Preliminary Examination, Numerical Analysis, January 2010

Instructions: This exam is closed books and notes. The time allowed is three hours and you need to work on any three out of questions 1-5 and any two out of questions 6-8. All questions have equal weights and the passing score will be determined after all the exams are graded. Indicate clearly the work that you wish to be graded.

Note: In problems 6-8, the notations $k = \Delta t$ and $h = \Delta x$ are used. Note also that at the end of the exam there is a list of Facts some of which may be useful to you.

1. Matrix Factorizations:
(a) Prove any two of the following statements:

(i) Schur Decomposition: Any matrix $A \in \mathbb{C}^{m\times m}$ can be factored as $A = Q^*TQ$, where Q is unitary and T is upper triangular.

(ii) Singular Value Decomposition: Any matrix $A \in \mathbb{C}^{m\times n}$ can be factored as $A = U\Sigma V^*$, where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary and $\Sigma \in \mathbb{R}^{m\times n}$ is a rectangular matrix whose only nonzero entries are non-negative entries on its diagonal.

(iii) QR Factorization: Any full-rank matrix $A \in \mathbb{R}^{m\times n}$ for $m \geq n$ can be factored $A = QR$ where $Q \in \mathbb{R}^{m\times m}$ is orthogonal and $R \in \mathbb{R}^{m\times n}$ is upper triangular with positive diagonal entries.

b) Discuss situations in which each of these factorizations is useful in numerical analysis and explain why the factorizations are useful in those situations.

2) Least Squares Problems

a) For a full rank real $m \times n$ matrix A, show that $X = A^{\dagger}$, the pseudoinverse of A, minimizes $\|AX - I\|_F$ over all $n \times m$ matrices X. What is the value of the minimum? (Hint: Relate the problem to a set of least-squares problems).

b) For a real full rank $m \times n$ matrix A and vector $\mathbf{b} \in \mathbb{R}^m$, explain how to solve the least-squares problem of finding $\mathbf{x} \in \mathbb{R}^n$ that minimizes $\|A\mathbf{x} - \mathbf{b}\|_2$ using i) the normal equations, and b) a QR factorization of the matrix A. What are the advantages and disadvantages of each of these methods?

3) Iterative Methods for Linear Systems

Consider the boundary value problem

$$-u''(x) + \beta u(x) = f(x), \quad \text{for } 0 \leq x \leq 1$$

where $\beta > 0$, and with $u(0) = u(1) = 0$, and the following discretization of it:

$$-U_{j-1} + (2 + \beta h^2)\, U_j - U_{j+1} = F_j$$

for $j = 1, 2, \ldots, N-1$ where $Nh = 1$, $F_j \equiv h^2\, f(jh)$, and $U_0 = U_N = 0$.

Analyze the convergence properties of the Jacobi iterative method for this problem. In particular, express the speed of convergence as a function of the discretization stepsize h. How does the number of iterations required to reduce the initial error by a factor δ depend on h? In practice, would you use this method to solve the given problem? If so, explain why this is a good idea? If not, how would you solve it in practice?

4) Interpolation and Integration:

a) Consider equally spaced points $x_j = a + jh$, $j = 0, \ldots, n$ on the interval $[a, b]$, where $nh = b - a$. Let $f(x)$ be a smooth function defined on $[a, b]$. Show that there is a unique polynomial $p(x)$ of degree $n + 1$ which interpolates f at all of the points x_j. Derive the formula for the interpolation error at an arbitrary point x in the interval $[a, b]$:

$$f(x) - p(x) \equiv E(x) = \frac{1}{(n+1)!}(x - x_0)(x - x_1) \cdots (x - x_n) f^{n+1}(\eta).$$

for some $\eta \in [a, b]$.

b) Let $I_n(f)$ denote the result of using the composite Trapezoidal rule to approximate $I(f) \equiv \int_a^b f(x)dx$ using n equally sized subintervals of length $h = (b - a)/n$. It can be shown that the integration error $E_n(f) \equiv I(f) - I_n(f)$ satisfies

$$E_n(f) = d_2 h^2 + d_4 h^4 + d_6 h^6 + \ldots$$

where $d_2, d_4, d_6, \ldots$ are numbers that depend only on the values of f and its derivatives at a and b. Suppose you have a black-box program that, given f, a, b, and n, calculates $I_n(f)$. Show how to use this program to obtain an $O(h^4)$ approximation and an $O(h^6)$ approximation to $I(f)$.

5) Sensitivity:
Consider a 6×6 symmetric positive definite matrix A with singular values $\sigma_1 = 1000$, $\sigma_2 = 500$, $\sigma_3 = 300$, $\sigma_4 = 20$, $\sigma_5 = 1$, $\sigma_6 = 0.01$.
a) Suppose you use a Cholesky factorization package on a computer with a machine epsilon 10^{-14} to solve the system $Ax = b$ for some nonzero vector b. How many digits of accuracy do you expect in the computed solution? Justify your answer in terms of condition and stability. You may assume that the entries of A and b are exactly represented in the computer's floating-point system.

b) Suppose that instead you use an iterative method to find an approximate solution to $Ax = b$ and you stop iterating and accept iterate $x^{(k)}$ when the residual $r^{(k)} = Ax^{(k)} - b$ has 2-norm less than 10^{-9}. Give an estimate of the maximum size of the relative *error* in the final iterate? Justify your answer.

6) Elliptic Problems:

Consider the standard five-point difference approximation (centered difference for both the gradient and divergence operators) for the variable coefficient Poisson equation

$$-\nabla \cdot (a \nabla v) = f$$

with Dirichlet boundary conditions, in a two-dimensional rectangular region. We assume that $a(x, y) \geq a_0 > 0$. The approximate solution $\{u_{i,j}\}$ satisfies a linear system $Au = b$.

1. State and prove the maximum principle for the numerical solution $u_{i,j}$.

2. Derive the matrix A in the one-dimensional case and show that it is symmetric and positive definite.

3. For the one-dimensional and *constant-coefficient* case, show that the global error $e_j = v(x_j) - u_j$ satisfies $\|e\|_2 = O(h^2)$ as the space step $h \to 0$.

4. Discuss the advantages and disadvantages of trying to solve the system for the two-dimensional problem using (i) the SOR (Successive Over Relaxation) method and (ii) the (preconditioned) Conjugate Gradient method.

7) Heat Equation Stability:

a) Consider the initial value problem for the constant-coefficient diffusion equation

$$v_t = \beta v_{xx}, \ t > 0$$

with initial data $v(x,0) = f(x)$. A scheme for this problem is:

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{\beta}{h^2} \left\{ u_{j-1}^{n+1} - 2u_j^{n+1} + u_{j+1}^{n+1} \right\}.$$

Analyze the 2-norm stability of this scheme. For which values of $k > 0$ and $h > 0$ is the scheme stable? (Note that this there are no boundary conditions here.)

b) Consider the variable coefficient diffusion equation

$$v_t = (\beta v_x)_x, \qquad 0 < x < 1, \ t > 0$$

with Dirichlet boundary conditions

$$v(0,t) \ = \ 0, \qquad v(1,t) \ = \ 0$$

and initial data $v(x,0) = f(x)$. Assume that $\beta(x) \geq \beta_0 > 0$, and that $\beta(x)$ is smooth. Let $\beta_{j+1/2} = \beta(x_{j+1/2})$. A scheme for this problem is:

$$\frac{u_j^{n+1} - u_j^n}{k} = \frac{1}{h^2} \ \left\{ \beta_{j-1/2} u_{j-1}^{n+1} - (\beta_{j-1/2} + \beta_{j+1/2}) u_j^{n+1} + \beta_{j+1/2} u_{j+1}^{n+1} \right\}.$$

Analyze the 2-norm stability of this scheme for solving this initial boundary value problem. DO NOT NEGLECT THE FACT THAT THERE ARE BOUNDARY CONDITIONS!

8) Numerical Methods for ODEs: Consider the Linear Multistep Method

$$y_{n+2} - \frac{4}{3}y_{n+1} + \frac{1}{3}y_n = \frac{2}{3}kf_{n+2}$$

for solving an initial value problem $y' = f(y, x)$, $y(0) = \eta$. You may assume that f is Lipschitz continuous with respect to y uniformly for all x.

a) Analyze the consistency, stability, accuracy, and convergence properties of this method.

b) Sketch a graph of the solution to the following initial value problem.

$$y' = -10^8[y - \cos(x)] - \sin(x), \quad y(0) = 2.$$

Would it be more reasonable to use this method or Euler's method for this problem? What would you consider in choosing a timestep k for each of the methods? Justify your answer.

Fact 1: A real symmetric $n \times n$ matrix A can be diagonalized by an orthogonal similarity transformation, and A's eigenvalues are real.

Fact 2: The $(N-1) \times (N-1)$ matrix M defined by

```
[ -2  1  0  0  0  .  .  .  0  0  0  0 ]
[  1 -2  1  0  0  .  .  .  0  0  0  0 ]
[  0  1 -2  1  0  .  .  .  0  0  0  0 ]
[  0  0  1 -2  1  .  .  .  0  0  0  0 ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  0  0  0  0  0  .  .  .  1 -2  1  0 ]
[  0  0  0  0  0  .  .  .  0  1 -2  1 ]
[  0  0  0  0  0  .  .  .  0  0  1 -2 ]
```

has eigenvalues $\mu_l \;=\; -\,4\sin^2(\frac{\pi l}{2N})$, $l = 1, 2, \ldots, N-1$.

Fact 3: The $(N+1) \times (N+1)$ matrix:

```
[ -1  1  0  0  0  .  .  .  0  0  0  0 ]
[  1 -2  1  0  0  .  .  .  0  0  0  0 ]
[  0  1 -2  1  0  .  .  .  0  0  0  0 ]
[  0  0  1 -2  1  .  .  .  0  0  0  0 ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  .  .  .  .  .  .  .  .  .  .  .  . ]
[  0  0  0  0  0  .  .  .  1 -2  1  0 ]
[  0  0  0  0  0  .  .  .  0  1 -2  1 ]
[  0  0  0  0  0  .  .  .  0  0  1 -1 ]
```

has eigenvalues $\mu_l = -4\sin^2\left(\frac{\pi l}{2(N+1)}\right), \quad l = 0, 1, \ldots, N.$

Fact 4: For a real $n \times n$ matrix A, the Rayleigh quotient of a vector $x \in R^n$ is the scalar

$$r(x) = \frac{x^T A x}{x^T x}.$$

The gradient of $r(x)$ is

$$\nabla r(x) = \frac{2}{x^T x}(Ax - r(x)x).$$

If x is an eigenvector of A then $r(x)$ is the corresponding eigenvalue and $\nabla r(x) = 0$.