MATH 2089 JUNE 2009 Numerical Methods

1.

a) i) ans1 = 8 + 3 * eps - 8 = 0
as 8 + x where $|x| < 8$ eps is stored as 8 on a computer

ans2 = exp(3.2e+120) = Inf
as largest number in floating point number system is
realmax $\approx 1.8 \times 10^{308}$, so log(realmax) ≈ 709.8

ii) $A = \begin{bmatrix} 1 & -3 & 0 \\ 2 & 1 & -4 \\ 0 & 0 & 3 \end{bmatrix}$

x = A(:, 2) = $\begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix}$ is the second column of A

ans4 = norm(x, Inf) = 3 as $\|\underset{\sim}{x}\|_\infty = \max_{i=1,\dots,n} |x_i|$

ans5 = A ≥ 0 = $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$

Array of the same size as A, with 1 if $A_{ij} \geq 0$, 0 otherwise

1 b) i) Chemical plant n processes takes $6n^3 + O(n^2)$ flops to simulate

3 GHz dual core, 2 flops/clock cycle
$\Rightarrow$ speed of PC is $3 \times 10^9 \times 2 \times 2 = 1.2 \times 10^{10}$ flops/sec
$\Rightarrow$ 10 minutes = 600 secs = 7.2×10^{12} flops

Size of largest problem: $6n^3 = 7.2 \times 10^{12} \Rightarrow n^3 = 1.2 \times 10^{12}$
(must be integer $n \approx 10{,}600$ OK). $\Rightarrow n = 10{,}627$

ii) The physical memory (RAM) to store all quantities could also be a limiting factor.

MATH 2089 JUNE 2009 Numerical Methods

1 c) i) chk1 = norm (A - A', 1) $\Rightarrow$ $\|A-A^T\|_1 \approx 1.4 \times 10^{-15} \approx 7\varepsilon$

chk1 = 1.4052×10^{-15} where $\varepsilon = 2.2 \times 10^{-16}$ is relative machine precision

$\therefore \|A-A^T\|_1 \approx 0$ to within a small multiple of machine precision

$\therefore A = A^T$, i.e. A is symmetric within rounding error

ii) A is symmetric: A is positive definite $\Leftrightarrow$ <u>all</u> eigenvalues are positive

From Matlab min (ev) = 4.5×10^{-2}

$\Rightarrow$ all eigenvalues $\geq 4.5 \times 10^{-2} > 0$

$\therefore$ A is positive definite

iii) For a real symmetric matrix, the 2-norm condition number

$$\kappa_2(A) = \|A\|_2 \, \|A^{-1}\|_2 = \frac{|\lambda_{max}(A)|}{|\lambda_{min}(A)|} = \frac{9.12 \times 10^4}{4.51 \times 10^{-2}} = 2.0 \times 10^6$$

N.B. $\kappa(A) \geq 1$ for any A and any of $1, 2, \infty$ norms.

iv) $A, \underset{\sim}{b}$ known to 6 significant decimal digits

$\Rightarrow$ rel err (A) $\leq \frac{1}{2} \times 10^{-6}$, rel err ($\underset{\sim}{b}$) $\leq \frac{1}{2} \times 10^{-6}$

$$\begin{aligned}
\text{rel err}(\underset{\sim}{x}) &\approx \kappa_2(A)\left[\text{rel err}(A) + \text{rel err}(\underset{\sim}{b})\right] \\
&= 2 \times 10^6 \left[\tfrac{1}{2} \times 10^{-6} + \tfrac{1}{2} \times 10^{-6}\right] \\
&= 2.
\end{aligned}$$

rel err (x) ≥ 1 $\Rightarrow$ there are <u>NO</u> significant digits in computed $\underset{\sim}{x}$

v) Cholesky factorization $A = R^T R$

Linear system $A\underset{\sim}{x} = \underset{\sim}{b}$ $\Rightarrow$ $R^T R \underset{\sim}{x} = \underset{\sim}{b}$

$\Rightarrow$ $R^T \underset{\sim}{y} = \underset{\sim}{b}$ $\quad R\underset{\sim}{x} = \underset{\sim}{y}$

Solve $R^T \underset{\sim}{y} = \underset{\sim}{b}$ by forward substitution (R^T lower triangular) to get $\underset{\sim}{y}$

Solve $R\underset{\sim}{x} = \underset{\sim}{y}$ by back substitution (R upper triangular) to get $\underset{\sim}{x}$

2. a) $a > 1$

i) $x = a^{\frac{1}{3}} \Leftrightarrow x^3 = a \Leftrightarrow p(x) = x^3 - a = 0$

ii) p is a polynomial in x so is continuous on $\mathbb{R}$

$p(1) = 1 - a < 0$ as $a > 1$

$p(a+1) = (1+a)^3 - a > 1 > 0$ as $a > 1 \Rightarrow (1+a)^3 > 1 + a$

$\therefore$ $p(1)$, $p(a+1)$ have opposite signs and as p is continuous the interval $(1, a+1)$ has at least one zero of p (ie $(1, a+1)$ brackets a zero of p)

iii) $p'(x) = 3x^2 \geq 3 > 0$ for <u>all</u> $x \in (1, a+1)$

$\therefore$ p is strictly increasing on $(1, a+1)$ so p has at most one zero on the interval $(1, a+1)$.

iv) Newton's method:

$$x_{k+1} = x_k - \frac{p(x_k)}{p'(x_k)} = x_k - \frac{(x_k^3 - a)}{3x_k^2}$$

$$= x_k - \frac{1}{3}x_k + \frac{a}{3x_k^2}$$

$$= \frac{2}{3}x_k + \frac{a}{3x_k^2} = \frac{1}{3}\left(2x_k + \frac{a}{x_k^2}\right)$$

v) If $p \in C^2$ and $p'(x^*) \neq 0$ (both true here) then Newtons method converges quadratically (has second order rate of convergence) for x_1 close to x^*.

From the provided data, $e_k = |x^* - x_k|$

$e_k \to 0$ as $k \to \infty$, so $x_k \to x^*$ converges

$e_{k+1}/e_k \to 0$ as $k \to \infty$, so order of convergence $\nu > 1$

$e_{k+1}/e_k^2 \to 0.48$ finite constant as expected for order $\nu = 2$

N.B. Ignored last $e_8/e_7^2 = 1.2 \times 10^3$ as $e_8 \approx 2\varepsilon$ limited by machine precision

$e_{k+1}/e_k^3 \to \infty$ as $k \to \infty$ so order of convergence $\nu < 3$.

2 b) i) IVP standard form

$$\underset{\sim}{x}' = \underset{\sim}{f}(t, \underset{\sim}{x}) \quad t \in (t_0, t_{max}) \quad \underset{\sim}{x}(t_0) = \underset{\sim}{y}_0$$

Time interval: $t_0 = 0$ $t_{max} = 1.5$ minutes (convert 90 sec)

Initial conditions: $\underset{\sim}{x}(0) = \begin{bmatrix} 5 \\ 3 \\ 0 \end{bmatrix} = \begin{bmatrix} \text{Initial concentration chemical 1} \\ \text{Initial concentration chemical 2} \\ \text{Initial concentration of product} \end{bmatrix}$

ODE: $\dfrac{dx_1}{dt} = \dfrac{1}{V}\left(q u_1 - q x_1 - Vr\right)$ where $r = \alpha x_1 x_2$

$$\frac{dx_2}{dt} = \frac{1}{V}\left(q u_2 - q x_2 - Vr\right)$$

$$\frac{dx_3}{dt} = \frac{1}{V}\left(q x_3 + Vr\right)$$

$$\therefore \frac{d\underset{\sim}{x}}{dt} = \begin{bmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \\ \frac{dx_3}{dt} \end{bmatrix} = \underset{\sim}{f}(t, \underset{\sim}{x}) = \begin{bmatrix} \frac{1}{V}(q u_1 - q x_1 - V\alpha x_1 x_2) \\ \frac{1}{V}(q u_2 - q x_2 - V\alpha x_1 x_2) \\ \frac{1}{V}(q x_3 + V\alpha x_1 x_2 \end{bmatrix}$$

ii) MATLAB function M-file reaction.m

```
function f = reaction(t, x)
alpha = 2.6;
V = 2;
u = [3.2; 4.8];
q = 10;
f = [ (q*u(1)-q*x(1) -V*alpha*x(1)*x(2))/V ;
      (q*u(2)-q*X(2) - V*alpha*x(1)*x(2))/V;
      (q*x(3) + V*alpha*x(1)*x(2) ];
```

iii) Explicit Euler: $\underset{\sim}{x}_{n+1} = \underset{\sim}{x}_n + h\,\underset{\sim}{f}(t_n, \underset{\sim}{x}_n)$ uses fixed stepsizes h so $t_n = t_0 + nh$, is $O(h)$ accurate very simple explicit formula
ODE45 uses 4th and 5th order Runge-Kutta methods to find variable stepsize required to achieve requested accuracy.

MATH 2089 JUNE 2009 Numerical Methods

3. a) You are given

space domain $x \in [0, L]$, time domain $t \in [0, T]$

PDE: $\frac{\partial u}{\partial t} = D(x) \frac{\partial^2 u}{\partial x^2}$

Also need: Initial conditions: $u(x, 0) = u_0(x) \quad x \in [0, L]$

Boundary conditions: $u(0, t) = u_0(t) \quad t \in (0, T]$

$u(L, t) = u_L(t) \quad t \in (0, T)$

b) At time step $t_{\ell+1}$ space point x_j with $u_j^{\ell+1} \approx u(x_j, t_{\ell+1})$

i) $$\left.\frac{\partial^2 u(x,t)}{\partial x^2}\right|_{\substack{x=x_j \\ t=t_{\ell+1}}} = \frac{u_{j-1}^{\ell+1} - 2u_j^{\ell+1} + u_{j+1}^{\ell+1}}{(\Delta x)^2} + O(\Delta x^2)$$

ii) $$\left.\frac{\partial u(x,t)}{\partial t}\right|_{\substack{x=x_j \\ t=t_{\ell+1}}} = \frac{u_j^{\ell} - u_j^{\ell+1}}{(-\Delta t)} + O(\Delta t)$$

$$= \frac{u_j^{\ell+1} - u_j^{\ell}}{\Delta t} + O(\Delta t)$$

c) Substitute approximations (ignore $O(\Delta x^2)$, $O(\Delta t)$) into PDE

$$\frac{u_j^{\ell+1} - u_j^{\ell}}{\Delta t} = D(x_j) \left[\frac{u_{j-1}^{\ell+1} - 2u_j^{\ell+1} + u_{j+1}^{\ell+1}}{(\Delta x)^2} \right]$$

Multiply through by Δt, move all terms with superscript $^{\ell+1}$ to left

$$\left[1 + \frac{2D(x_j)\Delta t}{(\Delta x)^2}\right] u_j^{\ell+1} - \frac{D(x_j)\Delta t}{(\Delta x)^2} u_{j-1}^{\ell+1} - \frac{D(x_j)\Delta t}{(\Delta x)^2} u_{j+1}^{\ell+1} = u_j^{\ell}$$

To get desired form divide both sizes by

$$\frac{D(x_j)\Delta t}{(\Delta x)^2}$$

$$\left[2 + \frac{(\Delta x)^2}{D(x_j)\Delta t}\right] u_j^{l+1} - u_{j-1}^{l+1} - u_{j+1}^{l+1} = \frac{(\Delta x)^2}{D(x_j)\Delta t} u_j^{l}$$

Thus $\alpha_j = \dfrac{(\Delta x)^2}{D(x_j)\Delta t}$, $\beta_j^l = \dfrac{(\Delta x)^2}{D(x_j)\Delta t} u_j^l$

N.B. $D(x)$ must be evaluated at x_j

d) $u(0,t) = 40$, $u(L,t) = 20$ $t \in (0,T)$ Boundary conditions

$n = 20$

i) $j=1 \Rightarrow (2+\alpha_1)\, u_1^{l+1} - u_0^{l+1} - u_2^{l+1} = \beta_1^l$

But $j=0 \Rightarrow$ on boundary $x_0 = 0$ $\therefore u_0^{l+1} = 40$, so get

$$(2+\alpha_1)\, u_1^{l+1} - u_2^{l+1} = \beta_1^l + 40$$

ii) $j = 14 \Rightarrow x_j$ is in interior of space domain, so

$$(2+\alpha_{14})\, u_{14}^{l+1} - u_{13}^{l+1} - u_{15}^{l+1} = \beta_{14}^l$$

e) Writing as linear system $A\underset{\sim}{u}^{l+1} = \underset{\sim}{b}^l$

i) sparsity $= \dfrac{\text{number of non-zeros in } A}{\text{product dimensions of } A} \times 100 = \dfrac{58}{20 \times 20} \times 100 = 14.5\%$.

(number of non-zeros, dimensions from spy plot)

ii) A is tridiagonal (lower bandwidth $m_l = 1$, upper bandwidth $m_u = 1$)

$\Rightarrow$ can solve linear system in $O(n)$ flops not $O(n^3)$

A is symmetric as $A_{i,i+1} = A_{i+1,i} = -1$

A is independent of time t_l as $A_{i,i} = 2 + \alpha_i$ does not depend on l so can be factored once not at each time step.

A is positive definite as $A_{ii} = 2 + \alpha_i > |A_{i-1,i}| + |A_{i,i+1}| = 2$ or 1

A is not Toeplitz as $A_{ii} = 2 + \alpha_i$ varies with i.