Tutorial Letter 203/2/2012
Numerical Analysis II

APM3711

Semester 2

Department of Mathematical Sciences

IMPORTANT INFORMATION:

This tutorial letter contains important information about your module.

UNISA university of south africa

Learn without limits.

Dear Student,

This tutorial letter intends to help you prepare for the examination. It contains the following information:

1. Solutions to a previous examination paper - Semester 1.

We give here a few guidelines which may be helpful in your preparation for the Oct/Nov 2012 examination.
This semester's examination will again have 6 questions.

(a) Use your tutorial letter 101 and the course outline that was put on myUnisa to make sure that you know what does the syllabus cover.

(b) It is necessary to know the iteration formulae of the following methods.

- Fourth and fifth order Runge–Kutta
- Adams
- Milne
- The Chebyshev polynomials

(c) Make sure that you study the following methods.

1) The Taylor Method
2) The Runge–Kutta–Methods
3) The Milne & Adams–Moulton–Methods
4) The Euler Method
5) The Shooting–Method
6) The A.D.I.–Method
7) Chebyshev–series
8) Power Method

2. SOLUTIONS TO A PREVIOUS EXAMINATION PAPER

In the following, we will give first the question in the form it appeared in the examination paper, and then the solution to it. My comments appear in square brackets, in italics; they do not form a part of the solution.

QUESTION 1

(a) Use the second-order Taylor method to find $y(1.1)$ when y is the solution to the differential equation

$$\frac{dy}{dx} = xy + 1, \quad y(1) = 2.$$

(6)

(b) Euler's method is used to solve a first-order differential equation

$$\frac{dy}{dx} = f(y, x)$$

from $x = 0$ to $x = 4$.

(i) If the step length is $h = \Delta x = 0.1$, we get an error of 0.12 for $y(4)$. What step length should we use to reduce the error for $y(4)$ to 0.01?

(ii) How would this change in the step size influence the local error? (8)

(c) Explain briefly **why** the modified Euler method is more accurate than the Euler method. (2)

[16]

SOLUTION

(a) We have

$$\begin{aligned} y' &= \frac{dy}{dx} = f(x, y) = xy + 1, \qquad y(1) = 2, \\ y'' &= \frac{\partial}{\partial x} f(x, y) = y + xy' \end{aligned}$$

So $y'(1) = 1 \cdot 2 + 1 = 3$, $y''(1) = 2 + 1 \cdot 3 = 5$ and therefore the Taylor method gives

$$\begin{aligned} y(1.1) &\approx y(1) + (0.1)\, y'(1) + \frac{1}{2}(0.1)^2 y''(1) \\ &= 2 + (0.1) \cdot 3 + \frac{1}{2}(0.01) \cdot 5 \\ &= 2.325. \end{aligned}$$

(b) (i) With $h_1 = 0.1$, we got as global error of $y(4)$ the value $E_1 = 0.12$. What should h_2 be for us to get the global error $E_2 = 0.01$?
Since Euler's method is a first–order method, then the global error E when using a step size h is given by

$$|E| = K \cdot h.$$

This means that the equation

$$\frac{|E_1|}{|E_2|} = \frac{h_1}{h_2}$$

must hold for the errors E_1, E_2 when using two different step sizes h_1 and h_2. Thus, we should select h_2 as

$$h_2 = \left|\frac{E_2}{E_1}\right| \cdot h_1 = \frac{0.01}{0.12} \cdot (0.1) = 0.0083.$$

(ii) The local error in the Euler method is of the order $O\left(h^2\right)$, so since the new step size is

$$\left(\frac{0.12}{0.01}\right) = 12$$

times smaller, we expect the one–step error to be $(12)^2 = 144$ times smaller.

(c) Euler method uses only the slope at the initial point of the interval, whereas the modified Euler method attempts to use the slope values at the initial and final points.

[Note that it really is not enough here to reply "Because Euler method is first–order and modified Euler method is second–order method"!]

QUESTION 2

(a) (i) Why is a Runge-Kutta method usually preferred to a Taylor series method of the same order?

(ii) Is there an upper limit to the order of a Runge-Kutta method? Justify your answer! (4)

(b) Explain briefly how the accuracy of a Runge-Kutta method can be determined by:

(i) halving the step size at the end of each interval;

(ii) using two Runge-Kutta methods with different orders.

Which method is more efficient? Why? (8)

(c) The predictor and corrector formulas of the Adams-Moulton method are:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) + \frac{251}{720}h^5 y^{(5)}(\xi_1) \\ y_{n+1} &= y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) - \frac{19}{720}h^5 y^{(5)}(\xi_2). \end{aligned}$$

Apply the Adams-Moulton method to calculate the approximate value of $y(0.8)$ and $y(1.0)$ from the differential equation

$$y' = t + y$$

and the starting values

t	**y(t)**
0.0	0.95
0.2	0.68
0.4	0.55
0.6	0.30

Use 3 decimal digits with rounding at each step. (8)

[20]

SOLUTION

(a) (i) One does not have to find the derivatives. *[It is not true that the Runge–Kutta methods are more accurate.]*

(ii) No. An n–th order Runge–Kutta method is equivalent to an n–th order Taylor series method, and the value of n can be arbitrarily large in the Taylor series.

(b) (i) We can recompute the solution value at the end of the interval with two steps of half the step length, and compare the results.

(ii) We can do the calculations using two different methods (e.g. a 4th and 5th order one) and compare the end results.

Method (ii) is more efficient by since by selecting the methods suitably, we can use the same k–values in both methods, and therefore need less function evaluations.

(c) $y' = t + y = f(t.y)$: The given starting values are

$$\begin{array}{lllll} n = 1 & t_1 = 0.0 & y_1 = 0.95 & f_1 = 0.95 \\ n = 2 & t_2 = 0.2 & y_2 = 0.68 & f_2 = 0.88 \\ n = 3 & t_3 = 0.4 & y_3 = 0.55 & f_3 = 0.95 \\ n = 4 & t_4 = 0.6 & y_4 = 0.30 & f_4 = 0.9 \end{array}$$

At $t = 0.8$, the predictor is

$$\begin{aligned} y_5 &= y_4 + \frac{0.2}{24}(55f_4 - 59f_3 + 37f_2 - 9f_1) \\ &= 0.3 + \frac{0.2}{24}(55(0.9) - 59(0.95) + 37(0.88) - 9(0.95)) \\ &= 0.4455, \qquad f_5 = 1.2455 \end{aligned}$$

and the corrector is

$$\begin{aligned} y_5 &= y_4 + \frac{h}{24}(9f_5 + 19f_4 - 5f_3 + f_2) \\ &= y_4 + \frac{0.2}{24}(9(1.2455) + 19(0.9) - 5(0.95) + 0.88) \\ &= 0.5037, \qquad f_5 = 1.3037. \end{aligned}$$

At $t = 1.0$, the predictor is:

$$\begin{aligned} y_6 &= y_5 + \frac{0.2}{24}(55f_5 - 59f_4 + 37f_3 - 9f_2) \\ &= 0.5037 + \frac{0.2}{24}(55(1.3037) - 59(0.9) + 37(0.95) - 9(0.881)) \\ &= 0.8856, \qquad f_6 = 1.8856 \end{aligned}$$

and the corrector is:

$$\begin{aligned} y_6 &= y_5 + \frac{h}{24}(9f_6 + 19f_5 - 5f_4 + f_3) \\ y_6 &= 0.5037 + \frac{0.2}{24}\left(9\left(1.8856\right) + 19\left(1.3037\right) - 5\left(0.9\right) + 0.95\right) \\ &= 0.822. \end{aligned}$$

QUESTION 3

(a) Consider the boundary value problem

$$\begin{aligned} y'' &= y + x \\ & \quad y(1) = 1, \;\; y(3) = 2. \end{aligned}$$

Solve this system over the interval $1 \leq x \leq 3$ by using the shooting method. Use the Euler method with $\Delta x = 1$. (8)

(b) Compare briefly the shooting method and the method of finite differences in solving boundary value problems, by stating the main advantages and the main disadvantages of each method.(4)

[12]

SOLUTION

(a) The boundary value problem is

$$\begin{aligned} y'' &= y + x \\ y\left(1\right) &= 1, \; y\left(3\right) = 2 \end{aligned}$$

We must first convert the second–order differential equation into a system of two first–order equations, by introducing a new variable, $z = y'$. This gives us the system

$$\begin{cases} y' = z & y\left(1\right) = 1 \quad y\left(3\right) = 2 \\ z' = y + x & z\left(1\right) = ? \end{cases} \tag{1}$$

[The reason for this conversion is that the Euler method is a method for solving first–order differential equation, or systems of first–order differential equations. It can ***not*** *be used directly to solve a second–order differential equation!]*

To apply the Euler method, we need the value of $z\left(1\right)$. Since this value is not known, we will make a guess at its value and see how close the obtained $y\left(3\right)$–value is to the required value $D = 2$.

Let us take as the first guess $G1 = 0$ and solve the system (1) with $z(1) = G1 = 0$. The calculations are:

$$\begin{aligned} y(1) &= 1 \\ z(1) &= 0 \\ \\ y(2) &= y(1) + 1 \cdot z(1) = 1 + 0 = 1 \\ z(2) &= z(1) + 1 \cdot (y(1) + 1) = 0 + (1 + 1) = 2 \\ \\ y(3) &= y(2) + 1 \cdot z(2) = 1 + 2 = 3 \end{aligned}$$

[Note the order of calculations: we need to find both $y(2)$ and $z(2)$ before we can calculate $y(3)$! We don't need to find $z(3)$.]

The value obtained for $y(3)$ was $R_1 = 3$, which is not equal to the required value $D = 2$. Therefore, we must make another guess; let us take $G2 = -1$. With $z(1) = G2 = -1$, the calculations are

$$\begin{aligned} y(1) &= 1 \\ z(1) &= -1 \\ \\ y(2) &= 1 - 1 = 0 \\ z(2) &= -1 + (1 + 1) = 1 \\ \\ y(3) &= 0 + 1 = 1 \end{aligned}$$

Here we obtained a value $R_2 = 1$ for $y(3)$, but again it is not equal to $D = 2$, and therefore we need another guess. This time, since we have available for us the results of two guesses we use the technique of interpolation to make a better guess. The interpolation formula is

$$\frac{G3 - G1}{D - R1} = \frac{G2 - G1}{R2 - R1}$$

where $G3$ is the next guess; we therefore get

$$G3 = G1 + \frac{G2 - G1}{R2 - R1}(D - R1)$$

so the value of the next guess is

$$G3 = 0 + \frac{-1 - 0}{1 - 3}(2 - 3) = -0.5.$$

With $z(1) = G3 = -0.5$, the calculations are

$$\begin{aligned} y(1) &= 1 \\ z(1) &= -0.5 \\ \\ y(2) &= 1 - 0.5 = 0.5 \\ z(2) &= -0.5 + (1+1) = 1.5 \\ \\ y(3) &= 0.5 + 1.5 = 2 \end{aligned}$$

We see that the third guess leads to the correct value of $y(3)$.

[The differential equation here is linear, so we know that only one interpolation is needed! For a non–linear equation, we might need to keep interpolating and making new guesses until $y(3)$ is within a given tolerance of the correct value 2.]

The approximate solution to the boundary value problem is therefore

$$y(1) = 1, \quad y(2) = 0.5, \quad y(3) = 2.$$

[Typical mistakes made were:

- *Using the modified Euler method rather than the Euler method.*
- *Using the Euler method incorrectly, for instance by trying to apply it directly to the equation $y'' = y + x$.*
- *Trying to use direct mathematical methods to change the second–order equation $y'' = y + x$ into a first–order one, for instance by trying to integrate the right hand side. A word of warning about this kind of approach: Don't do it! This is a numerical analysis course, and none of the methods require us to do mathematical manipulation of the equations to solve them halfway before applying the numerical methods. (The only exception of a sort is the Taylor method where you may need to do some differentiation to find the higher derivatives.) So when you see a second or higher order differential equation, rather convert it to a first–order one by the standard method of introducing new variables for the derivatives of the original variable. This standard approach has the benefit that it works every time and does not depend on the differential equation being particularly nice so that direct mathematical methods can be applied. It is true that sometimes mathematical tricks can be used to simplify the calculations in the numerical method, but you should only attempt this if you are absolutely sure that you know what you are doing.]*

(b)

	shooting method	finite differences method
+	certain to give a solution, even in the non–linear case	less computational effort (especially for higher–order systems)
–	more computational effor t	may not converge (if the equation is non–linear, and the initial estimate is not good)

QUESTION 4

(a) Explain why the numerical solution of a characteristic value problem of the type

$$u'' + f(x)u' + k^2 u = 0, \quad u(0) = 0, \quad u(1) = 0$$

by the method of finite differences, leads to an eigenvalue problem. Here, $f(x)$ is an arbitrary function of x. (5)

(b) The matrix

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

has an eigenvalue near -2.5 and another one near $+1.5$.

(i) Use the power method to find the eigenvalue near -2.5. Do three iterations, starting with the vector $(1, 0)^\top$.

(ii) Explain how you can use suitable shifting with the power method to find the eigenvalue near -2.5 with fewer iterations, **without** having to find the inverse of any matrix. (9)

[14]

SOLUTION

(a) Characteristic value problem:

$$u'' + f(x)\, u' + k^2 u = 0, \qquad u(0) = 0, \qquad u(1) = 0.$$

[*This problem has the trivial solution $u \equiv 0$. For certain values of w, called characteristic values, it may also have non–trivial solutions. "Solving" the characteristic value problem consists of finding these characteristic values, and usually finding the corresponding solutions as well.*]

Let us assume that the grid points are t_i and let u_i be the approximation for $u(t_i)$. Substituting finite difference approximations for the derivatives u'' and u' leads to a linear set of equations for the u_i. This set of equations can be written in the matrix equation form

$$\left(A - bk^2 I\right) u = 0$$

for some matrix A and a real number $b \in \mathbb{R}$, where u is the vector of the u_i–values.

[Here the difference equation will be

$$\left(1 + \frac{h}{2} f(x_i)\right) u_{i+1} + \left(h^2 k^2 - 2\right) u_i + \left(1 - \frac{h}{2} f(x_i)\right) u'_{i-1} = 0$$

so in fact $b = h^2$.]

Finding the values of k^2 such that a non–trivial solution vector u exists is therefore equivalent to the problem of finding the eigenvalues of A, with the corresponding eigenvectors giving the non–trivial approximate solutions.

[It is **not** true that we are "using the eigenvalues to solve the equations"! Please do also use mathematically correct language: It is incorrect to say that "we solve the matrix" or that "the set of equations can be written as a matrix"!]

(b) Matrix

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

with one eigenvalue near -2.5, another one near $+1.5$.

(i) The eigenvalue near -2.5 has the largest magnitude, so we can find it by applying the power method to the matrix A.

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

$$\begin{pmatrix} -2 & 2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -\frac{1}{2} \end{pmatrix} = \begin{pmatrix} -3 \\ \frac{1}{2} \end{pmatrix} = -3\begin{pmatrix} 1 \\ -\frac{1}{6} \end{pmatrix}$$

$$\begin{pmatrix} -2 & 2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -\frac{1}{6} \end{pmatrix} = \begin{pmatrix} -\frac{7}{3} \\ \frac{5}{6} \end{pmatrix} = -\frac{7}{3}\begin{pmatrix} 1 \\ \frac{5}{14} \end{pmatrix}.$$

The eigenvalue is approximately $\lambda = -\frac{7}{3}$.
[I have done the calculations with fractions, but you are welcome to do the calculations with decimal numbers instead. The end result with decimals is $\lambda = -2.33$.]

(ii) In shifting, we consider the matrix $B = A - sI$ where s is a suitably chosen number. Then then eigenvalues of B are $\lambda_1 - s$ and $\lambda_2 - s$ if λ_1 and λ_2 were the eigenvalues of matrix A. The power method requires fewer iterations if the magnitude of the largest eigenvalue is not too close to the magnitude of the second–largest one. The values $|2.5|$ and $|1.5|$ are fairly close to each other. To accelerate convergence, we could for instance shift by the value $s = 2$. The eigenvalues of B would then be

$$-2.5 - 2 = -4.5$$

and

$$1.5 - 2 = -0.5,$$

and in this case the larger one is much larger than the smaller one and therefore convergence will be faster.

QUESTION 5

(a) (i) Write down the finite difference equation corresponding to the three-dimensional partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} + u + xyz = 0$$

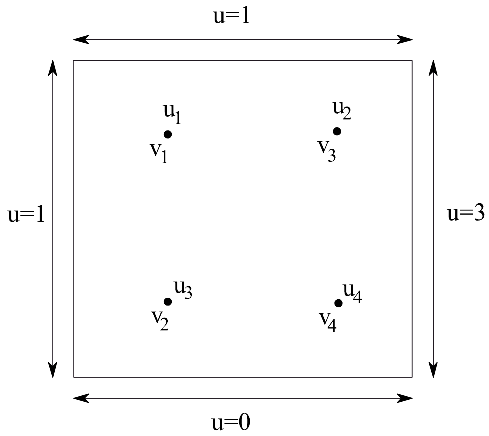
at the grid point (x_i, y_j, z_k) of a rectangular mesh with $\Delta x = \Delta y = \Delta z = h$. Denote $u(x_i, y_j, z_k)$ by $u_{i,j,k}$. (6)

(ii) Re-write the difference equation in (i) into an iteration formula appropriate for Liebmann's method. Use the superscripts u^k, u^{k+1} to show how new values are computed from previous ones. (7)

(b) The equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

is to be solved in a rectangular region. The grid (with $\Delta x = \Delta y = 1$), the boundary values and the numbering of the node points are indicated in the sketch below.



(i) Set up the two sets of equations for solving the values of u at the four node points by the ADI method. Use the notation of the sketch, with function values denoted by u_i in rowwise traverse and by v_i in columnwise traverse.

(ii) Explain how you would proceed to solve the problem by using the ADI method. (14)

[21]

SOLUTION

(a) (i)

$$\frac{u_{i+1,j,k} - 2u_{i,j.k} + u_{i-1,j,k}}{h^2} + \frac{u_{i,j+1,k} - 2u_{i,j,k} + u_{i,j-1,k}}{h^2} + \frac{u_{i,j,k+1} - 2u_{i,j,k} + u_{i,j,k-1}}{h^2} + x_i y_j z_k + u_{ijk} = 0$$

[Remember to replace x, y and z also by x_i, y_j and *z_k! The difference equation always approximates the differential equation at some grid point (x_i, y_j, z_k), so all terms must be consistently evaluated at that same point!]*

(ii)

$$u^{(k+1)}_{i,j,k} = \frac{1}{6-h^2}\left(u^{(k)}_{i+1,j,k} + u^{(k)}_{i-1,j,k} + u^{(k)}_{i,j+1,k} + u^{(k)}_{i,j-1,k} + u^{(k)}_{i,j,k+1} + u^{(k)}_{i,j,k-1} + h^2 x_i y_j z_k\right)$$

or

$$u^{(k+1)}_{i,j,k} = \frac{1}{6}\left(u^{(k)}_{i+1,j,k} + u^{(k)}_{i-1,j,k} + u^{(k)}_{i,j+1,k} + u^{(k)}_{i,j-1,k} + u^{(k)}_{i,j,k+1} + u^{(k)}_{i,j,k-1} + h^2 x_i y_j z_k + h^2 u_{i,j,k}\right)$$

[This is the Gauss–Seidel type approach; the alternative is to use a Jacobi method where on the right hand side, updated information is used when necessary, so that for instance we might use $u^{(k+1)}_{i-1,j,k}$, $u^{(k+1)}_{i,j-1,k}$ and $u^{(k+1)}_{i,j,k-1}$ instead of $u^{(k)}_{i-1,j,k}$, $u^{(k)}_{i,j-1,k}$ and $u^{(k)}_{i,j,k-1}$. Either one will give you full marks.]

(b) (i) Row–wise: We apply the equation

$$u_L - 2u_0 + u_R = -v_A + 2v_0 - v_B$$

at the nodes u_1, u_2, u_3, u_4 :

$$\begin{cases} 1 - 2u_1 + u_2 = -1 + 2v_1 - v_2 \\ u_1 - 2u_2 + 3 = -1 + 2v_3 - v_4 \\ 1 - 2u_3 + u_4 = -v_1 + 2v_2 = 0 \\ u_3 - 2u_4 + 3 = -v_3 + 2v_4 - 0 \end{cases}$$

$$\therefore \begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 0 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 2v_1 - v_2 - 2 \\ 2v_3 - v_4 - 4 \\ -v_1 + 2v_2 - 1 \\ -v_3 + 2v_4 - 3 \end{pmatrix} \tag{*}$$

Column–wise: We apply the equation

$$v_A - 2v_0 + v_B = -u_L + 2u_0 - u_R$$

at the nodes v_1, v_2, v_3, v_4 :

$$\begin{cases} 1 - 2v_1 + v_2 = -1 + 2u_1 - u_2 \\ v_1 - 2v_2 + 0 = -1 + 2u_3 - u_4 \\ 1 - 2v_3 + v_4 = -u_1 + 2u_2 - 3 \\ v_3 - 2u_4 + 0 = -u_3 + 2u_4 - 3 \end{cases}$$

$$\therefore \begin{pmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 0 & 0 \\ 0 & 0 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 2u_1 - u_2 - 2 \\ 2u_3 - u_4 - 1 \\ -u_1 + 2u_2 - 4 \\ -u_3 + 2u_4 - 3 \end{pmatrix} \tag{**}$$

(ii) – Start with a first guess at $v = v^0$ (the vector $(v_1^0, v_2^0, v_3^0, v_4^0)$)
– Repeat:

1) solve u^i from (*) using v^{i-1} on right–hand side
2) solve v^i from (**) using u^i on right–hand side

until the change $v^i - v^{i-1}$ is less than some given tolerance.
[Note that in (b)(i), you must write down the equation for this particular problem! It is not enough to just state the general equations.]

QUESTION 6

(a) The function e^x is to be approximated by a fifth-order polynomial over the interval $[-1, +1]$. Why is a Chebyshev series a better choice than a Taylor (or Maclaurin) expansion? (4)

(b) Given the power series

$$f(x) = 1 - x - 2x^3 - 4x^4$$

and the Chebyshev polynomials

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 + 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1, \end{aligned}$$

economize the power series $f(x)$ twice. (6)

(c) Find the Padé approximation $R3(x)$, with numerator of degree 2 and denominator of degree 1, to the function $f(x) = x^2 + x^3$. (7)

[17]

SOLUTION

(a) The maximum error over the whole interval is smaller. The Taylor series has zero error at $x = 0$, but the error can be quite large at $x = \pm 1$.

(b)

$$f(x) = 1 - x - 2x^3 - 4x^4$$

Economize once: we add/subtract T_4 suitably scaled, such that the x^4–term disappears. Here, we must add $\frac{1}{2}T_4 = 4x^4 - 4x^2 + \frac{1}{2}$, and will get the economized series

$$\begin{aligned} f^*(x) &= f(x) + \frac{1}{2}T_4 \\ &= \frac{3}{2} - x - 4x^2 - 2x^3. \end{aligned}$$

Economize again: Add T_3 suitably scaled, such that the x^3–term disappears. We must add

$$\frac{1}{2}T_3(x) = 2x^3 + \frac{3}{2}x,$$

and will get

$$\begin{aligned} f^{**}(x) &= f^*(x) + \frac{1}{2}T_3(x) \\ &= \frac{3}{2} - \frac{5}{2}x - 4x^2. \end{aligned}$$

(c) The Padé approximation is

$$R(x) = \frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x}.$$

This gives

$$R_3(x) - f(x) = \frac{a_0 + a_1 x + a_2 x^2}{1 + b_1 x} - \left(x^2 + x^3\right)$$

which simplifies to:

$$\begin{aligned} & \left(a_0 + a_1 x + a_2 x^2\right) - (1 - b_1 x)\left(x^2 + x^3\right) \\ = \; & a_0 + a_1 x + (a_2 - 1)x^2 - (1 + b_1)x^3 - b_1 x^4. \end{aligned}$$

To find the four unknown values a_0, a_1, a_3, b_1, we must set the coefficients of x^0, x^1, x^2, x^3 in the numerator to zero. This give us the equation

$$\begin{cases} a_0 = 0 \\ a_1 = 0 \\ a_2 - 1 = 0 \\ 1 + b_1 = 0 \end{cases} \quad \therefore \quad \begin{cases} a_0 = 0 \\ a_1 = 0 \\ a_2 = 1 \\ b_1 = -1 \end{cases}$$

Hence the Padé approximation is

$$R_3\left(x\right) = \frac{x^2}{1-x}.$$

[**Remarks:**

- *Note that we should not also set the coefficient of* x^4. *to zero! To find the four coefficients, we need 4 equations. To add one more equation would either be unnecessary or else lead to a contradiction.*
- *Please make sure that you can write down the general form of a Padé approximation* $R_n\left(x\right)$ *with the numerator and/or the denominator of a given degree! Remember that if the numerator (above the line) is of degree* n *(i.e. a polynomial of degree* n*) and if the denominator (below the line) is of degree* m*, then the Padé approximation is denoted by* R_N *where* $N = m + n$*. The constant term of the denominator (the "*b_0*" term) is always taken to be equal to* 1*, to ensure that* $R_N\left(0\right)$ *is well defined. The number of constants to be determined is then equal to* $N + 1$ *in the Padé approximation* $R_N\left(x\right)$. *To uniquely determine those* $N + 1$ *constants, we need* $N + 1$ *equations which are obtained by setting the coefficients of the powers of* x*, of orders* 0 *to* N*, to be equal to zero in the numerator of* $f\left(x\right) - R_N\left(x\right)$.]