MATH-4800 **NUMERICAL COMPUTING** **Summer 2023**

Assignment 1, due before class, Wednesday May 31, 2023.

Haowen He heh4@rpi.edu

1. Write a `Matlab` code to evaluate a polynomial using Horner's method as described on page 3 of the text (and discussed in class). Use your code to evaluate the polynomials in Text Exercises $1(b), 2(c)$ and $4(a)$ on page 5.

```
%Program 0.1 Nested multiplication
%Evaluates polynomial from nested form using Horner's Method
%Input: degree d of polynomial,
%       array of d+1 coefficients c (constant term first),
%       x-coordinate x at which to evaluate, and
%       array of d base points b, if needed
%Output: value y of polynomial at x
function y=nest(d,c,x,b)
if nargin<4, b=zeros(d,1); end
y=c(d+1);
for i=d:-1:1
  y = y.*(x-b(i))+c(i);
end
```

```
>> nest(4,[1 -5 5 4 -3],1/3)

ans =

     0

>> nest(6,[4 -2 0 0 -2 0 4],-1/2)

ans =

    4.9375

>> nest(3,[1 1/2 1/2 -1/2],5,[0 2 3])

ans =

    -4
```

2. Text Exercises 1(a) and 1(b) on page 20, and Computer Problem 1 on page 20.

$\frac{1-\sec x}{\tan^2 x}$: There is subtraction of nearly equal numbers when $1-\sec x \approx 0$. Setting $\sec x = 1$, we have $\boxed{x = 2k\pi}$ for $k \in \mathbb{Z}$. Reformulating by using the trig identity $\sec^2 x = 1 + \tan^2 x$, we have that

$$\frac{1-\sec x}{\tan^2 x} \cdot \frac{1+\sec x}{1+\sec^2 x} = \frac{\cancelto{-\tan^2 x}{1-\sec^2 x}}{\tan^2 x(1+\sec x)} = \boxed{-\frac{1}{1+\sec x}}$$

avoiding the cancellation problem for x near $2k\pi$.

$\frac{1-(1-x)^3}{x}$: There is substraction of nearly equal numbers when $1-(1-x)^3 \approx 0$. Setting $(1-x)^3 = 1$, we have $\boxed{x = 0}$. Simplifying the expression yields

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

$$= \frac{x^3-3x^2+3x}{x}$$

$$= \boxed{x^2-3x+3}$$

avoiding the cancellation problem for x near 0.

$$\frac{1-\sec x}{\tan^2 x} \qquad -\frac{1}{1+\sec x}$$

x	original	revised
0.10000000000000	−0.49874791371143	−0.49874791371143
0.01000000000000	−0.49998749979096	−0.49998749979166
0.00100000000000	−0.49999987501429	−0.49999987499998
0.00010000000000	−0.49999999362793	−0.49999999875000
0.00001000000000	−0.50000004133685	−0.49999999998750
0.00000100000000	−0.50004445029084	−0.49999999999987
0.00000010000000	−0.51070259132757	−0.50000000000000
0.00000001000000	0	−0.50000000000000
0.00000000100000	0	−0.50000000000000
0.00000000010000	0	−0.50000000000000
0.00000000001000	0	−0.50000000000000
0.00000000000100	0	−0.50000000000000
0.00000000000010	0	−0.50000000000000
0.00000000000001	0	−0.50000000000000

$$\frac{1-(1-x)^3}{x} \qquad x^2-3x+3$$

x	original	revised
0.10000000000000	2.71000000000000	2.71000000000000
0.01000000000000	2.97010000000001	2.97010000000000
0.00100000000000	2.99700100000000	2.99700100000000
0.00010000000000	2.99970000999905	2.99970001000000
0.00001000000000	2.99997000008379	2.99997000010000
0.00000100000000	2.99999700015263	2.99999700000100
0.00000010000000	2.99999969866072	2.99999970000001
0.00000001000000	2.99999998176759	2.99999997000000
0.00000000100000	2.99999991515421	2.99999999700000
0.00000000010000	3.00000024822111	2.99999999970000
0.00000000001000	3.00000024822111	2.99999999997000
0.00000000000100	2.99993363483964	2.99999999999700
0.00000000000010	3.00093283556180	2.99999999999970
0.00000000000001	2.99760216648792	2.99999999999997

3. Let $\tilde{f}(x)$ approximate $f(x)$, where

$$f(x) = \frac{1}{\sqrt{3+x}}, \qquad \tilde{f}(x) = \frac{1}{2} - \frac{x-1}{16}.$$

Find the absolute forward and backward errors in the approximation when

(a) $x = 0.9$,

(b) $x = 1.2$.

(a) Compute $f(0.9) = \frac{1}{\sqrt{3+0.9}} = 0.50637$

$\hat{f}(0.9) = \frac{1}{2} - \frac{0.9-1}{16} = 0.50625 = \hat{y}$

Forward absolute error $= |\hat{f}(0.9) - f(0.9)| = \boxed{0.00012}$.

Let $\hat{x}$ solve $\hat{y} = f(\hat{x})$, then $0.50625 = \frac{1}{\sqrt{3+\hat{x}}}$

$\hat{x} = \frac{1}{(0.50625)^2} - 3$

$= 0.90184$

Backward absolute error $= |\hat{x} - x| = \boxed{0.00184}$.

(b) Compute $f(1.2) = \dfrac{1}{\sqrt{3+1.2}} = 0.48795$

$$\hat{f}(1.2) = \frac{1}{2} - \frac{1.2-1}{16} = 0.48750 = \hat{y}$$

Forward absolute error $= |\hat{f}(1.2) - f(1.2)| = \boxed{0.00045}$.

Let $\hat{x}$ solve $\hat{y} = f(\hat{x})$, then $0.48750 = \dfrac{1}{\sqrt{3+\hat{x}}}$

$$\hat{x} = \frac{1}{(0.48750)^2} - 3$$

$$= 1.207758$$

Backward absolute error $= |\hat{x} - x| = \boxed{0.007758}$.

4. The roots x_1 and x_2 of the quadratic equation $x^2 + 2bx + c = 0$ may be computed using the results of the usual quadratic formula

$$x_1 = -b - \sqrt{b^2 - c}, \qquad x_2 = -b + \sqrt{b^2 - c}.$$

(a) Use 6-digit, base 10, floating-point arithmetic to compute the two roots using the formulas above for the cases (i) $b = 1.23456 \times 10^6, c = 9.87654 \times 10^8$ and (ii) $b = -2.46864 \times 10^{-2}, c = 1.35753 \times 10^{-8}$. Compute the "exact" values using full precision on a calculator or using `Matlab`. Determine the relative (forward) errors in the roots computed using 6-digit arithmetic. Are the 6-digit roots accurate? Explain the results in terms of round-off error.

(b) Write a `Matlab` function, called `myRoots` say, that takes as input the values of b and c in the quadratic equation above and returns x_1 and x_2 computed accurately. Base the calculation of the roots in your Matlab function on the formulas for x_1 and x_2 above, or the alternate forms.

$$x_1 = \frac{c}{-b + \sqrt{b^2 - c}}, \qquad x_2 = \frac{c}{-b - \sqrt{b^2 - c}}.$$

Your `Matlab` function should check the input values and decide which formulas give the most accurate results. Run your code for the two cases given in part (a).

(a) Quadratic formula $\Rightarrow$ $x_1 = -b - \sqrt{b^2 - c}$

$x_2 = -b + \sqrt{b^2 - c}$

(i) $b = 1.23456 \times 10^6$, $c = 9.87654 \times 10^8$

use the result of rounding in the next step till the end of computation $\Rightarrow$

$(b^2)_c = 1.52414 \times 10^{12}$

$(b^2 - c)_c = 1.52315 \times 10^{12}$

$(\sqrt{b^2 - c})_c = 1.23416 \times 10^6$

$$\boxed{\begin{aligned} (x_1)_c &= -2.46872 \times 10^6 \\ x_1 &= -2.46872 \times 10^6 \\ r_1 &= 0 \\ (x_2)_c &= -4 \times 10^2 \\ x_2 &= -400.067 \\ r_2 &\approx 1.67 \times 10^{-4} \end{aligned}}$$

//

Quadratic formula $\Rightarrow$ $x_1 = -b - \sqrt{b^2 - c}$

$x_2 = -b + \sqrt{b^2 - c}$

(ii) $b = -2.46864 \times 10^{-2}$, $c = 1.35753 \times 10^{-8}$

$$(b^2)_c = 6.09418 \times 10^{-4}$$

$$(b^2 - c)_c = 6.09404 \times 10^{-4}$$

$$(\sqrt{b^2 - c})_c = 2.46861 \times 10^{-2}$$

$$\boxed{\begin{aligned} (x_1)_c &= 3 \times 10^{-7} \\ x_1 &= 2.74957 \times 10^{-7} \\ r_1 &\approx 9.11 \times 10^{-2} \\ (x_2)_c &= 4.93725 \times 10^{-2} \\ x_2 &= 4.93725 \times 10^{-2} \\ r_2 &= 0 \end{aligned}}$$

//

Nonzero errors occur when we add two numbers with opposite sign : cancellation of the significant digits leaves the accumulation of the previous steps' round-off error as the result.

```
function x=myRoots(b,c)
%
% compute distinct real roots of the quadratic x^2+2*b*x+c=0
% using the given quadratic formula or the alternative
% based on the input value b
%

% calculate the discrimiant
discriminant=b^2-c;

% check the discriminant to make sure the roots are real
if discriminant<=0
    disp('Error: discriminant must be positive')
    x=[0 0];
    return
end

% By using the result from part A %
%
% if b positive
% -b and -sqrt(discriminant) have the same sign
% -b and sqrt(discriminant) have opposite sign
% thus we find x1 using the given formula
% and find x2 using the alternative
%
if b>0
    x1=-b-sqrt(discriminant);
    x2=c/x1;
%
% if b negative
% -b and -sqrt(discriminant) have opposite sign
% -b and sqrt(discriminant) have the same sign
% thus we find x2 using the given formula
% and find x1 using the alternative
%
else
    x2=-b+sqrt(discriminant);
    x1=c/x2;
end

% save the roots in the vector x
x=[x1 x2];
```

compare with results from part (a)

```
>> b=1.23456*10^6; c=9.87654*10^8;
>> [x1,x2]=myRoots(b,c)

x1 =

  -2.4687e+06


x2 =

 -400.0673


>> b=-2.46864*10^-2; c=1.35753*10^-8;
>> [x1,x2]=myRoots(b,c)

x1 =

   2.7496e-07


x2 =

    0.0494
```

5. Text Exercise 6 on page 24.

(a) $$P_4(x) = f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2}f''(x_0)(x-x_0)^2 + \frac{1}{6}f'''(x_0)(x-x_0)^3 + \frac{1}{24}f^{(4)}(x_0)(x-x_0)^4$$

$\Rightarrow$ $f(x) = x^{-2}$ and $x_0 = 1$

The first four derivatives of $f(x) = x^{-2}$ are

$f'(x) = -2x^{-3}$, $f''(x) = 6x^{-4}$, $f'''(x) = -24x^{-5}$, $f^{(4)}(x) = 120x^{-6}$.

At $x_0 = 1$, these derivatives are

$f(1) = 1$, $f'(1) = -2$, $f''(1) = 6$, $f'''(1) = -24$, $f^{(4)}(1) = 120$.

Substituting them into the formula for $P_4(x)$ yields

$$\boxed{P_4(x) = 1 - 2(x-1) + 3(x-1)^2 - 4(x-1)^3 + 5(x-1)^4}.$$

(b) $$P_4(0.9) = 1 - 2(0.9-1) + 3(0.9-1)^2 - 4(0.9-1)^3 + 5(0.9-1)^4 = \boxed{1.2345}.$$

$$P_4(1.1) = 1 - 2(1.1-1) + 3(1.1-1)^2 - 4(1.1-1)^3 + 5(1.1-1)^4 = \boxed{0.8265}.$$

(c) By the Lagrange's form of the remainder, the error bound is given by $\dfrac{6|x-1|^5}{c^7}$. To maximize error, we minimize the denominator by setting $c = 0.8$. Thus, for $x = 0.8$, the error is bounded by $\boxed{1.25445 \times 10^{-4}}$, and for $x = 1.1$, the error bound remains the same.

We expect the approximation to be $\boxed{\text{better for } x > 1}$ because the derivatives increase greatly as $x \to 0$.

(d) Actual error for $x = 0.8$: $1.23456790 1 - 1.2345$

$= 6.7901 \times 10^{-5}$ smaller than

$< 1.25445 \times 10^{-4}$ the error bound

Actual error for $x = 1.1$: $0.8265 - 0.82644628$

$= 5.372 \times 10^{-5}$ smaller than

$< 1.25445 \times 10^{-4}$ the error bound

MATH-4800 **NUMERICAL COMPUTING** **Summer 2023**

Assignment 2, due before class, Thursday June 8, 2023.

Haowen He heh4@rpi.edu

1. Let $P(x) = x^3 - 2x^2 - 5x + 8$.

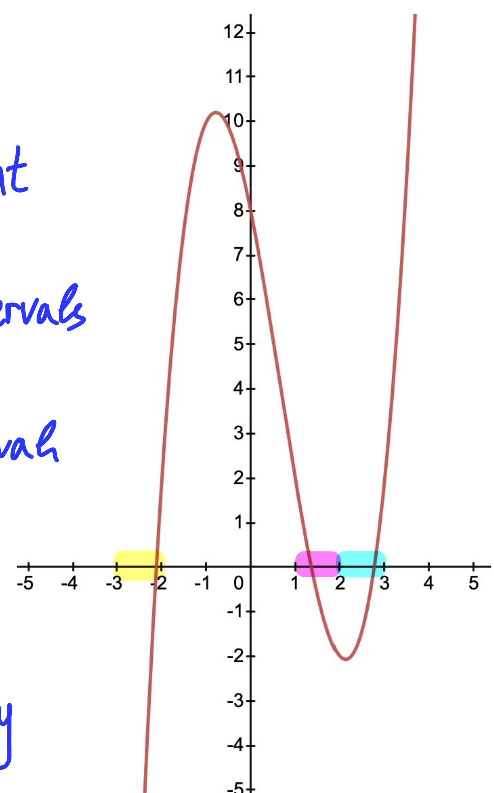
   (a) Show analytically (following the work in class) that $P(x)$ has three real roots, and find intervals of x that bracket each root individually.

   (b) Write a `Matlab` function that determines the root of a function using the bisection method. Use your function to determine the three roots of $P(x)$. For each root, use the stopping criterion $b - a < 10^{-6}$, where $[a, b]$ is the interval of the bracket at the n^{th} iteration. (Hint: see the example Matlab codes on LMS.)

   (c) Repeat the work in part (b), but now using the method of false position (as discussed in class). For each root, use the stopping criterion $|x_{n+1} - x_n| < 10^{-6}$, where x_n and x_{n+1} are successive approximations of the root during the iteration.

(a) Consider the cubic polynomial $P(x) = x^3 - 2x^2 - 5x + 8$.

It follows that $P(x)$ is continuous everywhere and has at most three distinct roots. We want to show that $P(x)$ changes sign on some intervals so that $P(x)$ must have a root on that interval by the intermediate value theorem. From the plot of $P(x)$ on the right, we can identify three such intervals:



check: setting $P'(x) = 3x^2 - 4x - 5 = 0$

$\Rightarrow x_1 \approx -0.8 \; , \; x_2 \approx 2.1$

$P''(x) = 6x - 4 \Rightarrow P''(x_1) < 0 \;\&\; P''(x_2) > 0$

- $[-3, -2]$ $\quad P(-3) = -22 < 0, \; P(-2) = 2 > 0$

- $[1, 2]$ $\quad P(1) = 2 > 0, \; P(2) = -2 < 0$

- $[2, 3]$ $\quad P(2) = -2 < 0, \; P(3) = 2 > 0$

Therefore, $P(x)$ has three distinct real roots, and there is a unique root in each of the three intervals above.

```
function x=myBisection(f,xspan,tol)

% Find a root x of the equation f(x)=0 using the
% bisection method.  The initial interval is xspan=[a,b]
% and the iteration stops when b-a<tol.

a=xspan(1);
b=xspan(2);

fa=feval(f,a);
if fa==0
  x=a; return
end
fb=feval(f,b);
if fb==0
  x=b; return
end
if fa*fb>0
  error('Error : f(a)*f(b)>0')
end

n=0;
while b-a>tol
  n=n+1;
  x=(a+b)/2; fx=feval(f,x);
  fprintf('n=%d  a=%1.8f b=%1.8f f(x)=%1.2e\n',n,a,b,fx)
  if fx==0
    return
  end
  if fx*fa<0
    b=x; fb=fx;
  else
    a=x; fa=fx;
  end
end
```

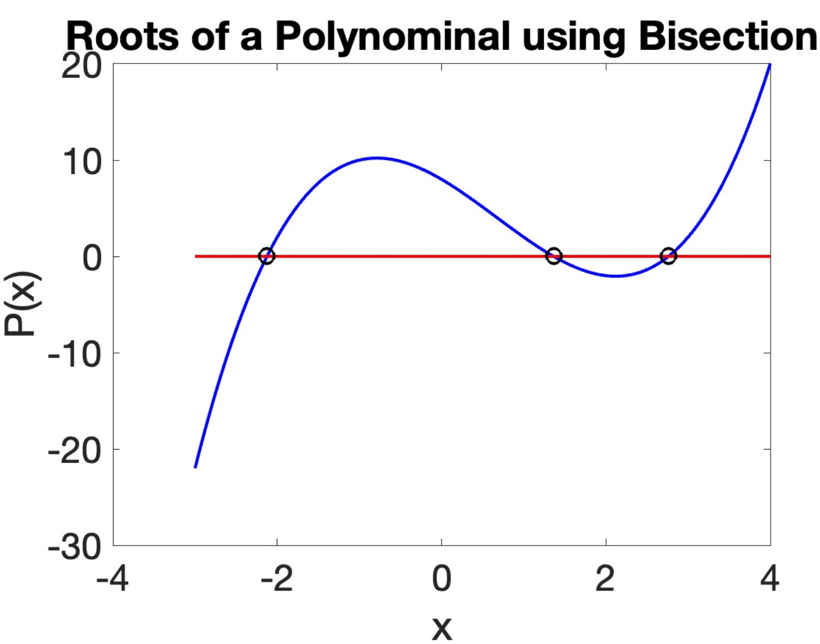
myBisection.m

from in-class example

use the midpoint of the interval to update the endpoints

```
% clear all variables and figures.
clear all
close all

% Set defaults for plotting
fontSize=24; lineWidth=2; markerSize=10;
set(0,'DefaultLineMarkerSize',markerSize);
set(0,'DefaultLineLineWidth',lineWidth);
set(0,'DefaultAxesFontSize',fontSize);
set(0,'DefaultLegendFontSize',fontSize);

% define and plot the polynomial y=x.^3-2*x.^2-5*x+8
P = @(x) x.^3-2*x.^2-5*x+8;
x=linspace(-3,4,300);
y=P(x);
plot(x,y,'b',x,zeros(size(x)),'r')

% Root on [-3,-2]
fprintf('Iteration for root on [-3,-2] :\n')
xspan=[-3,-2]; tol=10e-6;
x1=myBisection(P,xspan,tol);

% Root on [1,2]
fprintf('Iteration for root on [1,2] :\n')
xspan=[1,2]; tol=10e-6;
x2=myBisection(P,xspan,tol);

% Root on [2,3]
fprintf('Iteration for root on [2,3] :\n')
xspan=[2,3]; tol=10e-6;
x3=myBisection(P,xspan,tol);

% plot roots
hold on
roots=[x1 x2 x3];
plot(roots,zeros(3,1),'ko')
hold off
xlabel('x')
ylabel('P(x)')
title('Roots of a Polynominal using Bisection')
```

```
Iteration for root on [-3,-2] :
n=1  a=-3.00000000 b=-2.00000000 f(x)=-7.62e+00
n=2  a=-2.50000000 b=-2.00000000 f(x)=-2.27e+00
n=3  a=-2.25000000 b=-2.00000000 f(x)=-1.95e-03
n=4  a=-2.12500000 b=-2.00000000 f(x)=1.03e+00
n=5  a=-2.12500000 b=-2.06250000 f(x)=5.23e-01
n=6  a=-2.12500000 b=-2.09375000 f(x)=2.62e-01
n=7  a=-2.12500000 b=-2.10937500 f(x)=1.31e-01
n=8  a=-2.12500000 b=-2.11718750 f(x)=6.45e-02
n=9  a=-2.12500000 b=-2.12109375 f(x)=3.13e-02
n=10  a=-2.12500000 b=-2.12304688 f(x)=1.47e-02
n=11  a=-2.12500000 b=-2.12402344 f(x)=6.37e-03
n=12  a=-2.12500000 b=-2.12451172 f(x)=2.21e-03
n=13  a=-2.12500000 b=-2.12475586 f(x)=1.28e-04
n=14  a=-2.12500000 b=-2.12487793 f(x)=-9.13e-04
n=15  a=-2.12493896 b=-2.12487793 f(x)=-3.93e-04
n=16  a=-2.12490845 b=-2.12487793 f(x)=-1.32e-04
n=17  a=-2.12489319 b=-2.12487793 f(x)=-2.37e-06

Iteration for root on [1,2] :
n=1  a=1.00000000 b=2.00000000 f(x)=-6.25e-01
n=2  a=1.00000000 b=1.50000000 f(x)=5.78e-01
n=3  a=1.25000000 b=1.50000000 f(x)=-5.66e-02
n=4  a=1.25000000 b=1.37500000 f(x)=2.53e-01
n=5  a=1.31250000 b=1.37500000 f(x)=9.63e-02
n=6  a=1.34375000 b=1.37500000 f(x)=1.93e-02
n=7  a=1.35937500 b=1.37500000 f(x)=-1.88e-02
n=8  a=1.35937500 b=1.36718750 f(x)=2.29e-04
n=9  a=1.36328125 b=1.36718750 f(x)=-9.29e-03
n=10  a=1.36328125 b=1.36523438 f(x)=-4.53e-03
n=11  a=1.36328125 b=1.36425781 f(x)=-2.15e-03
n=12  a=1.36328125 b=1.36376953 f(x)=-9.61e-04
n=13  a=1.36328125 b=1.36352539 f(x)=-3.66e-04
n=14  a=1.36328125 b=1.36340332 f(x)=-6.85e-05
n=15  a=1.36328125 b=1.36334229 f(x)=8.03e-05
n=16  a=1.36331177 b=1.36334229 f(x)=5.91e-06
n=17  a=1.36332703 b=1.36334229 f(x)=-3.13e-05

Iteration for root on [2,3] :
n=1  a=2.00000000 b=3.00000000 f(x)=-1.38e+00
n=2  a=2.50000000 b=3.00000000 f(x)=-7.81e-02
n=3  a=2.75000000 b=3.00000000 f(x)=8.57e-01
n=4  a=2.75000000 b=2.87500000 f(x)=3.65e-01
n=5  a=2.75000000 b=2.81250000 f(x)=1.37e-01
n=6  a=2.75000000 b=2.78125000 f(x)=2.79e-02
n=7  a=2.75000000 b=2.76562500 f(x)=-2.55e-02
n=8  a=2.75781250 b=2.76562500 f(x)=1.10e-03
n=9  a=2.75781250 b=2.76171875 f(x)=-1.22e-02
n=10  a=2.75976562 b=2.76171875 f(x)=-5.56e-03
n=11  a=2.76074219 b=2.76171875 f(x)=-2.23e-03
n=12  a=2.76123047 b=2.76171875 f(x)=-5.64e-04
n=13  a=2.76147461 b=2.76171875 f(x)=2.70e-04
n=14  a=2.76147461 b=2.76159668 f(x)=-1.47e-04
n=15  a=2.76153564 b=2.76159668 f(x)=6.14e-05
n=16  a=2.76153564 b=2.76156616 f(x)=-4.29e-05
n=17  a=2.76155090 b=2.76156616 f(x)=9.23e-06
```



$\Rightarrow$ three roots of $P(x)$ using bisection method

$$\begin{cases} x_1 \approx -2.1249 \\ x_2 \approx 1.3633 \\ x_3 \approx 2.7616 \end{cases}$$

```matlab
function x=myFalsePosition(f,xspan,tol,nMax)
% Find a root x of the equation f(x)=0 using the
% method of false position.  The initial interval is xspan=[a,b]
% and the iteration stops when abs(x(n+1)-x(n))<tol.
a=xspan(1);
b=xspan(2);
fa=feval(f,a);
if fa==0
  x=a; return
end
fb=feval(f,b);
if fb==0
  x=b; return
end
if fa*fb>0
  error('Error : f(a)*f(b)>0')
end
for n=1:nMax
    x1=a-fa*(b-a)/(fb-fa);
    fx=feval(f,x1);
    fprintf('n=%d x=%1.13f f(x)=%1.2e\n',n,x1,fx)
    if fx==0
        x=x1;
        return
    end
    if fx*fa<0
        b=x1; fb=fx;
    else
        a=x1; fa=fx;
    end
    if n>1
        if abs(x1-x)<tol
            x=x1;
            return
        end
    end
    x=x1;
end
```

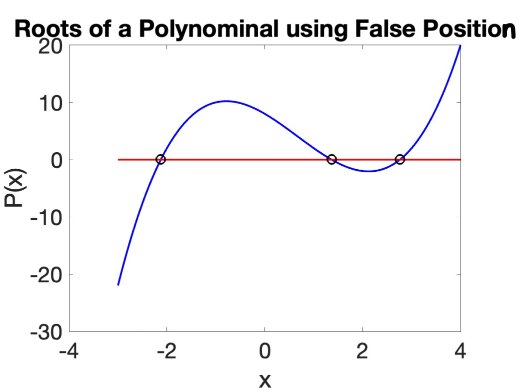
$\Rightarrow$ myFalsePosition.m

Instead of using the midpoint of the interval to update the interval endpoints, use the zero of the line connecting the points $(a, P(a))$ & $(b, P(b))$.

$\Rightarrow$ three roots of $P(x)$ using the method of false position

$$\begin{cases} x_1 \approx -2.1249 \\ x_2 \approx 1.3633 \\ x_3 \approx 2.7616 \end{cases}$$

```
Iteration for root on [-3,-2] :
n=1 x=-2.0833333333333 f(x)=6.94e-01
n=2 x=-2.1113604488079 f(x)=2.29e-01
n=3 x=-2.1205152088336 f(x)=7.43e-02
n=4 x=-2.1234766700117 f(x)=2.40e-02
n=5 x=-2.1244316547919 f(x)=7.73e-03
n=6 x=-2.1247392961840 f(x)=2.49e-03
n=7 x=-2.1248383681071 f(x)=8.02e-04
n=8 x=-2.1248702695645 f(x)=2.58e-04
n=9 x=-2.1248805415802 f(x)=8.31e-05
n=10 x=-2.1248838490514 f(x)=2.68e-05

Iteration for root on [1,2] :
n=1 x=1.5000000000000 f(x)=-6.25e-01
n=2 x=1.3809523809524 f(x)=-8.53e-02
n=3 x=1.3653686826843 f(x)=-9.94e-03
n=4 x=1.3635612027761 f(x)=-1.14e-03
n=5 x=1.3633547934626 f(x)=-1.30e-04
n=6 x=1.3633312644246 f(x)=-1.48e-05
n=7 x=1.3633285828500 f(x)=-1.68e-06
Iteration for root on [2,3] :
n=1 x=2.5000000000000 f(x)=-1.38e+00
n=2 x=2.7037037037037 f(x)=-3.74e-01
n=3 x=2.7504279356385 f(x)=-7.53e-02
n=4 x=2.7594789862117 f(x)=-1.42e-02
n=5 x=2.7611713093732 f(x)=-2.64e-03
n=6 x=2.7614856100351 f(x)=-4.89e-04
n=7 x=2.7615439092682 f(x)=-9.07e-05
n=8 x=2.7615547206036 f(x)=-1.68e-05
n=9 x=2.7615567254314 f(x)=-3.12e-06
```




```matlab
% clear all variables and figures.
clear all
close all

% Set defaults for plotting
fontSize=24; lineWidth=2; markerSize=10;
set(0,'DefaultLineMarkerSize',markerSize);
set(0,'DefaultLineLineWidth',lineWidth);
set(0,'DefaultAxesFontSize',fontSize);
set(0,'DefaultLegendFontSize',fontSize);

% define and plot the polynomial y=x.^3-2*x.^2-5*x+8
P = @(x) x.^3-2*x.^2-5*x+8;
x=linspace(-3,4,300);
y=P(x);
plot(x,y,'b',x,zeros(size(x)),'r')

% Root on [-3,-2]
fprintf('Iteration for root on [-3,-2] :\n')
xspan=[-3,-2]; tol=10e-6; nMax=100;
x1=myFalsePosition(P,xspan,tol,nMax);

% Root on [1,2]
fprintf('Iteration for root on [1,2] :\n')
xspan=[1,2]; tol=10e-6; nMax=100;
x2=myFalsePosition(P,xspan,tol,nMax);

% Root on [2,3]
fprintf('Iteration for root on [2,3] :\n')
xspan=[2,3]; tol=10e-6; nMax=100;
x3=myFalsePosition(P,xspan,tol,nMax);

% plot roots
hold on
roots=[x1 x2 x3];
plot(roots,zeros(3,1),'ko')
hold off
xlabel('x')
ylabel('P(x)')
title('Roots of a Polynominal using False Position')
```

2. Find all fixed points of the following g(x):

$$(a)\quad \frac{x+6}{3x-2}; \qquad (b)\quad \frac{8+2x}{2+x^2}; \qquad (c)\quad x^5.$$

(a) We have $g_1(x) = \frac{x+6}{3x-2}$. The fixed points of $g_1(x)$ are the solutions of $g_1(x) = x \Rightarrow \frac{x+6}{3x-2} = x$

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

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

$$x^2 - x - 2 = (x-2)(x+1) = 0$$

$$\Rightarrow \boxed{x = -1,\ 2}$$

NOT Required

The fixed points are found to be -1, 2. Local convergence requires that $|g_1'(x)| < 1$ at the fixed point. And since $g_1'(x) = -\frac{20}{(3x-2)^2}$, we note that $g_1'(-1) = -\frac{4}{5}$, $g_1'(2) = -\frac{5}{4}$.

Thus, the fixed-point iteration is locally convergent at -1, but not at 2.

(b) We have $g_2(x) = \frac{8+2x}{2+x^2}$. The fixed points of $g_2(x)$ are the solutions of $g_2(x) = x \Rightarrow \frac{8+2x}{2+x^2} = x$

$$8 + 2x = 2x + x^3$$

$$x^3 - 8 = 0$$

NOT Required

therefore, $\boxed{x = 2}$ is a fixed point and the other two fixed points, $x = -1 \pm \sqrt{3}i$ are complex. Now since $g_2'(x) = \frac{-2x^2 - 16x + 4}{(2+x^2)^2}$, we note that $g_2'(2) = -1$, thus, the fixed-point iteration is not locally convergent at 2.

(c) We have $g_3(x) = x^5$. The fixed points of $g_3(x)$ are the solutions of $g_3(x) = x \Rightarrow$

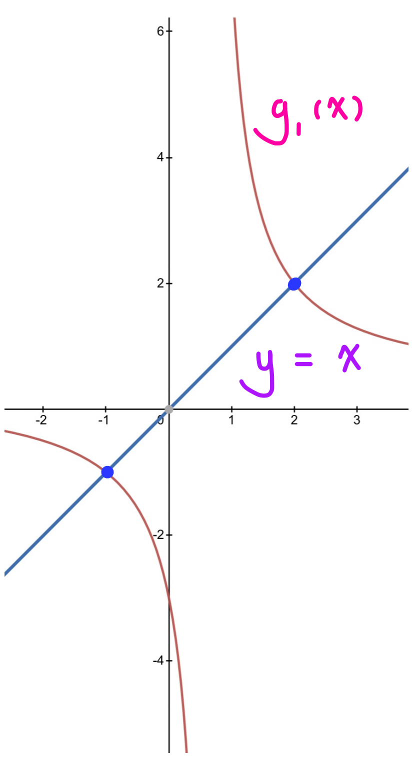
$$x^5 = x$$
$$x^5 - x = 0$$
$$x(x^4 - 1) = 0$$
$$x[(x^2+1)(x+1)(x-1)] = 0$$
$$\Rightarrow \boxed{x = 0, \ \pm 1}$$

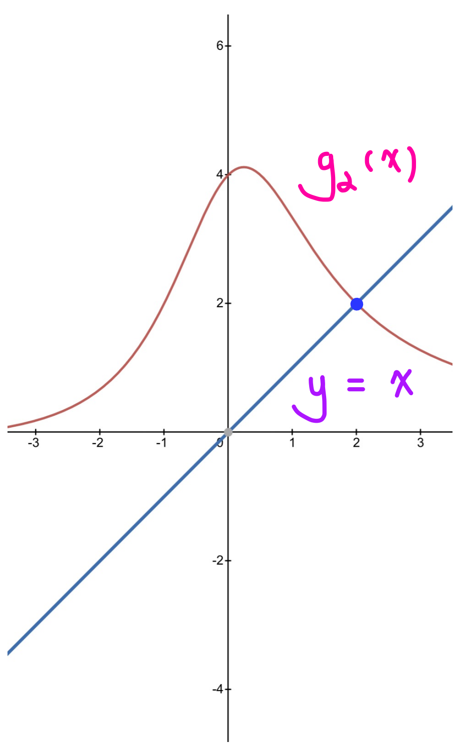
complex fixed points $\pm i$ omitted

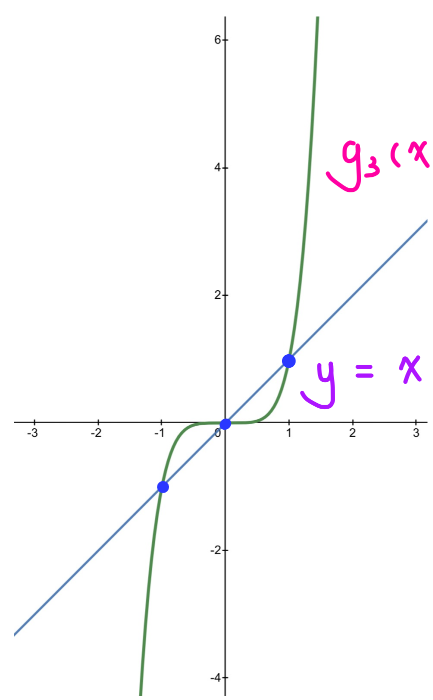
The fixed points are found to be $0, \pm 1$. Local convergence requires that $|g_3'(x)| < 1$ at the fixed point. And since $g_3'(x) = 5x^4$, we note that $g_3'(0) = 0$, $g_3'(\pm 1) = 5$.

NOT Required

Thus, the fixed-point iteration is locally convergent at 0, but not at ± 1.







3. Consider the fixed-point iteration $x_{n+1} = g(x_n)$, where $g(x) = x + \ln(2 - x)$, and define the interval $I = [1/2, 4/3]$.

(a) Show that $g(x) \in I$ whenever $x \in I$, i.e., $g(x)$ maps the interval I into itself, and thus a fixed point of $g(x)$ exists for $x \in I$.

(b) Consider $g'(x)$ for $x \in I$ to show that the fixed-point iteration converges to the unique fixed point of $g(x)$ assuming x_0 is chosen in the interval I.

(a) Given $g(x) = x + \ln(2-x)$, we have $g'(x) = 1 - \frac{1}{2-x} = \frac{1-x}{2-x}$.

For $x \in I = [\frac{1}{2}, \frac{4}{3}]$, we note that the denominator of $g'(x)$ is always positive. And the numerator is positive for $\frac{1}{2} \leq x < 1$, zero for $x = 1$, and negative for $1 < x < \frac{4}{3}$. It implies that $g(x)$ is increasing for $\frac{1}{2} \leq x < 1$ and decreasing for $1 < x < \frac{4}{3}$. Note that $\frac{1}{2} < g(\frac{1}{2}) \approx 0.9 < g(\frac{4}{3}) \approx 0.93 < g(1) = 1 < \frac{4}{3}$, which follows that $g(x)$ remains in the desired interval for $x \in I = [\frac{1}{2}, \frac{4}{3}]$.

Alternatively, to prove that $g(x)$ maps the interval I into itself, we can evaluate $g(x)$ at the endpoints of I and at $x = 1$ where $g'(x) = 0$. If the global extrema for I are all in I, then we can draw the conclusion that $g(x)$ maps I into I.

$$\text{At } x = \frac{1}{2}, \quad g(x) \approx 0.91 \in I.$$

$$\text{At } x = \frac{4}{3}, \quad g(x) \approx 0.93 \in I.$$

Setting $g'(x) = \frac{1-x}{2-x} = 0$, we have that $x = 1$, and that $g(1) = 1 \in I$. Therefore, we have shown that the global extrema for I are all in I, indicating that $g(x)$ remains in the desired interval for $x \in I = [1/2, 4/3]$.

(b) From above, we have $g'(x) = 1 - \frac{1}{2-x} = \frac{1-x}{2-x}$

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

indicating that $g''(x) < 0$ for all $x \in I = [\,1/2, 4/3\,]$, and

thus, $g'(x)$ is decreasing monotically in I, and the extrema

occurs at the two endpoints. Now since $g'(\frac{1}{2}) = \frac{1}{3}$

$$g'(\tfrac{4}{3}) = -\tfrac{1}{2}$$

which follows that $|g'(x)| \le \frac{1}{2} < 1$.

4. Text exercise 16 on page 44.

$\Rightarrow$ Check if cube root of 4 is a fixed point for the following functions:

setting $g_1(x) = \frac{2}{\sqrt{x}} = x \Rightarrow 2 = x^{3/2} \Rightarrow 4 = x^3$

setting $g_2(x) = \frac{3x}{4} + \frac{1}{x^2} = x \Rightarrow \frac{1}{x^2} = \frac{x}{4} \Rightarrow 4 = x^3$

setting $g_3(x) = \frac{2}{3}x + \frac{4}{3x^2} = x \Rightarrow \frac{4}{3x^2} = \frac{1}{3}x \Rightarrow 4 = x^3$

$\Rightarrow$ Check the rate of convergence for the following functions:

$g_1'(x) = -x^{-3/2}$, at $x = \sqrt[3]{4}$, $g_1'(x) = -\frac{1}{2}$

$g_2'(x) = \frac{3}{4} - 2x^{-3}$, at $x = \sqrt[3]{4}$, $g_2'(x) = \frac{1}{4}$

$g_3'(x) = \frac{2}{3} - \frac{8}{3x^3}$, at $x = \sqrt[3]{4}$, $g_3'(x) = 0$

$\Rightarrow$ C, B, A converge from fastest to slowest.

5. Consider the functions

$$f_1(x) = x^3 - 2x - 5; \qquad f_2(x) = e^{-x} - x; \qquad f_3(x) = x\sin(x) - 1.$$

(a) Write a `Matlab` function to compute a root of the function $f(x)$ using Newton's method. Use your code to compute the smallest positive root of each of the functions $f_i(x), i = 1, 2, 3,$ above. You will need to determine a suitable starting guess, x_0 , for each case, and use the stopping criterion $|x_{n+1} - x_n| < 10^{-12}$.

(b) Repeat the calculations in part (a) using the secant method. (You may use the values for x_1 obtained in part (a) along with x_0 for the two starting values needed for the secant method.) Compare the number of iterations needed to compute each roots using the secant method versus the number needed in part (a) using Newton's method.

```
function xStar=myNewton(f,x0,tol,nMax)

% find the root of f(x) using Newton's method with starting
% guess x0. The iteration stops if abs(x(n+1)-x(x))<tol or
% n>nMax

for n=1:nMax
    x1=x0-feval(f,x0,0)/feval(f,x0,1);
    err=abs(x1-x0);
    if err>tol
        fprintf(' n=%d x(n)=%1.4f err=%1.2e', n,x1,err)
        x0=x1;
    else
        fprintf(' n=%d x(n)=%1.4f err=%1.2e (converged)', n,x1,err)
        xStar=x1;
        return
    end
end

n=nMax+1;
fprintf(' n=%d x(n)=%1.4f err=%1.2e (iteration failed)', n,x1,err)
xStar=x1;
```

```
function y=NewtonFunction1(x, derivative)

if derivative==0
    y=x^3-2*x-5;
else
    y=3*x^2-2;
end
```

```
function y=NewtonFunction2(x, derivative)

if derivative==0
    y=exp(-x)-x;
else
    y=-exp(-x)-1;
end
```

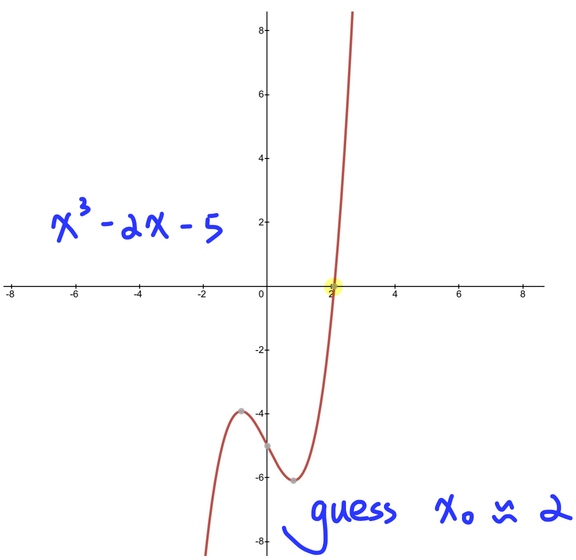
```
function y=NewtonFunction3(x, derivative)

if derivative==0
    y=x*sin(x)-1;
else
    y=sin(x)+x*cos(x);
end
```

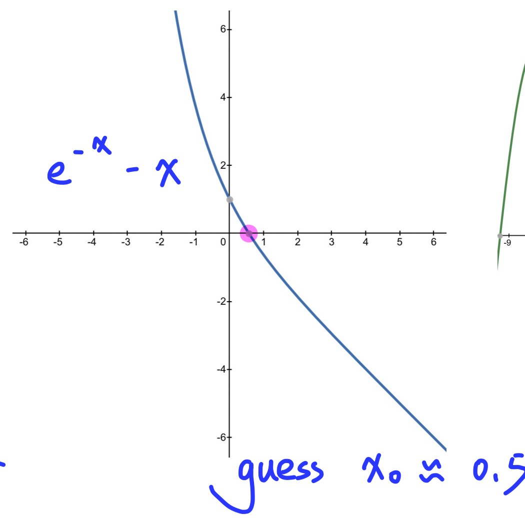
```
% Use Newton's method to find the smallest positive root
% of the following functions
%
% f1=x^3-2*x-5, f2=exp(-x)-x, fx=x*sin(x)-1
%
% with tolerance 10e-12

x1=myNewton('NewtonFunction1',2.0,10e-12,10);
x2=myNewton('NewtonFunction2',0.5,10e-12,10);
x3=myNewton('NewtonFunction3',1.0,10e-12,10);

fprintf('Roots = %0.15e %0.15e %0.15e \n',x1,x2,x3)
```

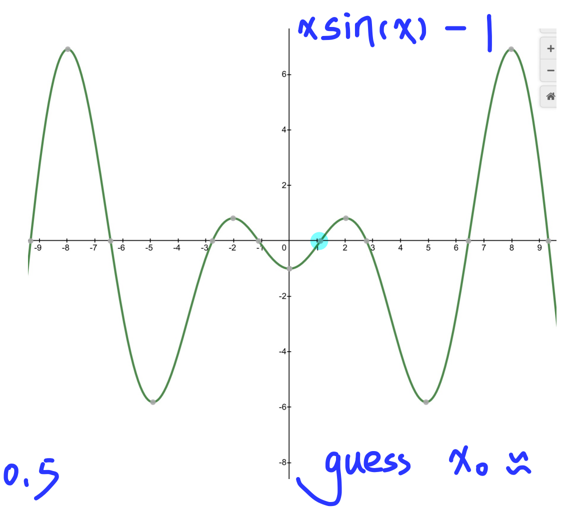
```
 n=1 x(n)=2.1000 err=1.00e-01
 n=2 x(n)=2.0946 err=5.43e-03
 n=3 x(n)=2.0946 err=1.66e-05
 n=4 x(n)=2.0946 err=1.56e-10
 n=5 x(n)=2.0946 err=0.00e+00 (converged)
 n=1 x(n)=0.5663 err=6.63e-02
 n=2 x(n)=0.5671 err=8.32e-04
 n=3 x(n)=0.5671 err=1.25e-07
 n=4 x(n)=0.5671 err=2.89e-15 (converged)
 n=1 x(n)=1.1147 err=1.15e-01
 n=2 x(n)=1.1142 err=5.72e-04
 n=3 x(n)=1.1142 err=1.40e-08
 n=4 x(n)=1.1142 err=2.22e-16 (converged)
Roots = 2.094551481542327e+00 5.671432904097838e-01 1.114157140871930e+00
```

```
function xStar=mySecant(f,x0,x1,tol,nMax)

% find the root of f(x) using secant method with starting
% guesses x0 and x1. The iteration stops if abs(x(n+1)-x(n))<tol or
% n>nMax

n=0;
while n<nMax+1
    n=n+1;
    fx0=feval(f,x0);
    fx1=feval(f,x1);
    x=x1-fx1*(x1-x0)/(fx1-fx0);
    err=abs(x-x1);
    if err>tol
        fprintf(' n = %d x0 = %1.8f x1 = %1.8f err = %1.2d \n', n,x0,x1,err)
        x0=x1;
        x1=x;
    else
        fprintf(' n = %d x0 = %1.8f x1 = %1.8f err = %1.2d (converged) \n',
 n,x0,x1,err)
        xStar=x;
        return
    end
end
```

```
>> mySecant(@(x) x^3-2*x-5, 2.0, 2.0946, 10e-12, 10)
 n = 1 x0 = 2.00000000 x1 = 2.09460000 err = 5.12e-05
 n = 2 x0 = 2.09460000 x1 = 2.09454880 err = 2.68e-06
 n = 3 x0 = 2.09454880 x1 = 2.09455148 err = 7.33e-11
 n = 4 x0 = 2.09455148 x1 = 2.09455148 err = 00 (converged)

ans =

    2.0946

>> mySecant(@(x) exp(1).^(-x)-x, 0.5, 0.5671, 10e-12, 10)
 n = 1 x0 = 0.50000000 x1 = 0.56710000 err = 4.28e-05
 n = 2 x0 = 0.56710000 x1 = 0.56714276 err = 5.31e-07
 n = 3 x0 = 0.56714276 x1 = 0.56714329 err = 4.16e-12 (converged)

ans =

    0.5671

>> mySecant(@(x) x*sin(x)-1, 1.0, 1.1142, 10e-12, 10)
 n = 1 x0 = 1.00000000 x1 = 1.11420000 err = 4.29e-05
 n = 2 x0 = 1.11420000 x1 = 1.11415714 err = 3.57e-09
 n = 3 x0 = 1.11415714 x1 = 1.11415714 err = 6.66e-15 (converged)

ans =

    1.1142
```







MATH-4800 **NUMERICAL COMPUTING** **Summer 2023**

Assignment 3, due before class, Thursday June 15, 2023.

Haowen He heh4@rpi.edu

1. Let

$$A = \begin{bmatrix} 1 & -2 & -1 \\ \alpha & 3 & 1 \\ -1 & \alpha & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} \beta \\ -4 \\ 2 \end{bmatrix}$$

and let $\mathbf{x}$ be the solution of $A\mathbf{x} = \mathbf{b}$, if it exists. Determine conditions on the constants α and β such that (a) a solution or solutions exist and (b) a unique solution exists.

(a)

$$\begin{aligned} \det(A) &= 1\cdot(3-\alpha) - (-2)(\alpha+1) - 1\cdot(\alpha^2+3) \\ &= 3 - \alpha + 2\alpha + 2 - \alpha^2 - 3 \\ &= -\alpha^2 + \alpha + 2 \end{aligned}$$

when $\det(A) \neq 0$, $-\alpha^2 + \alpha + 2 \neq 0$

$$(\alpha - 2)(\alpha + 1) \neq 0$$

$$\alpha \neq -1, 2$$

matrix A is non-singular, and thus, a solution exists.

when $\alpha = -1$, $\left(\begin{array}{ccc|c} 1 & -2 & -1 & \beta \\ -1 & 3 & 1 & -4 \\ -1 & -1 & 1 & 2 \end{array}\right) \xrightarrow[R_3 \to R_1 + R_3]{R_2 \to R_1 + R_2}$

$$\left(\begin{array}{ccc|c} 1 & -2 & -1 & \beta \\ 0 & 1 & 0 & \beta - 4 \\ 0 & -3 & 0 & \beta + 2 \end{array}\right) \xrightarrow{R_3 \to 3R_2 + R_3} \left(\begin{array}{ccc|c} 1 & -2 & -1 & \beta \\ 0 & 1 & 0 & \beta - 4 \\ 0 & 0 & 0 & 4\beta - 10 \end{array}\right)$$

in order for solutions to exist, $4\beta - 10 = 0 \Rightarrow \beta = 5/2$.

when $\alpha = 2$, $\left(\begin{array}{ccc|c} 1 & -2 & -1 & \beta \\ 2 & 3 & 1 & -4 \\ -1 & 2 & 1 & 2 \end{array}\right) \xrightarrow[R_3 \to R_1 + R_3]{R_2 \to R_2 - 2R_1}$

$\left(\begin{array}{ccc|c} 1 & -2 & -1 & \beta \\ 0 & 7 & 3 & -4-\beta \\ 0 & 0 & 0 & \beta+2 \end{array}\right)$ in order for solutions to exist,

$\beta + 2 = 0 \Rightarrow \beta = -2.$

Therefore, when $\boxed{\alpha \neq -1, 2 \text{ or } \alpha = -1, \beta = 5/2, \text{ or } \alpha = 2, \beta = -2}$, //

a solution or solutions exist.

(b) When $\boxed{\alpha \neq -1, 2}$, determinant of A is non-zero. It follows that matrix A is non-singular, which implies that A is invertible, and thus $A^{-1}b$ is a unique solution of $Ax = b$, regardless of the value of β we choose.

2. (*a*) Consider the two augmented matrices of the form $[A|\mathbf{b}]$ in text exercise 4 on page 81. For each matrix, use row operations to reduce it to the upper triangular form $[U|\mathbf{c}]$ and then find the solution of $U\mathbf{x} = \mathbf{c}$ using backwards substitution.

(*b*) Text exercise 6 and 7 on page 82.

(a) $$\left(\begin{array}{ccc|c} 3 & -4 & -2 & 3 \\ 6 & -6 & 1 & 2 \\ -3 & 8 & 2 & -1 \end{array}\right) \xrightarrow[R_3 \to R_3 + R_1]{R_2 \to R_2 - 2R_1} \left(\begin{array}{ccc|c} 3 & -4 & -2 & 3 \\ 0 & 2 & 5 & -4 \\ 0 & 4 & 0 & 2 \end{array}\right)$$

$$\xrightarrow{R_3 \to R_3 - 2R_2} \left(\begin{array}{ccc|c} 3 & -4 & -2 & 3 \\ 0 & 2 & 5 & -4 \\ 0 & 0 & -10 & 10 \end{array}\right) = [U|c]$$

$$\left.\begin{aligned} -10x_3 &= 10 \\ 2x_2 + 5x_3 &= -4 \\ 3x_1 - 4x_2 - 2x_3 &= 3 \end{aligned}\right\} \quad \begin{aligned} x_3 &= -1 \\ 2x_2 &= -4 + 5 = 1 \Rightarrow x_2 = 1/2 \\ 3x_1 &= 3 + 2 - 2 = 3 \Rightarrow x_1 = 1 \end{aligned}$$

$$\boxed{\text{Solution } x = \begin{pmatrix} 1 \\ 1/2 \\ -1 \end{pmatrix}}$$

//

$$\left(\begin{array}{ccc|c} 2 & 1 & -1 & 2 \\ 6 & 2 & -2 & 8 \\ 4 & 6 & -3 & 5 \end{array}\right) \xrightarrow[R_3 \to R_3 - 2R_1]{R_2 \to R_2 - 3R_1} \left(\begin{array}{ccc|c} 2 & 1 & -1 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & 4 & -1 & 1 \end{array}\right)$$

$$\xrightarrow{R_3 \to R_3 + 4R_1} \left(\begin{array}{ccc|c} 2 & 1 & -1 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & 0 & 3 & 9 \end{array}\right)$$

$$\left.\begin{aligned} 3x_3 &= 9 \\ -x_2 + x_3 &= 2 \\ 2x_1 + x_2 - x_3 &= 2 \end{aligned}\right\} \quad \begin{aligned} &x_3 = 3 \\ &-x_2 = 2 - 3 = -1 \Rightarrow x_2 = 1 \\ &2x_1 = 2 - 1 + 3 = 4 \Rightarrow x_1 = 2 \end{aligned}$$

$$\boxed{\text{Solution } x = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}}$$

//

(b) Given that the computer takes 0.005 seconds to perform back-substitution which has complexity of $O(n^2)$. Now as a full Gaussian elimination takes approximately $\frac{2}{3}n^3$ operations, with $n = 5000$, it follows that

$$n^2 \cdot \frac{2}{3} n$$

$$= 0.005 \cdot \frac{2 \cdot 5000}{3}$$

$$\approx \boxed{17 \text{ s}}.$$

//

Now given the computer takes 0.002 seconds to perform back-substitution on a 4000 by 4000 matrix, it follows that it can perform

$$\frac{4000^2}{0.002} = 8 \times 10^9 \text{ operations/second}$$

Solving a general system of 9000 equations with 9000 unknowns will take approximately $9000^2 + \frac{2}{3} 9000^3 = 4.86081 \times 10^{11}$ operations, and the computer will need approximately $\frac{4.86081 \times 10^{11}}{8 \times 10^9} \approx \boxed{61 \text{ s}}$ to compute.

3. Let $\hat{A}$ be an $n \times (n+2)$ matrix and consider the following steps in a `Matlab` code:

```
for kb=1:2
  x(n,kb)=Ahat(n,n+kb)/Ahat(n,n);
  for i=n-1:-1:1
    sum=Ahat(i,n+kb);
    for j=i+1:n
      sum=sum-Ahat(i,j)*x(j,kb);
    end
    x(i,kb)=sum/Ahat(i,i);
  end
end
```

Determine the number of `flops` used to compute the elements of `x` for a given positive interger n. (Find an expression for the number of `flops` exactly, without approximation for large n.)

Divisions $\quad \mathcal{F}_d = \sum_{kb=1}^{2} 1 + \sum_{kb=1}^{2} \sum_{i=1}^{n-1} 1$

$$= 2 + 2(n-1)$$

$$= 2n$$

Multiplications $\quad \mathcal{F}_m = \sum_{kb=1}^{2} 1 \sum_{j=i+1}^{n} = \sum_{kb=1}^{2} 1 \sum_{i=1}^{n-1} (n-(i+1)+1)$

$$= 2\sum_{i=1}^{n-1}(n-i) = 2\left(\sum_{i=1}^{n-1} n - \sum_{i=1}^{n-1} i\right)$$

$$= 2\left(n\sum_{i=1}^{n-1} 1 - \sum_{i=1}^{n-1} i\right)$$

$$= 2n(n-1) - 2\cdot\frac{n(n-1)}{2}$$

$$= 2n^2 - 2n - n^2 + n$$

$$= n^2 - n$$

Additions $\quad \mathcal{F}_a = n^2 - n \quad$ same as multiplication

one floating-point subtraction from each multiplication

Total $= 2n + n^2 - n + n^2 - n = \boxed{2n^2 \text{ flops}}$ //

4. (a) Write a Matlab code to compute the LU-decomposition of a given $n \times n$ matrix A following the algorithm discussed in class. Your code need not perform row interchanges (pivoting), but should check for very small pivot elements and provide a warning if one is encountered.

(b) Use your code to compute the L and U factors for the following matrices:

(i) `A=[6 3 2; -1 4 2; 1 3 -5];` (ii) `A=hilb(5);`

Print out L and U and compute `norm(A-L*U)` for each case. The latter is Matlab's calculation of the distance (i.e. norm) between A and the product LU. What are the expected values for this norm?

```
function [L,U] = myLU(A)

[m,n]=size(A);
if m~=n
    error("Matrix is not square!")
end
L=eye(n);
U=A;

for k=1:n-1
    if abs(U(k,k))<10e-16
        printf("Warning: value less than 1-e-16\n");
    end

    for i=k+1:n
        L(i,k)=U(i,k)/U(k,k);
        for j=k:n
            U(i,j)=U(i,j)-L(i,k)*U(k,j);
        end
    end
end

% HW 3 Problem 4 part (i)
[L,U] = myLU([6 3 2; -1 4 2; 1 3 -5])
norm([6 3 2; -1 4 2; 1 3 -5]- L*U)
% HW 3 Problem 4 part (ii)
[L,U] = myLU(hilb(5))
norm(hilb(5)- L*U)
```

Part i

```
L =

    1.0000         0         0
   -0.1667    1.0000         0
    0.1667    0.5556    1.0000


U =

    6.0000    3.0000    2.0000
         0    4.5000    2.3333
         0         0   -6.6296


ans =

     0
```

Part ii

```
L =

    1.0000         0         0         0         0
    0.5000    1.0000         0         0         0
    0.3333    1.0000    1.0000         0         0
    0.2500    0.9000    1.5000    1.0000         0
    0.2000    0.8000    1.7143    2.0000    1.0000


U =

    1.0000    0.5000    0.3333    0.2500    0.2000
         0    0.0833    0.0833    0.0750    0.0667
         0   -0.0000    0.0056    0.0083    0.0095
         0         0         0    0.0004    0.0007
         0         0         0         0    0.0000


ans =

   4.4909e-17
```

5. (a) An $n \times n$ elimination matrix M has the form

$$M = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & 0 & \\ & & 1 & & & \\ & & -m_{k+1,k} & & & \\ & 0 & \vdots & & 0 & \ddots \\ & & -m_{n,k} & & & 1 \end{bmatrix} \quad \text{for some integer } k \in [1, n-1]$$

Find 4×4 elimination matrices M_1 and M_2 satisfying

$$M_1 \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad M_2 \begin{bmatrix} 2 \\ -1 \\ 4 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 0 \\ 0 \end{bmatrix}$$

(b) An $n \times n$ permutation matrix P is a row or column permutation of the $n \times n$ identity matrix. Thus, P has exactly one 1 in every column and row. The following are examples of 3×3 permutation matrices

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

Find 4×4 permutation matrices P_1 and P_2 satisfying

$$P_1 \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 2 \end{bmatrix}, \qquad P_2 \begin{bmatrix} 2 \\ -1 \\ 4 \\ -3 \end{bmatrix} = \begin{bmatrix} 4 \\ -1 \\ -3 \\ 2 \end{bmatrix}$$

(a) $\quad 2 - 2 \cdot 1 = 3 - 3 \cdot 1 = 4 - 4 \cdot 1 = 0$

$$\Downarrow \qquad \Downarrow \qquad \Downarrow$$

$m_{2,1} = 2 \qquad m_{3,1} = 3 \qquad m_{4,1} = 4$

$$\boxed{M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -3 & 0 & 1 & 0 \\ -4 & 0 & 0 & 1 \end{pmatrix}}$$

//

$$4 + 4 \cdot (-1) = -3 - 3 \cdot (-1) = 0$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$m_{3,2} = -4 \qquad m_{4,2} = 3$$

$$\boxed{M_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 4 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{pmatrix}} //$$

By definition given above, elimination matrix M has the form of $\begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{pmatrix}$, where the bottom entries should be in the same column.

(b) $$\boxed{P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad P_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}} //$$

For permutation matrix P, the product PA is a new matrix whose rows consists of the rows of A rearranged in the new order.

MATH-4800 **NUMERICAL COMPUTING** **Summer 2023**

Assignment 4, due before class, Monday June 26, 2023.

Haowen He heh4@rpi.edu

1. Text exercises 2(a), 2(b), and 4 on page 97. Hint: it may be helpful to recall the following formulas valid for 2×2 matrices

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

(2a) $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ $\|A\|_\infty = \max\{3, 7\} = 7$

$A^{-1} = \begin{pmatrix} -2 & 1 \\ 3/2 & -1/2 \end{pmatrix}$ $\|A^{-1}\|_\infty = \max\{3, 2\} = 3$

$\Rightarrow \kappa_\infty(A) = 7 \times 3 = \boxed{21}$ //

(2b) $A = \begin{pmatrix} 1 & 2.01 \\ 3 & 6 \end{pmatrix}$ $\|A\|_\infty = \max\{3.01, 9\} = 9$

$A^{-1} = \begin{pmatrix} -200 & 67 \\ 100 & -100/3 \end{pmatrix}$ $\|A^{-1}\|_\infty = \max\{267, 400/3\} = 267$

$\Rightarrow \kappa_\infty(A) = 9 \times 267 = \boxed{2403}$ //

(4a) $\left(\begin{array}{cc|c} 1 & 2 & 1 \\ 2 & 4.01 & 2 \end{array}\right) \xrightarrow{R_2 \to R_2 - 2R_1} \left(\begin{array}{cc|c} 1 & 2 & 1 \\ 0 & 0.01 & 0 \end{array}\right) \Rightarrow \begin{cases} 0.01 x_2 = 0 \Rightarrow x_2 = 0 \\ x_1 + 2x_2 = 1 \Rightarrow x_1 = 1 \end{cases}$

Actual solution: $x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$\Rightarrow x_c = (-1, 1)$

Forward error: $\|x - x_c\|_\infty = \max\{2, 1\} = 2$

Backward error: $\|b - Ax_c\|_\infty = \max\{0, 0.01\} = 0.01$

$$\text{Error magnification factor} = \frac{\text{rel. forward err}}{\text{rel. backward err}}$$

$$= \frac{2/1}{0.01/2}$$

$$= \boxed{400}$$

(4b) $\Rightarrow x_c = (3, -1)$

Forward error: $\|x - x_c\|_\infty = \max\{|1-2|, 1\} = 2$

Backward error: $\|b - Ax_c\|_\infty = \max\{0, 0.01\} = 0.01$

$$\text{Error magnification factor} = \frac{\text{rel. forward err}}{\text{rel. backward err}}$$

$$= \frac{2/1}{0.01/2}$$

$$= \boxed{400}$$

(4c) $\Rightarrow x_c = (2, -1/2)$

Forward error : $\|x - x_c\|_\infty = \max\{|-1|, 1/2\} = 1$

Backward error : $\|b - Ax_c\|_\infty = \max\{0, 0.005\} = 0.005$

$$\text{Error magnification factor} = \frac{\text{rel. forward err}}{\text{rel. backward err}}$$

$$= \frac{1/1}{0.005/2}$$

$$= \boxed{400}$$

2. Computer problem 1 on page 98. For this problem, use $A_{ij} = 3/(i+2j-1)$ instead of the formula for A_{ij} given in the text. Be sure to include a brief comment on the result of the comparison. Is the result expected and why.

```
function [] = p2(n)
A=zeros(n);
x=ones(n,1);
for i=1:n
    for j=1:n
        A(i,j)=3/(i+2*j-1);
    end
end
b=A*x;
xc=A\b;

fe=norm(x-xc, "inf");
rfe=fe/norm(x, "inf");

be=norm(b-A*xc, "inf");
rbe=be/norm(b, "inf");

emf=rfe/rbe;

condnum=cond(A, inf);

disp(["forward error: ", num2str(fe)])
disp(["relative forward error: ", num2str(rfe)])
disp(["backward error: ", num2str(be)])
disp(["relative backward error: ", num2str(rbe)])
disp(["Error Magnification factor: ", num2str(emf)])
disp(["Condition number of A: ", num2str(condnum)])
```

```
p2(6)
p2(10)
```

n = 6

```
"forward error: "     "1.8479e-09"

"relative forward error: "     "1.8479e-09"

"backward error: "     "2.2204e-16"

"relative backward error: "     "6.042e-17"

"Error Magnification factor: "     "30584566.425"

"Condition number of A: "     "70342013.954"
```

n = 10

```
"forward error: "     "0.0028717"

"relative forward error: "     "0.0028717"

"backward error: "     "4.4409e-16"

"relative backward error: "     "1.0108e-16"

"Error Magnification factor: "     "28410438423259.55"

"Condition number of A: "     "131321934554336.8"
```

3. Text exercises $2(a)$, and $4(b)$ on page 106. These problems are to be done using hand calculations.

(2a) $A = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & -1 \\ -1 & 1 & -1 \end{pmatrix} \Rightarrow$ Row 2 has the largest pivot value $\Rightarrow P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

$$PA = \begin{pmatrix} 2 & 1 & -1 \\ -1 & 1 & -1 \\ 1 & 1 & 0 \end{pmatrix} \xrightarrow[R_3 \to R_3 - \frac{1}{2}R_1]{R_2 \to R_2 + \frac{1}{2}R_1} \begin{pmatrix} 2 & 1 & -1 \\ 0 & 3/2 & -3/2 \\ 0 & 1/2 & 1/2 \end{pmatrix} \xrightarrow{R_3 \to R_3 - \frac{1}{3}R_2}$$

$$\begin{pmatrix} 2 & 1 & -1 \\ 0 & 3/2 & -3/2 \\ 0 & 0 & 1 \end{pmatrix} = U \qquad \Rightarrow L = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{pmatrix}$$

$$PA = LU$$

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 \\ 2 & 1 & -1 \\ -1 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 1/2 & 1/3 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 & -1 \\ 0 & 3/2 & -3/2 \\ 0 & 0 & 1 \end{pmatrix}$$

(4b) $A = \begin{pmatrix} -1 & 0 & 1 \\ 2 & 1 & 1 \\ -1 & 2 & 0 \end{pmatrix} \Rightarrow$ Row 2 has the largest pivot value $\Rightarrow P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$

$$PA = \begin{pmatrix} 2 & 1 & 1 \\ -1 & 2 & 0 \\ -1 & 0 & 1 \end{pmatrix} \xrightarrow[R_3 \to R_3 + \frac{1}{2}R_1]{R_2 \to R_2 + \frac{1}{2}R_1} \begin{pmatrix} 2 & 1 & 1 \\ 0 & 5/2 & 1/2 \\ 0 & 1/2 & 3/2 \end{pmatrix} \xrightarrow{R_3 \to R_3 - \frac{1}{5}R_2}$$

$$\begin{pmatrix} 2 & 1 & -1 \\ 0 & 5/2 & 1/2 \\ 0 & 0 & 7/5 \end{pmatrix} = U \quad \Rightarrow \quad L = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ -1/2 & 1/5 & 1 \end{pmatrix}$$

back substitution: $Pb = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} -2 \\ 17 \\ 3 \end{pmatrix} = \begin{pmatrix} 17 \\ 3 \\ -2 \end{pmatrix}$

$$Ly = Pb \Rightarrow \left(\begin{array}{ccc|c} 1 & 0 & 0 & 17 \\ -1/2 & 1 & 0 & 3 \\ -1/2 & 1/5 & 1 & -2 \end{array}\right) \Rightarrow \begin{cases} y_1 = 17 \\ y_2 = 23/2 = 11.5 \\ y_3 = 21/5 = 4.2 \end{cases}$$

$$Ux = y \Rightarrow \left(\begin{array}{ccc|c} 2 & 1 & -1 & 17 \\ 0 & 5/2 & 1/2 & 11.5 \\ 0 & 0 & 7/5 & 4.2 \end{array}\right) \Rightarrow \begin{cases} x_3 = 3 \\ x_2 = 4 \\ x_1 = 5 \end{cases}$$

$$x = \begin{pmatrix} 5 \\ 4 \\ 3 \end{pmatrix}.$$

4. Write a `Matlab` code (following the algorithm discussed in class) that takes an $n \times n$ matrix A as input and returns P, L and U such that $PA = LU$. Your code should use row elimination with partial pivoting. Test your code using the matrix in text exercise $2(a)$ above and on a matrix generated using `A=rand(6,6)`. In both cases, compute `norm(P*A-L*U)` to verify that your code is working correctly.

```
function [P,L,U] = myPLU(A)

[m,n]=size(A);
if m~=n
    error("Matrix A is not square!")
end

P=eye(n);
L=eye(n);
U=A;

for k=1:n-1
    [pv,p] = max(abs(U(k:n, k)));
    p=p+k-1;
    if p~=k
        P([p,k],:)=P([k,p],:);
        U([p,k],:)=U([k,p],:);
        if k>1
            L([p,k],1:k-1)=L([k,p],1:k-1);
        end
    end

    for i=k+1:n
        m=U(i,k)/U(k,k);
        L(i,k)=m;
        for j=k:n
            U(i,j)=U(i,j)-U(k,j)*m;
        end
    end
end
```

```
% Problem 4 Part (i)
A=[1,1,0;2,1,-1;-1,1,-1];
[P,L,U]=myPLU(A)
fprintf("norm(P*A-L*U) = %g\n", norm(P*A-L*U))
```

```
P =

     0     1     0
     0     0     1
     1     0     0


L =

    1.0000         0         0
   -0.5000    1.0000         0
    0.5000    0.3333    1.0000


U =

    2.0000    1.0000   -1.0000
         0    1.5000   -1.5000
         0         0    1.0000

norm(P*A-L*U) = 0
```

```
% Problem 4 Part (iI)
A=rand(6,6);
[P,L,U]=myPLU(A)
fprintf("norm(P*A-L*U) = %g\n", norm(P*A-L*U))
```

```
P =

     0     0     1     0     0     0
     0     0     0     1     0     0
     0     1     0     0     0     0
     1     0     0     0     0     0
     0     0     0     0     0     1
     0     0     0     0     1     0


L =

    1.0000         0         0         0         0         0
    0.3766    1.0000         0         0         0         0
    0.2621    0.5457    1.0000         0         0         0
    0.8763    0.4409   -0.2799    1.0000         0         0
    0.2702    0.6511   -0.3456    0.6993    1.0000         0
    0.2116    0.7315    0.7278   -0.9116    0.4772    1.0000


U =

    0.9293    0.3517    0.7572    0.5308    0.0119    0.1656
         0    0.6984    0.4685    0.5793    0.3326    0.5396
    0.0000   -0.0000   -0.1683   -0.4013    0.2847    0.1906
         0         0         0   -0.7570    0.4915   -0.0185
         0         0         0         0    0.3292    0.3368
         0         0         0         0         0   -0.4831

norm(P*A-L*U) = 1.74528e-16
```

5. Consider the tridiagonal matrices

$$A = \begin{bmatrix} 2 & 3 & 0 \\ -1 & -5/2 & 2 \\ 0 & -1/4 & 3/2 \end{bmatrix}, \qquad B = \begin{bmatrix} 4 & -2 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

(a) Use direct factorization (as discussed in class) to find (by hand calculations) a lower triangular matrix L (with ones on the main diagonal) and an upper triangular matrix U such that $A = LU$.

(b) Show that the symmetric matrix B is positive definite by writing $\mathbf{x}^T B\mathbf{x}$ as a sum of squares involving the three components, (x_1, x_2, x_3) , of $\mathbf{x}$. Be sure to show that $\mathbf{x}^T B\mathbf{x} = 0$ implies that $x_1 = x_2 = x_3 = 0$.

(a) $$A = \begin{pmatrix} 2 & 3 & 0 \\ -1 & -5/2 & 2 \\ 0 & -1/4 & 3/2 \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & u_{22} + l_{21}u_{12} & u_{23} + l_{21}u_{13} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{pmatrix}$$

$$u_{11} = 2 \implies \begin{cases} l_{21}u_{11} = -1 \implies l_{21} = -1/2 \\ l_{31}u_{11} = 0 \implies l_{31} = 0 \end{cases}$$

$$u_{12} = 3 \implies \begin{cases} u_{22} + l_{21}u_{12} = -\frac{5}{2} \implies u_{22} = -1 \\ l_{31}u_{12} + l_{32}u_{22} = -\frac{1}{4} \implies l_{32} = \frac{1}{4} \end{cases}$$

$$u_{13} = 0 \implies \begin{cases} u_{23} + l_{21}u_{13} = 2 \implies u_{23} = 2 \\ l_{31}u_{13} + l_{32}u_{23} + u_{33} = \frac{3}{2} \implies u_{33} = 1 \end{cases}$$

$$L = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & 1/4 & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix} = \begin{pmatrix} 2 & 3 & 0 \\ 0 & -1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

(b) $$x^T B x = 4x_1^2 - 4x_1x_2 + 3x_2^2 + 2x_2x_3 + x_3^2$$

$$= (4x_1^2 - 4x_1x_2 + x_2^2) + x_2^2 + (x_2^2 + 2x_2x_3 + x_3^2)$$

$$= (2x_1 - x_2)^2 + x_2^2 + (x_2 + x_3)^2 \geqslant 0.$$

Now as $x^T B x$ is the sum of squares, it follows that $x^T B x = 0$ if and only if $2x_1 - x_2 = x_2 = x_2 + x_3 = 0$, which implies that all three components of x are zero.

6. Let

$$\mathbf{G}(\mathbf{x}) = \begin{bmatrix} \frac{1}{4}(x_1 - x_2 - 1) \\ \alpha(x_1^2 - 1) + x_2 \end{bmatrix}, \qquad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} \frac{1}{2}x_1 - x_2 - 1 \\ x_1^2 - x_2^2 - 1 \end{bmatrix}.$$

(a) Show that $\mathbf{G}(\mathbf{x})$ has a fixed point $\mathbf{x}^\star = (-1, 2)$ for any value of the constant α. Analyze the eigenvalues of the Jacobian matrix $J = \partial \mathbf{G}/\partial \mathbf{x}$ to determine whether the fixed-point iteration $\mathbf{x}_{k+1} = \mathbf{G}(\mathbf{x}_k)$ is locally convergent when $(i)\alpha = -1/4$ or when (ii) $\alpha = 1/2$.

(b) Let $f_1(x_1, x_2)$ and $f_2(x_1, x_2)$ be the two component functions of $\mathbf{F}(\mathbf{x})$. Plot the graphs of $f_1 = 0$ and $f_2 = 0$ in the $x_1 - x_2$ plane to estimate the roots $\mathbf{x} = \mathbf{r}$ of $\mathbf{F}$. Write a `Matlab` code to determine all roots of $\mathbf{F}$ using Newton's method. Output $\mathbf{x}_k$, the k^{th} iterate in Newton's method, and $||\mathbf{x}_k - \mathbf{x}_{k-1}||$ to verify quadratic convergence.

(a) $G(x^*) = \begin{pmatrix} \frac{1}{4}(-1-2-1) \\ \alpha(1-1)+2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix} = x^*$, thus x^* is a fixed point of $G(x)$ regardless of the value of constant α

$$J = \begin{pmatrix} \frac{\partial G_1}{\partial x_1} & \frac{\partial G_1}{\partial x_2} \\ \frac{\partial G_2}{\partial x_1} & \frac{\partial G_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} 1/4 & -1/4 \\ 2\alpha x_1 & 1 \end{pmatrix} \Rightarrow J_{(-1,2)} = \begin{pmatrix} 1/4 & -1/4 \\ -2\alpha & 1 \end{pmatrix}$$

At $\alpha = -\frac{1}{4}$, $J_1 = \begin{pmatrix} 1/4 & -1/4 \\ 1/2 & 1 \end{pmatrix}$

$$\det(J_1 - \lambda I) = \begin{vmatrix} \frac{1}{4}-\lambda & -1/4 \\ 1/2 & 1-\lambda \end{vmatrix} = \lambda^2 - \frac{5}{4}\lambda + \frac{1}{4} + \frac{1}{8} = \lambda^2 - \frac{5}{4}\lambda + \frac{3}{8}$$

set $\lambda^2 - \frac{5}{4}\lambda + \frac{3}{8} = 0 \Rightarrow (\lambda - \frac{1}{2})(\lambda - \frac{3}{4}) = 0$

$$\lambda_1 = \frac{1}{2}, \quad \lambda_2 = \frac{3}{4}$$

$$\max\{|\lambda_i|\} < 1$$

Therefore, when $\alpha = -1/4$, the fixed-point iteration is locally convergent.

At $\alpha = \frac{1}{2}$, $J_2 = \begin{pmatrix} 1/4 & -1/4 \\ -1 & 1 \end{pmatrix}$

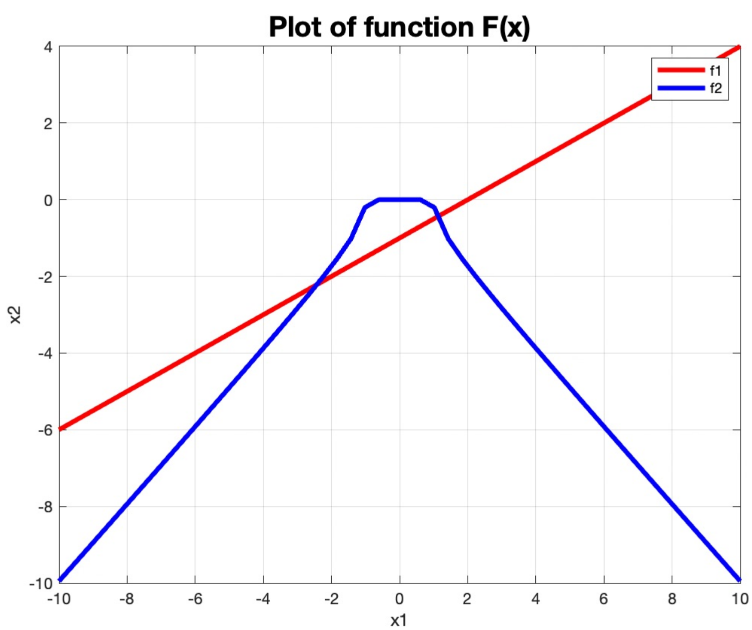
$$\det(J_2 - \lambda I) = \begin{vmatrix} \frac{1}{4} - \lambda & -1/4 \\ 1/2 & 1-\lambda \end{vmatrix} = \lambda^2 - \frac{5}{4}\lambda + \frac{1}{4} - \frac{1}{4} = \lambda^2 - \frac{5}{4}\lambda$$

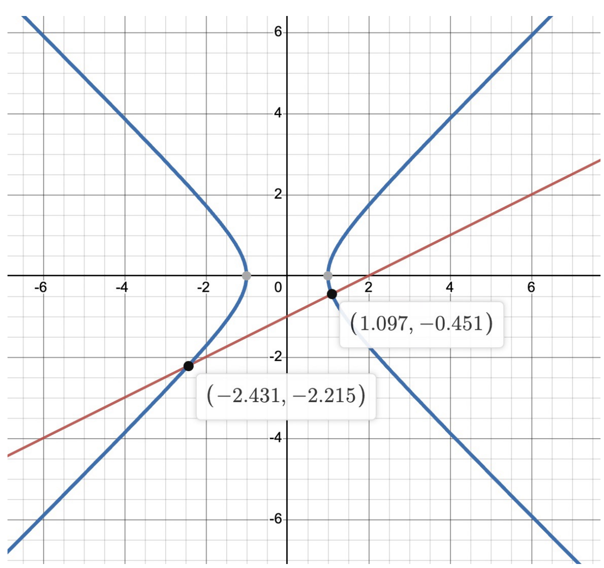
set $\lambda^2 - \frac{5}{4}\lambda = 0 \Rightarrow \lambda(\lambda - \frac{5}{4}) = 0$

$$\lambda_1 = 0, \quad \lambda_2 = \frac{5}{4}$$

$$\max\{|\lambda_i|\} > 1$$

Therefore, when $\alpha = \frac{1}{2}$, the fixed-point iteration is NOT locally convergent.





```
function x=myNewtonNL(f,fp,x0)

tol=1e-12;
nMax=10;
xpre=[0;0];
x=x0;
n=0;

while n<nMax && abs(norm(x-xpre))>tol
    xpre=x;
    s=fp(x)\-f(x);
    x=s+x;
    n=n+1;
    fprintf("iteration %d x=[%3.5f; %3.5f;] err=%8e\n",n,x(1,:),x(2,:),norm(x-
xpre))
end
end
```

```
f=@(x)[1/2*x(1)-x(2)-1; x(1)^2-x(2)^2-1];
fp=@(x)[1/2 -1; 2*x(1) -2*x(2)];
x0=[1; -0.5];
myNewtonNL(f,fp,x0)
x0=[-2.5; -2];
myNewtonNL(f,fp,x0)
```

```
iteration 1 x=[1.10000; -0.45000;] err=1.118034e-01
iteration 2 x=[1.09717; -0.45142;] err=3.164247e-03
iteration 3 x=[1.09717; -0.45142;] err=2.538619e-06
iteration 4 x=[1.09717; -0.45142;] err=1.633781e-12
iteration 5 x=[1.09717; -0.45142;] err=2.482534e-16

ans =

    1.0972
   -0.4514

iteration 1 x=[-2.41667; -2.20833;] err=2.243819e-01
iteration 2 x=[-2.43056; -2.21528;] err=1.552825e-02
iteration 3 x=[-2.43050; -2.21525;] err=6.113484e-05
iteration 4 x=[-2.43050; -2.21525;] err=9.476218e-10
iteration 5 x=[-2.43050; -2.21525;] err=4.440892e-16

ans =

   -2.4305
   -2.2153
```

MATH-4800
Numerical Computing

Summer 2023
Schwendeman/Traldi

Exam 1 Solutions

Thursday, June 29, 2023

Name (print clearly in the box):

RIN (print clearly in the box):

Instructions:

1. Please answer all 5 problems with sufficient supporting work.
2. No electronic gizmos (laptops, calculators, cell phones, etc.) permitted.
3. No books or notes permitted, but one crib sheet is allowed.
4. Write out and sign the **Honor Pledge** below.

Problem (pts)	Score
1 (12pts)	
2 (12pts)	
3 (12pts)	
4 (12pts)	
5 (12pts)	
Total (60pts)	

Honor Pledge: I have neither given nor received any illegal aid on this exam.

1. [12pts] (a) Let $S(x) = \sqrt{6-2x}$. An approximation of $S(x)$ is given by

$$\tilde{S}(x) = \frac{5}{2} - \frac{x}{2}$$

Compute (i) the absolute forward error in the approximation when $x = 5/2$ and (ii) the relative backward error in the approximation when $x = 3/2$.

Solution

The absolute forward error is

$$|\tilde{S}(5/2) - S(5/2)| = |5/4 - 1| = 1/4$$

The relative backward error is $|\tilde{x} - x|/|x|$, where $S(\tilde{x}) = \tilde{S}(x)$. Have

$$\tilde{S}(3/2) = \frac{7}{4} = S(\tilde{x}) = \sqrt{6-2\tilde{x}} \qquad \Rightarrow \qquad \tilde{x} = 47/32$$

Thus,

$$\frac{|\tilde{x} - x|}{|x|} = \frac{|47/32 - 3/2|}{|3/2|} = \frac{1}{48}$$

1. (b) Let $Q(x,y) = x + \sqrt{x^2+y^2}$. The function Q is computed using finite precision arithmetic for the following (x,y) pairs: (i) $(10^{-5}, 10^5)$ and (ii) $(-10^5, 10^{-5})$. For which of the two pairs, if any or both, does the calculation of Q lead to a large loss of significant digits (i.e. more than 2 or 3 significant digits say)? Explain.

Solution

For the first pair, we have $x = 10^{-5}$ and $y = 10^5$. Since $x \ll y$, we have $\sqrt{x^2+y^2} \approx |y| = 10^5$ with no loss in significant figures. Thus,

$$Q \approx x + |y| \approx |y| = 10^5$$

with no loss of significant figures.

For the second pair, we have $x = -10^5$ and $y = 10^{-5}$ and now $|x| \gg y$. Thus, $\sqrt{x^2+y^2} \approx |x| = 10^5$ with no loss in significant figures. However,

$$Q \approx x + |x| \approx 0$$

with loss of significant figures with nearly equal numbers are subtracted.

2. [12pts] (a) Let

$$f(x) = \frac{x}{x+1}$$

and suppose $P(x)$ is the second-degree (quadratic) Taylor polynomial approximation of $f(x)$ about the point $x_0 = 1$. Use the remainder term to estimate $|P(1.1) - f(1.1)|$.

Solution

Use

$$f'(x) = \frac{1}{(x+1)^2}, \qquad f''(x) = -\frac{2}{(x+1)^3}, \qquad f'''(x) = \frac{6}{(x+1)^4}$$

and

$$f(1) = \frac{1}{2}, \qquad f'(1) = \frac{1}{4}, \qquad f''(1) = -\frac{1}{4}.$$

to get

$$P(x) = \frac{1}{2} + \frac{1}{4}(x-1) - \frac{1}{8}(x-1)^2$$

The remainder term is

$$R(x) = \frac{f'''(c)}{3!}(x-1)^3 = \frac{1}{(c+1)^4}(x-1)^3$$

where c is between x and 1. For $x = 1.1$, use $c = 1$ to maximize the remainder. Thus,

$$|P(1.1) - f(1.1)| \le \frac{1}{16}10^{-3}$$

2. (b) Let $f(x) = x^3 - x^2 - 3x + 1$. A root of f is bracketed on the interval $x \in [0, 1]$. Use **two** steps of the method of false position to determine an approximation for the root.

Solution

Note that $f(0) = 1$ and $f(1) = -2$. The line $L_1(x)$ through the points $(x, y) = (0, 1)$ and $(1, -2)$ is

$$L_1(x) = 1 + \frac{-2 - 1}{1 - 0} x = 1 - 3x$$

Set $L_1 = 0$ to give the first approximation $r_1 = 1/3$. Since $f(r_1) = -2/27$, the new bracket is $[0, 1/3]$. The line $L_2(x)$ through the points $(x, y) = (0, 1)$ and $(1/3, -2/27)$ is

$$L_2(x) = 1 + \frac{-2/27 - 1}{1/3 - 0} x = 1 - \frac{29}{9} x$$

The approximation given by two steps is r_2 , where $L_2(r_2) = 0$, which implies $r_2 = 9/27$.

3. [12pts] (a) Let

$$g(x) = 5 + \frac{2}{3}\sin x, \qquad I = [\pi, 2\pi]$$

Student X claims that the fixed-point iteration, $x_{n+1} = g(x_n)$, converges to the unique fixed point in I for *any* starting value x_0 in I. Use the Fixed-Point theorem to support this claim or state why the claim may not be true.

Solution

First show that g maps I into itself. Check endpoints

$$g(\pi) = g(2\pi) = 5 \in I$$

Also, check for local extrema

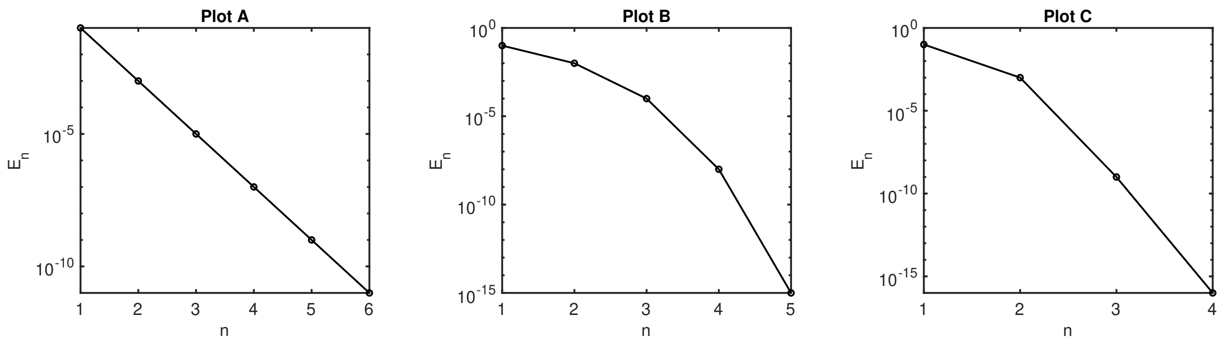
$$g'(x) = \frac{2}{3}\cos x = 0 \qquad \Rightarrow \qquad x = \frac{3\pi}{2} \in I$$

and

$$g(3\pi/2) = 5 - \frac{2}{3} \in I$$

Thus $g \in I$ whenever $x \in I$. Second, note that $|g'| \leq 2/3$ for all x including $x \in I$. Thus, according to the Fixed-Point Theorem, the iteration converges for all $x_0 \in I$. Student X is correct.

3. (b) A smooth function $F(x)$ has a simple root at $x = r$. Newton's method is used to compute the root using a starting value x_0 chosen close enough to r so that the iteration converges. Of the three plots below, which one best describes the behavior of the error, $E_n = |x_n - r|$, as a function of n. Explain.



Solution

Newton's method converges quadratically so that $E_{n+1} \approx CE_n^2$, where C is a constant. Consider the data...

For Plot A, the data appears to be

$$E_1 = 10^{-1}, \quad E_2 = 10^{-3}, \quad E_3 = 10^{-5}, \quad \ldots \qquad \Rightarrow \qquad E_{n+1} \approx \tfrac{1}{100} E_n \text{ , i.e. linear convergence}$$

For Plot B, the data appears to be

$$E_1 = 10^{-1}, \quad E_2 = 10^{-2}, \quad E_3 = 10^{-4}, \quad \ldots \qquad \Rightarrow \qquad E_{n+1} \approx E_n^2 \text{ , i.e. quadratic convergence}$$

Finally, the data for Plot C appears to be

$$E_1 = 10^{-1}, \quad E_2 = 10^{-3}, \quad E_3 = 10^{-9}, \quad \ldots \qquad \Rightarrow \qquad E_{n+1} \approx E_n^3 \text{ , i.e. cubic convergence}$$

The answer is Plot B.

4. [12pts] (a) Let

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

Use row elimination to find a lower triangular matrix L (with ones on the main diagonal) and an upper triangular matrix U such that $LU = A$.

Solution

Let

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \qquad \Rightarrow \qquad M_1 A = \begin{bmatrix} 2 & 1 & -1 \\ 0 & -1 & 4 \\ 0 & 2 & -5 \end{bmatrix}$$

and then

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \qquad \Rightarrow \qquad M_2 M_1 A = \begin{bmatrix} 2 & 1 & -1 \\ 0 & -1 & 4 \\ 0 & 0 & 3 \end{bmatrix}$$

So,

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

4. (b) Consider the following sequence of Matlab commands:

```
>> n=100;  A=rand(n,n);
>> x=ones(n,1);  b=A*x;
>> xc=myLinearSystemSolver(A,b);
>> norm(b-A*xc)/norm(b)

ans =

   4.2804e-16

>> delta=norm(xc-x)/norm(x);
```

Determine an estimate for the value of `delta` or explain why there is insufficient information given to provide an accurate estimate.

Solution

Since

$$\frac{\|\mathbf{b} - A\mathbf{x}_c\|}{\|\mathbf{b}\|} = O(\epsilon_{\text{mach}})$$

suggests that the linear solver is backwards stable. Thus,

$$\texttt{delta} = \frac{\|\mathbf{x}_c - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A)\epsilon_{\text{mach}}$$

Since the condition number $\kappa(A)$ is not given, there is insufficient information given to provide an accurate estimate.

5. [12pts] Let

$$A = \begin{bmatrix} \alpha & 2 & 0 \\ 2 & 4 & \beta \\ 0 & \beta & 5 \end{bmatrix}$$

where α and β are real constants.

(a) Determine all values of α such that A is symmetric, positive definite when $\beta = 0$.

(b) Find the Cholesky factorization of A when $\alpha = 3$ and $\beta = -2$.

Solutions

(a) Note that A is symmetric for all values of α and β. If $\beta = 0$, we have

$$\mathbf{x}^T A\mathbf{x} = \alpha x_1^2 + 4x_1x_2 + 4x_2^2 + 5x_3^2 = (\alpha - 1)x_1^2 + (x_1 + 2x_2)^2 + 5x_3^2 \geq 0 \qquad \text{if } \alpha \geq 1\,.$$

Also, $\mathbf{x}^T A\mathbf{x} = 0$ implies $x_3 = 0$, $x_1 + 2x_2 = 0$ and $(\alpha - 1)x_1^2 = 0$. If $\alpha > 1$, then $x_1 = 0$ which implies $x_2 = 0$. So, A is symmetric, positive definite if $\alpha > 1$.

(a) Let

$$L = \begin{bmatrix} \ell_{11} & 0 & 0 \\ \ell_{21} & \ell_{22} & 0 \\ 0 & \ell_{32} & \ell_{33} \end{bmatrix} \qquad \Rightarrow \qquad LL^T = \begin{bmatrix} \ell_{11}^2 & \ell_{11}\ell_{21} & 0 \\ \ell_{11}\ell_{21} & \ell_{22}^2 + \ell_{21}^2 & \ell_{22}\ell_{32} \\ 0 & \ell_{22}\ell_{32} & \ell_{33}^2 + \ell_{32}^2 \end{bmatrix}$$

Set $A = LL^T$ when $\alpha = 3$ and $\beta = -2$ gives

$$\ell_{11} = \sqrt{3}, \qquad \ell_{21} = 2/\sqrt{3}, \qquad \ell_{22} = 2\sqrt{2/3}, \qquad \ell_{32} = -\sqrt{3/2}, \qquad \ell_{33} = \sqrt{7/2}$$