Project 3 Numerical Methods

Preben Hast Sørli

April 30, 2020

1 Problem 1 Nonlinear equations

We shall reformulate the following equation into a fixed-point equation

$$e^{-x} - \arccos(2x) = 0 \tag{1}$$

Proof. By definition arccos only takes values $[-1, 1]$, so we restrict $x \in [-\frac{1}{2}, \frac{1}{2}]$. For fixed-point iteration we need an expression of the form $x = g(x)$, we rewrite 1:

$$\begin{aligned} & e^{-x} - \arccos(2x) = 0 \\ \Leftrightarrow \quad & e^{-x} = \arccos(2x) \\ \Leftrightarrow \quad & \cos\big(e^{-x}\big) = 2x \\ \Leftrightarrow \quad & \frac{1}{2}\cos\big(e^{-x}\big) = x \end{aligned}$$

giving us $x = g(x) = \frac{1}{2}\cos(e^{-x})$ as desired. It is clear that $g(x)$ is a real-valued continuous function (it is the composition of the two continous functions e^x, $\cos(x)$), hence, showing $g(x)$ is a contraction on $[-\frac{1}{2}, \frac{1}{2}]$ would imply it satisfies the contraction mapping theorem.(Thm. 1.3 in Suli & Meyers[1]
We need to prove the existence of a $L \in (0, 1)$ such that

$$|g(x) - g(y)| \leq L|x - y| \quad \forall \quad x, y \in [-\frac{1}{2}, \frac{1}{2}]. \tag{2}$$

By the mean value theorem we have

$$\forall x, y \quad \exists \quad c \in [x, y] \text{ s.t. } \frac{g(x) - g(y)}{x - y} = g'(c)$$

$$\begin{aligned} \Rightarrow \quad & |g(x) - g(y)| = |g'(c)||x - y| \\ \Rightarrow \quad & |\cos\big(e^{-x}\big) - \cos\big(e^{-y}\big)| = |\sin\big(e^{-x}\big)e^{-x}||x - y| \\ \Rightarrow \quad & |\cos\big(e^{-x}\big) - \cos\big(e^{-y}\big)| \leq |e^{-x}||x - y| \\ \Rightarrow \quad & |\cos\big(e^{-x}\big) - \cos\big(e^{-y}\big)| \leq |e^{\frac{1}{2}}||x - y| \\ \Rightarrow \quad & |g(x) - g(y)| \leq \frac{e^{\frac{1}{2}}}{2}|x - y| \end{aligned}$$

which proves 1 by choosing any L from $(\frac{e^{\frac{1}{2}}}{2}, 1)$. The last step comes from the fact that e^x is strictly increasing, while the preceding steps are rewrites of the implications of the mean value theorem.
This concludes the proof. □

2 Problem 2 Numerical linear algebra

a) We choose the following matrix for our singular value decomposition ($A = U\Sigma V^T$)

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

Before we continue, we note that we are deliberately making the task a little easier by choosing a symmetric matrix for this task as diagonalizibility of A enables some of the computation normally done when calculating SVDs redundant.
Now, let's calculate A^TA

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

This gives us

$$\begin{aligned} det(A^TA - \lambda I) = det \begin{bmatrix} 5 & 0 & -4 \\ 0 & 1 & 0 \\ -4 & 0 & 5 \end{bmatrix} &= (5-\lambda)(-\lambda+1)(-\lambda+5) - 4 \cdot 4(-\lambda+1) \\ &= -\lambda^3 + 11\lambda^2 - 19\lambda + 9 \\ &= -(\lambda - 9)(\lambda - 1)^2 \end{aligned}$$

Which has roots $\lambda_1 = 1$ and $\lambda_2 = 9$, hence, we have singular values $\sigma_1 = \sqrt{1} = 1$ and $\sigma_2 = \sqrt{9} = 3$. We will find the corresponding eigenvectors by row reduction, and normalize them:

$$\begin{bmatrix} 5-\lambda_1 & 0 & -4 \\ 0 & 1-\lambda_1 & 0 \\ -4 & 0 & 5-\lambda_1 \end{bmatrix} \sim \begin{bmatrix} 4 & 0 & -4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow v_1 = \begin{bmatrix} 1/2 \\ 1/\sqrt{2} \\ 1/2 \end{bmatrix}$$

$$\begin{bmatrix} 5-\lambda_2 & 0 & -4 \\ 0 & 1-\lambda_2 & 0 \\ -4 & 0 & 5-\lambda_2 \end{bmatrix} \sim \begin{bmatrix} -4 & 0 & -4 \\ 0 & -8 & 0 \\ 0 & 0 & 0 \end{bmatrix} \Rightarrow v_2 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix}$$

We must find a third vector v_3 completing an orthornormal basis for $\mathbb{R}^3$.
Let

$$v_3 = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}$$

Then we have

$$\begin{aligned} \langle v_2, v_3 \rangle &= \frac{\alpha}{\sqrt{2}} \qquad - \frac{\gamma}{\sqrt{2}} = 0 \Longrightarrow \alpha = \gamma \\ \langle v_1, v_3 \rangle &= \frac{\alpha}{2} + \frac{\beta}{\sqrt{2}} + \frac{\gamma}{2} = 0 \Longrightarrow \frac{\beta}{\sqrt{2}} = -\alpha \\ \Longrightarrow \quad v_3 &= \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{2} \end{bmatrix} \end{aligned}$$

Note that v_3 is also an eigenvector corresponding to $\lambda_1 = 1$. This set of orthonormal vectors gives the columns of the matrix V and further V^T in the SVD for A:

$$V = \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{2}} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{\sqrt{2}} & \frac{1}{2} \end{bmatrix} = V^T \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{2}} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{\sqrt{2}} & \frac{1}{2} \end{bmatrix}$$

Here V turned out to be symmetric, which is extremely convenient. We have $VV^T = I$ by construction, which now implies $V = V^{-1} = U$. Our Σ is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

note that the double appearance of 1 comes from the algebraic multiplicity of $\lambda_1 = 1$ being 2. For different reasons one usually want the singular values to appear in descending order on the diagonal, but I was lazy and sacrificed that construction for the convenience of a symmetric V matrix. Putting it all together, we have

$$A = U\Sigma V^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{2}} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{\sqrt{2}} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{2}} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{\sqrt{2}} & \frac{1}{2} \end{bmatrix}.$$

b) See the Jupyter notebook.

c) Let $A \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^n$. A typical least square problem will be:

$$\text{Find } x^* \text{ s.t. } \|Ax^* - b\|_2 = \min_{x\in\mathbb{R}^m} \|Ax - b\|_2. \tag{4}$$

We know of a few different ways to solve such a problem.

(1) Normal equations:[1] From linear algebra we can guarantee the existence of the vector x^* minimizing $\|Ax - b\|_2$ and that it is of the form $x^* = \left(A^TA\right)^{-1} A^Tb$ if A is of rank n. What we would call the normal equation would in this case be

$$A^TAx^* = A^Tb. \tag{5}$$

As we can see, the normal equation is not hard to find, as calculating A^T is cheap, but we would be at an immediate disadvantage if our matrix A is not of rank n. Another disadvantage with using normal equations appear when we have large condition numbers $\mathcal{K}_2(A)$ as $\mathcal{K}_2(A^TA)$ will grow with approximately the square of $\mathcal{K}_2(A)$ making the precision of our calculations increasingly problematic for matrices with large condition numbers.

(2) QR-factorization:[2] QR-factroization is the decomposition of A into the product $A = QR$, where Q orthonormal and R upper triangular. Unlike normal equations, QR-factorization don't rely as much on rank and invertibility, and it's precision does not suffer equally much when $\mathcal{K}_2(A)$ is large. Nevertheless it is not always as easy to do the factorization as simply transposing and computing A^TA as the factorization oftentimes will rely on orthogonalization. Orthogonalization can be problematic to do numerically because our inner products will usually compute to very small numbers close to zero, but seldom

[1] Fridberg, Insel, Spence, 2014, page 362, 2

[2] Suli & Meyers, 2006, page 78, 1

exactly zero. This unfortunately makes Gram-Schmidt an unstable algorithm, but it's fast and simple which is one of the advantages of QR-decomposition.

There are other factorization methods which are more stable, but these won't run as fast as the ones using Gram-Schmidt.

3 Problem 3 Condition numbers

Might replace this proof with one using the characteristic polynomials instead of diagonalizability.
The spectral theorem tells us that any normal matrix, and then especially any symmetric matrix, is diagonalizable. We get

$$A^n = PD^nP^{-1} \quad \forall \quad n \in \mathbb{Z}.$$

Hence, there exists a bijection between the eigenvalues of A and A^n by $\lambda \mapsto \lambda^n$.

$$\begin{aligned} \mathcal{K}_2 &:= \|A\|_2\|A^{-1}\|_2 \\ &= \rho(A^TA)^{\frac{1}{2}}\rho((A^{-1})^TA^{-1}) \end{aligned}$$

Because our matrix is symmetric and the inverse of a symmetric matrix is also symmetric, we get

$$A^TA = A^2, \qquad (A^{-1})^TA^{-1} = A^{-2}.$$

This gives us

$$\rho(A^TA)^{\frac{1}{2}}\rho((A^{-1})^TA^{-1})^{\frac{1}{2}} = \rho(A^2)^{\frac{1}{2}}\rho(A^{-2})^{\frac{1}{2}}$$

$$\begin{aligned} &= \left(\max_{\lambda\in\sigma(A^2)}|\lambda|\right)^{\frac{1}{2}} \cdot \left(\max_{\lambda\in\sigma(A^{-2})}|\lambda|\right)^{\frac{1}{2}} \\ &= \left(\max_{\lambda\in\sigma(A)}|\lambda^2|\right)^{\frac{1}{2}} \cdot \left(\max_{\lambda\in\sigma(A)}|\lambda^{-2}|\right)^{\frac{1}{2}} \\ &= \max_{\lambda\in\sigma(A)}|\lambda| \cdot \max_{\lambda\in\sigma(A)}|\lambda^{-1}| \\ &= \frac{\max_{\lambda\in\sigma(A)}|\lambda|}{\min_{\lambda\in\sigma(A)}|\lambda|} \end{aligned}$$

4 Problem 4

Generalization of the preceding problem. Writing A as SVD or using Jordan Blocks should help with identifying the eigenvalues in a similar manner as in the preceding problem.

Lemma 4.1. Let X, Y be two $n \times n$-matrices. Then XY and YX has the same eigenvalues.

Proof. Let $\lambda = 0$ be an eigenvalue of XY, then

$$0 = det(XY) = det(X)det(Y) = det(YX) = 0, \tag{6}$$

so 0 is an eigenvalue of YX. Hence, we assume $\lambda \neq 0$ with $\vec{v}$ it's corresponding eigenvector. Then $Y\vec{v} \neq 0$ and further

$$\lambda Y\vec{v} = Y(XY\vec{v}) = (YX)Y\vec{v}, \tag{7}$$

which means $Y\vec{v}$ is an eigenvector for YX with the same λ as it's eigenvalue. □

By definition the singular values of a positive, real matrix A are just the square roots of the eigenvalues of $B = A^TA$, and by Theorem 2.9 (Suli & Meyers), if λ_i are the eigenvalues of B, then $\|A\|_2 = \max_{i=1}^n \lambda_i^{1/2}$ or equivalently the biggest singular value of A, $\sigma_{\max}$. Now, the condition number $\mathcal{K}_2(A) := \|A^{-1}\|_2\|A\|_2$, so we need to find $\|A^{-1}\|_2$. Again, by theorem 2.9, we have

$$\begin{aligned}
\|A^{-1}\|_2 &\overset{\textbf{by }2.9}{=} \max_{\lambda\in\sigma(A^{-1^T}A^{-1})} \sqrt{\lambda} \\
&\overset{\textbf{by }4.1}{=} \max_{\lambda\in\sigma((A^TA)^{-1})} \sqrt{\lambda} \\
&= \max_{\lambda\in\sigma(A^TA)} \sqrt{\frac{1}{\lambda}} \\
&= \frac{1}{\min \lambda \in \sigma(A^TA)\sqrt{\lambda}} \\
&\overset{\textbf{by def}}{=} \frac{1}{\sigma_{\min}}
\end{aligned}$$

Putting this together, we get

$$\mathcal{K}_2(A) := \|A^{-1}\|_2\|A\|_2 = \sigma_{\max}\frac{1}{\sigma_{\min}},$$

completing the proof. □

5 Problem 5

Proposition 5.1. Let $A \in GL_n(\mathbb{R})$, with $\mathcal{K}_2(A) = \|A\|_2 \left\|A^{-1}\right\|_2$. Then

$$\min\left\{\frac{\|\delta A\|_2}{\|A\|_2} | \det(A + \delta A) = 0\right\} = \frac{1}{\mathcal{K}_2(A)}$$

where $\delta A \in \mathbb{R}^{n\times n}$

Proof. Assume $\det(A + \delta A) = 0$. This means we can find some vector $\vec{v} \neq 0$

such that $(A+\delta A)\vec{v}=0, \|\vec{v}\|_2=1$.

$$\begin{aligned}
& (A+\delta A)\vec{v}=0\\
\Rightarrow\quad & A\vec{v}=-\delta A\vec{v}\\
\Rightarrow\quad & \|A\vec{v}\|_2=\|\delta A\vec{v}\|_2\\
\Rightarrow\quad & \|\delta A\|_2 \geq \|\delta A\vec{v}\|_2=\|A\vec{v}\|_2 \geq \inf_{\|\vec{x}\|_2=1}\|A\vec{x}\|_2=\frac{1}{\|A^{-1}\|_2}\\
\Rightarrow\quad & \|\delta A\|_2 \geq \frac{1}{\|A^{-1}\|_2}\\
\Rightarrow\quad & \frac{\|\delta A\|_2}{\|A\|_2} \geq \frac{1}{\|A^{-1}\|_2\|A\|_2}=\frac{1}{\mathcal{K}_2(A)}.
\end{aligned}$$

Now that we have established a lower bound, all that's left is prove existence of a δA such that we have equality.
By theorem 2.14 in Suli & Mayers[1], A can be expressed as $A=U\Sigma V^T$, where Σ is a diagonal matrix with $\sigma_i i$ being the singular values of A on its diagonal and U, V are such that $U^TU=I_n=V^TV$. Now picking $\delta A=U\Sigma_\delta V^T$ where Σ_δ has entries:

$$\Sigma_{\delta ij}=\begin{cases}-\sigma_{nn} \text{ if } ij=nn\\ 0 \text{ elsewhere}\end{cases}.$$

Observe that we have picked δA such that

$$\begin{aligned}
\det(A+\delta A)&=\det(U\Sigma V^T+U\Sigma_\delta V^T)\\
&=\det(U(\Sigma V^T+\Sigma_\delta V^T))\\
&=\det(U\left((\Sigma+\Sigma_\delta)V^T\right))\\
&=\det(U)\det(\Sigma+\Sigma_\delta)\det(V^T)
\end{aligned}$$

but Σ and Σ_δ being diagonal matrices gives us

$$\det(\Sigma+\Sigma_\delta)=(\sigma_{11})(\sigma_{22})\cdots(\sigma_{nn}-\sigma_{nn})=0$$

so $\det(A+\delta A)=0$. □

As we have seen in the preceding problems, Theorem 2.9 $\|\delta A\|_2=\max_{\lambda\in\sigma((\delta A)^T\delta A)}\lambda^{\frac{1}{2}}$. Let's find these eigenvalues. We have

$$(\delta A)^T\delta A=\left(U\Sigma_\delta V^T\right)^T U\Sigma_\delta V^T=V\Sigma_\delta U^TU\Sigma_\delta V^T=V\Sigma_\delta^2V^T.$$

Furthermore

$$\begin{aligned}
\det(V\Sigma_\delta^2V^T-\lambda I)&=\det(V\Sigma_\delta^2V^T-\lambda VV^T)\\
&=\det(V(\Sigma_\delta^2-\lambda I)V^T)\\
&=\det(V)\det(\Sigma_\delta^2-\lambda I)\det(V^T)\\
&=\det(\Sigma_\delta^2-\lambda I)
\end{aligned}$$

from the fact that $VV^T=I_n$ and it's direct consequence $\det(V)\det(V^T)=1$. Now the fact that $(\delta A)^T\delta A)$ has the same characteristic polynomial as Σ_δ^2 means that the eigenvalues we are after are just the eigenvalues of Σ_δ^2, namely

$\{(-\sigma_{nn})^2\ \}$. Trivially this means $(-\sigma_{nn})^2$ is the biggest eigenvalue of $(\delta A)^T\delta A)$ making $\|\delta A\|_2 = \max_{\lambda\in\sigma((\delta A))^T\delta A)} \lambda^{\frac{1}{2}} = \sigma_{nn}$ which by construction equals the smallest singular value of A. To summarize we now have

$$\frac{\|\delta A\|_2}{\|A\|_2} = \frac{\sigma_{nn}}{\sigma_{11}} = \frac{1}{\frac{\sigma_{A_{\max}}}{\sigma_{A_{\min}}}} = \frac{1}{\mathcal{K}_2(A)}.$$

6 Problem 6 Divided differences

a) We want to interpolate using divided differences and the Newton form of the interpolation polynomial, $N(x)$,

x	-2	-1	0	1
y	1	2	3	0

$$N(x) := \sum_{j=0}^{k} a_j n_j(x) \tag{8}$$

where

$$n_j(x) := \prod_{i=0}^{j-1}(x - x_i) \text{ for } j > 0, \quad n_0(x) \equiv 1 \tag{9}$$

$$a_j := [y_0, \dots, y_j]. \tag{10}$$

Combining the definitions we get

$$N(x) = [y_0] + [y_0, y_1](x - x_0) + \cdots + [y_0, \dots, y_k](x - x_0)(x - x_1)\cdots(x - x_{k-1}). \tag{11}$$

We start by calculating all the divided differences needed:

$$\begin{aligned}
[y_0] &= 1\\
[y_0, y_1] &= \frac{y_1 - y_0}{x_1 - x_0} = \frac{2-1}{-1+2} = 1\\
[y_1, y_2] &= \frac{y_2 - y_1}{x_2 - x_1} = \frac{3-2}{0+1} = 1\\
[y_2, y_3] &= \frac{y_3 - y_2}{x_3 - x_2} = \frac{0-3}{1-0} = -3\\
[y_0, y_1, y_2] &= \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0} = \frac{1-1}{0+2} = 0\\
[y_1, y_2, y_3] &= \frac{[y_2, y_3] - [y_1, y_2]}{x_3 - x_1} = \frac{-3-1}{1+1} = -2\\
[y_0, y_1, y_2, y_3] &= \frac{[y_1, y_2, y_3] - [y_0, y_1, y_2]}{x_3 - x_0} = \frac{-2-0}{1+2)} = -\frac{2}{3}
\end{aligned}$$

Inserting in 11 yields:

$$\begin{aligned}
N(x) &= 1 + 1(x - x_0) + 0(x - x_0)(x - x_1) - \frac{2}{3}(x - x_0)(x - x_1)(x - x_2)\\
&= 1 + (x - 2) - \frac{2}{3}(x + 2)(x + 1)x\\
&= -\frac{2}{3}x^3 - 2x^2 - \frac{1}{3}x + 3.
\end{aligned}$$

b) We will use divided differences to find the polynomial of lowest degree such that

$$p(-1) = 1/2, \quad p'(1/2) = 3, \quad p(1) = -1/2.$$

Let $x_0 = -1, \quad x_1 = 1$, then by construction $y_0 = \frac{1}{2}, \quad y_1 = -\frac{1}{2}$. This gives Newton polynomial

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

Now, we can easily see that a polynomial of degree won't do the trick, so we expand our polynomial by adding the next term of the Newton interpolation polynomial, namely $a_2(x - x_0)(x - x_1)$:

$$\begin{aligned}
& N(x) = -\frac{1}{2}x + a_2(x - x_0)(x - x_1) \\
\Rightarrow \quad & N'(x) = 2a_2x - \frac{1}{2} \\
\Rightarrow \quad & N'\left(\frac{1}{2}\right) = a_2 - \frac{1}{2} \\
\Rightarrow \quad & N'\left(\frac{1}{2}\right) = 3 \Leftrightarrow a_2 = \frac{7}{2} \\
\Rightarrow \quad & N(x) = \frac{7}{2}x^2 - \frac{1}{2}x - \frac{7}{2}.
\end{aligned}$$

7 Problem 7 Divided differences

a) Firstly, let's check that the proposition works for $k = 0, \quad k = 1$:

$$S_0(n) = \sum_{i=0}^{n} i^0 = n$$

which surely has degree $1 = k + 1$ as a polynomial in $\mathbb{R}[n]$.

$$S_1(n) = \sum_{i=0}^{n} i^1 = \frac{n(n+1)}{2} = \frac{1}{2}(n^2 + n)$$

which once again has degree $2 = k + 1$. This establishes base cases for an induction proof. As we are looking to prove that S_k is a polynomial of degree $k + 1$, let's assume S_j is a polynomial of degree $j + 1 \quad \forall j \leq k - 1$. We are going to need the binomial theorem.

Theorem 7.1. Let $n \geq 0$ an integer. Then

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{12}$$

As a consequence of the binomial theorem (7.1) we have

$$\begin{aligned}
\sum_{i=0}^{n}(i+1)^{k+1} = \sum_{j=0}^{k+1}\binom{l+1}{j}\sum_{i=0}^{n} i^j 1^{k-j} &= \sum_{j=0}^{k+1}\binom{k+1}{j} S_j(n) \\
\Longrightarrow (n+1)^{k+1} &= \sum_{i=0}^{n}(i+1)^{k+1} - \sum_{i=0}^{n} i^{k+1} \\
&= \sum_{j=0}^{k+1}\binom{k+1}{j} S_j(n) - S_{k+1}(n) \\
&= \sum_{j=0}^{k}\binom{k+1}{j} S_j(n) \\
\Longrightarrow \binom{k+1}{k} S_k(n) = (n+1)^{k+1} &- \sum_{j=0}^{k-1}\binom{k+1}{j} S_j(n) \\
\Longrightarrow S_k(n) = \left(n^{k+1} - \sum_{j=0}^{k-1}\binom{k+1}{j} S_j(n) + 1\right) C,
\end{aligned}$$

with $C = \binom{k+1}{j}^{-1}$. By our induction hypothesis this means we can write

$$S_k(n) = \left(n^{k+1} - P(n)\right) C,$$

where $P(n)$ is just some polynomial of degree $k-1$ or lower, making it clear that $S_k(n)$ is a polynomial of degree $k+1$. □

b) Using the Newton interpolation polynomial as defined in the preceding problem (see equation 8), we can now express $S_k(n)$ in Newton form:

$$S_k(n) = \sum_{j=0}^{k} a_j n_j(x), \tag{13}$$

where $a_j = S_k[1, \ldots, 1+j]$, $n_j(n) = (n-1)(n-2)\cdots(n-j-1)$. To calculate $S_4(n)$, we simply have to calculate all the required divided differences. First, as $S_4(n)$ will be a polynomial of degree 5, we will need 6 nodes:

$$S_4(1) = 1, S_4(2) = 17, S_4(3) = 98, S_4(4) = 354, S_4(5) = 979, S_4(6) = 2275$$

These are now our respective x and y values for the divided differences method. Let's present the differences as an upper triangular matrix for convenience:

$$\begin{bmatrix}
\cdot[y_0] & [y_0, y_1] & \cdots & [y_0, \ldots, y_n] \\
0 & [y_1] & \cdots & [y_1, \ldots, y_n] \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & [y_n]
\end{bmatrix}
=
\begin{bmatrix}
1 & 16 & \frac{65}{2} & \frac{55}{3} & \frac{14}{4} & \frac{1}{5} \\
0 & 17 & 81 & \frac{175}{2} & \frac{97}{3} & \frac{18}{4} \\
0 & 0 & 98 & 256 & \frac{369}{2} & \frac{151}{3} \\
0 & 0 & 0 & 354 & 625 & \frac{671}{2} \\
0 & 0 & 0 & 0 & 974 & 1296 \\
0 & 0 & 0 & 0 & 0 & 2275
\end{bmatrix}$$

Now we just insert in 13 and get

$$\begin{aligned}S_4(n) &= 1 + 16(n-1) + \frac{65}{2}(n-1)(n-2) + \frac{55}{3}(n-1)(n-2)(n-3)\\ &+ \frac{14}{4}(n-1)(n-2)(n-3)(n-4) + \frac{1}{5}(n-1)(n-2)(n-3)(n-4)(n-5)\\ &= \frac{1}{5}n^5 + \frac{1}{2}n^4 + \frac{1}{3}n^3 - \frac{1}{30}n\end{aligned}$$

8 Problem 8 Quadratic formulae

a) Extrapolation is mainly about estimation, and unlike interpolation, it allows for those estimates to be outside the range of the original observations. "Extrapolation may also mean extension of a method"[3], and a good example is Richardson extrapolation which is the method used in Romberg's algorithm. Richardson extrapolation is a method for improving the rate of convergence of some estimation method – in our case this is the Trapezoid rule.

Romberg's algorithm is a repeated application of Richardson extrapolation on the trapezium rule.(ref boka)

b) We have

$$f(\tau) = e^{(-\tau)^2} \tag{14}$$

and want to approximate the integral $\int_0^1 f(\tau)\tau$, and find $R(3,2)$.
As is standard for Romberg, we let $h_n = \frac{1}{2^n}(1-0) = \frac{1}{2^n}$.

$$\begin{aligned}R(0,0) &= h_1(f(0)+f(1)) = \frac{1}{2}\left(1+\frac{1}{e}\right) \approx 0.68394\\ R(1,0) &= \frac{1}{2}R(0,0) + h_1\sum_{i=1}^{2^{1-1}} f(0+(2i-1)h_1) = \frac{1}{2}R(0,0) + \frac{1}{2}f\left(\frac{1}{2}\right) \approx 0.73137\\ R(2,0) &= \frac{1}{2}R(1,0) + h_2\sum_{i=1}^{2^2} f(0+(2i-1)h_2) = \frac{1}{2}R(1,0) + \frac{1}{2}\sum_{i=1}^{2} f\left((2i-1)\frac{1}{4}\right) \approx 0.74298\\ R(3,0) &= \frac{1}{2}R(2,0) + h_3\sum_{i=1}^{2^2} f(0+(2i-1)h_3) = \frac{1}{2}R(2,0) + \frac{1}{8}\sum_{i=1}^{4} f\left((2i-1)\frac{1}{8}\right) \approx 0.74586\\ R(1,1) &= R(1,0) + \frac{1}{4^1-1}(R(1,0)-R(0,0)) \approx 0.74718\\ R(2,1) &= R(2,0) + \frac{1}{4^1-1}(R(2,0)-R(1,1)) \approx 0.74158\\ R(2,2) &= R(2,1) + \frac{1}{4^2-1}(R(2,1)-R(1,1)) \approx 0.74121\\ R(3,1) &= R(3,0) + \frac{1}{4^1-1}(R(3,0)-R(2,0)) \approx 0.74682\\ R(3,2) &= R(3,1) + \frac{1}{4^1-1}(R(3,1)-R(2,1)) \approx 0.74507\end{aligned}$$

9 Problem 9 Convergence of Runge-Kutta methods

Let's start with restating the definitions and assumptions given by the problem description to get a grasp of what we know before we begin the proof. We have the initial value problem

$$\dot{y} = f(y), y(0) = y_0, \text{ on } [0,T], \quad y(t) \in \mathbb{R}^m.$$

We assume $f : \mathbb{R}^m \to \mathbb{R}^m$ continous in t and y and satisfies the Lipschitz condition w.r.t. y on $\mathbb{R} \times \mathbb{R}^m$ with the Lipschitz constant L. Let N be the number of steps and consider the one-step method

$$y_{n+1} = y_n + h\Psi_{f,h}(y_n), \quad h = \frac{T}{N}. \tag{15}$$

a) Now we will assume the function $\Psi_{f,h}$ also satisfies the Lipschitz condition on $\mathbb{R} \times \mathbb{R}^m$ with constant M and that 15 is consistent of order p.

Proof. We use the hint and follow the proof for convergence of the Euler method given in the lecture notes[4], and write $e_N := y(t_N) - y_N$, wanting to prove $\lim_{\substack{N\to\infty \\ h\to 0}} \|e_N\| = 0$.

$$\begin{aligned}
\|e_{N+1}\| = y(t_{N+1}) - y_{N+1} &\leq \|y(t_{N+1}) - z_{N+1}\| + \|z_{N+1} - y_{N+1}\| \\
&\leq \left\|\sigma_{t_{N+1}}, h\right\| + \|e_N + h\Psi_{f,h}(y(t_N))\| \\
&\leq \left\|\sigma_{t_{N+1}}, h\right\| + \|y(t_N) + h\Psi_{f,h}(y(t_N)) - y_N - h\Psi_{f,h}(y_N)\| \\
&\leq \left\|\sigma_{t_{N+1}}, h\right\| + \|e_N\| + h\|\Psi_{f,h}(y(t_N)) - \Psi_{f,h}(y_N)\| \\
&\leq \left\|\sigma_{t_{N+1}}, h\right\| + \|e_N\| + hM\|e_N\| \\
&\leq Ch^{p+1} + (1 + hM)\|e_N\|
\end{aligned}$$

Here we have used the triangle inequality for the first and fourth inequality, before using the Lipschitz condition and the consistency of the method for the final inequality. Now we apply the lemma and corresponding corollary from the lecture notes on the convergence of the Euler method, and get

$$\begin{aligned}
\|e_N\| &\leq e^{MT}\|e_0\| + D\frac{e^{MT} - 1}{DM}h^{p+1} \\
&= \frac{e^{MT} - 1}{M}h^{p+1}
\end{aligned}$$

$$\begin{aligned}
& \|e_N\| \leq \frac{e^{MT} - 1}{M}h^{p+1} \\
\Rightarrow & \lim_{\substack{N\to\infty \\ h\to 0}} \|e_N\| \leq \lim_{\substack{N\to\infty \\ h\to 0}} \frac{e^{MT} - 1}{M}h^{p+1} \\
\Rightarrow & \lim_{\substack{N\to\infty \\ h\to 0}} \|e_N\| \leq 0 \\
\Rightarrow & \lim_{\substack{N\to\infty \\ h\to 0}} \|e_N\| = 0
\end{aligned}$$

We have convergence. □

b) Now, we assume 15 is explicit Runge-Kutta with 2 stages and order p. We want to show

$$\Psi_{f,h}\left(t_n, y_n\right) = b_1 f\left(t_n, y_n\right) - b_2 f\left(t_n + ch, y_n + haf\left(t_n, y_n\right)\right) \tag{16}$$

is Lipschitz.

Proof. We begin with applying the triangle inequality like in a):

$$\begin{aligned}
&\|\Psi_{f,h}(t_N, y_N) - \Psi_{f,h}(t_N, z_N)\| \\
&= \|b_1 f(t_N, y_N) + b_2 f(t_N + ch, y_N + ahf(t_N, y_N)) - b_1 f(t_N, z_N) - b_2 f(t_N + ch, z_N + ahf(t_N, z_N))\| \\
&= \|b_1(f(t_N, y_N) - f(t_N, z_N)) + b_2(f(t_N + ch, y_N + ahf(t_N, y_N)) - f(t_N + ch, z_N + ahf(t_N, z_N)))\| \\
&\leq |b_1|\|f(t_N, y_N) - f(t_N, z_N)\| + |b_2|\|f(t_N + ch, y_N + ahf(t_N, y_N)) - f(t_N + ch, z_N + ahf(t_N, z_N))\| \\
&\leq |b_1|L\|y_N - z_N\| + |b_2|L\|y_N + ahf(t_N, y_N) - z_N - ahf(t_N, z_N)\| \\
&\leq |b_1|L\|y_N - z_N\| + |b_2|L\|y_N - z_N\| + |b_2 a|hL_f\|f(t_N, y_N) - f(t_N, z_N)\| \\
&\leq (|b_1|L + |b_2|L + |b_2 a|hL^2)\|y_N - z_N\|
\end{aligned}$$

Here L is the Lipschitz constant of f. We see that $K = b_1 L + b_2 L + b_2 a h_L^2$ satisfies the Lipschitz condition for $\Psi_{f,h}$. □

10 Problem 10

See the Jupyter notebook.

11 Problem 11

a) We are given

$$A_h := \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad G_h =: A_h + \omega^2 I, \quad \Theta = \begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_M \end{bmatrix},$$

and asked to find

$$G_h \Theta = \mathbf{b}$$

We calculate the matrix product:

$$G_h\Theta = b = \begin{bmatrix} \frac{1}{h^2}\left(-2\Theta_1 + \Theta_2 + 0 \cdots + 0\right) + \omega^2\Theta_1 \\ \frac{1}{h^2}\left(\Theta_1 - 2\Theta_2 + \Theta_3 + 0 \cdots + 0\right) + \omega^2\Theta_2 \\ \vdots \\ \frac{1}{h^2}\left(0 + \cdots + \Theta_{i-1} - 2\Theta_i + \Theta_{i+1} + 0 + \cdots\right) + \omega^2\Theta_i \\ \vdots \\ \frac{1}{h^2}\left(0 + \cdots + \Theta_{M-2} - 2\Theta_{M-1} + \Theta_M\right) + \omega^2\Theta_{M-1} \\ \frac{1}{h^2}\left(0 + \cdots + \Theta_{M-1} - 2\Theta_M\right) + \omega^2\Theta_M \end{bmatrix} = \begin{bmatrix} \frac{-\alpha}{h^2} \\ 0 \\ \vdots \\ 0 \\ \frac{-\beta}{h^2} \end{bmatrix}$$

b)

$$\vec{\tau}_h := G_h\vec{\theta} - \mathbf{b}, \qquad \vec{\theta} := \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}, \quad \theta_j := \theta\left(t_j\right)$$

Let's write out $\vec{\tau}_h$ for clarity:

$$\vec{\tau}_h = \begin{bmatrix} \frac{1}{h^2}\left(\alpha - 2\theta_1 + \theta_2\right) + \omega^2\theta_1 \\ \frac{1}{h^2}\left(\theta_1 - 2\theta_2 + \theta_3\right) + \omega^2\theta_2 \\ \vdots \\ \frac{1}{h^2}\left(\theta_{M-1} - 2\theta_M + \beta\right) + \omega^2\theta_M \end{bmatrix} \tag{17}$$

Observe how $\mathbf{b}$ conveniently completes the divided differences by adding α, β in the first and last row, while having no contribution to the rows inbetween. Hence, we write

$$\tau_i = \frac{1}{h^2}\Big(\left(\theta_{i-1}\right) - 2\left(\theta_i\right) + \left(\theta_{i+1}\right)\Big) + \omega^2\left(\theta_i\right) \tag{18}$$

Want to show:

$$\tau_m = \frac{1}{12}h^2\theta^{(4)}\left(t_m\right) + \mathcal{O}\left(h^4\right), \quad m = 1, \dots, M$$

The ugly Taylor calculations that follows are very similar to theorem 13.1 in the book[1] and might not be necessary to include here, but I did a lot of it before discovering the theorem, so I continue.
We will Taylor expand τ_i, but first observe that $\theta_i = \theta(t_i) = \theta(ih)$, $t_{i-1} = ih - h$, $t_{i+1} = ih + h$. We can use this to get Taylor functions for $\theta_{i-1}, \theta_i, \theta_{i+1}$ to evaluate in the same variable instead of three different ones.
We assume $\theta(t)$ to be four times differentiable with continuous derivatives on $[a, b]$. Then, by Taylor's theorem, for each value x in $[a, b]$, there exists $\xi = \xi(x)$ in (a, b) such that

$$f(x) = f(a) + (x-a)f'(a) + \cdots + \frac{(x-a)^n}{n!}f^{(n)}(a) + \frac{(x-a)^{n+1}}{(n+1)!}f^{(n+1)}(\xi)$$

Now, observe that choosing intervals $[ih - h, ih]$ and $[ih, ih + h]$ imply, by Taylor's theorem that there exist ξ_1, ξ_2 in the two intervals respectively, such that

$$\theta(ih-h) = \theta(ih) - h\theta'(ih) + \frac{h^2}{2}\theta''(ih) - \frac{h^3}{6}\theta'''(ih) + \frac{h^4}{24}\theta^{(4)}\left(\xi_1\right) - \frac{h^5}{120}\theta^{(5)}\left(ih\right) + \mathcal{O}(h^6)$$

$$\theta(ih+h) = \theta(ih) + h\theta'(ih) + \frac{h^2}{2}\theta''(ih) + \frac{h^3}{6}\theta'''(ih) + \frac{h^4}{24}\theta^{(4)}\left(\xi_2\right) + \frac{h^5}{120}\theta^{(5)}\left(ih\right) + \mathcal{O}(h^6)$$

Adding the two equations we get:

$$\theta(ih-h) + \theta(ih-h) = 2\theta(ih) + h^2\theta''(ih) + \frac{1}{24}h^4\Big(\theta^{(4)}(\xi_1) + \theta^{(4)}(\xi_2)\Big) \tag{19}$$

We have assumed $\theta^{(4)}$ to be continuous on $[ih-h, ih+h]$, implying there is a number $\xi \in (\xi_1, \xi_2)$, and thus also in $(ih-h, ih+h)$, such that

$$\frac{1}{2}\left(\theta^{(4)}(\xi_1) + \theta^{(4)}(\xi_2)\right) = \theta^{(4)}(\xi)$$

This fact inserted in 19 yields

$$\theta(ih-h) + \theta(ih-h) = 2\theta(ih) + h^2\theta''(ih) + \frac{1}{12}h^4\theta^{(4)}(\xi) \tag{20}$$

Recall the original pendulum equations:

$$\theta''(t) + \omega^2\theta(t) = 0, \quad 0 < t < 1, \quad \theta(0) = \alpha, \theta(1) = \beta$$

Hence, we can write $\theta''(t) = -\omega^2\theta(t)$. Now we insert 20 into 18 and get

$$\begin{aligned}\tau_i &= \frac{1}{h^2}\left(h^2\theta''(ih) + \frac{1}{12}h^4\theta^{(4)}(\xi) + \mathcal{O}(h^6)\right) + \omega^2(\theta_i)\\ &= \theta''(ih) + \frac{1}{12}h^2\theta^{(4)}(\xi) + \mathcal{O}(h^4) - \theta''(ih)\\ &= \frac{1}{12}h^2\theta^{(4)}(\xi) + \mathcal{O}(h^4)\end{aligned}$$

By definition of the two-norm, we have

$$\|\vec{\tau}\|_2 = \left(\sum_{i=1}^{M}|\tau_i|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{M}\left|\frac{\theta^{(4)}}{12}h^2 + \mathcal{O}(h^4)\right|^2\right)^{\frac{1}{2}} \tag{21}$$

By Taylor's theorem we have the existence of some $X \in (0,1)$ such that

$$\left|\frac{h^2}{12}\theta^{(4)}(ih) + \mathcal{O}(h^4)\right| \leq \left|\frac{h^2}{12}\theta^{(4)}(X)\right| \quad \forall \quad i \quad (\in \mathbb{N})$$

which in turn implies

$$\|\vec{\tau}\|_2 \leq \sqrt{X}\left|\frac{\theta^{(4)}(X)}{12}h^2\right|$$

Now, let's consider the behaviour of the two-norm when we interpret $\vec{\tau}$ as a piecewise constant function,

$$\vec{\tau}_2 = \int_0^1 |\tau_h|^2 dt = \sum_{j=1}^{M}\int_{t_j}^{t_{j+1}} |\tau_j|^2 dt \leq \sum_{j=1}^{M}\int_{t_j}^{t_{j+1}}\left|\frac{\theta^{(4)}(X)h^2}{12}\right|^2 dt \tag{22}$$

$$= hX\left|\frac{\theta^{(4)}(X)h^2}{12}\right|^2 \xrightarrow[h\to 0]{M\to\infty} 0 \tag{23}$$

c) Here we want to show convergence. A method for a boundary value problem is said to be convergent, if $\vec{E_h} \to 0$ as $h \to 0$. Observe that $\vec{\tau_h} = G_h(-\vec{E_h})$

$$\vec{E_h} = -G_h^{-1}\vec{\tau_h} \tag{24}$$

Now we can write $\|\vec{E_h}\|_2 = \| - G_h^{-1}\vec{\tau_h}\|_2$, and we know $\|\vec{\tau_h}\|_2$ from *b)* and hence, finding $\|G_h^{-1}\|_2$ is all we need to find $\|\vec{E_h}\|_2$. From this we can rewrite and bound the error-vector

$$\begin{aligned}\vec{E_h} = -G_h^{-1}\vec{\tau_h} &\leq \|G_h^{-1}\|_2\|\vec{\tau_h}\|_2 \\ &= \max_{\lambda_h \in \sigma(G_h)} \frac{\|\vec{\tau_h}\|_2}{|\lambda_h|} \\ &\leq \max_{\lambda_h \in \sigma(G_h)} \frac{h^2|\theta^{(4)}(X)|}{12|\lambda_h|}\end{aligned} \tag{25}$$

To find the eigenvalues of G_h, we want to use that

$$G_h^{-1} = \left(A_h + \omega^2 I\right)^{-1}$$

and the following lemma.

Lemma 11.1. The eigenvalues of $\alpha I + B$ are $\alpha + \lambda(B)$, where $\lambda(B)$ are the eigenvalues of B.

Proof. Let λ be any eigenvalue of B with corresponding eigenvector $\vec{v}$, then $B\vec{v} = \lambda\vec{v}$. Now $(\alpha I + B)\vec{v} = \alpha I\vec{v} + \lambda\vec{v} = (\alpha + \lambda)\vec{v}$. □

G_h is clearly symmetric and as seen in problem 3, we don't need to calculate the inverse of G_h to say find its 2-norm since $\lambda(G_h^{-1}) = \frac{1}{\lambda(G_h)}$. In particular this means that we are looking for the eigenvalues of

$$(A_h + \omega^2 I)$$

which, by the lemma, reduces to

$$\omega^2 + \lambda(A_h).$$

Lecture note on BVD page xviii state the following eigenvalues for A_h:

$$\lambda_m = \frac{2}{h^2}(\cos(m\pi h) - 1), \quad m = 1, \ldots, M.$$

Now because $0 < m < M + 1$ and $\frac{M+1}{h} = 1$ we see that $m = 1$ minimizes the above cos-term. Henceforth

$$\begin{aligned}\lambda_{h,m} &= \frac{2}{h^2}\left(\cos(mh\pi) - 1\right) + \omega^2 \\ \implies \min_{\lambda_h \in \sigma(G_h)} |\lambda_h| &= \left|\frac{2}{h^2}\left(\cos(h\pi) - 1\right) + \omega^2\right|\end{aligned} \tag{26}$$

Now let's study what happens when sending h to zero:

$$\begin{aligned}\lim_{h \to 0} \min |\lambda_h| &= \lim_{h \to 0} \left|\frac{2(\cos(\pi h) - 1)}{h^2} + \omega^2\right| \\ &= \lim_{h \to 0} \left|\frac{-\pi \sin(\pi h)}{h} + \omega^2\right| \\ &= \lim_{h \to 0} \left|-\pi^2 \cos(\pi h) + \omega^2\right| = |\pi^2 + \omega^2|\end{aligned}$$

Here we have first used the bound from 26 and then the definition of the derivative twice. Now we simply insert 26 into 25 using the fact that minimizing the denominator maximizes the fraction:

$$\begin{aligned}\|\vec{E}_h\|_2 &\leq \max_{\lambda_h \in \sigma(G_h)} \frac{h^2|\theta^{(4)}(X)|}{12|\lambda_h|} \\ &\xrightarrow{h \to 0} \frac{h^2|\theta(4)|(X)}{12|\pi^2 - \omega^2|} = 0\end{aligned}$$

By the assumption $\omega^2 \leq \frac{\pi^2}{2}$, $|\pi^2 + \omega^2| \neq 0$, keeping our limit well-defined. As $h = \frac{1}{M+1}$, $h \to 0$ would mean $(M+1) \to \infty$ this completes our proof of convergence – the error vector goes to zero as the number of discretization points grows big.

d) See the Jupyter notebook.

References

[1] Endre Suli & David Mayers. An Introduction to Numerical Analysis. Cambridge University Press, 2003.

[2] S. Fridberg, A.Insel, L.Spence. Linear Algebra. Pearson, 2014.

[3] https://en.wikipedia.org/wiki/Extrapolation

[4] Elena Celledoni. Lecture notes on consistency and convergence of one-step methods.