Numerical Analysis — Solutions for sheet 2

LU decomposition, bisection method

1. (a) The augmented matrix for the system given is

$$\tilde{A} = \left[\begin{array}{rrrr} 1 & -4 & 1 & 0 \\ -3 & 2 & 3 & 0 \\ 4 & -1 & 6 & 1 \end{array}\right].$$

(i) with no pivoting we perform the 1st Gaussian elimination step directly to give

$$\tilde{A} = \left[\begin{array}{rrrr} 1 & -4 & 1 & 0 \\ 0 & -10 & 6 & 0 \\ 0 & 15 & 2 & 1 \end{array}\right].$$

After the next step we have

$$\tilde{A} = \left[\begin{array}{rrrr} 1 & -4 & 1 & 0 \\ 0 & -10 & 6 & 0 \\ 0 & 0 & 11 & 1 \end{array}\right]$$

and solutions are $\mathbf{x} = (\frac{7}{55}, \frac{3}{55}, \frac{1}{11})^T$.]

(ii) with partial pivoting we first swap rows 1 and 3

$$\left[\begin{array}{rrrr} 4 & -1 & 6 & 1 \\ -3 & 2 & 3 & 0 \\ 1 & -4 & 1 & 0 \end{array}\right]$$

and then step 1 of Gaussian elimination is

$$\left[\begin{array}{rrrr} 4 & -1 & 6 & 1 \\ 0 & \frac{5}{4} & \frac{15}{2} & \frac{3}{4} \\ 0 & -\frac{15}{4} & -\frac{1}{2} & -\frac{1}{4} \end{array}\right];$$

The next steps involve swapping rows 2 and 3 (which has the largest leading column entry) and then doing another Gaussian elimination which results in

$$\left[\begin{array}{rrrr} 4 & -1 & 6 & 1 \\ 0 & -\frac{15}{4} & -\frac{1}{2} & -\frac{1}{4} \\ 0 & 0 & \frac{22}{3} & \frac{2}{3} \end{array}\right];$$

(iii) with scaled partial pivoting, we first scale each row by the largest matrix element in the row

$$\tilde{A} = \left[\begin{array}{rrrr} -\frac{1}{4} & 1 & -\frac{1}{4} & 0 \\ -1 & \frac{2}{3} & 1 & 0 \\ \frac{2}{3} & -\frac{1}{6} & 1 & \frac{1}{6} \end{array}\right]$$

Then we see that partial pivoting requires we swap rows 1 and 2 to give

$$\tilde{A} = \left[\begin{array}{rrrr} -1 & \frac{2}{3} & 1 & 0 \\ -\frac{1}{4} & 1 & -\frac{1}{4} & 0 \\ \frac{2}{3} & -\frac{1}{6} & 1 & \frac{1}{6} \end{array}\right]$$

and now we can do step 1 of Gaussian elimination

$$\tilde{A} = \begin{bmatrix} -1 & \frac{2}{3} & 1 & 0 \\ 0 & \frac{5}{6} & -\frac{1}{2} & 0 \\ 0 & \frac{5}{18} & \frac{5}{3} & \frac{1}{6} \end{bmatrix}.$$

The next steps involve rescaling rows 2 and 3 by the largest matrix entry which gives

$$\tilde{A} = \begin{bmatrix} -1 & \frac{2}{3} & 1 & 0 \\ 0 & 1 & -\frac{3}{5} & 0 \\ 0 & \frac{1}{6} & 1 & \frac{1}{10} \end{bmatrix}.$$

and noting we do not need to swap rows, so Gaussian step 2 gives

$$\tilde{A} = \begin{bmatrix} -1 & \frac{2}{3} & 1 & 0 \\ 0 & 1 & -\frac{3}{5} & 0 \\ 0 & 0 & \frac{11}{10} & \frac{1}{10} \end{bmatrix}.$$

(b) Everything above is done with exact fractions. I have also made all these computations with 2-digit precision arithmetic (e.g. $\frac{2}{3} = 0.67$ and after every arithmetic computation we round to 2 digits). I find the resulting numerical solutions using: no pivoting method (i) to be $\mathbf{x} = (0.091, 0.055, 0.13)^T$; using partial pivoting method (ii) to be $\mathbf{x} = (0.11, 0.053, 0.085)^T$; and using scaled partial pivoting method (iii) to be $\mathbf{x} = (0.091, 0.055, 0.12)^T$.

The exact answer (rounded to 2 digits) is $\mathbf{x} = (0.091, 0.055, 0.13)^T$. Here, the most accurate method, contrary to what I tell you in lectures, is without pivoting. The reason is that this is a poor example to showcase this since method (i) doesn't require any numerical rounding during the Gaussian elimination, only during back substitution. Method (ii) and (iii) happens to involve lots of calculations which require rounding. What we note is that scaled partial pivoting is almost exact.

2. Starting from the matrix

$$A = \begin{bmatrix} -1 & 1 & 1 \\ 2 & -1 & 1 \\ 1 & 1 & 2 \end{bmatrix},$$

the row operations $R_2 \to R_2 - (-2)R_1$ and $R_3 \to R_3 - (-1)R_1$ lead to

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 2 & 3 \end{bmatrix}.$$

The final row operation $R_3 \to R_3 - 2R_2 \to$ results in

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 2 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & -3 \end{bmatrix}.$$

This is the required LU decomposition of the matrix A. We can now solve the system $A\mathbf{x} = LU\mathbf{x} = \mathbf{b}$ by setting $U\mathbf{x} = \mathbf{y}$ and solving first $L\mathbf{y} = \mathbf{b}$ by forward substitution and then $U\mathbf{x} = \mathbf{y}$ by backward substitution. The three cases for the vector $\mathbf{b}$ are

$$\mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{array}{lll} y_1 & = 1 & = 1 \\ y_2 & = 0 + 2y_1 & = 2 \\ y_3 & = 0 + y_1 - 2y_2 & = -3 \end{array} \quad \begin{array}{lll} x_3 & = -y_3/3 & = 1 \\ x_2 & = y_2 - 3x_3 & = -1 \\ x_1 & = -y_1 + x_2 + x_3 & = -1 \end{array} \quad \mathbf{x} = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \begin{array}{lll} y_1 & = 0 & = 0 \\ y_2 & = 1 + 2y_1 & = 1 \\ y_3 & = 0 + y_1 - 2y_2 & = -2 \end{array} \quad \begin{array}{lll} x_3 & = -y_3/3 & = 2/3 \\ x_2 & = y_2 - 3x_3 & = -1 \\ x_1 & = -y_1 + x_2 + x_3 & = -1/3 \end{array} \quad \mathbf{x} = \begin{bmatrix} -1/3 \\ -1 \\ 2/3 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \begin{array}{lll} y_1 & = 0 & = 0 \\ y_2 & = 0 + 2y_1 & = 0 \\ y_3 & = 1 + y_1 - 2y_2 & = 1 \end{array} \quad \begin{array}{lll} x_3 & = -y_3/3 & = -1/3 \\ x_2 & = y_2 - 3x_3 & = 1 \\ x_1 & = -y_1 + x_2 + x_3 & = 2/3 \end{array} \quad \mathbf{x} = \begin{bmatrix} 2/3 \\ 1 \\ -1/3 \end{bmatrix}$$

The three results provide the column vectors of the inverse matrix A^{-1} which has the form

$$A^{-1} = \begin{bmatrix} -1 & -\frac{1}{3} & \frac{2}{3} \\ -1 & -1 & 1 \\ 1 & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}.$$

3. The pivot element a_{11} for the row operations of the first step in the LU-decomposition for the matrix A is zero. Hence the first row has to be interchanged either with row 2 or with row 3. (In partial pivoting we would swap with row 2 since it has the bigger element in the first column). Below we list both possible solutions.

   If we swap row 1 with row 2 then we obtain after the further row operations $R_3 \to R_3 - \frac{1}{2}R_1$ and $R_3 \to R_3 + 2R_2$ the following decomposition

$$PA = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 4 & 2 & 3 \\ 2 & -1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 3 \\ 0 & 1 & 2 \\ 2 & -1 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & -2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & \frac{13}{2} \end{bmatrix} = LU.$$

   If we swap row 1 with row 3 then we obtain after the further row operations $R_2 \to R_2 - 2R_1$ and $R_3 \to R_3 - \frac{1}{4}R_2$ the following decomposition

$$PA = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2 \\ 4 & 2 & 3 \\ 2 & -1 & 4 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 4 \\ 4 & 2 & 3 \\ 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 4 \\ 0 & 4 & -5 \\ 0 & 0 & \frac{13}{4} \end{bmatrix} = LU.$$

4. Simple: $P^{-1} = (P_1P_2)^{-1} = P_2^{-1}P_1^{-1} = P_2^TP_1^T = (P_1P_2)^T = P^T$ using standard results for matrices.

5. Perfect candidate for a proof by contradiction. So assume $A = L_1U_1 = L_2U_2$ where $\operatorname{diag}\{L_i\} = (1, 1, \ldots, 1)^T$ for $i = 1, 2$ and $L_1 \neq L_2$, $U_1 \neq U_2$. It follows that

$$L_2^{-1}L_1 = U_1^{-1}U_2$$

   Now, from Q5, we infer U_1^{-1} is upper triangular and the product $U_1^{-1}U_2$ is upper triangular also. We similarly infer that $L_2^{-1}L_1$ is lower triangular. Now L_2^{-1} must have ones on its diagonal, otherwise $L_2^{-1}L_2$ would not be I the Identity, since L_2 has ones on it's diagonal. Therefore

$$L_2^{-1}L_1 = I = U_1^{-1}U_2$$

   and so

$$L_1 = L_2, \qquad \text{and} \qquad U_1 = U_2$$

   which is a contradiction.

6. (a) Since A is non-singular, we can write it as $A = LU$ where L is lower triangular with $\text{diag}\{L\} = (1, 1, \ldots, 1)^T$ and U is upper triangular with $\text{diag}\{U\} = (u_{11}, u_{22}, \ldots, u_{nn})^T$. We can write

$$U = D\tilde{U}$$

where $D = \text{diag}\{u_{ii}\}$ is a diagonal matrix with entries $(D)_{ii} = u_{ii}$ and it follows that

$$(\tilde{U})_{ij} = u_{ij}/u_{ii}$$

remains an upper triangular matrix but with $\text{diag}\{U\} = (1, 1, \ldots, 1)^T$. Now, $A = LD\tilde{U}$ and since A is symmetric $A = A^T$ and so

$$LD\tilde{U} = (LD\tilde{U})^T = \tilde{U}^T DL^T.$$

Using the fact that $\tilde{U}^T$ is lower triangular with ones along the diagonal and that the LU-decomposition is unique (Q3) it must be that $\tilde{U}^T = L$ and so $A = LDL^T$ as required.

(b) We have

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T LDL^T\mathbf{x} = (L^T\mathbf{x})^T D(L^T\mathbf{x}) = \sum_{i=1}^{n} l_i^2 u_{ii}$$

where we have written the ith component of $(L^T\mathbf{x})_i = l_i$. This can only be positive for all $\mathbf{x}$ if $u_{ii} > 0$ for all i.

(c) If u_{ii}, the diagonal elements of D are all positive, then we can write, from (a) $A = L\sqrt{D}\sqrt{D}L^T = QQ^T$ where $Q = L\sqrt{D}$. Here $\sqrt{D}$ is the diagonal matrix with entries $\sqrt{u_{ii}}$.

Note: This decomposition is known as a “Cholesky decomposition”.

7. (a) The function $f(x) = x^3 - x - 1/4$ is a cubic polynomial and has at most 3 real roots. The curve of $x^3 - x$ is easy to plot: it goes through -1,0,1. Subtracting 1/4 from this moves the curve down by 1/4 and we can see from an overlaid sketch that the root at +1 will move right and the two roots at -1,0 will moved towards each other (and possibly vanish !)

It is helpful to locate possible maxima and minima of the function. The first two derivatives are $f'(x) = 3x^2 - 1$ and $f''(x) = 6x$. The first derivative vanishes if $x = \pm 1/\sqrt{3}$. Furthermore $f''(1/\sqrt{3}) > 0$ and $f''(-1/\sqrt{3}) < 0$. We conclude that:

$f(x)$ has a maximum at $x = -1/\sqrt{3}$ where $f(-1/\sqrt{3}) = (8 - 3\sqrt{3})/(12\sqrt{3}) > 0$.

$f(x)$ has a minimum at $x = 1/\sqrt{3}$ where $f(1/\sqrt{3}) = (-8 - 3\sqrt{3})/(12\sqrt{3}) < 0$.

We try some other values of x and find $f(-1) < 0$, $f(1) < 0$, and $f(2) > 0$. Hence there is one sign change in each of the intervals $[-1, -1/\sqrt{3}]$, $[-1/\sqrt{3}, 1/\sqrt{3}]$, and $[1, 2]$. This shows that the polynomial has indeed three roots, and one of them is in the interval $[1, 2]$.

(b) The starting interval $[1, 2]$ has length one, and after n steps of the bisection method the maximal error is 2^{-n}. We require $2^{-n} < 10^{-4}$. Taking the logarithm of this relation results in

$$n > \frac{4\ln(10)}{\ln(2)} \approx 13.29\,.$$

We conclude that 14 steps of the bisection method guarantee that the error is smaller than 10^{-4}.

(c) The required intervals were already given in part (a) where it was found that each of the intervals $[-1, -1/\sqrt{3}]$ and $[-1/\sqrt{3}, 1/\sqrt{3}]$ contains exactly one root.

8. The following table shows the first five steps of the bisection method for finding the root of $f(x) = x^2 - 2 = 0$ in the interval $[1, 2]$.

n	a_n	x_n	b_n	$f(a_n)$	$f(x_n)$	$f(b_n)$	max. err.
1	1.00000	1.50000	2.00000	-1.00	0.25	2.00	2^{-1}
2	1.00000	1.25000	1.50000	-1.00	-0.44	0.25	2^{-2}
3	1.25000	1.37500	1.50000	-0.44	-0.11	0.25	2^{-3}
4	1.37500	1.43750	1.50000	-0.11	0.07	0.25	2^{-4}
5	1.37500	1.40625	1.43750	-0.11	-0.02	0.07	2^{-5}

We conclude that the root has to be in the interval $[1.40625, 1.43750]$, and we can specify only the first two digits of the root after five iterations of the method. This shows that the convergence of the method is quite slow.

9. Since $\tanh(x) < 1$ the two curves of $y = \tanh(x)$ and $y = \mu/x$ intersect at $x = x^*$ when $\mu/x^* < 1$ implying $x^* > \mu$. We also know that $\tanh(x) < x$ for $x > 0$. The root x^* is to the right of the point where $y = \mu/x$ intersects $y = x$ which is when $x = \sqrt{\mu}$. I.e. $x^* > \sqrt{\mu}$ as well as $x^* > \mu$. So we have established two lower bounds and, to make Bisection as efficient as possible we want the largest lower bound. So if $\mu \leq 1$, this is $a = \sqrt{\mu}$ and if $\mu > 1$ this is $a = \mu$.

An upper bound is harder. We can also use the definition of tanh to write

$$\tanh(x) = \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x + \mathrm{e}^{-x}} > \frac{\mathrm{e}^x - \mathrm{e}^{-x}}{\mathrm{e}^x} = 1 - \mathrm{e}^{-2x}$$

and, for $x > 0$, $\mathrm{e}^{-2x} < 1/x$ so that $\tanh(x) > 1 - 1/x$. The intersection of the lines $y = \mu/x$ with $\tanh(x)$ is therefore to the left of the intersection of the lines $y = \mu/x$ with $y = 1 - 1/x$ which occurs at $x = \mu + 1$.

10. (a) After one step of Gaussian elimination the elements of $A^{(1)}$ are, for $i, j = 2, \ldots, n$

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}.$$

We are asked to show that

$$|a_{ii}^{(1)}| \geq \sum_{\substack{j=2 \\ \neq i}}^{n} |a_{ij}^{(1)}|$$

for $i = 2, \ldots, n$. In other words we need to show

$$\left| a_{ii} - \frac{a_{i1}}{a_{11}} a_{1i} \right| \geq \sum_{\substack{j=2 \\ \neq i}}^{n} \left| a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \right| .$$

Start with the RHS, and use $|a + b| \leq |a| + |b|$ to get

$$\sum_{\substack{j=2 \\ \neq i}}^{n} \left| a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \right| \leq \sum_{\substack{j=2 \\ \neq i}}^{n} |a_{ij}| + \frac{|a_{i1}|}{|a_{11}|} \sum_{\substack{j=2 \\ \neq i}}^{n} |a_{1j}|.$$

Now since row diagonal dominance of A implies both

$$\sum_{\substack{j=2\\ \neq i}}^{n} |a_{ij}| + |a_{i1}| \leq |a_{ii}|, \qquad \text{and} \qquad \sum_{\substack{j=2\\ \neq i}}^{n} |a_{1j}| + |a_{1i}| \leq |a_{11}|,$$

(recalling that $i \geq 2$) then we have

$$\sum_{\substack{j=2\\ \neq i}}^{n} \left| a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \right| \leq |a_{ii}| - |a_{i1}| + \frac{|a_{i1}|}{|a_{11}|} \left(|a_{11}| - |a_{1i}| \right) = |a_{ii}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}|$$

But $|a| - |b| \leq |a - b|$ and so

$$|a_{ii}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| \leq \left| a_{ii} - \frac{a_{i1}}{a_{11}} a_{1i} \right|$$

and we are done.

(b) If one step of Gaussian elimination preserves diagonal dominance, as we have shown, then it is preserved through every step. This means the fully reduced upper triangular matrix U is diagonally dominant.