Foundations of Computational Math I Exam 2
Take-home Exam
Open Notes, Textbook, Homework Solutions Only
Calculators Allowed
Due beginning of Class Wednesday, 2 December 2008

Question	Points Possible	Points Awarded
1. Iterative Methods for $Ax = b$	25	
2. Iterative Methods for $Ax = b$	25	
3. Solving a Nonlinear Equation	25	
4. Solving a System of Nonlinear Equations	25	
Total Points	100	

Name:_____

The alias given on Exam 1 will be used when posting anonymous grade list.

Problem 1

(25 points)

1.a

(10 points) Consider solving $Ax = b$ where the matrix A must be nonsingular by a linear stationary iterative method

$$x_{k+1} = Gx_k + f$$

(i) **(5 points)** Give an example of an iterative method for which G is singular. Justify your answer.

Solution:

The matrix $G_{gs} = (D - L)^{-1}U$ is always singular for Gauss-Seidel.. U is strictly upper triangular and therefore $Ue_1 = 0$.

(ii) **(5 points)** Does G being singular affect the asymptotic rate of the iteration compared to another iteration defined by $\tilde{G}$, that differs only in that $\tilde{G}$ has a nonzero eigenvalue when G has a zero eigenvalue with the rest of the eigenvalues the same for both matrices?

Solution:

Under these assumptions the spectral radii $\rho(\tilde{G})$ and $\rho(G)$ are the same and nonzero. Since they determine asymptotic behavior there is no difference in the methods.

Note as an aside that all of the eigenvalues of a matrix can be 0 and yet multiplying it with a vector can still take multiple steps to create a 0 vector. Consider the matrix

$$B = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

So even if we had an iterative method applied to a linear system where G had all of its eigenvalues 0 it does not guarantee convergence in one step.

1.b

(15 points) When attempting to solve $Ax = b$ where A is known to be nonsingular via an iterative method, we have seen various theorems that give sufficient conditions on A to guarantee the convergence of various iterative methods. It is not always easy to verify these conditions for a given matrix A. Let P and Q be two permutation matrices. Rather than solving $Ax = b$ we could solve $(PAQ)(Q^T x) = Pb$ using an iterative method. Sometimes it is possible to examine A and choose P and/or Q so that it is easy to apply one of our sufficient condition theorems.

(i) **(7 points)** Can you choose P and/or Q so that the permuted system converges for one or both of Gauss-Seidel and Jacobi with

$$A = \begin{pmatrix} 3 & -2 & 7 \\ 1 & 6 & -1 \\ 10 & -2 & 7 \end{pmatrix}?$$

Solution:

$$Q = I \quad P = \begin{pmatrix} e_3^T \\ e_2^T \\ e_1^T \end{pmatrix}$$

$$PA = \begin{pmatrix} 10 & -2 & 7 \\ 1 & 6 & -1 \\ 3 & -2 & 7 \end{pmatrix}$$

PAQ is strictly row diagonally dominant therefore both methods converge. The system with A does not converge for Jacobi and does not coverge for Gauss-Seidel.

(ii) **(8 points)** Can you choose P and/or Q so that the permuted system converges for one or both of Gauss-Seidel and Jacobi with

$$A = \begin{pmatrix} 3 & 7 & -1 \\ 7 & 4 & 1 \\ -1 & 1 & 2 \end{pmatrix}?$$

Solution:

$$P = I \quad Q = \begin{pmatrix} e_2 & e_1 & e_3 \end{pmatrix}$$

$$AQ = \begin{pmatrix} 7 & 3 & -1 \\ 4 & 7 & 1 \\ 1 & -1 & 2 \end{pmatrix}$$

PAQ is irreducibly diagonally dominant therefore both methods converge. The system with A does not converge for Jacobi and does not coverge for Gauss-Seidel.

Problem 2

(25 points)

Consider the block tridiagonal matrix associated with an $n \times n$ grid discretization of the partial differential $u_{\xi,\xi} + u_{\eta,\eta} = g$ on a two-dimensional domain.

The matrix is $n^2 \times n^2 = N \times N$ with $N = n^2$, $n \times n$ blocks $T_i \in \mathbb{R}^{n \times n}$ $1 \leq i \leq n$ $E_i = -I_n \in \mathbb{R}^{n \times n}$ $1 \leq i \leq n$ with block tridiagonal structure given by

$$A = \begin{pmatrix} T_1 & E_1 & 0 & \cdots & \cdots & \cdots & 0 \\ E_2 & T_2 & E_2 & 0 & & & \vdots \\ 0 & E_3 & T_3 & E_3 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & E_{n-1} & T_{n-2} & E_{n-2} & 0 \\ 0 & & \cdots & 0 & E_{n-1} & T_{n-1} & E_{n-1} \\ 0 & & & \cdots & 0 & E_n & T_n \end{pmatrix}$$

where T_i are tridiagonal and E_i are diagonal and dimensions $n \times n$

$$T_i = \begin{pmatrix} 4 & -1 & 0 & 0 & 0 & \ldots & 0 \\ -1 & 4 & -1 & 0 & 0 & \ldots & 0 \\ 0 & -1 & 4 & -1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \ldots & 0 & -1 & 4 & -1 & 0 \\ 0 & \ldots & 0 & 0 & -1 & 4 & -1 \\ 0 & \ldots & 0 & 0 & 0 & -1 & 4 \end{pmatrix}$$

$$E_i = -I_n = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & -1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & -1 & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \ldots & 0 & 0 & -1 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 & -1 & 0 \\ 0 & \ldots & 0 & 0 & 0 & 0 & -1 \end{pmatrix}$$

Note: Be careful with N vs. n in your answer to this problem.

2.a

(15 points)

i. Determine the computational complexity the matrix-vector product $Av \rightarrow z$ where $v,\ z \in \mathbb{R}^N$.

ii. Determine the computational complexity of one step of CG without preconditioning to solve $Ax = b$.

For both answers express the complexity as $CN^k + O(N^{k-1})$ where k is an appropriate integer and C is a constant. (You must give both k and C)

Solution:

The matrix A has 5 nonzero elements per row so $Av \rightarrow z$ has a computational complexity of $9N + O(1)$ operations. Approximately $5N$ multiplications and $4N$ additions if no other structure is exploited. However, 4 elements in each typical row are -1 and therefore do not require multiplications during the matrix vector product. So we have approximately N multiplications and $4N$ additions if all available structure is exploited. This yields a computational complexity of $5N + O(1)$ operations.

The CG without preconditioning iteration in the notes is

x_0 arbitrary; $r_0 = b - Ax_0$; $p_0 = r_0$
$k = 0, 1, \cdots$

$$
\begin{aligned}
v_k &= Ap_k \\
\alpha_k &= r_k^T r_k / p_k^T v_k \\
x_{k+1} &= x_k + \alpha_k p_k \\
r_{k+1} &= r_k - \alpha_k v_k \\
\beta_k &= r_{k+1}^T r_{k+1} / r_k^T r_k \\
p_{k+1} &= r_{k+1} + \beta_k p_k
\end{aligned}
$$

end

Each step requires one matrix vector product, 3 vector triads, and 3 inner products ($r_k^T r_k$, $p_k^T v_k$, and $r_{k+1}^T r_{k+1}$). This is a total complexity of $(5 + 3 * 2 + 3 * 2)N + O(1) = 17N + O(1)$. You can reduce it by one inner product if you notice that you can store $r_{k+1}^T r_{k+1}$ for use in the next step, i.e., where it is $r_k^T r_k$. This would make the complexity $15N + O(1)$.

2.b

(10 points)

Suppose you use an incomplete Cholesky preconditioner where $M = LL^T$. L is a lower triangular matrix with a nonzero structure identical to the diagonal and strict lower triangular part of A, i.e.,

$$L = \begin{pmatrix} L_1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ D_2 & L_2 & 0 & 0 & & & \vdots \\ 0 & D_3 & L_3 & 0 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & D_{n-1} & L_{n-2} & 0 & 0 \\ 0 & & \cdots & 0 & D_{n-1} & L_{n-1} & 0 \\ 0 & & & \cdots & 0 & D_n & L_n \end{pmatrix}$$

where L_i are lower triangular and D_i are diagonal and dimensions $n \times n$ where the nonzero positions are marked with $*$ (the actual values are not important).

$$L_i = \begin{pmatrix} * & 0 & 0 & 0 & 0 & \dots & 0 \\ * & * & 0 & 0 & 0 & \dots & 0 \\ 0 & * & * & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & * & * & 0 & 0 \\ 0 & \dots & 0 & 0 & * & * & 0 \\ 0 & \dots & 0 & 0 & 0 & * & * \end{pmatrix}, \quad D_i = \begin{pmatrix} * & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & * & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & * & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ 0 & \dots & 0 & 0 & * & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & * & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & * \end{pmatrix}$$

i. Determine the computational complexity of one step of CG **with** preconditioning to solve $Ax = b$ using $M = LL^T$. Express the complexity as $CN^k + O(N^{k-1})$ where k is an appropriate integer and C is a constant. (You must give both k and C)

ii. Compare the complexity of one step of CG with preconditioning and one step without preconditioning. How effective does the preconditioner have to be in reducing the number of iterations in order for CG with preconditioning to take fewer computations to reach the same accuracy?

Solution:

The preconditioned CG iteration given in the notes with $M = LL^T$ is:

x_0 arbitrary; $r_0 = b - Ax_0$;
solve $Mz_0 = r_0$; $p_0 = z_0$
$k = 0, 1, \cdots$

$$
\begin{aligned}
&v_k = Ap_k \\
&\alpha_k = r_k^T z_k / p_k^T v_k \\
&x_{k+1} = x_k + \alpha_k p_k \\
&r_{k+1} = r_k - \alpha_k v_k \\
&\text{solve } Mz_{k+1} = r_{k+1} \\
&\beta_k = r_{k+1}^T z_{k+1} / r_k^T z_k \\
&p_{k+1} = z_{k+1} + \beta_k p_k
\end{aligned}
$$

end

We have a forward and backward solve with L and L^T respectively. Since we are assuming nothing about the actual values in the matrices we have 3 elements in most rows which yields 2 multiplications, 2 additions, and 1 division per typical row in a foward or backward solve. Therefore, solving the preconditioning equation for z requires $10N + O(1)$ per step.

Additionally, each step requires one matrix vector product, 3 vector triads, and 3 inner products ($r_k^T z_k$, $p_k^T v_k$, and $r_{k+1}^T z_{k+1}$). This is additional complexity of $(5 + 3 * 2 + 3 * 2)N + O(1) = 17N + O(1)$. The total complexity of a preconditioned step is therefore $(10 + 17)N + O(1) = 27N + O(1)$. This can be reduced by $2N + O(1)$ if we store $r_{k+1}^T z_{k+1}$) for use on the next step. To give $25N + O(1)$.

The preconditioned step is 1.6 times more costly and therefore just to break even the number of iterations must be 1.6 times smaller due to the improvement in the spectrum caused by preconditioning. In practice of course we would want to see a larger factor.

Problem 3

(25 points)

Consider the fixed point iteration by the function

$$\phi(x) = x - \frac{(x^2 - 3)}{(x^2 + 2x - 3)}$$

The value $\alpha = \sqrt{3}$ is a fixed point for this iteration. Provide justification to all of your answers for the following:

(i) **(5 points)** Show that there exists a nontrivial interval $\alpha - \delta < x < \alpha + \delta$ with $\delta > 0$ such that the iteration defined by $\phi(x)$ converges to α for any x_0 in the interval.

(ii) **(5 points)** Is the order of convergence on this interval linear ($p = 1$) or higher ($p \geq 2$)?

(iii) **(10 points)** Show that for $x > \alpha$ we have

$$\alpha < \phi(x) < x$$

and use this to explain the convergence of the iteration when $x_0 > \alpha$. (**Suggestion:** let $x = \alpha + \epsilon$ with $\epsilon > 0$ and substitute into the desired relationships.)

(iv) **(5 points)** Plot or otherwise enumerate values of the curves $y = \phi(x)$ and $y = x$ on the interval $0 < x < \beta$ for $\beta > \alpha$. Use the information to examine the behavior of the iteration for $0 < x_0 < \alpha$. For what subinterval, if any, do you expect convergence to α?

You need not turn in the plot or enumeration. Simply state the important characteristics that support your conclusion. The plot can also assist in supporting your answers for the earlier parts of the question.

Solution:

$$\phi(x) = x - \frac{(x^2 - 3)}{(x^2 + 2x - 3)}$$

$$\phi'(x) = 1 - \frac{(2x)(x^2 + 2x - 3) - (x^2 - 3)(2x + 2)}{(x^2 + 2x - 3)^2}$$

$$= \frac{(x^2 + 2x - 3)^2 - (2x)(x^2 + 2x - 3) + (x^2 - 3)(2x + 2)}{(x^2 + 2x - 3)^2}$$

It is easily seen that $\phi'(\alpha) = 0$. From this we conclude two things:

- $\phi(x)$ is a contraction mapping on some interval around $\alpha = \sqrt{3}$ and therefore the iteration converges to α for any x_0 in the interval.
- The iteration converges at least quadratically. In fact examining $\phi''(\alpha)$, although not required for this problem, reveals it is nonzero and therefore convergence is quadratic.

To extend the convergence region we first consider $x > \alpha$. If

$$\alpha < \phi(x) < x$$

then iteration remains in the interval but moves closer to α on each iteration. It therefore converges to α. To verify that the ordering is correct let $\epsilon > 0$ and note that

$$\phi(\sqrt{3} + \epsilon) = \sqrt{3} + \epsilon - \gamma$$
$$\gamma = \frac{(\sqrt{3} + \epsilon)^2 - 3}{(\sqrt{3} + \epsilon)^2 - 3 + 2(\sqrt{3} + \epsilon)} = \epsilon \left(\frac{(\epsilon + 2\sqrt{3})}{\epsilon(\epsilon + 2\sqrt{3}) + 2(\epsilon + \sqrt{3})} \right) = \epsilon\mu$$
$$0 < \mu < 1$$

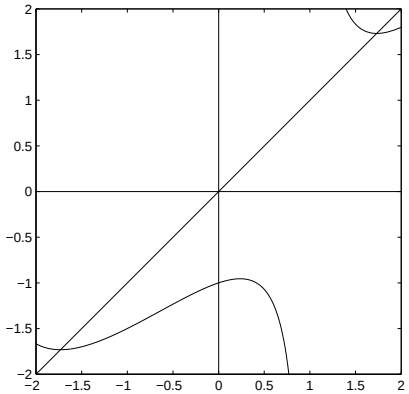
We have that

$$\sqrt{3} < \sqrt{3} + \epsilon(1 - \mu) < \sqrt{3} + \epsilon$$
$$\sqrt{3} < \phi(\sqrt{3} + \epsilon) < \sqrt{3} + \epsilon$$

$$\therefore \forall x > \alpha, \quad \alpha < \phi(x) < x$$

and our convergence result follows.

The remainder of the real line can be best seen from examining the standard plot of $y = \phi(x)$ and $y = x$. We have



$\phi(x)$ has a positive vertical asymptote when approaching $x = 1$ from above and a negative vertical asymptote when approaching $x = 1$ from below due to the root of the denominator at 1. On the interval $\alpha - \delta < x < \alpha$ we know it is a contraction mapping. Clearly, for x_0

near $x = 1$, the value of $x_1 = \phi(x_0) > \alpha$ and convergence to α results from there. It is clear from the lower left part of the plot that between $0 < x < 1$ we have convergence to $-\alpha$ that can be shown rigorously in a manner to that above. There is also a vertical asymptote of concern due to the second root of the denominator at $x = -3$.

Problem 4

(25 points)

Consider the system of equations in $\mathbb{R}^2$

$$\xi^2 + \eta^2 = 4$$
$$e^{\xi} + \eta = 1$$

The system has two solution vectors.

Consider the fixed point iteration by the function

$$G(x) = \begin{pmatrix} \gamma_1(\xi,\eta) \\ \gamma_2(\xi,\eta) \end{pmatrix} = \begin{pmatrix} \log(1-\eta) \\ -\sqrt{4-\xi^2} \end{pmatrix}$$

$$\begin{pmatrix} \xi_{k+1} \\ \eta_{k+1} \end{pmatrix} = \begin{pmatrix} \log(1-\eta_k) \\ -\sqrt{4-\xi_k^2} \end{pmatrix}$$

One of the solutions is in the rectangle $(0,\sqrt{2})\times(0,-2)$. Show that the iteration converges to that solution for any

$$(\xi_0,\eta_0) \in (0,\sqrt{2}) \times (0,-2)$$

(**Hint:** Consider the Jacobian of $G(x)$ carefully when proving convergence on the region.)

Solution:

It is easy to see that the two equations are a circle of radius 2 center at the origin and exponential opening down and to the left. The exponential, passes through the origin, has an asymptote at $\eta = 1$ as $\xi \to -\infty$ and it is $\eta \to -\infty$ as $\xi \to -\infty$. There are two intersections: the leftmost is in the upper left quadrant and the rightmost is in the lower right quadrant. The intersection is contained in the box $(0,\sqrt{2})\times(0,-2)$ and this can be used as the domain in which the initial condition is chosen.

The key to proving convergence is to note that the 2×2 Jacobian is very simple for functional iterations that are easily derived by separating ξ and η. There are 0's in the diagonal elements and nonzero off-diagonals. By Gershgorin, if the magnitudes of the $(1,2)$ and $(2,1)$ elements are less than 1 the iteration is a contraction.

The Jacobian is

$$J(x) = \begin{pmatrix} 0 & \frac{1}{1-\eta} \\ \frac{\xi}{\sqrt{4-\xi^2}} & 0 \end{pmatrix}$$

Clearly, $|1/1-\eta| < 1$ for $\eta < 0$. We have

$$\frac{\xi}{\sqrt{4-\xi^2}} = \begin{cases} 0 & \text{when } \xi = 0 \\ \to \infty & \text{when } \xi \to 2 \\ 1 & \text{when } \xi = \sqrt{2} \end{cases}$$

It is easy to see that

$$\frac{\alpha\xi}{\sqrt{4-\alpha^2\xi^2}} > \frac{\xi}{\sqrt{4-\xi^2}}$$

for $\alpha > 1$ and $\alpha\xi \in (0, \sqrt{2})$. So both offdiagonal elements of the Jacobian are less than one in magnitude $\forall(\xi, \eta) \in (0, \sqrt{2}) \times (0, -2)$. Since the domain also contains the root we have the desired result of a contraction mapping with constraints on the initial condition.