MASSACHUSETTS INSTITUTE OF TECHNOLOGY
DEPARTMENT OF MECHANICAL ENGINEERING
CAMBRIDGE, MASSACHUSETTS 02139

2.29 NUMERICAL FLUID MECHANICS— SPRING 2007

Solution of Quiz 1

Time 1 hour and 15 min, Totally 25 points
Thursday 11 a.m. 03/22/07, Focused on Lecture 1 to 11

Problem 1 (6 points):

State which of the following statements are true and which are false. You do not have to justify your answer.

1. The number of significant digits achievable by a specific floating point representation is not dependent on exponent length.
2. If $f = ax^2y^{-2}$, then the relative error of f is more sensitive to relative error of x, compared to relative error of y.
3. Bi-section method is capable of predicting the maximum number of iterations required for a specific error level ahead in time.
4. If the Newton-Raphson's method converges for a root finding problem, then the absolute error in each step will be less than the square of absolute error in previous step.
5. The Jacobi iterative method for a linear problem will always converge for a positive definite matrix.
6. The numerical stability of Gaussian elimination is guaranteed provided that we do full pivoting and equilibration.

Solution:

1. **True**. Indeed if the length of mantissa is "t", then it can distinguish up to 2^t states and the number of significant digits will be $\log_{10} 2^t = t\log_{10} 2$ in decimal base.
2. **False**. The relative error in multiplication is dependent on absolute value of power. The sign of power only determines the sign of error.

$$f = ax^2y^{-2}$$

$$\log f = \log a + 2\log x - 2\log y$$

$$\frac{df}{f} = 2\frac{dx}{x} - 2\frac{dy}{y}$$

$$\varepsilon_f = 2\varepsilon_x - 2\varepsilon_y,\ \ \varepsilon: \text{ Relative Error}$$

3. **True**. If the first guesses are x_l, x_u, then the maximum error at the beginning is $|x_l - x_u|$. Also in each step, the maximum error is divided by 2. Consequently for a specific level of absolute error (e), we can find the maximum number of iterations by $\frac{|x_l - x_u|}{2^n} \le e$.
4. **False**. Quadratic convergence means that the error decays such that the relation $|e_{n+1}| \le a(x_r)e_n^2$ holds for a constant "$a = a(x_r)$". In general "a" is not equal to "1" and the statement will be false.
5. **False**. However, the Gauss-Seidel iterative method for a linear problem will always converge for a positive definite matrix.
6. **False**. Refer to notes and recall that pivoting and equilibration will improve the solution accuracy. However if we do not have enough significant digits, compared to matrix condition number the solution will be unstable and inaccurate.

Problem 2 (3 points):

The steady state potential flow past a cylinder is given by below formula.

$$\phi = U_\infty(r + \frac{a^2}{r})\cos\theta$$

Here r, θ are cylindrical coordinates, "a" is cylinder radius and U_∞ is the far field uniform velocity.

1. Derive the expression for the velocity field.
2. Ignore the gravity and derive the expression for the pressure field assuming zero far field pressure.

Solution:

1.

$$\vec{V} = \nabla\phi$$

$$V_r = \frac{\partial\phi}{\partial r} = U_\infty\left(1 - \left(\frac{a}{r}\right)^2\right)\cos\theta$$

$$V_\theta = \frac{1}{r}\frac{\partial\phi}{\partial\theta} = -U_\infty\left(1 + \left(\frac{a}{r}\right)^2\right)\sin\theta$$

2. Gravity term is ignored and steady flow with Bernoulli relation is assumed. It is alos taken into account that every streamline passes through infinity, and there it has zero pressure and U_∞ velocity.

$$const = \frac{1}{2}|\vec{V}|^2 + \frac{P}{\rho}$$

$$\frac{1}{2}U_\infty^2 + 0 = \frac{P}{\rho} + \frac{1}{2}U_\infty^2\{[(1-(\frac{a}{r})^2)\cos\theta]^2 + [U(1+(\frac{a}{r})^2)\sin\theta]^2\}$$

$$\frac{1}{2}U_\infty^2 + 0 = \frac{P}{\rho} + \frac{1}{2}U_\infty^2\{1+(\frac{a}{r})^4 + 2(\frac{a}{r})^2(\sin^2\theta - \cos^2\theta)\}$$

$$P = P(r,\theta) = -\frac{1}{2}\rho U_\infty^2\{(\frac{a}{r})^4 - 2(\frac{a}{r})^2\cos 2\theta\}$$

Problem 3 (2 points):

In a special floating point representation we have 3 bytes, with base 2:

Mantissa length	15 bits
Mantissa sign	1 bit
Exponent length	7 bits
Exponent sign	1 bit

Now answer below questions:

1. What numerical range is covered by this floating point? How many significant digits do we have?
2. What is the smallest nonzero number?
3. What is the largest relative error due to rounding of the mantissa?
4. What is the largest absolute error due to rounding of the mantissa?

Solution:

1. The maximum mantissa is almost "1", because we have a lengthy mantissa [1]. The maximum value of exponent is also $2^7 - 1 = 127$. Consequently the numbers between -2^{127} and $+2^{127} \cong 1.70\times 10^{38}$ are covered. On the other hand the significant digits are governed by mantissa length. The mantissa stores up to 2^{15} = 32768 states, and we will have about 4 to 5 significant digits (more formally $\log_{10}(2^{15}-1) \cong 4.51$ significant digits).
2. Among negative numbers the smallest one is -2^{127}. On the other hand, among positive numbers the smallest number corresponds to $2^{-128}\times 2^{-1} = 2^{-129} \cong 1.47\times 10^{-39}$ [2,3].

[1] The maximum mantissa is $2^{-1} + 2^{-2} + \cdots + 2^{-14} + 2^{-15} = 1 - 2^{-15} \cong 0.99997$.

[2] The answer refers to the conventional case, when by normalizing the largest bit of non-zero mantissa is "1". In this case the minimum non-zero mantissa will be 2^{-1}. If we relax this condition the minimum nonzero mantissa will be 2^{-15} and consequently the minimum nonzero positive number will be $2^{-128}\times 2^{-15} = 2^{-143} \cong 8.96\times 10^{-44}$.

[3] A signed byte covers from -127 to 127. However, with usual sign bias, it will cover from -128 to 127.

3. The maximum error of rounding in mantissa is $\frac{2^{-15}}{2}=2^{-16}$. The minimum mantissa is 2^{-1}. Consequently the maximum relative error due to rounding will be $\frac{2^{-16}}{2^{-1}}=2^{-15}\cong 3.05\times 10^{-5}$. Note that exponent does not play any role in relative error.
4. The maximum error of rounding in mantissa is $\frac{2^{-15}}{2}=2^{-16}$. The maximum exponent is 127. Consequently the maximum absolute error will be $2^{127}\times 2^{-16}=2^{111}\cong 2.60\times 10^{33}$.

Problem 4 (2 points):

Find one of the roots of the following equation with your method of choice.

$$f(x)=x+\frac{1}{2}-\tan x$$

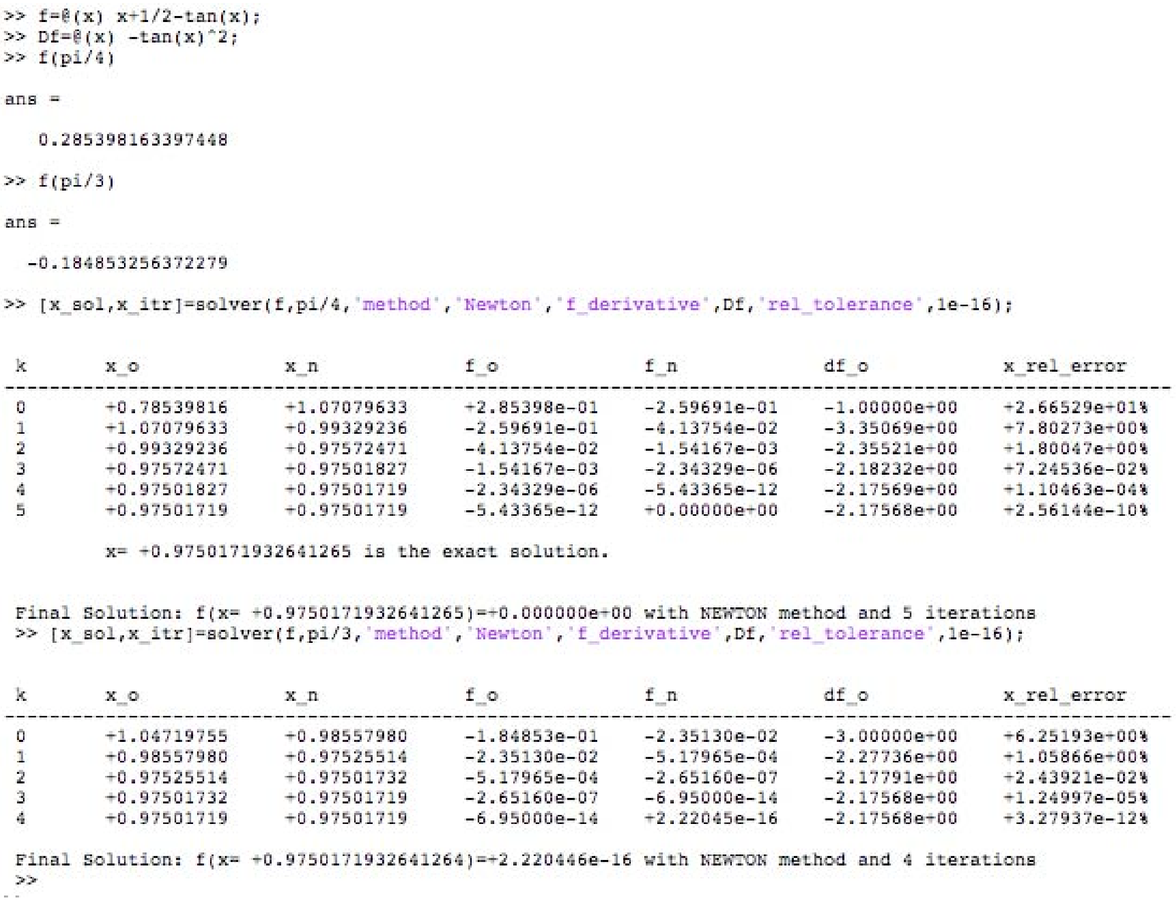
The relative error in the root, between consecutive steps, should be less than 10^{-6}.

Solution:

The roots are the intersection of $x+\frac{1}{2}$ line and $\tan x$. The $\tan x$ is a periodic function with the $(-\infty,+\infty)$ range in each of its continuous periods equal to π. So the equation has infinite number of roots with a distance almost equal to π from each other. So each number has at most a distance almost equal to $\frac{\pi}{2}$ from a root. Consequently any number larger than $\frac{\pi}{2}\times 10^{6}$ is a root with at most 10^{-6} relative error!

Now we focus on the smallest positive root. We know that $f(0)=\frac{1}{2}, f(\frac{\pi}{2})=-\infty$; consequently since the function is continuous in the $(0,\frac{\pi}{2})$ interval it has a root in this interval. Furthermore $f'(x)=1-(1+\tan^2 x)=-\tan^2 x\le 0$ and since it is strictly decreasing in this interval it has just one root in the given interval.

We can use the bisection algorithm to find this root, but since the required error is rather tight it needs about 20 iterations ($\frac{\frac{\pi}{2}}{2^{20+1}}\cong 0.75\times 10^{-6}\le 10^{-6}$). Consequently we use the Newton's method to find the solution accurately and quickly, as can be seen from below print. Indeed starting from either $\frac{\pi}{3}$ or $\frac{\pi}{4}$ we can find the root to be $\mathrm{x_r}\cong 0.97501719$ by at most 6 iterations.



```
>> f=@(x) x+1/2-tan(x);
>> Df=@(x) -tan(x)^2;
>> f(pi/4)

ans =

   0.285398163397448

>> f(pi/3)

ans =

  -0.184853256372279

>> [x_sol,x_itr]=solver(f,pi/4,'method','Newton','f_derivative',Df,'rel_tolerance',1e-16);

 k       x_o             x_n             f_o             f_n             df_o            x_rel_error
----------------------------------------------------------------------------------------------------------
 0       +0.78539816     +1.07079633     +2.85398e-01    -2.59691e-01    -1.00000e+00    +2.66529e+01%
 1       +1.07079633     +0.99329236     -2.59691e-01    -4.13754e-02    -3.35069e+00    +7.80273e+00%
 2       +0.99329236     +0.97572471     -4.13754e-02    -1.54167e-03    -2.35521e+00    +1.80047e+00%
 3       +0.97572471     +0.97501827     -1.54167e-03    -2.34329e-06    -2.18232e+00    +7.24536e-02%
 4       +0.97501827     +0.97501719     -2.34329e-06    -5.43365e-12    -2.17569e+00    +1.10463e-04%
 5       +0.97501719     +0.97501719     -5.43365e-12    +0.00000e+00    -2.17568e+00    +2.56144e-10%

         x= +0.9750171932641265 is the exact solution.

 Final Solution: f(x= +0.9750171932641265)=+0.000000e+00 with NEWTON method and 5 iterations
 >> [x_sol,x_itr]=solver(f,pi/3,'method','Newton','f_derivative',Df,'rel_tolerance',1e-16);

 k       x_o             x_n             f_o             f_n             df_o            x_rel_error
----------------------------------------------------------------------------------------------------------
 0       +1.04719755     +0.98557980     -1.84853e-01    -2.35130e-02    -3.00000e+00    +6.25193e+00%
 1       +0.98557980     +0.97525514     -2.35130e-02    -5.17965e-04    -2.27736e+00    +1.05866e+00%
 2       +0.97525514     +0.97501732     -5.17965e-04    -2.65160e-07    -2.17791e+00    +2.43921e-02%
 3       +0.97501732     +0.97501719     -2.65160e-07    -6.95000e-14    -2.17568e+00    +1.24997e-05%
 4       +0.97501719     +0.97501719     -6.95000e-14    +2.22045e-16    -2.17568e+00    +3.27937e-12%

 Final Solution: f(x= +0.9750171932641264)=+2.220446e-16 with NEWTON method and 4 iterations
 >>
```

Problem 5 (8 points):

Consider the following system of equations:

$$Ax = b, \quad A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 8 & 0 \\ -1 & 0 & 4 \end{bmatrix}, \; b = \begin{bmatrix} 0 \\ 8 \\ 4 \end{bmatrix}$$

1. Cholesky factorize A (Note that A is positive definite).
2. Find an LU factorization form for A.
3. Use LU factorization of A to find x.
4. Compute the x by two iterations of successive over-relaxation scheme. Use relaxation parameter $\omega = 1.5$ and initial guess of zero.

Solution:

1. We need to find the lower triangular matrix L such that $A = LL^*$. However, since A is positive definite the "L" elements are real and we have $A = LL^* = LL^T$. So we need to solve the below equations:

$$Find\ \ L=\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix},\ such\ that\ A=LL^T\ and\ l_{ii}>0$$

$$\begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{11}l_{21} & l_{22}^2+l_{21}^2 & l_{21}l_{31}+l_{22}l_{32} \\ l_{11}l_{31} & l_{21}l_{31}+l_{22}l_{32} & l_{33}^2+l_{32}^2+l_{31}^2 \end{bmatrix}=\begin{bmatrix} 1 & 2 & -1 \\ 2 & 8 & 0 \\ -1 & 0 & 4 \end{bmatrix}$$

The above equation can be solved very easily:

$$\begin{aligned} &l_{11}=1,\quad l_{21}=2,\quad l_{31}=-1 \\ &l_{22}=\sqrt{8-l_{21}^2}=2 \\ &l_{32}=-\frac{l_{21}l_{31}+0}{l_{22}}=-\frac{2\times -1}{2}=1 \\ &l_{33}=\sqrt{4-l_{32}^2-l_{31}^2}=\sqrt{2} \end{aligned}$$

So we have:

$$L=\begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ -1 & 1 & \sqrt{2} \end{bmatrix}$$

Alternatively we could use the below formula

$$l_{i,j}=\frac{1}{l_{j,j}}\left(a_{i,j}-\sum_{k=1}^{j-1} l_{i,k}l_{j,k}\right), \qquad \text{for } i>j$$

$$l_{i,i}=\sqrt{a_{i,i}-\sum_{k=1}^{i-1} l_{i,k}^2}.$$

2. The Cholesky decomposition is already a "LU" factorization form so[4]:

$$L=\begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ -1 & 1 & \sqrt{2} \end{bmatrix}, U=L^T=\begin{bmatrix} 1 & 2 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & \sqrt{2} \end{bmatrix}$$

[4] Otherwise we can use the Gaussian Elimination and find L and U accordingly:

$$L=\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 0.5 & 1 \end{bmatrix}, U=\begin{bmatrix} 1 & 2 & -1 \\ 0 & 4 & 2 \\ 0 & 0 & 2 \end{bmatrix}$$

3. We have to find y such that $Ly = b$ and then we have to find x such that $Ux = y$:

$$Ly = b \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ -1 & 1 & \sqrt{2} \end{bmatrix} y = \begin{bmatrix} 0 \\ 8 \\ 4 \end{bmatrix} \Rightarrow y = \begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix}$$

$$Ux = y \Rightarrow \begin{bmatrix} 1 & 2 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & \sqrt{2} \end{bmatrix} x = \begin{bmatrix} 0 \\ 4 \\ 0 \end{bmatrix} \Rightarrow x = \begin{bmatrix} -4 \\ 2 \\ 0 \end{bmatrix}$$

4. Since the matrix is positive definite, we do not need to impose diagonally dominant condition. As a result we can use the below format for Gauss-Seidel:

$$Ax = b \Rightarrow \begin{cases} x_1 = -2x_2 + x_3 \\ x_2 = -\dfrac{x_1}{4} + 1 \\ x_3 = \dfrac{x_1}{4} + 1 \end{cases}$$

Also for the relaxation we have:

$x_i^{(n+1)} = (1-\omega)x_i^{(n)} + \omega\bar{x}_i^{(n+1)}$, *where* $\bar{x}_i^{n+1}$ *is the* "$n+1$" *iterate on* x_i *computed by Gauss* $-$ *Seidel from x*

So we have:

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\bar{x}_1^{(1)} = -2 \times 0 + 0 = 0 \Rightarrow x_1^{(1)} = (1-1.5) \times 0 + 1.5 \times 0 = 0$$

$$\bar{x}_2^{(1)} = -\frac{0}{4} + 1 = 1 \quad \Rightarrow x_2^{(1)} = (1-1.5) \times 0 + 1.5 \times 1 = 1.5$$

$$\bar{x}_3^{(1)} = +\frac{0}{4} + 1 = 1 \quad \Rightarrow x_3^{(1)} = (1-1.5) \times 0 + 1.5 \times 1 = 1.5$$

$$x^{(1)} = \begin{bmatrix} 0 \\ 1.5 \\ 1.5 \end{bmatrix}$$

$$\bar{x}_1^{(2)} = -2 \times 1.5 + 1.5 = -1.5 \quad \Rightarrow x_1^{(2)} = (1-1.5) \times 0 + 1.5 \times -1.5 \quad = -2.25$$

$$\bar{x}_2^{(2)} = -\frac{-2.25}{4} + 1 \quad = 1.5625 \Rightarrow x_2^{(2)} = (1-1.5) \times 1.5 + 1.5 \times 1.5625 = 1.59375$$

$$\bar{x}_3^{(2)} = +\frac{-2.25}{4} + 1 \quad = 0.4375 \Rightarrow x_3^{(2)} = (1-1.5) \times 1.5 + 1.5 \times 0.4375 = -0.09375$$

$$x^{(2)} = \begin{bmatrix} -2.25 \\ 1.59375 \\ -0.09375 \end{bmatrix}$$

Problem 6 (4 points):

Consider the below (x,y) pairs:

$$x = \begin{bmatrix} -2 \\ 0 \\ 1 \\ 2 \end{bmatrix}, \quad y = f(x) = \begin{bmatrix} 2 \\ 0 \\ 1 \\ -2 \end{bmatrix}$$

1. Find the Lagrange polynomial for above points.
2. Interpolate that polynomial at x=-1.
3. Find the ordered polynomial for above points with Newton's formula.
4. Interpolate the ordered polynomial at x=-1.

Solution:

1.

$$\begin{aligned} L(x) = 2\times\frac{(x-0)(x-1)(x-2)}{(-2-0)(-2-1)(-2-2)} + 0\times\frac{(x-(-2))(x-1)(x-2)}{(0-(-2))(0-1)(0-2)} \\ +1\times\frac{(x-(-2))(x-0)(x-2)}{(1-(-2))(1-0)(1-2)} - 2\times\frac{(x-(-2))(x-0)(x-1)}{(2-(-2))(2-0)(2-1)} \end{aligned}$$

$$L(x) = -\frac{x(x-1)(x-2)}{12} - \frac{(x+2)x(x-2)}{3} - \frac{(x+2)x(x-1)}{4}$$

$$L(x) = \frac{-2x^3+5x}{3}$$

2.

$$L(-1) = \frac{2-5}{3} = -1$$

3.

$$\begin{array}{lllll}
x & f(x) & & & \\
-2 & 2 & & & \\
 & & \dfrac{0-2}{0-(-2)}=-1 & & \\
 & & & \dfrac{1-(-1)}{1-(-2)}=\dfrac{2}{3} & \\
0 & 0 & & & \\
 & & \dfrac{1-0}{1-0}=1 & & \dfrac{(-2)-\frac{2}{3}}{2-(-2)}=-\dfrac{2}{3} \\
1 & 1 & & & \\
 & & & \dfrac{-3-1}{2-0}=-2 & \\
 & & \dfrac{-2-1}{2-1}=-3 & & \\
2 & -2 & & &
\end{array}$$

$$N(x)=2+(-1)\times(x-(-2))+\frac{2}{3}\times(x-(-2))x+(-\frac{2}{3})\times(x-(-2))x(x-1)$$

$$N(x)=2-1\times(x+2)+\frac{2}{3}\times(x+2)x-\frac{2}{3}\times(x+2)x(x-1)$$

$$N(x)=\frac{-2x^3+5x}{3}$$

4. Note that L(x)=N(x) and they pass from the same 4 pairs of points. Indeed both Newton's scheme and Lagrange' scheme refer to the same ordered polynomial and they are both two different method to find the same polynomial.

$$N(-1)=L(-1)=-1$$