CS3911 Introduction to Numerical Methods with Fortran Exam 2 Solutions

1. **Accuracy and Reliability**

(a) **[20 points]** Let ϵ be a very small positive number (*i.e.*, $\epsilon \approx 0$) but $1/\epsilon$ will not cause overflow. Consider the following system of linear equations:

$$\begin{bmatrix} \epsilon & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

First, solve this system of linear equations *with* **and** *without* partial pivoting using finite precision arithmetics. Then, discuss the major reason or reasons that can explain the difference(s) between the two solutions. **You should provide a general argument rather than an argument based on a fixed precision. A convincing and to-the-point argument is required. As a result, just stating a "reason" such as "**it is because of cancelation**" or "**overflow**" will receive zero point. Note also that "prove-by-example" is not acceptable.**

Solution: Without pivoting, one multiplies $-1/\epsilon$ to the first equation and adds the result to the second. This yields the following:

$$\begin{bmatrix} \epsilon & 1 \\ 0 & 1 - 1/\epsilon \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 2 - 1/\epsilon \end{bmatrix}$$

Since ϵ is small, $1/\epsilon$ is large. As a result, -$1/\epsilon$ is the dominating term in both $1 - 1/\epsilon$ and $2 - 1/\epsilon$. In other word, the equation $(1 - 1/\epsilon)y = 2 - 1/\epsilon$ would numerically become

$$\left(-\frac{1}{\epsilon}\right) y \approx -\frac{1}{\epsilon}$$

Backward substitution yields $y = 1$. Plugging $y = 1$ into the first equation $\epsilon x + y = 1$ yields $x = 0$.

With pivoting, the system becomes the following after a row swap:

$$\begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Multiplying the first equation by $-\epsilon$ and adding the result to the second yields:

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 - \epsilon \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 - 2\epsilon \end{bmatrix}$$

Since ϵ is very small, it does not contribute much to $1 - \epsilon$ and $1 - 2\epsilon$. As a result, $1 - \epsilon \approx 1$ and $1 - 2\epsilon \approx 1$, and the second equation $(1 - \epsilon)y = 1 - 2\epsilon$ numerically becomes $y = 1$. Plugging $y = 1$ into the first equation gives $x = 1$.

Obviously, $x = y = 1$ is the correct numerical solution if ϵ is small. The reason for the non-pivoting solution to go wrong is the rounding error in computing $1 - 1/\epsilon$ and $2 - 1/\epsilon$ if $1/\epsilon$ is very large. In this case, rounding error makes both terms nearly equal to $-1/\epsilon$ numerically. Consequently, we have $y = 1$ which is still correct. But, since ϵ is very small, $x = 1 - \epsilon$ will have rounding error again. This time, the impact of ϵ is so small that becomes insignificant compared with 1. Therefore, $x = 0$!

For example, on a 7-digit computer, if $\epsilon = 0.00000001$, which is not a very small number, we have $1/\epsilon = 100,000,000$. Then, $1 - 1/\epsilon = -99,999,999$ and $2 - 1/\epsilon = -99,999,998$. Both would be rounded to 7 digits and the result is 0.1×10^9.
This example shows that pivoting is necessary. ∎

2. **Linear Algebra**

(a) **[5 points]** Suppose a matrix A has the following LU-decomposition using 2 row swaps and 3 column swaps. Compute the determinant of A. **You should show all computation steps. Only providing an answer and/or using a wrong method receives zero point.**

$$L = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ 1 & -1 & 1 & \\ 0 & 1 & 2 & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 1 & 0 & 2 \\ & 2 & 1 & 0 \\ & & 3 & 1 \\ & & & 4 \end{bmatrix}$$

Solution: The determinant is the product of all diagonal entries if no pivoting is used. If pivoting is used, this product should be multiplied by $(-1)^{(\text{no. of row and column swaps})}$. Since the product is $24 = 1 \times 2 \times \times 3 \times 4$ and since the total number of row and column swaps is 5, the desired answer is $-24 = (-1)^5 \times (24)$. ∎

(b) **[15 points]** Find the LU-decomposition of the following matrix without pivoting. **You should show all computation steps. Only providing an answer and/or using a wrong method receives zero point.**

$$A = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 2 & 2 & 1 & -2 \\ 0 & 4 & 5 & 0 \\ -1 & 0 & 6 & 5 \end{bmatrix}$$

Solution: A 4×4 lower triangular matrix L has the following form, where $\times$ denotes a value to be determined in the process of Gaussian elimination.

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \times & 1 & 0 & 0 \\ \times & \times & 1 & 0 \\ \times & \times & \times & 1 \end{bmatrix}$$

Multiplying -2, 0 and 1 to the first row of A and adding the results to the second, third and fourth rows, respectively, we have L and U as follows:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & \times & 1 & 0 \\ -1 & \times & \times & 0 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 2 & 1 & 0 \\ 0 & 4 & 5 & 0 \\ 0 & 0 & 6 & 4 \end{bmatrix}$$

The multipliers (*i.e.*, 2, 0 and -1) are saved to the first column of L with opposite signs.
Then, multiplying -2 and 0 to the second row of A (or U above) and adding the results to the third and fourth rows, respectively, yields:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ -1 & 0 & \times & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 6 & 4 \end{bmatrix}$$

The multipliers (*i.e.*, 2 and 0) are saved to the second column of L with opposite signs.

Finally, multiplying the third row by -2 and adding the result to the fourth, and saving the multiplier with opposite sign to column 3 of L yields:

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ -1 & 0 & 2 & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

The above L and U are the desired results. ∎

(c) **[12 points]** Use Gauss-Seidel method to solve the following system of linear equations, and fill the table below with your results. The initial value (*i.e.*, iteration 0) is $x = y = z = 0$, and you only do two iterations (*i.e.*, iterations 1 and 2).

$$\begin{aligned} 3x + y - z &= 6 \\ -x + 4y + 2z &= 10 \\ x + y + 6z &= 16 \end{aligned}$$

Solution: The solution is shown below:

Iteration	x	y	z
0	0	0	0
1	2	3	$\frac{11}{6} = 1.8333333\ldots$
2	$\frac{29}{18} = 1.6555555\ldots$	$\frac{143}{72} = 1.9861111\ldots$	$\frac{893}{432} = 2.0671296\ldots$

First, the equations must be transformed to the following:

$$x = \frac{1}{3}(6 - y + z) \qquad (1)$$

$$y = \frac{1}{4}(10 + x - 2z) \qquad (2)$$

$$z = \frac{1}{6}(16 - x - y) \qquad (3)$$

- **Iteration 1**:
  - Since the initial values are $x = 0$, $y = 0$ and $z = 0$, Equation (1) gives the new $x = (6 - 0 + 0)/3 = 2$.
  - Now we have $x = 2$, $y = 0$ and $z = 0$, they are used in Equation (2) to compute the new $y = (10 + 2 - 2 \times 0)/4 = 3$.
  - This gives $x = 2$, $y = 3$ and $z = 0$. They are used in Equation (3) to compute the new $z = (16 - 2 - 3)/6 = 11/6$. This completes the first iteration.
- **Iteration 2**:
  - Iteration 2 starts with $x = 2$, $y = 3$ and $z = 11/6$, the new x from Equation (1) is $x = (6 - 3 + \frac{11}{6})/3 = \frac{29}{18}$.

- Now we have $x = \frac{29}{18}$, $y = 3$ and $z = \frac{11}{6}$, Equation (2) gives $y = (10+\frac{29}{18}-2\times\frac{11}{6})/4 = \frac{143}{72}$.
- So far we have $x = \frac{29}{18}$, $y = \frac{143}{72}$ and $z = \frac{11}{6}$. They are used in Equation (3) to compute the new $z = (16 - \frac{29}{18} - \frac{143}{72})/6 = \frac{893}{432}$. This completes the second iteration.

At the end of the second iteration, we have $x = \frac{29}{18}$, $y = \frac{143}{72}$ and $z = \frac{893}{432}$. ∎

(d) **[15 points]** Suppose a program read in the following system of linear equations:

$$A\cdot x = B \quad \text{where } A = \begin{bmatrix} 1 & 2 \\ -1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 4 \\ 0 \end{bmatrix} \quad \text{and} \quad A = L\cdot U = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 0 & 4 \end{bmatrix}$$

and computed $x = 3$ and $y = 2$. This solution is inaccurate. Use the iterative refinement method to improve the accuracy of this "solution." **You have to show <u>all</u> computation steps using the given LU-decomposition, and explain how you get the results. Otherwise (*e.g.*, only providing an answer and/or asking me to guess your intention from a bunch of numbers), you will receive <u>zero</u> point.**

<u>**Solution:**</u> The following shows all computation steps:

- <u>**Compute the error vector r**</u>:

$$r = B - A \cdot X = \begin{bmatrix} 4 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} -3 \\ -1 \end{bmatrix}$$

- <u>**Forward substitution to find T in $L \cdot T = r$**</u>:
The equation of $L \cdot T = r$ is the following:

$$\begin{bmatrix} 1 & \\ -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} -3 \\ -1 \end{bmatrix}$$

Forward substitution gives $t_1 = -3$. Plugging $t_1 = -3$ into the second equation $-t_1+t_2 = -1$ yields $t_2 = -4$.

- <u>**Backward substitution to find Δ in $U \cdot \Delta = T$**</u>:
The equation of $U \cdot \Delta = T$ is shown below:

$$\begin{bmatrix} 1 & 2 \\ & 4 \end{bmatrix} \cdot \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} = \begin{bmatrix} -3 \\ -4 \end{bmatrix}$$

Backward substitution gives $\delta_2 = -1$. Plugging $\delta_2 = -1$ into the first equation $\delta_1+2\delta_2 = -3$ yields $\delta_1 = -1$. Therefore, $\Delta = [\delta_1, \delta_2]^T = [-1, -1]^T$.

- <u>**Compute the new X**</u>:
The new X is computed as $X + \Delta$: new$X = [3, 2]^T + [-1, -1]^T = [2, 1]^T$.
- <u>**Verify the computed result**</u>:
Since $B - A \cdot [2, 1]^T = [0, 0]^T$, we have computed the correct solution to the system of linear equations. ∎

3. **Polynomial Interpolation**

(a) **[10 points]** Given a polynomial of degree n as follows,

$$P_n(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n$$

Develop a way to evaluate $P_n(x)$ with n multiplications. **You should provide an algorithm and its complexity analysis. A method that does not achieve $O(n)$ receives <u>zero</u> point.**

<u>**Solution:**</u> The given polynomial can be rewritten in a *nested* form as follows:

$$P_n(x) = a_0 + (a_1 + (a_2 + (a_3 + (\cdots(a_{n-1} + a_nx)x)\cdots)x)x)$$

For example, $P_4(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 = a_0 + (a_1 + (a_2 + (a_3 + a_4x)x)x)x$. In this way, only n multiplications are needed (*i.e.*, $O(n)$). On the other hand, since a_ix^i requires i multiplications, the total number of multiplications with a direct evaluation of the *power* form is $0 + 1 + 2 + \cdots + n = n(n+1)/2$ (*i.e.*, $O(n^2)$).

The following is a possible algorithm, which assumes that coefficients a_i's are in array `a( )`, x is in variable `x`, and `Px` has the result:

```
Px = a(n)
DO i = n-1, 0, -1
  Px = a(i) + Px*x
END DO
```

Since this `DO` loop iterates n times, each of which uses exactly one multiplication, the order of complexity is $O(n)$.

One may suggest the following $O(n)$ implementation:

```
Px    = a(0)
Power = x
DO i = 1, n
  Px    = Px + a(i)*Power
  Power = Power*x
END DO
```

Although it is an $O(n)$ method, it uses two multiplications per iteration and the total number of multiplications is $2n$. This is certainly slower than the nested form. ∎

(b) **[15 points]** Find the Lagrange interpolating polynomial for the data points $(x_0, y_0) = (-2, 0)$, $(x_1, y_1) = (0, 4)$ and $(x_2, y_2) = (2, 0)$. **You should show <u>all</u> computation steps. Only providing an answer and/or using a wrong method receives <u>zero</u> point.**

<u>**Solution:**</u> The degree 2 Lagrange interpolating polynomial $P_2(x)$ is

$$P_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}y_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}y_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}y_2$$

Since $y_0 = y_2 = 0$, the above immediately reduces to the following:

$$P_2(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}y_1$$

Plugging $x_0 = -2$, $x_1 = 0$, $x_2 = 2$ and $y_1 = 4$ into the above yields:

$$P_2(x) = \frac{(x-(-2))(x-2)}{(0-(-2))(0-2)}4 = -(x+2)(x-2)$$

Hence, the desired degree 2 Lagrange interpolating polynomial is $P_x(2) = -(x+2)(x-2)$ ∎

(c) **[8 points]** Add a new data point $(x_3, y_3) = (3, -5)$ to this interpolating polynomial. That is, use this newly available data point to **update** the interpolating polynomial obtained in the previous problem. **You should show all computation steps. Only providing an answer and/or using a wrong method receives zero point. You will also receive zero point if you do not use the update technique.**

Solution: Since the degree increases from 2 to 3, one more term is needed:

$$\text{new term} = \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} y_3$$

Since $x_3 = 3$ and $y_3 = -5$, the new term is

$$\text{new term} = -\frac{1}{3}(x + 2)x(x - 2)$$

Since only one term (*i.e.*, the y_1 term) has to be updated, we have

$$\text{new } y_1 \text{ term} = (\text{old } y_1 \text{ term}) \times \frac{x - x_3}{x_1 - x_3} = \frac{1}{3}(x + 2)(x - 2)(x - 3)$$

Finally, the new degree 3 Lagrange interpolating polynomial $P_3(x)$ is

$$P_3(x) = \frac{1}{3}(x + 2)(x - 2)(x - 3) - \frac{1}{3}(x + 2)x(x - 2)$$