MT414: Numerical Analysis

Homework 1

Answers

1. Let \( f(x) = xe^{-2} \).

(a) Find the fourth Taylor polynomial \( P_4(x) \) for \( f(x) \) about \( x_0 = 0 \).
(b) Find an upper bound for \( |f(x) - P_4(x)| \) for \( x \in [0, 0.4] \).
(c) Approximate \( \int_0^{0.4} f(x) \, dx \) using \( \int_0^{0.4} P_4(x) \, dx \).
(d) Find an upper bound for the error in the computation in part (c) by using your answer to part (b).
(e) Approximate \( f'(0.2) \) by computing \( P_4'(0.2) \). Use the correct answer for \( f'(0.2) \) (to 5 decimal places) to compute the relative error in your computation.

**Answer:**

(a) We have

\[
f(x) = xe^{-2}
\]

\[
f'(x) = e^{-2} + 2xe^{-2} = (1 + 2x^2)e^{-2}
\]

\[
f''(x) = 2xe^{-2} + (1 + 2x^2)(2x)e^{-2} = (6x + 4x^3)e^{-2}
\]

\[
f^{(3)}(x) = (6 + 12x^2)e^{-2} + (6x + 4x^3)(2x)e^{-2} = (6 + 24x^2 + 8x^4)e^{-2}
\]

\[
f^{(4)}(x) = (48x + 32x^3)e^{-2} + (6 + 24x^2 + 8x^4)(2x)e^{-2} = (60x + 80x^3 + 16x^5)e^{-2}
\]

\[
f^{(5)}(x) = (60 + 240x + 80x^3 + 32x^5 + 480x^5)(2x)e^{-2} = (60 + 360x^4 + 120x^5)e^{-2}
\]

We have \( f(0) = 0, \ f'(0) = 1, \ f''(0) = 0, \ f^{(3)}(0) = 6 \) and \( f^{(4)}(0) = 0 \). This means that \( P_4(x) = x + x^3 \).

(b) We know that \( f(x) = P_4(x) + E_4(x) \), where \( E_4(x) = f^{(5)}(\xi)x^5/120 \), where \( 0 \leq \xi \leq x \). Because \( x \in [0, 0.4] \), the largest possible value of \( E_4(x) \) is given by \( f^{(5)}(0.4)(0.4)^5/120 \). We can also say that \( f^{(5)}(0.4) \leq 124e^{0.42} \leq 146 \), using 12-digit arithmetic, so \( f^{(5)}(\xi) \leq 146 \). Therefore, \( |f(x) - P_4(x)| = |E_4(x)| \leq 146 \cdot 0.4^5/120 \leq 0.0125 \).

(c) We can approximate \( \int_0^{0.4} f(x) \, dx \) by

\[
\int_0^{0.4} P_4(x) \, dx = \int_0^{0.4} (x + x^3) \, dx = \frac{x^2}{2} + \frac{x^4}{4}\bigg|_0^{0.4} = \frac{0.4^2}{2} + \frac{0.4^4}{4} = 0.0864.
\]

(d) The absolute error in this computation is

\[
\left| \int_0^{0.4} f(x) \, dx - \int_0^{0.4} P_4(x) \, dx \right| \leq \int_0^{0.4} |f(x) - P_4(x)| \, dx
\]

\[
= \int_0^{0.4} |E_4(x)| \, dx \leq \int_0^{0.4} 0.0125 \, dx = 0.0125 \cdot 0.4 = 0.005
\]

However, the relative error is possibly as large as \( \frac{0.005}{0.0864} \leq 0.07 \), so our answer is not quite correct to 2 decimal places.

(e) We have \( P_4'(x) = 1 + 3x^2 \), so \( P_4'(0.2) = 1.12 \). We can compute that \( f'(0.2) = 1.1241 \), so the relative error is 0.0036.

2. Use the Intermediate Value Theorem and Rolle’s Theorem to show that the equation \( x^3 + 2x + k = 0 \) has exactly one real solution, regardless of the value of the constant \( k \).

**Answer:** Let \( f(x) = x^3 + 2x + k \). Suppose that there are two unequal numbers \( a \) and \( b \) so that \( f(a) = f(b) = 0 \). Rolle’s Theorem then says that there is a value \( c \) between \( a \) and \( b \) so that \( f'(c) = 0 \). However, \( f'(c) = 3c^2 + 2 \), and the equation \( 3c^2 + 2 = 0 \) has no solutions for any value of \( c \).
Therefore, there is at most one real solution. How can we be sure that there is at least one solution? Here’s an argument that is probably much too detailed: If \(|k| \leq 1\), then \(f(2) \geq 0\) and \(f(-2) \leq 0\), so the Intermediate Value Theorem can be applied to deduce that there must be a solution. If \(|k| > 1\), then \(f(|k|) > 0\), and \(f(-|k|) < 0\), so we again can apply the Intermediate Value Theorem.

3. Perform the following calculations

\(\text{(i)}\) exactly,
\(\text{(ii)}\) using three-digit chopping arithmetic, and
\(\text{(iii)}\) using three-digit rounding arithmetic.

\(\text{(iv)}\) Compute the relative errors in parts \((\text{ii})\) and \((\text{iii})\).

\[\begin{array}{cccc}
(a) & \frac{4}{5} + \frac{1}{3} & (b) & \frac{4}{5} \cdot \frac{1}{3} \\
(c) & \left(\frac{1}{3} - \frac{3}{11}\right) + \frac{3}{20} & (d) & \left(\frac{1}{3} + \frac{3}{11}\right) - \frac{3}{20}
\end{array}\]

**Answer:** We have

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>17 \frac{17}{15}</td>
<td>4 \frac{4}{15}</td>
<td>139 \frac{139}{15}</td>
<td>301 \frac{301}{15}</td>
</tr>
<tr>
<td>3-digit chopping</td>
<td>1.13</td>
<td>0.266</td>
<td>0.211</td>
<td>0.455</td>
</tr>
<tr>
<td>Relative error</td>
<td>0.003</td>
<td>0.0025</td>
<td>0.002</td>
<td>0.00233</td>
</tr>
<tr>
<td>3-digit rounding</td>
<td>1.13</td>
<td>0.266</td>
<td>0.21</td>
<td>0.456</td>
</tr>
<tr>
<td>Relative error</td>
<td>0.003</td>
<td>0.0025</td>
<td>0.0029</td>
<td>0.000133</td>
</tr>
</tbody>
</table>

4. Suppose that two points \((x_0, y_0)\) and \((x_1, y_1)\) are on a straight line with \(y_1 \neq y_0\). Two formulas are available to compute the \(x\)-intercept of the line:

\[x = \frac{x_0 y_1 - x_1 y_0}{y_1 - y_0}\quad\text{and}\quad x = x_0 - \frac{(x_1 - x_0)y_0}{y_1 - y_0}.

\(\text{(a)}\) Show that both formulas are algebraically correct.

\(\text{(b)}\) Suppose that \((x_0, y_0) = (1.31, 3.24)\) and \((x_1, y_1) = (1.93, 4.76)\). Use three-digit rounding arithmetic to compute the \(x\)-intercept using both of the formulas. Which method is better and why?

**Answer:** (a) We should be a bit careful here to avoid dividing by 0. It is potentially unsafe to write that the equation of the line is \(y - y_0 = \frac{y_1 - y_0}{x_1 - x_0} x - x_0\), because potentially \(x_0 = x_1\). However, we are told that \(y_0 \neq y_1\),

so we can instead write the equation of the line as \(x - x_0 = \frac{x_1 - x_0}{y_1 - y_0} y - y_0\). We can cross-multiply and write this instead as \(x - x_0 = y - y_0 \left(\frac{x_1 - x_0}{y_1 - y_0}\right)\).

The \(x\)-intercept is the point on the line at which \(y = 0\), so we can substitute \(y = 0\) into this equation and get \(x - x_0 = (-y_0) \left(\frac{x_1 - x_0}{y_1 - y_0}\right)\), or \(x = x_0 - \frac{(x_1 - x_0)y_0}{y_1 - y_0}\), which is the given formula.

Now we can simplify:

\[x = x_0 - \frac{(x_1 - x_0)y_0}{y_1 - y_0} = \frac{x_0(y_1 - y_0)}{y_1 - y_0} - \frac{(x_1 - x_0)y_0}{y_1 - y_0} = \frac{x_0y_1 - x_1y_0}{y_1 - y_0}.
\]
(b) The first formula gives the answer \(-0.00658\), while the second formula gives the answer \(-0.0100\). In this case, the second formula is better. The first one involved subtracting \(x_0y_1 - x_1y_0\). Because \(x_0y_1 = 6.24\) and \(x_1y_0 = 6.25\), the result of the subtraction has only one significant digit.

We can check this by working to 10 significant digits. In that case, the first formula gives \(-0.0115789474\) and the second gives \(-0.0115789470\). Surely the answer is closer to \(-0.01\) than to \(-0.00658\).

5. The Taylor polynomial of degree \(n\) for \(f(x) = e^x\) is \(\sum_{i=0}^{n} \frac{x^i}{i!}\). Use the Taylor polynomial of degree 9 and three-digit chopping arithmetic to find an approximation to \(e^{-5}\) using each of the following methods:

\[
(a)\quad e^{-5} \approx \sum_{i=0}^{9} \frac{(-5)^i}{i!} = \sum_{i=0}^{9} \frac{(-1)^i5^i}{i!}.
\]

\[
(b)\quad e^{-5} = \frac{1}{e^5} \approx \frac{1}{\sum_{i=0}^{9} \frac{5^i}{i!}}.
\]

(c) Use your calculator to approximate \(e^{-5}\) to 8 places. Which formula, \((a)\) or \((b)\), gave the most accuracy, and why?

**Answer:** We have

<table>
<thead>
<tr>
<th>(i)</th>
<th>(i!)</th>
<th>(5^i)</th>
<th>(\frac{5^i}{i!})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5.00</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>25</td>
<td>12.5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>125</td>
<td>20.8</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>625</td>
<td>26.0</td>
</tr>
<tr>
<td>5</td>
<td>120</td>
<td>3120</td>
<td>26.0</td>
</tr>
<tr>
<td>6</td>
<td>720</td>
<td>15600</td>
<td>21.6</td>
</tr>
<tr>
<td>7</td>
<td>5040</td>
<td>78000</td>
<td>15.4</td>
</tr>
<tr>
<td>8</td>
<td>40300</td>
<td>390000</td>
<td>9.67</td>
</tr>
<tr>
<td>9</td>
<td>362000</td>
<td>1950000</td>
<td>5.38</td>
</tr>
</tbody>
</table>

The formula in part \((a)\) gives \(1.00 - 5.00 + 12.5 - 20.8 + 26.0 - 26.0 + 21.6 - 15.4 + 9.67 - 5.38 = -1.81\). This is obviously incorrect, because \(e^{-5} > 0\). The formula in part \((b)\) gives \(1/(1.00 + 5.00 + 12.5 + 20.8 + 26.0 + 26.0 + 21.6 + 15.4 + 9.67 + 5.38) = 1/141 = 0.00709\). To 8 places, \(e^{-5} \approx 0.0067379470\).

The formula in part \((b)\) becomes a bit better if we add the numbers in increasing order, to avoid losing precision. We have \(1/(1.00 + 5.00 + 5.38 + 9.67 + 12.5 + 15.4 + 20.8 + 21.6 + 26.0 + 26.0) = 1/143 = 0.00699\).

The difficulty with the formula in part \((a)\) is that it can involve subtracting nearly equal numbers, resulting in a loss of precision.

6. Suppose that \(fl(y)\) is a \(k\)-digit rounding approximation to \(y\). Show that

\[
\left| \frac{y - fl(y)}{y} \right| \leq 0.5 \times 10^{-k+1}.
\]

**Answer:** Suppose that \(y = 0.d_1d_2d_3 \ldots \times 10^n\). If \(d_{k+1} < 5\), then \(fl(y) = 0.d_1d_2 \ldots d_k \times 10^n\), and therefore

\[
\left| \frac{y - fl(y)}{y} \right| = \left| 0.0 \ldots 0d_{k+1} \ldots \right| < 5 \times 10^{-k-1} \times 0.1 = 5 \times 10^{-k} = 0.5 \times 10^{-k+1}.
\]

If \(d_{k+1} \geq 5\), then \(fl(y) = (0.d_1d_2 \ldots d_k \times 10^{-k}) \times 10^n\), and then

\[
\left| \frac{y - fl(y)}{y} \right| = \left| 0.0 \ldots 0d_{k+1} - 10^{-k} \right| \leq 5 \times 10^{-k-1} \times 0.1 = 0.5 \times 10^{-k+1}.
\]