Problem 1 Let $I$ denote the closed interval $I = [-1, 1]$. A continuous function $f : I \to \mathbb{R}$ has been carefully sampled on 75 equidistant points spread across $I$ and the results are given in the table below.

<table>
<thead>
<tr>
<th>n</th>
<th>x(n)</th>
<th>f(x(n))</th>
<th>n</th>
<th>x(n)</th>
<th>f(x(n))</th>
<th>n</th>
<th>x(n)</th>
<th>f(x(n))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.0000</td>
<td>-1.0290</td>
<td>26</td>
<td>-0.3243</td>
<td>-0.0295</td>
<td>51</td>
<td>0.3514</td>
<td>0.0439</td>
</tr>
<tr>
<td>2</td>
<td>-0.9730</td>
<td>-1.0095</td>
<td>27</td>
<td>-0.2973</td>
<td>0.0032</td>
<td>52</td>
<td>0.3784</td>
<td>0.0199</td>
</tr>
<tr>
<td>3</td>
<td>-0.9459</td>
<td>-0.9855</td>
<td>28</td>
<td>-0.2703</td>
<td>0.0336</td>
<td>53</td>
<td>0.4054</td>
<td>-0.0050</td>
</tr>
<tr>
<td>4</td>
<td>-0.9189</td>
<td>-0.9577</td>
<td>29</td>
<td>-0.2432</td>
<td>0.0616</td>
<td>54</td>
<td>0.4324</td>
<td>-0.0306</td>
</tr>
<tr>
<td>5</td>
<td>-0.8919</td>
<td>-0.9262</td>
<td>30</td>
<td>-0.2162</td>
<td>0.0872</td>
<td>55</td>
<td>0.4595</td>
<td>-0.0567</td>
</tr>
<tr>
<td>6</td>
<td>-0.8649</td>
<td>-0.8916</td>
<td>31</td>
<td>-0.1892</td>
<td>0.1103</td>
<td>56</td>
<td>0.4865</td>
<td>-0.0829</td>
</tr>
<tr>
<td>7</td>
<td>-0.8378</td>
<td>-0.8543</td>
<td>32</td>
<td>-0.1622</td>
<td>0.1308</td>
<td>57</td>
<td>0.5135</td>
<td>-0.1090</td>
</tr>
<tr>
<td>8</td>
<td>-0.8108</td>
<td>-0.8145</td>
<td>33</td>
<td>-0.1351</td>
<td>0.1487</td>
<td>58</td>
<td>0.5405</td>
<td>-0.1349</td>
</tr>
<tr>
<td>9</td>
<td>-0.7838</td>
<td>-0.7727</td>
<td>34</td>
<td>-0.1081</td>
<td>0.1640</td>
<td>59</td>
<td>0.5676</td>
<td>-0.1600</td>
</tr>
<tr>
<td>10</td>
<td>-0.7568</td>
<td>-0.7291</td>
<td>35</td>
<td>-0.0811</td>
<td>0.1765</td>
<td>60</td>
<td>0.5946</td>
<td>-0.1843</td>
</tr>
<tr>
<td>11</td>
<td>-0.7297</td>
<td>-0.6842</td>
<td>36</td>
<td>-0.0541</td>
<td>0.1863</td>
<td>61</td>
<td>0.6216</td>
<td>-0.2073</td>
</tr>
<tr>
<td>12</td>
<td>-0.7027</td>
<td>-0.6382</td>
<td>37</td>
<td>-0.0270</td>
<td>0.1935</td>
<td>62</td>
<td>0.6486</td>
<td>-0.2288</td>
</tr>
<tr>
<td>13</td>
<td>-0.6757</td>
<td>-0.5915</td>
<td>38</td>
<td>0.0000</td>
<td>0.1980</td>
<td>63</td>
<td>0.6757</td>
<td>-0.2483</td>
</tr>
<tr>
<td>14</td>
<td>-0.6486</td>
<td>-0.5443</td>
<td>39</td>
<td>0.0270</td>
<td>0.1998</td>
<td>64</td>
<td>0.7027</td>
<td>-0.2656</td>
</tr>
<tr>
<td>15</td>
<td>-0.6216</td>
<td>-0.4969</td>
<td>40</td>
<td>0.0541</td>
<td>0.1991</td>
<td>65</td>
<td>0.7297</td>
<td>-0.2803</td>
</tr>
<tr>
<td>16</td>
<td>-0.5946</td>
<td>-0.4496</td>
<td>41</td>
<td>0.0811</td>
<td>0.1958</td>
<td>66</td>
<td>0.7568</td>
<td>-0.2920</td>
</tr>
<tr>
<td>17</td>
<td>-0.5676</td>
<td>-0.4025</td>
<td>42</td>
<td>0.1081</td>
<td>0.1900</td>
<td>67</td>
<td>0.7838</td>
<td>-0.3044</td>
</tr>
<tr>
<td>18</td>
<td>-0.5405</td>
<td>-0.3561</td>
<td>43</td>
<td>0.1351</td>
<td>0.1818</td>
<td>68</td>
<td>0.8108</td>
<td>-0.3049</td>
</tr>
<tr>
<td>19</td>
<td>-0.5135</td>
<td>-0.3105</td>
<td>44</td>
<td>0.1622</td>
<td>0.1713</td>
<td>69</td>
<td>0.8378</td>
<td>-0.3054</td>
</tr>
<tr>
<td>20</td>
<td>-0.4865</td>
<td>-0.2658</td>
<td>45</td>
<td>0.1892</td>
<td>0.1587</td>
<td>70</td>
<td>0.8649</td>
<td>-0.3011</td>
</tr>
<tr>
<td>21</td>
<td>-0.4595</td>
<td>-0.2224</td>
<td>46</td>
<td>0.2162</td>
<td>0.1439</td>
<td>71</td>
<td>0.8919</td>
<td>-0.2918</td>
</tr>
<tr>
<td>22</td>
<td>-0.4324</td>
<td>-0.1804</td>
<td>47</td>
<td>0.2432</td>
<td>0.1272</td>
<td>72</td>
<td>0.9189</td>
<td>-0.2771</td>
</tr>
<tr>
<td>23</td>
<td>-0.4054</td>
<td>-0.1399</td>
<td>48</td>
<td>0.2703</td>
<td>0.1087</td>
<td>73</td>
<td>0.9459</td>
<td>-0.2563</td>
</tr>
<tr>
<td>24</td>
<td>-0.3784</td>
<td>-0.1012</td>
<td>49</td>
<td>0.2973</td>
<td>0.0885</td>
<td>74</td>
<td>0.9730</td>
<td>-0.2291</td>
</tr>
<tr>
<td>25</td>
<td>-0.3514</td>
<td>-0.0643</td>
<td>50</td>
<td>0.3243</td>
<td>0.0669</td>
<td>75</td>
<td>1.0000</td>
<td>-0.1950</td>
</tr>
</tbody>
</table>

1. (5 points) Show that the function $f$ has at least two zeros in $I$.

Solution By assumption, the function $f$ is continuous. Therefore, there is at least one zero between each pair of points $x$ and $y$ where $f(x)f(y) < 0$. Therefore there is at least one zero $z_1$ in the interval $(x_{26}, x_{27})$ and at least one zero $z_2$ in the interval $(x_{52}, x_{53})$.

2. (10 points) Compute each of the zeros with a relative error than $\tau = 0.05$.

Solution Let $c_1 = \frac{x_{26} + x_{27}}{2}$ and let $c_2 = \frac{x_{52} + x_{53}}{2}$ be the midpoints of the two intervals bracketing the roots. Then the absolute errors are

$$|z_i - c_i| \leq \frac{1}{2} \frac{2}{74} = \frac{1}{74}, \quad i = 1, 2.$$
simply because the interval $I$ has length 2 and has been cut into 74 subintervals of equal length. In order to estimate the relative errors we proceed as follows

$$\frac{|z_1 - c_1|}{|z_1|} \leq \frac{1/74}{|x_{27}|} = \frac{1/74}{|x_{27}|} = \frac{1/74}{|x_{27}|} = \frac{1}{74 - 52} = \frac{1}{22} < \frac{1}{20} = 0.05,$$

because $c_1 < x_{27} < 0$ and

$$\frac{|z_2 - c_2|}{|z_2|} \leq \frac{1/74}{|x_{52}|} = \frac{1/74}{|x_{52}|} = \frac{1/74}{|x_{52}|} = \frac{1}{102 - 74} = \frac{1}{28} < \frac{1}{20} = 0.05,$$

because $0 < x_{52} < z_2.$

3. (10 points) It is known that the function $f$ is twice differentiable and

$$\forall x \in [-1, 1] : |f''(x)| \leq 10.5.$$

Compute the value of $f(0.72)$ with an absolute error less than $\nu = 0.05$.

**Solution** By inspection we find that $x_{64} < 0.72 < x_{65}$. Let $p$ be denote the polynomial which interpolates $f$ at these two points. Let $x \in I$. Since $f$ is twice differentiable with a continuous second derivative, there exists a $\xi$ such that

$$f(x) - p(x) = \frac{f^{(2)}(\xi)}{2} \omega(x)$$

where $\omega(x) = (x - x_{64})(x - x_{65})$. In particular, we have

$$\omega(0.72) = (0.72 - 0.7027)(0.72 - 0.7297) = -0.00016781 = -1.6781 \times 10^{-4}$$

It follows that

$$|f(x) - p(x)| \leq \frac{10.5}{2} \cdot 1.6781 \times 10^{-4} = 8.3905 \times 10^{-4} \ll 0.05$$

It remains to evaluate $p(x)$ at $x = 0.72$. In general, we have

$$p(x) = \frac{x - x_{64}}{x_{65} - x_{64}} f(x_{65}) + \frac{x - x_{65}}{x_{64} - x_{65}} f(x_{64}).$$

In particular, we have

$$p(0.72) = 37[(0.72 - 0.7027)(-0.2656) + (0.7297 - 0.72)(-0.2803)]$$

$$= -0.27061023 = -270.6123 \times 10^{-4} \approx 0.27$$
**Problem 2** The integral $\int_0^1 f(x)dx$ of a function $f : [0,1] \to \mathbb{R}$ has been computed numerically using Simpson's rule and many different stepsizes $h = \frac{1}{N}$. The results along with some auxiliary values are given below. It is known that $f$ is infinitely often differentiable.

<table>
<thead>
<tr>
<th>N</th>
<th>Sh</th>
<th>$(Sh-S2h)$</th>
<th>$(S2h-S4h)/(Sh-S2h)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>524288</td>
<td>2.45837007000238e-01</td>
<td>0.0000e+00</td>
<td>Inf</td>
</tr>
<tr>
<td>262144</td>
<td>2.45837007000238e-01</td>
<td>1.2212e-15</td>
<td>4.545455e-01</td>
</tr>
<tr>
<td>131072</td>
<td>2.45837007000237e-01</td>
<td>5.5511e-16</td>
<td>-3.050000e+00</td>
</tr>
<tr>
<td>65536</td>
<td>2.45837007000236e-01</td>
<td>-1.6931e-15</td>
<td>-5.245902e-01</td>
</tr>
<tr>
<td>32768</td>
<td>2.45837007000238e-01</td>
<td>8.8818e-16</td>
<td>-7.500000e-01</td>
</tr>
<tr>
<td>16384</td>
<td>2.45837007000237e-01</td>
<td>-6.6613e-16</td>
<td>4.166667e-02</td>
</tr>
<tr>
<td>8192</td>
<td>2.45837007000238e-01</td>
<td>-2.7756e-17</td>
<td>-9.000000e+00</td>
</tr>
<tr>
<td>4096</td>
<td>2.45837007000238e-01</td>
<td>2.4980e-16</td>
<td>6.666667e-01</td>
</tr>
<tr>
<td>2048</td>
<td>2.45837007000237e-01</td>
<td>1.6653e-16</td>
<td>2.933333e+01</td>
</tr>
<tr>
<td>1024</td>
<td>2.45837007000237e-01</td>
<td>4.8850e-15</td>
<td>1.526136e+01</td>
</tr>
<tr>
<td>512</td>
<td>2.45837007000232e-01</td>
<td>7.4551e-14</td>
<td>1.599442e+01</td>
</tr>
<tr>
<td>256</td>
<td>2.45837007000158e-01</td>
<td>1.1924e-12</td>
<td>1.600126e+01</td>
</tr>
<tr>
<td>128</td>
<td>2.45837006998965e-01</td>
<td>1.9080e-11</td>
<td>1.600174e+01</td>
</tr>
<tr>
<td>64</td>
<td>2.45837006979885e-01</td>
<td>3.0531e-10</td>
<td>1.600662e+01</td>
</tr>
<tr>
<td>32</td>
<td>2.45837006674572e-01</td>
<td>4.8870e-09</td>
<td>1.602645e+01</td>
</tr>
<tr>
<td>16</td>
<td>2.45837001787536e-01</td>
<td>7.8322e-08</td>
<td>1.610547e+01</td>
</tr>
<tr>
<td>8</td>
<td>2.45836923465701e-01</td>
<td>1.2614e-06</td>
<td>1.641672e+01</td>
</tr>
<tr>
<td>4</td>
<td>2.45835662055614e-01</td>
<td>2.0708e-05</td>
<td>1.758315e+01</td>
</tr>
<tr>
<td>2</td>
<td>2.45814953836298e-01</td>
<td>3.6412e-04</td>
<td>0.000000e+00</td>
</tr>
<tr>
<td>1</td>
<td>2.45450838083890e-01</td>
<td>0.0000e+00</td>
<td>0.000000e+00</td>
</tr>
</tbody>
</table>

1. (5pt) Explain why the computed value of the fraction $\frac{S_{2h}-S_{4h}}{S_h-S_{2h}}$ will always deviate dramatically from the real value as $h$ tends to zero.

**Solution** By assumption the function $f$ is infinitely differentiable, so in exact arithmetic we have

$$S_h, S_{2h}, S_{4h} \to \int_0^1 f(x)dx, \quad h \to 0_+.$$  

As a result, the we will experience catastrophic cancellation when calculating $S_h - S_{2h}$ and $S_{2h} - S_{4h}$. The computed fraction is therefore the ratio of two numbers which have been calculated with large relative errors. Therefore, it is extremely unlikely that we will get a value which is close to the correct one which is 16 in this particular case.

2. (10pt) Determine the range of $N$ where the *computed* value of the fraction

$$\frac{S_{2h} - S_{4h}}{S_h - S_{2h}}$$

is close to the correct value.
displays the same behavior as if it had been computed in exact arithmetic.

**Solution** There are two possible behaviors. The fractions must converge monotonically to 16 either from below or from above. In our case we see monotone convergence down to 16 for the values $N = 4, 8, \ldots, 256$. At $N = 512$ the fraction is still close to 16, but it has jumped to the other side of 16 indicating that the rounding errors are starting to make their presence felt.

3. (10pt) Find the smallest value of $N$ for which you are certain the relative error is less than $\tau = 10^{-11}$.

**Solution** We are handed the values $S_h - S_{2h}$. It is clear that the integral is greater than 0.24583. Therefore a good relative error estimate is given by

$$\frac{\left| \int_0^1 f(x)dx - S_h \right|}{\int_0^1 f(x)dx} \leq \frac{|S_h - S_{2h}|}{0.24583 \cdot 15}$$

Scanning the table from bottom to top, we see that for $N \leq 64$ the relative error estimate is too large, but for $N = 128$ the relative error estimate is smaller than $\tau$. Since we are well inside the range of $N$ where the computed fraction is exhibiting monotone convergence to 16 we can trust the error estimates. Hence $N = 128$ is the smallest value which is acceptable.
Problem 3 Consider the function $f : [2, \infty) \to \mathbb{R}$ given by

$$f(x) = \sqrt{x+1} - \sqrt{x-1}.$$ 

The following MATLAB commands have been used to generate a plot of the graph of $f$ for $x \in [2^{20}, 2^{21}]$:

```matlab
>> f=@(x)sqrt(x+1)-sqrt(x-1);
>> x=single(linspace(2^20,2^21,1025));
>> plot(log2(x),log2(f(x)))
```

The plot is presented in Figure 1.

![Figure 1: An inferior plot of log₂(f(x)) as a function of log₂(x).](image)

1. (5pt) List as many differences between this MATLAB plot and the true graph of $f$ as you can.

Solution The real function $f$ is clearly differentiable with

$$f'(x) = \frac{1}{2}(x+1)^{-\frac{1}{2}} - \frac{1}{2}(x-1)^{-\frac{1}{2}} < 0$$

Hence the function is monotone decreasing and there are no solutions of the equation $f'(x) = 0$. 

5
Regardless, the plot shows a function which highly oscillatory on some intervals and constant on other intervals.

2. (10pt) Show that the condition number of \( f \) is given by

\[
\kappa_f(x) = \frac{x}{2\sqrt{x + 1}\sqrt{x - 1}}
\]

and explain why it is at least not theoretically impossible to compute \( f \) with a relative error which is less than the unit roundoff error \( u \).

**Solution** By definition, the condition number of a function \( f \) at a point \( x \neq 0 \) where \( f(x) \neq 0 \) is given by

\[
\kappa_f(x) = \left| \frac{x f'(x)}{f(x)} \right|
\]

We therefore manipulate the expression for \( f'(x)/f(x) \). We have

\[
2 \frac{f'(x)}{f(x)} = \frac{\frac{1}{\sqrt{x+1}} - \frac{1}{\sqrt{x-1}}}{\sqrt{x+1} - \sqrt{x-1}}
\]

With an eye on the target, we decide to compute

\[
2 \frac{f'(x)}{f(x)} \sqrt{x - 1}\sqrt{x + 1} = \frac{\sqrt{x - 1} - \sqrt{x + 1}}{\sqrt{x + 1} - \sqrt{x - 1}} = -1.
\]

It follows immediately that

\[
\frac{xf'(x)}{f(x)} = -\frac{x}{2\sqrt{x - 1}\sqrt{x + 1}} < 0
\]

and

\[
\kappa_f(x) = \frac{x}{2\sqrt{x - 1}\sqrt{x + 1}} = \frac{x}{2\sqrt{x^2 - 1}}, \quad x \geq 2.
\]

We observe that this is a monotone decreasing function, because

\[
\frac{d}{dx} \kappa_f(x) = \frac{2\sqrt{x^2 - 1} - 2x \frac{2x}{\sqrt{x^2 - 1}}}{(2(x^2 - 1))^2} = 2\sqrt{x^2 - 1} - 1 - 4 \frac{x^2}{4(x^2 - 1)} < 0, \quad x \geq 2
\]

Therefore, the condition number is bounded by the value at \( x = 2 \),

\[
\kappa_f(x) \leq \frac{2}{2\sqrt{3}} = \frac{1}{\sqrt{3}} < 1
\]

It follows, that a perfect routine which does no rounding errors during the calculation can compute \( f(x) \), where \( x \) is a real number in the representable range, with a relative error which is less than the unit roundoff \( u \).
3. (10pt) Find a reliable way of computing $f$ in MATLAB.

**Solution** We must find a way to avoid the catastrophic cancellation which occurs for large values of $x$. We have

$$f(x) = \sqrt{x+1} - \sqrt{x-1} = \left(\sqrt{x+1} - \sqrt{x-1}\right) \left(\frac{\sqrt{x+1} + \sqrt{x-1}}{\sqrt{x+1} + \sqrt{x-1}}\right)$$

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

The final expression, i.e.

$$f(x) = \frac{2}{\sqrt{x+1} + \sqrt{x-1}}$$

is mathematically equivalent to the original, but there is no catastrophic cancellation as $x$ tends to infinity. We notice that the expression $x - 1$ can not cancel catastropically either as we have assumed that $2 \leq x$. 


Problem 4 This problem centers on the rapid calculation of reciprocal values on a binary computer with no hardware division. Let $\alpha \neq 0$ be a nonzero machine number. Our goal is to compute the reciprocal value $\frac{1}{\alpha}$.

1. (5 points) Explain, how we can easily compute reciprocal values for all non-zero machine numbers, if we can handle all positive machine numbers in the interval $[1,2)$.

Solution Any floating nonzero floating point number $\alpha$ can be written in the form $\alpha = (-1)^s(1.f)_2 \times 2^m$. The reciprocal value is given by

$$\frac{1}{\alpha} = (-1)^s \frac{1}{(1.f)_2} \times 2^{-m}.$$ 

The real problem is therefore the computation of the reciprocal value of $\alpha = (1.f)_2 \in [1,2)$.

2. (10 points) Find a function $g : \mathbb{R} \to \mathbb{R}$ such that the fixpoint iteration given by

$$x_0 \in \mathbb{R}, \quad \text{and} \quad x_{n+1} = g(x_n), \quad n = 0,1,2\ldots$$

satisfies

$$1 - \alpha x_{n+1} = (1 - \alpha x_n)^3.$$ 

Moreover, it must be possible to evaluate $g$ without doing any divisions.

Solution We desire a relation of the type

$$1 - \alpha x_{n+1} = (1 - \alpha x_n)^3.$$ 

Therefore we expand the right hand side in order to obtain

$$1 - \alpha x_{n+1} = (1 - \alpha x_n)^3 = 1 - \alpha x_n + \alpha^2 x_n^2 - \alpha^3 x_n^3.$$ 

It follows that we should pick

$$x_{n+1} = x_n - \alpha x_n^2 + \alpha^2 x_n^3$$

which corresponds to the choice of

$$g(x) = x - \alpha x^2 + \alpha^2 x^3.$$ 

We notice that this function $g$ can be evaluated without doing any divisions.

3. (10 points) Let $\alpha \in [1,2)$. Show that if $x_0$ is chosen such that

$$0 < x_0 < \frac{2}{\alpha}$$

then not only is the sequence $\{x_n\}_{n=0}^{\infty}$ convergent, but

$$x_n \to \frac{1}{\alpha}, \quad n \to \infty, \quad n \in \mathbb{N}.$$ 

**Solution** We observe that

\[
0 < x_0 < \frac{2}{a} \iff 0 < ax_0 < 2 \iff -1 < 1 - ax_0 < 1 \iff |1 - ax_0| < 1.
\]

Moreover, since

\[
1 - ax_{n+1} = (1 - ax_n)^3
\]

for all \(n\), we have

\[
|1 - ax_n| = |1 - ax_0|^{3^n}.
\]

It follows that the sequence \(\{y_n\}_{n=0}^{\infty}\) given by

\[
y_n = 1 - ax_n
\]

is convergent with limit 0, whenever \(0 < x_0 < \frac{2}{a}\). Since

\[
x_n = \frac{1 - y_n}{a}
\]

it follows that the sequence \(\{x_n\}_{n=0}^{\infty}\) is convergent with limit \(\frac{1}{a}\).