}$$

i.e., find $\alpha_i, \beta_i, \gamma_i \in R$ in terms of h and ϵ.

(b) Think of (5) as a difference equation and show that

$$U_i = 1, \qquad U_i = \Big(\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}\Big)^i, \qquad 0 \le i \le N+1,$$

are two linearly independent solutions of (5). Find the solution of (5) in terms of these two solutions.

(c) Find the exact solution of (4) and note that it does not oscillate. Determine the relation between ϵ and h that ensures there are no oscillations in $\mathbf{U}$. Hint: Consider the sign of $\frac{\frac{2\epsilon}{h}+1}{\frac{2\epsilon}{h}-1}$. Would you use this method to approximate (4) when $\epsilon << 1$ (the advection dominated case)?

(d) Replace the centered difference approximation of the first order term $u'(x_i)$ by the *up-wind* difference $u'(x_i) \approx \big(u(x_i) - u(x_{i-1})\big)/h$. Repeat (a) and (b) and show that the resulting matrix A satisfies (2) and (3). Draw conclusions.

5 (15 pts) *Discrete supersolutions and error estimates*: Let $\alpha, \beta \in \mathbb{R}$ and u be the solution of the boundary value problem

$$Lu = -u'' + u = f(x) \quad x \in (0,1), \qquad u(0) = \alpha, \quad u'(1) = \beta. \tag{6}$$

(a) Write a *finite difference* method on a uniform partition $\mathcal{T} = \{x_i\}_{i=0}^{N+1}$ with $0 = x_0 < x_1 < \cdots < x_N = 1$ and meshsize h. Write 2 discretizations of the Neumann condition $u'(1) = \beta$ that amount to truncation errors of order $O(h)$ and $O(h^2)$. Let $\mathbf{U} = (U_i)_{i=1}^N$ denote the discrete solution.

(b) Write out the resulting $N \times N$ matrix A and right-hand side $\mathbf{F} \in \mathbb{R}^N$. Show that A is nonsingular and $A^{-1} \ge 0$ (use Pb2). Show that A can be symmetrized by suitable rescaling.

(c) Derive a bound for the truncation error at interior nodes (use Pb1).

(d) Let w be the solution of (6) with $f(x) = 1, \alpha = 0$ and $\beta = 1$; w is a supersolution of the homogeneous problem. Let $\mathbf{W} = (W_i)_{i=1}^N$ with $W_i = w(x_i)$ be the corresponding discrete counterpart. Show that there exists a constant $h_0 > 0$ such that $(A\mathbf{W})_i \ge \frac{h^\gamma}{2}$, where $\gamma = 2$ for $i < N$ and $\gamma = 1$ or $\gamma = 2$ for $i = N$ depending on the discretization of the Neumann condition.

(e) Use $\mathbf{W}$ as a barrier to show the error estimate

$$\max_{1 \le i \le N} |u(x_i) - U_i| \le C h^\gamma,$$

where $C > 0$ depends on u and $\gamma = 1$ or 2 depending on how the Neumann condition is approximated.

6 (15 pts) *Finite element method*: (a) Convert (6) into a problem with homogeneous Dirichlet boundary condition at $x = 0$ upon substracting a suitable linear function $g(x)$; in this case g may just be a constant. Write a variational (or weak) formulation of (6) in terms of the new unknown $z = u - g$ of the form

$$z \in V: \quad B(z, v) = F(v) \quad \forall\, v \in V.$$

Determine the functional space V, the bilinear form B and the linear functional F. Show that B is continuous and coercive in V.
(b) Write the piecewise linear finite element discretization over the uniform partition $\mathcal{T}$ of Pb.5. Find the resulting (stiffness) matrix $\mathbf{A}$ and forcing term $\mathbf{F}$, and compare with Pb.5(b).
(c) Show the existence of a constant $C > 0$ independent of $\mathcal{T}$ and z such that (*best approximation*)

$$\|z - z_h\|_V \leq C \inf_{v \in V_h} \|z - v\|_V,$$

and deduce a similar one for u. Obtain an optimal error estimate for $\|u - u_h\|_V$.
(d) *Quadrature*: Assume that the integrals involving p are computed using the mid-point rule in each element, and those involving q and f using instead the trapezoidal rule. Determine the resulting matrix A and forcing $\mathbf{F}$, and compare with Pb.5(b). The finite element method with quadrature leads to finite difference formulas.

7 (20 pts) *MATLAB*: Consider the nonhomogeneous Dirichlet boundary value problem in $I = (1, 2)$

$$Lu = -(xu')' + \frac{1}{x}u = 0, \qquad u(1) = 1, \quad u(2) = \frac{1}{2}. \tag{7}$$

(a) Find the exact solution. Hint: try a function of the form x^r and find r.

(b) Write a MATLAB program that solves (7) with centered finite differences and uniform meshsizes $h = 0.1 \times 2^{-k}$ for $k = 0, 1, 2, 3$ (use the command `diag` to construct the stiffness matrix). Compute the error at the nodes, verify its dependence on h and relate to theory. Plot the error functions.

(c) Write a MATLAB program that solves (7) with piecewise linear finite elements and $h = 0.1 \times 2^{-k}$ for $k = 0, 1, 2, 3$. Organize the program so that the integrals for the stiffness matrix and right-hand side are computed elementwise. Compare with (b). Hint: First convert (7) into a problem with homogeneous Dirichlet condition by setting

$$z(x) = u(x) - \left(\frac{3}{2} - \frac{1}{2}x\right).$$

Check that z satisfies

$$Lz = -\frac{3}{2}x, \qquad z(1) = z(2) = 0.$$

Write the weak form in terms of z, and solve for z. Finally add $\frac{3}{2} - \frac{1}{2}x$ back on to the discrete solution. Compute the integrals in each element using the two-point Gauss quadrature rule.

(d) Reformulate the finite element problem to solve directly for u without introducing z. Explain. This approach is particularly useful in several dimensions where finding z is problematic.