Roman Z. Morawski
Warsaw University of Technology
Faculty of Electronics and Information Technology
room: #445, phone: (22) 234-7721
e-mail: *r.morawski@ire.pw.edu.pl*

Numerical Methods (ENUME)
SOLVED PROBLEMS
Spring Semester 2018/2019

Warsaw 2019

ENUME: SOLVED PROBLEMS

1. ACCURACY OF NON-ITERATIVE ALGORITHMS

1.1. Propagation of errors in the data

Problem: The function $T(x) \equiv \delta[\tilde{y}]/\delta[\tilde{x}]$, characterising propagation of the relative error of the variable x to the variable $y = f(x)$, may be computed according to the formula: $T(x) = \frac{x}{y}\frac{dy}{dx}$.

Demonstrate that $T(x) = \frac{d\ln(y)}{d\ln(x)}$.

Solution #1: One may compute the derivative $\frac{d\ln(y)}{d\ln(x)}$ by substituting $x = e^z$; then:

$$\frac{d\ln(y)}{d\ln(x)} = \frac{d\ln(f(x))}{d\ln(x)} = \frac{d\ln\left(f\left(e^z\right)\right)}{dz} = \frac{1}{f\left(e^z\right)}\frac{df(x)}{dx}\frac{de^z}{dz} = \frac{1}{y}\frac{dy}{dx}x = T(x)$$

Solution #2: Alternatively. one may compute the derivative $\frac{d\ln(y)}{d\ln(x)}$ in the following way; since:

$$d\ln(y) = \frac{d\ln(y)}{dy}dy = \frac{1}{y}dy \quad \text{and} \quad d\ln(x) = \frac{d\ln(x)}{dx}dx = \frac{1}{x}dx$$

the ratio of both sides is:

$$\frac{\frac{d\ln(y)}{dy}}{\frac{d\ln(x)}{dx}} = \frac{\frac{1}{y}dy}{\frac{1}{x}dx} = \frac{x}{y}\frac{dy}{dx} = T(x)$$

Problem: Assess the relative error of computing:

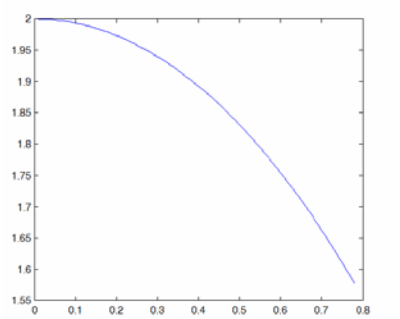
$$y = \sin^2(x) \ \text{ for } \ x \in \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$$

caused by the relative error of the datum x, not exceeding 1%.

Solution: The coefficient of error propagation:

$$T(x) = \frac{dy}{dx}\cdot\frac{x}{y} = 2\sin(x)\cos(x)\frac{x}{\sin^2(x)} = 2\cos(x)\frac{x}{\sin(x)}$$

is an even function whose shape is shown in the figure. Since $\sup\{T(x)\} = 2$, the error of computation does not exceed 2%.

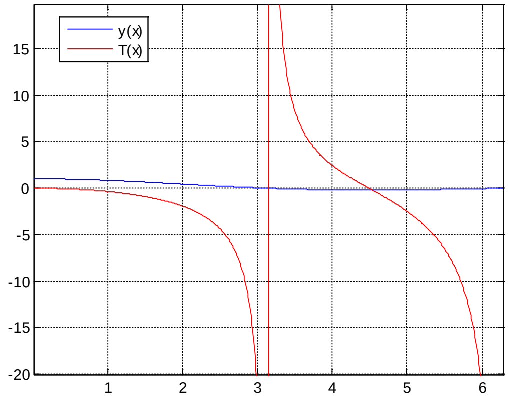


Problem: Determine the functions $T(x) \equiv \delta[\tilde{y}]/\delta[\tilde{x}]$ characterizing the propagation of relative errors in the data, for the following operators:

$$y=\frac{\sin(x)}{x} \quad \text{for } x\in[0,10\pi]$$

Solution:

$$T(x)=\frac{x}{y(x)}\frac{dy(x)}{dx}=\frac{x^2}{\sin(x)}\cdot\frac{\cos(x)\cdot x-\sin(x)}{x^2}=\frac{\cos(x)\cdot x-\sin(x)}{\sin(x)}=x\cdot\operatorname{ctg}(x)-1$$



Problem: Determine the functions $T(x)\equiv\delta[\tilde{y}]/\delta[\tilde{x}]$, characterizing the propagation of relative errors in the data, for the following operators:

$$y=\tan(x) \qquad \text{for } x\in\left[0,\frac{\pi}{2}\right]$$

$$y=e^x\sin(x) \qquad \text{for } x\in[0,10\pi]$$

$$y=e^x\sin(x) \qquad \text{for } x\in[0,10\pi]$$

$$y=xe^{-x}\sin(x) \qquad \text{for } x\in[0,2\pi]$$

$$y=x^2e^{-x}\sin(x) \qquad \text{for } x\in[0,2\pi]$$

$$y=\frac{1+x+x^2+x^3}{1-x+x^2-x^3} \qquad \text{for } x\in[0,10]$$

Verify the results by numerical simulation of errors in MATLAB according to the following formula:

$$T(\dot{x})=1000\frac{\tilde{y}-\dot{y}}{\dot{y}},$$

where $\dot{y}=y(\dot{x})$ and $\tilde{y}=y(1.001\dot{x})$. Draw the graphs of all $y(x)$ and $T(x)$.

Problem: Determine the function $T(x)$, characterising the propagation of the relative error in the variable x during computing the value of the function $y=x^{\frac{1}{x}}$.

Solution #1: The direct differentiation of the RHS should follow the rule:

$$\frac{d}{dx}F\left(f_1(x),f_2(x)\right)=\left.\frac{\partial F(y_1,y_2)}{\partial y_1}\right|_{\substack{y_1=f_1(x)\\ y_2=f_2(x)}}\frac{df_1(x)}{dx}+\left.\frac{\partial F(y_1,y_2)}{\partial y_2}\right|_{\substack{y_1=f_1(x)\\ y_2=f_2(x)}}\frac{df_2(x)}{dx}$$

In the considered case:

$$f_1(x)\equiv x,\ \ f_2(x)\equiv\frac{1}{x} \quad \text{and} \quad F(y_1,y_2)\equiv y_1^{y_2}$$

Thus:

$$\frac{df_1(x)}{dx}=1,\ \frac{df_2(x)}{dx}=-\frac{1}{x^2},\ \frac{\partial F(y_1,y_2)}{\partial y_1}=y_2 y_1^{y_2-1}\ \text{ and }\ \frac{\partial F(y_1,y_2)}{\partial y_2}=\ln(y_1)\,y_1^{y_2}$$

and consequently:

$$\frac{dy}{dx}=\tfrac{1}{x}x^{\frac{1}{x}-1}+\ln(x)\,x^{\frac{1}{x}}\left(-\tfrac{1}{x^2}\right)=x^{\frac{1}{x}-2}\left(1-\ln(x)\right)$$

$$T(x)=\frac{x}{y}\frac{dy}{dx}=\frac{x\left[x^{\frac{1}{x}-2}\left(1-\ln(x)\right)\right]}{x^{\frac{1}{x}}}=\frac{1-\ln(x)}{x}$$

<u>Solution #2:</u> The same result may be obtained in a more efficient way by differentiation of the RHS logarithm:

$$\frac{d}{dx}\ln(y)=\frac{d}{dx}\ln\left(x^{\frac{1}{x}}\right)=\frac{d}{dx}\left[\frac{1}{x}\ln(x)\right]=\frac{d}{dx}\left(\frac{1}{x}\right)\ln(x)+\frac{1}{x}\frac{d}{dx}\left[\ln(x)\right]$$

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

Since

$$\frac{d}{dx}\ln(y)=\frac{1}{y}\frac{dy}{dx}$$

the sought-for function may be given the form:

$$T(x)=\frac{x}{y}\frac{dy}{dx}=x\frac{1-\ln(x)}{x^2}=\frac{1-\ln(x)}{x}$$

Problem: Assess the relative error of computing:

$$\tilde{y}=\frac{d}{dx}\left(\frac{x+\tilde{a}}{x+\tilde{b}}\right)\quad\text{for }\ x\in[0,1]$$

caused by the relative errors in the data: $\tilde{a}=1+\alpha$ and $\tilde{b}=2(1+\beta)$, where $|\alpha|\le 1\%$ and $|\beta|\le 1\%$.

<u>Solution:</u> Since the propagation of errors in the data does not depend on the numerical algorithm, the formula defining $\tilde{y}$ may be simplified by execution of differentiation:

$$\tilde{y}=\frac{\tilde{b}-\tilde{a}}{\left(x+\tilde{b}\right)^2}=\frac{2(1+\beta)-(1+\alpha)}{\left(x+2(1+\beta)\right)^2}=\frac{1+2\beta-\alpha}{\left((x+2)+2\beta\right)^2}=\frac{1+2\beta-\alpha}{(x+2)^2\left(1+\frac{2\beta}{x+2}\right)^2}$$

Hence:

$$\delta[\tilde{y}]=2\beta-\alpha-\frac{4\beta}{x+2}=-\alpha+\left(2-\frac{4}{x+2}\right)\beta=-\alpha+\frac{2x}{x+2}\beta$$

and consequently:

$$\left|\delta[\tilde{y}]\right|\le|\alpha|+\frac{2x}{x+2}|\beta|\le\left(1+\frac{2x}{x+2}\right)1\%=\frac{3x+2}{x+2}1\%$$

The function $F(x)\equiv\dfrac{3x+2}{x+2}$ is increasing for $x\in[0,1]$ because:

$$\frac{dF(x)}{dx}\equiv\frac{4}{(x+2)^2}>0$$

Therefore:

$$\left|\delta[\tilde{y}]\right| \le F(1)\cdot 1\% = \frac{3+2}{1+2}1\% \simeq 1.66\%$$

Problem: Determine the function $T(x)$, characterising the propagation of relative errors in the variable x, for the following operator:

$$y = \left(\frac{1-x}{1+x}\right)^x \text{ for } x \in (0,1)$$

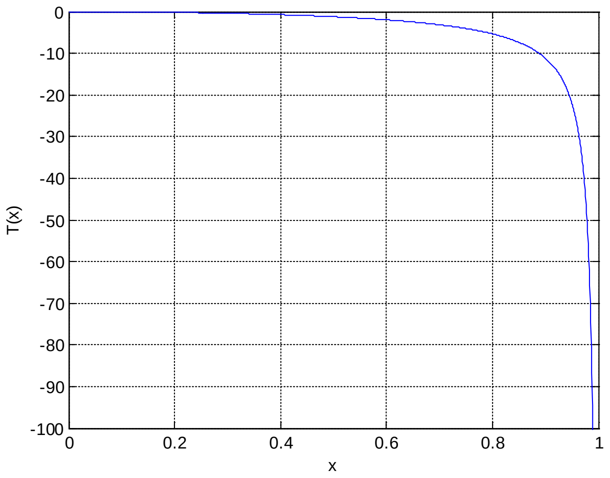
<u>Solution:</u>

$$\ln(y) = x\ln\left(\frac{1-x}{1+x}\right) = x\left[\ln(1-x) - \ln(1+x)\right]$$

$$\left[\ln(y)\right]' = \left[\ln(1-x) - \ln(1+x)\right] + x\left[-\frac{1}{1-x} - \frac{1}{1+x}\right]$$

$$\frac{y'}{y} = \ln\left(\frac{1-x}{1+x}\right) - \frac{2x}{1-x^2}$$

$$T(x) \equiv x\frac{y'}{y} = x\ln\left(\frac{1-x}{1+x}\right) - \frac{2x^2}{1-x^2}$$




```
clear all
x=linspace(0,1,1000);
T=x.*log((1-x)./(1+x))-2*x.*x./(1-x.*x);
plot(x,T)
xlabel('x')
ylabel('T(x)')
axis([0 1 -100 0])
grid on
```

Problem: Determine the function $T(x)$, characterizing the propagation of relative errors in the variable x, for the following operator:

$$y = \left[\tan(x)\right]^x \text{ for } x \in \left(0, \frac{\pi}{2}\right)$$

<u>Solution #1:</u>

$$\ln(y) = x\ln\left[\tan(x)\right]$$

$$\frac{y'}{y} = \ln\left[\tan(x)\right] + x\frac{1}{\tan(x)}\frac{1}{\cos^2(x)} = \ln\left[\tan(x)\right] + \frac{2x}{\sin(2x)}$$

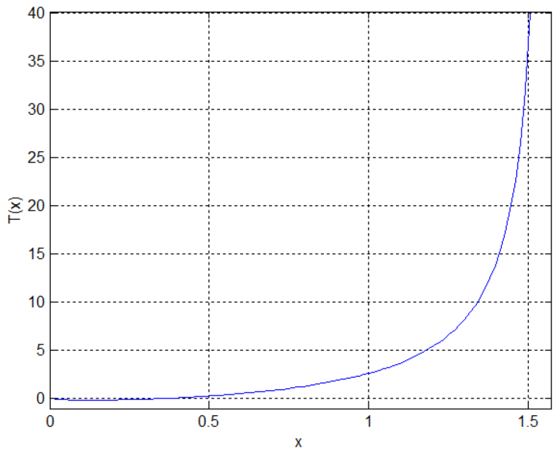
$$T(x) \equiv x\frac{y'}{y} = x\ln\left[\tan(x)\right] + \frac{2x^2}{\sin(2x)}$$

Solution #2:

$$T(x) \equiv \frac{d(\ln(y))}{d(\ln(x))} = \frac{dv}{du}, \text{ where } u \equiv \ln(x),\ v \equiv \ln(y)$$

$$v \equiv \ln(y) = e^u \tan(e^u)$$

$$T(x) = \frac{dv}{du} = \ldots = e^u \ln(\tan(e^u)) + e^u \frac{1}{\tan(e^u)} \frac{1}{\cos^2(e^u)} e^u = x \ln[\tan(x)] + \frac{2x^2}{\sin(2x)}$$




```
clear all
x=linspace(0,pi/2,100);
T=x.*log(tan(x))+2*x.*x./sin(2*x);
plot(x,T)
xlabel('x')
ylabel('T(x)')
axis([0 pi/2 -1 40])
grid on
```

Problem: Assess the absolute error in the argument of the complex variable:

$$\tilde{z} = \frac{j\tilde{b}_1}{\tilde{a}_2 + j\tilde{b}_2}$$

caused by the floating-point representation of the data: $\tilde{b}_1 \cong 2$, $\tilde{a}_2 \cong 1$ i $\tilde{b}_2 \cong 1$.

Solution: On the one hand, we have:

$$\tilde{z} = \frac{j\tilde{b}_1(\tilde{a}_2 - j\tilde{b}_2)}{(\tilde{a}_2 + j\tilde{b}_2)(\tilde{a}_2 - j\tilde{b}_2)} = \frac{\tilde{b}_1\tilde{b}_2 + j\tilde{b}_1\tilde{a}_2}{\tilde{a}_2^2 + \tilde{b}_2^2}$$

which means that:

$$\mathrm{tg}(\tilde{\phi}) = \frac{\tilde{b}_1\tilde{a}_2}{\tilde{b}_1\tilde{b}_2} = \frac{\tilde{a}_2}{\tilde{b}_2} \text{ (where } \tilde{\varphi} \text{ is the argument of } \tilde{z}\text{)}$$

and – after substitution $\tilde{a}_2 = 1 + \alpha_2$ and $\tilde{b}_2 = 1 + \beta_2$:

$$\mathrm{tg}(\tilde{\phi}) = \frac{1+\alpha_2}{1+\beta_2} \cong 1 + \alpha_2 - \beta_2$$

On the other hand, the Taylor's expansion of $\mathrm{tg}(\tilde{\varphi})$ yields:

$$\mathrm{tg}(\tilde{\varphi}) = \mathrm{tg}(\varphi + \Delta\varphi) \cong \mathrm{tg}(\varphi) + \frac{1}{\cos^2(\varphi)}\Delta\varphi$$

The comparison of both expressions for $\mathrm{tg}(\tilde{\varphi})$ leads to the conclusion that:

$$\mathrm{tg}(\phi) = 1$$

$$\frac{1}{\cos^2(\varphi)}\Delta\varphi \cong \alpha_2 - \beta_2$$

which – after substitution $\dfrac{1}{\cos^2(\phi)} = 1 + \mathrm{tg}^2(\phi) = 2$ – provides the solution to the problem:

$$\Delta\phi \cong \cos^2(\phi)(\alpha_2 - \beta_2) = 0.5(\alpha_2 - \beta_2)$$
$$|\Delta\phi| \cong 0.5|\alpha_2 - \beta_2| \le eps$$

Problem: Assess the absolute error in the argument of the complex variable:

$$\tilde{z} = \frac{\tilde{b}_1}{\tilde{a}_2 + j\tilde{b}_2}$$

caused by the floating-point representation of the data: $\tilde{b}_1 \cong 1$, $\tilde{a}_2 \cong 1$ i $\tilde{b}_2 \cong 1$.

Solution: On the one hand, we have:

$$\tilde{z} = \frac{\tilde{b}_1(\tilde{a}_2 - j\tilde{b}_2)}{(\tilde{a}_2 + j\tilde{b}_2)(\tilde{a}_2 - j\tilde{b}_2)} = \frac{\tilde{b}_1\tilde{a}_2 - j\tilde{b}_1\tilde{b}_2}{\tilde{a}_2^2 + \tilde{b}_2^2}$$

which means that:

$$\mathrm{tg}(\tilde{\phi}) = -\frac{\tilde{b}_1\tilde{b}_2}{\tilde{b}_1\tilde{a}_2} = -\frac{\tilde{b}_2}{\tilde{a}_2} \quad \text{(where } \tilde{\varphi} \text{ is the argument of } \tilde{z}\text{)}$$

and – after substitution $\tilde{a}_2 = 1 + \alpha_2$ and $\tilde{b}_2 = 1 + \beta_2$:

$$\mathrm{tg}(\tilde{\phi}) = -\frac{1+\beta_2}{1+\alpha_2} \cong -(1 - \alpha_2 + \beta_2)$$

On the other hand, the Taylor's expansion of $\mathrm{tg}(\tilde{\varphi})$ yields:

$$\mathrm{tg}(\tilde{\varphi}) = \mathrm{tg}(\varphi + \Delta\varphi) \cong \mathrm{tg}(\varphi) + \frac{1}{\cos^2(\varphi)}\Delta\varphi$$

The comparison of both expressions for $\mathrm{tg}(\tilde{\varphi})$ leads to the conclusion that:

$$\mathrm{tg}(\phi) = -1$$
$$\frac{1}{\cos^2(\varphi)}\Delta\varphi \cong \alpha_2 - \beta_2$$

which – after substitution $\dfrac{1}{\cos^2(\phi)} = 1 + \mathrm{tg}^2(\phi) = 2$ – provides the solution to the problem:

$$2\Delta\phi \cong \alpha_2 - \beta_2 \;\Rightarrow\; |\Delta\phi| \cong 0.5|\alpha_2 - \beta_2| \le eps$$

Problem: Assess the absolute error in the argument of the complex variable:

$$\tilde{z} = \frac{\tilde{a}_1 + j\tilde{b}_1}{\tilde{a}_2 + j\tilde{b}_2}$$

caused by the floating-point representation of the data: $\tilde{a}_1 \cong 1$, $\tilde{b}_1 \cong 2$, $\tilde{a}_2 \cong 1$ i $\tilde{b}_2 \cong 1$.

Solution: On the one hand, we have:

$$\tilde{z} = \frac{(\tilde{a}_1 + j\tilde{b}_1)(\tilde{a}_2 - j\tilde{b}_2)}{(\tilde{a}_2 + j\tilde{b}_2)(\tilde{a}_2 - j\tilde{b}_2)} = \frac{(\tilde{a}_1\tilde{a}_2 + \tilde{b}_1\tilde{b}_2) + j(\tilde{b}_1\tilde{a}_2 - \tilde{a}_1\tilde{b}_2)}{\tilde{a}_2^2 + \tilde{b}_2^2}$$

which means that:

$$\mathrm{tg}(\tilde{\varphi}) = \frac{\tilde{b}_1\tilde{a}_2 - \tilde{a}_1\tilde{b}_2}{\tilde{a}_1\tilde{a}_2 + \tilde{b}_1\tilde{b}_2}$$ (where $\tilde{\varphi}$ is the argument of $\tilde{z}$)

and – after substitution $\tilde{a}_1 = 1+\alpha_1$, $\tilde{b}_1 = 2(1+\beta_1)$, $\tilde{a}_2 = 1+\alpha_2$ and $\tilde{b}_2 = 1+\beta_2$:

$$\mathrm{tg}(\tilde{\varphi}) = \frac{2(1+\beta_1)(1+\alpha_2)-(1+\alpha_1)(1+\beta_2)}{(1+\alpha_1)(1+\alpha_2)+2(1+\beta_1)(1+\beta_2)} \cong \frac{1+2\beta_1+2\alpha_2-\alpha_1-\beta_2}{3+\alpha_1+\alpha_2+2\beta_1+2\beta_2} \cong$$

$$\cong \frac{1}{3}\left(1+\frac{4}{3}\beta_1+\frac{5}{3}\alpha_2-\frac{4}{3}\alpha_1-\frac{5}{3}\beta_2\right)$$

On the other hand, the Taylor's expansion of $\mathrm{tg}(\tilde{\varphi})$ yields:

$$\mathrm{tg}(\tilde{\varphi}) = \mathrm{tg}(\varphi+\Delta\varphi) \cong \mathrm{tg}(\varphi) + \frac{1}{\cos^2(\varphi)}\Delta\varphi$$

The comparison of both expressions for $\mathrm{tg}(\tilde{\varphi})$ leads to the conclusion that:

$$\mathrm{tg}(\varphi) = \frac{1}{3}$$

$$\frac{1}{\cos^2(\varphi)}\Delta\varphi \cong \frac{1}{3}\left(\frac{4}{3}\beta_1+\frac{5}{3}\alpha_2-\frac{4}{3}\alpha_1-\frac{5}{3}\beta_2\right)$$

which – after substitution $\frac{1}{\cos^2(\varphi)} = 1+\mathrm{tg}^2(\varphi) = \frac{10}{9}$ – provides the solution to the problem:

$$\Delta\varphi \cong \frac{9}{10}\frac{1}{3}\left(\frac{4}{3}\beta_1+\frac{5}{3}\alpha_2-\frac{4}{3}\alpha_1-\frac{5}{3}\beta_2\right) = 0.4\beta_1+0.5\alpha_2-0.4\alpha_1-0.5\beta_2$$

$$|\Delta\varphi| \cong |0.4\beta_1+0.5\alpha_2-0.4\alpha_1-0.5\beta_2| \le (0.4eps+0.5eps+0.4eps+0.5eps) = 1.8eps$$

Problem: Determine the dependence of the absolute error $\Delta\phi$ of the variable:

$$\phi = Arg\left(\frac{a\cdot e^{jx}+1}{e^{jx}-1}\right)$$

on the absolute error Δx of the variable x for $a=1$. Draw the graph of the function $\frac{\Delta\phi}{\Delta x} = f(x)$.

<u>Solution:</u> The following equalities hold:

$$\frac{a\cdot e^{jx}+1}{e^{jx}-1} = \frac{(a\cdot e^{jx}+1)\cdot(e^{-jx}-1)}{(e^{jx}-1)\cdot(e^{-jx}-1)} = \frac{a-a\cdot e^{jx}+e^{-jx}-1}{2-2\cos(x)}$$

$$= \frac{[a-1-a\cos(x)+\cos(x)]+j[-a\sin(x)-\sin(x)]}{2-2\cos(x)} = \frac{(a-1)[1-\cos(x)]-j(a+1)\sin(x)}{2-2\cos(x)}$$

$$\tan(\phi) = \frac{-(a+1)\sin(x)}{(a-1)[1-\cos(x)]} = -\frac{a+1}{a-1}\cdot\frac{2\sin\left(\frac{x}{2}\right)\cos\left(\frac{x}{2}\right)}{2\sin^2\left(\frac{x}{2}\right)} = -\frac{a+1}{a-1}\cdot\cot\left(\frac{x}{2}\right)$$

By differentiating LHS and RHS with respect to x, one obtains:

$$\frac{1}{\cos^2(\phi)}\frac{d\phi}{dx} = -\frac{a+1}{a-1}\cdot\left[-\frac{1}{\sin^2\left(\frac{x}{2}\right)}\right]\cdot\frac{1}{2} \quad \text{and} \quad \frac{d\phi}{dx} = \frac{1}{2}\cdot\frac{a+1}{a-1}\cdot\frac{\cos^2(\phi)}{\sin^2\left(\frac{x}{2}\right)}$$

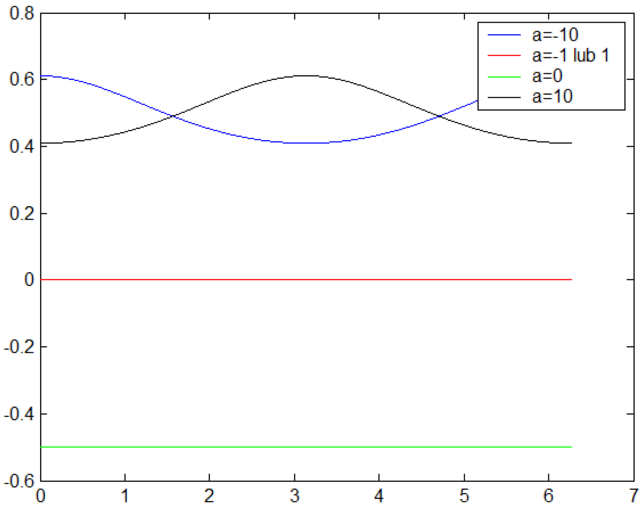
After substituting:

$$\cos^2(\phi) = \frac{1}{1+\tan^2(\phi)} = \frac{1}{1+\left[-\frac{a+1}{a-1}\cdot\cot\left(\frac{x}{2}\right)\right]^2}$$

and obvious algebraic simplifications, one gets the result:

$$\frac{d\phi}{dx} = \frac{1}{2}\cdot\frac{a^2-1}{(a-1)^2\sin^2\left(\frac{x}{2}\right)+(a+1)^2\cos^2\left(\frac{x}{2}\right)} = \frac{1}{2}\cdot\frac{a^2-1}{a^2+1+2a\cos(x)}$$

which is shown in figure below for several values of a.




```
clear all
x=[0:0.001:2*pi];
a=-10;
defipodeix=0.5*(a*a-1)./(a*a+1+2*a*cos(x));
plot(x,defipodeix,'b');hold on
a=-1;
defipodeix=0.5*(a*a-1)./(a*a+1+2*a*cos(x));
plot(x,defipodeix,'r');hold on
a=0;
defipodeix=0.5*(a*a-1)./(a*a+1+2*a*cos(x));
plot(x,defipodeix,'g');hold on
a=10;
defipodeix=0.5*(a*a-1)./(a*a+1+2*a*cos(x));
plot(x,defipodeix,'k');hold on
legend('a=-10','a=-1 lub 1','a=0','a=10');
hold off
```

This figure does not take into account a singularity which appears for $a=1$, when $x=\pi$, because $\cos(x)=-1$, and consequently:

$$\frac{d\phi}{dx} = \frac{1}{2}\cdot\frac{a^2-1}{a^2+1-2a} = \frac{1}{2}\cdot\frac{a^2-1}{(a-1)^2} = \frac{1}{2}\cdot\frac{a+1}{a-1}\xrightarrow[a\to 1]{}\infty$$

Problem: Assess the relative error of the solution to the equation:

$$ax - x^a = 0 \quad \text{for } a\in(0,1)$$

caused by the relative error of the parameter a, not exceeding $p=1\%$.

<u>Solution:</u> The differentiation of the LHS and RHS of the equation with respect to a yields:

$$x + a\frac{dx}{da} - ax^{a-1}\frac{dx}{da} = 0$$

Hence:

$$T(a) = \frac{dx}{da}\cdot\frac{a}{x} = \frac{x}{ax^{a-1}-a}\cdot\frac{a}{x} = \frac{1}{x^{a-1}-1}$$

Since by definition of the solution $x^a = ax$, the coefficient of error amplification may be given the form:

$$T(a) = \frac{1}{x^{-1}x^{a} - 1} = \frac{1}{x^{-1}ax - 1} = \frac{1}{a-1}$$

Hence the assessment:

$$|\delta x| \le |T(a)| \cdot |\delta a| \le |T(a)| \cdot p = \frac{10^{-2}}{a-1}$$

1.2. Propagation of rounding errors

Problem: Assess the relative error in the result of computing:

$$y = \ln\left(\frac{x_1}{x_2}\right) \text{ for } x_1 > e \cdot x_2,\ x_2 > 1$$

by means of the algorithm:

$$A: \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \left[v = \frac{x_1}{x_2} \right] \to \left[y = \ln(v) \right]$$

Neglect the errors in the data.

<u>Solution:</u> The analysis of the algorithm A yields:

$$\tilde{y} = y\left(1 + \frac{\eta_d}{y}\right)(1+\eta_l) = y\left(1 + \frac{\eta_d}{y} + \eta_l\right)$$

$$\delta[\tilde{y}] = \frac{\eta_d}{y} + \eta_l \quad \Rightarrow \quad |\delta[\tilde{y}]| \le \left(\frac{1}{|y|} + 1\right) eps = \left(\frac{1}{y} + 1\right) eps < \left(\frac{1}{\inf[y]} + 1\right) eps < 2eps$$

Problem: Assess the relative error of computing:

$$y = \frac{d}{dx}\left(\frac{x+a}{x+b}\right) \text{ for } x \in [0,1],$$

caused by rounding of the results of the following operations: $x+a$ and $x+b$.

<u>Solution:</u> Since the propagation of rounding errors does depend on the numerical algorithm, the differentiation should not be performed before introducing the rounding errors:

$$\tilde{y} = \frac{d}{dx}\left(\frac{(x+a)(1+\eta_a)}{(x+b)(1+\eta_b)}\right) = \frac{d}{dx}\left(\frac{x+a}{x+b}\right)(1+\eta_a-\eta_b) = y(1+\eta_a-\eta_b)$$

Hence:

$$\delta[\tilde{y}] = \eta_a - \eta_b \text{ and } |\delta[\tilde{y}]| \le |\eta_a| + |\eta_b| \le 2eps$$

Problem: Assess the relative error in the result of computing:

$$y = \ln\left(\frac{x_1}{x_2}\right) \text{ for } e \cdot x_2 < x_1 < \frac{e^2}{x_2} \text{ i } x_2 > 1$$

by means of the algorithm:

$$A: \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \begin{bmatrix} v_1 = \ln(x_1) \\ v_2 = \ln(x_2) \end{bmatrix} \to \left[y = v_1 - v_2 \right]$$

Neglect the errors in the data.

<u>Solution:</u> The analysis of the algorithm A yields:

$$\tilde{y} = \left[\ln(x_1)(1+\eta_1) - \ln(x_2)(1+\eta_2)\right](1+\eta_o)$$

$$= \left[\left(\ln(x_1) - \ln(x_2)\right) + \left(\ln(x_1)\eta_1 - \ln(x_2)\eta_2\right)\right](1+\eta_o)$$

$$\tilde{y} = \left[y + \left(\ln(x_1)\eta_1 - \ln(x_2)\eta_2\right)\right](1+\eta_o) = y\left[1 + \frac{\ln(x_1)\eta_1 - \ln(x_2)\eta_2}{y} + \eta_o\right]$$

$$\delta[\tilde{y}] = \frac{\ln(x_1)\eta_1 - \ln(x_2)\eta_2}{y} + \eta_o$$

$$\left|\delta[\tilde{y}]\right| \le \left(\frac{\left|\ln(x_1)\right| + \left|\ln(x_2)\right|}{|y|} + 1\right) eps = \left(\frac{\ln(x_1) + \ln(x_2)}{y} + 1\right) eps = \left(\frac{\ln(x_1 x_2)}{\ln\left(\frac{x_1}{x_2}\right)} + 1\right) eps$$

$$\left|\delta[\tilde{y}]\right| \le \left(\frac{\sup\left[\ln(x_1 x_2)\right]}{\inf\left[\ln\left(\frac{x_1}{x_2}\right)\right]} + 1\right) eps = \left(\frac{\ln(e^2)}{\ln(e)} + 1\right) eps = 3eps$$

Problem: Compare the relative errors in the results of computing:

$$y = \ln\left(\frac{x_1}{x_2}\right) \text{ for } x_1, x_2 > 1$$

by means of two algorithms:

$$A_1 : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to v = \frac{x_1}{x_2} \to y_1 = \ln(v)$$

$$A_2 : \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \begin{bmatrix} v_1 = \ln(x_1) \\ v_2 = \ln(x_2) \end{bmatrix} \to y_2 = v_1 - v_2$$

Draft the borderline of a set of values of $x_1, x_2 > 1$, for which the algorithm A_1 is more accurate.

<u>Solution:</u> The analysis of the algorithm A_1 yields:

$$\tilde{y}_1 = \ln\left(\frac{x_1}{x_2}(1+\eta_d)\right)(1+\eta_l) = \left[\ln\left(\frac{x_1}{x_2}\right) + \ln(1+\eta_d)\right](1+\eta_l) = \left[y + \eta_d\right](1+\eta_l)$$

$$\tilde{y}_1 = y\left(1 + \frac{\eta_d}{y}\right)(1+\eta_l) = y\left(1 + \frac{\eta_d}{y} + \eta_l\right)$$

$$\delta[\tilde{y}_1] = \frac{\eta_d}{y} + \eta_l \quad \Rightarrow \quad \left|\delta[\tilde{y}_1]\right| \le \left(\frac{1}{|y|} + 1\right) eps = \left(\frac{1}{y} + 1\right) eps$$

The analysis of the algorithm A_2 yields:

$$\tilde{y}_2 = \left[\ln(x_1)(1+\eta_1) - \ln(x_2)(1+\eta_2)\right](1+\eta_o) = \left[\left(\ln(x_1) - \ln(x_2)\right) + \left(\ln(x_1)\eta_1 - \ln(x_2)\eta_2\right)\right](1+\eta_o)$$

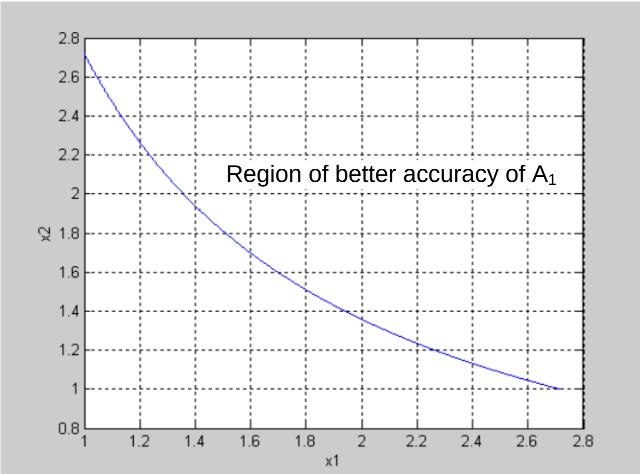
$$\tilde{y}_2 = \left[y + \left(\ln(x_1)\eta_1 - \ln(x_2)\eta_2\right)\right](1+\eta_o) = y\left[1 + \frac{\ln(x_1)\eta_1 - \ln(x_2)\eta_2}{y} + \eta_o\right]$$

$$\delta[\tilde{y}_2] = \frac{\ln(x_1)\eta_1 - \ln(x_2)\eta_2}{y} + \eta_o$$

$$\Rightarrow \quad \left|\delta[\tilde{y}_2]\right| \le \left(\frac{|\ln(x_1)| + |\ln(x_2)|}{|y|} + 1\right) eps = \left(\frac{\ln(x_1) + \ln(x_2)}{y} + 1\right) eps$$

Thus, better accuracy guarantees provides A_1 if $\left(\frac{1}{y} + 1\right) eps < \left(\frac{\ln(x_1) + \ln(x_2)}{y} + 1\right) eps$, *i.e.*:

$$\frac{1}{y} + 1 < \frac{\ln(x_1) + \ln(x_2)}{y} + 1 \text{ or } \ln(x_1) + \ln(x_2) > 1$$



Problem: Determine and draw the function $K(x)$, characterizing the propagation of the relative errors resulting from rounding the result of computing the logarithm in the following operator:

$$y = \sqrt[3]{\frac{\ln(x) + 1}{\ln(x) - 1}} \quad \text{for } x > e$$

Determine $K(x)$ using: a) the differentiation method, b) the "epsilon" calculus.

<u>Solution:</u> The differentiation method yields:

$$y = \sqrt[3]{\frac{\ln(x) + 1}{\ln(x) - 1}} = \left(\frac{v+1}{v-1}\right)^{\frac{1}{3}}, \text{ where } v \equiv \ln(x)$$

$$\frac{\partial y}{\partial v} = \frac{1}{3}\left(\frac{v+1}{v-1}\right)^{-\frac{2}{3}} \frac{(v-1) - (v+1)}{(v-1)^2} = -\frac{2}{3}\left(\frac{v-1}{v+1}\right)^{\frac{2}{3}} \frac{1}{(v-1)^2}$$

$$K(x) \equiv \frac{v}{y}\frac{\partial y}{\partial v} = -\frac{2}{3} v \left(\frac{v-1}{v+1}\right)^{\frac{1}{3}} \left(\frac{v-1}{v+1}\right)^{\frac{2}{3}} \frac{1}{(v-1)^2} = -\frac{2}{3} v \frac{v-1}{v+1} \frac{1}{(v-1)^2}$$

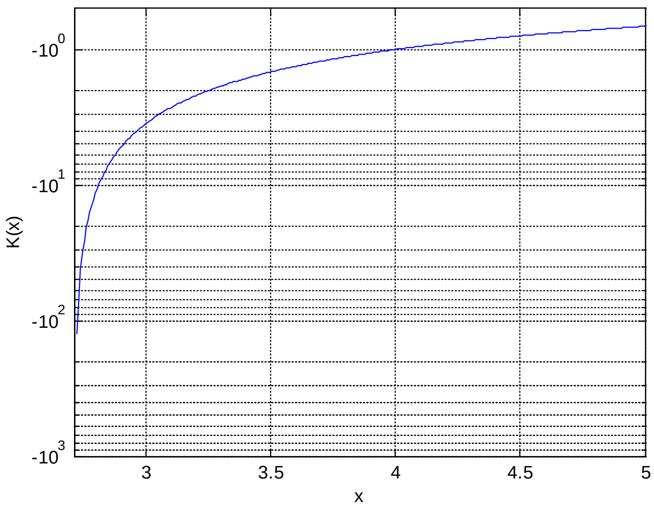
$$K(x) = -\frac{2}{3}\frac{v}{v^2 - 1} = -\frac{2}{3}\frac{\ln(x)}{[\ln(x)]^2 - 1}$$

The "epsilon" calculus yields:

$$\tilde{y} = \sqrt[3]{\frac{\ln(x)(1+\eta)+1}{\ln(x)(1+\eta)-1}} = \sqrt[3]{\frac{\ln(x)+1+\ln(x)\eta}{\ln(x)-1+\ln(x)\eta}} = y\left[\frac{1+\dfrac{\ln(x)}{\ln(x)+1}\eta}{1+\dfrac{\ln(x)}{\ln(x)-1}\eta}\right]^{\frac{1}{3}}$$

$$\delta[\tilde{y}] = \frac{1}{3}\frac{\ln(x)}{\ln(x)+1}\eta - \frac{1}{3}\frac{\ln(x)}{\ln(x)-1}\eta = \frac{1}{3}\ln(x)\left[\frac{1}{\ln(x)+1} - \frac{1}{\ln(x)-1}\right]\eta$$

$$K(x) = -\frac{2}{3}\frac{\ln(x)}{\left[\ln(x)\right]^2-1}$$




```
clear all
x=linspace(exp(1),10,1000);
v=log(x);
K=(-2/3)*v./(v.*v-1);
semilogy(x,K)
xlabel('x')
ylabel('K(x)')
axis([exp(1.00001) 5 -1000 -0.5])
grid on
```

Problem: Determine and draw the function $T(x)$, characterizing the propagation of the relative errors of rounding the result of computing the logarithm in the following operator:

$$y = \sqrt{\frac{1+\sin(x)}{1-\sin(x)}} \quad \text{for } x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

Determine $K(x)$ using: a) the differentiation method, b) the "epsilon" calculus.

Solution: The differentiation method yields:

$$y = \left(\frac{1+v}{1-v}\right)^{\frac{1}{2}}, \text{ where } v \equiv \sin(x)$$

$$\frac{\partial y}{\partial v} = \frac{1}{2}\left(\frac{1+v}{1-v}\right)^{-\frac{1}{2}}\frac{(1-v)+(1+v)}{(1-v)^2} = \left(\frac{1-v}{1+v}\right)^{\frac{1}{2}}\frac{1}{(1-v)^2}$$

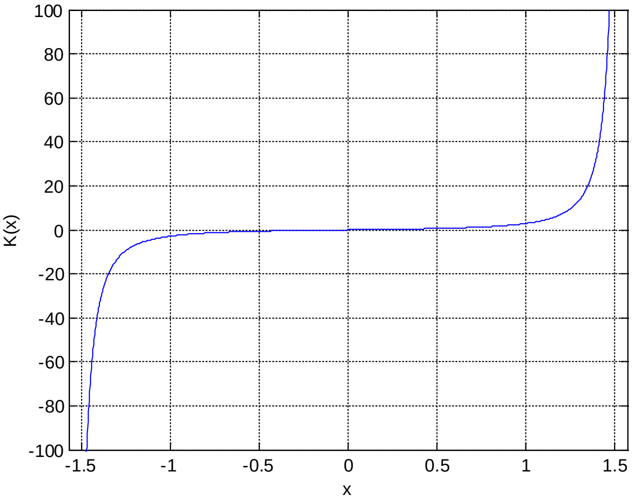
$$K(x) \equiv \frac{v}{y}\frac{\partial y}{\partial v} = v\left(\frac{1-v}{1+v}\right)^{\frac{1}{2}}\left(\frac{1-v}{1+v}\right)^{\frac{1}{2}}\frac{1}{(1-v)^2} = v\frac{1-v}{(1+v)(1-v)^2} = \frac{v}{1-v^2} = \frac{\sin(x)}{1-\left[\sin(x)\right]^2}$$

The "epsilon" calculus yields:

$$\tilde{y} = \sqrt{\frac{1+\sin(x)(1+\eta)}{1-\sin(x)(1+\eta)}} = \sqrt{\frac{1+\sin(x)+\sin(x)\eta}{1-\sin(x)-\sin(x)\eta}} = y\left[\frac{1+\dfrac{\sin(x)}{1+\sin(x)}\eta}{1-\dfrac{\sin(x)}{1-\sin(x)}\eta}\right]^{\frac{1}{2}}$$

$$\delta[\tilde{y}] = \frac{1}{2}\frac{\sin(x)}{1+\sin(x)}\eta + \frac{1}{2}\frac{\sin(x)}{1-\sin(x)}\eta = \frac{1}{2}\sin(x)\left[\frac{1}{1+\sin(x)}+\frac{1}{1-\sin(x)}\right]\eta$$

$$K(x) = \frac{\sin(x)}{1-\left[\sin(x)\right]^2}$$




```
clear all
x=linspace(-pi/2,pi/2,1000);
v=sin(x);
K=v./(1-v.*v);
plot(x,K)
xlabel('x')
ylabel('K(x)')
axis([-pi/2 pi/2 -100 100])
grid on
```

Problem: Assess the relative error of computing:

$$y = \sqrt{1+\frac{1}{x}} - 1 \text{ for } x \in \left(10^6, 10^9\right),$$

caused by the rounding error in the result of square-root calculation.

<u>Solution:</u> Let's denote $v = \sqrt{1+\frac{1}{x}}$; then:

$$\tilde{y} = v(1+\eta) - 1 = y + v\eta = y\left(1+\frac{v\eta}{y}\right)$$

$$\delta[\tilde{y}] = \frac{v\eta}{y} = \frac{\sqrt{1+\frac{1}{x}}}{\sqrt{1+\frac{1}{x}}-1}\eta \cong \frac{1+\frac{1}{2x}}{1+\frac{1}{2x}-1}\eta = (2x+1)\eta$$

$$\left|\delta[\tilde{y}]\right| \le \left|(2x+1)\right| eps \le 2\cdot 10^9 eps$$

Problem: Assess the error of computing:

$$y = \frac{x^2}{x^2+1} \quad \text{dla } x \in (-\infty, +\infty)$$

caused by the rounding of the results of floating-point operations.

Solution:

$$\tilde{y} = \frac{x^2\left(1+\eta_p\right)}{\left[x^2\left(1+\eta_p\right)+1\right]\left(1+\eta_s\right)}\left(1+\eta_d\right)$$

$$\tilde{y} = \frac{x^2\left(1+\eta_p+\eta_d-\eta_s\right)}{\left(x^2+1\right)\left[1+\frac{x^2}{x^2+1}\eta_p\right]} = \frac{x^2\left(1+\eta_p+\eta_d-\eta_s-\frac{x^2}{x^2+1}\eta_p\right)}{\left(x^2+1\right)}$$

$$\delta\left[\tilde{y}\right] = \eta_p+\eta_d-\eta_s-\frac{x^2}{x^2+1}\eta_p = \eta_d-\eta_s+\frac{1}{x^2+1}\eta_p$$

$$\left|\delta\left[\tilde{y}\right]\right| \le \left|\eta_d\right|+\left|\eta_s\right|+\frac{1}{x^2+1}\left|\eta_p\right| \quad \le 3eps$$

Problem: Determine the functions $K(x)$, characterizing the propagation of the relative error caused by rounding x^2 during evaluation of the following expression:

$$y = \frac{1+x+x^2+x^3}{1-x+x^2-x^3} \quad \text{for } x \in [0,10].$$

Solution #1: Let's denote: $a \equiv 1+x+x^3$, $b \equiv 1-x-x^3$ and $v \equiv x^2$. Then:

$$y = \frac{a+v}{b+v} \Rightarrow \frac{dy}{dv} = \frac{(b+v)-(a+v)}{(b+v)^2} = \frac{b-a}{(b+v)^2} \Rightarrow K = \frac{v}{y}\frac{dy}{dv} = v\frac{b+v}{a+v}\frac{b-a}{(b+v)^2} = \frac{v(b-a)}{(a+v)(b+v)}$$

$$K = \frac{x^2\left[\left(1-x-x^3\right)-\left(1+x+x^3\right)\right]}{\left(1+x+x^2+x^3\right)\left(1-x+x^2-x^3\right)} = \frac{-2x^2\left(x+x^3\right)}{\left[\left(1+x^2\right)+\left(x+x^3\right)\right]\left[\left(1+x^2\right)-\left(x+x^3\right)\right]}$$

$$K = \frac{-2x^2\left(x+x^3\right)}{\left(1+x^2\right)^2-\left(x+x^3\right)^2} = \frac{-2x^3\left(1+x^2\right)}{\left(1+x^2\right)^2-x^2\left(1+x^2\right)^2} = \frac{-2x^3}{\left(1+x^2\right)-x^2\left(1+x^2\right)} = \frac{-2x^3}{\left(1-x^2\right)\left(1+x^2\right)}$$

$$K(x) = \frac{-2x^3}{1-x^4}$$

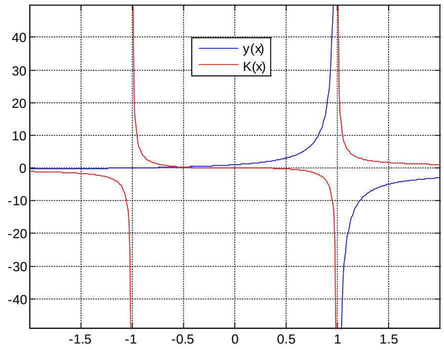
Solution #2: Let's denote with η the rounding error of the result of computing x^2. Then:

$$\tilde{y} = \frac{1+x+x^2(1+\eta)+x^3}{1-x+x^2(1+\eta)-x^3} = \frac{1+x+x^2+x^3+x^2\eta}{1-x+x^2-x^3+x^2\eta} = \frac{1+x+x^2+x^3}{1-x+x^2-x^3}\cdot\frac{1+\frac{x^2}{1+x+x^2+x^3}\eta}{1+\frac{x^2}{1-x+x^2-x^3}\eta}$$

$$\tilde{y} = y\cdot\frac{1+\frac{x^2}{1+x+x^2+x^3}\eta}{1+\frac{x^2}{1-x+x^2-x^3}\eta} = y\left(1+\frac{x^2}{1+x+x^2+x^3}\eta-\frac{x^2}{1-x+x^2-x^3}\eta\right)$$

$$K(x) = \frac{x^2}{1+x+x^2+x^3}-\frac{x^2}{1-x+x^2-x^3} = x^2\frac{\left(1-x+x^2-x^3\right)-\left(1+x+x^2+x^3\right)}{\left(1+x+x^2+x^3\right)\left(1-x+x^2-x^3\right)}$$

$$K(x) = x^2\frac{\left(1-x+x^2-x^3\right)-\left(1+x+x^2+x^3\right)}{\left(1+x+x^2+x^3\right)\left(1-x+x^2-x^3\right)} = \ldots = \frac{-2x^3}{1-x^4}$$



Problem: Determine the functions $K_{NO}\left(x\right)$, characterizing the propagation of rounding errors related to all elementary operations the following functions are composed of:

$$y=\frac{\sin\left(x\right)}{x} \quad \text{for } x\in\left[0,10\pi\right],$$
$$y=x\sin\left(x\right) \quad \text{for } x\in\left[0,10\pi\right],$$
$$y=e^{x}\sin\left(x\right) \quad \text{for } x\in\left[0,10\pi\right],$$
$$y=xe^{-x}\sin\left(x\right) \quad \text{for } x\in\left[0,2\pi\right],$$
$$y=x^{2}e^{-x}\sin\left(x\right) \quad \text{for } x\in\left[0,2\pi\right],$$

Verify the results by numerical simulation of errors in MATLAB. Assess the total relative error in the result of computation $\tilde{y}$, caused by representation errors and rounding errors. Draw the graphs of all $y\left(x\right)$ and $K_{NO}\left(x\right)$.

Problem: Assess the error of computing:

$$y=\frac{\ln\left(x\right)}{\ln\left(x\right)+1} \quad \text{dla } x\in\left(1,+\infty\right)$$

caused by rounding the results of floating-point operations.

<u>Solution:</u>

$$\tilde{y}=\frac{\ln\left(x\right)\left(1+\eta_{l}\right)}{\left[\ln\left(x\right)\left(1+\eta_{l}\right)+1\right]\left(1+\eta_{s}\right)}\left(1+\eta_{d}\right)$$

$$\tilde{y}=\frac{\ln\left(x\right)\left(1+\eta_{l}+\eta_{d}-\eta_{s}\right)}{\left(\ln\left(x\right)+1\right)\left[1+\frac{\ln\left(x\right)}{\ln\left(x\right)+1}\eta_{l}\right]}=\frac{\ln\left(x\right)\left(1+\eta_{l}+\eta_{d}-\eta_{s}-\frac{\ln\left(x\right)}{\ln\left(x\right)+1}\eta_{l}\right)}{\ln\left(x\right)+1}$$

$$\delta\left[\tilde{y}\right]=\eta_{l}+\eta_{d}-\eta_{s}-\frac{\ln\left(x\right)}{\ln\left(x\right)+1}\eta_{l}=\eta_{d}-\eta_{s}+\frac{1}{\ln\left(x\right)+1}\eta_{l}$$

$$\left|\delta\left[\tilde{y}\right]\right|\le\left|\eta_{d}\right|+\left|\eta_{s}\right|+\frac{1}{\left|\ln\left(x\right)+1\right|}\left|\eta_{l}\right|\le 3eps$$

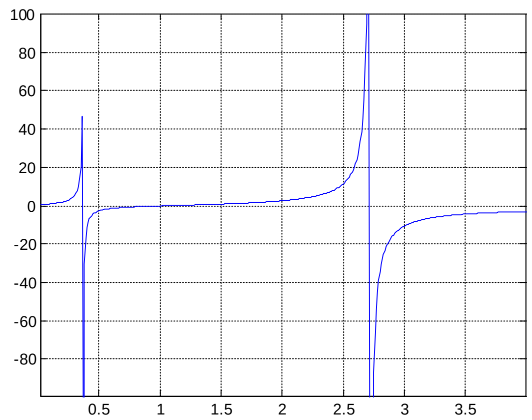
Problem: Determine the function $K\left(x\right)$ characterising the propagation of the relative rounding error associated with the $\ln\left(x\right)$ computation in the following expression:

$$y=\frac{1+\ln(x)}{1-\ln(x)} \quad \text{for } x\in(0,\ +\infty)$$

Solution:

$$K(x)=\frac{dy}{dv}\cdot\frac{v}{y}\bigg|_{v=\ln(x)}=\left(\frac{1+v}{1-v}\right)'\cdot v\cdot\frac{1-v}{1+v}\bigg|_{v=\ln(x)}=\frac{2}{(1-v)^2}\cdot v\cdot\frac{1-v}{1+v}\bigg|_{v=\ln(x)}=\frac{2v}{1-v^2}\bigg|_{v=\ln(x)}$$

$$K(x)=\frac{2\ln(x)}{1-\left[\ln(x)\right]^2}$$



1.3. Propagation of data errors and rounding errors

Problem: Construct a numerically correct algorithm for computing:

$$y=\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}-x^{\frac{1}{6}} \quad \text{dla } \ x>>1$$

Assess the error of computing y by means of this algorithm, caused by the rounding of the results of floating-point operations.

Solution:

$$y=\frac{\left[\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}-x^{\frac{1}{6}}\right]\left[\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}+x^{\frac{1}{6}}\right]}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}+x^{\frac{1}{6}}}=\frac{x^{\frac{1}{3}}+1-x^{\frac{1}{3}}}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}+x^{\frac{1}{6}}}=\frac{1}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}+x^{\frac{1}{6}}}$$

$$\tilde{y}=\frac{1+\eta_d}{\left[\left\{\left[x^{\frac{1}{3}}(1+\eta_3)+1\right](1+\eta_s')\right\}^{\frac{1}{2}}(1+\eta_2)+x^{\frac{1}{6}}(1+\eta_6)\right](1+\eta_s'')}=$$

$$=\frac{1+\eta_d-\eta''}{\left[x^{\frac{1}{3}}(1+\eta_3)+1\right]^{\frac{1}{2}}\left(1+\frac{1}{2}\eta_s'+\eta_2\right)+x^{\frac{1}{6}}(1+\eta_6)}=$$

$$=\frac{1+\eta_d-\eta_s''}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}\left(1+\frac{x^{\frac{1}{3}}\eta_3}{x^{\frac{1}{3}}+1}\right)^{\frac{1}{2}}\left(1+\frac{1}{2}\eta_s'+\eta_2\right)+x^{\frac{1}{6}}\left(1+\eta_6\right)}=$$

$$=\frac{1+\eta_d-\eta_s''}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}\left(1+\eta\right)+x^{\frac{1}{6}}\left(1+\eta_6\right)}$$

where $\eta\equiv\frac{1}{2}\frac{x^{\frac{1}{3}}\eta_3}{x^{\frac{1}{3}}+1}+\frac{1}{2}\eta_s'+\eta_2\cong\frac{1}{2}\eta_3+\frac{1}{2}\eta_s'+\eta_2$, because $x>>1$. Hence:

$$\tilde{y}=\frac{1+\eta_d-\eta_s''}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}\left(1+\eta\right)+x^{\frac{1}{6}}\left(1+\eta_6\right)}=\frac{1+\eta_d-\eta_s''}{\left[\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}+x^{\frac{1}{6}}\right]\left[1+\frac{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}\eta+x^{\frac{1}{6}}\eta_6}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}+x^{\frac{1}{6}}}\right]}$$

$$\delta\left[\tilde{y}\right]=\eta_d-\eta_s''-\frac{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}\eta+x^{\frac{1}{6}}\eta_6}{\left(x^{\frac{1}{3}}+1\right)^{\frac{1}{2}}+x^{\frac{1}{6}}}\cong\eta_d-\eta_s''-\frac{x^{\frac{1}{6}}\eta+x^{\frac{1}{6}}\eta_6}{2x^{\frac{1}{6}}}=\eta_d-\eta_s''-\frac{1}{2}\eta-\frac{1}{2}\eta_6$$

because $x>>1$. After substituting η:

$$\delta\left[\tilde{y}\right]=\eta_d-\eta_s''-\frac{1}{2}\left(\frac{1}{2}\eta_3+\frac{1}{2}\eta_s'+\eta_2\right)-\frac{1}{2}\eta_6$$

$$\delta\left[\tilde{y}\right]=\eta_d-\eta_s''-\frac{1}{4}\eta_3-\frac{1}{4}\eta_s'-\frac{1}{2}\eta_2-\frac{1}{2}\eta_6$$

$$\left|\delta\left[\tilde{y}\right]\right|\le\left(1+1+\frac{1}{4}+\frac{1}{4}+\frac{1}{2}+\frac{1}{2}\right)eps=3.5eps$$

Problem: Demonstrate that the algorithm:

$$A:\begin{bmatrix}x_1\\x_2\end{bmatrix}\rightarrow\begin{bmatrix}v_1=x_1^3\\v_2=x_2^2\end{bmatrix}\rightarrow y=\frac{v_1}{v_2+1}$$

is numerically correct (stable).

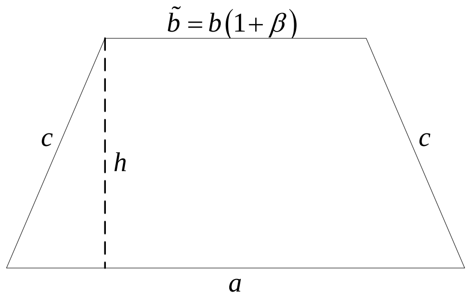
<u>Solution:</u> The analysis of A yields:

$$\tilde{y}=\frac{x_1^3\left(1+\eta_1\right)}{\left[x_2^2\left(1+\eta_2\right)+1\right]\left(1+\eta_s\right)}\left(1+\eta_d\right)=\frac{x_1^3\left(1+\eta_1-\eta_s+\eta_d\right)}{x_2^2\left(1+\eta_2\right)+1}=\frac{\left[x_1\left(1+\frac{\eta_1-\eta_s+\eta_d}{3}\right)\right]^3}{\left[x_2\left(1+\frac{\eta_2}{2}\right)\right]^2+1}$$

which means that the effect of rounding errors is equivalent to the effect of data errors that do not exceed *eps*.

1.4. Non-numerical applications

Problem: Assess the relative change of the area of a trapezoid (shown in the figure below), caused by a small relative change of its upper basis. Make calculations for $a=3$, $b=1$, $c=2$ and $|\beta| \le 0.12\%$.



Solution: The squared hight h of the trapezoid depends on the basis b in the following way:

$$h^2 = c^2 - \left(\frac{a-b}{2}\right)^2 = 4 - \left(\frac{3-b}{2}\right)^2 = \tfrac{7}{4} + \tfrac{6}{4}b - \tfrac{1}{4}b^2 \xrightarrow{b=1} 3$$

Consequently, the squared area of the trapezoid may be related to b by the equation:

$$P^2 = \tfrac{1}{4}(a+b)^2 h^2 = \tfrac{1}{4}\left(9+6b+b^2\right)\left(\tfrac{7}{4}+\tfrac{6}{4}b-\tfrac{1}{4}b^2\right) \xrightarrow{b=1} 12$$

Hence: $16P^2 = \left(9+6b+b^2\right)\left(7+6b-b^2\right)$, and:

$$16\left(2P\frac{dP}{db}\right) = (6+2b)\left(7+6b-b^2\right) + \left(9+6b+b^2\right)(6-2b) \xrightarrow{b=1} 160$$

This means that $P\dfrac{dP}{db} = 5$, the corresponding coefficient of error propagation is:

$$T_b = \frac{b}{P}\frac{dP}{db} = \frac{b}{P}\frac{5}{P} = \frac{5}{P^2} = \tfrac{5}{12}$$

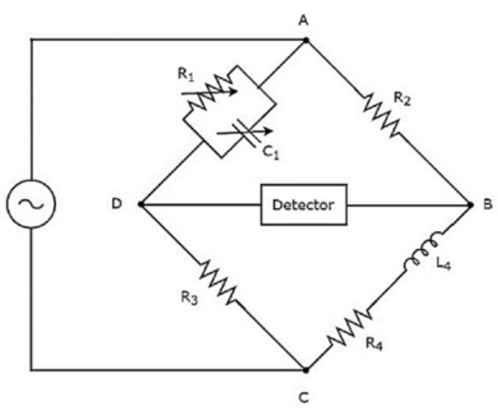
and the relative error of the area is subject to the assessment:

$$\left|\delta\left[\tilde{P}\right]\right| \cong \left|T_b \beta\right| \le \tfrac{5}{12} 0.12\% = 0.05\%$$

The computed value of T_b has been checked by means of the following MATLAB script:

```
clear all
a=3; b=1; c=2;
h=sqrt(c^2-0.25*(a-b)^2);
P=0.5*(a+b)*h;
b=1.0001;
h=sqrt(c^2-0.25*(a-b)^2);
P1=0.5*(a+b)*h;
Tb=(P1-P)/P/0.0001
```

Problem: The Maxwell's bridge, shown below, is used to measure the value of the resistance R_4 and inductance L_4. Their values are determined according to the formula:



$$\hat{R}_4 = \frac{\dot{R}_2 \dot{R}_3}{\dot{R}_1}, \quad \hat{L}_4 = \dot{C}_1 \dot{R}_2 \dot{R}_3$$

under the assumption that the voltage between the points D and B is zero. Assess the absolute errors of measurement implied by the error $\Delta\dot{U}$ corrupting the measured value of this voltage. Carry out computation for: $\dot{R}_1 = \dot{R}_2 = \dot{R}_3 = 1\,\mathrm{k}\Omega$, $\dot{C}_1 = 1\,\mu\mathrm{F}$ and the amplitude of the source signal $\dot{E} = 1\,\mathrm{V}$.

Solution: Using the following symbols of impedances:

$$Z_1 = \frac{R_1}{1 + j\omega R_1 C_1} \text{ and } Z_4 = R_4 + j\omega L_4$$

one may express the equilibrium condition by means of the equation:

$$Z_4 = \frac{R_2 R_3}{Z_1} = \frac{R_2 R_3}{R_1} + j\omega C_1 R_2 R_3 \Rightarrow R_4 = \frac{R_2 R_3}{R_1},\ L_4 = C_1 R_2 R_3 \Rightarrow \dot{R}_4 = 1\,\text{k}\Omega,\ \dot{L}_4 = 1\,\text{H}$$

The difference of the potentials determined by two voltage dividers, *i.e.* the voltage between the points D and B:

$$\dot{U} = \left(\frac{\dot{Z}_4}{\dot{R}_2 + \dot{Z}_4} - \frac{\dot{R}_3}{\dot{Z}_1 + \dot{R}_3} \right) \dot{E}$$

after substitution of $\dot{R}_2 = \dot{R}_3 = 1\,\text{k}\Omega$ and $\dot{E} = 1\,\text{V}$ takes on the form:

$$\dot{U} = \frac{\dot{Z}_4}{1 + \dot{Z}_4} - \frac{1}{\dot{Z}_1 + 1}$$

Hence:

$$\dot{Z}_4 = \frac{\dot{U}\left(\dot{Z}_1 + 1\right) + 1}{\left(1 - \dot{U}\right)\left(\dot{Z}_1 + 1\right) - 1} \text{ and } \hat{Z}_4 = \frac{1}{\dot{Z}_1}$$

The errors of measurement may be, therefore, assessed as follows:

$$\Delta\hat{Z}_4 = \frac{1}{\dot{Z}_1} - \frac{\dot{U}\left(\dot{Z}_1 + 1\right) + 1}{\left(1 - \dot{U}\right)\left(\dot{Z}_1 + 1\right) - 1} - = -\frac{\left(\dot{Z}_1 + 1\right)^2 \Delta\dot{U}}{\dot{Z}_1\left(\dot{Z}_1 - \dot{Z}_1\Delta\dot{U} - \Delta\dot{U}\right)} \cong -\frac{\left(\dot{Z}_1 + 1\right)^2}{\dot{Z}_1^2}\Delta\dot{U} = -\left(1 + \frac{1}{\dot{Z}_1}\right)^2 \Delta\dot{U}$$

$$\cong -\left(1 + \frac{1}{\dot{R}_1} + j\omega\dot{C}_1\right)^2 \Delta\dot{U} = \left[\left(1 + \frac{1}{\dot{R}_1}\right)^2 - \left(\omega\dot{C}_1\right)^2\right]\Delta\dot{U} + j\omega\left[2\left(1 + \frac{1}{\dot{R}_1}\right)\dot{C}_1\right]\Delta\dot{U}$$

or:

$$\Delta\hat{R}_4 \cong \left[\left(1 + \frac{1}{\dot{R}_1}\right)^2 - \left(\omega\dot{C}_1\right)^2\right]\Delta\dot{U} = \left(4 - \omega^2\right)\Delta\dot{U}\ [\text{k}\Omega] \text{ and } \Delta\hat{L}_4 \cong 2\left(1 + \frac{1}{\dot{R}_1}\right)\dot{C}_1\Delta\dot{U} = 4\Delta\dot{U}\ [\text{H}]$$

Problem: Assess the relative deviation of the cut-off frequency of a low-pass filter caused by imperfection of the thermostat, under the following assumptions:

– The main cause of the frequency deviation is the thermal instability of a resistor and of a capacitor.

– The dependence of the cut-off frequency $f(T)$ on the temperature-varying values of the resistance $R(T)$ and capacitance $C(T)$, where T stands for temperature, may be adequately modelled by the following equation:

$$f(T) = \frac{1}{2\pi\, R(T)\, C(T)}$$

– The dependence of the resistance and capacitance on temperature, in the vicinity of the reference value of temperature T_0, may be adequately modelled by the following equations:

$$R(T) = \frac{R_0}{1 + 2\cdot 10^{-6}\left(T - T_0\right)} \text{ and } C(T) = C_0\left(1 + 10^{-2}\left(T - T_0\right)^3\right)$$

– The guaranteed range of temperature stabilisation is $T_0 \pm 0.01^{\circ}$.

<u>Solution:</u> The substitution yields:

$$f(T)=\frac{1}{2\pi\,R(T)C(T)}=\frac{1+2\cdot10^{-6}(T-T_0)}{2\pi\,R_0\,C_0\left(1+10^{-2}(T-T_0)^3\right)}=f(T_0)\frac{1+2\cdot10^{-6}(T-T_0)}{1+10^{-2}(T-T_0)^3}$$

Since $|T-T_0|\le 0.01$, the "epsilon" calculus may be applied for error evaluation:

$$\frac{f(T)}{f(T_0)}=\frac{1+2\cdot10^{-6}(T-T_0)}{1+10^{-2}(T-T_0)^3}\cong 1+2\cdot10^{-6}(T-T_0)-10^{-2}(T-T_0)^3$$

This means that the relative deviation of $f(T)$ may be assessed in the following way:

$$\delta\left[f(T)\right]\cong\underbrace{2\cdot10^{-6}}_{a_1}\underbrace{(T-T_0)}_{x}-\underbrace{10^{-2}}_{a_3}\underbrace{(T-T_0)^3}_{x^3}\equiv\varphi(x)$$

The following values of the function $\varphi(x)\equiv a_1x-a_3x^3$ for $|x|\le10^{-2}$ have to be compared in order to find $\sup\left|\delta\left[f(T)\right]\right|$:

$$\varphi\left(-10^{-2}\right),\ \varphi\left(10^{-2}\right),\ \varphi(x_{\min})\ \text{and}\ \varphi(x_{\max})$$

where $x_{\min}$ and $x_{\min}$ are abscissas of the minimum and maximum of $\varphi(x)$, which may be determined from the necessary condition for an extremum, *viz.*: $a_1-3a_3x^2=0$; their values are:

$$x_{\min}=-\sqrt{\frac{a_1}{3a_3}}=-\sqrt{\frac{2\cdot10^{-6}}{3\cdot10^{-2}}}=-\sqrt{\tfrac{2}{3}}\cdot10^{-2}\cong-0.0082\ \text{and}\ x_{\max}=\sqrt{\tfrac{2}{3}}\cdot10^{-2}=0.0082$$

Since:

$$\varphi\left(-10^{-2}\right)=10^{-8},\ \varphi\left(10^{-2}\right)=-10^{-8},\ \varphi(x_{\min})\cong-1.0887\cdot10^{-8}\ \text{and}\ \varphi(x_{\max})\cong1.0887\cdot10^{-8}$$

the relative error $\delta\left[f(T)\right]$ may be assessed as follows:

$$\left|\delta\left[f(T)\right]\right|\le1.0887\cdot10^{-8}.$$

ENUME: SOLVED PROBLEMS

2. METHODS FOR SOLVING LINEAR ALGEBRAIC EQUATIONS

2.1. Matrix factorisation

Problem: Apply the Cholesky-Banachiewicz decomposition to the following matrix.

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

Verify the obtained result by multiplying $\mathbf{L}$ and $\mathbf{L}^T$.

<u>Solution:</u> The equality:

$$\mathbf{L}\mathbf{L}^T = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \cdot \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 2 & 2 \end{bmatrix}$$

implies the following:

$$l_{11}^2 = 1 \;\Rightarrow\; l_{11} = 1$$

$$l_{11} \cdot l_{21} = 1 \;\Rightarrow\; l_{21} = 1$$

$$l_{11} \cdot l_{31} = 0 \;\Rightarrow\; l_{31} = 0$$

$$l_{21}^2 + l_{22}^2 = 3 \;\Rightarrow\; l_{22} = \sqrt{3 - l_{21}^2} = \sqrt{2}$$

$$l_{21} \cdot l_{31} + l_{22} \cdot l_{32} = 2 \;\Rightarrow\; l_{32} = \frac{2 - l_{21} \cdot l_{31}}{l_{22}} = \sqrt{2}$$

$$l_{31}^2 + l_{32}^2 + l_{33}^2 = 2 \;\Rightarrow\; l_{33} = \sqrt{2 - l_{31}^2 - l_{32}^2} = 0$$

Thus:

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

Problem: Provide the result of the Cholesky-Banachiewicz factorisation applied to the matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{bmatrix}$$

<u>Solution:</u> $\mathbf{A} = \mathbf{L} \cdot \mathbf{L}^T$, *i.e.*:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \cdot \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

$$= \begin{bmatrix} (l_{11})^2 & l_{11}\cdot l_{21} & l_{11}\cdot l_{31} \\ l_{21}\cdot l_{11} & (l_{21})^2+(l_{22})^2 & l_{21}\cdot l_{31}+l_{22}\cdot l_{32} \\ l_{31}\cdot l_{11} & l_{31}\cdot l_{21}+l_{32}\cdot l_{22} & (l_{31})^2+(l_{32})^2+(l_{33})^2 \end{bmatrix}$$

Hence:

$$(l_{11})^2=1 \;\Rightarrow\; l_{11}=1;\; l_{11}\cdot l_{21}=2 \;\Rightarrow\; l_{21}=2;\; l_{11}\cdot l_{31}=3 \;\Rightarrow\; l_{31}=3;$$
$$(l_{21})^2+(l_{22})^2=5 \;\Rightarrow\; l_{22}=\sqrt{5-4}=1;\; l_{21}\cdot l_{31}+l_{22}\cdot l_{32}=8 \;\Rightarrow\; l_{32}=8-6=2;$$
$$(l_{31})^2+(l_{32})^2+(l_{33})^2=14 \;\Rightarrow\; l_{33}=\sqrt{14-9-4}=1.$$

Thus:

$$\mathbf{L}=\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix}$$

Problem: Perform the LU factorization of the matrices:

$$\mathbf{A}=\begin{bmatrix} 1 & 2 & 4 \\ 2 & 4 & 8 \\ 1 & 2 & 16 \end{bmatrix},\; \mathbf{A}=\begin{bmatrix} 1 & 2 & 4 \\ 2 & 0 & 8 \\ 1 & 2 & 16 \end{bmatrix},\; \mathbf{A}=\begin{bmatrix} 0 & 2 & 4 \\ 2 & 0 & 8 \\ 1 & 2 & 16 \end{bmatrix} \text{ and } \mathbf{A}=\begin{bmatrix} 0 & 2 & 4 \\ 2 & 0 & 8 \\ 1 & 2 & 0 \end{bmatrix}.$$

Verify the results by multiplying the corresponding matrices $\mathbf{L}$ and $\mathbf{U}$ (do not forget about the permutation matrix).

2.2. Error propagation

Problem: Let's assume that $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T=\tilde{\mathbf{A}}$, where:

$$\tilde{\mathbf{A}}=\begin{bmatrix} 1+\alpha_{11} & 1+\alpha_{12} \\ 1+\alpha_{12} & 2(1+\alpha_{22}) \end{bmatrix} \text{ and } |\alpha_{11}|,|\alpha_{12}|,|\alpha_{22}|\le eps$$

Assess the relative errors $\delta[\tilde{l}_{11}]$, $\delta[\tilde{l}_{21}]$ and $\delta[\tilde{l}_{22}]$ of the elements of the matrix:

$$\tilde{\mathbf{L}}=\begin{bmatrix} \tilde{l}_{11} & 0 \\ \tilde{l}_{21} & \tilde{l}_{22} \end{bmatrix}$$

caused by the relative disturbances α_{11}, α_{12} and α_{22} of the elements of the matrix $\tilde{\mathbf{A}}$.

<u>Solution:</u> The equality:

$$\tilde{\mathbf{L}}\tilde{\mathbf{L}}^T=\begin{bmatrix} \tilde{l}_{11} & 0 \\ \tilde{l}_{21} & \tilde{l}_{22} \end{bmatrix}\cdot\begin{bmatrix} \tilde{l}_{11} & \tilde{l}_{21} \\ 0 & \tilde{l}_{22} \end{bmatrix}=\begin{bmatrix} \tilde{l}_{11}^2 & \tilde{l}_{11}\cdot\tilde{l}_{21} \\ \tilde{l}_{11}\cdot\tilde{l}_{21} & \tilde{l}_{21}^2+\tilde{l}_{22}^2 \end{bmatrix}=\begin{bmatrix} 1+\alpha_{11} & 1+\alpha_{12} \\ 1+\alpha_{12} & 2(1+\alpha_{22}) \end{bmatrix}=\tilde{\mathbf{A}}$$

implies:

$$\tilde{l}_{11}^2=1+\alpha_{11} \;\Rightarrow \tilde{l}_{11}=\sqrt{1+\alpha_{11}}=1+\frac{\alpha_{11}}{2} \;\Rightarrow \left|\delta\left[\tilde{l}_{11}\right]\right|\le\frac{1}{2}eps$$
$$\tilde{l}_{11}\cdot\tilde{l}_{21}=1+\alpha_{12} \;\Rightarrow \tilde{l}_{21}=\frac{1+\alpha_{12}}{\tilde{l}_{11}}=\frac{1+\alpha_{12}}{1+\frac{\alpha_{11}}{2}}=1+\alpha_{12}-\frac{\alpha_{11}}{2} \Rightarrow \left|\delta\left[\tilde{l}_{21}\right]\right|\le\frac{3}{2}eps$$

$$\tilde{l}_{21}^2+\tilde{l}_{22}^2=2(1+\alpha_{22})\Rightarrow\tilde{l}_{22}=\sqrt{2(1+\alpha_{22})-\tilde{l}_{21}^2}=\sqrt{2(1+\alpha_{22})-\left(1+\alpha_{12}-\frac{\alpha_{11}}{2}\right)^2}$$
$$\Rightarrow\tilde{l}_{22}=\sqrt{2+2\alpha_{22}-1-2\alpha_{12}+\alpha_{11}}=\sqrt{1+2\alpha_{22}-2\alpha_{12}+\alpha_{11}}$$
$$\Rightarrow\tilde{l}_{22}=1+\alpha_{22}-\alpha_{12}+\frac{\alpha_{11}}{2}\Rightarrow\left|\delta\left[\tilde{l}_{22}\right]\right|\le\frac{5}{2}eps$$

Problem: Assess the relative error of the numerical solution of the following system of linear algebraic equations:

$$\begin{bmatrix}100 & 101\\ 102 & 103\end{bmatrix}\cdot\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}1\\ 1\end{bmatrix}$$

caused by one-percent error in the diagonal elements of the LHS matrix.

<u>Solution #1:</u> The consecutive steps are as follows:

$$\begin{bmatrix}100\cdot(1+\alpha_{11}) & 101\\ 102 & 103\cdot(1+\alpha_{22})\end{bmatrix}\cdot\begin{bmatrix}x_1\cdot(1+\delta_1)\\ x_2\cdot(1+\delta_2)\end{bmatrix}=\begin{bmatrix}1\\ 1\end{bmatrix}$$

$$100\cdot(1+\alpha_{11})\cdot x_1\cdot(1+\delta_1)+101\cdot x_2\cdot(1+\delta_2)=1$$
$$102\cdot x_1\cdot(1+\delta_1)+103\cdot(1+\alpha_{22})\cdot x_2\cdot(1+\delta_2)=1$$
$$100\cdot x_1\cdot(1+\alpha_{11}+\delta_1)+101\cdot x_2\cdot(1+\delta_2)=1$$
$$102\cdot x_1\cdot(1+\delta_1)+103\cdot x_2\cdot(1+\alpha_{22}+\delta_2)=1$$

From the first equation:

$$x_1=\frac{1-101\cdot x_2\cdot(1+\delta_2)}{100\cdot(1+\alpha_{11}+\delta_1)}=\frac{1-\alpha_{11}-\delta_1}{100}-\frac{101\cdot x_2\cdot(1+\delta_2-\alpha_{11}-\delta_1)}{100}$$

and

$$102\cdot\left[\frac{1-\alpha_{11}-\delta_1}{100}-\frac{101\cdot x_2\cdot(1+\delta_2-\alpha_{11}-\delta_1)}{100}\right]\cdot(1+\delta_1)+103\cdot x_2\cdot(1+\alpha_{22}+\delta_2)=1$$
$$102\cdot\left[\frac{1-\alpha_{11}}{100}-\frac{101\cdot x_2\cdot(1+\delta_2-\alpha_{11})}{100}\right]+103\cdot x_2\cdot(1+\alpha_{22}+\delta_2)=1$$
$$-102\cdot\frac{101\cdot x_2\cdot(1+\delta_2-\alpha_{11})}{100}+103\cdot x_2\cdot(1+\alpha_{22}+\delta_2)=1-102\cdot\frac{1-\alpha_{11}}{100}$$

Hence equations with respect to x_2 and δ_2:

$$-102\cdot\frac{101\cdot x_2}{100}+103\cdot x_2=1-\frac{102}{100}\;\Rightarrow\;x_2=1$$
$$-102\cdot\frac{101\cdot(\delta_2-\alpha_{11})}{100}+103\cdot(\alpha_{22}+\delta_2)=-102\cdot\frac{-\alpha_{11}}{100}$$
$$-102\cdot\frac{101\cdot\delta_2}{100}+103\cdot\delta_2=102\cdot\frac{\alpha_{11}}{100}-102\cdot\frac{101\cdot\alpha_{11}}{100}-103\cdot\alpha_{22}$$
$$-102\cdot101\cdot\delta_2+100\cdot103\cdot\delta_2=102\cdot\alpha_{11}-102\cdot101\cdot\alpha_{11}-100\cdot103\cdot\alpha_{22}$$
$$\delta_2=\frac{(102-102\cdot101)\cdot\alpha_{11}-100\cdot103\cdot\alpha_{22}}{-102\cdot101+100\cdot103}=\frac{-100\cdot102\cdot\alpha_{11}-100\cdot103\cdot\alpha_{22}}{-2}\cong5000\cdot(\alpha_{11}+\alpha_{22})$$
$$|\delta_2|\le10^4\cdot10^{-2}=10^4\%$$

…

$$\left|\delta_1\right| \le 10^4 \cdot 10^{-2} = 10^4\%$$

Solution #2: The reasoning is as follows:

$$\mathbf{A}\cdot\dot{\mathbf{x}} = \mathbf{b} \;\Rightarrow\; \dot{\mathbf{x}} = \mathbf{A}^{-1}\cdot\mathbf{b}$$

$$(\mathbf{A}+\Delta\mathbf{A})\cdot(\dot{\mathbf{x}}+\Delta\mathbf{x}) = \mathbf{b} \;\Rightarrow\; \mathbf{A}\cdot\dot{\mathbf{x}}+\mathbf{A}\cdot\Delta\mathbf{x}+\Delta\mathbf{A}\cdot\dot{\mathbf{x}}+\Delta\mathbf{A}\cdot\Delta\mathbf{x} = \mathbf{b} \;\Rightarrow\; \mathbf{b}+\mathbf{A}\cdot\Delta\mathbf{x}+\Delta\mathbf{A}\cdot\dot{\mathbf{x}} \cong \mathbf{b}$$

$$\Rightarrow\; \mathbf{A}\cdot\Delta\mathbf{x} \cong -\Delta\mathbf{A}\cdot\dot{\mathbf{x}} \;\Rightarrow\; \Delta\mathbf{x} \cong -\mathbf{A}^{-1}\cdot\Delta\mathbf{A}\cdot\dot{\mathbf{x}}$$

Where:

$$\mathbf{A} = \begin{bmatrix} 100 & 101 \\ 102 & 103 \end{bmatrix} \Rightarrow \mathbf{A}^{-1} = \begin{bmatrix} -51.5 & 50.5 \\ 51.0 & -50.0 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \dot{\mathbf{x}} = \begin{bmatrix} -51.5 & 50.5 \\ 51.0 & -50.0 \end{bmatrix}\cdot\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\Delta\mathbf{A} = \begin{bmatrix} \Delta a_{11} & 0 \\ 0 & \Delta a_{22} \end{bmatrix} \Rightarrow \Delta\mathbf{x} = \begin{bmatrix} -51.5 & 50.5 \\ 51.0 & -50.0 \end{bmatrix}\cdot\begin{bmatrix} \Delta a_{11} & 0 \\ 0 & \Delta a_{22} \end{bmatrix}\cdot\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\Rightarrow \Delta\mathbf{x} = \begin{bmatrix} -51.5 & 50.5 \\ 51.0 & -50.0 \end{bmatrix}\cdot\begin{bmatrix} -\Delta a_{11} \\ \Delta a_{22} \end{bmatrix} \Rightarrow \Delta\mathbf{x} = \begin{bmatrix} 51.5\cdot\Delta a_{11} - 50.5\cdot\Delta a_{22} \\ -51.0\cdot\Delta a_{11} + 50.0\cdot\Delta a_{22} \end{bmatrix}$$

$$\Rightarrow \Delta\mathbf{x} = \begin{bmatrix} 51.5\cdot 100\cdot\alpha_{11} - 50.5\cdot 103\cdot\alpha_{22} \\ -51.0\cdot 100\cdot\alpha_{11} + 50.0\cdot 103\cdot\alpha_{22} \end{bmatrix}$$

Hence:

$$\left|\Delta x_1\right| \le 51.5\cdot 100\cdot 10^{-2} + 50.5\cdot 103\cdot 10^{-2} = 103.52$$

$$\left|\Delta x_2\right| \le 51.0\cdot 100\cdot 10^{-2} + 50.0\cdot 103\cdot 10^{-2} = 102.50$$

Since $\left|x_1\right| = \left|x_2\right| = 1$:

$$\left|\delta_1\right| \le 103.52 = 10352\% \cong 10^4\% \text{ and } \left|\delta_2\right| \le 102.50 = 10250\% \cong 10^4\%$$

Problem: Assess the errors of the solution of the system of equations:

$$\begin{bmatrix} \tilde{a}_{1,1} & 0 & 0 \\ \tilde{a}_{2,1} & \tilde{a}_{2,2} & 0 \\ 0 & \tilde{a}_{3,2} & \tilde{a}_{3,3} \end{bmatrix}\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \tilde{y}_3 \end{bmatrix} = \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \tilde{b}_3 \end{bmatrix}$$

caused by the floating-point representation of the data: $\tilde{a}_{1,1} \cong 1$, $\tilde{a}_{2,1} \cong 1$, $\tilde{a}_{2,2} \cong 2$, $\tilde{a}_{3,2} \cong 2$, $\tilde{a}_{3,3} \cong 3$, $\tilde{b}_1 \cong 1$, $\tilde{b}_2 \cong -1$ and $\tilde{b}_3 \cong 1$.

Solution: After substitution of $\tilde{a}_{1,1} = 1+\alpha_{1,1}$ and $\tilde{b}_1 = 1+\beta_1$, the first equation takes on the form:

$$(1+\alpha_{1,1})\,\tilde{y}_1 = 1+\beta_1$$

Its solution:

$$\tilde{y}_1 = \frac{1+\beta_1}{1+\alpha_{1,1}} = 1+\beta_1-\alpha_{1,1}$$

is subject to the error:

$$\delta[\tilde{y}_1] = \beta_1-\alpha_{1,1} \;\Rightarrow \left|\delta[\tilde{y}_1]\right| = \left|\beta_1-\alpha_{1,1}\right| \le (1+1)\,eps = 2eps$$

After substitution $\tilde{a}_{2,1} = 1+\alpha_{2,1}$, $\tilde{a}_{2,2} = 2(1+\alpha_{2,2})$ and $\tilde{b}_2 = -(1+\beta_2)$, the second equation takes on the form:

$$\left(1+\alpha_{2,1}\right)\tilde{y}_1+2\left(1+\alpha_{2,2}\right)\tilde{y}_2=-1-\beta_2$$

Its solution:

$$\tilde{y}_2=\frac{-1-\beta_2-\left(1+\alpha_{2,1}\right)\tilde{y}_1}{2\left(1+\alpha_{2,2}\right)}=\frac{-1-\beta_2-\left(1+\alpha_{2,1}\right)\left(1+\beta_1-\alpha_{1,1}\right)}{2\left(1+\alpha_{2,2}\right)}\cong$$

$$\cong\frac{-1-\beta_2-\left(1+\alpha_{2,1}+\beta_1-\alpha_{1,1}\right)}{2\left(1+\alpha_{2,2}\right)}\cong\frac{-2\left(1+\frac{1}{2}\beta_2+\frac{1}{2}\alpha_{2,1}+\frac{1}{2}\beta_1-\frac{1}{2}\alpha_{1,1}\right)}{2\left(1+\alpha_{2,2}\right)}\cong$$

$$\cong(-1)\left(1+\frac{1}{2}\beta_2+\frac{1}{2}\alpha_{2,1}+\frac{1}{2}\beta_1-\frac{1}{2}\alpha_{1,1}-\alpha_{2,2}\right)$$

is subject to the error:

$$\delta\left[\tilde{y}_2\right]\cong\frac{1}{2}\beta_2+\frac{1}{2}\alpha_{2,1}+\frac{1}{2}\beta_1-\frac{1}{2}\alpha_{1,1}-\alpha_{2,2}$$

$$\left|\delta\left[\tilde{y}_2\right]\right|\cong\left|\frac{1}{2}\beta_2+\frac{1}{2}\alpha_{2,1}+\frac{1}{2}\beta_1-\frac{1}{2}\alpha_{1,1}-\alpha_{2,2}\right|\le\left(\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+\frac{1}{2}+1\right)eps=3eps$$

After substitution $\tilde{a}_{3,2}=2\left(1+\alpha_{3,2}\right)$, $\tilde{a}_{3,3}=3\left(1+\alpha_{3,3}\right)$ and $\tilde{b}_3=1+\beta_3$, the third equation takes on the form:

$$2\left(1+\alpha_{3,2}\right)\tilde{y}_2+3\left(1+\alpha_{3,3}\right)\tilde{y}_3=1+\beta_3$$

Its solution:

$$\tilde{y}_3=\frac{1+\beta_3-2\left(1+\alpha_{3,2}\right)\tilde{y}_2}{3\left(1+\alpha_{3,3}\right)}=\frac{1+\beta_3-2\left(1+\alpha_{3,2}\right)(-1)\left(1+\delta\left[\tilde{y}_2\right]\right)}{3\left(1+\alpha_{3,3}\right)}$$

$$\cong\frac{1+\beta_3+2\left(1+\alpha_{3,2}+\delta\left[\tilde{y}_2\right]\right)}{3\left(1+\alpha_{3,3}\right)}\cong\frac{3\left(1+\frac{1}{3}\beta_3+\frac{2}{3}\alpha_{3,2}+\frac{2}{3}\delta\left[\tilde{y}_2\right]\right)}{3\left(1+\alpha_{3,3}\right)}$$

$$\cong1+\frac{1}{3}\beta_3+\frac{2}{3}\alpha_{3,2}+\frac{2}{3}\delta\left[\tilde{y}_2\right]-\alpha_{3,3}$$

is subject to the error:

$$\delta\left[\tilde{y}_3\right]\cong\frac{1}{3}\beta_3+\frac{2}{3}\alpha_{3,2}+\frac{1}{3}\beta_2+\frac{1}{3}\alpha_{2,1}+\frac{1}{3}\beta_1-\frac{1}{3}\alpha_{1,1}-\frac{2}{3}\alpha_{2,2}-\alpha_{3,3}$$

$$\left|\delta\left[\tilde{y}_3\right]\right|\cong\left|\frac{1}{3}\beta_3+\frac{2}{3}\alpha_{3,2}+\frac{1}{3}\beta_2+\frac{1}{3}\alpha_{2,1}+\frac{1}{3}\beta_1-\frac{1}{3}\alpha_{1,1}-\frac{2}{3}\alpha_{2,2}-\alpha_{3,3}\right|$$

$$\le\left(\frac{1}{3}+\frac{2}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{1}{3}+\frac{2}{3}+1\right)eps=4eps$$

Problem: Assess the aggregated relative error of the numerical solution of the following system of linear algebraic equations:

$$\begin{bmatrix}1 & 2\\ 3 & 5\end{bmatrix}\cdot\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}1\\ 2\end{bmatrix}$$

caused by one-percent error in the elements of the RHS vector.

Solution #1: The linearity of the system $\mathbf{A}\cdot\dot{\mathbf{x}}=\mathbf{b}$ implies: $\mathbf{A}\cdot\Delta\mathbf{x}=\Delta\mathbf{b}$. Thus, the exact solution is:

$$\dot{\mathbf{x}} = \mathbf{A}^{-1} \cdot \mathbf{b} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

and the absolute error of the solution:

$$\Delta\mathbf{x} = \mathbf{A}^{-1} \cdot \Delta\mathbf{b} = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 \cdot \beta_1 \\ 2 \cdot \beta_2 \end{bmatrix} = \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ 2\beta_2 \end{bmatrix}$$

Hence:

$$\Delta x_1 = -5\beta_1 + 4\beta_2 \text{ and } \Delta x_2 = 3\beta_1 - 2\beta_2$$

and

$$\|\Delta\mathbf{x}\|_\infty = \sup\left\{|-5\beta_1 + 4\beta_2|, |3\beta_1 - 2\beta_2|\right\} \le 0.09$$

Consequently:

$$\frac{\|\Delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le \frac{0.09}{\sup\left\{|-1|, |+1|\right\}} = 0.09 = 9\%$$

Solution #2: The worst-case assessment obtained for the errors in the RHS vector has the form:

$$\frac{\|\Delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le \|\mathbf{A}\|_\infty \cdot \|\mathbf{A}^{-1}\|_\infty \cdot \frac{\|\Delta\mathbf{b}\|_\infty}{\|\mathbf{b}\|_\infty}$$

The values of the norm in the considered case are:

$$\|\mathbf{A}\|_\infty = \sup\{1+2, 3+5\} = 8 \quad \text{and} \quad \|\mathbf{A}^{-1}\|_\infty = \sup\{5+2, 3+1\} = 7$$

$$\|\mathbf{b}\|_\infty = \sup\{1, 2\} = 2 \quad \text{and} \quad \|\Delta\mathbf{b}\|_\infty = \sup\{1 \cdot \beta_1, 2 \cdot \beta_2\} \le 0.02$$

Hence:

$$\frac{\|\Delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le 8 \cdot 7 \cdot \frac{0.02}{2} = 0.56 = 56\%$$

Problem: Assess the relative errors of the numerical solutions of the following systems of linear algebraic equations:

$$\begin{bmatrix} 10 & 101 \\ 102 & 10 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 111 \\ 112 \end{bmatrix}, \begin{bmatrix} 100 & 101 \\ 102 & 103 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 100 & 10 \\ 10 & 103 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 110 \\ 113 \end{bmatrix}$$

caused by one-percent error in the elements of the LHS matrices. Verify the results by numerical simulation of errors in MATLAB.

Problem: Assess the aggregated relative errors of the numerical solutions of the following systems of linear algebraic equations:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 15 \\ 25 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 8 \\ 3 & 8 & 14 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 25 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 8 \\ 3 & 8 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 11 \end{bmatrix}$$

caused by one-percent error in the elements of the LHS matrices. Verify the results by numerical simulation of errors in MATLAB. Repeat calculations for both norms: $\| \; \|_2$ and $\| \; \|_\infty$.

Problem: Assess the aggregated relative errors of the numerical solutions of the following systems of linear algebraic equations:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 15 \\ 25 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 8 \\ 3 & 8 & 14 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 25 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 8 \\ 3 & 8 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 11 \end{bmatrix}$$

caused by one-percent error in the elements of the LHS matrices and RHS vectors. Verify the results by numerical simulation of errors in MATLAB. Repeat calculations for both norms: $\| \; \|_2$ and $\| \; \|_\infty$.

Problem: Assess the relative error of the determinant of the matrix:

$$\tilde{\mathbf{A}} = \begin{bmatrix} \tilde{a}_{1,1} & 2 & 3 & 4 & 5 \\ 1 & 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

caused by the one-percent uncertainty of the element $\tilde{a}_{1,1} \cong 1$.

<u>Solution:</u> The following formula may be useful in this case for computing $\det(\tilde{\mathbf{A}})$:

$$\det(\tilde{\mathbf{A}}) = \det(\tilde{\mathbf{L}}) \cdot \det(\tilde{\mathbf{U}}) = \tilde{u}_{1,1} \cdot \ldots \cdot \tilde{u}_{5,5}$$

where the matrices $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{U}}$ result from the LU decomposition of the matrix $\tilde{\mathbf{A}}$. The matrix $\tilde{\mathbf{U}}$ may be determined via one-step elimination of the element $a_{2,1} = 1$ by means of the first row of $\tilde{\mathbf{A}}$. After this operation the second row takes on the form:

$$\begin{bmatrix} 0 & 1 - \dfrac{2}{\tilde{a}_{1,1}} & 2 - \dfrac{3}{\tilde{a}_{1,1}} & 3 - \dfrac{4}{\tilde{a}_{1,1}} & 4 - \dfrac{5}{\tilde{a}_{1,1}} \end{bmatrix} \text{ where } \tilde{a}_{1,1} = 1 + \alpha \text{ and } |\alpha| \le 1\%$$

Thus:

$$\det(\tilde{\mathbf{A}}) = \tilde{a}_{1,1} \cdot \left(1 - \frac{2}{\tilde{a}_{1,1}}\right) \cdot 1 \cdot 1 \cdot 1 = (1+\alpha) \cdot \left(1 - \frac{2}{1+\alpha}\right) = 1 + \alpha - 2 = -1 + \alpha = -1 \cdot (1 - \alpha)$$

$$\left| \delta\left[\det(\tilde{\mathbf{A}})\right] \right| = |-\alpha| \le 1\%$$

Alternatively, $\det(\tilde{\mathbf{A}})$ may be calculated using the formula:

$$\det(\tilde{\mathbf{A}}) = \tilde{a}_{1,1} \cdot \det\left(\begin{bmatrix} & & & \\ 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right) - 1 \cdot \det\left(\begin{bmatrix} 2 & 3 & 4 & 5 \\ & & & \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}\right) = \tilde{a}_{1,1} \cdot 1 - 1 \cdot 2$$

Problem: Given the matrix:

$$\tilde{\mathbf{Y}} \equiv \tilde{\mathbf{X}}^T \cdot \tilde{\mathbf{X}} \text{ with } \tilde{\mathbf{X}} \equiv \begin{bmatrix} 1+\varepsilon & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } |\varepsilon| \le 1\%$$

assess the relative error of the element $\tilde{l}_{2,2}$ of the lower triangular matrix $\tilde{\mathbf{L}}$ resulting from the Cholesky decomposition of the matrix $\tilde{\mathbf{Y}}$ – the error caused by the one-percent uncertainty of the element $\tilde{x}_{1,1}$ of the matrix $\tilde{\mathbf{X}}$.

Solution: It follows from the definition of the matrix $\tilde{\mathbf{Y}}$ that:

$$\tilde{\mathbf{Y}} = \begin{bmatrix} 1+\varepsilon & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1+\varepsilon & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4+2\varepsilon & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Thus, the elements of the matrix $\tilde{\mathbf{L}}$ have to satisfy the equation:

$$\begin{bmatrix} \tilde{l}_{1,1} & 0 & 0 & 0 \\ \tilde{l}_{2,1} & \tilde{l}_{2,2} & 0 & 0 \\ \tilde{l}_{3,1} & \tilde{l}_{3,2} & \tilde{l}_{3,3} & 0 \\ \tilde{l}_{4,1} & \tilde{l}_{4,2} & \tilde{l}_{4,3} & \tilde{l}_{4,4} \end{bmatrix} \cdot \begin{bmatrix} \tilde{l}_{1,1} & \tilde{l}_{2,1} & \tilde{l}_{3,1} & \tilde{l}_{4,1} \\ 0 & \tilde{l}_{2,2} & \tilde{l}_{3,2} & \tilde{l}_{4,2} \\ 0 & 0 & \tilde{l}_{3,3} & \tilde{l}_{4,3} \\ 0 & 0 & 0 & \tilde{l}_{4,4} \end{bmatrix} = \begin{bmatrix} 4+2\varepsilon & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Hence:

$$\left(\tilde{l}_{1,1}\right)^2 = 4+2\varepsilon = 4\left(1+\frac{1}{2}\varepsilon\right) \Rightarrow \tilde{l}_{1,1} \cong 2\left(1+\frac{1}{4}\varepsilon\right)$$

$$\tilde{l}_{1,1}\tilde{l}_{2,1} = 3 \Rightarrow \tilde{l}_{2,1} = \frac{3}{\tilde{l}_{1,1}} \cong \frac{3}{2}\left(1-\frac{1}{4}\varepsilon\right)$$

$$\left(\tilde{l}_{2,1}\right)^2 + \left(\tilde{l}_{2,2}\right)^2 = 3 \Rightarrow \tilde{l}_{2,2} = \sqrt{3-\left(\tilde{l}_{2,1}\right)^2} = \sqrt{3-\frac{9}{4}\left(1-\frac{1}{2}\varepsilon\right)} \cong \sqrt{\frac{3}{4}}\left(1+\frac{3}{4}\varepsilon\right)$$

Thus:

$$\left|\delta\left[\tilde{l}_{2,2}\right]\right| = \left|\frac{3}{4}\varepsilon\right| \le 0.75\%$$

Problem: Check the convergence of the iterative algorithm (IA) defined by the formula:

$$\begin{bmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{3} & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} + \begin{bmatrix} -\frac{1}{2} \\ \frac{2}{3} \end{bmatrix}$$

to the solution of the following system of linear algebtraic equations:

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}.$$

Solution: For the vector $\dot{\mathbf{x}} = \begin{bmatrix} -1 & 1 \end{bmatrix}^T$ being the exact solution of the system of equations, the IA satisfies the coherence condition, *i.e.*:

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{3} & 0 \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \begin{bmatrix} -\frac{1}{2} \\ \frac{2}{3} \end{bmatrix}$$

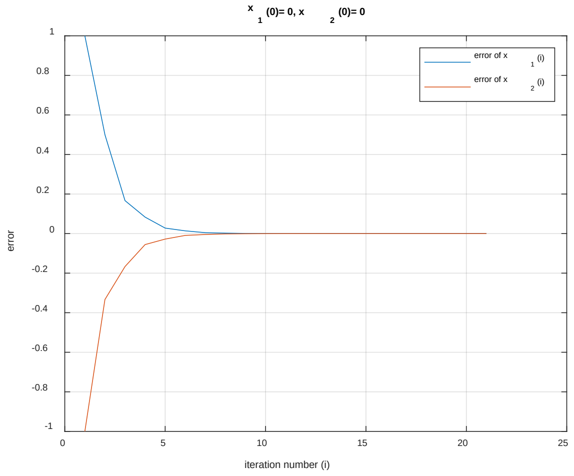
The IA will be convergent to $\dot{\mathbf{x}}$ if:

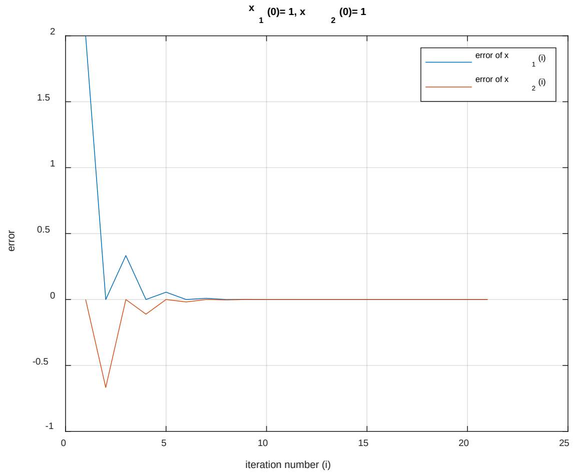
$$\mathrm{sr}\left(\begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{3} & 0 \end{bmatrix}\right) < 1$$

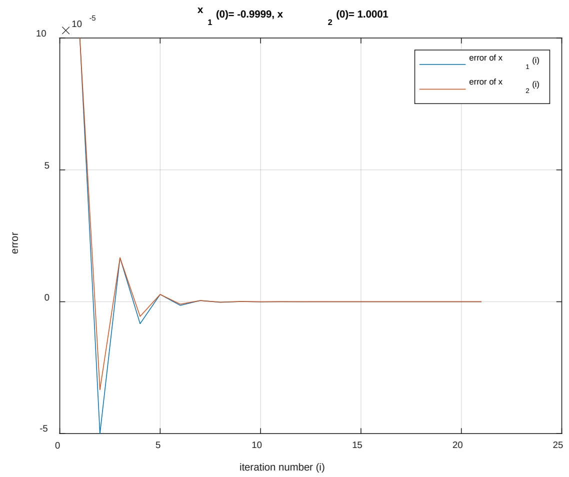
The eigenvalues of the matrix satisfy the inequalities $|\lambda_1| < 1$ and $|\lambda_2| < 1$ since:

$$\det\left(\begin{bmatrix} -\lambda & -\frac{1}{2} \\ -\frac{1}{3} & -\lambda \end{bmatrix}\right) = 0 \Rightarrow \lambda^2 - \frac{1}{6} = 0$$

This means that the IA is convergent as shown in the following figures:








```
clear all
M=[0 -0.5;-1/3 0];w=[-0.5 2/3]';
x0=[-0.9999 1.0001]';x=x0;errX=x-[-1 1]';
for i=1:20
x=M*x+w;
errX=[errX x-[-1 1]'];
end
plot(errX');grid;
xlabel('iteration number (i)');
ylabel('error');
legend('error of x_1(i)',' error of x_2(i)')
title(['x_1(0)= ',num2str(x0(1)),', x_2(0)=
',num2str(x0(2))])
```

ENUME: SOLVED PROBLEMS

3. SOLVING NONLINEAR ALGEBRAIC EQUATIONS

3.1. Preliminary exercises

Problem: Compute the derivative of the vector function:

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2^2 + x_3^2 \\ x_1 x_2^2 \\ \dfrac{x_2^2}{x_1 + x_2 + x_3} \end{bmatrix}$$

with respect to the vector $\mathbf{x}$.

<u>Solution:</u>

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^T} = \begin{bmatrix} 0 & 2x_2 & 2x_3 \\ x_2^2 & 2x_1 x_2 & 0 \\ \dfrac{-x_2^2}{(x_1 + x_2 + x_3)^2} & \dfrac{2x_1 x_2 + x_2^2 + 2x_2 x_3}{(x_1 + x_2 + x_3)^2} & \dfrac{-x_2^2}{(x_1 + x_2 + x_3)^2} \end{bmatrix}$$

Problem: Compute the following ratios of polynomials:

$\dfrac{x^2 - 1}{x + 1}$ (<u>Solution:</u> $x - 1$)

$\dfrac{x^4 - 2x^3 + x^2 - 1}{-x^2 + x + 1}$ (<u>Solution:</u> $-x^2 + x - 1$)

$\dfrac{x^5 + x^4 - 4x^3 - 1}{-x^2 + x + 1}$ (<u>Solution:</u> $-x^3 - 2x^2 + x - 1$)

$\dfrac{x^7 - x^6 - 2x^5 + x^4 + 2x^3 - x^2 - x}{-x^2 + x + 1}$ (<u>Solution:</u> $-x^5 + x^3 - x$)

$\dfrac{x^7 - 12x^6 + 20x^5 + 2x^4 - 20x^3 + 8x^2 + x}{-x^2 + 2x - 1}$ (<u>Solution:</u> $-x^5 + 10x^4 + x^3 - 10x^2 - x$)

3.2. Analysis of one-point iterative algorithms

Problem: Determine the parameters of local convergence , C and ρ, of the following iterative algorithm:

$$y_{i+1} = y_i - \frac{2}{15}\left(y_i^3 - x\right) \qquad \text{for } x \in [1, 8]$$

in the vicinity of the point $\dot{y} = \sqrt[3]{x}$.

<u>Solution:</u> The only stationary point is $\dot{y} = \sqrt[3]{x}$. The RHS of the algorithm:

$$\phi(y_i) = y_i - \frac{2}{15}\left(y_i^3 - x\right)$$

may be developed at this point in the following Taylor series:

$$\phi(y_i) = \phi(\dot{y}) + \phi'(\dot{y})\Delta_i + \frac{1}{2}\phi''(\dot{y})\Delta_i^2 + \ldots$$

where:

$$\phi(\dot{y}) = \dot{y} \text{ and } \phi'(\dot{y}) = 1 - \frac{2}{15}\left(3\dot{y}^2 - 0\right) = 1 - \frac{2}{5}\dot{y}^2 \neq 0$$

Thus:

$$\dot{y} + \Delta_{i+1} \cong \dot{y} + \phi'(\dot{y})\Delta_i \text{ and } \Delta_{i+1} \cong \phi'(\dot{y})\Delta_i \equiv \left(1 - \frac{2}{5}\dot{y}^2\right)\Delta_i$$

Hence:

$$\rho = 1 \text{ and } C = 1 - \frac{2}{5}\dot{y}^2 = 1 - \frac{2}{5}x^{\frac{2}{3}}$$

The convergence is guaranteed because $x \in [1, 8]$ implies $C \in \left[-\frac{3}{5}, \frac{3}{5}\right]$, *i.e.* $|C| < 1$.

Problem 3: Assess the attainable accuracy of the following iterative algorithm (AI):

$$y_{i+1} = y_i - \tfrac{1}{18}\left(y_i^4 - x\right) \text{ for } x \in [1, 16]$$

<u>Solution:</u> The value $\dot{y} = \sqrt[4]{x}$ is the only stationary point of AI. The function $\phi(y)$ defining AI has the form:

$$\phi(y) \equiv y - \tfrac{1}{18}\left(y^4 - x\right)$$

Since:

$$\phi(\dot{y}) = \dot{y} \text{ and } \phi'(y) = 1 - \tfrac{1}{18}\left(4y^3 - 0\right) = 1 - \tfrac{4}{18}y^3 \xrightarrow[y \to \dot{y}]{} 1 - \tfrac{2}{9}x^{\frac{3}{4}} \neq 0$$

one can gather that $\rho = 1$ and $C(x) = 1 - \frac{2}{9}x^{\frac{3}{4}}$. The function $C(x)$ is decreasing from $\frac{7}{9}$ to $-\frac{7}{9}$ in the interval when x is growing from 1 to 16. Consequently:

$$\sup\left\{\left|C(x)\right| \,\middle|\, x \in [1, 16]\right\} = \tfrac{7}{9}$$

which means that AI is convergent because $|C(x)| < 1$ in the whole interval $[1, 16]$. During a single iteration, when $y_i \to \dot{y}$, the rounding errors generate the error component $\Delta\theta$ which may be assessed in the following way:

$$\tilde{\phi}(\dot{y}) = \left\{\dot{y} - \tfrac{1}{18}\left[\dot{y}^4\left(1 + \eta_p\right) - x\right]\left(1 + \eta_m\right)\left(1 + \eta_o'\right)\right\}\left(1 + \eta_o''\right)$$

$$\tilde{\phi}(\dot{y}) = \left\{\dot{y} - \tfrac{1}{18}\left[x\left(1 + \eta_p\right) - x\right]\left(1 + \eta_m + \eta_o'\right)\right\}\left(1 + \eta_o''\right)$$

$$\tilde{\phi}(\dot{y}) = \left\{\dot{y} - \tfrac{1}{18}\left[x\eta_p\right]\left(1 + \eta_m + \eta_o'\right)\right\}\left(1 + \eta_o''\right) = \left\{\dot{y} - \tfrac{1}{18}x\eta_p\right\}\left(1 + \eta_o''\right) = \dot{y}\left\{1 - \tfrac{1}{18}x^{\frac{3}{4}}\eta_p\right\}\left(1 + \eta_o''\right)$$

$$\Delta\theta = -\tfrac{1}{18}x^{\frac{3}{4}}\eta_p + \eta_o'' \Rightarrow \left|\Delta\theta\right| \leq \left(\tfrac{1}{18}x^{\frac{3}{4}} + 1\right)eps \leq \left(\tfrac{8}{18} + 1\right)eps = \tfrac{13}{9}eps$$

The latter result may be used for computing the parameter characterising the attainable accuracy of AI:

$$K = \frac{\frac{\Delta\theta}{eps}}{1-\sup\left|C(x)\right|} = \frac{\frac{13}{9}}{1-\frac{7}{9}} = \frac{13}{2} = 6.5$$

Problem: Determine the parameters of local convergence , C and ρ, of the following iterative algorithm:

$$y_{i+1} = y_i - \frac{1}{6}y_i^4 + \frac{1}{6}y_i^{-2}$$

Solution: There are two real-valued solutions of the equation:

$$\dot{y} = \dot{y} - \frac{1}{6}\dot{y}^4 + \frac{1}{6}\dot{y}^{-2}$$

viz.: $\dot{y} = 1$ and $\dot{y} = -1$ (the stationary points of the iterative algorithm under study).

The function defining the algorithm and its derivatives have the form:

$$\phi(y) = y - \frac{1}{6}y^4 + \frac{1}{6}y^{-2}$$

$$\phi'(y) = 1 - \frac{2}{3}y^3 - \frac{1}{3}y^{-3} \begin{cases} \xrightarrow[y\to 1]{} 1 - \frac{2}{3} - \frac{1}{3} = 0 \\ \xrightarrow[y\to -1]{} 1 + \frac{2}{3} + \frac{1}{3} = 2 \end{cases}$$

$$\phi''(y) = -2y^2 + y^{-4} \xrightarrow[y\to 1]{} -2 + 1 = -1$$

Thus, the algorithm is converging only to the point $\dot{y} = 1$ with $C = -0.5$ and $\rho = 2$.

Problem: Determine the parameters of local convergence, C and ρ, of the following iterative algorithm:

$$y_{i+1} = y_i - \frac{1}{2}\left[\exp(y_i) - x\right] \quad \text{for } x \in (0, 4)$$

in the vicinity of the point $\dot{y} = \ln(x)$.

Solution: The only stationary point is $\dot{y} = \ln(x)$. The RHS of the algorithm:

$$\phi(y_i) = y_i - \frac{1}{2}\left[\exp(y_i) - x\right]$$

may be developed at this point in the following Taylor series:

$$\phi(y_i) = \phi(\dot{y}) + \phi'(\dot{y})\Delta_i + \frac{1}{2}\phi''(\dot{y})\Delta_i^2 + \ldots$$

where:

$$\phi(\dot{y}) = \dot{y} \text{ and } \phi'(\dot{y}) = 1 - \frac{1}{2}\left[\exp(\dot{y}) - 0\right] = 1 - \frac{1}{2}\exp(\dot{y}) \neq 0$$

Thus:

$$\dot{y} + \Delta_{i+1} \cong \dot{y} + \phi'(\dot{y})\Delta_i$$

$$\Delta_{i+1} \cong \phi'(\dot{y})\Delta_i \equiv \left(1 - \frac{1}{2}\exp(\dot{y})\right)\Delta_i$$

Hence:

$$\rho = 1 \text{ and } C = 1 - \frac{1}{2}\exp(\dot{y}) = 1 - \frac{1}{2}x$$

The convergence is guaranteed because $x \in (0, 4)$ implies $C \in (-1, 1)$, *i.e.* $|C| < 1$.

Problem: Determine the parameters of local convergence, ρ and C, of the following iterative algorithm (IA):

$$y_{i+1} = y_i + \alpha \frac{\sin(y_i) - \frac{1}{2}}{\cos(y_i)}$$

in the vicinity of its stationary point $\dot{y}$ belonging to the interval $[0, 1]$. Indicate the value of the parameter α guaranteeing the quickest convergence of that IA.

<u>Solution:</u> The only stationary point of IA, belonging to the interval $[0, 1]$, is $\dot{y} = \frac{\pi}{6}$. The convergence parameters, corresponding to this point, may be determined in the standard way:

$$\varphi(y) = y + \alpha \frac{\sin(y) - \frac{1}{2}}{\cos(y)},$$

$$\varphi'(y) = 1 + \alpha \frac{1 - \frac{1}{2}\sin(y)}{\cos^2(y)} \xrightarrow[y \to \frac{\pi}{6}]{} 1 + \alpha$$

Hence:

$$C(\alpha) = 1 + \alpha \text{ and } \rho(\alpha) = 1$$

The convergence is guaranteed if $|C(\alpha)| < 1$, *i.e.*:

$$|1 + \alpha| < 1 \Rightarrow -1 < 1 + \alpha < 1 \Rightarrow -2 < \alpha < 0$$

The quickest convergence of IA is guaranteed for $\alpha = -1$ since $C(-1) = 0$.

Problem: Determine the parameters of local convergence, C and ρ, of the following iterative algorithm (IA):

$$x_{i+1} = x_i \frac{1 - \ln(x_i)}{1 + x_i}$$

designed for solvig the equation: $\ln(x) + x = 0$.

<u>Solution #1:</u> The only stationary point of IA, by definition, satisfies the equality $\ln(\dot{x}) = -\dot{x}$. The function defining IA has the form:

$$\phi(x) = x \frac{1 - \ln(x)}{1 + x}$$

Its first derivative is:

$$\phi'(x) = \frac{-x - \ln(x)}{(1 + x)^2} \xrightarrow[x \to \dot{x}]{} 0$$

Its second derivative is:

$$\phi''(x) = -\frac{1}{x(1 + x)} + 2\frac{x + \ln(x)}{(1 + x)^3} \xrightarrow[x \to \dot{x}]{} -\frac{1}{\dot{x}(1 + \dot{x})} \neq 0$$

Thus: $\rho = 2$ and $C = -\dfrac{1}{2\dot{x}(1+\dot{x})}$.

Solution #2: The only stationary point of IA, by definition, satisfies the equality $\ln(\dot{x}) = -\dot{x}$. After substitution $x_i = \dot{x} + \Delta_i$ and $x_{i+1} = \dot{x} + \Delta_{i+1}$, IA takes on the form:

$$\dot{x} + \Delta_{i+1} = (\dot{x}+\Delta_i)\frac{1-\ln(\dot{x}+\Delta_i)}{1+\dot{x}+\Delta_i} = (\dot{x}+\Delta_i)\frac{1-\left[\ln(\dot{x}) + \frac{1}{\dot{x}}\Delta_i - \frac{1}{2\dot{x}^2}\Delta_i^2 + \ldots\right]}{1+\dot{x}+\Delta_i}$$

Hence:

$$\Delta_{i+1} = (\dot{x}+\Delta_i)\frac{1-\left[-\dot{x} + \frac{1}{\dot{x}}\Delta_i - \frac{1}{2\dot{x}^2}\Delta_i^2 + \ldots\right]}{1+\dot{x}+\Delta_i} - \dot{x} = \ldots = -\frac{1}{2\dot{x}(1+\dot{x})}\Delta_i^2$$

Thus: $\rho = 2$ and $C = -\dfrac{1}{2\dot{x}(1+\dot{x})} \cong -0.5626$ $(\dot{x} \cong 0.5671)$.

Problem: Determine the parameters of local convergence, C and ρ, of the following iterative algorithm (IA):

$$x_{i+1} = \frac{\ln(x_i)}{\ln(x_i) + x_i - 1}$$

designed for solvig the equation: $\ln(x) + x = 0$.

Solution #1: The only stationary point of IA, by definition, satisfies the equality $\ln(\dot{x}) = -\dot{x}$. The function defining IA has the form:

$$\phi(x) = \frac{\ln(x)}{\ln(x) + x - 1}$$

Its first derivative is:

$$\phi'(x) = \frac{\frac{1}{x}\left[\ln(x) + x - 1\right] - \ln(x)\left(\frac{1}{x}+1\right)}{\left[\ln(x)+x-1\right]^2} \xrightarrow[x\to\dot{x}]{} -\frac{1}{\dot{x}} + 1 + \dot{x} \neq 0$$

Thus: $\rho = 1$ and $C = -\dfrac{1}{\dot{x}} + 1 + \dot{x}$.

Solution #2: The only stationary point of IA, by definition, satisfies the equality $\ln(\dot{x}) = -\dot{x}$. After substitution $x_i = \dot{x} + \Delta_i$ and $x_{i+1} = \dot{x} + \Delta_{i+1}$, IA takes on the form:

$$\dot{x} + \Delta_{i+1} = \frac{\ln(\dot{x}+\Delta_i)}{\ln(\dot{x}+\Delta_i) + \dot{x} + \Delta_i - 1} = \frac{\ln(\dot{x}) + \frac{1}{\dot{x}}\Delta_i - \frac{1}{2\dot{x}^2}\Delta_i^2 + \ldots}{\ln(\dot{x}) + \frac{1}{\dot{x}}\Delta_i - \frac{1}{2\dot{x}^2}\Delta_i^2 + \ldots + \dot{x} + \Delta_i - 1}$$

Hence:

$$\Delta_{i+1} = \frac{-\dot{x} + \frac{1}{\dot{x}}\Delta_i - \frac{1}{2\dot{x}^2}\Delta_i^2 + \ldots}{-\dot{x} + \frac{1}{\dot{x}}\Delta_i - \frac{1}{2\dot{x}^2}\Delta_i^2 + \ldots + \dot{x} + \Delta_i - 1} - \dot{x} = \ldots = \left(-\frac{1}{\dot{x}} + 1 + \dot{x}\right)\Delta_i$$

Thus: $\rho=1$ and $C=-\frac{1}{\dot{x}}+1+\dot{x}$.

Problem: Determine the parameter of attainable accuracy, K, of the following iterative algorithm (IA):

$$x_{i+1}=x_i\frac{1-\ln\left(x_i\right)}{1+x_i}$$

being quadratically convergent to the solution of the equation: $\ln\left(x\right)+x=0$.

Solution: The quadratic convrgence of AI enables us to ignore the propagation of errors from one iteration to another. Thus:

$$\tilde{x}_{i+1}=\dot{x}\frac{\left[1-\ln\left(\dot{x}\right)\left(1+\eta_l\right)\right]\left(1+\eta_o\right)}{\left(1+\dot{x}\right)\left(1+\eta_s\right)}\left(1+\eta_m\right)\left(1+\eta_d\right)$$

$$=\dot{x}\frac{\left[1-\ln\left(\dot{x}\right)\left(1+\eta_l\right)\right]}{1+\dot{x}}\left(1+\eta_o+\eta_m+\eta_d-\eta_s\right)$$

$$\tilde{x}_{i+1}=\dot{x}\frac{1-\ln\left(\dot{x}\right)}{1+\dot{x}}\left[1-\frac{\ln\left(\dot{x}\right)}{1-\ln\left(\dot{x}\right)}\eta_l\right]\left(1+\eta_o+\eta_m+\eta_d-\eta_s\right)$$

$$=\dot{x}\left[1-\frac{\ln\left(\dot{x}\right)}{1-\ln\left(\dot{x}\right)}\eta_l+\eta_o+\eta_m+\eta_d-\eta_s\right]$$

$$\left|\vartheta_{i+1}\right|\le\left|\frac{\ln\left(\dot{x}\right)}{1-\ln\left(\dot{x}\right)}\right|\left|\eta_l\right|+\left|\eta_o\right|+\left|\eta_m\right|+\left|\eta_d\right|+\left|\eta_s\right|\le\left(\left|\frac{\ln\left(\dot{x}\right)}{1-\ln\left(\dot{x}\right)}\right|+4\right)eps$$

$$K=\left|\frac{\ln\left(\dot{x}\right)}{1-\ln\left(\dot{x}\right)}\right|+4\cong 4.36\quad\left(\dot{x}\cong 0.5671\right)$$

Problem: Demonstrate that the iterative algorithm (IA):

$$y_{i+1}=\frac{1+y_i}{1+\exp(y_i)}\quad\text{for } i=0,1,\ldots$$

may converge to the solution of the equation $y\cdot\exp(y)-1=0$. Determine the parameters of local convergence (C, ρ) and attainable accuracy (K).

Solution: The solution $\dot{y}$ of the equation $y\cdot\exp(y)-1=0$ is a stationary point of IA because:

$$\exp(\dot{y})=\frac{1}{\dot{y}}$$

and consequently:

$$RHS=\frac{1+\dot{y}}{1+\exp(\dot{y})}=\frac{1+\dot{y}}{1+\frac{1}{\dot{y}}}=\dot{y}$$

The parameters of local convergence may be determined using the derivatives of the function defining IA:

$$\phi\left(y_i\right)\equiv\frac{1+y_i}{1+\exp(y_i)}$$

for $y_i \xrightarrow[i\to\infty]{} \dot{y}$. The first derivative is:

$$\phi'(y_i) = \frac{1 - y_i \cdot \exp(y_i)}{\left[1+\exp(y_i)\right]^2} = \xrightarrow[y_i \to \dot{y}]{} 0$$

The second derivative is:

$$\phi''(y_i) = \frac{-\left[\exp(y_i) + y_i \cdot \exp(y_i)\right]\cdot\left[1+\exp(y_i)\right]^2 - \left[1 + y_i \cdot \exp(y_i)\right]\cdot 2 \cdot \left[1+\exp(y_i)\right]\cdot \exp(y_i)}{\left[1+\exp(y_i)\right]^4}$$

$$\xrightarrow[y_i \to \dot{y}]{} \frac{-\left[1+\exp(\dot{y})\right]\cdot\left[1+\exp(\dot{y})\right]^2 - 0}{\left[1+\exp(\dot{y})\right]^4} = -\frac{\dot{y}}{\dot{y}+1}$$

Thus, the parameters of local convergence are:

$$\rho = 2 \text{ and } C = -\frac{\dot{y}}{2\cdot(\dot{y}+1)} \cong -0.18 \text{ because } \dot{y} \cong 0.57$$

Since $\rho = 2$, the effect of error propagation may be neglected, and the parameter of attainable accuracy may be determined on the basis of the relative error committed during a single iteration only:

$$\tilde{y}_{i+1} = \frac{(1+\dot{y})\cdot(1+\eta_a')}{\left[1+\exp(\dot{y})\cdot(1+\eta_e)\right]\cdot(1+\eta_a'')}\cdot(1+\eta_d)$$

$$= \frac{(1+\dot{y})\cdot(1+\eta_a' + \eta_d - \eta_a'')}{\left[1+\exp(\dot{y})\right]\cdot\left[1+\dfrac{\exp(\dot{y})}{1+\exp(\dot{y})}\cdot\eta_e\right]} = \dot{y}\cdot\left(1+\eta_a'+\eta_d-\eta_a''-\frac{\exp(\dot{y})}{1+\exp(\dot{y})}\cdot\eta_e\right)$$

Hence:

$$\Delta\vartheta_i = \eta_a' + \eta_d - \eta_a'' - \frac{\exp(\dot{y})}{1+\exp(\dot{y})}\cdot\eta_e$$

$$\left|\Delta\vartheta_i\right| \le \left|\eta'\right|_a + \left|\eta_d\right| + \left|\eta_a''\right| + \frac{\exp(\dot{y})}{1+\exp(\dot{y})}\cdot\left|\eta_e\right| \le eps + eps + eps + \frac{1}{1+\dot{y}}\cdot eps$$

Thus:

$$K = 3 + \frac{1}{1+\dot{y}} = 3.64$$

Problem: Taking into account that the convergence exponent ρ of the Newton algorithm is 2, assess the attainable accuracy of this algorithm when applied to the equation:

$$\sin(y) = \cos(y) \quad \text{dla} \quad y \in [0, \pi/2]$$

<u>Solution:</u> Since $\rho = 2$, one may neglect the transmission of errors from one iteration to the next, and to limit the assessment to the computing errors that appear during one iteration:

$$f(y) = \sin(y) - \cos(y)$$

$$f(y) = 0 \;\Rightarrow\; y_\infty = \pi/4 \;\Rightarrow\; \sin(y_\infty) = \cos(y_\infty) = 1\big/\sqrt{2}$$

$$f(y) = \sin(y) - \cos(y) \;\Rightarrow\; f'(y) = \sin(y) + \cos(y)$$

$$y_{i+1} = y_i - \frac{\sin(y_i) - \cos(y_i)}{\sin(y_i) + \cos(y_i)} \text{ for } i = 0, 1, \ldots$$

$$\mathrm{fl}\left(y_{i+1}\right)=\left\{\frac{\pi}{4}-\frac{\left[\frac{1}{\sqrt{2}}\left(1+\eta_{\sin}\right)-\frac{1}{\sqrt{2}}\left(1+\eta_{\cos}\right)\right]\left(1+\eta_o\right)}{\left[\frac{1}{\sqrt{2}}\left(1+\eta_{\sin}\right)+\frac{1}{\sqrt{2}}\left(1+\eta_{\cos}\right)\right]\left(1+\eta_s\right)}\left(1+\eta_d\right)\right\}\left(1+\eta_{oo}\right)\text{ for } i\to\infty$$

$$\mathrm{fl}\left(y_{i+1}\right)=\left\{\frac{\pi}{4}-\frac{\left[\eta_{\sin}-\eta_{\cos}\right]\left(1+\eta_o+\eta_d-\eta_s\right)}{\left[2+\eta_{\sin}+\eta_{\cos}\right]}\right\}\left(1+\eta_{oo}\right)\text{ for } i\to\infty$$

$$\mathrm{fl}\left(y_{i+1}\right)=\left\{\frac{\pi}{4}-\frac{\left[\eta_{\sin}-\eta_{\cos}\right]}{2}\right\}\left(1+\eta_{oo}\right)\text{ for } i\to\infty$$

$$\mathrm{fl}\left(y_{i+1}\right)=\frac{\pi}{4}\left\{1-\frac{4}{\pi}\frac{\left[\eta_{\sin}-\eta_{\cos}\right]}{2}\right\}\left(1+\eta_{oo}\right)\text{ for } i\to\infty$$

$$\mathrm{fl}\left(y_{i+1}\right)=\frac{\pi}{4}\left\{1-\frac{2}{\pi}\left(\eta_{\sin}-\eta_{\cos}\right)+\eta_{oo}\right\}\text{ for } i\to\infty$$

$$\Delta\upsilon_i=-\frac{2}{\pi}\left(\eta_{\sin}-\eta_{\cos}\right)+\eta_{oo}\text{ for } i\to\infty$$

$$\left|\Delta\upsilon_i\right|\le\left(\frac{4}{\pi}+1\right)eps\text{ for } i\to\infty$$

Problem: The following iterative algorithm:

$$x_1=0;\ \ x_{i+1}=\frac{\frac{5}{4}\pi\left[\sin\left(x_i\right)+\cos\left(x_i\right)\right]+\sqrt{2}x_i}{\sin\left(x_i\right)+\cos\left(x_i\right)+\sqrt{2}}\text{ for } i=1,2,\ldots$$

may be used for solving the equation:

$$\sin\left(x\right)+\cos\left(x\right)=0\text{ for } x\in\left(0,\tfrac{5}{4}\pi\right)$$

Determine the parameters of its local convergence (ρ and C).

<u>Solution:</u> The only solution of the equation $\sin\left(x\right)+\cos\left(x\right)=0$ in the interval $\left(0,\tfrac{5}{4}\pi\right)$ is $\dot{x}=\tfrac{3}{4}\pi$. The function defining the algorithm has the form:

$$\phi\left(x\right)=\frac{N\left(x\right)}{D\left(x\right)}$$

where:

$$N\left(x\right)=\tfrac{5}{4}\pi\left[\sin\left(x\right)+\cos\left(x\right)\right]+\sqrt{2}x\text{ and } D\left(x\right)=\sin\left(x\right)+\cos\left(x\right)+\sqrt{2}$$

Thus, its derivative may be calculated according to the scheme:

$$\phi'\left(\dot{x}\right)=\frac{N'\left(\dot{x}\right)D\left(\dot{x}\right)-N\left(\dot{x}\right)D'\left(\dot{x}\right)}{D^2\left(\dot{x}\right)}$$

where:

$$N\left(\dot{x}\right)=\tfrac{5}{4}\pi\left[\sin\left(\dot{x}\right)+\cos\left(\dot{x}\right)\right]+\sqrt{2}\dot{x}=\sqrt{2}\dot{x}=\sqrt{2}\tfrac{3}{4}\pi$$

$$N'\left(\dot{x}\right)=\tfrac{5}{4}\pi\left[\cos\left(\dot{x}\right)-\sin\left(\dot{x}\right)\right]+\sqrt{2}=\sqrt{2}\left(1-\tfrac{5}{4}\pi\right)$$

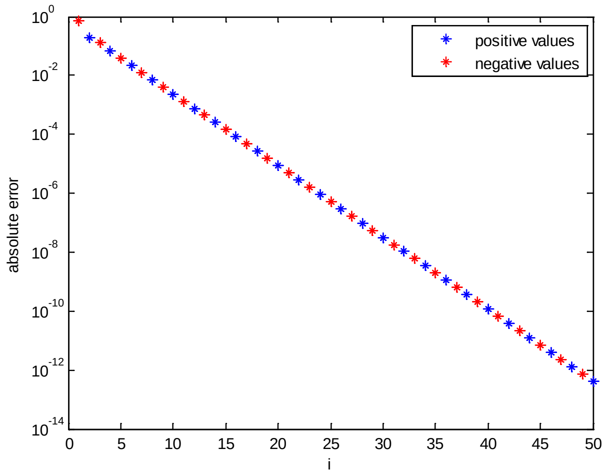
$$D\left(\dot{x}\right)=\sin\left(\dot{x}\right)+\cos\left(\dot{x}\right)+\sqrt{2}=\sqrt{2}$$

$$D'\left(\dot{x}\right)=\cos\left(\dot{x}\right)-\sin\left(\dot{x}\right)=-\sqrt{2}$$

After substitution:

$$\phi'\left(\dot{x}\right)=1-\tfrac{\pi}{2}$$

Thus: $\rho = 1$ and $C = 1 - \frac{\pi}{2}$.




```
clear all
x0=5*pi/4;y0=sin(x0)+cos(x0);
x1=0;
for i=1:50
    y1=sin(x1)+cos(x1);
    num=x0*y1-x1*y0;
    den=y1-y0;
    x2=num/den;
    x1=x2;
    del(i)=x1-3*pi/4;
end
semilogy(del,'b*');hold on
semilogy(-del,'r*');
xlabel('i');
ylabel('absolute error');
legend('positive values','negative
values')
```

Problem: The maximum of the function:

$$J(y) = \frac{1}{1+y^2}$$

may be found using the following iterative algorithm:

$$y_{i+1} = y_i - \frac{y_i}{1+y_i^2} \quad \text{for } i = 0, 1, 2, \ldots$$

Determine the parameters ρ and C, characterizing the local convergence of that algorithm.

<u>Solution:</u> The only maximum of the function $J(y)$ is located at $\dot{y} = 0$. This is a stationary point of the algorithm because:

$$\dot{y} = \dot{y} - \frac{\dot{y}}{1+\dot{y}^2}$$

Thus, the error equation has the form:

$$\Delta_{i+1} = \Delta_i - \frac{\Delta_i}{1+\Delta_i^2} \quad \text{for } i = 0, 1, 2, \ldots$$

$$\Delta_{i+1} = \frac{\Delta_i + \Delta_i^3 - \Delta_i}{1+\Delta_i^2} \cong \Delta_i^3 \quad \text{for } i = 0, 1, 2, \ldots$$

and consequently $\rho = 3, C = 1$.

3.3. Analysis of multiple-point algorithms

Problem: Compute the parameters of local convergence, C and ρ, of the following iterative algorithm:

$$x_{i+1} = \frac{x_i x_{i-1}\left(x_i + x_{i-1}\right) + 2}{x_i^2 + x_i x_{i-1} + x_{i-1}^2 + 1} \quad \text{for } i = 0, 1, \ldots$$

designed for solving a nonlinear algebraic equation of the form $f(x) = 0$.

<u>Solution:</u> The only stationary point is $\dot{x} = 1$. For this point:

$$1+\Delta_{i+1}=\frac{(1+\Delta_i)(1+\Delta_{i-1})(1+\Delta_i+1+\Delta_{i-1})+2}{(1+\Delta_i)^2+(1+\Delta_i)(1+\Delta_{i-1})+(1+\Delta_{i-1})^2+1}\ \text{ dla }\ i=0,1,\dots$$

$$\Delta_{i+1}=\frac{2(1+\Delta_i)(1+\Delta_{i-1})+(1+\Delta_i)(1+\Delta_{i-1})(\Delta_i+\Delta_{i-1})+2}{(1+\Delta_i)^2+(1+\Delta_i)(1+\Delta_{i-1})+(1+\Delta_{i-1})^2+1}-1\ \text{ dla }\ i=0,1,\dots$$

$$\Delta_{i+1}\cong\frac{(1+\Delta_i)(1+\Delta_{i-1})+(1+\Delta_i)(1+\Delta_{i-1})(\Delta_i+\Delta_{i-1})-(1+\Delta_i)^2-(1+\Delta_{i-1})^2+1}{4}$$

$$\Delta_{i+1}\cong\frac{(1+\Delta_i)(1+\Delta_{i-1})(1+\Delta_i+\Delta_{i-1})-(1+\Delta_i)^2-(1+\Delta_{i-1})^2+1}{4}$$

$$\Delta_{i+1}\cong\frac{\Delta_i\Delta_{i-1}(3+\Delta_i+\Delta_{i-1})}{4}\cong 0.75\Delta_i\Delta_{i-1}$$

This is an expression similar to that obtained for the secant method ($\hat{C}=0.75$). Thus:

$$\rho=0.5\left(1+\sqrt{5}\right)\cong 1.618\ \text{ and }\ C=0.75^{0.5\left(\sqrt{5}-1\right)}\cong 0.837$$

Problem: Determine the parameters of local convergence, ρ and C, and the attainable accuracy for the following iterative algorithm:

$$y_{i+1}=\frac{y_i y_{i-1}(y_i+y_{i-1})-1}{y_i(y_i+y_{i-1})+y_{i-1}^2-2}\quad\text{for }\ i=1,2,\dots$$

in the vicinity of the point $\dot{y}=1$.

Solution: The convergence parameters are determined in the standard way as follows:

$$\Delta_{i+1}=\frac{(1+\Delta_i)(1+\Delta_{i-1})(1+\Delta_i+1+\Delta_{i-1})-1}{(1+\Delta_i)(1+\Delta_i+1+\Delta_{i-1})+(1+\Delta_{i-1})^2-2}-1=$$

$$=\frac{(1+\Delta_i+\Delta_{i-1}+\Delta_i\Delta_{i-1})(2+\Delta_i+\Delta_{i-1})-1}{(1+\Delta_i)(2+\Delta_i+\Delta_{i-1})+(1+\Delta_{i-1})^2-2}-1=$$

$$=\frac{(1+\Delta_i+\Delta_{i-1}+\Delta_i\Delta_{i-1}-1-\Delta_i)(2+\Delta_i+\Delta_{i-1})-1-(1+\Delta_{i-1})^2+2}{(1+\Delta_i)(2+\Delta_i+\Delta_{i-1})+\Delta_{i-1}^2+2\Delta_{i-1}-1}=$$

$$=\frac{\Delta_i^2\Delta_{i-1}+3\Delta_i\Delta_{i-1}+\Delta_i\Delta_{i-1}^2}{1+\dots}\cong\Delta_i\Delta_{i-1}(3+\Delta_i+\Delta_{i-1})\cong 3\Delta_i\Delta_{i-1}$$

$$\Delta_{i+1}=3\Delta_i\left(\frac{1}{C}\Delta_i\right)^{\frac{1}{\rho}}=\frac{3}{C^{\frac{1}{\rho}}}\Delta_i^{1+\frac{1}{\rho}}=C\Delta_i^{\rho}\Rightarrow\begin{cases}\rho=1+\dfrac{1}{\rho}\Rightarrow\rho=\dfrac{1}{2}\left(1+\sqrt{5}\right)\cong 1.62\\ C=3C^{-\frac{1}{\rho}}\Rightarrow C^{1+\frac{1}{\rho}}=3\Rightarrow C^{\rho}=3\Rightarrow C\cong 1.97\end{cases}$$

The attainable accuracy may be assessed taking into account the negligibility of the transmission of errors from one iteration to the next:

$$\tilde{y}_{i+1}=\frac{\left[\dot{y}^2(1+\eta_m')2\dot{y}(1+\eta_s')(1+\eta_m'')-1\right](1+\eta_o')}{\left\{\left[\dot{y}(\dot{y}+\dot{y})(1+\eta_m''')(1+\eta_s'')+y_\infty^2(1+\eta_p)\right](1+\eta_s''')-2\right\}(1+\eta_o'')}(1+\eta_d)=$$

$$=\frac{\left[2(1+\eta_m'+\eta_s'+\eta_m'')-1\right](1+\eta_o'+\eta_d-\eta_o'')}{\left\{\left[2(1+\eta_m'''+\eta_s'')+(1+\eta_p)\right](1+\eta_s''')-2\right\}}=\frac{1+2\eta_m'+2\eta_s'+2\eta_m''+\eta_o'+\eta_d-\eta_o''}{1+2\eta_m'''+2\eta_s''+\eta_p+3\eta'''}$$

$$=1+2\eta_m'+2\eta_s'+2\eta_m''+\eta_o'+\eta_d-\eta_o''-2\eta_m'''-2\eta_s''-\eta_p-3\eta'''$$

$$\left|\delta\left[\tilde{y}_{i+1}\right]\right| \le 17eps$$

Problem: Find the value of a parameter α guaranteeing the convergence of the following iterative algorithm (IA):

$$y_{i+1} = \frac{y_i y_{i-1} + x}{\alpha y_i + y_{i-1}} \quad (i = 1, 2, \ldots)$$

to $\dot{y} = \sqrt{x}$ for $x > 0$.

<u>Solution:</u> The definition of the absolute error of the successive approximations of the solution implies the following sequence of relationships:

$$\Delta_{i+1} \equiv y_{i+1} - \dot{y} = \frac{(\dot{y} + \Delta_i)(\dot{y} + \Delta_{i-1}) + x}{\alpha(\dot{y} + \Delta_i) + (\dot{y} + \Delta_{i-1})} - \dot{y} = \frac{2\dot{y}^2 + \dot{y}\Delta_i + \dot{y}\Delta_{i-1} + \Delta_i\Delta_{i-1}}{\alpha\dot{y} + \alpha\Delta_i + \dot{y} + \Delta_{i-1}} - \dot{y}$$

$$\cong \frac{2\dot{y}^2 + \dot{y}\Delta_i + \dot{y}\Delta_{i-1} + \Delta_i\Delta_{i-1} - \alpha\dot{y}^2 - \alpha\dot{y}\Delta_i - \dot{y}^2 - \dot{y}\Delta_{i-1}}{(1+\alpha)\dot{y}}$$

$$= \frac{(1-\alpha)\dot{y}^2 + (1-\alpha)\dot{y}\Delta_i + \Delta_i\Delta_{i-1}}{(1+\alpha)\dot{y}} = \frac{1-\alpha}{1+\alpha}\dot{y} + \frac{1-\alpha}{1+\alpha}\Delta_i + \frac{1}{(1+\alpha)\dot{y}}\Delta_i\Delta_{i-1}$$

It follows from the latter equality that IA may converge only for $\alpha = 1$ because only for this value of α the constant term of the RHS is zero. For this value of α:

$$\Delta_{i+1} = \frac{1}{2\sqrt{x}}\Delta_i\Delta_{i-1}$$

like for the secant method. Thus: $\rho \cong 1.618$, and IA is convergent for any $x > 0$.

Problem: Determine the parameters of local convergence, ρ and C, and the attainable accuracy for the following iterative algorithm:

$$y_{i+1} = y_i + \frac{\cos(y_i)}{\sin(y_{i-1})} \quad \text{for } i = 1, 2, \ldots$$

in the vicinity of the point $\dot{y} = \pi/2$.

<u>Solution:</u> The convergence parameters are determined in the standard way as follows:

$$\cos(y_i) = \cos\left(\tfrac{\pi}{2} + \Delta_i\right) \cong \cos\left(\tfrac{\pi}{2}\right) - \sin\left(\tfrac{\pi}{2}\right)\Delta_i - \tfrac{1}{2}\cos\left(\tfrac{\pi}{2}\right)\Delta_i^2 = -\Delta_i$$

$$\sin(y_{i-1}) = \sin\left(\tfrac{\pi}{2} + \Delta_{i-1}\right) \cong \sin\left(\tfrac{\pi}{2}\right) + \cos\left(\tfrac{\pi}{2}\right)\Delta_{i-1} - \tfrac{1}{2}\sin\left(\tfrac{\pi}{2}\right)\Delta_{i-1}^2 = 1 - \tfrac{1}{2}\Delta_{i-1}^2$$

$$\Delta_{i+1} = \Delta_i + \frac{-\Delta_i}{1 - \frac{1}{2}\Delta_{i-1}^2} = \Delta_i\left(1 - \frac{1}{1 - \frac{1}{2}\Delta_{i-1}^2}\right) = \frac{-\frac{1}{2}\Delta_i\Delta_{i-1}^2}{1 - \frac{1}{2}\Delta_{i-1}^2} \cong -\tfrac{1}{2}\Delta_i\Delta_{i-1}^2$$

$$\Delta_{i+1} = -\tfrac{1}{2}\Delta_i\left[\left(\tfrac{1}{C}\Delta_i\right)^{\frac{1}{\rho}}\right]^2 = -\tfrac{1}{2}C^{-\frac{2}{\rho}}\Delta_i^{1+\frac{2}{\rho}} = C\Delta_i^{\rho} \Rightarrow \begin{cases} \rho = 1 + \frac{2}{\rho} \Rightarrow \rho = 2 \\ |C| = \frac{1}{2}|C|^{-\frac{2}{\rho}} = \frac{1}{2}|C|^{-1} \Rightarrow |C| = \frac{1}{\sqrt{2}} \end{cases}$$

The attainable accuracy may be assessed taking into account the negligibility of the transmission of errors from one iteration to the next:

$$\tilde{y}_{i+1} = \left[\tfrac{\pi}{2} + \frac{\cos\left(\frac{\pi}{2}\right)(1+\eta_{\cos})}{\sin\left(\frac{\pi}{2}\right)(1+\eta_{\sin})}(1+\eta_d)\right](1+\eta_s) = \tfrac{\pi}{2}(1+\eta_s)$$

$$\left|\delta\left[\tilde{y}_{i+1}\right]\right| \le eps$$

3.4. Estimation of the roots of polynomials

Problem: Assess the relative error of an estimate $\hat{Q}$ of the quality indicator:

$$Q = \frac{f_3 - f_1}{f_2}$$

characterising an electronic circuit, if it is calculated on the basis of the approximate values $\tilde{f}_1 = 1.001$, $\tilde{f}_2 = 2.002$ and $\tilde{f}_3 = 2.997$ of the frequencies f_1, f_2 and f_3, respectively, obtained by means of the Newton's method applied to the equation:

$$f^3 - 6f^2 + 11f - 6 = 0$$

<u>Solution:</u> The exact solutions of the above equation are: $f_1 = 1$, $f_2 = 2$ and $f_3 = 3$; thus, the absolute errors in the frequency values obtained by numerically solving this equation are: $\Delta\tilde{f}_1 = 0.001$, $\Delta\tilde{f}_2 = 0.002$ and $\Delta\tilde{f}_3 = -0.003$. The computed value of Q is:

$$\hat{Q} = \frac{3 + \Delta\tilde{f}_3 - 1 - \Delta\tilde{f}_1}{2 + \Delta\tilde{f}_2} = \frac{2 + \Delta\tilde{f}_3 - \Delta\tilde{f}_1}{2 + \Delta\tilde{f}_2} \cong 1 + \tfrac{1}{2}\left(\Delta\tilde{f}_3 - \Delta\tilde{f}_1 - \Delta\tilde{f}_2\right)$$

$$\delta\left[\hat{Q}\right] \cong \tfrac{1}{2}\left(\Delta\tilde{f}_3 - \Delta\tilde{f}_1 - \Delta\tilde{f}_2\right) = \tfrac{1}{2}\left(-0.003 - 0.001 - 0.002\right) = -0.003$$

Problem: The estimates $\hat{y}_1$, $\hat{y}_2$ and $\hat{y}_3$ of the roots $\dot{y}_1 = 1$, $\dot{y}_2 = j$ and $\dot{y}_3 = -j$, respectively, of a third-order polynomial have been computed in the following way:

- an estimate $\hat{y}_1 = \dot{y}_1\left(1 + \eta_1\right)$ of $\dot{y}_1$ has been determined by means of an iterative method guaranteeing $\left|\eta_1\right| \le 10^{-6}$;
- the estimates of $\hat{y}_2$ and $\hat{y}_3$ of $\dot{y}_2$ and $\dot{y}_3$, respectively, have been obtained by solving the quadratic equation resulting from linear deflation.

Assess the relative error the estimate $\hat{y}_2$.

<u>Solution:</u>

$$y^3 - y^2 + y - 1 = 0$$

The coefficients of the quadratic equation resulting from linear deflation:

$$\tilde{b}_2 y^2 + \tilde{b}_1 y + \tilde{b}_0 = 0$$

(subject to errors inherited from $\hat{y}_1$) may be obtained by comparing the third-order polynomial of the form:

$$\left(\tilde{b}_2 y^2 + \tilde{b}_1 y + \tilde{b}_0\right)\left(y - \hat{y}_1\right) = \tilde{b}_2 y^3 + \left(\tilde{b}_1 - \tilde{b}_2\hat{y}_1\right)y^2 + \left(\tilde{b}_0 - \tilde{b}_1\hat{y}_1\right)y - \tilde{b}_0\hat{y}_1$$

with the third-order polynomial whose roots are estimated:

$$y^3 - y^2 + y - 1$$

i.e. by solving the following set of linear algebraic equations:

$$\tilde{b}_2 = 1, \quad \tilde{b}_1 - \tilde{b}_2\hat{y}_1 = -1, \quad \tilde{b}_0 - \tilde{b}_1\hat{y}_1 = 1$$

Hence:

$$\tilde{b}_1 = \tilde{b}_2\hat{y}_1 - 1 = 1 + \eta_1 - 1 = \eta_1 \text{ and } \tilde{b}_0 = \tilde{b}_1\hat{y}_1 + 1 = \eta_1\left(1 + \eta_1\right) + 1 \cong 1 + \eta_1$$

Thus, $\hat{y}_2$ is a root of the following polynomial:

$$y^2+\eta_1 y+(1+\eta_1)=0$$

It may be therefore determined by means of the so-called school method:

$$\Delta=\eta_1^2-4(1+\eta_1)\cong-4(1+\varepsilon)\eta_1 \;\Rightarrow \sqrt{\Delta}=\sqrt{-4(1+\eta_1)}\cong 2j\left(1+\tfrac{1}{2}\eta_1\right)$$

$$\tilde{y}_2=\frac{-\eta_1+2j\left(1+\frac{1}{2}\eta_1\right)}{2}=-\tfrac{1}{2}\eta_1+j\left(1+\tfrac{1}{2}\eta_1\right)=j-\tfrac{1}{2}\eta_1+\tfrac{1}{2}j\eta_1 \;=j\left(1+\tfrac{1}{2}j\eta_1+\tfrac{1}{2}\eta_1\right)$$

$$\Rightarrow\left|\delta\left[\tilde{y}_2\right]\right|=\left|\tfrac{1}{2}j\eta_1+\tfrac{1}{2}\eta_1\right|=\tfrac{1}{2}|\eta_1||1+j|=\tfrac{1}{\sqrt{2}}|\eta_1|\le\tfrac{1}{\sqrt{2}}10^{-6}$$

The same result may be obtained without referring to any particular method for estimation of the root $\dot{y}_2=j$, *viz.*:

$$\hat{y}_2^2+\eta_1\hat{y}_2+(1+\eta_1)=0 \text{ with } \hat{y}_2=j(1+\eta_2)$$

$$\left[j(1+\eta_2)\right]^2+\eta_1 j(1+\eta_2)+(1+\eta_1)=0$$

$$-1(1+2\eta_2)+j\eta_1+(1+\eta_1)=0$$

$$\eta_2=\tfrac{1}{2}(1+j)\eta_1$$

3.5. Accuracy of numerical solutions

Problem: Assess the relative error of the solution to the equation:

$$ax-x^a=0$$

caused by the relative error of the parameter a, not exceeding $p=1\%$.

<u>Solution #1:</u> The direct differentiation of the LHS of the equation yields:

$$(ax)'=x+ax'$$

$$\left(x^a\right)'=ax^{a-1}x'+\ln(x)x^a$$

The latter have been obtained using the following rule of differentiation:

$$\frac{d}{dx}F\left(f_1(x),f_2(x)\right)=\left.\frac{\partial F(y_1,y_2)}{\partial y_1}\right|_{\substack{y_1=f_1(x)\\ y_2=f_2(x)}}\frac{df_1(x)}{dx}+\left.\frac{\partial F(y_1,y_2)}{\partial y_2}\right|_{\substack{y_1=f_1(x)\\ y_2=f_2(x)}}\frac{df_2(x)}{dx}$$

Taking into account that by definition of the solution $x^a=ax$, one may simplify the second term in the following way:

$$\left(x^a\right)'=a^2x'+ax\ln(x)$$

Hence the equation with respect to x' whose solution is:

$$x'=\frac{x(1-a\ln(x))}{a(a-1)}$$

Consequently:

$$T(a)\equiv\frac{a}{x}x'=\frac{1-a\ln(x)}{a-1}$$

Taking into account that by definition of the solution $\ln\left(x^a\right)=\ln(ax)$, one may get rid of $\ln(x)$:

$$T(a)=\frac{(a-1)-a\ln(a)}{(a-1)^2}$$

Hence the assessment:

$$\left|\delta x\right| \le \left|T(a)\right| \cdot \left|\delta a\right| \le \left|T(a)\right| \cdot p = \frac{\left|a\left(1-\ln\left(a\right)\right)-1\right|}{\left(a-1\right)^2} 10^{-2}$$

Solution #2: The logarithm of the equation, rewritten in the form $ax = x^a$, has the form:

$$\ln\left(ax\right) = \ln\left(x^a\right) \;\Rightarrow\; \ln\left(a\right)+\ln\left(x\right) = a\ln\left(x\right) \;\Rightarrow \ln\left(x\right) = \frac{\ln\left(a\right)}{a-1}$$

The differentiation of the LHS and RHS with respect to a:

$$\frac{1}{x}x' = \frac{\frac{1}{a}\left(a-1\right)-\ln\left(a\right)}{a-1} \;\Rightarrow\; \frac{1}{x}x' = \frac{a\left(1-\ln\left(a\right)\right)-1}{a\left(a-1\right)^2}$$

enables one to quickly determine the function characterising the relative error propagation:

$$T\left(a\right) \equiv \frac{a}{x}x' = \frac{\left(a-1\right)-a\ln\left(a\right)}{\left(a-1\right)^2}$$

Hence the assessment:

$$\left|\delta x\right| \le \left|T(a)\right| \cdot \left|\delta a\right| \le \left|T(a)\right| \cdot p = \frac{\left|a\left(1-\ln\left(a\right)\right)-1\right|}{\left(a-1\right)^2} 10^{-2}$$

Problem: Assess the absolute error of the solution to the equation:

$$\sin\left(x+a\right)+\cos(x) = 0 \;\text{ for }\; a \in \left[0, \tfrac{\pi}{2}\right]$$

caused by the relative error of the parameter a, not exceeding $p = 1\%$.

Solution: The equation may be given the form:

$$\left[\sin\left(x\right)\cos(a)+\sin\left(a\right)\cos(x)\right]+\cos(x) = 0$$

which implies:

$$\tan\left(x\right) = -\frac{1+\sin\left(a\right)}{\cos(a)}$$

The differentiation of the LHS and RHS of this formula with respect to a yields:

$$\frac{1}{\cos^2\left(x\right)}\frac{dx}{da}\cdot = -\frac{\cos^2\left(a\right)+\left[1+\sin\left(a\right)\right]\sin\left(a\right)}{\cos^2\left(a\right)} = -\frac{1+\sin\left(a\right)}{\cos^2\left(a\right)}$$

Taking into account that $\cos^2\left(x\right)$ may be expressed by $\tan\left(x\right)$:

$$\cos^2\left(x\right) = \frac{1}{1+\tan^2\left(x\right)} = \frac{1}{1+\left[\dfrac{1+\sin\left(a\right)}{\cos(a)}\right]^2} = \frac{\cos^2\left(a\right)}{\cos^2\left(a\right)+1+2\sin\left(a\right)+\sin^2\left(a\right)} = \frac{\cos^2\left(a\right)}{2+2\sin\left(a\right)}$$

one can get:

$$\frac{dx}{da}\cdot = -\frac{1+\sin\left(a\right)}{\cos^2\left(a\right)}\cdot\frac{\cos^2\left(a\right)}{2+2\sin\left(a\right)} = -\tfrac{1}{2}$$

Hence the assessment:

$$\left|\Delta x\right| \le \left|\frac{dx}{da}\right| \cdot \left|\Delta a\right| \le -\frac{1}{2}1\%\left|a\right| = 5\cdot 10^{-3}\left|a\right|$$

Problem: Assess the absolute error of the solution to the equation:

$$\int_0^1 \left[\sin(\omega x) - \cos(\omega x + a)\right] d\omega = \frac{1+\sin(a)}{x} \quad \text{for } a \in \left[0, \tfrac{\pi}{2}\right]$$

caused by the relative error of the parameter a, not exceeding $p = 1\%$.

<u>Solution:</u> LHS, after integration, takes on the form:

$$\text{LHS} = \left[-\frac{1}{x}\cos(\omega x) - \frac{1}{x}\sin(\omega x + a)\right]_0^1 = -\frac{1}{x}\left[\cos(x) + \sin(x+a) - 1 - \sin(a)\right]$$

Consequently the whole equation reduces to:

$$\sin(x+a) + \cos(x) = 0$$

The latter may be given the form:

$$\left[\sin(x)\cos(a) + \sin(a)\cos(x)\right] + \cos(x) = 0$$

which implies:

$$\tan(x) = -\frac{1+\sin(a)}{\cos(a)}$$

The differentiation of the LHS and RHS of this formula with respect to a yields:

$$\frac{1}{\cos^2(x)}\frac{dx}{da} \cdot = -\frac{\cos^2(a) + \left[1+\sin(a)\right]\sin(a)}{\cos^2(a)} = -\frac{1+\sin(a)}{\cos^2(a)}$$

Taking into account that $\cos^2(x)$ may be expressed by $\tan(x)$:

$$\cos^2(x) = \frac{1}{1+\tan^2(x)} = \frac{1}{1+\left[\dfrac{1+\sin(a)}{\cos(a)}\right]^2} = \frac{\cos^2(a)}{\cos^2(a) + 1 + 2\sin(a) + \sin^2(a)} = \frac{\cos^2(a)}{2+2\sin(a)}$$

one can get:

$$\frac{dx}{da} \cdot = -\frac{1+\sin(a)}{\cos^2(a)} \cdot \frac{\cos^2(a)}{2+2\sin(a)} = -\tfrac{1}{2}$$

Hence the assessment:

$$|\Delta x| \le \left|\frac{dx}{da}\right| \cdot |\Delta a| \le -\tfrac{1}{2} 1\% |a| = 5 \cdot 10^{-3} |a|$$

ENUME: SOLVED PROBLEMS

4. INTERPOLATION AND APPROXIMATION

4.1. Interpolation

Problem: Compute the coefficients of the third-order algebraic polynomial, $y = \hat{f}(x)$, interpolating the following data:

n	0	1	2	3
x_n	–1	0	1	2
y_n	–4	–1	0	5

Compare the values $\hat{f}(x_n)$ with the corresponding values y_n in the above table.

<u>Solution:</u> The interpolating Lagrange polynomial is:

$$\hat{f}(x) = \sum_{n=0}^{3} y_n L_n(x) \text{ where } L_n(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq n}}^{3} \frac{x - x_\nu}{x_n - x_\nu}$$

The elementary Lagrange polynomials $L_n(x)$, after substitution of the data, take on the form:

$$L_0(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq 0}}^{3} \frac{x - x_\nu}{x_0 - x_\nu} = -\frac{1}{6}x^3 + \frac{1}{2}x^2 - \frac{1}{3}x, \qquad L_1(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq 1}}^{3} \frac{x - x_\nu}{x_1 - x_\nu} = \frac{1}{2}x^3 - x^2 - \frac{1}{2}x + 1,$$

$$L_2(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq 2}}^{3} \frac{x - x_\nu}{x_2 - x_\nu} = -\frac{1}{2}x^3 + \frac{1}{2}x^2 + x, \qquad L_3(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq 3}}^{3} \frac{x - x_\nu}{x_3 - x_\nu} = \frac{1}{6}x^3 - \frac{1}{6}x$$

Hence:

$$\hat{f}(x) = (-4)\cdot\left(-\frac{1}{6}x^3 + \frac{1}{2}x^2 - \frac{1}{3}x\right) + (-1)\cdot\left(\frac{1}{2}x^3 - x^2 - \frac{1}{2}x + 1\right) + 0\cdot\left(-\frac{1}{2}x^3 + \frac{1}{2}x^2 + x\right)$$
$$+ 5\cdot\left(\frac{1}{6}x^3 - \frac{1}{6}x\right) = x^3 - x^2 + x - 1$$

Problem: Compute the coefficients of the second-order algebraic polynomial, $y = \hat{f}(x)$, interpolating the following data:

n	0	1	2
x_n	–1	0	1
y_n	1	–1	1

Under an assumption that the data y_n are corrupted with random errors, which may be adequately modelled with statistically independent random variables following the distribution $\mathcal{N}\left(0; \sigma^2\right)$, determine the variance of the random absolute error of $\hat{f}(x)$.

<u>Solution:</u> The interpolating Lagrange polynomial is:

$$\hat{f}(x) = \sum_{n=0}^{2} y_n L_n(x)$$

where $L_n(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq n}}^{2} \frac{x - x_\nu}{x_n - x_\nu}$ are elementary Lagrange polynomials which after substitution of the data, take on the form:

$$L_0(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq 0}}^{2} \frac{x - x_\nu}{x_0 - x_\nu} = \frac{1}{2}x^2 - \frac{1}{2}x, \; L_1(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq 1}}^{2} \frac{x - x_\nu}{x_1 - x_\nu} = 1 - x^2, \; L_2(x) \equiv \prod_{\substack{\nu=0 \\ \nu \neq 2}}^{2} \frac{x - x_\nu}{x_2 - x_\nu} = \frac{1}{2}x^2 + \frac{1}{2}x$$

Hence:

$$\hat{f}(x) = 1 \cdot \left(\frac{1}{2}x^2 - \frac{1}{2}x \right) + (-1) \cdot (1 - x^2) + 1 \cdot \left(\frac{1}{2}x^2 + \frac{1}{2}x \right) = 2x^2 - 1$$

The error-corrupted data are modelled with the following random variables:

$$\underline{y_0} = \dot{y}_0 + \underline{\Delta y_0}, \; \underline{y_1} = \dot{y}_1 + \underline{\Delta y_1} \text{ and } \underline{y_2} = \dot{y}_2 + \underline{\Delta y_2}$$

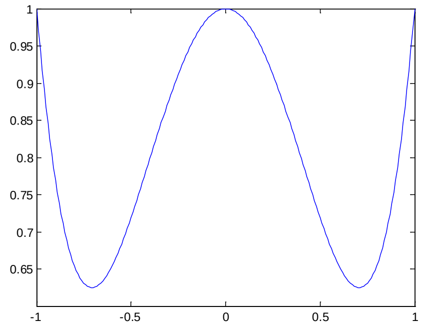
where $\dot{y}_0$, $\dot{y}_1$ and $\dot{y}_2$ are exact values of those data, and $\underline{\Delta y_0}$, $\underline{\Delta y_1}$ and $\underline{\Delta y_2}$ random variables following the distribution $\mathcal{N}(0; \sigma^2)$. The result of interpolation is a random function:

$$\underline{\hat{f}}(x) = \sum_{n=0}^{2} \underline{y}_n L_n(x)$$

whose variance may be calculated as follows:

$$\mathrm{Var}\left[\underline{\hat{f}}(x) \right] = \sum_{n=0}^{2} \mathrm{Var}\left[\underline{y}_n \right] \cdot \left[L_n(x) \right]^2 = \sigma^2 \sum_{n=0}^{2} \left[L_n(x) \right]^2 = \left(\frac{3}{2}x^4 - \frac{3}{2}x^2 + 1 \right) \sigma^2$$

The graph of the function $\dfrac{\mathrm{Var}\left[\underline{\hat{f}}(x) \right]}{\sigma^2} = \dfrac{3}{2}x^4 - \dfrac{3}{2}x^2 + 1$ is shown in the figure below.



Problem: Assess the variance of the random component of the absolute error of the second-order polynomial, $y = \hat{f}(x)$, interpolating the following points on the x-y plane:

n	0	1	2
x_n	–1	0	1
y_n	1	–1	1

under an assumption that the data x_0 and x_2 are subject to zero-mean random relative errors with the variance σ^2.

<u>Solution #1:</u> The error-corrupted data x_0 and x_2 may be adequately modelled with the following random variables:

$$\underline{x_0} = -\left(1 + \underline{\varepsilon_0}\right) \text{ and } \underline{x_2} = \left(1 + \underline{\varepsilon_2}\right)$$

where $\underline{\varepsilon_0}$ and $\underline{\varepsilon_2}$ are independent zero-mean random variables with the variance σ^2. The corresponding Lagrange polynomial has the form:

$$\underline{\hat{f}(x)} = \sum_{n=0}^{2} y_n \underline{L}_n(x)$$

where:

$$\underline{L}_0(x) \equiv \prod_{\substack{\nu=0\\ \nu\neq 0}}^{2} \frac{x-\underline{x}_\nu}{\underline{x}_0-\underline{x}_\nu} = \frac{x-x_1}{\underline{x}_0-x_1}\cdot\frac{x-\underline{x}_2}{\underline{x}_0-\underline{x}_2} = \frac{x-0}{-(1+\underline{\varepsilon}_0)-0}\cdot\frac{x-(1+\underline{\varepsilon}_2)}{-(1+\underline{\varepsilon}_0)-(1+\underline{\varepsilon}_2)}$$

$$= \frac{1}{2}x(x-1)\left(1-\frac{3}{2}\underline{\varepsilon}_0-\frac{x+1}{2(x-1)}\underline{\varepsilon}_2\right)$$

$$\underline{L}_1(x) \equiv \prod_{\substack{\nu=0\\ \nu\neq 1}}^{2} \frac{x-\underline{x}_\nu}{x_1-\underline{x}_\nu} = \frac{x-\underline{x}_0}{x_1-\underline{x}_0}\cdot\frac{x-\underline{x}_2}{x_1-\underline{x}_2} = \frac{x+(1+\underline{\varepsilon}_0)}{0+(1+\underline{\varepsilon}_0)}\cdot\frac{x-(1+\underline{\varepsilon}_2)}{0-(1+\underline{\varepsilon}_2)}$$

$$= -(x-1)(x+1)\left(1-\frac{x}{x+1}\underline{\varepsilon}_0-\frac{x}{x-1}\underline{\varepsilon}_2\right)$$

$$\underline{L}_2(x) \equiv \prod_{\substack{\nu=0\\ \nu\neq 2}}^{2} \frac{x-\underline{x}_\nu}{x_2-\underline{x}_\nu} = \frac{x-\underline{x}_0}{\underline{x}_2-\underline{x}_0}\cdot\frac{x-x_1}{\underline{x}_2-x_1} = \frac{x+(1+\underline{\varepsilon}_0)}{(1+\underline{\varepsilon}_2)+(1+\underline{\varepsilon}_0)}\cdot\frac{x-0}{(1+\underline{\varepsilon}_2)-0}$$

$$= \frac{1}{2}x(x+1)\left(1-\frac{x-1}{2(x+1)}\underline{\varepsilon}_0-\frac{3}{2}\underline{\varepsilon}_2\right)$$

Thus, the absolute errors of the elementary Lagrange functions may be given the form:

$$\underline{\Delta L_0}(x) = -\frac{3}{4}x(x-1)\underline{\varepsilon}_0-\frac{1}{4}x(x+1)\underline{\varepsilon}_2$$

$$\underline{\Delta L_1}(x) = x(x-1)\underline{\varepsilon}_0+x(x+1)\underline{\varepsilon}_2$$

$$\underline{\Delta L_2}(x) = -\frac{1}{4}x(x-1)\underline{\varepsilon}_0-\frac{3}{4}x(1+x)\underline{\varepsilon}_2$$

and, consequently, the absolute error of the interpolating polynomial is:

$$\begin{aligned}\underline{\Delta\hat{f}}(x) = \sum_{n=0}^{2} y_n \underline{\Delta L}_n(x) &= -\frac{3}{4}x(x-1)\underline{\varepsilon}_0-\frac{1}{4}x(x+1)\underline{\varepsilon}_2\\ &\quad -x(x-1)\underline{\varepsilon}_0-x(x+1)\underline{\varepsilon}_2\\ &\quad -\frac{1}{4}x(x-1)\underline{\varepsilon}_0-\frac{3}{4}x(x+1)\underline{\varepsilon}_2 = -2x(x-1)\underline{\varepsilon}_0-2x(x+1)\underline{\varepsilon}_2\end{aligned}$$

Its variance may be assessed as follows:

$$\text{Var}\left[\underline{\Delta\hat{f}(x)}\right] = \left[-2x(1-x)\right]^2\sigma^2+\left[-2x(1+x)\right]^2\sigma^2 = 8x^2\left(x^2+1\right)\sigma^2 \le 16\sigma^2$$

Solution #2: The alternative solution is based on the following form of the interpolating polynomial:

$$\underline{\hat{f}(x)} = \underline{a}x^2+\underline{b}x+\underline{c}$$

The random variables modelling the coefficients of this polynomial should satisfy the following equations:

$$\underline{a}\underline{x}_0^2+\underline{b}\underline{x}_0+\underline{c} = y_0$$

$$\underline{a}x_1^2+\underline{b}x_1+\underline{c} = y_1$$

$$\underline{a}\underline{x}_2^2+\underline{b}\underline{x}_2+\underline{c} = y_2$$

which after substitution of $\underline{x}_0 = -(1+\underline{\varepsilon}_0)$ and $\underline{x}_2 = (1+\underline{\varepsilon}_2)$ take on the form:

$$\underline{a}\left(1+\underline{\varepsilon_0}\right)^2-\underline{b}\left(1+\underline{\varepsilon_0}\right)+\underline{c}=1$$
$$\underline{c}=-1$$
$$\underline{a}\left(1+\underline{\varepsilon_2}\right)^2+\underline{b}\left(1+\underline{\varepsilon_2}\right)+\underline{c}=1$$

Hence:

$$\underline{a}\left(1+2\underline{\varepsilon_0}\right)-\underline{b}\left(1+\underline{\varepsilon_0}\right)\cong 2 \quad\Rightarrow\quad \underline{a}\left(1+\underline{\varepsilon_0}\right)-\underline{b}\cong 2\left(1-\underline{\varepsilon_0}\right)$$
$$\underline{a}\left(1+2\underline{\varepsilon_2}\right)+\underline{b}\left(1+\underline{\varepsilon_2}\right)\cong 2 \quad\Rightarrow\quad \underline{a}\left(1+\underline{\varepsilon_2}\right)+\underline{b}\cong 2\left(1-\underline{\varepsilon_2}\right)$$

and:

$$\underline{a}\cong 2\left(1-\underline{\varepsilon_0}-\underline{\varepsilon_2}\right)$$
$$\underline{b}\cong 2\left(\underline{\varepsilon_0}-\underline{\varepsilon_2}\right)$$

Thus:

$$\underline{\hat{f}\left(x\right)}=\underline{a}x^2+\underline{b}x+\underline{c}=\left(2x^2+1\right)+\left[-2x\left(x-1\right)\underline{\varepsilon_0}-2x\left(x+1\right)\underline{\varepsilon_2}\right]$$

and:

$$\underline{\Delta\hat{f}\left(x\right)}=-2x\left(x-1\right)\underline{\varepsilon_0}-2x\left(x+1\right)\underline{\varepsilon_2}$$

i.e. the same as in the Solution #1.

Problem: Compute the estimates $\hat{a}$, $\hat{b}$ and $\hat{c}$ of the parameters a, b and c of the function:

$$y=\hat{f}\left(x;a,b,c\right)\equiv a\cdot\sin\left[c\left(x-x_n\right)\right]+b\cdot\cos\left[c\left(x-x_n\right)\right],$$

interpolating the data: $\left(x_{n-1}=1,\ y_{n-1}=1\right)$, $\left(x_n=2,\ y_n=4\right)$ and $\left(x_{n+1}=3,\ y_{n+1}=0\right)$.

Solution: The interpolation condition has the form:

$$a\cdot\sin\left[c\left(x_{n-1}-x_n\right)\right]+b\cdot\cos\left[c\left(x_{n-1}-x_n\right)\right]=y_{n-1}$$
$$a\cdot\sin\left[c\left(x_n-x_n\right)\right]+b\cdot\cos\left[c\left(x_n-x_n\right)\right]=y_n$$
$$a\cdot\sin\left[c\left(x_{n+1}-x_n\right)\right]+b\cdot\cos\left[c\left(x_{n+1}-x_n\right)\right]=y_{n+1}$$

The second equation implies: $\hat{b}=y_n$; thus:

$$-a\cdot\sin\left(c\cdot h\right)+y_n\cdot\cos\left(c\cdot h\right)=y_{n-1}$$
$$a\cdot\sin\left(c\cdot h\right)+y_n\cdot\cos\left(c\cdot h\right)=y_{n+1}$$

with $h\equiv x_n-x_{n-1}=x_{n+1}-x_n$. The sum of the above equations is a nonlinear algebraic equation:

$$2\cdot y_n\cdot\cos\left(c\cdot h\right)=y_{n-1}+y_{n+1}$$

whose solution is:

$$\hat{c}=\frac{1}{h}\arccos\left(z_n\right)\ \text{ with }\ z_n\equiv\frac{y_{n-1}+y_{n+1}}{2\cdot y_n}$$

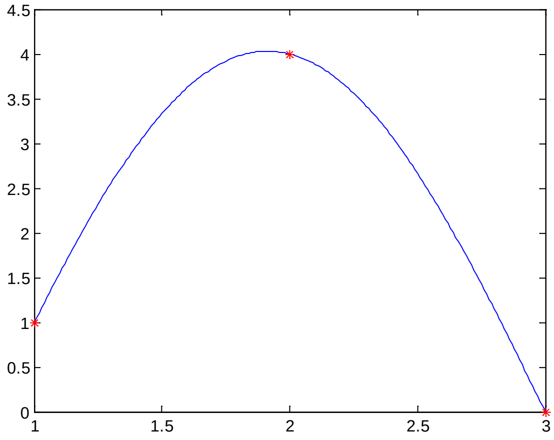
The difference of the above equations is a nonlinear algebraic equation:

$$2\cdot a\cdot\sin\left(c\cdot h\right)=y_{n+1}-y_{n-1}$$

whose solution is:

$$\hat{a}=\frac{y_{n+1}-y_{n-1}}{2\cdot\sin\left(\hat{c}\cdot h\right)}=\frac{y_{n+1}-y_{n-1}}{2\cdot\sqrt{1-z_n^2}}$$

After substitution of the data: $\hat{a}\cong-0.50$, $\hat{b}\cong 4.00$ and $\hat{c}=1.45$.




```
clear all
xw=[1 2 3]';yw=[1 4 0]';
z=0.5*(yw(1)+yw(3))/yw(2);
a=0.5*(yw(3)-yw(1))/sqrt(1-z*z)
b=yw(2)
c=acos(z)
x=[1:0.01:3];
y=a*sin(c*(x-xw(2)))+b*cos(c*(x-xw(2)));
plot(x,y);hold on
plot(xw,yw,'*')
hold off
```

Problem: Determine the coefficients of a cubic spline function $y = s(x)$, interpolating the data:

n	0	1	2
x_n	-1	0	1
y_n	0	1	-1

and satisfying the following boundary conditions: $s'(-1_+) = 2$ and $s'(1_-) = -9$.

Solution: The function $s(x)$ should have the form:

$$s(x) = \begin{cases} a_0(x+1)^3 + b_0(x+1)^2 + c_0(x+1) + d_0 & \text{for } x \in (-1, 0) \\ a_1 x^3 + b_1 x^2 + c_1 x + d_1 & \text{for } x \in (0, 1) \end{cases}$$

The conditions of interpolation and continuity od $s(x)$ are the following:

$$s(-1_+) = d_0 = 0$$

$$s(0_-) = a_0 + b_0 + c_0 + d_0 = 1 \;\Rightarrow\; a_0 + b_0 + c_0 = 1$$

$$s(0_+) = d_1 = 1$$

$$s(1_-) = a_1 + b_1 + c_1 + d_1 = -1 \;\Rightarrow\; a_1 + b_1 + c_1 = -2$$

The continuity condition for $s'(x)$ has the form:

$$3a_0(x+1)^2 + 2b_0(x+1) + c_0\Big|_{x=0} = 3a_1 x^2 + 2b_1 x + c_1\Big|_{x=0}$$

or:

$$3a_0 + 2b_0 + c_0 = c_1$$

The continuity condition for $s''(x)$ has the form:

$$6a_0(x+1) + 2b_0\Big|_{x=0} = 6a_1 x + 2b_1\Big|_{x=0}$$

or:

$$6a_0 + 2b_0 = 2b_1 \;\Rightarrow\; 3a_0 + b_0 = b_1$$

The boundary conditions may be expressed in the form:

$$3a_0(x+1)^2 + 2b_0(x+1) + c_0\Big|_{x=-1} = 2 \;\Rightarrow\; c_0 = 2$$

$$3a_1 x^2 + 2b_1 x + c_1\Big|_{x=1} = -9 \;\Rightarrow\; 3a_1 + 2b_1 + c_1 = -9$$

Since three coefficients ($c_0 = 2$, $d_0 = 0$ and $d_1 = 1$) are already known, the other five may be obtained by solving the following equations:

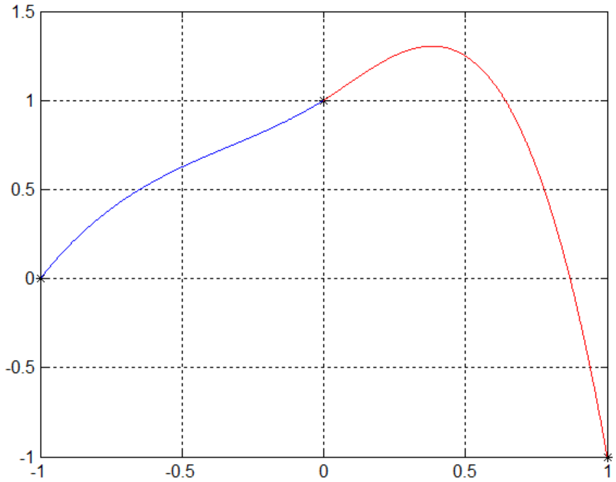
$$a_0 + b_0 = -1$$
$$a_1 + b_1 + c_1 = -2$$
$$3a_0 + b_0 = b_1$$
$$3a_0 + 2b_0 + 2 = c_1$$
$$3a_1 + 2b_1 + c_1 = -9$$

The result is:

$$a_0 = 1,\ b_0 = -2,\ a_1 = -4,\ b_1 = 1,\ c_1 = 1$$

Thus:

$$s(x) = \begin{cases} (x+1)^3 - 2(x+1)^2 + 2(x+1) & \text{for } x \in (-1, 0) \\ -4x^3 + x^2 + x + 1 & \text{for } x \in (0, 1) \end{cases}$$




```
xw=[-1 0 1];
yw=[0 1 -1];
plot(xw,yw,'k*');hold on
grid on

x=[-1:0.001:0];
x1=x+1
y=x1.^3-2*x1.^2+2*x1;
plot(x,y,'b');hold on

x=[0:0.001:1];
y=-4*x.^3+x.^2+x+1;
plot(x,y,'r');hold on
```

4.2. Least-squares approximation

Problem: Compute the estimates $\hat{p}_0$ and $\hat{p}_1$ of the parameters p_0 and p_1 of the function: $y = \hat{f}(x; p_0, p_1) \equiv p_0 + p_1 x$, approximating the data:

n	1	2	3	4
x_n	-1	0	1	2
y_n	0.1	0.9	1.9	3.1

in the sense of the criterion: $J(p_0, p_1) = \sum_{n=1}^{4} \left[y_n - (p_0 + p_1 x_n) \right]^2$. Draw the function $\hat{f}(x; \hat{p}_0, \hat{p}_1)$; indicate the points (x_n, y_n) for $n = 1, 2, 3, 4$. Assess the absolute errors of the estimates $\hat{p}_0$ and $\hat{p}_1$, implied by the absolute errors of the data y_n; assume that the magnitudes of those errors are not greater than 0.01.

<u>Solution:</u> The necessary condition for the minimum of $J(p_0, p_1)$ is:

$$\frac{\partial J(p_0, p_1)}{\partial p_0} = 2\sum_{n=1}^{4} \left[y_n - (p_0 + p_1 x_n) \right](-1) = 0$$
$$\frac{\partial J(p_0, p_1)}{\partial p_1} = 2\sum_{n=1}^{4} \left[y_n - (p_0 + p_1 x_n) \right](-x_n) = 0$$

or:

$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{p} = \mathbf{\Phi}^T\mathbf{y}$$

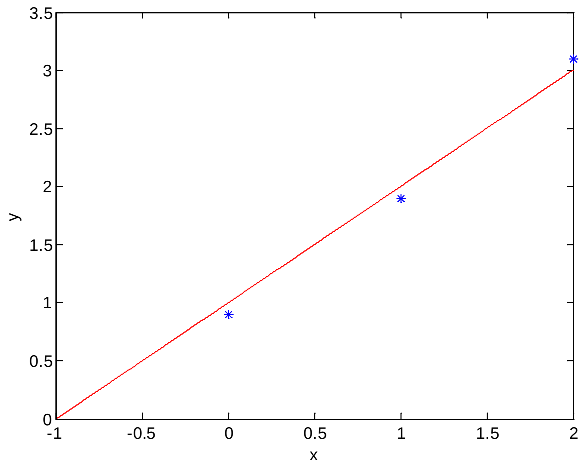
with:

$$\mathbf{\Phi} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \ \mathbf{p} = \begin{bmatrix} p_0 \\ p_1 \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.9 \\ 1.9 \\ 3.1 \end{bmatrix}$$

Hence the equation:

$$\begin{bmatrix} 4 & 2 \\ 2 & 6 \end{bmatrix}\mathbf{p} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}$$

whose solution is: $\hat{p}_0 = 1$, $\hat{p}_1 = 1$.




```
clear all
xw=[-1 0 1 2]';yw=[0.1 0.9 1.9 3.1]';
plot(xw,yw,'*');hold on
Fi=[1 1 1 1;-1 0 1 2]';
p=Fi\yw;
x=linspace(-1,2,1000);y=p(1)+p(2)*x;
plot(x,y,'r');hold off;
xlabel('x');ylabel('y');
```

The equation modelling the relationship between errors $\Delta\mathbf{y}$ and $\Delta\mathbf{p}$ has the form:

$$\mathbf{\Phi}^T\mathbf{\Phi}\Delta\mathbf{p} = \mathbf{\Phi}^T\Delta\mathbf{y}$$

Thus:

$$\Delta\mathbf{p} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\Delta\mathbf{y} = \begin{bmatrix} 0.4 & 0.3 & 0.2 & 0.1 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{bmatrix}\Delta\mathbf{y}$$

Hence the required assessment:

$$\left|\Delta p_0\right| \le (0.4+0.3+0.2+0.1)\cdot 0.01 = 0.010$$

$$\left|\Delta p_1\right| \le (0.3+0.1+0.1+0.3)\cdot 0.01 = 0.008$$

Problem: Compute the estimates $\hat{p}_0$ and $\hat{p}_1$ of the parameters p_0 and p_1 of the function $y = \hat{f}(x; p_0, p_1) \equiv p_0 + p_1 x$, approximating the data:

n	1	2	3	4
x_n	0	1	2	3
y_n	1.1	–0.3	–0.7	–2.1

in the sense of the criterion: $J(p_0, p_1) = \sum_{n=1}^{4}\left[y_n - \left(p_0 + p_1 x_n\right)\right]^2$. Draw the function $\hat{f}(x; \hat{p}_0, \hat{p}_1)$; indicate the points (x_n, y_n) for $n = 1, 2, 3, 4$. Assess the absolute errors of the estimates $\hat{p}_0$ and $\hat{p}_1$, implied by the absolute errors of the data y_n; assume that the magnitudes of those errors are not greater than 0.01.

Solution: The necessary condition for the minimum of $J(p_0, p_1)$ is:

$$\frac{\partial J(p_0, p_1)}{\partial p_0} = 2\sum_{n=1}^{4}\left[y_n - \left(p_0 + p_1 x_n\right)\right](-1) = 0$$
$$\frac{\partial J(p_0, p_1)}{\partial p_1} = 2\sum_{n=1}^{4}\left[y_n - \left(p_0 + p_1 x_n\right)\right](-x_n) = 0$$

or:

$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{p} = \mathbf{\Phi}^T\mathbf{y}$$

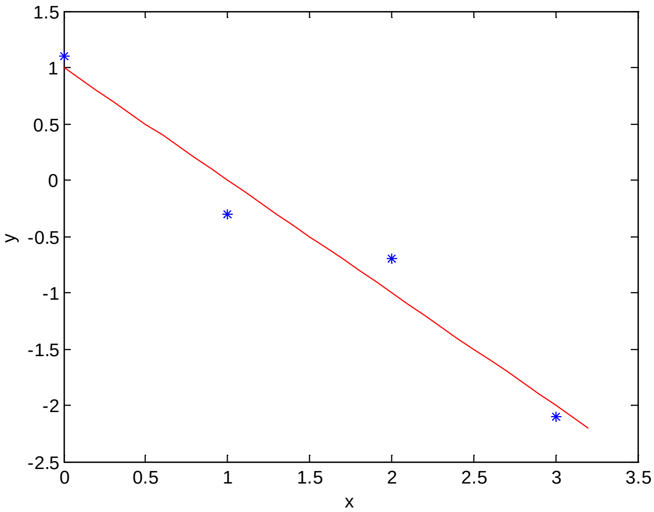
with:

$$\mathbf{\Phi} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \ \mathbf{p} = \begin{bmatrix} p_0 \\ p_1 \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1.1 \\ -0.3 \\ -0.7 \\ -2.1 \end{bmatrix}$$

Hence the equation:

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}\mathbf{p} = \begin{bmatrix} -2 \\ -8 \end{bmatrix}$$

whose solution is: $\hat{p}_0 = 1$, $\hat{p}_1 = -1$.




```
clear all
x=[0:0.1:3.2]; y=1-x;
xw=[0 1 2 3];yw=[1.1 -0.3 -0.7 -2.1];
plot(x,y,'r');hold on;
xlabel('x');ylabel('y');
plot(xw,yw,'*');
```

The equation modelling the relationship between errors $\Delta\mathbf{y}$ and $\Delta\mathbf{p}$ has the form:

$$\mathbf{\Phi}^T\mathbf{\Phi}\Delta\mathbf{p} = \mathbf{\Phi}^T\Delta\mathbf{y}$$

Thus:

$$\Delta\mathbf{p} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\Delta\mathbf{y} = \begin{bmatrix} 0.7 & 0.4 & 0.1 & -0.2 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{bmatrix}\Delta\mathbf{y}$$

Hence the required assessment:

$$\left|\Delta p_0\right| \le (0.7+0.4+0.1+0.2)\cdot 0.01 = 0.014$$
$$\left|\Delta p_1\right| \le (0.3+0.1+0.1+0.3)\cdot 0.01 = 0.008$$

Problem: Compute the estimates $\hat{p}_0$ and $\hat{p}_1$ of the parameters p_0 and p_1 of the function $y = \hat{f}(x;\, p_0,\, p_1) \equiv p_0 + p_1 x^2$, approximating the data:

n	1	2	3
x_n	1	$\sqrt{2}$	$\sqrt{3}$
y_n	1	5	3

in the sense of the criterion: $J(p_0,\, p_1) = \sum_{n=1}^{3} \left[y_n - \left(p_0 + p_1 x_n^2 \right) \right]^2$. Draw the function $\hat{f}(x;\, \hat{p}_0,\, \hat{p}_1)$; indicate the points $(x_n,\, y_n)$ for $n = 1, 2, 3$. Assess the absolute errors of the estimates $\hat{p}_0$ and $\hat{p}_1$, implied by the absolute errors of the data y_n; assume that the magnitudes of those errors are not greater than 0.01.

Solution: The necessary condition for the minimum of $J(p_0,\, p_1)$ is:

$$\frac{\partial J(p_0,\, p_1)}{\partial p_0} = 2 \sum_{n=1}^{3} \left[y_n - \left(p_0 + p_1 x_n^2 \right) \right] (-1) = 0$$

$$\frac{\partial J(p_0,\, p_1)}{\partial p_1} = 2 \sum_{n=1}^{3} \left[y_n - \left(p_0 + p_1 x_n^2 \right) \right] \left(-x_n^2 \right) = 0$$

or:

$$\mathbf{\Phi}^T \mathbf{\Phi} \mathbf{p} = \mathbf{\Phi}^T \mathbf{y}$$

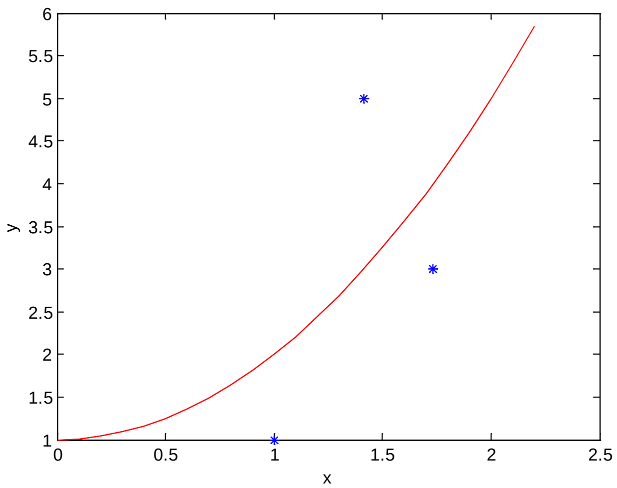
with:

$$\mathbf{\Phi} = \begin{bmatrix} 1 & x_1^2 \\ 1 & x_2^2 \\ 1 & x_3^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix},\ \mathbf{p} = \begin{bmatrix} p_0 \\ p_1 \end{bmatrix},\ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 3 \end{bmatrix}$$

Hence the equation:

$$\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \mathbf{p} = \begin{bmatrix} 9 \\ 20 \end{bmatrix}$$

whose solution is: $\hat{p}_0 = \hat{p}_1 = 1$.




```
x=[0:0.1:2.2]; y=1+x.*x;
xw=[1 sqrt(2) sqrt(3)];yw=[1 5 3];
plot(x,y,'r');hold on;
xlabel('x');ylabel('y');
plot(xw,yw,'*');
```

The equation modelling the relationship between errors $\Delta\mathbf{y}$ and $\Delta\mathbf{p}$ has the form:

$$\mathbf{\Phi}^T \mathbf{\Phi} \Delta\mathbf{p} = \mathbf{\Phi}^T \Delta\mathbf{y}$$

Thus:

$$\Delta\mathbf{p} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\Delta\mathbf{y} = \begin{bmatrix} \frac{4}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}\Delta\mathbf{y}$$

Hence the required assessment:

$$\left|\Delta p_0\right| \le \left(\frac{4}{3}+\frac{1}{3}+\frac{2}{3}\right)\cdot 0.01 = 0.0233...$$

$$\left|\Delta p_1\right| \le \left(\frac{1}{2}+\frac{1}{2}\right)\cdot 0.01 = 0.01$$

Problem: Compute the estimates $\hat{p}_0$ and $\hat{p}_1$ of the parameters p_0 and p_1 of the function $y = \hat{f}\left(x;\, p_0,\, p_1\right) \equiv p_0 \sin\left(x\right) + p_1 \cos\left(x\right)$, approximating the data:

n	1	2	3	4	5
x_n	$-\frac{\pi}{2}$	$-\frac{\pi}{4}$	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$
y_n	-1	-1	0	1	1

in the sense of the criterion: $J\left(p_0,\, p_1\right) = \sum_{n=1}^{5}\left[y_n - \left(p_0 \sin\left(x_n\right) + p_1 \cos\left(x_n\right)\right)\right]^2$. Draw the function $\hat{f}\left(x;\, \hat{p}_0,\, \hat{p}_1\right)$; indicate the points $\left(x_n,\, y_n\right)$ for $n = 1, 2, 3, 4, 5$.

<u>Solution:</u> The necessary condition for the minimum of $J\left(p_0,\, p_1\right)$ is:

$$\frac{\partial J\left(p_0,\, p_1\right)}{\partial p_0} = 2\sum_{n=1}^{5}\left[y_n - \left(p_0 \sin\left(x_n\right) + p_1 \cos\left(x_n\right)\right)\right]\left(-\sin\left(x_n\right)\right) = 0$$

$$\frac{\partial J\left(p_0,\, p_1\right)}{\partial p_1} = 2\sum_{n=1}^{5}\left[y_n - \left(p_0 \sin\left(x_n\right) + p_1 \cos\left(x_n\right)\right)\right]\left(-\cos\left(x_n\right)\right) = 0$$

or:

$$\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{p} = \mathbf{\Phi}^T\mathbf{y}$$

with:

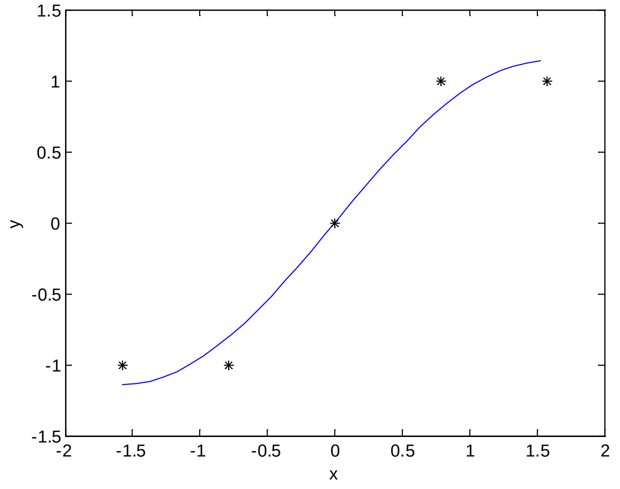
$$\mathbf{\Phi} = \begin{bmatrix} \sin\left(x_1\right) & \cos\left(x_1\right) \\ \sin\left(x_2\right) & \cos\left(x_2\right) \\ \sin\left(x_3\right) & \cos\left(x_3\right) \\ \sin\left(x_4\right) & \cos\left(x_4\right) \\ \sin\left(x_5\right) & \cos\left(x_5\right) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 1 & 0 \end{bmatrix},\ \mathbf{p} = \begin{bmatrix} p_0 \\ p_1 \end{bmatrix},\ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Hence the equation:

$$\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}\mathbf{p} = \begin{bmatrix} 2+\sqrt{2} \\ 0 \end{bmatrix}$$

whose solution is:

$$\hat{\mathbf{p}} = \begin{bmatrix} \frac{2+\sqrt{2}}{3} \\ 0 \end{bmatrix}$$




```
clear all
xw=[-pi/2 -pi/4 0 pi/4 pi/2]';
yw=[-1 -1 0 1 1]';
x=[-pi/2:0.1:pi/2];
y=((2+sqrt(2))/3)*sin(x);
plot(x,y,'b');hold on;
xlabel('x');ylabel('y');
plot(xw,yw,'k*');
hold off
```

Problem: Compute the estimates $\hat{p}_0$ and $\hat{p}_1$ of the parameters p_0 and p_1 of the function $y = \hat{f}(x; p_1, p_2) \equiv p_1 x + p_2 2^x$, approximating the data:

n	1	2	3
x_n	0	1	2
y_n	1	2	6.5

in the sense of the criterion: $J(p_1, p_2) = \sum_{n=1}^{3} \left[y_n - \left(p_1 x_n + p_2 2^{x_n} \right) \right]^2$. Draw the function $\hat{f}(x; \hat{p}_0, \hat{p}_1)$; indicate the points (x_n, y_n) for $n = 1, 2, 3$. Assess the absolute errors of the estimates $\hat{p}_0$ and $\hat{p}_1$, implied by the absolute errors of the data y_n; assume that the magnitudes of those errors are not greater than 0.01.

<u>Solution:</u> The necessary condition for the minimum of $J(p_0, p_1)$ is:

$$\frac{\partial J(p_1, p_2)}{\partial p_1} = 2\sum_{n=1}^{3} \left[y_n - \left(p_1 x_n + p_2 2^{x_n} \right) \right] (-x_n) = 0$$

$$\frac{\partial J(p_1, p_2)}{\partial p_2} = 2\sum_{n=1}^{3} \left[y_n - \left(p_1 x_n + p_2 2^{x_n} \right) \right] (-2^{x_n}) = 0$$

or:

$$\mathbf{\Phi}^T \mathbf{\Phi} \mathbf{p} = \mathbf{\Phi}^T \mathbf{y}$$

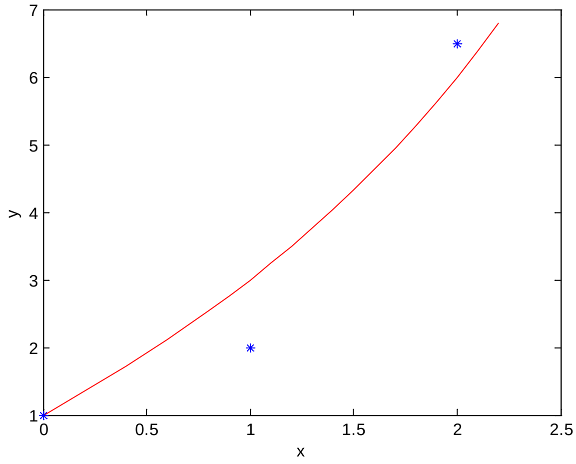
with:

$$\mathbf{\Phi} = \begin{bmatrix} x_1 & 2^{x_1} \\ x_2 & 2^{x_2} \\ x_3 & 2^{x_3} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 2 \\ 2 & 4 \end{bmatrix}, \ \mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 6.5 \end{bmatrix}$$

Hence the equation:

$$\begin{bmatrix} 5 & 10 \\ 10 & 21 \end{bmatrix} \mathbf{p} = \begin{bmatrix} 15 \\ 31 \end{bmatrix}$$

whose solution is: $\hat{p}_1 = \hat{p}_2 = 1$.




```
clear all
x=[0:0.1:2.2]; y=x+2.^x;
xw=[0 1 2];yw=[1 2 6.5];
plot(x,y,'r');hold on;
xlabel('x');ylabel('y');
plot(xw,yw,'*');
```

The equation modelling the relationship between errors $\Delta\mathbf{y}$ and $\Delta\mathbf{p}$ has the form:

$$\mathbf{\Phi}^T\mathbf{\Phi}\Delta\mathbf{p} = \mathbf{\Phi}^T\Delta\mathbf{y}$$

Thus:

$$\Delta\mathbf{p} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\Delta\mathbf{y} = \begin{bmatrix} -2.0 & 0.2 & 0.4 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}\Delta\mathbf{y}$$

Hence the required assessment:

$$|\Delta p_1| \le (2.0+0.2+0.4)\cdot 0.01 = 0.026$$

$$|\Delta p_2| \le (1.0+0.0+0.0)\cdot 0.01 = 0.01$$

Problem: Assess the bias of the estimate $\hat{p}_1$ of the parameter p_1 of the static characteristics of a sensor:

$$y = p_0 + p_1 x$$

obtained by means of the least-squares method using the following data:

$$\left\{\tilde{x}_n^{cal}, y_n^{cal} \mid n = 1,.., N\right\}$$

where the data $\tilde{x}_n^{cal}$ are corrupted with random errors which may be adequately modelled with statistically independent random variables following the distribution $\mathcal{N}\left(0;\sigma_x^2\right)$. Assume that:

$$S_x = \frac{1}{N}\sum_{n=1}^{N} x_n^{cal} = 1 \qquad S_y = \frac{1}{N}\sum_{n=1}^{N} y_n^{cal} = 1$$

$$S_{xy} = \frac{1}{N}\sum_{n=1}^{N} x_n^{cal} y_n^{cal} = 2 \quad S_{xx} = \frac{1}{N}\sum_{n=1}^{N}\left(x_n^{cal}\right)^2 = 2$$

<u>Solution:</u> The solution of the system of normal equations:

$$\begin{cases} \hat{p}_0 + \tilde{S}_x\hat{p}_1 = S_y \\ \tilde{S}_x\hat{p}_0 + \tilde{S}_{xx}\hat{p}_1 = \tilde{S}_{xy} \end{cases}$$

with respect to $\hat{p}_1$ has the form:

$$\hat{p}_1 = \frac{\tilde{S}_{xy} - \tilde{S}_x S_y}{\tilde{S}_{xx} - \tilde{S}_x^2}$$

Since:

$$\underline{\hat{p}}_1 = p_1 + \sum_n \frac{\partial p_1}{\partial x_n^{cal}} \underline{\Delta\tilde{x}}_n^{cal} + \frac{1}{2}\sum_n \frac{\partial^2 p_1}{\partial\left(x_n^{cal}\right)^2}\left(\underline{\Delta\tilde{x}}_n^{cal}\right)^2 + \ldots$$

where $\underline{\Delta\tilde{x}}_n^{cal} \equiv \underline{\tilde{x}}_n^{cal} - x_n^{cal}$, $\mathrm{E}\left[\underline{\Delta\tilde{x}}_n^{cal}\right] = 0$ and $\mathrm{E}\left[\left(\underline{\Delta\tilde{x}}_n^{cal}\right)^2\right] = \sigma_x^2 \neq 0$, the second derivatives $\frac{\partial^2 p_1}{\partial\left(x_n^{cal}\right)^2}$ are necessary for the assessment of the bias. They may be determined in the following way:

$$p_1 = \frac{S_{xy} - S_x S_y}{S_{xx} - S_x^2} = 1, \quad S_x = S_y = 1, \quad S_{xy} = S_{xx} = 2 \;\Rightarrow\; p_1 = \frac{S_{xy} - S_x}{S_{xx} - S_x^2}$$

$$\frac{\partial p_1}{\partial x_n^{cal}} = \frac{\left(\frac{1}{N} y_n^{cal} - \frac{1}{N}\right)\left(S_{xx} - S_x^2\right) - \left(S_{xy} - S_x\right)\left(\frac{2}{N} x_n^{cal} - \frac{2}{N} S_x\right)}{\left(S_{xx} - S_x^2\right)^2}$$

$$N\frac{\partial p_1}{\partial x_n^{cal}} = \frac{y_n^{cal} - 1}{S_{xx} - S_x^2} - 2\frac{\left(S_{xy} - S_x\right)\left(x_n^{cal} - S_x\right)}{\left(S_{xx} - S_x^2\right)^2}$$

$$N\frac{\partial^2 p_1}{\partial\left(x_n^{cal}\right)^2} = -\frac{y_n^{cal} - 1}{\left(S_{xx} - S_x^2\right)^2}\left(\frac{2}{N} x_n^{cal} - \frac{2}{N} S_x\right) +$$

$$-2\frac{\left[\left(\frac{1}{N} y_n^{cal} - \frac{1}{N}\right)\left(x_n^{cal} - S_x\right) + \left(S_{xy} - S_x\right)\left(1 - \frac{1}{N}\right)\right]\left(S_{xx} - S_x^2\right)^2}{\left(S_{xx} - S_x^2\right)^4} +$$

$$+2\frac{\left(S_{xy} - S_x\right)\left(x_n^{cal} - S_x\right)\cdot 2\left(S_{xx} - S_x^2\right)\left(\frac{2}{N} x_n^{cal} - \frac{2}{N} S_x\right)}{\left(S_{xx} - S_x^2\right)^4} =$$

$$= -\frac{2}{N}\left(y_n^{cal} - 1\right)\left(x_n^{cal} - 1\right) - \frac{2}{N}\left[\left(y_n^{cal} - 1\right)\left(x_n^{cal} - 1\right) + \left(N - 1\right)\right] + \frac{8}{N}\left(x_n^{cal} - 1\right)^2$$

$$-\frac{N^2}{2}\frac{\partial^2 p_1}{\partial\left(x_n^{cal}\right)^2} = 2x_n^{cal} y_n^{cal} - 2x_n^{cal} - 2y_n^{cal} + 2 + N - 1 - 4\left(x_n^{cal}\right)^2 + 8x_n^{cal} - 4 =$$

$$= 2x_n^{cal} y_n^{cal} + 6x_n^{cal} - 2y_n^{cal} - 4\left(x_n^{cal}\right)^2 + N - 3$$

Hence the bias:

$$\mathrm{E}\left[\underline{\hat{p}}_1\right] - p_1 = \frac{1}{2}\sum_n \frac{\partial^2 p_1}{\partial\left(x_n^{cal}\right)^2}\mathrm{E}\left[\left(\underline{\Delta\tilde{x}}_n^{cal}\right)^2\right] =$$

$$= -\frac{1}{N}\left[\frac{1}{N}\sum_n\left(2x_n^{cal} y_n^{cal} + 6x_n^{cal} - 2y_n^{cal} - 4\left(x_n^{cal}\right)^2 + N - 3\right)\right]\sigma_x^2 =$$

$$= -\frac{1}{N}\left[2S_{xy} + 6S_x - 2S_y - 4S_{xx} + N - 3\right]\sigma_x^2 =$$

$$= -\frac{1}{N}\left[4 + 6 - 2 - 8 + N - 3\right]\sigma_x^2 = -\frac{1}{N}\left(N - 3\right)\sigma_x^2 = -\frac{N - 3}{N}\sigma_x^2$$

4.3. Padé approximation

Problem: Determine the parameters a_0, a_1, a_2, b_1 and b_2 of the rational function:

$$\hat{f}(x) = \frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x + b_2 x^2}$$

approximating the function $f(x) = e^x$ in the Padé sense for $|x| \le 1$.

<u>Solution:</u> The MacLaurin expansion of $f(x) = e^x$ has the form:

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \ldots$$

The Padé approximation is based on the equality:

$$a_0 + a_1 x + a_2 x^2 = \left(1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 + \ldots\right) \cdot \left(1 + b_1 x + b_2 x^2\right)$$

generating equations with respect to the parameters a_0, a_1, a_2, b_1 and b_2. The right-hand side of this equation has the form:

$$\begin{aligned}
\mathrm{RHS} = 1 \; & + b_1 x + b_2 x^2 \\
& + x + b_1 x^2 + b_2 x^3 \\
& + \frac{1}{2}x^2 + \frac{1}{2}b_1 x^3 + \frac{1}{2}b_2 x^4 \\
& + \frac{1}{6}x^3 + \frac{1}{6}b_1 x^4 + \frac{1}{6}b_2 x^5 \\
& + \frac{1}{24}x^4 + \frac{1}{24}b_1 x^5 + \frac{1}{24}b_2 x^6 + \ldots
\end{aligned}$$

Thus:

$$a_0 = 1$$

$$a_1 = b_1 + 1$$

$$a_2 = b_2 + b_1 + \frac{1}{2}$$

$$0 = b_2 + \frac{1}{2}b_1 + \frac{1}{6}$$

$$0 = \frac{1}{2}b_2 + \frac{1}{6}b_1 + \frac{1}{24}$$

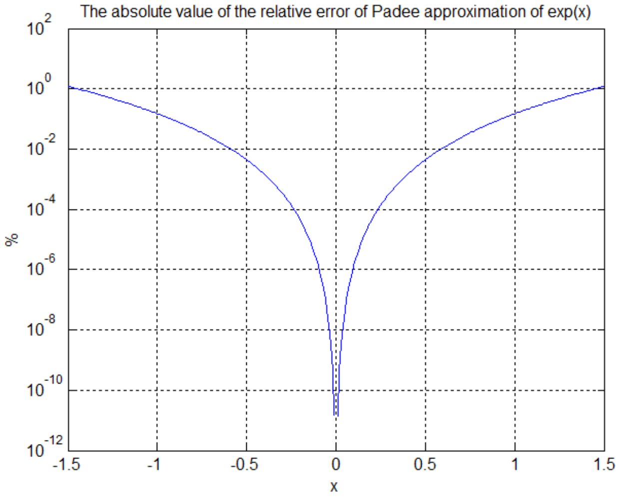
Hence:

$$a_1 = \frac{1}{2},\; a_2 = \frac{1}{12},\; b_1 = -\frac{1}{2} \text{ and } b_2 = \frac{1}{12}$$

which means that the function $\hat{f}(x)$ has the form:

$$\hat{f}(x) = \frac{1 + \frac{1}{2}x + \frac{1}{12}x^2}{1 - \frac{1}{2}x + \frac{1}{12}x^2}$$

The relative error of approximation is shown in the figure below.




```
clear all
x=[-1.5:0.01:1.5];
y=exp(x);
ya=(1+x/2+x.*x/12)./(1-x/2+x.*x/12);
semilogy(x,100*abs(ya-y)./y)
title('The absolute value of the relative
error of Pade approximation of exp(x)');
xlabel('x')
ylabel('%')
grid on
```

Problem: Determine the parameters a_0, a_1, a_2, b_1 and b_2 of the rational function:

$$\hat{f}(x) = \frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x + b_2 x^2}$$

approximating the function $f(x) = \ln(1+x)$ in the Padé sense for $|x| \le 1$.

<u>Solution:</u> The MacLaurin expansion of $f(x) = \ln(1+x)$ has the form:

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \ldots$$

The Padé approximation is based on the equality:

$$a_0 + a_1 x + a_2 x^2 = \left(x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \ldots \right) \cdot \left(1 + b_1 x + b_2 x^2 \right)$$

generating equations with respect to the parameters a_0, a_1, a_2, b_1 and b_2. The right-hand side of this equation has the form:

$$\begin{array}{llllll}
\text{RHS} = x & +b_1 x^2 & +b_2 x^3 & & & \\
& -\frac{1}{2}x^2 & -\frac{1}{2}b_1 x^3 & -\frac{1}{2}b_2 x^4 & & \\
& & +\frac{1}{3}x^3 & +\frac{1}{3}b_1 x^4 & +\frac{1}{3}b_2 x^5 & \\
& & & -\frac{1}{4}x^4 & -\frac{1}{4}b_1 x^5 & -\frac{1}{4}b_2 x^6 - \ldots
\end{array}$$

Thus:

$$a_0 = 0$$

$$a_1 = 1$$

$$a_2 = b_1 - \frac{1}{2}$$

$$0 = b_2 - \frac{1}{2}b_1 + \frac{1}{3}$$

$$0 = -\frac{1}{2}b_2 + \frac{1}{3}b_1 - \frac{1}{4}$$

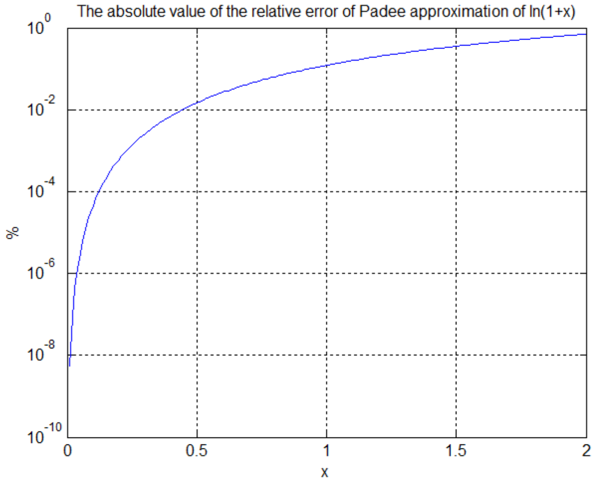
Hence:

$$a_1 = 1,\ a_2 = \frac{1}{2},\ b_1 = 1 \text{ and } b_2 = \frac{1}{6}$$

which means that the function $\hat{f}(x)$ has the form:

$$\hat{f}(x) = \frac{x + \frac{1}{2}x^2}{1 + x + \frac{1}{6}x^2}$$

The relative error of approximation is shown in the figure below.



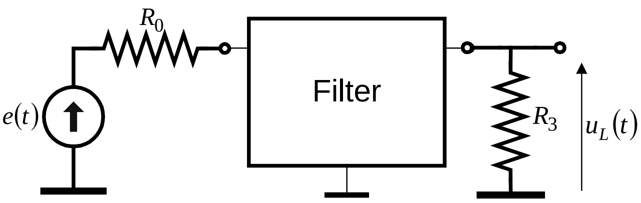

```
clear all
x=[-1:0.01:2];
y=log(1+x);
ya=(x+x.*x/2)./(1+x+x.*x/6);
semilogy(x,100*abs(ya-y)./y)
title('The absolute value of the relative
error of Pade approximation of ln(1+x)');
xlabel('x')
ylabel('%')
grid on
```

ENUME: SOLVED PROBLEMS

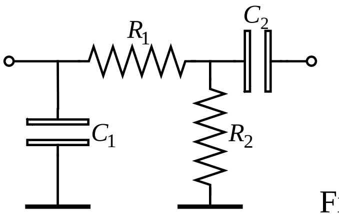
5. SOLVING ORDINARY DIFFERENTIAL EQUATIONS

5.1. Formulation of ODE systems

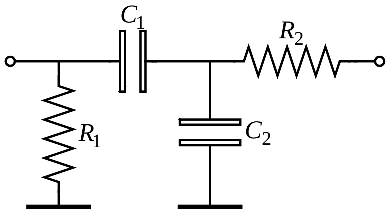
Problem: Formulate the system of ordinary differential equations (ODE) modelling the following circuit:



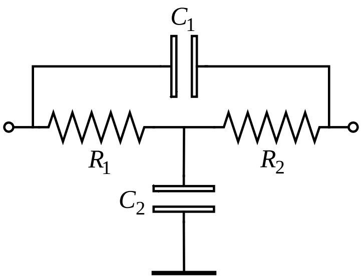
for each of the filters provided below:



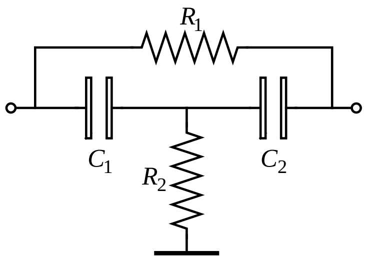
Filter #1



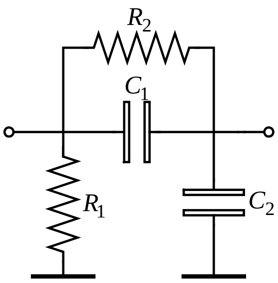
Filter #2



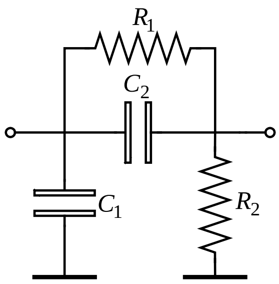
Filter #3



Filter #4



Filter #5



Filter #16

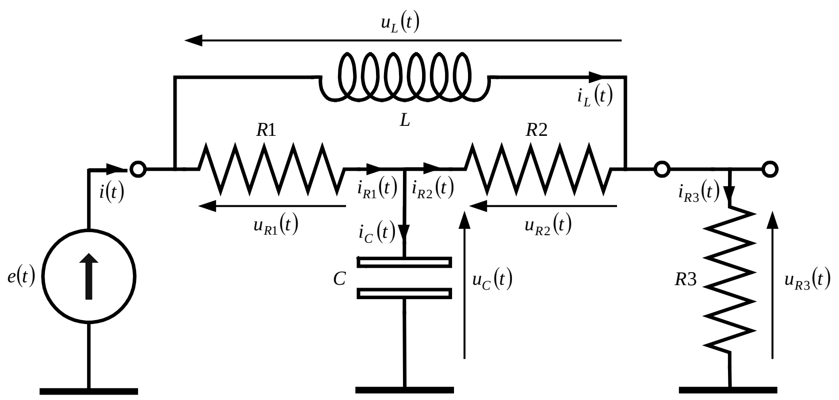
$$e(t) = \begin{cases} 0\,V & \text{for } t < 0 \\ 1\,V & \text{for } t > 0 \end{cases}$$

$R_0 = R_1 = R_2 = R_3 = 1\,k\Omega$

$C_1 = C_2 = C_3 = 1\,\mu F$

$L_3 = 1\,H$

Problem: Formulate the state equations (*i.e.* two ordinary differential equations with respect to $u_C(t)$ and $i_L(t)$) for the following circuit with $R1 = R2 = R3 = 1\,k\Omega$, $C = 1\,\mu F$ and $L = 1\,H$.



Solution:

$$\begin{aligned}
&e - u_{R1} - u_C = 0\\
&u_C - u_{R2} - u_{R3} = 0\\
&u_L - u_{R1} - u_{R2} = 0\\
&i - i_L - i_{R1} = 0\\
&i_{R1} - i_C - i_{R2} = 0\\
&i_{R2} + i_L - i_{R3} = 0\\
&u_{R1} = R1 \cdot i_{R1}\\
&u_{R2} = R2 \cdot i_{R2}\\
&u_{R3} = R3 \cdot i_{R3}\\
&i_C = C \cdot u_C'\\
&u_L = L \cdot i_L'
\end{aligned}
\quad \Rightarrow \quad
\begin{aligned}
&e - R1 \cdot i_{R1} - u_C = 0 \Rightarrow i_{R1} = \frac{e - u_C}{R1}\\
&u_C - R2 \cdot i_{R2} - R3 \cdot i_{R3} = 0\\
&L \cdot i_L' - R1 \cdot i_{R1} - R2 \cdot i_{R2} = 0\\
&i_{R1} - C \cdot u_C' - i_{R2} = 0\\
&i_{R2} + i_L - i_{R3} = 0
\end{aligned}
\quad \Rightarrow$$

$$\begin{aligned}
&u_C - R2 \cdot i_{R2} - R3 \cdot i_{R3} = 0\\
&L \cdot i_L' - e + u_C - R2 \cdot i_{R2} = 0 \Rightarrow i_{R2} = \frac{L \cdot i_L' - e + u_C}{R2}\\
&\frac{e - u_C}{R1} - C \cdot u_C' - i_{R2} = 0\\
&i_{R2} + i_L - i_{R3} = 0
\end{aligned}
\quad \Rightarrow$$

$$\begin{aligned}
&e - L \cdot i_L' - R3 \cdot i_{R3} = 0 \Rightarrow i_{R3} = \frac{e - L \cdot i_L'}{R3}\\
&\frac{e - u_C}{R1} - C \cdot u_C' - \frac{L \cdot i_L' - e + u_C}{R2} = 0\\
&\frac{L \cdot i_L' - e + u_C}{R2} + i_L - i_{R3} = 0
\end{aligned}
\quad \Rightarrow$$

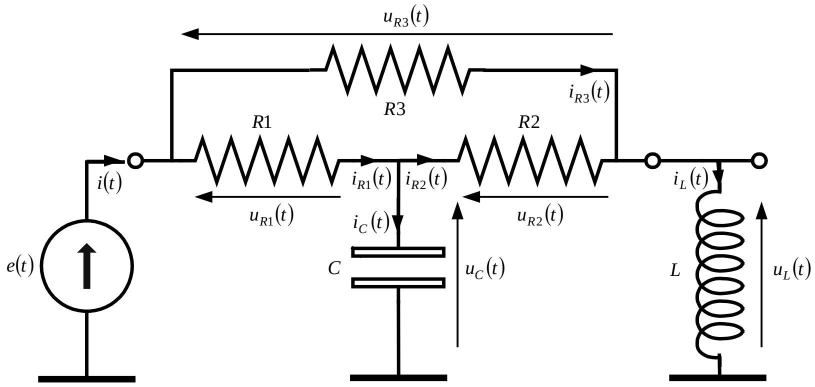
$$\begin{aligned}
&\frac{e - u_C}{R1} - C \cdot u_C' - \frac{L \cdot i_L' - e + u_C}{R2} = 0\\
&\frac{L \cdot i_L' - e + u_C}{R2} + i_L - \frac{e - L \cdot i_L'}{R3} = 0 \Rightarrow \; i_L' = -\frac{R3}{L \cdot (R2 + R3)} \cdot u_C - \frac{R2 \cdot R3}{L \cdot (R2 + R3)} \cdot i_L + \frac{1}{L} \cdot e
\end{aligned}
\quad \Rightarrow$$

$$u_C' = -\frac{R1 + R2 + R3}{C \cdot R1 \cdot (R2 + R3)} \cdot u_C + \frac{R3}{C \cdot (R2 + R3)} \cdot i_L - \frac{1}{C \cdot R1} \cdot e$$

After substitution of $R1 = R2 = R3 = 1\,k\Omega$, $C = 1\,\mu F$ and $L = 1\,H$:

$$u_C'(t) = -\frac{3}{2}\cdot u_C(t) + \frac{1}{2}\cdot i_L(t) - e(t) \text{ and } i_L'(t) = -\frac{1}{2}\cdot u_C(t) - \frac{1}{2}\cdot i_L(t) + e(t)$$

Problem: Formulate the state equations (*i.e.* two ordinary differential equations with respect to $u_C(t)$ and $i_L(t)$) for the following circuit with $R1 = R2 = R3 = 1\,k\Omega$, $C = 1\,\mu F$ and $L = 1\,H$.



Solution:

$$\begin{aligned}
&e - u_{R1} - u_C = 0\\
&u_C - u_{R2} - u_L = 0\\
&u_{R3} - u_{R1} - u_{R2} = 0\\
&i - i_{R3} - i_{R1} = 0\\
&i_{R1} - i_C - i_{R2} = 0\\
&i_{R2} - i_L + i_{R3} = 0\\
&u_{R1} = R1\cdot i_{R1}\\
&u_{R2} = R2\cdot i_{R2}\\
&u_{R3} = R3\cdot i_{R3}\\
&i_C = C\cdot u_C'\\
&u_L = L\cdot i_L'
\end{aligned}
\quad\Rightarrow\quad
\begin{aligned}
&e - R1\cdot i_{R1} - u_C = 0 \Rightarrow i_{R1} = \frac{e - u_C}{R1}\\
&u_C - R2\cdot i_{R2} - L\cdot i_L' = 0\\
&R3\cdot i_{R3} - R1\cdot i_{R1} - R2\cdot i_{R2} = 0\\
&i_{R1} - C\cdot u_C' - i_{R2} = 0\\
&i_{R2} - i_L + i_{R3} = 0
\end{aligned}
\quad\Rightarrow$$

$$\begin{aligned}
&u_C - R2\cdot i_{R2} - L\cdot i_L' = 0\\
&R3\cdot i_{R3} - e + u_C - R2\cdot i_{R2} = 0 \Rightarrow i_{R2} = \frac{R3\cdot i_{R3} - e + u_C}{R2}\\
&\frac{e - u_C}{R1} - C\cdot u_C' - i_{R2} = 0\\
&i_{R2} - i_L + i_{R3} = 0
\end{aligned}
\quad\Rightarrow$$

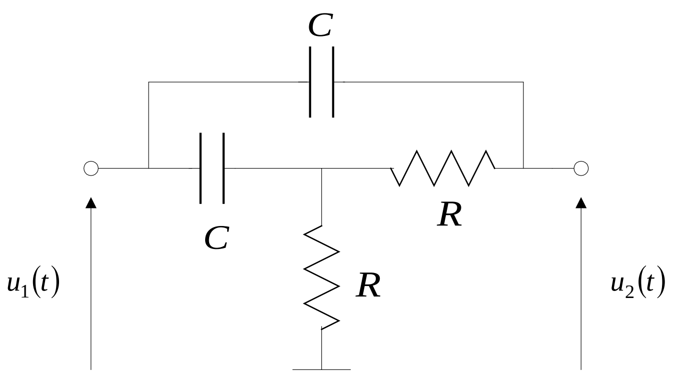
$$\begin{aligned}
&e - L\cdot i_L' - R3\cdot i_{R3} = 0 \Rightarrow i_{R3} = \frac{e - L\cdot i_L'}{R3}\\
&\frac{e - u_C}{R1} - C\cdot u_C' - \frac{R3\cdot i_{R3} - e + u_C}{R2} = 0
\end{aligned}
\quad\Rightarrow$$

$$\frac{R3\cdot i_{R3}-e+u_C}{R2}-i_L+i_{R3}=0$$
$$\frac{e-u_C}{R1}-C\cdot u_C'-\frac{u_C-L\cdot i_L'}{R2}=0$$
$$\frac{u_C-L\cdot i_L'}{R2}-i_L+\frac{e-L\cdot i_L'}{R3}=0\Rightarrow \qquad\Rightarrow$$
$$i_L'=\frac{R3}{L\cdot(R2+R3)}\cdot u_C-\frac{R2\cdot R3}{L\cdot(R2+R3)}\cdot i_L+\frac{R2}{L\cdot(R2+R3)}\cdot e$$

$$u_C'=-\frac{R1+R2+R3}{C\cdot R1\cdot(R2+R3)}\cdot u_C-\frac{R3}{C\cdot(R2+R3)}\cdot i_L+\frac{R1+R2+R3}{C\cdot R1\cdot(R2+R3)}\cdot e$$

After substitution of $R1=R2=R3=1\,k\Omega$, $C=1\,\mu F$ and $L=1\,H$:
$$u_C'(t)=-\frac{3}{2}\cdot u_C(t)-\frac{1}{2}\cdot i_L(t)+\frac{3}{2}\cdot e(t) \text{ and } i_L'(t)=\frac{1}{2}\cdot u_C(t)-\frac{1}{2}\cdot i_L(t)+\frac{1}{2}\cdot e(t)$$

Problem: Formulate the system of ordinary differential equations (ODE) modelling the following circuit for $R=1\,\text{k}\Omega$ and $C=1\,\mu\text{F}$:



Express the output voltage $u_2(t)$ as a linear combination of the voltages on the capacitors and of the input voltage $u_1(t)$. Solve the system of ODEs by means of the explicit Euler method for $u_1(t)=\mathbf{1}(t)$. Assess the range of step values (h) which are guaranteeing the stability of the numerical solution.

5.2. Analytical solution of ODE systems

Problem: Prove that the solution of the following initial-value problem (IVP):
$$y_1'(t)=a_{11}\cdot y_1(t)+a_{12}\cdot y_2(t)+b_1\cdot x_1(t);\; y_1(0)=0$$
$$y_2'(t)=a_{21}\cdot y_1(t)+a_{22}\cdot y_2(t)+b_2\cdot x_2(t);\; y_2(0)=0$$
may be expressed in the form:
$$y_1(t)=g_{11}(t)*x_1(t)+g_{12}(t)*x_2(t)$$
$$y_2(t)=g_{21}(t)*x_1(t)+g_{22}(t)*x_2(t)$$
Express $g_{11}(t)$, $g_{12}(t)$, $g_{21}(t)$ and $g_{22}(t)$ on terms of a_{11}, a_{12}, a_{21} and a_{22}.

Solution: In the domain of Laplace transforms, the IVP takes on the form of a system of algebraic equations:

$$sY_1(s) = a_{11} \cdot Y_1(s) + a_{12} \cdot Y_2(s) + b_1 \cdot X_1(s)$$

$$sY_2(s) = a_{21} \cdot Y_1(s) + a_{22} \cdot Y_2(s) + b_2 \cdot X_2(s)$$

whose solution is:

$$Y_1(s) = \frac{b_1 \cdot (s - a_{22})}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} X_1(s) + \frac{b_2 \cdot a_{12}}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} X_2(s)$$

$$Y_1(s) = \frac{b_1 \cdot a_{21}}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} X_1(s) + \frac{b_2 \cdot (s - a_{11})}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} X_2(s)$$

Its original in the time domain has the target form with:

$$g_{11}(t) = \mathscr{L}^{-1}\left\{ \frac{b_1 \cdot (s - a_{22})}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} \right\}$$

$$g_{12}(t) = \mathscr{L}^{-1}\left\{ \frac{b_2 \cdot a_{12}}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} \right\}$$

$$g_{21}(t) = \mathscr{L}^{-1}\left\{ \frac{b_1 \cdot a_{21}}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} \right\}$$

$$g_2(t) = \mathscr{L}^{-1}\left\{ \frac{b_2 \cdot (s - a_{11})}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} \right\}$$

The inverse Laplace transforms may be determined analytically, taking into account that:

$$f(t) = \mathscr{L}^{-1}\left\{ \frac{1}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} \right\} = \frac{1}{s_1 - s_2}\left(e^{s_1 t} - e^{s_2 t}\right)$$

and

$$f'(t) = \mathscr{L}^{-1}\left\{ \frac{s}{(s - a_{11}) \cdot (s - a_{22}) - a_{12} \cdot a_{21}} \right\} = \frac{1}{s_1 - s_2}\left(s_1 e^{s_1 t} - s_2 e^{s_2 t}\right)$$

Hence:

$$g_{11}(t) = b_1 \cdot \left[f'(t) - a_{22} \cdot f(t) \right]$$

$$g_{12}(t) = b_2 \cdot a_{12} \cdot f(t)$$

$$g_{21}(t) = b_1 \cdot a_{21} \cdot f(t)$$

$$g_{22}(t) = b_2 \cdot \left[f'(t) - a_{11} \cdot f(t) \right]$$

Problem: Solve the following equation:

$$y'(t) = -y(t) + \sin(t); \quad y(0) = 0$$

using the method of variable constants.

Solution: The solution of the homogenous equation $y'(t) = -y(t)$ has the form $y(t) = C \cdot e^{-t}$. Thus, the general solution of the non-homogenous equation has the form:

$$y(t) = C(t) \cdot e^{-t} \quad \Rightarrow \quad y'(t) = C'(t) \cdot e^{-t} - C(t) \cdot e^{-t}$$

By substituting this solution and its derivative to the original equation, we obtain:

$$C'(t) \cdot e^{-t} - C(t) \cdot e^{-t} = -C(t) \cdot e^{-t} + \sin(t)$$

or:

$$C'(t)=e^t\cdot\sin(t)\ \Rightarrow\ C(t)=\int e^t\cdot\sin(t)\,dt=\frac{1}{2}e^t\left[\sin(t)-\cos(t)\right]+C_0$$

where the constant C_0 should satisfy the initial condition $y(0)=0$:

$$C(0)\cdot e^{-0}=0\Rightarrow C(0)=0\Rightarrow C_0=\frac{1}{2}$$

Thus, the solution of the original equation has the form:

$$y(t)=\left\{\frac{1}{2}e^t\left[\sin(t)-\cos(t)\right]+\frac{1}{2}\right\}e^{-t}=\frac{1}{2}e^{-t}+\frac{1}{2}\left[\sin(t)-\cos(t)\right]=\frac{1}{2}e^{-t}+\sqrt{2}\sin\left(t+\frac{\pi}{4}\right)$$

Problem: Solve the following system of ordinary differential equations:

$$y_1'(t)=-y_1(t)+y_2(t);\quad y_1(0)=0$$
$$y_2'(t)=y_2(t)+\sin(t);\quad y_2(0)=0$$

by transforming them into a set of two scalar equations and applying the method of "variable constants" to each of them.

Solution: The system of ODEs may be given the matrix form:

$$\mathbf{y}'(t)=\mathbf{A}\cdot\mathbf{y}(t)+\mathbf{b}\cdot\sin(t)\ \text{ for } t\in[0,T]$$

where $\mathbf{y}(t)=\left[y_1(t)\ y_2(t)\right]^T$, $\mathbf{A}=\begin{bmatrix}-1 & 1\\ 0 & 1\end{bmatrix}$, $\mathbf{b}=\begin{bmatrix}0\\ 1\end{bmatrix}$

The matrix $\mathbf{A}$ may be factorised: $\mathbf{A}=\mathbf{V}\cdot\mathbf{\Lambda}\cdot\mathbf{V}^{-1}$, where $\mathbf{V}=\begin{bmatrix}v_{11} & v_{12}\\ v_{21} & v_{22}\end{bmatrix}$ and $\mathbf{\Lambda}=diag\left\{\lambda_1,...,\lambda_M\right\}$ is the diagonal matrix of eigenvalues:

$$\det(\mathbf{A}-\lambda\mathbf{I})=(-1-\lambda)(1-\lambda)=\lambda^2-1=0\Rightarrow\lambda_1=1,\ \lambda_2=-1$$

The equality $\mathbf{A}=\mathbf{V}\cdot\mathbf{\Lambda}\cdot\mathbf{V}^{-1}$ is equivalent to the equality $\mathbf{A}\cdot\mathbf{V}=\mathbf{V}\cdot\mathbf{\Lambda}$ *i.e.*:

$$\begin{bmatrix}-1 & 1\\ 0 & 1\end{bmatrix}\cdot\begin{bmatrix}v_{11} & v_{12}\\ v_{21} & v_{22}\end{bmatrix}=\begin{bmatrix}v_{11} & v_{12}\\ v_{21} & v_{22}\end{bmatrix}\cdot\begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}\ \Rightarrow\ \begin{bmatrix}-v_{11}+v_{21} & -v_{12}+v_{22}\\ v_{21} & v_{22}\end{bmatrix}=\begin{bmatrix}v_{11} & -v_{12}\\ v_{21} & -v_{22}\end{bmatrix}$$

or:

$$\Rightarrow\begin{bmatrix}-v_{11}+v_{21}-v_{11} & -v_{12}+v_{22}+v_{12}\\ v_{21}-v_{21} & v_{22}+v_{22}\end{bmatrix}=\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$$

Hence:

$$v_{21}=2v_{11}\qquad v_{22}=0$$
$$0=0\qquad 2v_{22}=0$$

The "simplest" solution of that set of these algebraic equations is:

$$\mathbf{V}=\begin{bmatrix}1 & -\frac{1}{2}\\ 2 & 0\end{bmatrix}\ \text{and}\ \mathbf{V}^{-1}=\begin{bmatrix}0 & \frac{1}{2}\\ -2 & 1\end{bmatrix}$$

The sought-for set of independent scalar equations has the form:

$$\mathbf{V}^{-1}\cdot\mathbf{y}'(t)=\mathbf{\Lambda}\cdot\mathbf{V}^{-1}\cdot\mathbf{y}(t)+\mathbf{V}^{-1}\cdot\mathbf{b}\cdot\sin(t)$$

$$z_1'(t)=z_1(t)+\frac{1}{2}\sin(t);\quad z_1(0)=0$$

$$z_2'(t)=-z_2(t)+\sin(t);\quad z_2(0)=0$$

The general solution of the first equation, $z_1'(t) = z_1(t) + \frac{1}{2}\sin(t)$, has the form:

$$z_1(t) = C(t)\cdot e^t \quad \Rightarrow \quad z_1'(t) = C'(t)\cdot e^t + C(t)\cdot e^t$$

By substituting this solution and its derivative to the original equation, we obtain:

$$C'(t)\cdot e^t + C(t)\cdot e^t = C(t)\cdot e^t + \frac{1}{2}\sin(t)$$

or:

$$C'(t) = \frac{1}{2}e^{-t}\cdot\sin(t) \quad \Rightarrow \quad C(t) = \frac{1}{2}\int e^{-t}\cdot\sin(t)\,dt = -\frac{1}{4}e^{-t}\left[\sin(t)+\cos(t)\right] + C_0$$

where the constant C_0 should satisfy the initial condition $z_1(0) = 0$:

$$C(0)\cdot e^0 = 0 \Rightarrow C(0) = 0 \Rightarrow -\frac{1}{4}e^0\left[\sin(0)+\cos(0)\right] + C_0 = 0 \Rightarrow C_0 = \frac{1}{4}$$

Thus, the solution of the second equation has the form:

$$z_1(t) = \left\{-\frac{1}{4}e^{-t}\left[\sin(t)+\cos(t)\right] + \frac{1}{4}\right\}e^t = -\frac{1}{4}\left[\sin(t)+\cos(t)\right] + \frac{1}{4}e^t$$

The general solution of the second equation, $z_2'(t) = -z_2(t) + \sin(t)$, has the form:

$$z_2(t) = C(t)\cdot e^{-t} \quad \Rightarrow \quad z_2'(t) = C'(t)\cdot e^{-t} - C(t)\cdot e^{-t}$$

By substituting this solution and its derivative to the original equation, we obtain:

$$C'(t)\cdot e^{-t} - C(t)\cdot e^{-t} = -C(t)\cdot e^{-t} + \sin(t)$$

or:

$$C'(t) = e^t\cdot\sin(t) \quad \Rightarrow \quad C(t) = \int e^t\cdot\sin(t)\,dt = \frac{1}{2}e^t\left[\sin(t)-\cos(t)\right] + C_0$$

where the constant C_0 should satisfy the initial condition $z_2(0) = 0$:

$$C(0)\cdot e^{-0} = 0 \Rightarrow C(0) = 0 \Rightarrow C_0 = \frac{1}{2}$$

Thus, the solution of the second equation has the form:

$$z_2(t) = \left\{\frac{1}{2}e^t\left[\sin(t)-\cos(t)\right] + \frac{1}{2}\right\}e^{-t} = \frac{1}{2}e^{-t} + \frac{1}{2}\left[\sin(t)-\cos(t)\right]$$

Thus, the solutions of the original system are as follows:

$$\mathbf{y}(t) = \mathbf{V}\cdot\mathbf{z}(t) = \begin{bmatrix} 1 & -\frac{1}{2} \\ 2 & 0 \end{bmatrix}\cdot\begin{bmatrix} -\frac{1}{4}\left[\sin(t)+\cos(t)\right] - \frac{1}{4}e^t \\ \frac{1}{2}e^{-t} + \frac{1}{2}\left[\sin(t)-\cos(t)\right] \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\sin(t) - \frac{1}{4}e^t + \frac{1}{4}e^{-t} \\ -\frac{1}{2}\left[\sin(t)+\cos(t)\right] - \frac{1}{2}e^t \end{bmatrix}$$

Problem: Solve the following system of ordinary differential equations:

$$y_1'(t) = -\frac{3}{2}y_1(t) - \frac{1}{2}y_2(t); \quad y_1(0) = 0$$

$$y_2'(t) = -\frac{1}{2}y_1(t) - \frac{3}{2}y_2(t) + \mathbb{1}(t); \quad y_2(0) = 0$$

by transforming them into a set of two scalar equations and applying the method of "variable constants" to each of them.

<u>Solution:</u> The system of ODEs may be given the matrix form:

$$\mathbf{y}'(t) = \mathbf{A}\cdot\mathbf{y}(t) + \mathbf{b}\cdot\mathbb{1}(t) \text{ for } t\in[0,T]$$

where $\mathbf{y}(t) = [y_1(t)\; y_2(t)]^T$, $\mathbf{A} = \begin{bmatrix} -\frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{3}{2} \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

The matrix $\mathbf{A}$ may be factorized: $\mathbf{A} = \mathbf{V}\cdot\mathbf{\Lambda}\cdot\mathbf{V}^{-1}$, where $\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$ and $\mathbf{\Lambda} = diag\{\lambda_1, ..., \lambda_M\}$ is the diagonal matrix of eigenvalues:

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \left(-\frac{3}{2} - \lambda\right)^2 - \left(-\frac{1}{2}\right)^2 = 0 \Rightarrow \lambda_1 = -1, \lambda_2 = -2$$

The equality $\mathbf{A} = \mathbf{V}\cdot\mathbf{\Lambda}\cdot\mathbf{V}^{-1}$ may be given the form $\mathbf{A}\cdot\mathbf{V} = \mathbf{V}\cdot\mathbf{\Lambda}$ *i.e.*:

$$\begin{bmatrix} -\frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{3}{2} \end{bmatrix}\cdot\begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}\cdot\begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \Rightarrow \begin{bmatrix} \frac{3}{2}v_{11} + \frac{1}{2}v_{21} & \frac{3}{2}v_{12} + \frac{1}{2}v_{22} \\ \frac{1}{2}v_{11} + \frac{3}{2}v_{21} & \frac{1}{2}v_{12} + \frac{3}{2}v_{22} \end{bmatrix} = \begin{bmatrix} v_{11} & 2v_{12} \\ v_{21} & 2v_{22} \end{bmatrix}$$

or:

$$\frac{3}{2}v_{11} + \frac{1}{2}v_{21} - v_{11} = 0 \qquad \frac{3}{2}v_{12} + \frac{1}{2}v_{22} - 2v_{12} = 0$$

$$\frac{1}{2}v_{11} + \frac{3}{2}v_{21} - v_{21} = 0 \qquad \frac{1}{2}v_{12} + \frac{3}{2}v_{22} - 2v_{22} = 0$$

Hence:

$$v_{21} = -v_{11} \qquad v_{12} = v_{22}$$

$$v_{21} = -v_{11} \qquad v_{12} = v_{22}$$

The "simplest" solution of that set of algebraic equations:

$$\mathbf{V} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \text{ and } \mathbf{V}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

The sought-for set of independent scalar equations has the form:

$$\mathbf{V}^{-1}\cdot\mathbf{y}'(t) = \mathbf{\Lambda}\cdot\mathbf{V}^{-1}\cdot\mathbf{y}(t) + \mathbf{V}^{-1}\cdot\mathbf{b}\cdot\mathbb{1}(t)$$

$$z_1'(t) = -z_1(t) - \frac{1}{2}\mathbb{1}(t); \quad z_1(0) = 0$$

$$z_2'(t) = -2z_2(t) + \frac{1}{2}\mathbb{1}(t); \quad z_2(0) = 0$$

Their solutions have the form:

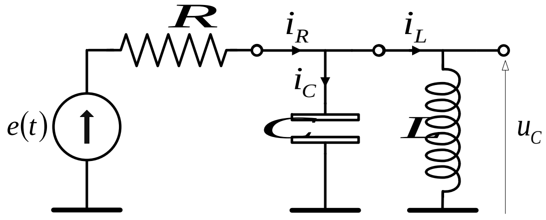
$$z_1(t) = -\frac{1}{2}\left(1 - e^{-t}\right)$$

$$z_2(t) = \frac{1}{4}\left(1 - e^{-2t}\right)$$

Thus, the solutions of the original system are as follows:

$$\mathbf{y}(t) = \mathbf{V}\cdot\mathbf{z}(t) = \mathbf{V} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\cdot\begin{bmatrix} -\frac{1}{2}\left(1 - e^{-t}\right) \\ \frac{1}{4}\left(1 - e^{-2t}\right) \end{bmatrix} = \begin{bmatrix} -\frac{1}{4} + \frac{1}{2}e^{-t} - \frac{1}{4}e^{-2t} \\ \frac{3}{4} - \frac{1}{2}e^{-t} - \frac{1}{4}e^{-2t} \end{bmatrix}$$

Problem: Compile and solve the system of ordinary differential equations modeling the following circuit:



Assume: $e(t)=\mathbb{1}(t)$, $R=4\ \mathrm{k\Omega}$, $C=1\ \mathrm{nF}$, $L=100\ \mathrm{mH}$ and zero initial conditions.

<u>Solution:</u> The Kirchhoff's laws yield:

$$e(t)-R\cdot i_R(t)-u_C(t)=0$$

$$i_R(t)-i_C(t)-i_L(t)=0$$

They should be satisfied together with the elemental equations:

$$u_C(t)=L\cdot\frac{di_L(t)}{dt}$$

$$i_C(t)=C\cdot\frac{du_C(t)}{dt}$$

Two state equations are obtained by elimination of two variables, *viz.* $i_R(t)$ and $i_C(t)$:

$$\left.\begin{array}{l} e(t)-R\cdot\left(i_C(t)+i_L(t)\right)-u_C(t)=0 \\ u_C(t)=L\cdot\dfrac{di_L(t)}{dt} \\ i_C(t)=C\cdot\dfrac{du_C(t)}{dt} \end{array}\right\}\Rightarrow\left\{\begin{array}{l} e(t)-R\cdot\left(C\cdot\dfrac{du_C(t)}{dt}+i_L(t)\right)-u_C(t)=0 \\ u_C(t)=L\cdot\dfrac{di_L(t)}{dt} \end{array}\right.$$

Hence:

$$\begin{cases} \dfrac{du_C(t)}{dt}=-\dfrac{1}{RC}u_C(t)-\dfrac{1}{C}i_L(t)+\dfrac{1}{RC}e(t) \\ \dfrac{di_L(t)}{dt}=\dfrac{1}{L}u_C(t) \end{cases}$$

or after substitution of the parameters:

$$\begin{cases} \dfrac{du_C(t)}{dt}=-2.5\cdot10^5\cdot u_C(t)-10^9\cdot i_L(t)+2.5\cdot10^5\cdot\mathbb{1}(t) \\ \dfrac{di_L(t)}{dt}=10\cdot u_C(t) \end{cases}$$

This system of ODEs may be given the matrix form:

$$\mathbf{y}'(t)=\mathbf{A}\cdot\mathbf{y}(t)+\mathbf{b}\cdot\mathbb{1}(t)\ \text{ for } t\in[0,T]$$

where $\mathbf{y}(t)=\left[u_C(t)\ i_L(t)\right]^T$, $\mathbf{A}=\begin{bmatrix}-2.5\cdot10^5 & -10^9 \\ 10 & 0\end{bmatrix}$, $\mathbf{b}=\begin{bmatrix}2.5\cdot10^5 \\ 0\end{bmatrix}$

The matrix $\mathbf{A}$ may be factorized: $\mathbf{A} = \mathbf{V}\cdot\mathbf{\Lambda}\cdot\mathbf{V}^{-1}$, where $\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$ and $\mathbf{\Lambda} = diag\left\{\lambda_1, ..., \lambda_M\right\}$ is the diagonal matrix of eigenvalues:

$$\mathbf{V} = \begin{bmatrix} -0.999999998750000 & 0.999999980000 \\ 0.000049999999938 & -0.000199999996 \end{bmatrix} \quad \lambda_1 = -2\cdot 10^5 \text{ and } \lambda_2 = -0.5\cdot 10^5$$

$$\mathbf{V}^{-1} = \begin{bmatrix} -1.333333335 & -6666.666675 \\ -0.333333340 & -6666.666800 \end{bmatrix}$$

The sought-for set of independent scalar equations has the form:

$$\mathbf{V}^{-1}\cdot\mathbf{y}'(t) = \mathbf{D}\cdot\mathbf{V}^{-1}\cdot\mathbf{y}(t) + \mathbf{V}^{-1}\cdot\mathbf{b}\cdot\mathbb{1}(t)$$

where $\mathbf{V}^{-1}\cdot\mathbf{b} = \begin{bmatrix} -3.3333333375\cdot 10^5 \\ -0.8333333500\cdot 10^5 \end{bmatrix} \equiv \boldsymbol{\beta}$

$$z_1'(t) = -2\cdot 10^5\cdot z_1(t) - \beta_1\cdot\mathbb{1}(t); \quad z_1(0) = 0$$

$$z_2'(t) = -0.5\cdot 10^5\cdot z_1(t) - \beta_2\cdot\mathbb{1}(t); \quad z_1(0) = 0$$

Their solutions have the form:

$$z_1(t) = \frac{\beta_1}{\lambda_1}\left(e^{\lambda_1 t} - 1\right) = 1.66666667\left(e^{-2\cdot 10^5 t} - 1\right)$$

$$z_2(t) = \frac{\beta_2}{\lambda_2}\left(e^{\lambda_2 t} - 1\right) = 1.6666667\left(e^{-0.5\cdot 10^5 t} - 1\right)$$

Thus, the solutions of the original system are as follows:

$$\mathbf{y}(t) = \mathbf{V}\cdot\mathbf{z}(t)$$

5.3. Determination of error indicators

Problem: Determine the coefficients and local-error indicators for the following methods for solving ODEs:

- the Heun method,
- the mid-point method,
- the explicit Adams methods of order 2-4,
- the implicit Adams methods of order 2-4,
- the explicit Gear methods of order 2-4,
- the implicit Gear methods of order 2-4.

Solution: Verify the results by comparing them with the corresponding lecture slides.

Problem: Assess the local error of the following method for solving ODEs:

$$y_n = 2y_{n-1} - y_{n-2} + h\left(y_{n-1}' - y_{n-2}'\right)$$

Sketch the border of the region of absolute stability on the λh plane.

Solution: The development of the RHS of:

$$y_n = 2\dot{y}_{n-1} - \dot{y}_{n-2} + h\left(\dot{y}_{n-1}' - \dot{y}_{n-2}'\right)$$

into the Taylor series yields:

$$RHS = 2\left(\dot{y}_n - \dot{y}_n' h + \frac{1}{2}\dot{y}_n'' h^2 - \frac{1}{6}\dot{y}_n''' h^3 + \ldots \right) - \left(\dot{y}_n - 2\dot{y}_n' h + 2\dot{y}_n'' h^2 - \frac{4}{3}\dot{y}_n''' h^3 + \ldots \right)$$
$$+ \left(\dot{y}_n' h - \dot{y}_n'' h^2 + \frac{1}{2}\dot{y}_n''' h^3 - \ldots \right) - \left(\dot{y}_n' h - 2\dot{y}_n'' h^2 + 2\dot{y}_n''' h^3 - \ldots \right)$$
$$RHS = (2-1)\dot{y}_n + (-2+2+1-1)\dot{y}_n' h + (1-2-1+2)\dot{y}_n'' h^2 + \left(-\frac{1}{3} + \frac{4}{3} + \frac{1}{2} - 2 \right)\dot{y}_n''' h^3 + \ldots$$
$$RHS = \dot{y}_n - \frac{1}{2}\dot{y}_n''' h^3 + \ldots$$
This means that the method is of order 2 and the local error is $e_n \cong -\frac{1}{2}\dot{y}_n''' h^3$.

The absolute stability analysis is based on the application of the tested method to the test equation:
$$y'(t) = \lambda y(t)$$
which yields in the time domain:
$$y_n = 2y_{n-1} - y_{n-2} + h\lambda\left(y_{n-1} - y_{n-2} \right)$$
and in the Z-transform domain:
$$1 = 2z^{-1} - z^{-2} + h\lambda\left(z^{-1} - z^{-2} \right)$$
Hence:
$$h\lambda = \frac{1-2z^{-1}+z^{-2}}{z^{-1}-z^{-2}} = \frac{\left(1-z^{-1}\right)^2}{z^{-1}\left(1-z^{-1}\right)} = \frac{1-z^{-1}}{z^{-1}} = z-1$$
Thus, the equation of the border of the region of absolute stability is:
$$h\lambda = e^{j\phi} - 1 \quad \text{for} \quad \phi \in [0, 2\pi]$$
This is the unit-radius circle whose centre is located at (-1,0).

Problem: Determine the parameters α_1 and α_2 of the following method for solving ODEs: $y_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + h\left(y_{n-1}' - y_{n-2}' \right)$, as to make it to be of highest possible order. Provide the order of the method and the local error assessment.

<u>Solution:</u> The development of the RHS of:
$$y_n = \alpha_1 \dot{y}_{n-1} + \alpha_2 \dot{y}_{n-2} + h\left(\dot{y}_{n-1}' - \dot{y}_{n-2}' \right)$$
into the Taylor series yields:
$$RHS = \alpha_1\left(\dot{y}_n - \dot{y}_n' h + \frac{1}{2}\dot{y}_n'' h^2 - \frac{1}{6}\dot{y}_n''' h^3 + \ldots \right) + \alpha_2\left(\dot{y}_n - 2\dot{y}_n' h + 2\dot{y}_n'' h^2 - \frac{4}{3}\dot{y}_n''' h^3 + \ldots \right)$$
$$+ \left(\dot{y}_n' h - \dot{y}_n'' h^2 + \frac{1}{2}\dot{y}_n''' h^3 - \ldots \right) - \left(\dot{y}_n' h - 2\dot{y}_n'' h^2 + 2\dot{y}_n''' h^3 - \ldots \right)$$
$$RHS = (\alpha_1 + \alpha_2)\dot{y}_n + (-\alpha_1 - 2\alpha_2 + 1 - 1)\dot{y}_n' h + \left(\frac{\alpha_1}{2} + 2\alpha_2 - 1 + 2 \right)\dot{y}_n'' h^2$$
$$+ \left(-\frac{\alpha_1}{6} - \frac{4}{3}\alpha_2 + \frac{1}{2} - 2 \right)\dot{y}_n''' h^3 + \ldots$$
Hence two equations:
$$\alpha_1 + \alpha_2 = 1$$
$$-\alpha_1 - 2\alpha_2 + 1 - 1 = 0$$

whose solution is: $\alpha_1 = 2$ and $\alpha_2 = -1$. For these parameters:

$$\frac{\alpha_1}{2} + 2\alpha_2 - 1 + 2 = 0$$

and therefore the method is of order 2 with the local error assessment:

$$e_n \cong \left(-\frac{\alpha_1}{6} - \frac{4}{3}\alpha_2 + \frac{1}{2} - 2 \right) \dot{y}_n''' h^3 = -\frac{1}{2} \dot{y}_n''' h^3$$

Problem: Assess the local error of the midpoint method (modified Euler's method), defined by the formula:

$$y_n = y_{n-1} + hf\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} f\left(t_{n-1}, y_{n-1}\right) \right)$$

Solution: Let's notice that the formula may be re-written in the form:

$$y_n = y_{n-1} + hf\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} y'_{n-1} \right)$$

Then the function $f\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} y'_{n-1} \right)$ may be developed in the following Taylor series:

$$\frac{y_n - y_{n-1}}{h} = f\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} y'_{n-1} \right) = f\left(t_{n-1}, y_{n-1}\right) + f_t\left(t_{n-1}, y_{n-1}\right)\frac{h}{2} + f_y\left(t_{n-1}, y_{n-1}\right)\frac{h}{2} y'_{n-1} +$$

$$+\frac{1}{2} f_{tt}\left(t_{n-1}, y_{n-1}\right)\frac{h^2}{4} + \frac{1}{2} f_{yy}\left(t_{n-1}, y_{n-1}\right)\frac{h^2}{4}\left(y'_{n-1}\right)^2 +$$

$$+\frac{1}{2} f_{ty}\left(t_{n-1}, y_{n-1}\right)\frac{h^2}{4} y'_{n-1} + \frac{1}{2} f_{yt}\left(t_{n-1}, y_{n-1}\right)\frac{h^2}{4} y'_{n-1} + \ldots$$

Let's notice that $y'(x) = f\left(x, y(x)\right)$ implies:

$$y''(t) = f_t\left(t, y(t)\right) + f_y\left(t, y(t)\right) y'(t)$$

$$y'''(t) = \left[f_{tt}\left(t, y(t)\right) + f_{ty}\left(t, y(t)\right) y'(t) \right] + \left[f_{yt}\left(t, y(t)\right) y'(t) + f_{yy}\left(t, y(t)\right)\left(y'(t)\right)^2 + f_y\left(t, y(t)\right) y''(t) \right]$$

Hence:

$$\frac{y_n - y_{n-1}}{h} = f\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} y'_{n-1} \right)$$

$$= y'\left(t_{n-1}\right) + y''\left(t_{n-1}\right)\frac{h}{2} + \left[y'''\left(t_{n-1}\right) - f_y\left(t_{n-1}, y\left(t_{n-1}\right)\right) y''\left(t_{n-1}\right) \right]\frac{h^2}{8} + o\left(h^3\right)$$

and

$$y_n = y_{n-1} + hf\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} y'_{n-1} \right)$$

$$= y_{n-1} + y'\left(t_{n-1}\right) h + y''\left(t_{n-1}\right)\frac{h^2}{2} + \left[y'''\left(t_{n-1}\right) - f_y\left(t_{n-1}, y\left(t_{n-1}\right)\right) y''\left(t_{n-1}\right) \right]\frac{h^3}{8} + o\left(h^4\right)$$

Taking into account that:

$$y\left(t_n\right) = y\left(t_{n-1}\right) + y'\left(t_{n-1}\right) h + \frac{1}{2} y''\left(t_{n-1}\right) h^2 + \frac{1}{6} y'''\left(t_{n-1}\right) h^3 + o\left(h^4\right)$$

$$= y_{n-1} + y'_{n-1}h + \frac{1}{2}y''_{n-1}h^2 + \frac{1}{6}y'''(t_{n-1})h^3 + o(h^4)$$

one may gather that the local error is:

$$y(t_n) - y_n = \left[\frac{1}{6}y'''(t_{n-1}) - \frac{1}{8}\left(y'''_{n-1} - f_y(t_{n-1}, y_{n-1})\,y''_{n-1}\right)\right]h^3 + o(h^4)$$

$$= \left[\frac{1}{24}y'''_{n-1} + \frac{1}{8}f_y(t_{n-1}, y_{n-1})\,y''_{n-1}\right]h^3 + o(h^4)$$

Problem: Determine the coefficients and the indicators of local accuracy for the implicit Adams method of order 2

$$y_n = y_{n-1} + h\cdot\left(\beta_0^*\cdot f_n + \beta_1^*\cdot f_{n-1}\right)$$

a) by integration of Lagrange polynomial,
b) by expansion of RHS into a Taylor series.

Solution: The Lagrange polynomial has the form:

$$\hat{f}(t) = f_n\frac{t - t_{n-1}}{t_n - t_{n-1}} + f_{n-1}\frac{t - t_n}{t_{n-1} - t_n}$$

Hence:

$$h\cdot\beta_0^* = \int_{t_{n-1}}^{t_n}\frac{t - t_{n-1}}{t_n - t_{n-1}}dt = \left|\begin{matrix}\tau = t - t_{n-1}\\ t = \tau + t_{n-1}\end{matrix}\right| = \int_0^h\frac{\tau}{h}d\tau = \frac{1}{2}h \Rightarrow \beta_0^* = \frac{1}{2}$$

$$h\cdot\beta_1^* = \int_{t_{n-1}}^{t_n}\frac{t - t_n}{t_{n-1} - t_n}dt = \left|\begin{matrix}\tau = t - t_{n-1}\\ t = \tau + t_{n-1}\end{matrix}\right| = \int_0^h\frac{\tau - h}{(-h)}d\tau = \frac{1}{2}h \Rightarrow \beta_1^* = \frac{1}{2}$$

The expansion of RHS into a Taylor series yields:

$$\text{RHS} = \left(y_n - y'_nh + \frac{1}{2}y''_nh^2 - \frac{1}{6}y'''_nh^3 + \ldots\right) + h\beta_0^*y'_n + h\beta_1^*\left(y'_n - y''_nh + \frac{1}{2}y'''_nh^2 + \ldots\right)$$

$$= y_n + \left(-1 + \beta_0^* + \beta_1^*\right)y'_nh + \left(\frac{1}{2} - \beta_1^*\right)y''_nh^2 + \left(-\frac{1}{6} + \frac{1}{2}\beta_1^*\right)y'''_nh^3 + \ldots$$

Hence the equations:

$$-1 + \beta_0^* + \beta_1^* = 0,\quad \frac{1}{2} - \beta_1^* = 0$$

whose solution is provided above, and an estimate of the local error:

$$r_n = \left(-\frac{1}{6} + \frac{1}{2}\beta_1^*\right)y'''_nh^3 = \left(-\frac{1}{6} + \frac{1}{4}\right)y'''_nh^3 = \frac{1}{12}y'''_nh^3$$

Thus, finally: $\beta_0^* = \frac{1}{2}$, $\beta_1^* = \frac{1}{2}$ and $c_3^* = \frac{1}{12}$.

Problem: Determine the coefficients (α_1 , α_2 , β_1) and the indicators of local accuracy for the following explicit Gear's method of order 2:

$$y_{n+1} = \alpha_1\cdot y_n + \alpha_2\cdot y_{n-1} + h\cdot\beta_1\cdot f_n$$

a) by integration of Lagrange polynomial,
b) by expansion of RHS into a Taylor series.

Solution: The Lagrange polynomial has the form:

$$\hat{y}(t) = y_{n+1}\frac{(t-t_n)}{(t_{n+1}-t_n)}\frac{(t-t_{n-1})}{(t_{n+1}-t_{n-1})} + y_n\frac{(t-t_{n+1})}{(t_n-t_{n+1})}\frac{(t-t_{n-1})}{(t_n-t_{n-1})} + y_{n-1}\frac{(t-t_{n+1})}{(t_{n-1}-t_{n+1})}\frac{(t-t_n)}{(t_{n-1}-t_n)} =$$

$$= y_{n+1}\frac{(t-t_n)}{(h)}\frac{(t-t_{n-1})}{(2h)} + y_n\frac{(t-t_{n+1})}{(-h)}\frac{(t-t_{n-1})}{(h)} + y_{n-1}\frac{(t-t_{n+1})}{(-2h)}\frac{(t-t_n)}{(-h)}$$

and its derivative is:

$$\hat{y}'(t) = \frac{1}{2h^2}\left\{y_{n+1}\left[(t-t_n)+(t-t_{n-1})\right] - 2y_n\left[(t-t_{n+1})+(t-t_{n-1})\right] + y_{n-1}\left[(t-t_{n+1})+(t-t_n)\right]\right\}$$

$$\hat{y}'(t_n) = \frac{1}{2h^2}\left\{y_{n+1}[0+h] + 2y_n[-h+h] + y_{n-1}[-h+0]\right\} = \frac{1}{2h}\left\{y_{n+1}-y_{n-1}\right\} = f_n$$

Hence:

$$y_{n+1} = y_{n-1} + 2hf_n$$

which means that:

$$\alpha_1 = 0,\ \alpha_2 = 1 \text{ and } \beta_1 = 2.$$

The indicators of local accuracy result from the comparison of the terms of order 3:

$$r_{n+1} = \frac{1}{6}y_n'''h^3 - \left(-\frac{1}{6}y_n'''h^3\right) = \frac{1}{3}y_n'''h^3$$

Problem: Determine the values of the parameters α_1, α_2 and β_1 of the following method for solving ODEs:

$$\hat{y}_n = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + h\beta_1 f(x_{n-1}, y_{n-1})$$

in such a way as to make this method be of order 2. Assess the local error of this method.

<u>Solution:</u> The development of the components into the Taylor series yields:

$$RHS = \alpha_1 y_{n-1} + \alpha_2 y_{n-2} + h\beta_1 y_{n-1}'$$

$$RHS = \alpha_1\left(\dot{y}_n - \dot{y}_n'h + \frac{1}{2}\dot{y}_n''h^2 - \frac{1}{6}\dot{y}_n'''h^3 + \ldots\right)$$

$$+\alpha_2\left(\dot{y}_n - 2\dot{y}_n'h + 2\dot{y}_n''h^2 - \frac{4}{3}\dot{y}_n'''h^3 + \ldots\right)$$

$$+h\beta_1\left(\dot{y}_n' - \dot{y}_n''h + \frac{1}{2}\dot{y}_n'''h^2 - \right)$$

Hence the set of equations:

$$\alpha_1 + \alpha_2 = 1$$

$$-\alpha_1 - 2\alpha_2 + \beta_1 = 0$$

$$\frac{1}{2}\alpha_1 + 2\alpha_2 - \beta_1 = 0$$

whose solution is: $\alpha_1 = 0, \alpha_2 = 1, \beta_1 = 2$. The local error may be assessed by means of the term:

$$r_n = \alpha_1\left(-\frac{1}{6}\dot{y}_n'''h^3\right) + \alpha_2\left(-\frac{4}{3}\dot{y}_n'''h^3\right) + h\beta_1\left(\frac{1}{2}\dot{y}_n'''h^2\right) = 1\cdot\left(-\frac{4}{3}\dot{y}'''h^3\right) + 2\cdot\left(\frac{1}{2}\dot{y}_n'''h^3\right) = -\frac{1}{3}\dot{y}_n'''h^3$$

Problem: The following ODE:

$$y'(t) = -y(t) + 1;\quad y(0) = 0$$

has been solved by means of the open Euler's method: $y_n = y_{n-1} + h \cdot f\left(t_{n-1}, y_{n-1}\right)$. Assess the total (global) error of the obtained solution for $n \to \infty$ and $|h| << 1$.

Solution: Since $y(t) = C \cdot e^{-t}$ is the solution of the homogenous equation $y'(t) = -y(t)$, the general solution of the nonhomogeneous equation and its derivative have the form:

$$y(t) = C(t) \cdot e^{-t} \quad \text{and} \quad y'(t) = C'(t) \cdot e^{-t} - C(t) \cdot e^{-t}$$

After substitution to the ODE under consideration:

$$C'(t) \cdot e^{-t} - C(t) \cdot e^{-t} = -C(t) \cdot e^{-t} + 1$$

and:

$$C'(t) = e^{t} \quad \text{and} \quad C(t) = \int e^{t} dt = e^{t} + C_0$$

where C_0 is a constant to be determined on the basis of the intial condition $y(0) = 0$:

$$C(0) \cdot e^{-0} = 0 \Rightarrow 1 + C_0 = 0 \Rightarrow C_0 = -1$$

Thus $C(t) = e^{t} - 1$, and consequently:

$$y(t) = C(t) \cdot e^{-t} = \left(e^{t} - 1\right) \cdot e^{-t} = 1 - e^{-t}$$

The expansion of the open Euler's formula into the Taylor series:

$$y_n = \left(\dot{y}_n - \dot{y}_n^{(1)} h + \frac{1}{2} \dot{y}_n^{(2)} h^2 + \ldots \right) + h\left(\dot{y}_n^{(1)} - \dot{y}_n^{(2)} h + \ldots \right) = \dot{y}_n - \frac{1}{2} \dot{y}_n^{(2)} h^2 + \ldots$$

yields the following estimate of the local error:

$$r_n \cong -\frac{1}{2} \dot{y}_n^{(2)} h^2 = \frac{1}{2} e^{-nh} h^2 > 0$$

Thus the total (global) error may be assessed for $|h| << 1$ in the following way:

$$e_n \cong \sum_{i=1}^{n} r_i = \frac{h^2}{2} \sum_{i=1}^{n} e^{-ih} \ldots \to \frac{h^2}{2} \frac{e^{-h}}{1 - e^{-h}} \cong \frac{h^2}{2} \frac{1-h}{h} \cong \frac{h}{2}$$

Problem: The open Gear's method:

$$y_n = y_{n-2} + 2hy'_{n-1}$$

has been applied for solving the ODE modelling an electronic circuit. When using the step h the approximate voltage value $u_{10}^{(1)} = 10.01\,\text{V}$ has been obtained, while when making two steps of length $\frac{h}{2}$ – the value $u_{10}^{(2)} = 9.99\,\text{V}$. Assess the local error the solution $u_{10}^{(1)}$ is subject to.

Solution: The applied Gear's method is of order 2 ($p = 2$) because:

$$y_n = y_{n-2} + 2hy'_{n-1} = \dot{y}_n - 2\dot{y}'_n h + 2\dot{y}''_n h^2 - \tfrac{4}{3} \dot{y}'''_n h^3 + \ldots$$

$$2\dot{y}'_n h - 2\dot{y}''_n h^2 + \dot{y}'''_n h^3 + \ldots = \dot{y}_n - \tfrac{4}{3} \dot{y}'''_n h^3 + \ldots$$

The error to be assessed is of order $r_{10}^{(1)} \cong \gamma h^{p+1}$; so:

$$u_{10}^{(1)} \cong \dot{y}_n + \gamma h^{p+1} = \dot{y}_n + r_{10}^{(1)}$$

Since:

$$u_{10}^{(2)} \cong \dot{y}_n + 2\gamma \left(\frac{h}{2} \right)^{p+1} = \dot{y}_n + \frac{r_{10}^{(1)}}{2^p}$$

the solution of the problem has the form:

$$u_{10}^{(1)} - u_{10}^{(2)} \cong r_{10}^{(1)} - \frac{r_{10}^{(1)}}{2^p} = r_{10}^{(1)}\left(1-\frac{1}{2^p}\right) \text{ and } r_{10}^{(1)} \cong \frac{u_{10}^{(1)} - u_{10}^{(2)}}{1-\frac{1}{2^p}} = \frac{0.02}{0.75} \cong 0.027\,[\mathrm{V}]$$

Problem: The local error of the second-order Adams-Bashforth (AB) method may be estimated by:

$$r_n^{(AB)} \cong -\frac{5}{12} y_n''' h^3$$

The local error of the second-order Adams-Moulton (AM) method may be estimated by:

$$r_n^{(AM)} \cong \frac{1}{12} y_n''' h^3$$

An ODE has been solved by means of both methods and the following results have been obtained: $y_n^{(AB)}$ for the AB method and $y_n^{(AM)}$ for the AM method. Assess the local error corrupting $y_n^{(AM)}$.

Solution: Since:

$$y_n^{(AB)} \cong \dot{y}_n - \frac{5}{12} y_n''' h^3 \text{ and } y_n^{(AM)} \cong \dot{y}_n + \frac{1}{12} y_n''' h^3$$

the unknown derivative value y_n''' may be found on the basis of the difference:

$$y_n^{(AM)} - y_n^{(AB)} \cong \frac{1}{12} y_n''' h^3 + \frac{5}{12} y_n''' h^3 = \frac{1}{2} y_n''' h^3$$

The substitution of the solution of the above equation:

$$y_n''' h^3 \cong 2\left(y_n^{(AM)} - y_n^{(AB)}\right)$$

to the formula defining local error of the AM method yields:

$$r_n^{(AM)} \cong \frac{1}{6}\left(y_n^{(AM)} - y_n^{(AB)}\right)$$

Problem: Assess the local error of the solution of the equation $y'(t) = -y(t)$, which has been obtained by means of the Lobatto IIIA method defined by the formula:

$$y_n = y_{n-1} + \frac{1}{2} h\left(f_1 + f_2\right), \text{ where: } f_1 = f\left(t_{n-1}, y_{n-1}\right) \text{ and } f_2 = f\left(t_n, y_{n-1} + \frac{1}{2} h\left(f_1 + f_2\right)\right)$$

Solution: The application of the Lobatto IIIA method to the equation $y'(t) = -y(t)$ yields:

$$y_n = y_{n-1} + \frac{1}{2} h\left(f_1 + f_2\right), \text{ where: } f_1 = -y_{n-1} \text{ and } f_2 = -y_{n-1} - \frac{1}{2} h\left(f_1 + f_2\right)$$

The sum of the equations defining f_1 and f_2 is:

$$f_1 + f_2 = -2y_{n-1} - \frac{1}{2} h\left(f_1 + f_2\right)$$

Hence:

$$f_1 + f_2 = -\frac{2}{1+\frac{h}{2}} y_{n-1}$$

After substitution of this result to the Lobatto IIIA formula, the following recursive equation is obtained:

$$y_n = y_{n-1} - \frac{h}{1+\frac{h}{2}} y_{n-1} = \frac{1-\frac{h}{2}}{1+\frac{h}{2}} y_{n-1}$$

which may be used for assessing he local error in the following way:

$$y_n = \frac{1-\frac{h}{2}}{1+\frac{h}{2}} \dot{y}_{n-1} = \frac{1-\frac{h}{2}}{1+\frac{h}{2}} \dot{y}\left(t_n - h\right) = \sum_{i=0}^{\infty} c_i h^i$$

where c_i are coefficients of the corresponding Taylor series. The latter may be obtained directly or by multiplying the series representative of $\dot{y}\left(t_n - h\right)$ and the series:

$$\frac{1-\frac{h}{2}}{1+\frac{h}{2}} = 1 - h + \frac{1}{2}h^2 - \frac{1}{4}h^3 + \ldots$$

The Taylor series representative of $\dot{y}\left(t_n - h\right)$ has the form:

$$\dot{y}\left(t_n - h\right) = \dot{y}_n - \dot{y}_n' \, h + \frac{1}{2}\dot{y}_n'' \, h^2 - \frac{1}{6} y_n''' \, h^3 + \ldots$$

$$= \dot{y}_n + \dot{y}_n h + \frac{1}{2}\dot{y}_n h^2 + \frac{1}{6}\dot{y}_n h^3 + \ldots = \dot{y}_n \left(1 + h + \frac{1}{2}h^2 + \frac{1}{6}h^3 + \ldots\right)$$

because in the considered case: $\dot{y}_n' = -\dot{y}_n$, $\dot{y}_n'' = -\dot{y}_n'$ and $\dot{y}_n''' = -\dot{y}_n''$. The product of both series divided by $\dot{y}_n$:

$$1 + h + \frac{1}{2}h^2 + \frac{1}{6}h^3 + \ldots$$

$$-h - h^2 - \frac{1}{2}h^3 - \frac{1}{6}h^4 + \ldots$$

$$+\frac{1}{2}h^2 + \frac{1}{2}h^3 + \frac{1}{4}h^4 + \frac{1}{12}h^5 + \ldots = 1 + \frac{1}{6}h^3 + \ldots$$

enables one to conclude that: $r_n \cong \frac{1}{6}\dot{y}_n h^3$.

5.4. Testing absolute stability

Problem: Check whether the following methods for solving ODEs:

- the Heun method,
- the mid-point method,
- the explicit Adams methods of order 2–4,
- the implicit Adams methods of order 2–4,
- the explicit Gear methods of order 2–4,
- the implicit Gear methods of order 2–4.

are absolutely stable for the following values of $h\lambda$:

- $-3, -2, -1$;
- $j3,\ j2,\ j1$;
- $1+j3, 1+j2, 1+j1$;

Solution: Verify the results by comparing them with the corresponding graphs on the lecture slides.

Problem: The following ODE:

$$y''(t)+21\cdot y'(t)+20\cdot y(t)=\sin(t) \text{ for } t>0$$

with the initial conditions:

$$y(0)=0 \text{ and } y'(0)=0$$

has been solved using the open Euler method: $y_n = y_{n-1}+h\cdot f(t_{n-1},y_{n-1})$. What are the values of the integration step h guarantying the absolute stability in this case?

Solution: The stability conditions may be derived from the solution obtained for the zero RHS and non-zero initial conditions. This solution has the following general form:

$$y(t)=C_1\cdot e^{\lambda_1 t}+C_2\cdot e^{\lambda_2 t}$$

And its derivatives – the form:

$$y'(t)=C_1\cdot\lambda_1\cdot e^{\lambda_1 t}+C_2\cdot\lambda_2\cdot e^{\lambda_2 t} \text{ i } y''(t)=C_1\cdot\lambda_1^2\cdot e^{\lambda_1 t}+C_2\cdot\lambda_2^2\cdot e^{\lambda_2 t}$$

After substitution of $y(t)$, $y'(t)$ i $y''(t)$ to the ODE with the zero RHS, we obtain:

$$C_1\cdot\left(\lambda_1^2+21\cdot\lambda_1+20\right)\cdot e^{\lambda_1 t}+C_2\cdot\left(\lambda_2^2+21\cdot\lambda_2+20\right)\cdot e^{\lambda_2 t}=0$$

Hence the conclusion that λ_1 and λ_2 are solutions of the quadratic equation:

$$\lambda^2+21\cdot\lambda+20=0$$

i.e. $\lambda_1=-1$ and $\lambda_2=-20$. Since $\mathrm{Im}(\lambda_1)=\mathrm{Im}(\lambda_2)=0$, the condition of the absolute stability takes on the form:

$$h\cdot\sup\left\{|\lambda_1|,|\lambda_2|\right\}<2 \text{, or } h\cdot 20<2 \text{, or } h<0.1.$$

Problem: Verify whether the midpoint method for solving ODEs, defined by the formula:

$$y_n=y_{n-1}+hf\left(x_{n-1}+\frac{h}{2},y_{n-1}+\frac{h}{2}f(x_{n-1},y_{n-1})\right)$$

is absolutely stable for $h\lambda=-1-j$ and $h\lambda=-1+j$.

Solution: For the test equation:

$$y_n=y_{n-1}+h\lambda\left(y_{n-1}+\frac{1}{2}h\lambda y_{n-1}\right)=\left(1+h\lambda+\frac{1}{2}(h\lambda)^2\right)y_{n-1}$$

The stability condition is: $\left|1+h\lambda+\frac{1}{2}(h\lambda)^2\right|<1$. Thus, for $h\lambda=-1-j$, the method is stable because:

$$\left|1-1-j+\frac{1}{2}(-1-j)^2\right|=\left|-j+\frac{1}{2}(1+2j-1)\right|=0<1$$

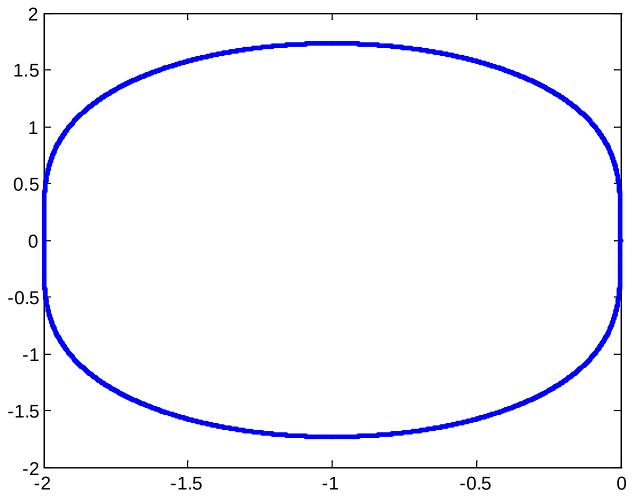
and for $h\lambda=-1+j$, it is stable also because:

$$\left|1-1+j+\frac{1}{2}(-1+j)^2\right|=\left|j+\frac{1}{2}(1-2j-1)\right|=0<1$$

In fact, the borderline delimiting the region of absolute stability:

$$1+h\lambda+\frac{1}{2}(h\lambda)^2=e^{j\phi} \text{ for } \phi\in[0,2\pi]$$

has the form shown in the figure below.



Problem: Determine the region of absolute stability for the midpoint method:

$$y_n = y_{n-1} + h \cdot f\left(t_{n-1} + \frac{h}{2}, y_{n-1} + \frac{h}{2} f\left(t_{n-1}, y_{n-1}\right)\right)$$

<u>Solution:</u> The method may be rewritten in the form:

$$y_n = y_{n-1} + h \cdot f\left(t_{n-\frac{1}{2}}, y_{n-\frac{1}{2}}\right)$$

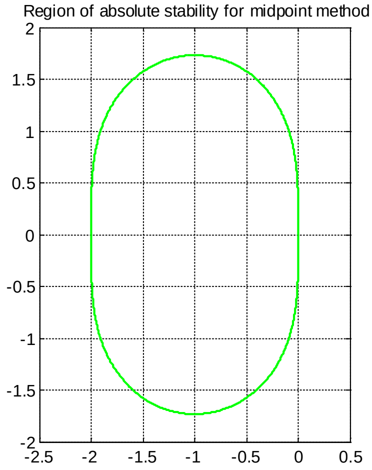
where $t_{n-\frac{1}{2}} = t_{n-1} + \frac{h}{2}$ and $y_{n-\frac{1}{2}} = y_{n-1} + \frac{h}{2} f\left(t_{n-1}, y_{n-1}\right)$. Its application to the test equation $y'(t) = \lambda \cdot y(t)$ yields:

$$y_{n-\frac{1}{2}} = y_{n-1} + \frac{h\lambda}{2} y_{n-1} = \left(1 + \frac{h\lambda}{2}\right) y_{n-1}$$

$$y_n = y_{n-1} + h\lambda\left(1 + \frac{h\lambda}{2}\right) y_{n-1} = \left(1 + h\lambda + \frac{1}{2}(h\lambda)^2\right) y_{n-1}$$

The absolute stability condition has the form:

$$\left|1 + h\lambda + \frac{1}{2}(h\lambda)^2\right| < 1 \text{ or } 1 + h\lambda + \frac{1}{2}(h\lambda)^2 = e^{j\phi}$$



Problem: Determine the equation of the border of the region of absolute stability for the Heun method:

$$y_n = y_{n-1} + \frac{1}{2}h\left[f\left(t_{n-1}, y_{n-1}\right) + f\left(t_{n-1}+h, y_{n-1} + hf\left(t_{n-1}, y_{n-1}\right)\right)\right]$$

Draw this border on the $h\lambda$ plane.

<u>Solution:</u> The method may be rewritten in the form:

$$y_n = y_{n-1} + h \cdot \frac{\left[f\left(t_{n-1}, y_{n-1}\right) + f\left(t_n, \hat{y}_n\right)\right]}{2}$$

where $\hat{y}_n = y_{n-1} + hf\left(t_{n-1}, y_{n-1}\right)$. Its application to the test equation $y'(t) = \lambda \cdot y(t)$ yields:

$$\hat{y}_n = y_{n-1} + h\lambda \cdot y_{n-1} = (1+h\lambda) \cdot y_{n-1}$$

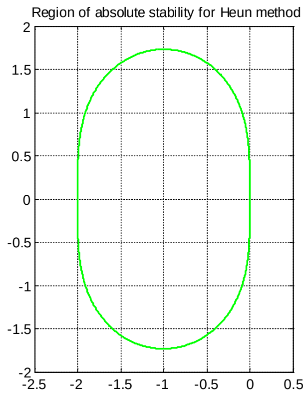
$$y_n = y_{n-1} + h \cdot \frac{\left[\lambda \cdot y_{n-1} + \lambda \cdot \hat{y}_n\right]}{2} = y_{n-1} + \frac{h\lambda}{2}\left[y_{n-1} + (1+h\lambda)\cdot y_{n-1}\right] = \left[1 + h\lambda + \frac{1}{2}(h\lambda)^2\right] y_{n-1}$$

The absolute stability condition has the form:

$$\left|1 + h\lambda + \frac{1}{2}(h\lambda)^2\right| < 1 \text{ or } 1 + h\lambda + \frac{1}{2}(h\lambda)^2 = e^{j\phi}$$

The solutions of this equation for selected values of ϕ are:

- for $\phi = 0$: $h\lambda = (0,0)$ or $(-2,0)$;
- for $\phi = \pi$: $h\lambda = \left(-1, \sqrt{3}\right)$ or $\left(-1, -\sqrt{3}\right)$.



Problem: Check the absolute stability of the Shichman formula:

$$y_n = \frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2} + \frac{2}{3}hy'_n$$

for (a) $h\lambda = 1$, (b) $h\lambda = -1$, (c) $h\lambda = j$, (d) $h\lambda = \frac{3}{2}$.

<u>Solution:</u> When applied to the test equation $y'(t) = \lambda y(t)$, the Shichman formula yields:

$$y_n = \frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2} + \frac{2}{3}h\lambda y_n \quad \text{or} \quad \left(1 - \frac{2}{3}h\lambda\right) y_n = \frac{4}{3}y_{n-1} - \frac{1}{3}y_{n-2}$$

After z-transformation, the latter difference equation takes on the form:

$$(3 - 2h\lambda)Y\left(z^{-1}\right) = 4Y\left(z^{-1}\right)z^{-1} - Y\left(z^{-1}\right)z^{-2}$$

Hence the characteristic equation:

$$(3-2h\lambda)=4z^{-1}-z^{-2} \quad \text{or} \quad (3-2h\lambda)z^2-4z+1=0$$

Its solutions are:

$$z_1=\frac{4+\sqrt{\Delta}}{2(3-2h\lambda)} \quad \text{and} \quad z_1=\frac{4-\sqrt{\Delta}}{2(3-2h\lambda)} \quad \text{with} \quad \Delta=16-4(3-2h\lambda)=4(1+2h\lambda)$$

or:

$$z_1=\frac{2+\sqrt{1+2h\lambda}}{3-2h\lambda} \quad \text{and} \quad z_2=\frac{2-\sqrt{1+2h\lambda}}{3-2h\lambda}$$

(a) For $h\lambda=1$:

$$|z_1|=\left|2+\sqrt{3}\right|>1 \quad \text{and} \quad |z_2|=\left|2-\sqrt{3}\right|<1 \text{ (stability not guaranteed)}$$

(b) For $h\lambda=-1$:

$$|z_1|=\left|\frac{2+j}{6}\right|<1 \quad \text{and} \quad |z_2|=\left|\frac{2-j}{6}\right|<1 \text{ (stability guaranteed)}$$

(c) For $h\lambda=j$:

$$|z_1|=\left|\frac{2+\sqrt{1+2j}}{3-2j}\right|\cong 0.933<1 \quad \text{and} \quad |z_2|=\left|\frac{2-\sqrt{1+2j}}{3-2j}\right|\cong 0.297<1 \text{ (stability guaranteed)}$$

(d) For $h\lambda=\frac{3}{2}$:

$$z_1=\frac{2+\sqrt{1+2h\lambda}}{3-2h\lambda}\xrightarrow[h\lambda\to\frac{3}{2}]{}+\infty \quad \text{(stability not guaranteed)}$$

$$z_2=\frac{\left(2-\sqrt{1+2h\lambda}\right)\left(2+\sqrt{1+2h\lambda}\right)}{(3-2h\lambda)\left(2+\sqrt{1+2h\lambda}\right)}=\frac{3-2h\lambda}{(3-2h\lambda)\left(2+\sqrt{1+2h\lambda}\right)}=\frac{1}{2+\sqrt{1+2h\lambda}}\xrightarrow[h\lambda\to\frac{3}{2}]{}\frac{1}{4}$$

Problem: The forward Euler method is to be used for solving the following system of ordinary differential equations:

$$y_1'(t)=cy_1(t)+y_2(t)+e(t)$$
$$y_2'(t)=y_1(t)+cy_2(t)$$

Determine the interval of values of the real-valued parameter c for which the absolute stability of the numerical solution may be guaranteed for the integration step $h<1$.

Solution: The eigenvalues λ_1 and λ_2 of the transition matrix $\mathbf{A}=\begin{bmatrix} c & 1 \\ 1 & c \end{bmatrix}$ satisfy the following algebraic equation:

$$\det(\mathbf{A}-\lambda\mathbf{I})=\det\left(\begin{bmatrix} c-\lambda & 1 \\ 1 & c-\lambda \end{bmatrix}\right)=(c-\lambda)^2-1=0$$

Its solutions are: $\lambda_1=c-1$ and $\lambda_2=c+1$. Since the interval of stability for the forward Euler method is $(-2, 0)$, the following inequalities should be jointly satisfied:

$$\lambda_1 h\in(-2, 0) \quad \text{and} \quad \lambda_2 h\in(-2, 0)$$

or:

$$c-1\in\left(-\frac{2}{h}, 0\right) \quad \text{and} \quad c+1\in\left(-\frac{2}{h}, 0\right)$$

or:

$$c \in \left(1-\frac{2}{h},1\right) \text{ and } c \in \left(-1-\frac{2}{h},-1\right)$$

which means that the absolute stability cannot be guaranteed for any value of the parameter c.

Problem: The following method (called *TM2* hereinafter):

$$y_n = \left[y(t) + hf\left(t, y(t)\right) + \frac{1}{2}h^2 \frac{d}{dt} f\left(t, y(t)\right) \right]_{t=t_{n-1}}$$

is designed for solving ordinary differential equations of the form:

$$y'(t) = f\left(t, y(t)\right)$$

Determine the order of the TM2. Assess the local error of the solution obtained by means of the TM2. Check whether the points (–1,0) and (0, 1) belong to the region of absolute stability of the TM2.

<u>Solution:</u> For a single step of integration, the TM2 may be rewritten in the abridged notation as:

$$y_n = \dot{y}_{n-1} + h\dot{y}'_{n-1} + \frac{1}{2}h^2 \dot{y}''_{n-1}$$

After the development of the RHS into Taylor series, the above formula takes on the form:

$$y_n = \dot{y}_n - h\dot{y}'_n + \frac{1}{2}h^2\dot{y}''_n - \frac{1}{6}h^3\dot{y}'''_n + \ldots$$

$$+h\dot{y}'_n - \quad h^2\dot{y}''_n + \frac{1}{2}h^3\dot{y}'''_n - \ldots$$

$$+\frac{1}{2}h^2\dot{y}''_{n-1} - \frac{1}{2}h^3\dot{y}'''_n + \ldots = -\frac{1}{6}h^3\dot{y}''' + \ldots$$

Hence: $p = 2$ and $r_n \cong -\frac{1}{6}h^3\dot{y}'''$. The application of the TM2 formula (in the abridged notation) to the test equation $y'(t) = \lambda y(t)$ yields:

$$y_n = \dot{y}_{n-1} + h\lambda\dot{y}_{n-1} + \frac{1}{2}h^2\lambda^2\dot{y}_{n-1} = \left(1 + h\lambda\dot{y}_{n-1} + \frac{1}{2}h^2\lambda^2\right)\dot{y}_{n-1}$$

which means that the condition of the absolute stability has the form:

$$\left|1 + h\lambda + \frac{1}{2}h^2\lambda^2\right| < 1$$

This inequality is satisfied for $h\lambda = -1$ because $\left|1-1+\frac{1}{2}\right| < 1$, and not for $h\lambda = j$ because:

$$\left|1 + j - \frac{1}{2}\right| = \left|\frac{1}{2} + j\right| > 1$$

5.5. Other problems

Problem: The following ODE:

$$y'(t) = -y(t) + 1; \quad y(0) = 0$$

has been solved by means of the explicit Euler method: $y_n = y_{n-1} + h \cdot f\left(t_{n-1}, y_{n-1}\right)$. Assess the local error of the solution for $n = 10$ (r_{10}) if $h = 0.01\,h_{\max}$ where $h = 0.01\,h_{\max}$ is the maximum step guaranteeing the absolute stability of the solution.

Solution: The development of the RHS of the explicit Euler scheme in the Taylor series yields:

$$y_n = \left(\dot{y}_n - \dot{y}_n^{(1)} h + \frac{1}{2} \dot{y}_n^{(2)} h^2 + \ldots \right) + h\left(\dot{y}_n^{(1)} - \dot{y}_n^{(2)} h + \ldots \right) = \dot{y}_n - \frac{1}{2} \dot{y}_n^{(2)} h^2 + \ldots$$

Hence the estimate of its local error:

$$r_n \cong -\frac{1}{2} \dot{y}_n^{(2)} h^2$$

The needed derivative may be determined using the analytical solution of the ODE under consideration. Since the homogeneous equation $y'(t) = -y(t)$ has the solution of the form $y(t) = C \cdot e^{-t}$, the general solution should have the form $y(t) = C(t) \cdot e^{-t}$ which implies:

$$y'(t) = C'(t) \cdot e^{-t} - C(t) \cdot e^{-t}$$

The function $C(t)$ may be determined by substituting $y(t)$ and $y'(t)$ to the original equation:

$$C'(t) \cdot e^{-t} - C(t) \cdot e^{-t} = -C(t) \cdot e^{-t} + 1$$

Hence:

$$C'(t) = e^t \quad \text{and} \quad C(t) = \int e^t dt = e^t + C_0$$

where C_0 is a constant to be determined using the initial value of the solution $y(0) = 0$:

$$C(0) \cdot e^{-0} = 0 \;\Rightarrow\; 1 + C_0 - 0 \Rightarrow\; C_0 = -1$$

Thus $C(t) = e^t - 1$; consequently:

$$y(t) = C(t) \cdot e^{-t} = \left(e^t - 1\right) \cdot e^{-t} = 1 - e^{-t}$$

and:

$$r_n \cong -\frac{1}{2} \dot{y}_n^{(2)} h^2 = \frac{1}{2} e^{-nh} h^2 > 0$$

The maximum step guaranteeing the absolute stability of the solution is $h_{\max} = 2$ because the explicit Euler method when applied to the homogeneous equation $y'(t) = -y(t)$ yields:

$$y_n = y_{n-1} - h \cdot y_{n-1} = (1 - h) \cdot y_{n-1}$$

and, therefore, the stability is guaranteed if $|1-h| < 1$ or $h < 2$. For $h = 0.01\, h_{\max} = 0.02$ the local error r_{10} may be assessed as follows:

$$r_{10} \cong -\frac{1}{2} \dot{y}_{10}^{(2)} h^2 = \frac{1}{2} e^{-10h} h^2 \cong \frac{1}{2}(1 - 10h) h^2 = \frac{1}{2}(1 - 0.2) 0.0004 = 1.6 \cdot 10^{-4}$$

Problem: Assess the local error of the following method for solving ODEs:

$$y_n = y_{n-2} + 2hf\left(x_{n-1}, y_{n-1}\right)$$

Check its stability for $h\lambda = 1$.

Solution: The Taylor series of RHS is:

$$RHS = \left(\dot{y}_n - 2\dot{y}_n' h + 2\dot{y}_n'' h^2 - \frac{4}{3} \dot{y}_n''' h^3 + \ldots \right) + 2h\left(\dot{y}_n' - \dot{y}_n'' h + \frac{1}{2} \dot{y}_n''' h^2 + \ldots \right)$$

$$RHS = \dot{y}_n - 2\dot{y}_n' h + 2\dot{y}_n'' h^2 - \frac{4}{3} \dot{y}_n''' h^3 + \ldots + 2\dot{y}_n' h - 2\dot{y}_n'' h^2 + \dot{y}_n''' h^3 + \ldots = \dot{y}_n - \frac{4}{3} \dot{y}_n''' h^3 + y_n''' h^3 + \ldots$$

$$\Rightarrow e_n \cong -\frac{1}{3} y_n''' h^3$$

The stability condition is resulting from the following implication:

$$\hat{y}_n = y_{n-2} + 2h\lambda y_{n-1} \Rightarrow 1 = z^{-2} + 2h\lambda z^{-1} \Rightarrow z^2 - 2h\lambda z - 1 = 0 \Rightarrow z^2 - 2z - 1 = 0 \Rightarrow z_{1/2} = 1 \pm \sqrt{2}$$

Thus, the method is not stable for $h\lambda = 1$.

Problem: For the following method of ODEs integration:

$$y_n = y_{n-1} + \frac{h}{2}\left(y_n' + y_{n-1}'\right)$$

Determine the indicators of local accuracy, the region of absolute stability and the one-step rounding error in the solution of the test equation, obtained for $h \to 0$.

Solution: The order of the method and an estimate of the local error may be determined in the following way:

$$y_n = \dot{y}_{n-1} + \frac{h}{2}\left(\dot{y}_n^{(1)} + \dot{y}_{n-1}^{(1)}\right) = \dot{y}_n + \dot{y}_n^{(1)}(-h) + \frac{1}{2}\dot{y}_n^{(2)}(-h)^2 + \frac{1}{6}\dot{y}_n^{(3)}(-h)^3 + \ldots$$

$$+\frac{h}{2}\left[\dot{y}_n^{(1)} + \dot{y}_n^{(1)} + \dot{y}_n^{(2)}(-h) + \frac{1}{2}\dot{y}_n^{(3)}(-h)^2 + \ldots\right]$$

$$y_n = \dot{y}_n - \dot{y}_n^{(1)}h + \frac{1}{2}\dot{y}_n^{(2)}h^2 - \frac{1}{6}\dot{y}_n^{(3)}h^3 + \ldots + \dot{y}_n^{(1)}h - \frac{1}{2}\dot{y}_n^{(2)}h^2 + \frac{1}{4}\dot{y}_n^{(3)}h^3 \ldots = \dot{y}_n + \frac{1}{12}\dot{y}_n^{(3)}h^3 + \ldots$$

$$y_n - \dot{y}_n = \frac{1}{12}\dot{y}_n^{(3)}h^3 + \ldots$$

The region of absolute stability is defined by the condition:

$$\left|\frac{1+\frac{\lambda h}{2}}{1-\frac{\lambda h}{2}}\right| \le 1$$

because:

$$y_n = y_{n-1} + \frac{\lambda h}{2}\left(y_n + y_{n-1}\right) \Rightarrow y_n = \frac{1+\frac{\lambda h}{2}}{1-\frac{\lambda h}{2}} y_{n-1}$$

This condition is satisfied in the whole left half-plane of λh because:

$$|2+\lambda h| \le |2-\lambda h| \Rightarrow |2+a+jb| \le |2-a-jb| \Rightarrow |2+a+jb|^2 \le |2-a-jb|^2$$

$$\Rightarrow (2+a)^2 + b^2 \le (2-a)^2 + b^2 \Rightarrow 4a \le -4a \Rightarrow a \le 0$$

where $a = \mathrm{Re}(\lambda h)$ and $b = \mathrm{Im}(\lambda h)$. Alternatively, the method of $\mathscr{Z}$-transform may be applied:

$$\mathscr{Z}\{y_n\} = \mathscr{Z}\left\{\frac{1+\frac{\lambda h}{2}}{1-\frac{\lambda h}{2}} y_{n-1}\right\} \Rightarrow 1 = \frac{1+\frac{\lambda h}{2}}{1-\frac{\lambda h}{2}} z^{-1} \Rightarrow \lambda h = 2\frac{z-1}{z+1}$$

Thus, the border of the region of absolute stability is determined by the equation:

$$\lambda h = 2\left.\frac{z-1}{z+1}\right|_{z=\exp(j\phi)} = 2\frac{\exp(j\phi)-1}{\exp(j\phi)+1} = j2\frac{\sin(\phi)}{1+\cos(\phi)}$$

which means that the $\mathrm{Im}[\lambda h]$ axis is the border of absolute stability.

The one-step rounding error in the solution of the test equation (for $h \to 0$)may be assessed as follows:

$$\tilde{y}_n = \frac{\left[1+0.5\left(1+\eta_w\right)\right]\left(1+\eta_s\right)}{\left[1-0.5\left(1+\eta_w\right)\right]\left(1+\eta_o\right)}\left(1+\eta_d\right)y_{n-1}\left(1+\eta_m\right) = y_{n-1}\frac{\left[1.5+0.5\eta_w\right]}{\left[0.5-0.5\eta_w\right]}\left(1+\eta_s-\eta_o+\eta_d+\eta_m\right)$$

where: $\tilde{w} = \frac{\lambda h}{2}\left(1+\eta_d'+\eta_m'\right) \Rightarrow \left|\eta_w\right| = \left|\delta\left[\tilde{w}\right]\right| \le 2eps$

$$\tilde{y}_n = 3y_{n-1}\frac{\left[1+\frac{1}{3}\eta_w\right]}{\left[1-\eta_w\right]}\left(1+\eta_s-\eta_o+\eta_d+\eta_m\right) = 3y_{n-1}\left(1+\frac{4}{3}\eta_w+\eta_s-\eta_o+\eta_d+\eta_m\right)$$

$$\left|\delta\left[\tilde{y}_n\right]\right| = \left|\frac{4}{3}\eta_w+\eta_s-\eta_o+\eta_d+\eta_m\right| \le \frac{4}{3}\cdot 2eps+4eps = 6\frac{2}{3}eps$$

Problem: Apply the implicit Euler method:

$$y_{n+1} = y_n + h\cdot f\left(t_{n+1}, y_{n+1}\right)$$

to the following nonlinear ODE:

$$y'(t) = -0.5\cdot y^3(t);\ y(0)=1$$

Design an iterative Newton-method-based algorithm for computing y_{n+1} on the basis of y_n.

<u>Solution:</u> The application of the implicit Euler method yields:

$$y_{n+1} = y_n - \frac{h}{2}\cdot y_{n+1}^3$$

Thus, at each step, $t_n \to t_{n+1}$, the following nonlinear algebraic equation is to be solved:

$$F\left[y_{n+1}^{(i)}\right] \equiv y_{n+1}^{(i)} - y_n + \frac{h}{2}\cdot\left[y_{n+1}^{(i)}\right]^3 = 0$$

with respect to $y_{n+1}^{(i)}$. The use of the Newton method for this purpose yields:

$$y_{n+1}^{(i+1)} = y_{n+1}^{(i)} - \frac{F\left[y_{n+1}^{(i)}\right]}{F'\left[y_{n+1}^{(i)}\right]} \quad \text{for } i = 0, 1, \ldots$$

where $F'\left[y_{n+1}^{(i)}\right] = 1+\frac{3h}{2}\cdot\left[y_{n+1}^{(i)}\right]^2$. Thus, the desired algorithm has the form

$$y_{n+1}^{(i+1)} = \frac{h\cdot\left[y_{n+1}^{(i)}\right]^3 + y_n}{1+\frac{3h}{2}\cdot\left[y_{n+1}^{(i)}\right]^2} \quad \text{for } i = 0, 1, \ldots$$

Problem: The forward Euler method has been used for solving an ODE whose exact solution is $y(t) = e^{-t}$. Assess the maximum size of the integration step, $h_{\max}(t)$, guaranteeing the local error of the numerical solution smaller than $0.5\cdot 10^{-6}$.

<u>Solution:</u> The local error for the forward Euler method is assessed according to the formula:

$$r_n(h) \cong -0.5\dot{y}_n''h^2$$

In the considered case:

$$\dot{y}_n'' = y''(t_n) = \exp(-t_n)$$

consequently:

$$\left|r_n(h)\right| \cong 0.5\exp(-t_n)h^2$$

This error is smaller than $0.5\cdot 10^{-6}$ if:

$$0.5\exp\left(-t_n\right)h^2 < 0.5\cdot 10^{-6}$$

i.e. for $h < 10^{-3}\exp\left(\frac{t_n}{2}\right) \equiv h_{\max}\left(t_n\right)$.

Problem: Solve the following equation:

$$y'(t) = -y(t) + \sin(t); \quad y(0) = 0$$

using the method of variable constants. Under the assumption that the forward Euler method is to be used for solving this equation, assess the maximum size of the integration step, $h_{\max}(t)$, guaranteeing the local error of the numerical solution smaller than 10^{-6}.

Solution: The solution of the homogenous equation $y'(t) = -y(t)$ has the form $y(t) = C\cdot e^{-t}$. Thus, the general solution of the non-homogenous equation has the form:

$$y(t) = C(t)\cdot e^{-t} \quad\Rightarrow\quad y'(t) = C'(t)\cdot e^{-t} - C(t)\cdot e^{-t}$$

By substituting this solution and its derivative to the original equation, we obtain:

$$C'(t)\cdot e^{-t} - C(t)\cdot e^{-t} = -C(t)\cdot e^{-t} + \sin(t)$$

or:

$$C'(t) = e^{t}\cdot\sin(t) \;\Rightarrow\; C(t) = \int e^{t}\cdot\sin(t)\,dt = \frac{1}{2}e^{t}\left[\sin(t) - \cos(t)\right] + C_0$$

where the constant C_0 should satisfy the initial condition $y(0) = 0$:

$$C(0)\cdot e^{-0} = 0 \Rightarrow C(0) = 0 \Rightarrow C_0 = \frac{1}{2}$$

Thus, the solution of the original equation has the form:

$$y(t) = \left\{\frac{1}{2}e^{t}\left[\sin(t) - \cos(t)\right] + \frac{1}{2}\right\}e^{-t} = \frac{1}{2}e^{-t} + \frac{1}{2}\left[\sin(t) - \cos(t)\right] = \frac{1}{2}e^{-t} + \sqrt{2}\sin\left(t + \frac{\pi}{4}\right)$$

The local error for the forward Euler method is assessed according to the formula:

$$r_n(h) \cong -\frac{1}{2}\dot{y}_n'' h^2$$

In the considered case:

$$\dot{y}_n'' \equiv y''(t_n) = \frac{1}{2}\exp(-t_n) - \sqrt{2}\sin\left(t_n + \frac{\pi}{4}\right)$$

consequently:

$$\left|r_n(h)\right| \cong \left|\frac{1}{4}\exp(-t_n) - \frac{1}{\sqrt{2}}\sin\left(t_n + \frac{\pi}{4}\right)\right| h^2$$

This error is smaller than 10^{-6} if:

$$\left|\frac{1}{4}\exp(-t_n) - \frac{1}{\sqrt{2}}\sin\left(t_n + \frac{\pi}{4}\right)\right| h^2 < 10^{-6}$$

i.e. for:

$$h < \frac{10^{-3}}{\sqrt{\left|\frac{1}{4}\exp(-t_n) - \frac{1}{\sqrt{2}}\sin\left(t_n + \frac{\pi}{4}\right)\right|}} \equiv h_{\max}(t_n)$$