FAMILY NAME: Solutions
OTHER NAME(S):
STUDENT NUMBER:
SIGNATURE:

THE UNIVERSITY OF NEW SOUTH WALES

SCHOOL OF MATHEMATICS AND STATISTICS

Example Class Test 1

# MATH2089
# Numerical Methods Example Class Test 1

(1) TIME ALLOWED – 50 minutes

(2) TOTAL NUMBER OF QUESTIONS – 4

(3) ANSWER ALL QUESTIONS

(4) THE QUESTIONS ARE OF EQUAL VALUE

(5) THIS PAPER MAY **NOT** BE RETAINED BY THE CANDIDATE

(6) **ONLY** CALCULATORS WITH AN AFFIXED "UNSW APPROVED" STICKER MAY BE USED

(7) Write your answers on this test paper in the space provided.
Ask your tutor if you need more paper.

All answers must be written in ink. Except where they are expressly required pencils may only be used for drawing, sketching or graphical work.

**1.** a) **[3 marks]** Give the results of the following MATLAB commands when executed on a computer:

```
a = [-1:1]
b = a./(a.^3-a)
```

*Answer:*

$a = [-1 \quad 0 \quad 1]$

$b = [-1 \quad 0 \quad 1]./([-1 \quad 0 \quad 1].\wedge 3 - [-1 \quad 0 \quad 1])$

$= [-1 \quad 0 \quad 1]./([-1 \quad 0 \quad 1] - [-1 \quad 0 \quad 1])$

$= [-1 \quad 0 \quad 1]./[0 \quad 0 \quad 0]$

$= [-\text{Inf} \quad \text{NaN} \quad \text{Inf}]$

b) **[3 marks]**

i) **Define** the **relative error** in a computed approximation $\bar{x}$ to $x \neq 0$.
*Answer:*

$$\text{Relative error} = \frac{|x - \bar{x}|}{|x|}$$

ii) Estimate the **absolute error** in storing $y = (8.01 \times 10^{12})^{\frac{1}{3}}$ on a computer using double precision floating point arithmetic.
*Answer:*

Absolute error in storing $y$ on a computer is $|y|\,\varepsilon$ where

$\varepsilon$ = relative machine precision
$= 2.2 \times 10^{-16}$ in double precision

Absolute error $= |(8.01 \times 10^{12})^{1/3}| \times 2.2 \times 10^{-16}$
$\approx 2 \times 10^{4} \times 2.2 \times 10^{-16}$
$\approx 4 \times 10^{-12}$

c) [**4 marks**] You are asked to calculate the expression

$$D = b + \sqrt{b^2 + \alpha}$$

when $b < 0$ and $\alpha$ is much smaller in magnitude than $b$

i) Explain why this expression is/is not good for implementation on a computer.
*Answer:*

$\alpha$ much smaller than $b$ in magnitude

$\Rightarrow b^2 + \alpha \approx b^2$

$\Rightarrow \sqrt{b^2+\alpha} \approx |b|$

Thus if $b < 0$, the expression $b + \sqrt{b^2+\alpha}$ is subtracting two nearly equal values, an example of catastophic cancellation (significant increase in relative error in result)

ii) Find a mathematically equivalent, but numerically preferable, expression for $D$.
*Answer:*

$$D = \left(b + \sqrt{b^2+\alpha}\right) \times \frac{\left(b - \sqrt{b^2+\alpha}\right)}{\left(b - \sqrt{b^2+\alpha}\right)}$$

$$= \frac{b^2 - (b^2+\alpha)}{b - \sqrt{b^2+\alpha}}$$

$$= \frac{-\alpha}{b - \sqrt{b^2+\alpha}} = \frac{\alpha}{\sqrt{b^2+\alpha} - b}$$

This does <u>not</u> exhibit catastophic cancellation when $b < 0$ as then $-b > 0$, and $\sqrt{b^2+\alpha} - b$ is the addition of two numbers.

**2.** The computational complexity of some common operations with $n$ by $n$ matrices are given in the Table below.

| Operation | Flops |
|---|---|
| Matrix multiplication | $2n^3$ |
| LU factorization | $\frac{2n^3}{3} + O(n^2)$ |
| Cholesky factorization | $\frac{n^3}{3} + O(n^2)$ |
| Back/forward substitution | $n^2 + O(n)$ |
| Tridiagonal solve | $8n + O(1)$ |

a) [**4 marks**] You have a 3GHz quad core computer where each core can do two floating point operations per clock cycle. Estimate how long it will take to solve the $n$ by $n$ linear system $A\mathbf{x} = \mathbf{b}$ where $A$ has no special structure and $n = 10^4$.

*Answer:*

Speed of computer $= \text{GHz} \times \text{cores} \times \text{flops/cycle}$

$= 3 \times 10^9 \times 4 \times 2$

$= 2.4 \times 10^{10}$ flops / sec.

A has no special structure, so need LU factorization.

$\Rightarrow \frac{2n^3}{3}$ flops is dominant cost

$n = 10^4 \Rightarrow \frac{2 \times 10^{12}}{3}$ flops

$$\text{Time} = \frac{\text{number of flops}}{\text{speed (flops/sec)}}$$

$$= \frac{2 \times 10^{12}}{3 \times 2.4 \times 10^{10}} \text{ secs}$$

$\approx 28$ seconds

b) **[3 marks]** Estimate the size $n$ of the largest $n$ by $n$ tridiagonal matrix that can be stored in 1Gb RAM using double precision floating point arithmetic.

*Answer:*

$n \times n$ tridiagonal matrix requires $3n-2$ elements
( $n$ on main diagonal, $n-1$ for immediate sub and super diagonals)

Double precision $\Rightarrow$ 8 bytes / element
$\Rightarrow$ $24n$ bytes (ignoring lower order terms)

$1\ \text{Gb} = 2^{30}$ bytes $\Rightarrow 24n = 2^{30} \Rightarrow n = 44{,}739{,}000$

<u>or</u> $1\ \text{Gb} = 10^{9}$ bytes $\Rightarrow 24n = 10^{9} \Rightarrow n = 41{,}667{,}000$

In either case $n \approx$ 42- 45 million

c) **[3 marks]** A programmer claims that as solving a linear system $A\mathbf{x} = \mathbf{b}$ takes around 10 seconds, solving ten linear systems $A\mathbf{x}_j = \mathbf{b}_j$ for $j = 1, \ldots, 10$ will take 100 seconds. Justify or refute this statement.

*Answer:*

The claim is false.
The dominant cost in solving a linear system $A\underset{\sim}{x} = \underset{\sim}{b}$ is the LU factorization taking $\frac{2}{3}n^3 + O(n^2)$ flops
Forward + back substitution at $n^2 + O(n)$ flops is insignificant in comparison.
To solve $A\underset{\sim}{x}_j = \underset{\sim}{b}_j$ for $j = 1, \ldots, 10$ you do the LU factorization of $A$ ONCE taking $\approx$ 10 seconds.
Doing forward and back substitution 10 times for each different RHS $\underset{\sim}{b}_j$ is not significant.
Total solution time $\approx$ 10 secs, NOT $10 \times 10 = 100$ secs.

**3.** a) [**4 marks**] Give MATLAB commands for **EITHER** an anonymous function `osc` **OR** a function M-file `osc.m` to calculate

$$f(x) = x \sin\left(\frac{1}{x}\right).$$

Your function should work for an array of inputs `x`, producing an array of output values of the same size.

*Answer:*

Anonymous function:

```
osc = @(x) x .* sin(1./x);
```

or Function M-file in osc.m

```
function f = osc(x)
f = x .* sin(1./x);
```

b) [**6 marks**] Consider the function $f(x) = x^3 - \cos(x)$.

i) Prove that $f$ has at least one zero in the interval $[0, \pi]$

*Answer:*

$f$ is continuous on $\mathbb{R}$ and hence on $[0, \pi]$

$f(0) = 0 - \cos(0) = -1 < 0$

$f(\pi) = \pi^3 - \cos(\pi) = \pi^3 + 1 > 0$

As $f$ is continuous and $f(0)f(\pi) < 0$ (opposite signs)

then $f$ has at least one zero on $[0, \pi]$

ii) Prove that $f$ has at most one zero in the interval $[0, \pi]$

*Answer:*

$f$ is continuously differentiable

$f'(x) = 3x^2 + \sin(x) > 0$ for all $x \in (0, \pi]$

$\therefore f$ is strictly increasing on $[0, \pi]$

$\Rightarrow f$ has at most one zero on $[0, \pi)$

iii) Let `err` be a vector containing the values $e^{(k)} = |x^{(k)} - x^*|$ for $k = 0, 1, \ldots, 10$ produced by an iterative method. The MATLAB commands

```
cv1 = err(2:end) ./ err(1:end-1)
cv2 = err(2:end) ./ err(1:end-1).^2
```

produce

```
cv1 =
  0.8000  0.6400  0.4096  0.1678  0.0281  0.0008  0.0000  0.0000  0.0000
cv2 =
  1.0890  1.1479  1.2500  1.4348  1.7936  2.5735  4.6156 11.8781 54.8251
```

Giving reasons, estimate the rate of convergence.

*Answer:*

CV1 has values of $\dfrac{e^{(k+1)}}{e^{(k)}} \to 0$ as $k$ grows

$\Rightarrow$ rate of convergence is faster than linear (order of convergence $\nu > 1$)

CV2 has values of $\dfrac{e^{(k+1)}}{(e^{(k)})^2}$ which grow as $k$ increases

$\Rightarrow$ order of convergence $\nu < 2$

Thus the rate of convergence is superlinear (order of convergence $\nu$ : $1 < \nu < 2$)

**4.** You are given the results of the following MATLAB commands and the spy plots in Figure 4.1.

```
size(A)
ans =
     121    121
symchk = norm(A-A',1)
symchk =
     4.5056e-016
ev = eig(A);
evmin = min(ev)
evmin =
    0.9916
evmax = max(ev)
evmax =
          5.5605
p = amd(A);
```

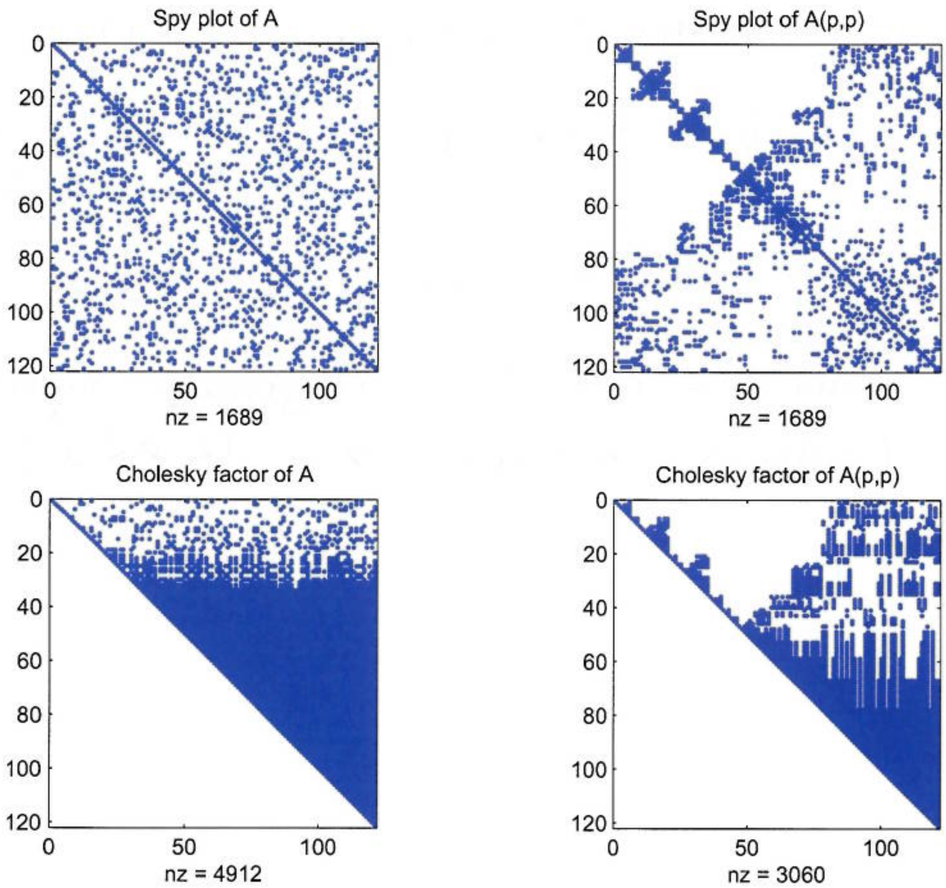


Figure 4.1: Spy plot of $A$ and $A(p,p)$

a) **[2 marks]** A student claims the matrix is not symmetric. Justify or refute this claim.
*Answer:*

$A$ is symmetric $\Leftrightarrow A^T = A \Leftrightarrow \|A - A^T\| = 0$

From the MATLAB output

symchk $= \|A - A^T\|_1 = 4.5 \times 10^{-16} \approx 2\varepsilon$

Thus $A$ is symmetric within the limits of double precision floating point arithmetic.

b) **[2 marks]** Calculate the sparsity of $A$ as a percentage.
*Answer:*

$$\text{Sparsity} = \frac{\text{number of non-zeros in } A}{\text{total number of elements in } A} \times 100\ \%$$

$$= \frac{1689}{121 \times 121} \times 100\ \%$$

$$= 11.5\ \%.$$

Number of non-zeros in $A$ is nz under the spy plot of $A$.

c) **[2 marks]** Calculate the condition number $\kappa(A)$ of $A$.
*Answer:*

For a real symmetric matrix

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{|\lambda_{max}|}{|\lambda_{min}|}$$

From the MATLAB output

$$\kappa_2(A) = \frac{5.5605}{0.9916} \approx 5.6$$

Note : $\kappa(A) \geq 1$ for any norm.

d) **[4 marks]** The elements of the coefficient matrix $A$ are known exactly and the right-hand-side vector $\mathbf{b}$ is known to 6 significant figures.

i) Estimate the relative error in the computed solution to $A\mathbf{x} = \mathbf{b}$.
*Answer:*

$\underset{\sim}{b}$ is known to 6 significant figures
$\Rightarrow$ relative error in $\underset{\sim}{b} \leq \frac{1}{2} \times 10^{-6}$

A "exact" $\Rightarrow$ relative error in A $\approx \varepsilon$

Relative error in computed solution $\underset{\sim}{x}$
$\leq \kappa(A)$ [Rel. error in A + Rel. error in $\underset{\sim}{b}$]
$= 5.6 \left[ \varepsilon + \frac{1}{2} \times 10^{-6} \right]$
$\approx 2.8 \times 10^{-6}$

ii) What confidence do you have in the computed solution?
*Answer:*

$$\text{Rel error} = 2.8 \times 10^{-6} = 0.28 \times 10^{-5} \leq \frac{1}{2} \times 10^{-5}$$

$\Rightarrow$ computed solution has at least 5 significant figures