Foundations of Computational Math 1 Final
In-class Exam
Open Notes, Textbook, Homework Solutions Only
Calculators may be used.
Tuesday 8 December, 2009

Question	Points Possible	Points Awarded
1. Iterative Methods Nonlinear equation	30	
2. Backward Error for Linear Systems	30	
3. Matrix Factorization	30	
4. Iterative Methods for Linear Systems	25	
5. Quasi-Newton Methods	30	
6. Finite Precision	25	
Total Points	170	

Name:

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

Are you registered for an S/U course grade ?

YES NO

Problem 1

(30 points)

Recall, Steffensen's method for finding the roots of a nonlinear scalar function from the homework:

$$x_{k+1} = g(x_k)$$

$$g(x) = x - \frac{f(x)}{\phi(x)}, \qquad \phi(x) = \frac{f(x + f(x)) - f(x)}{f(x)}$$

where $x_k \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$. Consider applying Steffensen's method to the function

$$f(x) = (x - \alpha)^2$$

The function has a double root at α.

1.a

(15 points)

i. Newton's method and Steffensen's method can have difficulties when evaluating their iteration at a double root α. For example, since $f'(\alpha) = 0$ and $\phi(\alpha) = 0$ both methods produce an undefined update of

$$x_{k+1} = \alpha - \frac{0}{0}.$$

Rewrite, $g(x)$, for Steffensen's method for $f(x) = (x - \alpha)^2$ so that the update is well-defined at the fixed point α, i.e.,

$$\alpha = \alpha - \frac{0}{C + 0}, \qquad C \neq 0$$

ii. Is your rewritten expression valid for all $x \in \mathbb{R}$ or is it undefined at some other point $\tilde{x} \neq \alpha$?

Solution: We have

$$g(x) = x - \frac{f(x)}{\phi(x)}, \qquad \phi(x) = \frac{f(x + f(x)) - f(x)}{f(x)}$$
$$f(x) = (x - \alpha)^2$$

$$f(x + f(x)) = (x + (x - \alpha)^2 - \alpha)^2 = (\hat{x} + \hat{x}^2)^2, \quad \hat{x} = x - \alpha$$

$$g(x) = x - \frac{\hat{x}}{2 + \hat{x}} = x - \frac{(x - \alpha)}{2 + (x - \alpha)}$$

So we have

$$\alpha = g(\alpha) = \alpha - \frac{0}{2}$$

and $\tilde{x} = \alpha - 2$ implies an undefined $g(\tilde{x})$.

1.b

(15 points)

Show that Steffensen's method converges only linearly to the double root α.

Solution:

$$g(x) = x - \frac{\hat{x}}{2+\hat{x}} = x - \frac{(x-\alpha)}{2+(x-\alpha)}$$
$$g'(x) = 1 - \frac{2+(x-\alpha)-(x-\alpha)}{(2+(x-\alpha))^2} = 1 - \frac{2}{(2+(x-\alpha))^2}$$

We have

$$g'(x) = 1 - \frac{2+(x-\alpha)-(x-\alpha)}{(2+(x-\alpha))^2} = 1 - \frac{2}{(2+(x-\alpha))^2}$$
$$g'(\alpha) = 1 - 1/2 = 1/2$$
$$\therefore |g'(\alpha)| < 1$$

So we have a nontrivial interval around α on which g is a contraction mapping and therefore the iteration converges. For determining the asymptotic rate of convergence this is all that is needed. We also note that $g'(\alpha) \neq 0$ so convergence is linear.

Problem 2

(30 points)

2.a

(15 points)
Let $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$, and $b \in \mathbb{R}^n$ be given with

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}, \quad b = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}.$$

Suppose that

$$Ax \neq b \quad \text{and} \quad \xi_i \neq 0 \quad 1 \leq i \leq n$$

Show that there exists a backward error $E \in \mathbb{R}^{n \times n}$ such that

$$(A + E)x = b$$

where E is a diagonal matrix, i.e., all of the off-diagonal elements are 0 and any nonzero element must be on the diagonal.
Solution:

$$(A + E)x = b$$
$$Ex = b - Ax = r$$
$$\epsilon_i \xi_i = \rho_i, \quad 1 \leq i \leq n$$

$$\therefore \epsilon_i = \rho_i / \xi_i, \quad 1 \leq i \leq n$$

All ϵ_i exist since $\xi_i \neq 0$.

2.b

(15 points)
Suppose that

$$Ax \neq b \quad \text{and} \quad A = A^T$$

i.e., the matrix A is symmetric. Let $r = b - Ax$.
Show that if $r^T x \neq 0$ then there exists a backward error $E \in \mathbb{R}^{n \times n}$ such that

$$(A + E)x = b$$

where E is a symmetric rank-1 matrix, i.e.,

$$E = \sigma v v^T, \qquad v \in \mathbb{R}^n, \quad \sigma \in \mathbb{R}.$$

Solution:

$$\begin{gathered}(A+E)x = b \\ Ex = b - Ax = r \\ \therefore E = r v^T, \quad v^T x = 1\end{gathered}$$

We must also enforce symmetry so we have

$$\begin{gathered}E = \sigma v v^T \\ \sigma = \frac{1}{r^T x} \\ v = r \\ Ex = \sigma v v^T x = \frac{1}{r^T x} r r^T x = \frac{r^T x}{r^T x} r = r\end{gathered}$$

Problem 3

(30 points)

Let $A \in \mathbb{R}^{n\times n}$ be a diagonally dominant tridiagonal matrix, e.g., for $n = 10$

$$A = \begin{pmatrix} \alpha_1 & \gamma_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_2 & \alpha_2 & \gamma_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_3 & \alpha_3 & \gamma_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_4 & \alpha_4 & \gamma_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_5 & \alpha_5 & \gamma_5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta_6 & \alpha_6 & \gamma_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_7 & \alpha_7 & \gamma_7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \beta_8 & \alpha_8 & \gamma_8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_9 & \alpha_9 & \gamma_9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{10} & \alpha_{10} \end{pmatrix}$$

There exists a nonsingular transformation matrix T^{-1} such that

$$T^{-1}A = S \rightarrow A = TS$$

where

$$S = \begin{pmatrix} I_{n/2} & M \\ N & I_{n/2} \end{pmatrix}$$

where $I_{n/2}$ is the identity matrix in $\mathbb{R}^{n/2 \times n/2}$, $M \in \mathbb{R}^{n/2 \times n/2}$ with all of its nonzero elements in its first column and $N \in \mathbb{R}^{n/2 \times n/2}$ with all of its nonzero elements in its last column. For $n = 10$ we have

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \sigma_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \sigma_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \sigma_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \sigma_5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_6 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_7 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_8 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_9 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{10} & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

3.a

(15 points)

Show the form of the matrix T. (**Hint:** It is very simply related to the matrix A) You may give your answer for $n = 10$ if it helps the presenation.

Solution:

$$T = \begin{pmatrix} \alpha_1 & \gamma_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_2 & \alpha_2 & \gamma_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_3 & \alpha_3 & \gamma_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_4 & \alpha_4 & \gamma_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_5 & \alpha_5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha_6 & \gamma_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_7 & \alpha_7 & \gamma_7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \beta_8 & \alpha_8 & \gamma_8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_9 & \alpha_9 & \gamma_9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \beta_{10} & \alpha_{10} \end{pmatrix}$$

More generally

$$T = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

where $A_1 \in \mathbb{R}^{n/2 \times n/2}$ $A_2 \in \mathbb{R}^{n/2 \times n/2}$ are the block diagonal submatrices of $A \in \mathbb{R}^{n \times n}$.

3.b

(15 points)

Show how to compute S from A. You may give your answer for $n = 10$ if it helps the presenation.

Solution:

Take $n = 2$ and the pattern of solution is suggested.

$$\begin{pmatrix} \alpha_1 & \gamma_1 \\ \beta_2 & \alpha_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} 1 & \alpha_1^{-1}\gamma_1 \\ \alpha_2^{-1}\beta_2 & 1 \end{pmatrix}$$

$$\alpha_1\sigma_1 = \gamma_1 \qquad \alpha_2\sigma_2 = \beta_2$$

Generalizing we have

$$\begin{pmatrix} \alpha_1 & \gamma_1 & 0 & 0 & 0 \\ \beta_2 & \alpha_2 & \gamma_2 & 0 & 0 \\ 0 & \beta_3 & \alpha_3 & \gamma_3 & 0 \\ 0 & 0 & \beta_4 & \alpha_4 & \gamma_4 \\ 0 & 0 & 0 & \beta_5 & \alpha_5 \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \gamma_5 \end{pmatrix}$$

$$\begin{pmatrix} \alpha_6 & \gamma_6 & 0 & 0 & 0 \\ \beta_7 & \alpha_7 & \gamma_7 & 0 & 0 \\ 0 & \beta_8 & \alpha_8 & \gamma_8 & 0 \\ 0 & 0 & \beta_9 & \alpha_9 & \gamma_9 \\ 0 & 0 & 0 & \beta_{10} & \alpha_{10} \end{pmatrix} \begin{pmatrix} \sigma_6 \\ \sigma_7 \\ \sigma_8 \\ \sigma_9 \\ \sigma_{10} \end{pmatrix} = \begin{pmatrix} \beta_6 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Therefore, the two vectors that define S are the result of solving two diagonally dominant tridiagonal systems of size 5 or more generally $n/2$. This can be in $O(n)$ computations using an LU factorization without pivoting.

This can be written more generally as

$$A_1 M = \gamma_{n/2} e_{n/2} e_1^T$$

$$A_2 N = \beta_{(n/2)+1} e_1 e_{n/2}^T$$

where everything is in $\mathbb{R}^{n/2 \times n/2}$. The structure of the righthand side matrices shows that M and N have only one nonzero column each (the first and the last respectively).

Problem 4

(25 points) Consider simple accelerated Richardson's method to solve $Ax = b$

$$\begin{gathered} \text{Given,} \quad x_0 \\ x_{k+1} = x_k + \alpha r_k \\ r_k = b - Ax_k \\ \alpha > 0 \end{gathered}$$

The nonsymmetric matrix

$$A = \begin{pmatrix} 10 & 2 & 3 \\ 2 & 2 & 1 \\ 0 & 1 & 3 \end{pmatrix}$$

has real positive eigenvalues. What value would choose for $\alpha > 0$ so that simple accelerated Richardson's method will converge for any x_0? Justify your answer.

Solution:

Use Gershgorin's second theorem and consider the column disks. We have

$$\begin{aligned} \mathcal{C}_1 : &\quad 8 \leq \lambda \leq 12 \\ \mathcal{C}_2 : &\quad -1 \leq \lambda \leq 5 \\ \mathcal{C}_3 : &\quad -1 \leq \lambda \leq 7 \end{aligned}$$

Therefore, we have an isolated disk and $8 \leq \lambda_{max} \leq 12$. Since positive real eigenvalues and $\alpha < 2/\lambda_{max}$ guarantees convergence we can take $\alpha < 1/6$.

From MATLAB the eigenvalues are

$$10.568136060691241, \quad 1.351700746414633, \quad 3.080163192894140$$

An alternative is to simply use the norm $\|A\|_1$ which yields the same bound. Note that the information that A has positive eigenvalues is crucial. If it is indefinite then the spectral radius of A is meaningless since we know the iteration cannot converge. The disks do not guarantee the definiteness of A. Extra information is required (and provided here).

Problem 5

(30 points)

Consider solving the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ using a Quasi-Newton method.

5.a

(10 points)

Suppose the Quasi-Newton method guarantees that the symmetric matrix $B_k \in \mathbb{R}^{n \times n}$ used to define the local quadratic model of $f(x)$ is symmetric positive definite.

Show that one can only find a symmetric positive definite matrix B_{k+1} that satisfies the secant condition

$$B_{k+1} s_k = y_k$$

if s_k and y_k satisfy a simple constraint. (**Hint:** Consider the angle between s_k and y_k.)

Solution:

Since B_{k+1} is positive definite we have

$$\forall x, \quad x^T B_{k+1} x > 0$$

The secant condition requires

$$B_{k+1} s_k = y_k$$

Therefore

$$s_k^T B_{k+1} s_k > 0 \to s_k^T y_k > 0$$

This is called the curvature condition.

5.b

(20 points)

Consider solving the minimization problem by Newton's methods ($\alpha_k = 1$) and BFGS using the H_k update form. Suppose the problem size n is moderate so all linear systems are solved using Cholesky factorization and there is no sparsity to exploit in any matrix operation.

i. Describe the computational complexity of one step of Newton's method.

ii. Describe the computational complexity of one step of BFGS using the H_k update form.

iii. Assuming the search for α_k satisfying Wolfe's conditions is not computationally significant, discuss how much the number of iterations must be reduced to justify using Newton's method.

Solution:

The main observation comparing the two methods is that every computation in the BFGS step is $O(n^2)$ at most. Since the H_k update form is used there is no system to solve. All matrix vectors are $2n^2 + O(n)$ and the update is $O(n^2)$.

On the other hand, Newton's method requires the formation of the Hessian which requires

$$\frac{\gamma}{2}n^2$$

computations where γ is the average number of operations per element of the Hessian. (We assume here that γ is independent of n.) Since it is symmetric, only half of the elements must be computed. The Hessian must then be factored and since it symmetric positive definite we can use Cholesky. This requires

$$\frac{n^3}{3} + O(n^2)$$

computations and it is the dominant part of each step. We therefore have a complexity for a step of Newton's method of

$$\frac{n^3}{3} + O(n^2)$$

versus $O(n^2)$ for BFGS. The convergence must be accelerated considerably (by a factor of Cn) in order to profit from Newton. This can be done by having a very good initial guess (which is often the case when exact Newton is used). However, clearly the need for an inexact Newton to remove the $O(n^3)$ term and have two $O(n^2)$ methods is crucial if a Newton method other than the Quasi-Newton family is to be practical.

To see that the H_k update is $O(n^2)$ note that

$$\begin{aligned} H_{k+1} &= (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T \\ &= H_k - (H_k y_k) s_k^T - s_k (H_k y_k)^T + (\rho_k - y_k^T H_k y_k) s_k s_k^T \end{aligned}$$

Only half the elements must be computed due to symmetry. Note $H_k y_k$ can be computed once with $O(n^2)$ complexity. Each of the terms added to H_k requires $O(n^2)$ complexity and since there is a constant number of updates, the total cost is $O(n^2)$.

The assumption that the search for α_k that satisfies the Wolfe conditions is insignificant is not completely reasonable. It can be a significant portion of the BFGS step's complexity. Of course, it varies from step to step.

Since Newton's method is usually used in its damped form, i.e., $\alpha_k \neq 1$ it will also have such a cost but the major problem with Newton in practice is the evaluation and factorization of the Hessian on every step.

Problem 6

(25 points)

Let $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^n$ be two vectors with

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}, \quad y = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}.$$

$$|\xi_i| \geq 1 \qquad |\eta_i| \geq 1$$

Consider the evaluation of the two inner products

$$\mu = x^T x$$

$$\gamma = x^T y$$

Which of the two inner products would you expect to be less sensitive to the perturbations caused by the finite precision of IEEE floating point arithmetic? Justify your answer.

Solution:

We know that there is a backward error analysis for these inner products. We looked carefully at the summation portion. It is the summation that is of concern here. Recall that the sensitivity of the summation to perturbation in the inputs (which can be used to bound roundoff error) is determined by the condition number of the sum. We showed this to be

$$\kappa_{rel} = \frac{|\rho_1| + \cdots + |\rho_n|}{|\rho_1 + \cdots + \rho_n|}$$

for computing

$$\sigma = \rho_1 + \cdots + \rho_n$$

Now consider the two sums associated with μ and γ. For μ we have

$$\rho_i = \xi_i^2 > 0 \rightarrow \kappa_{rel} = 1$$

This is perfectly conditioned.

For γ we have

$$\rho_i = \eta_i \xi_i$$

Since the signs of these elements cannot be determined we can have well-conditioned and ill-conditioned problems.

So we would expect the computation of $\mu = x^T x$ to be less sensitive to finite precision errors.