UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 1999

BEng Honours Degree in Computing Part III
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute

PAPER 3.14

NUMERICAL ANALYSIS
Wednesday, May 5th 1999, 10.00 – 12.00

Answer THREE questions

For admin. only:
paper contains 4 questions

1a Explain briefly the fundamental difference between C and FORTRAN in the ordering of two-dimensional arrays and its implication for linear algebra algorithms.

b Let L be a non-singular $n \times n$ lower triangular matrix, U a non-singular $n \times n$ upper triangular matrix and $\mathbf{b}$ an n-vector.

i) Explain clearly what is meant by forward-substitution and backward-substitution for solving the systems of equations

$$L\mathbf{x} = \mathbf{b} \quad \text{and} \quad U\mathbf{x} = \mathbf{b}$$

respectively. State precisely the operation count in each case.

ii) Construct two algorithms for solving $L\mathbf{x} = \mathbf{b}$ with $\mathbf{x}$ overwriting $\mathbf{b}$; one being preferable if it is to be implemented in C, the other being preferable if it is to be implemented in FORTRAN.

iii) Construct the analogous pair of algorithms for solving $U\mathbf{x} = \mathbf{b}$ with $\mathbf{x}$ overwriting $\mathbf{b}$.

c If a lower triangular matrix L is non-singular, show, by considering the equation $L\mathbf{x} = \mathbf{e}_j$ where $\mathbf{e}_j$ is the j^{th} unit vector, that L^{-1} is also lower triangular. Describe clearly the corresponding result and proof for a non-singular upper triangular matrix U.

The five parts carry, respectively, 10%,30%,20%,20%,20% of the total marks.

2a Let z be an arbitrary non-zero real number and $fl(z)$ its closest representable floating point approximation, using base β and finite length mantissa with precision m.

i) State the relative error for $fl(z)$ in terms of β and m.

ii) Define the unit round-off $\bar{u}$ and express $fl(z)$ as a perturbation of z.

iii) State, in terms of $\bar{u}$, the relative error bounds and perturbation results that the four basic arithmetic operations must satisfy in IEEE standard arithmetic.

b For this section it is to be assumed that n is an integer such that the unit round-off $\bar{u}$ satisfies $n\bar{u} < 1$, and you are given that $\forall k \leq n$

$$|\delta_i| \leq \bar{u} \quad i = 1, \ldots, k \qquad \Rightarrow \qquad \prod_{i=1}^{k}(1+\delta_i) = 1 + \theta_k$$

with $|\theta_k| \leq \frac{k\bar{u}}{1-k\bar{u}}$. Consider the two algorithms below.

```
s ← 0;
for i = 1,...,n do
    s ← s + x_i y_i;
end i.
```

```
for i = 1,...,n do
    c_i ← 0;
    for j = 1,...,n do
        c_i ← c_i + a_ij x_j;
    end j;
end i.
```

i) If the scalar product $s \equiv \mathbf{x}^T\mathbf{y} \quad \mathbf{x}, \mathbf{y} \in \Re^n$ is obtained from the algorithm on the left, explain carefully why the computed approximation $\hat{s}$ satisfies

$$\hat{s} = (\mathbf{x} + \Delta\mathbf{x})^T\mathbf{y}, \qquad \text{where} \qquad |\Delta\mathbf{x}| \leq \frac{n\bar{u}}{1-n\bar{u}}|\mathbf{x}|.$$

State the resulting bound on the difference between s and $\hat{s}$.

ii) Suppose the matrix-vector product $\mathbf{c} = A\mathbf{x}$, for $\mathbf{x} \in \Re^n$ and $A \in \Re^{n\times n}$, is obtained from the algorithm on the right. Use the result in i) to deduce that the computed approximation $\hat{\mathbf{c}}$ satisfies

$$\hat{\mathbf{c}} = (A + \Delta A)\mathbf{x}, \qquad \text{where} \qquad |\Delta A| \leq \frac{n\bar{u}}{1-n\bar{u}}|A|.$$

State the resulting bound on the difference between $\mathbf{c}$ and $\hat{\mathbf{c}}$.

The five parts carry, respectively, 10%,15%,15%,40%,20% of the total marks.

Turn over ...

3a Give the definition for an $n \times n$ matrix Q to be orthogonal and deduce that, when Q is orthogonal,

$$\|Q\mathbf{x}\|_2 = \|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \Re^n.$$

b Verify that the $n \times n$ matrix

$$H(\mathbf{w}) \equiv I - 2\frac{\mathbf{w}\mathbf{w}^T}{\mathbf{w}^T\mathbf{w}}$$

is orthogonal for every non-zero $\mathbf{w} \in \Re^n$ and that, if the first $k-1$ components of $\mathbf{w}$ are zero, then

i) $H(\mathbf{w})\mathbf{x}$ leaves the first $k-1$ components of $\mathbf{x}$ unchanged,

ii) $H(\mathbf{w})\mathbf{x} = \mathbf{x}$ if the last $n-k+1$ components of $\mathbf{x}$ are zero,

c If $\mathbf{y} \in \Re^n$ satisfies $\sum_{i=2}^n y_i^2 \neq 0$, verify that

$$H(\|\mathbf{y}\|_2\mathbf{e}_1 + \mathbf{y})\mathbf{y} = -\|\mathbf{y}\|_2\mathbf{e}_1,$$

where $\mathbf{e}_1 \in \Re^n$ is the first unit vector, by calculating

$$(\|\mathbf{y}\|_2\mathbf{e}_1 + \mathbf{y})^T\mathbf{y} \quad \text{and} \quad (\|\mathbf{y}\|_2\mathbf{e}_1 + \mathbf{y})^T(\|\mathbf{y}\|_2\mathbf{e}_1 + \mathbf{y}).$$

d If $\mathbf{y} \in \Re^n$ satisfies $\sum_{i=k+1}^n y_i^2 \neq 0$, verify that

$$H(\|\hat{\mathbf{y}}\|_2\mathbf{e}_k + \hat{\mathbf{y}})\mathbf{y} = (y_1, \ldots, y_{k-1}, -\|\hat{\mathbf{y}}\|_2, 0, \ldots, 0)^T,$$

where $\mathbf{e}_k \in \Re^n$ is the k^{th} unit vector and $\hat{\mathbf{y}} \equiv (0, \ldots, 0, y_k, y_{k+1}, \ldots, y_n)^T$.

e Use parts c and d to explain how matrices of the form $H(\mathbf{w})$ can be used to simplify and then solve the system of equations

$$A\mathbf{x} = \mathbf{b},$$

where A is a given non-singular $n \times n$ matrix and $\mathbf{b}$ is a given n-vector.

The five parts carry, respectively, 10%,20%,15%,15%,40% of the total marks.

4. Let $A \equiv \{a_{ij}\}$ be a given $m \times n$ matrix with $m > n$, $\mathbf{b} \equiv \{b_i\}$ a given vector in $\Re^m$, $\mathbf{x} \equiv \{x_i\}$ a general vector in $\Re^n$ and

$$g(\mathbf{x}) \equiv \|\mathbf{b} - A\mathbf{x}\|_2^2$$
$$\equiv \sum_{i=1}^{m}\left(b_i - \sum_{j=1}^{n} a_{ij}x_j\right)^2 .$$

a Deduce that

$$(\dagger) \qquad \frac{\partial}{\partial x_k} g(\mathbf{x}) = 0$$

if and only if

$$\mathbf{a}_k^T (\mathbf{b} - A\mathbf{x}) = 0,$$

where $\mathbf{a}_k$ is the k^{th} column of A, and hence establish that ($\dagger$) holds for $k = 1, \ldots, n$ if and only if

$$(\ddagger) \qquad A^T A\mathbf{x} = A^T\mathbf{b}.$$

b Assume that $A^T A$ is non-singular.

i) Deduce that $A\mathbf{z} = \mathbf{0}$ if and only if $\mathbf{z} = \mathbf{0}$.

ii) Use i) to prove that $\mathbf{z}^T A^T A\mathbf{z} > \mathbf{0}$ unless $\mathbf{z} = \mathbf{0}$.

iii) If $\mathbf{x}^\star$ denotes the solution of ($\ddagger$), use ii) to verify that

$$g(\mathbf{x}) > g(\mathbf{x}^\star) \quad \text{unless} \quad \mathbf{x} = \mathbf{x}^\star.$$

c For the particular case

$$A \equiv \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{b} \equiv \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

i) Apply the normal equations approach to find the linear least squares solution of $A\mathbf{x} = \mathbf{b}$.

ii) Draw the three lines

$$a_{i1}x_1 + a_{i2}x_2 = b_i \qquad i = 1, 2, 3$$

on an $x_1 : x_2$ graph, together with your least squares solution.

The six parts carry, respectively, 20%,10%,10%,30%,20%,10% of the total marks.

End of Paper