Numerical Analysis **Solutions for sheet 4**

Aitken's Δ^2 Method, Newton-Raphson in higher dimensions, Lagrange interpolating polynomials

---

1. The fixed point iteration is $x_{n+1} = \frac{1}{2}\sqrt{10 - x_n^3}$. Aitken's Δ^2 method uses the sequence of the x_n to define a new sequence in the form

$$\hat{x}_n = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} = x_n - \frac{(x_{n+1} - x_n)^2}{(x_{n+2} - 2x_{n+1} + x_n)}\,.$$

The following table shows the numerical results for the two sequences. The starting point is $x_0 = 1.5$. One can see that the sequence obtained from Aitken's method converges much

n	x_n	$\hat{x}_n$
0	1.5000000	1.361886
1	1.2869538	1.364329
2	1.4025408	1.364999
3	1.3454584	1.365169
4	1.3751703	1.365214
5	1.3600942	
6	1.3678470	

quicker to the exact solution $x^* = 1.3652300$.

2. Newton's method for a system $\mathbf{f}(\mathbf{x}) = 0$ consists of the iteration

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} - J^{-1}(\mathbf{x}^{(m)})\mathbf{f}(\mathbf{x}^{(m)}),$$

where $J(\mathbf{x})$ is the Jacobian matrix. Its matrix elements are $J_{ij} = \partial f_i/\partial x_j$. Let A be a non singular $n \times n$ matrix. The linear system $A\mathbf{x} = \mathbf{b}$ can be written in the form $\mathbf{f}(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = 0$. Alternatively, we can write it in component form as

$$f_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} A_{ij}x_j - b_i = 0\,, \qquad i = 1, \ldots, n\,,$$

where A_{ij} are the matrix elements of A, and x_i and b_i are the components of $\mathbf{x}$ and $\mathbf{b}$, respectively. The Jacobian matrix J for this system has matrix elements

$$J_{ij}(\mathbf{x}) = \frac{\partial f_i(\mathbf{x})}{\partial x_j} = A_{ij}\,.$$

One sees that the Jacobian matrix J is identical to the matrix A, and it does not dependent on $\mathbf{x}$. Let us denote the initial point for Newton's method by $\mathbf{x}^{(0)}$. One step of Newton's method results in

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - J^{-1}(\mathbf{x}^{(0)})\,\mathbf{f}(\mathbf{x}^{(0)}) = \mathbf{x}^{(0)} - A^{-1}(A\mathbf{x}^{(0)} - \mathbf{b}) = A^{-1}\mathbf{b}\,.$$

This is the exact solution of the linear system $A\mathbf{x} = \mathbf{b}$.

3. The system of equations is

$$f(x,y) = ax^2 + by + c = 0\,, \qquad g(x,y) = dx + e = 0\,,$$

from which we obtain the solution as $x^* = -e/d$ and $y^* = -ae^2/(bd^2) - c/b$.

The Jacobian matrix J for this system and its inverse J^{-1} are given by

$$J = \begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} 2ax & b \\ d & 0 \end{bmatrix}, \qquad J^{-1} = \frac{1}{(-bd)} \begin{bmatrix} 0 & -b \\ -d & 2ax \end{bmatrix}.$$

One step of Newton's method with initial point (x_0, y_0) results in

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \frac{1}{(-bd)} \begin{bmatrix} 0 & -b \\ -d & 2ax_0 \end{bmatrix} \begin{bmatrix} ax_0^2 + by_0 + c \\ dx_0 + e \end{bmatrix} = \frac{1}{bd} \begin{bmatrix} -be \\ adx_0^2 - cd + 2aex_0 \end{bmatrix}.$$

The second step in Newton's iteration can be obtained from this result by replacing x_1 and y_1 by x_2 and y_2, and also x_0 and y_0 by x_1 and y_1. We find

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \frac{1}{bd} \begin{bmatrix} -be \\ adx_1^2 - cd + 2aex_1 \end{bmatrix} = \begin{bmatrix} -e/d \\ -ae^2/(bd^2) - c/b \end{bmatrix},$$

where we inserted the value $x_1 = -e/d$ from the previous equation. This agrees with the exact solution given before.

4. Here we consider the system

$$f(x,y) = x^2 - y^2 = 0\,, \qquad g(x,y) = 1 + xy = 0\,.$$

From the first equation we obtain $y = \pm x$ and from the second equation $y = -1/x$. These relations are only compatible if we chose the negative sign in the first relation: $y = -x$. Then we obtain from the second relation $x^2 = 1$. We conclude that the solutions are $(x,y) = (1,-1)$ and $(x,y) = (-1,1)$. The Jacobian matrix J for this system and its inverse J^{-1} are given by

$$J = \begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} 2x & -2y \\ y & x \end{bmatrix}, \qquad J^{-1} = \frac{1}{2(x^2+y^2)} \begin{bmatrix} x & 2y \\ -y & 2x \end{bmatrix}.$$

The initial point is $(x_0, y_0) = (\alpha, \alpha)$, and one step of Newton's method results in

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix} - \frac{1}{4\alpha^2} \begin{bmatrix} \alpha & 2\alpha \\ -\alpha & 2\alpha \end{bmatrix} \begin{bmatrix} 0 \\ 1 + \alpha^2 \end{bmatrix} = \frac{\alpha^2 - 1}{2\alpha} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

This shows that if one iteration point is on the line $y = x$ then the following iteration point is also on the line $y = x$. Hence the iteration can never converge to one of the two solutions because they are not on this line.

5. (a) The first relation you can plot as $y = -\mathrm{e}^x$ and the second you can plot as $y = \ln(x)$ and where they intersect will represent a root of the two non-linear equations. Then you see that $\ln(x)$ is less than zero for $0 < x < 1$ and in this range $-\mathrm{e} < y < -1$ so this defines a bounding box in which the root will lie.

   (b) The Newton scheme is defined in the notes and we need the Jacobian first

$$J = \begin{pmatrix} \mathrm{e}^x & 1 \\ -1 & \mathrm{e}^y \end{pmatrix}$$

   and then its inverse

$$J^{-1} = \frac{1}{1+\mathrm{e}^{x+y}} \begin{pmatrix} \mathrm{e}^y & -1 \\ 1 & \mathrm{e}^x \end{pmatrix}$$

   So the Newton iteration step is

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{1+\mathrm{e}^{x_n+y_n}} \begin{pmatrix} \mathrm{e}^{y_n} & -1 \\ 1 & \mathrm{e}^{x_n} \end{pmatrix} \begin{pmatrix} \mathrm{e}^{x_n} + y_n \\ \mathrm{e}^{y_n} - x_n \end{pmatrix}$$

   where $\mathbf{x}_n = (x_n, y_n)^T$. Or

$$x_{n+1} = x_n - (\mathrm{e}^{y_n+x_n} + y_n\mathrm{e}^{y_n} - \mathrm{e}^{y_n} + x_n)/(1+\mathrm{e}^{x_n+y_n})$$

   and

$$y_{n+1} = y_n - (\mathrm{e}^{x_n} + y_n + \mathrm{e}^{x_n+y_n} - x_n\mathrm{e}^{x_n})/(1+\mathrm{e}^{x_n+y_n})$$

   (c) With $(x_0, y_0) = (1, -1)$ we have, upon substituting in

$$x_1 = \mathrm{e}^{-1}, \qquad y_1 = -1$$

   Next we have

$$x_2 = \mathrm{e}^{-1} - (\mathrm{e}^{-1+1/\mathrm{e}} - \mathrm{e}^{-1})/(1+\mathrm{e}^{-1+1/\mathrm{e}}) = 0.2610638\ldots = 0.26106$$

   to five digit precision. Also

$$y_2 = -1.29035\ldots = -1.2904$$

   rounded to five digits.

6. There are three parts of the question.

   (a) We use the formula in the notes for the Lagrange polynomial with $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ and $f(x) = \mathrm{e}^x$:

$$P_2(x) = \frac{(x-1)(x-2)}{2}\mathrm{e}^0 + \frac{x(x-2)}{-1}\mathrm{e}^1 + \frac{x(x-1)}{2}\mathrm{e}^2$$

   and find we can simply to

$$P_2(x) = 1 + \frac{1}{2}x^2(1-\mathrm{e})^2 - \frac{1}{2}x(\mathrm{e}-3)(\mathrm{e}-1)$$

(b) We use the formula for the error in the notes

$$E = \max_{0\le x\le 2} |f(x) - P_2(x)| \le \max_{0\le x\le 2} \frac{|f'''(x)|}{3!} \max_{0\le x\le 2} \left| \prod_{i=0}^{n} (x - x_i) \right|$$

We can find the max of $f'''(x) = \mathrm{e}^x$ which is e^2. Next we let

$$W(x) = \prod_{i=0}^{n} (x - x_i) = x(x-1)(x-2) = x^3 - 3x^2 + 2x$$

and to find maximum values we differentiate to give

$$W'(x) = 3x^2 - 6x + 2$$

Setting this to zero gives a quadratic with roots

$$x = \frac{6 \pm \sqrt{36 - 24}}{6} = 1 \pm \sqrt{3}3 = 1 \pm 1/\sqrt{3}$$

These two maxima are both in the interval $0 \le x \le 2$ so we need to calculate $W(x)$ at both values (or do we ?). Then

$$W(1 + 1/\sqrt{3}) = \left(1 + \frac{1}{\sqrt{3}}\right) \frac{1}{\sqrt{3}} \left(1 - \frac{1}{\sqrt{3}}\right) = -\frac{2\sqrt{3}}{9}$$

and

$$W(1 - 1/\sqrt{3}) = \frac{2\sqrt{3}}{9}$$

(why ? because the cubic is odd about $x = 1$). So the maximum W is $2\sqrt{3}$ and the bound on the error is

$$E \le \frac{\mathrm{e}^2\sqrt{3}}{27}$$

(c) The actual maximum error require a calculation of

$$E = \max_{0\le x\le 2} |f(x) - P_2(x)|$$

and

$$f(x) - P_2(x) = \mathrm{e}^x - 1 - \frac{1}{2}x^2(1-\mathrm{e})^2 + \frac{1}{2}x(\mathrm{e}-3)(\mathrm{e}-1)$$

setting its derivative to zero gives

$$\mathrm{e}^x - x(1-\mathrm{e})^2 + \frac{1}{2}(\mathrm{e}-3)(\mathrm{e}-1) = 0$$

and we are given the two roots, both in the interval $0 \le x \le 2$. When we take $x = 0.448304$ we find the modulus of $f - P_2$ is $0.16046\ldots$ and for $x = 1.60644$ we find $f - P_2 = -0.21346\ldots$ and so the maximum error is the 0.21346. The bound on the error in (ii) evaluates to 0.47401 and so we are inside that bound.

7. (a) Just have to use the formula in the notes, and $x_0 = 0$, $x_1 = 1$, $x_2 = 2$ with $f(x_0) = 1$, $f(x_1) = 2$, $f(x_2) = 4$. We get

$$P_2(x) = \frac{(x-1)(x-2)}{(-1)(-2)}1 + \frac{(x-0)(x-2)}{(1)(-1)}2 + \frac{(x-0)(x-1)}{(2)(1)}4.$$

Simplifying gives

$$P_2(x) = \frac{1}{2}(x-1)(x-2) - 2x(x-2) + 2x(x-1) = \frac{x^2}{2} + \frac{x}{2} + 1.$$

(b) The error is defined (see notes) by

$$|f(x) - P_2(x)| = \frac{1}{6}|f'''(\xi)||x(x-1)(x-2)|$$

for some $\xi \in (0,2)$. First, from $f(x) = 2^x$, we have $f'(x) = \ln(2)2^x$, and so $f'''(\xi) = (\ln(2))^3 2^\xi$. The maximum value of $|f'''(\xi)|$ for $\xi \in (0,2)$ is therefore $4(\ln(2))^3$. Also, we need to find a bound on the second term $|x(x-1)(x-2)|$ and this requires

$$\frac{d}{dx}\left(x(x-1)(x-2)\right) = 0$$

which gives

$$3x^2 - 6x + 2 = 0.$$

Solving gives $3(x-1)^2 - 1 = 0$ and so $x = 1 \pm 1/\sqrt{3}$ at which

$$x(x-1)(x-2) = \pm\left(1 \pm 1/\sqrt{3}\right)\frac{1}{\sqrt{3}}\left(-1 \pm 1/\sqrt{3}\right) = \pm\frac{1}{\sqrt{3}}\left(\frac{1}{3} - 1\right) = \mp\frac{2}{3\sqrt{3}}.$$

We conclude, therefore, that $|x(x-1)(x-2)| < 2/3\sqrt{3}$. Putting everything together, we have that

$$|f(x) - P_2(x)| \le \frac{1}{6} \times 4(\ln(2))^3 \times \frac{2}{3\sqrt{3}} \approx 0.085.$$

8. Note that L and U are defined differently to in the LU-decomposition (see question).

(a) Since $A = L + U$ and $(L+U)\mathbf{x}^* = \mathbf{b}$ it follows from the iterative scheme that

$$L\mathbf{x}^{(k+1)} - (L+U)\mathbf{x}^* = \mathbf{b} - U\mathbf{x}^{(k)} - \mathbf{b}$$

and

$$L(\mathbf{x}^{(k+1)} - \mathbf{x}^*) = -U(\mathbf{x}^{(k)} - \mathbf{x}^*).$$

(b) This is harder. We approach this by induction. First, we have from the top line of the equation in (a)

$$a_{11}\mathbf{e}_1^{(k+1)} = -\sum_{j=2}^{n} a_{1j}\mathbf{e}_j^{(k)}$$

and so

$$|a_{11}\mathbf{e}_1^{(k+1)}| = \left|\sum_{j=2}^{n} a_{1j}\mathbf{e}_j^{(k)}\right|$$

and then we have

$$|a_{11}||\mathbf{e}_1^{(k+1)}| \leq \sum_{j=2}^{n} |a_{1j}||\mathbf{e}_j^{(k)}|$$

but, according to the definition, $|\mathbf{e}_j^{(k)}| \leq \|\mathbf{e}^{(k)}\|_\infty$ so it follows that

$$|a_{11}||\mathbf{e}_1^{(k+1)}| \leq \|\mathbf{e}^{(k)}\|_\infty \sum_{j=2}^{n} |a_{1j}| < \|\mathbf{e}^{(k)}\|_\infty |a_{11}|$$

using diagonal dominance. And so $|\mathbf{e}_1^{(k+1)}| < \|\mathbf{e}^{(k)}\|_\infty$. Now assume $|\mathbf{e}_j^{(k+1)}| < \|\mathbf{e}^{(k)}\|_\infty$ for $j = 1, \ldots, i-1$, say, for some $i \geq 2$. Then the ith row in result in (a) reads

$$\sum_{j=1}^{i-1} a_{ij}\mathbf{e}_j^{(k+1)} + a_{ii}\mathbf{e}_i^{(k+1)} = -\sum_{j=i+1}^{n} a_{ij}\mathbf{e}_j^{(k)}$$

which we rearrange to

$$a_{ii}\mathbf{e}_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij}\mathbf{e}_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\mathbf{e}_j^{(k)}.$$

Then it follows as above that

$$|a_{ii}||\mathbf{e}_i^{(k+1)}| \leq \sum_{j=1}^{i-1} |a_{ij}||\mathbf{e}_j^{(k+1)}| + \sum_{j=i+1}^{n} |a_{ij}||\mathbf{e}_j^{(k)}|.$$

Using the assumption, we have

$$|a_{ii}||\mathbf{e}_i^{(k+1)}| < \|\mathbf{e}^{(k)}\|_\infty \sum_{\substack{j=1 \\ \neq i}}^{n} |a_{ij}|$$

and it follows after using diagonal dominance that

$$|\mathbf{e}_i^{(k+1)}| < \|\mathbf{e}^{(k)}\|_\infty.$$

By induction, we have shown that it is true for $i = 1$ and true for i assuming true for all $j = 1, \ldots, i-1$ and therefore true for all $1 \leq i \leq n$. Therefore $\|\mathbf{e}^{(k+1)}\|_\infty < \|\mathbf{e}^{(k)}\|_\infty$.

Remark: Direct methods for solving linear systems such as LU-decomposition find the answer in a finite number of steps, and typically require $O(n^3)$ floating point operations. Iterative methods such as Gauss-Seidel only approximate solutions which, under diagonally dominant conditions, will convergence but require a stopping condition. However, each iteration only requires $O(n^2)$ floating point operations and so, if the number of iterations required is less than $O(n)$, they can be quicker than direct methods. In particular, if the matrix A is *sparse* (that is, they are mainly full of zeros) then the number of operations required is reduced further (no need to multiply by zero in your algorithm !).